METHODS OF CORRELATION ANALYSIS

BY

MORDECAI EZEKIEL

Economic Adviser to the Secretary of Agriculture
Fellow of the American Statistical Association
Fellow of the Econometric Society

SECOND EDITION

NEW YORK
JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED

SECOND EDITION
Fourth Printing, December, 1947

PRINTED IN U. S. A.

PREFACE TO SECOND EDITION

Twice since the first edition of *Methods of Correlation Analysis* appeared there have been reprintings in which minor errors in computations or typography were corrected. Now, a decade after the publication of the first edition, I am making the first general revision.

There have been many refinements and developments in the application of correlation methods to social and economic data during this period, and a beginning has been made in their application to engineering and other technological problems. The general technique has been but little changed during the period, and the main body of methods still seems useful. The major changes during the decade have been, first, in the interpretation of the meaning of standard errors and, second, in the application of logical limitations to the flexibility of graphic curves. Other significant developments have been in the perfection of new and speedier methods of calculation and in the development of methods of estimating the reliability of an individual estimate or forecast. All these are covered in this revision.

One completely new chapter has been added to this edition. That is Chapter 19, dealing with the reliability of an individual forecast and also with the applicability of error formulas to time series. The conclusion is reached there that these formulas are more serviceable in connection with time series than has generally been believed. Chapter 16, dealing with the short-cut (Bean) method of graphic correlation, has been almost entirely rewritten and materially enlarged. Increased emphasis is placed upon the precautions which need to be taken to get dependable results by this method and upon the way in which logical analysis should be used to place limitations upon the shape of the curves fitted, and thus prevent undue flexibility in their fitting. The chapters dealing with sampling theory, Chapter 2 for means and Chapter 18 for correlation results, have been materially revised to bring the explanation of the significance of standard error computations up to the modern interpretation. The section on the sampling significance of graphic regression curves has been moved from the technical appendix to this section and has also been materially expanded, with fuller illustrations. After a decade of use, it is now believed that this technique provides a valuable check on the significance of graphic regression and net regression curves.

Other chapters have been less extensively revised. Chapter 23, on examples of correlation applications, has been briefly brought up to date. One time-series analysis has been extrapolated to date in Chapter 14. A new explanatory example, which it is believed will aid the student in comprehending the meaning of partial regression coefficients, has been added at the beginning of Chapter 10; and Chapter 11 has been expanded somewhat. Although the analysis of variance is introduced here, no attempt is made to provide a complete treatment for it, as it was felt to lie outside the major field of this book. Chapters 7, 13, and 15, dealing with the measurement of standard error of estimate and degree of correlation, have also been revised to state more precisely the meaning of the adjustment of the crude coefficients to obtain unbiased estimates of the probable value in the universe. Other chapters have been corrected or expanded in various details. The appendix on methods of computation has been expanded to cover the most expeditious methods of computing partial correlation coefficients, the standard error of an individual forecast, and of making graphic transfers in the graphic short-cut method; and the explanations on the charts in Appendix 3 have been modified in line with the changes in Chapters 2 and 18.

With respect to the perennial debate as between the use of elaborate mathematical curves or transformations or the use of freehand curves in representing curvilinear regressions, my basic position remains unchanged in favoring freehand curves unless there are logical reasons for the selection of a particular mathematical equation. Much more attention is given to the logical meaning of freehand curves, however, and to the use of logical limitations in drawing in the curves. As before, the techniques for both methods are described and illustrated. The cross-referencing from one method to the other, and the discussion of the proper place for each, has also been somewhat expanded.

To aid instructors and others who may wish to use this revised edition along with the old, the table numbers have been left unchanged throughout the body of the book, new tables being designated by an A or B after the number. Figure numbers similarly are left unchanged up to Chapter 16, where the considerable number of new figures added made it seem better to begin renumbering. Equation numbers have been left unchanged throughout most of the body of the book, equations being renumbered only from Chapter 21 on. Prior to that point, equations numbered with whole numbers stand exactly as in the first edition; when the previous equations were changed or new equations were added, they are numbered with decimal fractions.

I hope that with these changes and additions the book will prove more useful than heretofore for classroom purposes and individual study. Naturally I am grateful that so specialized a book as this has found so wide an application in teaching and research, and I am always interested in hearing of applications of these methods to new fields.

During recent years I have had to devote myself primarily to matters of economic policy and have not been able to follow the developments in statistical methods as closely as during the period when this book was first taking shape. In preparing this revision I have had to lean heavily on the advice of those who in recent years have been closer to statistical teaching and practice than I have been myself. Valuable suggestions as to desirable revisions and new content have been received from Frederick V. Waugh, Charles F. Sarle, Elmer J. Working, Louis H. Bean, O. C. Stine, and Clarence M. Purves. I am indebted to my first teacher, Howard R. Tolley, for many suggestions noted during the period he was using the book for classroom teaching at the University of California. In addition, much of the revision, especially in the more mathematical sections, has been guided by the advice of two expert mathematical statisticians, W. Edwards Deming and Meyer A. Girshick. I am deeply indebted to them both for helpful suggestions and criticisms and for reading much of the revised manuscript, especially the sections dealing with the sampling significance of results. The increased precision and clarity of these sections are largely attributable to their aid. R. G. Hainsworth has again helped me with the figures, maintaining consistency with the excellence of those he prepared for the first edition. Any errors or misstatements remain my own responsibility, and not that of those who have aided with suggestions or criticisms.

To these and to many others who, over the years, have called my attention to errors or suggested revisions I express my appreciation and gratitude.

Although the new material has been carefully checked, some errors of computation or notation have no doubt crept in. Again I shall be grateful if any student or reader will inform me of any such errors he notices.

Mordecai Ezekiel

Washington, D. C.
June 15, 1941

PREFACE TO FIRST EDITION

This book is not intended to cover the entire field of statistics, but rather, as its name indicates, that part of the field which is concerned with studying the relations between variables. The first two chapters are devoted to a brief review of the central elements in the measurement of variability in a statistical series, and to the essential concepts in judging the reliability of conclusions. These chapters are not to be regarded as a full statement, but instead as brief summaries to clarify the basic ideas which are involved in the subsequent development.

No attempt is made in the body of the text to present the mathematical theory on which the art of statistical analysis is based. Instead, the aim throughout has been to show how the various methods may be employed in practical research work, what their limitations are, and what the results really mean. Only the simplest of algebraic statements have been employed, and the practical procedure for each operation has been worked out step by step. It is believed that the material will be readily comprehensible to anyone who has had courses in elementary algebra.

Although the examples which are used in presenting the several methods are drawn very largely from the author's own field of agricultural economics, the methods themselves are explained in sufficiently general terms so that they can be applied in any field. In addition, two chapters are devoted to a discussion of the types of problems in a great many different fields of work to which correlation analysis has been successfully applied, and to research methods and the place of correlation analysis in research. It is hoped that this presentation will assist research workers in many fields to appreciate both the possibilities and the limitations of correlation analysis, and so gain from their data knowledge of all the relations which so frequently lie hidden beneath the surface.

Where the methods presented are the well-established ones developed by the fathers of the modern science, mainly the English statisticians, no attempt is made to prove or derive the various formulas. On a few crucial points, however, or where derivations not generally accessible are involved, the derivations of the formulas are shown in notes in the technical appendix, in the simplest manner possible.

The methods presented in this book, insofar as they constitute an advance over those previously available, represent largely the joint product of a group of young researchers in the Bureau of Agricultural Economics of the United States Department of Agriculture during the past decade. The new methods include (a) the application of the Doolittle method to the solution of multiple correlation problems, greatly reducing the labor of obtaining multiple correlation results, and making feasible the use of multiple correlation in actual research work; (b) the development of approximate methods for determining curvilinear multiple correlations, and, more recently, very rapid graphic methods for their determination; (c) the recognition of "joint" correlation, and the gradual development of methods of treating it; and (d) by extensive use in actual investigations, concrete demonstration of the possibilities of these methods in research work. These recent developments in correlation analysis are as yet largely unavailable except in the original articles in technical journals. One object of this book is to present them in organized form, and with such interpretation that their significance and application may be fully understood.

During the last two decades, the English statisticians "Student" and R. A. Fisher have been developing more exact methods of judging the reliability of conclusions, particularly where those conclusions involve correlation or are based on small samples. These new methods have as yet received but little recognition from American statisticians. They are presented here as simply as possible, and the discussion of the reliability of conclusions gives them full consideration.

So many persons have helped in the years during which this book has been growing that it is difficult for me to enumerate them all. First of all I should like to mention Howard R. Tolley, from whom I received my introduction to statistics, and with whom it has been a constant joy to work. I give him credit for much that is included here. The very order of presentation reflects that which he worked out for his classes. In a very real sense this book is a product of the spirit of research with which the Bureau of Agricultural Economics was imbued by the broad vision of Henry C. Taylor. John D. Black was the first to point out some of the undeveloped phases of statistical analysis, and then aided with encouragement and counsel in their solution. Bradford B. Smith aided in the beginning of the new developments, and his vivid imagination and logical mind have been a constant help. Among others who have collaborated in various stages, or who have independently worked out various phases of the problem, may be mentioned Sewall Wright, Donald Bruce, Fred Waugh, Louis Bean, and Andrew Court. Susie White, Helen L. Lee, and Della E. Merrick have given intelligent, conscientious, and loyal assistance in the clerical work in the development and testing of each new step.

In the preparation of the book itself I have had generous and willing help. Dorothea Kittredge and Bruce Mudgett have given the very substantial assistance of a detailed reading of the entire text, and many improvements in presentation and in material are due to their suggestions. For two terms the mimeographed manuscript has been used as a text in the United States Department of Agriculture Graduate School, and the members of the class have helped me in working out the illustrations, in clarifying the text, and in eliminating errors. R. G. Hainsworth, who prepared the figures, deserves credit for the excellence of the graphic illustrations. O. V. Wells helped in computing many of the illustrative problems, and Corrine F. Kyle in verifying the arithmetic. For the laborious and exacting work of typing the preliminary stencils, the many revisions, and the final manuscript, and for her care, patience, and suggestions, I am indebted to my mother, Rachel Brill Ezekiel; and for editing the manuscript and helping in the lengthy task of proofreading, to my wife, Lucille Finsterwald Ezekiel.

To all these, and to the many others who have helped me in the development of this work, I take this opportunity of expressing my obligation and my gratitude.

For any errors in the statements made and in the theories advanced, I alone am of course responsible. Although the text has been checked painstakingly, it is hardly to be hoped that a publication of this character will appear without some errors creeping in, in mathematics, in arithmetic, or in spelling. When such errors, or any ambiguities of statement, are noted by any reader, I would be very grateful if he would inform me of them.

MORDECAI EZEKIEL.

WASHINGTON, D. C.,
April 20, 1930.

CONTENTS

CHAPTER 1

PAGE

MEASURING THE VARIABILITY OF A STATISTICAL SERIES 1

The Arithmetic Average 2
Frequency Tables 4
Average Deviation 6
Standard Deviation 8

CHAPTER 2

JUDGING THE RELIABILITY OF STATISTICAL RESULTS 14

Assumptions in Sampling 15
Computing the Standard Error 19
Reliability of Small Samples 22
Meaning and Use of the Standard Error 25
Universes, Past and Present 30

CHAPTER 3

THE RELATION BETWEEN TWO VARIABLES, AND THE IDEA OF FUNCTION 34

Relations between Variables 34
Graphic Representation of Relation between Two Variables 36
Expressing a Functional Relation Mathematically 39
Determining a Functional Relation Statistically 42

CHAPTER 4

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN ANOTHER CHANGES: (1) BY THE USE OF AVERAGES 47

Independent and Dependent Variables 50
Reliability of Group Averages 54
Range within Which True Relation May Fall 56

CHAPTER 5

PAGE

DETERMINING THE WAY ONE VARIABLE CHANGES WITH ANOTHER: (2) ACCORDING TO THE STRAIGHT-LINE FUNCTION 59

The Equation of a Straight Line 59
Fitting the Equation by Least Squares 64
Interpreting the Linear Equation 71

CHAPTER 6

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN ANOTHER CHANGES: (3) FOR CURVILINEAR FUNCTIONS ... 75

Different Types of Equations 76
Fitting a Simple Parabola 83
Fitting a Cubic Parabola 89
Fitting a Logarithmic Curve 93
Expressing a Curvilinear Relation by a Free-hand Curve 105
The Logical Significance of Mathematical Functions 113
A Mathematical Equation Used in an Economic Problem 121
Limitations in Estimating One Variable from Known Values of Another 125

CHAPTER 7

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION 128

The Closeness of Estimate—Standard Error of Estimate 128
For Linear Relations 129
For Curvilinear Relations 131
Adjustment of Standard Error of Estimate for Number of Observations 133
The Relative Importance of the Relation—Correlation 136
Linear—Coefficient of Correlation 137
Curvilinear—Index of Correlation 138
Adjustments for Number of Observations 141

CHAPTER 8

PRACTICAL METHODS FOR WORKING TWO-VARIABLE CORRELATION PROBLEMS 146

Terms to be Used 146
Working Out a Linear Correlation 147
Interpreting the Results 151
Working Out a Curvilinear Correlation 152
Interpreting the Results 157

CHAPTER 9

PAGE

THREE MEASURES OF CORRELATION: THE MEANING AND USE FOR EACH 159

CHAPTER 10

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE VARIABLES CHANGE: (1) BY SUCCESSIVE ELIMINATION 163

Theoretical Example 164
Practical Example 169
Eliminating the Approximate Influence of One Variable 172
Eliminating the Approximate Influence of Both Variables 174
Correcting Results by Successive Eliminations 176

CHAPTER 11

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (2) BY CROSS-CLASSIFICATION AND AVERAGES 181

Cross-classification for Three Variables 181
Differences between Matched Sub-groups 185
Limitations of Cross-classification for Many Variables 186

CHAPTER 12

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (3) BY USING A LINEAR REGRESSION EQUATION 190

Determining a Regression Equation for Two Independent Variables 191
Determining a Regression Equation for Three Independent Variables 198
Determining the Regression Equation for any Number of Independent Variables 203
Interpreting the Multiple Regression Equation 205

CHAPTER 13

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION FOR LINEAR MULTIPLE CORRELATION 208

Standard Error of Estimate 208
Multiple Correlation 210
Measuring the Separate Effect of Individual Variables 213
Partial Correlation 213
"Beta" Coefficients 217

CHAPTER 14

PAGE

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (4) USING CURVILINEAR REGRESSIONS 220

Multiple Regression Curves Mathematically Determined 221
Multiple Regression Curves by Successive Approximations 222
Determining the First Approximation Net Regression Curves 228
Estimating X_1 from the First Approximation Curves 235
Determining the Second Approximation Regression Curves 239
Estimating X_1 from the Second Approximation Curves 243
Correcting the Curves by Further Successive Approximations 247
Stating the Final Conclusions 247
Limitations on the Use of the Results 254
A Test in Actual Forecasting of Yield 255
Reliability of Regression Curves 258

CHAPTER 15

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION FOR CURVILINEAR MULTIPLE CORRELATION ... 259

Standard Error of Estimate 259
Index of Multiple Correlation 264
Measuring the Net Curvilinear Importance of Individual Factors 267

CHAPTER 16

SHORT-CUT METHODS OF DETERMINING NET REGRESSION LINES AND CURVES 268

Linear Net Regressions 269
The Short-cut Method Applied to Curvilinear Regressions 277
Identifying Joint Relations by the Short-cut Process 296
Application of the Short-cut Method to Large Samples 298

CHAPTER 17

MEASURING THE WAY A DEPENDENT VARIABLE CHANGES WITH CHANGES IN A NON-QUANTITATIVE INDEPENDENT FACTOR 302

Eliminating the Influence of Other Variables 302
Determining the Net Influence of the New Variable 305
Making Further Successive Approximations 308

CHAPTER 18

PAGE

DETERMINING THE RELIABILITY OF CORRELATION CONCLUSIONS 312

Simple Correlation 312
- Regression Coefficients 312
- Correlation Coefficients 318
- Correlation Indexes 320

Multiple Correlation 321
- Coefficients of Multiple Correlation and Net Regression 321
- Multiple Curvilinear Correlation 327

CHAPTER 19

THE RELIABILITY OF AN INDIVIDUAL FORECAST AND OF TIME-SERIES ANALYSES 341

Reliability of an Individual Forecast 341
- Simple Correlation 342
- Multiple Correlation 344

Extrapolation of a Regression Equation beyond the Observed Range 347

Error Formulas for Time Series 349

Practical Procedures for Judging Reliability of Forecasts 356

CHAPTER 20

INFLUENCE OF SELECTION OF SAMPLE AND ACCURACY OF OBSERVATIONS ON CORRELATION RESULTS 359

Selection of Sample 359
- With Respect to Values of the Independent Variable 360
- With Respect to Values of the Dependent Factor 361
- With Reference to Values of Both Variables 362

Accuracy of Observations 364
- Errors in the Dependent Variable 365
- Errors in the Independent Variable 366
- Errors in Both Variables 366
- Errors of Observation in Multiple Correlations 367

CHAPTER 21

MEASURING THE RELATION BETWEEN ONE VARIABLE AND TWO OR MORE OTHERS OPERATING JOINTLY 372

Determining a Joint Function for Two Independent Variables 376

Determining a Joint Function for Two Independent Variables, Holding Other Independent Variables Constant 390

Measuring Correlation with Respect to Joint Functions 391

Determining Joint Influence of Three or More Independent Variables 391

CHAPTER 22

PAGE

SUPPLEMENTARY METHODS FOR DETERMINING CURVILINEAR AND JOINT RELATIONS 396

Determining Net Regression Curves by Mathematical Functions 396
Supplementary Methods of Determining the Final Shape of Net Regression Curves 401
Determining Joint Relations by Contours 404
Determining Joint Relations by Definite Mathematical Functions .. 407
Measures of Correlation for Mathematically Determined Regressions 412

CHAPTER 23

TYPES OF PROBLEMS TO WHICH CORRELATION ANALYSIS HAS BEEN APPLIED 415

Land Values 415
Physical Relations between Input and Output 416
Weather Conditions and Crop Yields 418
Relation of Physical Characteristics to Chemical Characteristics 420
Relation of Farm Organization to Farm Income 421
Relation of Economic Conditions to Market Price for a Commodity ... 422
Relation of Characteristics of Different Lots of a Commodity to Prices at which They Sell 424
Other Price Studies 427
Relation of Changes in Production to Prices and Other Factors 428
Miscellaneous Agricultural Problems 429
Correlation in Psychology and Education 429
Correlation Analysis in Other Fields 433
More Recent Applications of Correlation Analysis 433

CHAPTER 24

STEPS IN RESEARCH WORK, AND THE PLACE OF STATISTICAL ANALYSIS 442

Relation of Statistical Analysis to Research 442
Stating the Objective 442
Developing an Hypothesis 443
Measuring the Factors 444
Studying the Apparent Relations 445
Running a Correlation Analysis 446
Meaning of Correlation Results 450

APPENDIX 1

PAGE

METHODS OF COMPUTATION 455

Coefficients of Correlation and Regression 455

Coefficients of Multiple Correlation and Net Regression 459

Use of the Check Sum 461

The Doolittle Method for Solving Normal Equations 464

Standard Errors of Partial Regression Coefficients and of an Individual Estimate 469

Coefficients of Partial Correlation 474

Graphic Processes with the Short-cut Method 479

APPENDIX 2

TECHNICAL NOTES 486

APPENDIX 3

GRAPHIC CHARTS FOR INTERPRETING OR ADJUSTING CORRELATION CONSTANTS 504

Reliability of Small Samples 504

Reliability of Observed Correlations 504

Adjustment of Correlation for Size of Sample 511

APPENDIX 4

LIST OF IMPORTANT EQUATIONS 512

APPENDIX 5

GLOSSARY 521

REFERENCES 522

INDEX 523

CHAPTER 1

MEASURING THE VARIABILITY OF A STATISTICAL SERIES

Statistical analysis is used where the thing to be studied can be reduced to or stated in terms of numbers. Not all the undertakings that rely on measurements ordinarily employ statistical analyses. In surveying, physics, and chemistry, for example, the particular thing being studied can usually be measured so closely, and varies over such a small range, that the true value can be established within narrow limits. In fact, the concept of true value owes its existence to the reproducibility of measurements in certain fields. In many natural sciences, likewise, the problem to be studied can be simplified by the use of controlled experimental conditions, which permit the influence of various factors to be studied one at a time. Even in such sciences, statistical methods can be used to plan experiments in such a way as to make the conclusions most significant with a minimum of effort. In the social sciences, there are fewer opportunities for the use of controlled experiments. Such sciences have to rely on statistical analysis, both to judge the significance of observed differences and to untangle the separate effects of multiple factors. Statistical analysis is used in the study of occurrences where the true value or relation cannot be measured directly or is hidden by other things. The numerical statement of the occurrence or of the relationship cannot be obtained directly from the original or "raw" figures. Instead, the data must be analyzed to determine the values desired.

Use of
Stat.
Analysis

The especial need for analytical methods in the social sciences has been clearly stated by an eminent Englishman, as follows:[1]

> Causation in social science is never simple and single as in physics or biology, but always multiple and complex. It is of course true that one-to-one causation is an artificial affair, only to be unearthed by isolating phenomena from their total background. Nonetheless, this method is the most powerful weapon in the armory of natural science: it disentangles the chaotic field of influence and reduces it to a series of single causes, each of which can then be given due weight when the isolates are put

[1] Julian Huxley, The science of society, *Virginia Quarterly Review*, Vol. 16, No. 3, pp. 348–65, summer, 1940.

back into their natural interrelatedness, or when they are deliberately combined (as in modern electrical science and its applications) into new complexes unknown in nature. This method of analysis is impossible in social science. Multiple causation here is irreducible.

The problem is a two-fold one. In the first place, the human mind is always looking for single causes for phenomena. The very idea of multiple causation is not only difficult, but definitely antipathetic. And secondly, even when the social scientist has overcome this resistance, extreme practical difficulties remain. Somehow he must disentangle the single causes from the multiple field of which they form an inseparable part. And for this a new technique is necessary.

The arithmetic average. The basic forms of statistical analysis have to do with organizing quantitative information as a basis for drawing inferences. Some of the basic work involves averaging and classifying data. Thus if one were studying the yield of corn in one year in some area, say a county, for example, he might talk with 20 farmers picked at random and obtain figures, such as those in Table 1, showing the yield of corn which each farmer had obtained.

The most natural first step in reducing such a series of observations to more usable shape is to find the arithmetic average—to add all the yields reported and divide by the number of items. The 20 reports total 600 bushels, or an average of 30 bushels.[2] This provides a single figure into which is condensed one characteristic of the whole group.

[2] Bushels are used here to represent any other quantity in which one might be interested in a particular case. If we let X' represent the number of bushels reported by farmer 1, X'' the bushels reported by farmer 2, X''' the bushels by farmer 3, and so on, we can then represent the sum of all the reports by the expression ΣX (read "summation of the X's"). Similarly, if we use n to represent the number of observations we have obtained and use M_x to represent the *average* (or *mean*) number of bushels for all reports we can define the *arithmetic mean* by the formula:

$$M_x = \frac{\Sigma X}{n} \tag{1}$$

This formula can be applied to anything we are studying, no matter whether X means bushels of corn, inches in height, degrees of temperature, or any other measurable quantity; or whether there are 2 cases or 2 million. This is a perfectly general formula which can be applied to any given problem. As statistics is a study of general methods, so stated that they can be applied to particular problems as desired, it will be necessary to use many general formulas of this sort. The student should therefore familiarize himself with the definitions given above and with the way they are used in formula (1), so that he will be able to understand and use each formula as it occurs.

demerit of average-

But the average is not the only characteristic of the group which might be of interest. The average would still be 30 if every one of the 20 farmers had had a yield of 30 bushels per acre; yet there certainly would be a significant difference between 20 reports each of 30 bushels, and 20 reports ranging from 23 to 38 bushels, even though both did have the same average.

TABLE 1

YIELDS OF CORN OBTAINED BY TWENTY FARMERS*

33 = ≥32·5 but <≠ 33·5

Farmer	Yield	Farmer	Yield	Farmer	Yield	Farmer	Yield
	Bushels per acre		*Bushels per acre*		*Bushels per acre*		*Bushels per acre*
1	29	6	33	11	29	16	33
2	25	7	26	12	35	17	31
3	38	8	28	13	26	18	37
4	30	9	30	14	23	19	28
5	27	10	29	15	31	20	32

* In making entries in a table such as this, the actual values may be "rounded off" to any desired extent. In this case they are rounded to the nearest whole bushel. For example, "33 bushels" represents any report of 32.5 bushels or more, and any up to but not including 33.5 bushels. If the original reports were secured to the nearest tenth bushel, this might be indicated by writing "32.5–33.4" instead of "33"; or if secured to the nearest hundredth bushel, by writing "32.50–33.49." The entry "32.5 to 33.5" will be used to indicate "from 32.5 *up to but not including* 33.5," whereas "32.5–33.4" will be used to mean "from 32.5 to 33.4, both inclusive."

W. S.

Classifying the data. One way of showing the differences in the individual reports is to arrange them in some regular order. If the farmers interviewed have simply been visited at random, and not selected so that those visited first represent one portion of the county and those visited later another portion, the order in which the records stand has nothing to do with their meaning. As a first step to seeing just what the data do show they can be rearranged in order from smallest to largest, as shown in Table 2.

TABLE 2

YIELDS OF CORN ON 20 FARMS, ARRANGED IN ORDER OF INCREASING YIELDS

Bushels per acre

23	28	30	33
25	28	30	33
26	29	31	35
26	29	31	37
27	29	32	38

38
−23
15

It is now easier to tell from the series something about the group of reports. One can now see that only 1 farmer had yields of less than 25 bushels per acre, and only 2 had more than 35, so that 17 out of the 20 had 25 to 35, inclusive. The series shows, too, that 10 of the farmers had less than 30 bushels of corn per acre and 10 had 30 or more, so that the figures 29 and 30 mark the middle of the number of yields reported. If we divide each half into halves again, we see that 5 men had yields of 27 bushels or less, 5 had yields of 33 bushels or more, whereas 10 men—half of those reporting—had yields of 28 to 32 bushels, inclusive. This tells something about how variable yields were from farm to farm in the area from which the reports were secured—half the reports fell within this 5-bushel range.[3]

Even as rearranged in Table 2, the 20 reports still constitute a large tabulation. If there were several hundred, such a listing would be so unwieldy that it would be difficult to use.

Frequency tables. The records can be studied more easily if, instead of writing "29" three times when there are 3 farmers with 29 bushels each, we simply show that each of 3 men reported 29 bushels. Similarly, instead of putting "30" down twice, we can show that 30 bushels were reported by 2 men. If this operation is performed for all the reports, the data can then be assembled into what is known as a "frequency table." It shows the frequency, that is, the number of times each yield of corn was reported.

In preparing a frequency table such as Table 3, spaces are put in for all yields (such as 24 bushels) for which no reports were received, but which lie between the largest and the smallest report, to show clearly that no such yields were reported.

Table 3 is an improvement on Table 2, but it is still pretty long—and if the lowest yield had happened to be 15, say, and the highest 60, it would have been longer still. For that reason it is frequently desirable to group the reports, not only for a yield of a specified number of bushels but for yields within a certain range of bushels. Thus Table 4 is just the same as Table 3, except that, instead of showing the number of reports by individual bushel groups, it shows the number of reports for groups covering 3 bushels.

The presentation is now condensed enough so that it can be readily

[3] In statistical terminology, the figure that divides the number of reports into halves—as 29.5 in this case—is termed the *median*; and the figures that divide the numbers into quarters—as 27.5 and 32.5—are termed the *lower* and *upper quartiles*. The difference between the two quartiles, within which the central half of the reports fall, is termed the *interquartile range*.

understood. It is easy to see that most of the reports fell around 25.5 to 34.4 bushels and that more fell near 30 bushels than anywhere else. Of course, the 3-bushel group is purely arbitrary, and any other convenient "class interval," as it is called in statistical terminology, could have been used. Thus, if a 5-bushel class interval had been selected, the convenient groups 19.5–24.4, 24.5–29.4, 29.5–34.4, and 34.5–39.4 bushels could have been established, giving frequencies of 1, 9, 7, and 3 for the four groups. Just what class interval makes the most satisfactory table for any given set of data

TABLE 3

FREQUENCY TABLE, SHOWING NUMBER OF TIMES EACH YIELD WAS REPORTED, BY INDIVIDUAL BUSHELS

Yield of Corn	Number of times reported	Yield of Corn	Number of times reported
Bushels		*Bushels*	
23	1	31	2
24	0	32	1
25	1	33	2
26	2	34	0
27	1	35	1
28	2	36	0
29	3	37	1
30	2	38	1

TABLE 4

FREQUENCY TABLE, SHOWING NUMBER OF TIMES EACH YIELD WAS REPORTED, BY 3-BUSHEL GROUPS

Yield of corn	Number of times reported
Bushels	
22.5–25.4	2
25.5–28.4	5
28.5–31.4	7
31.5–34.4	3
34.5–37.4	2
37.5–40.4	1

depends upon how the data run and how much detail it is desired to show. Where convenient, class intervals of 10 or some fraction or multiple of 10 are most convenient—the example just given shows how much easier it is to comprehend the 5-bushel classes than the 3-bushel.[4]

Measures of Deviation

The average deviation. Table 4 shows, in fairly compact form, the way that the several individual reports fall on each side of the average value. For some uses, however, it is desirable to have a single figure which expresses the "scatteration" of the whole group of reports, in just the same way that the arithmetic mean expresses the average yield of the whole group.

One way in which the tendency of the group to scatter either faı from, or close to, the mean may be measured is by finding out how far, on the average, each report lies from the mean. The following tabulation illustrates the way in which this can be done:

TABLE 5

Computation of Average Deviation from the Mean

Original report	Mean	Report minus the mean
Bushels	*Bushels*	*Bushels*
29	30	−1
25	30	−5
38	30	8
30	30	0
27	30	−3
. . *		
Total.	. . . .	60†

* The remaining 15 reports are not shown in this table, though included in the total.
† The plus and minus signs are disregarded in making this total.

$$\text{Average deviation} = \frac{60 \text{ bushels}}{20} = 3 \text{ bushels}$$

[4] Where there is a tendency for the reports to be grouped around certain values, such as 5, 10, it is desirable to take the class intervals so as to make these values fall in the middle of the groups. Thus, with a concentration on even 5's and 10's, the groups 2.5–7.4, 7.5–12.4, 12.5–17.4, etc., may be used.

In computing the average deviation, the plus and minus signs are disregarded in adding up the individual differences from the mean.[5]

The new figure, 3 bushels, is the *average deviation* of all the reports. It shows that the 20 individual reports differed from the mean yield of 30 bushels by an average of 3 bushels each. This furnishes a single figure which expresses how much or how little the individual yields differed from the average yield. If the group of 20 reports were being compared with another group of 20, all of 30 bushels each, the *average deviations* of the two sets would indicate at once the difference in their make-up, even though both sets had exactly the same average value of 30 bushels. The second set, with all the reports exactly equal to the average, would have an average deviation of 0, as compared to the 3-bushel average deviation for the first set.

Very

[5] Before writing the general formula for the average deviation it is first necessary to have some way of writing *any* deviation. Using X to indicate any given report, as before, and M_x to indicate the arithmetic average of all such reports, the small x will be used to indicate the deviation of each report from the mean of all, thus:

$$\begin{aligned} X \ - M_x &= x \\ X' - M_x &= x' \\ X'' - M_x &= x'' \end{aligned} \qquad (2)$$

and so on.

Similar to the previous usage, Σx (read "summation of all the small x's") is used to indicate the sum of the values such as x, x', x'', etc.

The average deviation, denoted by the sign δ, is then defined by the following equation:

$$\delta = \frac{\Sigma x \text{ (taken without regard to sign)}}{n} \qquad (3)$$

It is necessary to disregard the signs in taking this sum, as otherwise the sum would be zero. If the signs were not disregarded, the values added would be as follows:

it tells why it is necessary to disregard the sign

$$\begin{aligned} &\text{For item 1, } x \ \ (= X \ \ - M_x) \\ &\text{item 2, } x' \ \ (= X' \ - M_x) \\ &\text{item 3, } x'' \ (= X'' - M_x) \end{aligned}$$

and so on to the last item

$$\text{item } n,\ x_n \ (= X_n - M_x)$$

So when the deviations were summed,

$$\Sigma x = \Sigma X - nM_x$$

but

$$M_x = \frac{\Sigma X}{n}, \text{ so } nM_x = \Sigma X$$

hence

$$\Sigma x = 0$$

Whereas the arithmetic average is a measure of the central tendency of a group of reports, the average deviation is instead a measure of the "scatteration" of the individual reports—of their tendency to lie near to, or far from, the central value.

The standard deviation. How far a group of reports tends to scatter from the mean of the group may also be measured by another coefficient which has certain advantages from a mathematical point of view. This measure is based on the deviation of each report from the mean, just as is the average deviation. After the individual deviations are computed, each one is then squared. These squared values are added together to give the sum. This sum is then divided by the number of items, and the square root extracted of this average of the squared deviations.

TABLE 6

Computation of Standard Deviation from the Mean

Original report	Mean	Report minus the mean (= deviation)	Deviations squared
Bushels	*Bushels*	*Bushels*	*Bushels*
29	30	−1	1
25	30	−5	25
38	30	8	64
30	30	0	0
27	30	−3	9
. . *			
Total..........	....	....	288

* The remaining 15 reports are not shown in this table, though included in the total.

The sum of the squared deviations, as shown in Table 6, is then divided by the number of items included in the group, and the square root of the result computed. The computation is as follows:

$$\frac{288}{20} = 14.4$$

$$\text{Standard deviation} = \sqrt{14.4} = 3.79 \text{ bushels}$$ [6]

[6] The Greek letter σ is used as the sign for the standard deviation. Using x to represent individual differences from the mean, as before, x^2 for the square of each

The new value, 3.79 bushels, is called the standard deviation.[7] (It is sometimes called the root-mean-square deviation, because it is the square root of the mean of the squares of the individual deviations.) In comparison to the average deviation, which was found to be 3 bushels, it is somewhat larger. That is a relation which always holds —the process of squaring the deviations tends to emphasize the largest deviations more than does merely averaging them together. With well-distributed observations, so that the distribution is "normal" or nearly "normal," the standard deviation is about one and a quarter times as large as the average deviation.[8]

Imp

1¼

of such deviations, and Σx^2 for the sum of all such values, the standard deviation is defined mathematically by the formula

$$\checkmark \quad \sigma_x = \sqrt{\frac{\Sigma x^2}{n}} \tag{4}$$

Where the arithmetic average is a fraction, so that computing each individual deviation and squaring it would take much arithmetic for accurate work, the standard deviation may be computed more easily by the following formula:

[illegible]

$$\checkmark \quad \sigma_x = \sqrt{\frac{\Sigma X^2}{n} - M_x^2} \tag{5}$$

Here the original X values are squared instead of the deviations from mean, or x, values. It can be readily demonstrated algebraically that the two formulas give identical values for σ_x.

proof

Thus $\quad$ each $x = X - M_x$

each $x^2 = X^2 - 2XM_x + M_x^2$

hence

$$\Sigma x^2 = \Sigma X^2 - 2\Sigma XM_x + \Sigma M_x^2$$

But

$$\Sigma X = nM_x$$

and

$$\Sigma M_x^2 = nM_x^2$$

hence

$$\Sigma x^2 = \Sigma X^2 - 2nM_x^2 + nM_x^2$$

and

$$\Sigma x^2 = \Sigma X^2 - nM_x^2$$

[7] For a shorter method of computing the standard deviation, when there is a large number of observations, see Note 1 at the end of this chapter.

[8] A "normal distribution" is such a one as will be obtained from a series of observations of a variable influenced only by a large number of random or chance causes, each one small in proportion to the total. Thus the values secured by tossing a number of dice, and noting the spots at each reading, tend to conform to a "normal curve." Variables composed of a large number of small, independent elements also tend to have a normal distribution. Since this distribution can be studied mathematically, it is possible to work out theoretically many of its properties. These theoretical characteristics of the normal curve are valuable in studying data where the distributions are nearly normal.

definition of a normal distribution

The distribution of the observations shown in Table 3 is fairly regular. Most of the reports come at about the middle values and then thin out to both ends (that is, the distribution approximates normality). In such cases the standard deviation gives a measure of the range within which a definite proportion of the cases will be included. Specifically, if we take the range from the distance of the standard deviation below the mean to the distance of the standard deviation above the mean, about 68 per cent of the records will be included. In this particular case the mean is 30.00 bushels, and the standard deviation is 3.79 bushels, so the range will be from 3.79 less than 30.00, or 26.21, to 3.79 more than 30.00, or 33.79. Comparing this with Table 3, we find that 13 farmers reported yields between 26.5 and 33.4 bushels, whereas 4 reported 26.4 or less, and 3 reported 33.5 or more. The range 26.5 to 33.4 thus included 13 out of the 20 cases, or 65 per cent. This comes as close to the 68 per cent which would be expected for the range 26.21 to 33.79 provided the distribution of the data were normal as would be anticipated with only 20 observations.

[illegible]

variance [illegible]

For some uses, the square of the standard deviation has advantages over the standard deviation itself. Just as the standard deviation, 3.79 bushels in this case, may be thought of as measuring "variability," so the standard deviation squared, 14.4, may be thought of as measuring "average squared variability." The term "variance" has been suggested by R. A. Fisher, an eminent English statistician, to designate this squared variability, and that term will be used hereafter in this book when the standard deviation squared is to be referred to.

The relation of the three measures which have been discussed in this chapter—the mean, the average deviation, and the standard deviation—is illustrated graphically in Figure 1. Here the frequency distribution shown in Table 4 has been charted, showing the yield in bushels of corn along the bottom of the chart, and the number of reports falling in each group along the sides.[9]

[9] Mathematically, the quantities which are measured from left to right, and shown along the bottom of the chart, as the bushels of corn are here, are called the "abscissas," whereas the quantities which are measured from bottom to top, and shown along the sides as the number of reports are here, are called the "ordinates." Since any point in the whole chart can be located by telling how far it is from the left side, and how high it is from the bottom, these two items tell exactly where any particular point in the figure should fall. Thus the line for the group from 28.5 to 31.5 bushels has for ordinate the height 7 farms, and the abscissas of the ends of the line are 28.5 and 31.5 bushels. The ordinate and abscissa, taken together, are called the "coordinates" of a point.

Besides showing the number of reports included in each 3-bushel group by the height of the continuous line, the position of the mean in about the center of the group of reports is indicated, and likewise the number of reports included within a range of both one average deviation and of one standard deviation on each side of the mean.

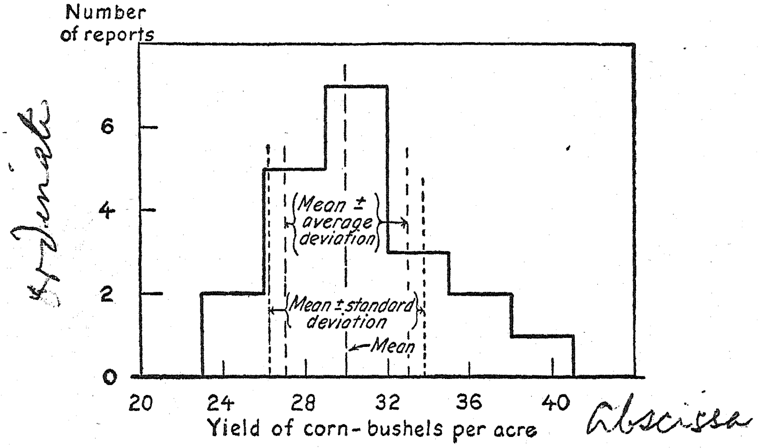


FIG. 1. Frequency distribution of corn yields, and range above and below the mean included by average and standard deviations.

Summary. This chapter has shown (1) how a series of measurements of any one variable, such as the yield of corn from farm to farm, may be classified into a frequency distribution which shows how the individual reports are distributed from high to low; (2) how an arithmetic average may be computed which shows the value around which all the reports center; and (3) how the variation of the individual reports from the average may be summarized by computing the average deviation or the standard deviation, either one of which serves as an indication of the variability of the items included in the particular series. Although these statistical constants, especially the arithmetic average, are frequently of value for themselves alone, they are discussed here because it is necessary to know how they are computed and what they mean before the next propositions to be discussed can be fully understood.

Shorter method calculating S.D.

N

Note 1, Chapter 1. Where the number of observations is large, the standard deviations may be computed more readily from a grouped frequency table than from the individual items. This process is illustrated in the following tabulation.

Yield	Number of reports (F)	Deviation from assumed mean (d)	Extensions	
			dF	d^2F
22.5 to 25.5...........	2	−2	−4	8
25.5 to 28.5...........	5	−1	−5	5
28.5 to 31.5...........	7	0	0	0
31.5 to 34.5...........	3	+1	3	3
34.5 to 37.5...........	2	+2	4	8
37.5 to 40.5...........	1	+3	3	9
Sums...............	20	...	+1	33

The standard deviation is then calculated from the grouped data by the formula

$$\sigma_u = \sqrt{\frac{\Sigma(d^2F)}{n} - \left[\frac{\Sigma(dF)}{n}\right]^2 - \frac{c^2}{12}} \qquad (6)$$

$\sigma x = (\sigma u)$ unit see p. 13

Substituting the values shown in the tabulation

$$\sigma_u = \sqrt{\frac{33}{20} - \left(\frac{1}{20}\right)^2 - \frac{1}{12}} = \sqrt{1.65 - (0.05)^2 - 0.0833} = 1.25$$

In making this computation, any convenient group may be selected as the assumed mean, and the deviations of the other groups (d) cal~~culated as departures~~ from it. This method assumes that all the cases in each group fall at the center of the group. With most variables, with a tendency toward a normal distribution, the average of the items in each group will fall somewhat nearer the center of the distribution than the midpoint of the group, so the use of this method tends to give too large a value for the standard deviation. The correction $-\frac{c^2}{12}$ called "Sheppard's correction" after its originator, makes an approximate allowance for this tendency. The c of the formula stands for the number of units of d in each class interval. Where a unit of 1 is used for each class interval, as in this problem, the correction becomes simply $-\frac{1}{12}$, to be applied to σ_u^2.

In computing the standard deviation from a grouped frequency table, the σ calculated will be in terms of the units in which d is expressed. In the illustration, each unit in d—one class interval—represents 3 units in X, since the yields were grouped in 3-bushel classes. The standard deviation computed in terms of class intervals, σ_u, is therefore only one-third as large as is the standard deviation in terms of X.

The latter may be calculated from the former by multiplying σ_u by the number of units in each group. That is,

$$\sigma_x = (\text{units of } X \text{ per class interval})\ \sigma_u$$

In this problem

$$\sigma_x = 3\ (1.25) = 3.75$$

The resulting value, 3.75, found by the short-cut method, is seen to be almost the same as the exact value of 3.79 bushels, previously found by the longer method. The greater the number of cases, and the more nearly normal the distribution, the more time will the short-cut method save, and the more nearly will its approximate result agree with the exact value found by the longer method.

Imp

[illegible]

CHAPTER 2

JUDGING THE RELIABILITY OF STATISTICAL RESULTS

Almost without exception, the object of a statistical study is to furnish a basis for generalization. In a case like that discussed in the preceding chapter, for example, no one would be likely to visit 20 farms scattered all over a county simply for the purpose of finding out what the yield of corn was on *those particular farms*. Instead, he might be studying the yield on those farms as a basis for determining what the average yield of corn was for all the farms in the county. Stated in statistical terms, he would be finding out what was the average yield in a *sample* of farms, picked at random, with a view to determining what was about the average yield in the *universe* in which he was interested, that is, on all the farms in the county.[1]

Of course it would be possible to visit all the farmers in the county, find out exactly what yield each one obtained, and so get an average of all the yields in the whole county. But this process would not only be expensive but also in most cases would be a pure waste of time and energy. We need only take a large enough sample by a well-designed sampling method to satisfy ourselves to any desired degree of accuracy concerning the actual average for all the farms of the county. In this case, 100 records may enable one to determine the average yield quite as accurately as is necessary. Obtaining records from all the several thousand farmers in the county might add nothing to the significance of the results.

Before considering ways of finding out how many records would be needed in any given case, we might well discuss a little more fully what the process of statistical inference involves. Really, all that we do is to examine or measure a certain group of objects, and *infer* from the size or measurement of those objects, or from the way those objects behave, what will be the size of other objects of the

[1] These two terms, "universe," meaning the whole group of cases about which one is interested in finding out certain facts, and "sample," meaning a certain number of those cases, picked at random or otherwise from all those in the particular universe, are both used frequently in statistical work, and should be clearly understood.

same sort, or how other objects of the same kind will behave. This process is also called *induction,* because from particular facts about particular objects we lead out (*in duct*) *general* conclusions as to what will be the facts for all such objects in general. Now of course we do not really know what the particular facts are for any particular object without actually examining that individual object. All that we can do is to separate off certain groups of objects which we know to be alike in one or more particulars, and then assume that they will be alike in other particulars too, even though we do not examine every one to prove it. In the case of our farms, all that we know about them is that they are in the same county. Now because they are in the same county, we may expect that the temperature will be about the same, the rainfall will be similar, and the growing season will probably not be much different from farm to farm. We may also expect that the kind of soil will not be very greatly different from farm to farm, and that the fertility will be somewhere near the same. Finally, we may expect that the fields are equally well drained on the farms within the county.[2] But these expectations are not necessarily matters of known fact—we may expect that they are so from our general knowledge of the particular situation and of other similar situations. If the conditions agree with our expectations, generalizations from the facts of our sample to the facts of the universe as a whole may be correct; if conditions do not agree with expectations, then our general conclusions may be incorrect. In either case it is not merely a matter of statistical technique but also of prior or additional knowledge of the subject. All that the statistical technique can do is to provide us with an average (or other measure or description of our facts) and a statement of how much confidence we can place in that average *under certain given assumptions.* Those assumptions may not be correct in any given case, and then our conclusion will be incorrect also; but that is not the fault of the statistics, but of the statistician; not of the facts, but of the use to which we try to put them.

Assumptions in sampling. The basic assumptions upon which the theory of sampling rests apply both to the way in which the sample is obtained and to the material which is being sampled. With respect to the material sampled, the assumption is that there is a large "uni-

[2] Obviously, these things would not be true in many sections. In hilly or mountainous areas temperature, rainfall, and length of growing season may differ very greatly within short distances, whereas in other regions, such as the Coastal Plains areas, the soils may be so varied that very fertile and very infertile soils are jumbled together in a veritable crazy-quilt.

verse" of uniform conditions, in that throughout the universe the individual items vary among themselves in response to the same causes and with about the same variability. With respect to the selection of sample, the values must be so selected (*a*) that there will not be any relation between the size of successive observations, that is, that the chances of a high observation being followed by another high observation will be just the same as of a low or a medium observation being followed by a high observation; (*b*) that the successive items in the sample are not definitely selected from different portions of the universe in regular order, but are simply picked at random so that the chance of the occurrence of any particular value is the same with each successive observation in the sample; and (*c*) that the sample is not picked all from one portion of the universe, but that the observations are scattered through the universe by purely chance selection.[3] Where these assumptions are fulfilled, the sample is designated a "random sample," and its reliability may be estimated by the methods now to be described.

Taking up the question of how reliable a statistical average really is, we must first consider, "What is the meaning of *reliable?*" If we are interested in corn yield, for example, it is obvious that a perfectly reliable sample would be one whose average agreed exactly with the average yield in the county. But if we are interested in knowing the average yield to within one bushel, then for that purpose the sample would be sufficiently reliable if its average came within one bushel of the average for the whole county.

Variations in successive samples. Suppose that 20 farms had been visited at random, with the results already presented. If we wanted to find out how near we could expect the average from that sample to come to the average for the county as a whole, we might try taking another sample—visiting 20 other farms at random, and getting the average yield for those 20. If the average yield of the second sample differed from the average of the first sample by, say, 3 bushels, we should know that both could not come within one bushel of the true average; if, however, the average of the second sample came within a

[3] Where the items are so selected as to represent different portions of the universe, it may be called a "stratified sample"; where they are all selected from one portion of the universe, it may be called a "spot" sample.

Where the universe is not completely uniform, a "stratified" sample tends to be more reliable than a random sample, while a "spot" sample tends to be less reliable than a random sample. See G. U. Yule, *Introduction to the Theory of Statistics*, pp. 347 to 349 of sixth edition, for formulas as to the reliability of stratified and spot samples.

half bushel of the first average, we should be inclined to place more confidence in it. If we repeated the process several times over, and all the different samples had averages falling within one bushel of each other—say between 29.0 and 30.0 bushels—then we should feel pretty certain that the average yield for the county as a whole was 29.5 bushels, or very close to it.

Let us suppose that 15 more samples had been made, each from 20 farms selected at random, and that when we tabulate the 16 averages from the 16 different samples, we have the following 16 values:

TABLE 7

Average Yield of Corn in One County, as Determined by 16 Different Samples of 20 Farms Each

Sample	Yield	Sample	Yield
	Bushels per acre		*Bushels per acre*
1	30.0	9	30.3
2	27.5	10	28.9
3	29.3	11	29.3
4	30.6	12	28.0
5	29.8	13	29.2
6	31.1	14	30.9
7	28.3	15	29.1
8	29.6	16	30.4

Although the 16 averages range all the way from 27.5 bushels for the smallest to 31.1 bushels for the largest, we can see that most of them fall around 29 or 30 bushels. This is even more evident when we arrange the 16 reports in a frequency table as shown in Table 8.

Although there is some tendency for the averages to cluster around 29 and 30 bushels, still there are several below 28.5 and several above 30.5. The average for the whole group is 29.5 bushels, and the standard deviation is 0.99 bushel, or, for practical purposes, 1 bushel.

The fact that the standard deviation of the group of averages is 1 bushel tells us one thing about the way they scatter, from what we already know about the meaning of *standard deviation.* It tells us that about 68 per cent of them will fall in the range between one standard deviation below the mean of all the averages and one standard deviation above the mean. In this particular case, the mean is 29.5 bushels, and the standard deviation is approximately 1 bushel, so the range of one standard deviation above and below the mean includes approximately 28.5 bushels to 30.5 bushels. Checking this against the array of averages shown in Table 8, we find that this range does include 10 out of the 16 cases, or close to the proportion expected.

TABLE 8

Frequency Table Showing the Number of Times Various Average Yields Were Obtained out of 16 Samples, by One-half Bushel Groups

Yield of corn	Number of averages in group	Yield of corn	Number of averages in group
Bushels		*Bushels*	
27.5–27.9	1	29.5–29.9	2
28.0–28.4	2	30.0–30.4	3
28.5–28.9	1	30.5–30.9	2
29.0–29.4	4	31.0–31.4	1

Now let us go back to our single original average of 30 bushels, based on visits to the original 20 farms. What we want to know is how reliable that one average is. Stated another way, how much is that average likely to be changed if the study were made over again—if another sample of the same size were taken?

In Tables 7 and 8 we have seen how it might actually work out if we *did* do the study over several times. We have seen that, in case the new averages did fall as shown in those tables, two-thirds of the new averages would fall within a range of 2 bushels. Furthermore, those figures showed that *all* the different averages fell within a range of 4 bushels (27.5 to 31.5). But those conclusions were obtained only *after* getting 15 more samples of 20 cases each, and making 15 new averages, one for each sample. Is there any way to find out how much the single original is likely to vary from the true average without going to all the work of taking a number of new samples?

Estimating the Reliability of a Sample

If we could estimate the extent to which the averages from new samples would be likely to vary, *without ever getting the new samples,* then we should know something more about how much faith we could put in the particular average which we had already. For example, if in the present case we knew that, if we did go out and get a large number of new averages (such as those shown in Tables 7 and 8), those new averages would have a standard deviation of 1 bushel, this fact would tell us at once *something* about how much our one average was likely to be different from the real average on all the farms. For example, we should know that about 68 per cent of the averages would lie in a range of 2 bushels (one standard deviation on each side of the mean of the samples). The one particular average which we had obtained might be any one of all those in a distribution like that shown in Table 8. If we assume that the mean of all the samples would coincide with the true average, then, as we have just seen, the chances would be about 68 out of 100 that our average was one of the averages falling within *one bushel* of the true mean. If on the other hand we knew that the standard deviation of a group of new averages would probably be, say, 5 bushels, then we should know that we only had about 2 chances out of 3 of the mean of any one sample coming within *five bushels* of the true average. Obviously, when an average has 2 chances out of 3 of coming within one bushel of the true average it is much more reliable than if it had 2 chances out of 3 of coming within *five bushels* of the true average.

Whether we can judge how reliable a given average really is depends, therefore, on whether we can tell what would be the standard deviation of a number of similar averages, computed from random samples of the same number of items drawn from the same universe. If we could tell exactly what that standard deviation would be, we should know how much faith we could put in the average we had—we should know what the chances were of its being changed if the study were made over. Even if we did not know *exactly* what the standard deviation of the whole group of similar averages would be it would be some help if we knew approximately what it would be, or if we had a minimum or maximum value for its size, so that there would be some measure of how much trust to place in the particular average.

Computing the standard error. Fortunately, it is possible to estimate with some degree of accuracy what the standard deviation of a whole series of averages is likely to be, if each average is computed from a sample of the same size and drawn from the same universe.[4] Except under the exact assumed conditions, which are seldom completely obtained in practice, this estimate is not necessarily the best that could be made. Even so, the ability to make a rough estimate is a tremendous aid to statistical investigators, for it affords some check on the dependability of results, without going to the expense that would be

[4] Note 1 of Appendix 2 gives the derivation of this formula and shows the specific assumptions on which it is based.

involved in repeating every sample 15, 20, or more times, to make sure that a reliable result had been obtained.

The method for computing the estimated standard deviation of the average involves just two values. These are (1) the standard deviation of the items in the universe from which the sample was drawn and (2) the number of items in the sample. We do not know the standard deviation of the items in the universe, however, and can only estimate it from the standard deviation of the items in the sample. It has been determined that an unbiased estimate of the standard deviation in the universe can be made by adjusting the standard deviation observed in the sample as follows: [5]

Estimated stand. dev. of the universe

$$= (\text{observed stand. dev. in the sample})\left(\sqrt{\frac{n}{n-1}}\right)$$

In this case

$$= 3.79\sqrt{\tfrac{20}{19}} = (3.79)(1.026)$$

$$= 3.89$$

The standard deviation of the group of averages may next be estimated by dividing the estimated standard deviation in the uni-

[5] Using the symbol σ as before to mean the standard deviation observed in the sample, and $\bar{\sigma}$ to represent the estimated standard deviation in the universe from which the sample was drawn, we can define the estimated value as

$$\bar{\sigma} = \sigma\sqrt{\frac{n}{n-1}} \tag{6.1}$$

It may more readily be computed by the equation

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{n-1}} \tag{6.2}$$

The two equations are identical, as may readily be proved by combining equations (4) and (6.1).

When equation (5) is used, $\bar{\sigma}$ may be computed

$$\bar{\sigma}_x = \sqrt{\frac{\Sigma X^2 - nM_x^2}{n-1}} \tag{6.3}$$

verse by the square root of the number of cases in the sample. Thus, for our original sample of 20 farms,[6]

Standard error of the average

$$= \frac{\text{estimated standard deviation of items in the universe}}{\text{square root of the number of cases in the sample}}$$

$$= \frac{3.89 \text{ bushels}}{\sqrt{20}}$$

$$= \frac{3.89 \text{ bushels}}{4.47}$$

$$= 0.87 \text{ bushel}$$

In comparison with the 15 other averages, all shown in Table 7, we see that in this case the standard deviation of all the averages was a trifle larger than we estimated it was likely to be—0.99 bushel, as compared to 0.87 bushel expected. It has already been noted that where a number of repeated samples are actually taken, this may easily occur. In practice, sampling rarely fulfills all the conditions on which the mathematical formula is based, and for that reason an average may be either less or more accurate than the estimated

[6] Here the symbol σ denotes the standard deviation as before, the subscript x indicates that it is the standard deviation of the individual items that go to make up our sample, and the subscript M indicates that it is the standard deviation of the means which is to be computed, thus:

$\bar{\sigma}_x$ = standard deviation of the items in the universe, estimated by equations (6.1), (6.2), or (6.3).

σ_M = estimated standard deviation of the group of averages if similar samples were repeated = *standard error* of the mean of X.

The standard error of the mean is then given by the formula

$$\sigma_M = \frac{\bar{\sigma}_x}{\sqrt{n}} \tag{7.1}$$

Here, just as in the previous formulas, n stands for the number of items in the original sample—the same items as those from which σ_x was computed.

In some statistical textbooks, a different notation is followed from that used here. In those books the Greek letters are used to represent the true values existing in the universe, whereas corresponding Latin letters represent the values for the same constants as determined from a particular sample. In this notation σ_x would mean the true standard deviation in the universe, whereas s_x would mean the standard deviation observed in a sample. This use is referred to here for the information of students who may have occasion to refer to other textbooks using this other notation.

standard deviation indicates that it is likely to be. Even so, this estimated "standard deviation of similar averages" is an exceedingly useful figure. Such an estimated standard deviation for an average (or any other statistical measure) is called the *standard error* of that average (or other statistical measure). It serves as a standard measure to give warning of about how much that sample may give results which vary from the true facts of the universe, solely as the result of chance fluctuations in sampling. It gives some indication of how much confidence can be placed in the measures computed from a sample.

Reliability of small samples. Where there are only a small number of observations in the sample, the standard deviation of the averages from a series of such samples tends to be somewhat larger than the standard deviation estimated by means of equation (7.1), and the distribution of the averages from such small samples tends to be somewhat different from that for large samples. If there are 30 or more observations in the sample, the difference is so small that it may be disregarded. The farther the number of observations falls below 30, the more serious the difference. A correction has therefore been worked out, by higher mathematics, to allow for this error in the estimated standard deviation where there are less than 30 observations. This correction is shown by comparing the difference between the sample mean and the true mean of the universe with the estimated standard error of the mean, and by indicating in what proportion of repeated samples of the same size this ratio will exceed given values. These proportions are shown in Table A and in Figure A on page 505.[7]

The table shows the proportion of the trials in which a sample of each given size will have an average which differs from the true average by more than the specified range. Thus, if there are a large

[7] Table A applies as stated only in the case of measures such as the arithmetic average, which are computed from the original data by the determination of a single constant. Where the computation of the statistical measure involves simultaneously determining two constants from the original data, $n - 1$ should be used for the "number of observations in the sample." This applies to the coefficient of regression. Where the computation of the statistical measure involves simultaneously determining a large number of constants, say j in number, from the original data, then $(n - j + 1)$ should be used for the "number of observations" in entering Table A or Figure A. Thus for a coefficient of partial regression, $b_{12.345}$ obtained from a sample of 20 observations, 5 constants are involved, so 16 would be used as the "number of observations" in using Table A to judge the reliability of the computed value. (Subsequent chapters will explain the meaning of the new coefficients mentioned here.)

number of observations in the sample, and we state that the true average lies within one standard error of the computed average, we should be wrong for 3 out of 10 such statements. (The exact proportion expected is 317 out of 1,000.) If there were 20 observations in the samples, and we made the same statement, we should be wrong 33 times out of 100. For samples with only 2 observations, such a statement would be wrong 50 times out of 100, on the average.

TABLE A

Proportion of Repeated Samples in Which the Ratio of the Error in the Mean to the Estimated Standard Error of the Mean Exceeds the Value Specified in the Left-Hand Column, for Various Sizes of Sample*

Ratio of the error in the mean to the estimated standard error of the mean	Size of sample (n)						
	2	4	6	10	16	20	30 or more
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
.50	.7048	.6514	.6382	.6290	.6244	.6228	.6171
1.00	.5000	.3910	.3632	.3434	.3332	.3298	.3173
1.50	.3744	.2306	.1940	.1678	.1544	.1500	.1336
2.00	.2952	.1394	.1020	.0766	.0640	.0600	.0455
2.50	.2422	.0878	.0544	.0338	.0246	.0218	.0124
3.00	.2048	.0576	.0300	.0150	.0090	.0074	.0027
3.50	.1772	.0394	.0172	.0068	.0032	.0024	.0005
4.00	.1560	.0280	.0104	.0032	.0012	.0008	
4.50	.1392	.0204	.0064	.0014			
5.00	.1256	.0154	.0042	.0008			

Based on article by "Student." New tables for testing the significance of observations. *Metron* V, No. 3, 105–120, 1925.

* See Figure A, Appendix 3, for full set of values.

The estimated standard error of 0.87 bushel from our single sample of 20 cases, with an average of 30.0 bushels, would therefore tell us that 67 per cent of such samples would have averages which fell within a range of 0.87 bushel of the true mean. If our sample is a true random sample, we should then have 2 chances out of 3 of being right if we estimated that the real average yield for all the farms in the county, the year the sample was taken, was within 0.87 bushel of the average shown by the sample.

It is important to keep in mind that the probabilities shown in Table A refer to the ratio between the error in the mean and the estimated standard error of that mean, and not to the error itself, The size of the ratio will depend both upon the size of the error and the size of the estimated standard error. At times the ratio may be very large, even when the error in the mean is small, merely because the sample happened to be one that showed an exceptionally small standard deviation. Conversely, the ratio will at times be small, not because the error in the mean is small but because the sample happened to be one that showed an exceptionally large standard deviation. For this reason it is well to be cautious in interpreting the average from a very small sample, even though that sample seems to be very reliable, as judged by the size of its estimated standard error and by the probabilities of various departures from the true mean, as read from Table A. This brings up the subject of the standard error of the standard error, which is treated in the next paragraph.

Standard error of the standard error. A small sample (say of 30 cases or less) cannot serve as a satisfactory guide to the facts of the universe, even with the aid of Table A. With a small sample, not only do we not know the true value of the mean, but also we do not know the true value of the standard deviation from which we estimate the standard error of the mean. Our estimate of the standard error of the mean is itself subject to error. With very small samples, say of 5 to 10 cases, this introduces a degree of unreliability which no amount of calculation can fully correct. The results are uncertain within wide limits, and only a larger sample, or several successive small samples, can reduce that uncertainty.

The standard error of the standard error, stated in relative terms, depends solely upon the number of cases in the sample. It is computed as follows:

Relative standard error of the standard error [8]

$$= \frac{1}{\text{square root of two times (number of cases in sample} - 1)}$$

[8] Using σ_{σ_M} to represent the relative standard error of the estimated standard error, we may define it

$$\sigma_{\sigma_M} = \frac{1}{\sqrt{2(n-1)}} \qquad (7.2)$$

A slightly more accurate estimate can be made by use of the equation

$$\sigma_{\sigma_M} = \frac{1}{\sqrt{n(n-1)}}$$

The differences between the two equations are, however, negligible. See W. Edwards Deming and Raymond T. Birge, On the statistical theory of errors, *Reviews of Modern Physics*, pp. 119–161, Vol. 6, July, 1934, For our sample of 20 cases

$$= \frac{1}{\sqrt{2(20-1)}} = \frac{1}{\sqrt{38}}$$

$$= 0.162$$

The standard error of the standard error, for the sample sizes shown in Table A, is given in Table B.

TABLE B*

RELATIVE STANDARD ERROR OF THE ESTIMATED STANDARD ERROR OF THE MEAN, FOR VARYING SIZES OF SAMPLE

Size of sample	Relative standard error†
2	0.707
4	0.408
6	0.316
10	0.236
16	0.183
20	0.162

* Footnote 7, on page 22, applies to Table B as well.

† Stated as a proportion of the estimated standard error.

Table B illustrates how, with very small samples, even our estimate of the standard error of the average is subject to a wide zone of uncertainty. With 4 cases, its own standard error is 41 per cent of the value computed.

Meaning and Use of the Standard Error

It is good statistical practice, whenever an average is cited, to give with that average its estimated standard error, so that the reader will know about how significant that average is and not be led into using it to make comparisons or to draw conclusions that are not justified by the number of observations which are summed up in that average. One way of doing this is to write the average followed by the statement "plus or minus the standard error." Thus, in the case we have been considering, with the single sample showing an average of 30.0 bushels with a standard error of 0.87 bushel, and with only 20 cases in the sample, the correct statement is to say "the average yield has been shown by the sample to be 30.0 ± 0.87 bushels (20 cases)."[9] If a similar sample from a different area has shown the average yield to be 28 ± 2.0 bushels (20 cases), the reader would know that there was a fair chance that the true average yield was really the same in both areas, in spite of the difference shown by the two averages.

The greatest value of the standard error does not lie in merely indicating how near the sample value may come to the true value, for two samples out of three, on the average of a number of such samples. In exactly the same way that we have seen that two-thirds of the averages from the samples usually fall within *one* standard deviation on either side of the true mean, mathematicians have determined for large samples that 19 out of 20 (95.45 per cent) of the samples will give averages which fall within *two* standard deviations of the mean, 369 out of 370 (99.73 per cent) will usually fall within *three* standard deviations of the mean, and all but one case out of 16,667 samples (99.994 per cent) will usually fall within *four* standard deviations of the mean.

When there are less than 30 observations in the sample, the tendency of the computed standard error to be misleading is even greater for high odds than it is for lower odds. Corrections to take this into account are also shown in Table A. Thus, with samples of 20 cases, 6 samples out of 100 will give averages differing from the true average by more than twice the computed standard deviation, and 7 samples out of 1,000 will miss the true average by more than three standard deviations. This last is three times the proportion of such failures which would occur in the long run with samples of over 30 observations. With very small samples, the failures for high odds occur even more frequently. Thus, for samples with only 4 observations, 14 samples out of 100 will differ from the true mean

[9] The most general practice is to write after the average ±.6745 times the standard error (0.59 bushel in this case, so the statement would read 30.0 ± 0.59 bushels). This value, $0.6745\sigma_M$, is called the *probable error* of the mean, since it gives the range within which the chances are even that the true mean lies, when there are more than 30 observations—and also the range *without* which the chances are even that the true mean lies. Since this tends to make the average appear rather more accurate than does the standard error, the practice suggested of using the standard error instead has been recommended by many competent statisticians. Wherever that is done, however, it would be well to insert a footnote explaining that it is the *standard error*, and not the *probable error*, which is being shown after the sign "±."

by twice the computed standard error, and about 6 out of 100 will differ by three times the standard error, on the average.

Where high reliability is desired, and only small samples are available, it is very important to take into account the corrections shown in Table A.

Interpreting the standard error in the illustrative problem. Ignoring for the time the lack of complete accuracy in our estimate of the standard error itself (page 24), we can interpret the statement that the average yield in the area studied was 30 ± 0.87 bushels in any of the following ways:[10]

a. If we state that the true mean lies within one standard error of the observed mean (between 29.13 and 30.87 bushels, in this case) each time we use a sample of this size, we shall be wrong in our statement one time out of three, on the average.

b. If we state that the true mean lies within two standard errors of the observed mean (between 28.26 and 31.74 bushels) each time we use a sample of this size, we shall be wrong in our statement one time out of 17, on the average.

c. If we state that the true mean lies within three standard errors of the observed mean (between 27.39 and 32.61 bushels) each time we use a sample of this size, we shall be wrong in our statement one time out of 135, on the average.

d. If we state that the true mean lies within four standard errors of the observed mean (between 26.52 and 33.48 bushels) each time we use a sample of this size, we shall be wrong in our statement only one time out of 1,250, on the average.

Comparing these conclusions with the 16 samples shown in Tables 7 and 8, we see that 2 of those samples did fall outside the limits given by twice the estimated standard error. If we had been so unlucky as to have got the worst one of these as our single sample, instead of the one which we actually did get, then we should not have hit the average even if we had used a range of twice the computed standard deviation as that within which we expected the true average to fall. On the other hand, every one of the averages fell within the range covered by three times the standard deviation. Even if, in picking our single sample, we had been unfortunate enough to draw the poorest one of the lot—the one which gave an average yield of 27.5 bushels—and had used a range of three times the standard error, we should have been correct in our statement as to the range within

[10] Figure A, page 505, which gives in more detailed form the corrections shown in Table A, may be used to work out these odds.

which we expected the true average to lie. Then we should have concluded that the true mean fell somewhere between 24.3 and 30.7 bushels, which would have been wide enough to include the real mean. Of course, if we had taken four times the standard error, we should have been almost absolutely certain of including the true mean in the stated range, with only one chance in over 1,000 of being wrong.

In most statistical work, three times the standard error is taken as the greatest extent to which a given observed constant is likely to miss the true value for the universe. Even though there is about one chance in 370 of being further off than this with samples of 30 or more, most scientists are willing to take the chance that their sample is not that one exceptional case. For exceedingly important work, or where absolute accuracy of comparison is essential, even four times the standard error might be used; but for the general run of statistical problems, and with fair-sized samples, it would seem safe to regard three times the standard error as about the largest extent to which the conclusions might be out *solely because of the chances of getting* an *unusual sample* in random sampling.

In view of the possibility of the standard error itself being in error, however, the number of observations should always be stated, as well as the standard error of the constant, particularly where the sample is small.

Bias in sampling. The figure as to standard error tells nothing at all of how much error there may be because of *bias* in sampling. Thus, if in taking our sample of 20 farms, we had visited only the largest farms with the most prosperous-looking buildings, we should be very likely to get a sample which was not representative of *all* the farmers in the county, but simply of the better ones, and so might get an average yield, say 10 bushels, above the true average for the county. Even if we only selected our farmers to the extent of including those which were most willing to give us the figures we wanted, we might have a badly biased sample, as usually the best farmers and the most intelligent ones are most willing to answer such questions. We must depend largely on common sense and on other knowledge of the situation we are studying, and not on statistical computations, to tell us whether or not our sample is really representative of the universe we want to study. Thus we might compare the average size or value of the farms in our sample with the averages for all the farms in the county, as shown by the census reports, to see whether they were representative or not. All that the computed standard error can tell us is about how closely it is likely to approach the average (or other characteristic) *of the group it does actually represent*—whether that group is the one we meant it to represent or only a part of that group. This caution must always be kept in mind in using samples: Computed standard errors tell us how far our results may be off solely because of the chance of getting a poor sample with a limited number of cases; but they do not tell us how far we may be off because of a *biased* sample, which is not a fair selection from the universe we wish to study.

Deciding on the size of sample necessary to obtain a stated reliability. One other application of the standard-error formula remains to be mentioned. The way in which this formula can be used to estimate the reliability of the average from a given sample, when the number of cases is known, has already been explained. The same formula can be used to determine how large a sample would have to be taken in order to secure results within any reasonable assigned limits of accuracy.

Thus it has already been shown that the records from 20 farms could be used to say that the true average yield lay somewhere between 27.39 and 32.61 bushels, with about one chance in 135 of that statement's being wrong. How many farms would one have to visit to state the same average yield to within one bushel, with the same chance of the statement's being wrong? The same formula which was used to determine the standard error of the average can be turned around to answer this question also.

If we know that we want to get an average reliable to within one bushel, for a range of three times its standard error, then we know that the standard error of that average would have to be only one-third of a bushel. We may also assume that when we take our larger sample, the standard deviation of the yields on the individual farms will be found to be not very different from what it was in our sample of 20 cases, and so use the same standard deviation as we did before.

Taking the relation which was used in computing the standard error before, we have:

$$\sigma_M = \frac{\bar{\sigma}_x}{\sqrt{n}}$$

In the new case we have the required standard error given, ⅓ bushel; we are assuming that the estimated standard deviation for the universe from our larger sample will be 3.89 bushels, just as it was from our sample of 20 cases. Substituting these values in our equation, and using n'' to represent the number of cases required in the new sample, we then have

$$\tfrac{1}{3} \text{ bushel} = \frac{3.89 \text{ bushels}}{\sqrt{n''}}$$

When the terms are shifted around, this becomes

$$\sqrt{n''} = \frac{3.89 \text{ bushels}}{\tfrac{1}{3} \text{ bushel}} = 11.67$$

Hence

$$n'' = 136.2$$

We therefore conclude that if a sample of 136 reports were obtained, we should probably get an average yield which would not differ from the true average yield for all the farms by more than one bushel in more than one such sample out of several hundreds of such samples. If any other limit of error was set, we could similarly determine how many reports would probably be necessary to satisfy that limit.

In these computations we have ignored the standard error of the standard error. If we took into account the possibility that the true standard error might be larger than our computed standard error, we should need a still larger sample to be sure of the accuracy specified.

Standard errors for other measures. This whole discussion has been in terms of determining how closely it was possible to approximate the *true average* from the *average shown by a sample.* In exactly the same way standard-error formulas have been worked out indicating how closely it is possible to approximate the true values of other statistical measures (such as standard deviations, for example) from the values for those measures determined from a sample[11] These are interpreted in much the same way as are the standard errors of averages; they will be referred to in subsequent chapters.

Universes, Past and Present

Any statistical measurement relates to something that is already past by the time the measurement can be analyzed. Thus our records

[11] The standard error of a standard deviation (σ_σ) may be approximately determined by the formula

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{2(n-1)}}$$

of the yield of corn obtained must relate to some crop that has already been harvested. Yields for a crop still growing could only be forecasts, and could never be precisely accurate until the crop was harvested and was weighed or measured. Yet human beings cannot live in the past. Our measurements of past events can be of meaning to us only when we project them into the future, and use them as a guide to future conduct. In studying the yield of corn, for example, the actual realized yield of corn in a county in a given year, no matter how accurately measured, is already a matter of history. The only thing that can be significant in human affairs is the average yield in some future year, still to be produced. If we are planning an A.A.A. control program, for example, and wish to estimate how many acres will produce a given total bushelage, we shall always be dealing with future years. We can do nothing to change the past. Only the future can be affected by our actions. When we take the average yield for a past year as our "universe" to be studied, what we are really interested in knowing is usually something about the yield most likely to be secured in one or in a series of years in the future. Even if we took a census of the yield on all the farms in the county, we should not have all the facts about our true universe. That universe, whose values we really wish to estimate, is composed of the yields next year and in other years still to come. Measurements of conditions in the past, no matter how accurately made, can serve only as one part of the basis for judging what the values in the future are likely to be. Analysis of what has happened in a succession of years in the past may help us to make a better estimate of the future. Such analysis may show a steady upward trend, or a variation from year to year with rainfall, or other variations whose cause we do not know. But before we can project the past trends into the future, we must understand what caused them, and judge whether those causes will continue to operate. These judgments are not a matter of statistical analysis as such but must be based upon scientific and technological study of all the forces at work. Thus a steady upward trend in cotton yields might reflect a rising price of cotton in the period studied, and a resulting increase in the quantities of fertilizer applied per acre. But equally well it might reflect a steady decrease in the total acreage (due to crop control or other causes) and a concentration of the remaining acreage on the better lands. Or it might reflect the gradual adoption of improved strains. A forecast of whether the upward trend would continue into the future would be materially different in the three cases. Besides the statistical facts, it would involve non-statistical judgments as to whether the increase in price or the limitation of acreage or the improvement in seed was likely to continue.

Whether we are dealing with the statistical characteristics of people or of crops or of prices or of atoms, the real universe for which we wish to estimate is the universe of future events. Our ability to forecast those events will differ widely from field to field. Presumably the characteristics of atoms or of chemical compounds will be less subject to change than will those of crops, and crops will be less subject to unpredictable change than will prices. In each case, however, the statistical information gained from the study of past samples must be tempered by other knowledge of the situation, based on study and analysis which may be quite non-statistical in nature. When we move from the facts of the past to forecast the unknown universe of the future it is not the statistics but the statistician who is on trial. Unless he mixes an ample measure of anthropology or agronomy or economics or other appropriate scientific information with his statistics —plus a liberal dash of common sense—he may find his analysis of past events a detriment, rather than an aid, in judging as to the future.

Summary. This chapter considers the question of how far statistical results derived from a selected "sample" drawn from a universe can be used to reach general conclusions as to the facts of the entire universe.

The confidence which can be placed in any measure computed from a sample, say an average, depends upon how closely that average is likely to come to the true average of the whole universe. One way of determining that would be to collect additional samples, each of the same size. From the way the averages from each of these different samples varied one could judge how near the average from any one sample was likely to come to the true average. For samples which meet the conditions of simple sampling, another much more rapid way is to compute the *standard error* of the average, which indicates the minimum extent to which the average is likely to be correct. With samples of over 30 cases, the true average will probably be within twice the standard error from the observed average for 19 samples out of 20, and within three times the standard error 369 times out of 370. This is the minimum error; where the number of observations is smaller, the possibility of error is larger, as is indicated by Tables A and B.

The same formula can be used to estimate how large a sample must be taken to secure any desired degree of accuracy in the final average.

The estimated standard error does not take into account bias in selecting the sample, but only shows the chances of reaching incorrect results even when an honest random sample is obtained.

Even after the values in the universe have been estimated from the facts shown by the sample, the statistician must still remember that that universe is a past universe. In applying that knowledge to problems of future action, he must give due allowance to the fact that the yet unborn universe of the future may never be identical with the past and dead universe from which his sample was obtained.

CHAPTER 3

THE RELATION BETWEEN TWO VARIABLES, AND THE IDEA OF FUNCTION

Relations are the fundamental stuff out of which all science is built. To say that a given piece of metal weighs so many pounds is to state a *relationship*. The weight simply means that there is a certain relationship between the pull of gravity on that piece of metal and the pull on another piece which has been named the "pound." We can tell what our "pound" is only by defining it in terms of still other units, or by comparing it to a master lump of metal carefully sheltered in the Bureau of Standards. If the pull is twice as great on the given piece of metal as it is on the standard pound, then we say that the lump weighs 2 pounds. If, further, we say it weighs 2 pounds per cubic inch, that is stating a composite relationship, involving at the same time the arbitrary units which we use to measure extent or distance in space and the units for measuring the gravitational force or attracting power of the earth.

Simple relation

Relations between variables. Besides these very simple relationships which are implicit in all our statements of numerical description—weight, length, temperature, size, age, and so on—there are more complicated relationships where two or more variables are concerned. A variable is any numerical value which can assume varying or different values in successive individual cases. The yield of corn on different farms is a variable, since it may differ widely from farm to farm. So is the length of time which a falling body takes to reach the earth, or the quantity of sugar that can be dissolved in a glass of water, or the distance it takes for an automobile to stop after the brakes are applied, or the quantity of milk that one cow will produce in a year, or the profit that a farm will pay in a year, or the length of time it takes a person to memorize a quotation. In contrast to these *variables* there are other numerical values called *constants*, because they never change. Thus one foot *always* contains 12 inches; one dollar *always* is equal to 100 cents; and a stone *always* falls 16 feet in the first second (under certain specified conditions). Science, of any sort, ultimately deals with the relation between variable factors and with the determination, where possible, of the constants which describe exactly what those relationships are.

The variables which have been mentioned may be used to illustrate the way in which changes in one variable can be related to changes in another. Thus the length of time which a falling body takes to reach the earth varies with—that is, is related to—the distance through which the body has to fall. etc The quantity of sugar which can be dissolved in a glass of water varies both with the size of the glass and the temperature of the water. The distance it takes for an automobile to stop after the brakes are applied varies with the speed with which the car is traveling when the brakes are applied, the area of braking surface on the drums, the area of tire surface on the road, how tightly the brakes are applied, how much the car weighs, the kind of road, and so on.

multiple factors

Imp.

Still more complicated relations

Then when we come to variables like the production of milk or the income on a given farm, or the time to memorize a quotation, we find the situation still more complicated. How much milk a cow will produce varies with her age, breed, inherent ability, and the richness of the milk, and with the kind, quality, amount, and composition of the feed she receives, the way she is stabled and cared for, and many other similar factors. Similarly the variables which may affect the income on a farm—the size, the equipment, the crops grown, the livestock kept, the methods followed, the costs paid, the prices received, the rainfall—are so numerous that it would take an entire book merely to list and discuss the different factors affecting this one single variable. The time it takes to memorize a quotation may be affected by its length, the subject's age, sex, training, fatigue or freshness, his familiarity with the material discussed, and his interest in the topic.

18

Yet it is precisely with relations between complex variables that many statistical studies must deal. The statistical methods which may be used to handle such problems can best be understood if presented first for the simplest cases, and then expanded to cover the more complicated ones.

Ex. 1 (physics)

Suppose a physicist, knowing nothing about the exact nature of the relation between the distance a body has to fall and the length of time it takes, made some experiments to determine the matter and obtained the results shown in Table 9.

Looking over these figures we see that there is some sort of general relation between the two. As the distance increases, the time increases also. But that is not uniformly true. In one case the distance increased without there being any increase in the recorded time; in some other cases the recorded time was not the same even though the distance was unchanged.

TABLE 9

Relation between Distance a Marble Drops and Time It Takes to Fall

Distance traveled	Time elapsed	Distance traveled	Time elapsed
Feet	*Seconds*	*Feet*	*Seconds*
5	0.6	20	1.1
5	0.5	20	1.1
5	0.6	20	1.2
10	0.9	20	1.1
10	0.8	25	1.2
10	0.7	25	1.3
15	1.0	25	1.2
15	0.9	25	1.3
15	1.0		

average .57+

average time elapsed .8

and so on

Graphic representation of relation between two variables. We can get a better idea of just exactly what the relation is if we "plot" it on cross-section paper, so that we can see graphically just how the time does change with the distance. Figure 2 illustrates the way this is usually done. The units of one variable, in this case the distance to be traversed, are measured off from the left, starting with zero in the lower left-hand corner and counting over toward the right. The

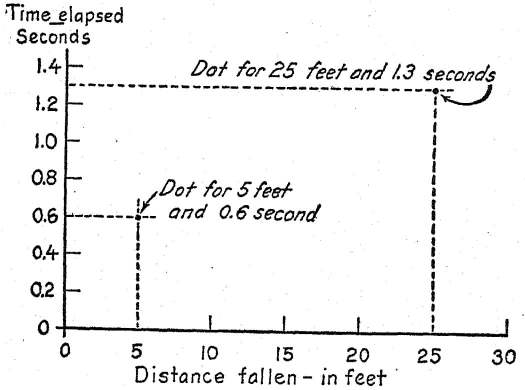


FIG. 2. Method of constructing a dot chart. Time elapsed is the dependent variable, and the distance is the independent variable.

units of the other variable, in this case the time elapsed, are measured off from the bottom, starting with zero and counting up toward the top. If negative values are present, then the counting is started with the *largest negative* value, decreasing from left to right or from bottom to top, until zero is reached and the positive values begin to appear.

Where one variable may be regarded as the cause and the other variable as the result, it is customary to put the causal variable along the bottom. In this case it may be said that the differences in distance traversed cause the differences in time elapsed. Distance, therefore, is measured in the horizontal direction, and time in the vertical. There is no particular reason for plotting data just this way except that this is the customary way of doing it and so it is most readily understood by other persons. Some relations of this sort can be reversed, so that either may be regarded as cause and either as effect.[1]

Having laid off the chart in the way indicated, we next "plot" the individual observations. The way this is done is illustrated in Figure 2. The first observation was that it took 0.6 second for the marble to fall 5 feet. This is indicated on the chart by counting over to the 5-foot line from the left of the chart, and then counting up along that line until 0.6 second is reached. A dot is placed on the chart at that point. As indicated, this dot is at the *intersection* of the line starting from the "0.6 second" at the left of the chart and extending parallel to the "0-second" line, with the other line starting from "5 feet" at the bottom of the chart and extending parallel to the "0-foot" line. Similarly, the last observation, 25 feet in 1.3 seconds, is indicated by a dot where the horizontal line representing 1.3 seconds crosses the vertical line representing 25 feet.

Entering a dot for each individual observation in the same way, we get the chart shown in Figure 3. This figure now gives a visual representation of the way in which the length of time changes as the distance traversed changes. Such a chart is known as a "dot chart" or a "scatter diagram."

But even this figure does not show the *exact* relation between the distance and the time. Both the first and the second trials were for exactly the same distance, yet the time was slightly different. Obviously that difference in time could not have been due to the difference in distance between the two, because there was no difference. The investigator must therefore assume that some outside cause, perhaps the accuracy with which the time was measured, may have been

[1] For a more extended discussion of this point, see pp. 50 and 51.

responsible for these slight differences. It will be noted, too, that when the different observations are plotted as in Figure 3, they come close to all lying along a continuous curve. We also see that the individual cases do not adhere absolutely to a continuous curve. ✓If we are willing to assume that all the differences between the different observations at the same point along the curve are due solely to extraneous factors, we can estimate the true effect of the distance, by itself, by averaging together the several observations as to time taken for each of the several tests for the same length of fall. A

Av. table p. 36 6

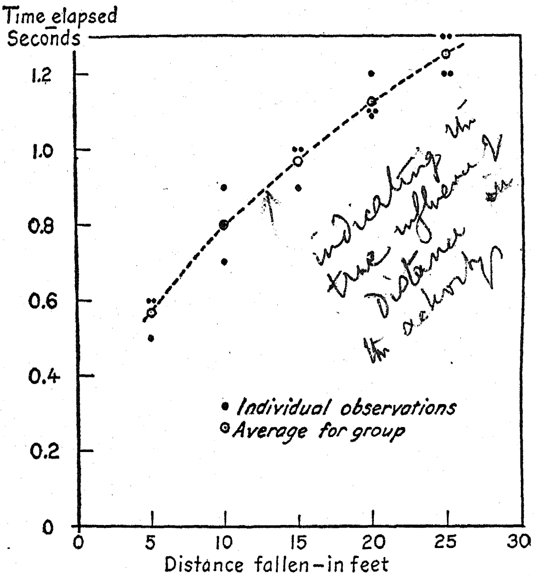


FIG. 3. Relation of distance a marble falls to time elapsed in falling, as shown by individual observations and curve of average time.

continuous curve drawn through these averages would then indicate the way in which the duration of fall varied with the distance, *on the average* of the cases studied. Although it might not hold true for any one individual case, as we have just seen, still it does indicate *about* what the time will be. For practical purposes we may say that under given conditions the time a body takes to fall *is determined* by the distance which it has to fall.

The average time for each distance is indicated by the small circles in Figure 3. It is evident that all these averages lie very close to the smooth freehand curve which has been drawn on the chart.

Expressing a functional relation mathematically. The relation shown by the curve in Figure 3 is what mathematicians call a *functional* relationship; the time it takes a body to fall *is a function* of the distance which it has to traverse.[2] All that this means is that for any particular distance-fallen, there is some corresponding time-required. The term "function" means that there is *some* definite relation between the two variables, number of feet and number of seconds, but it does not at all tell just *what* that relationship is. When, however, it is said that time is a function of distance according *to the curve shown in the figure,* then the statement has been made perfectly definite. The curve shows, for any given distance, exactly how long it will take a body to fall, on the average of a series of trials.

$y = f(x)$

In this particular case the function is defined only by the graphic curve. It may also be stated as a mathematical expression

Law of gravity → $$Y = \tfrac{1}{4}\sqrt{X}$$ → an equation relating time & Distance for a falling body

using X for distance in feet and Y for time in seconds. This equation corresponds to the curve in a peculiar way, in that if any value of X is substituted in it, and then the value of Y determined, that will be the value of Y—time in seconds—corresponding to that particular value of X—distance in feet—as shown by the curve in Figure 3. This equation is therefore *the equation of the function,* since this simple mathematical expression tells just as much about the relation between the two varying quantities—time and distance—as does the entire curve in the figure.

The way this equation is used may be illustrated by two examples. Suppose a marble falls 16 feet; how long should it take to fall? The value of X would then be 16; substituting this value in the equation, we have

$$Y = \tfrac{1}{4}\sqrt{16}$$
$$Y = \tfrac{1}{4}(4)$$
$$Y = 1$$

This gives a value of 1 for Y, which means that it would take 1 second to fall. Suppose again a bomb were dropped from an airplane

[2] Using Y for time and X for distance, we state this mathematically

$$y = f(X)$$

10,000 feet high. How long would it take to reach earth? The value of X is then 10,000; substituting this value in the equation, we have

$$Y = \tfrac{1}{4}\sqrt{10{,}000}$$
$$Y = \tfrac{1}{4}(100)$$
$$Y = 25$$

The result $Y = 25$ means that it would take 25 seconds for the bomb to fall.[3]

It is evident that the equation goes much further than does the graph of the curve. The latter gives the relation between distance and time only for the distances which are shown on the chart. The equation, on the other hand, gives the relation for any distance whatever, no matter what it may be. It is possible to state this *law of gravity*, as it is called, in an equation only because physicists have studied this relation in the past and determined exactly how the one quantity varies with the other. Having found that the same relation between the two variables held through their entire range of observation and having worked out on philosophical grounds a good reason why that relation should hold, they have felt safe in coming to the conclusion that it will continue to hold even beyond the range of the experimental verification.[4] Where only a graph of the function is available, on the contrary, only the relation within the stated range is known. The graph does not tell, of and by itself, the direction the curve would take if extended beyond the limits determined by the experiments.

P.+.2

(Chemistry)

Now if instead of the relation we have just been discussing we consider the relation between the quantity of sugar which can be dissolved in a glassful of water and the temperature of the water, we

Always note conditions

[3] Outside causes, such as friction with the air, may make the time of fall slightly different from the calculated time; therefore with so long a fall as this the time might differ quite perceptibly from the theoretical time given by the equation. This equation gives the time required *when no influence other than gravity* is taken into account. Obviously a marble would fall in air much faster than a feather—the resistance of the air has very little influence on the speed of the marble and a great deal of influence on the speed of the feather. In a vacuum they would fall at the same rate.

[4] It should be noted that for very great distances—say 10,000 miles—the formula might need to be modified, since then the pull of the earth would be less than it is at the surface. The equation ~~holds true~~ only for those distances from the earth within which its pull is practically a constant.

have quite a different problem, and yet one that is similar in many aspects. If we start to determine it experimentally, we must first make sure that the quantity of water with which we are working is the same in every trial; then we must measure accurately both the temperature of the water and the amount of sugar which could be dissolved in it. Water expands when it is heated, and it also has a tendency to evaporate; so we would have to decide whether we wanted the *same volume* of water, irrespective of the fact that at a higher temperature there would be actually *less* water in that volume, or whether we wanted the *volume* of water equivalent to what would be the same volume at a given fixed temperature. (This would necessitate determining the relation between volume and temperature for a given weight of water as a preliminary study, or else using weight instead of volume as our criterion.) Many other similar factors which might possibly influence the result would have to be considered before even the exact plan of the experiment could be drawn up.

conditions of experiment

Once the experiment had been run the numerical results would probably be somewhat similar in character to those in the gravity test. It would be found that *about the same* quantity of sugar was dissolved in a given quantity of water when repeated tests were made at the same temperature, but that the quantities varied slightly from each other. If the data were plotted on a scatter diagram like Figure 3, it would be found that the data fell in the general shape of a curve, but that very few of the dots fell exactly on the curve, some lying above and some below the continuous line which could be drawn about through the center of them. Again we might conclude that these slight differences from exact agreement were due to factors other than the temperature of the water—to slight experimental errors in the quantity or temperature of the water, or to slight errors of measurement in determining the quantity of sugar—and be willing to conclude that the line drawn through the center of the series of observations showed the *real* effect of differences in temperature on the quantity of sugar dissolved, when extraneous influences were removed. This again would be a *functional* relation. The curve would express the relation between changes in temperature and changes in quantity of sugar, showing for any given temperature exactly how much sugar could be dissolved. It might then be possible to determine a type of equation which would accurately specify the function by a mathematical formula, similar ~~to that~~ discussed for the gravity example, if the logical type of relation between the two variables could be worked out.[5]

Determining a functional relation statistically. In the two cases which have been discussed the relation between the two variables was sufficiently close so that by taking proper experimental precautions other influences which might affect the result could be largely removed and a series of observations obtained sufficiently consistent with each other so that the exact nature of the relation could be readily determined. In many other types of relations this cannot be done so easily. It is with this type of relation that statistical methods really become important.

Ex. 3. Can apply to Social Sciences

If we were making a traffic study in a given city, for example, we might wish to know what would be the safe speed limits to permit on different streets. In that connection we might need to know in what distance an automobile could be stopped when traveling at different speeds, so that by comparing this distance with the width of the different streets and the length of view at intersections we could judge how fast machines might be able to travel without risk of collisions at street intersections. One way to determine what is the relation between speed and stopping distance would be to make a number of tests in different portions of the city, taking different types of machines and different drivers. Let us suppose that as the result of such a series of tests we obtained the series of observations shown in Table 10.

[5] Some logical foundation *is* needed before a mathematical equation to a curve can be of any more value than merely the chart which graphs the curve. Thus in the gravity example it is evident that the farther a body falls, the faster it falls; in every successive instant the speed it has already attained is increased by the effect of the continued pull which is added to it. Purely mathematical investigations of the relation between such constantly growing magnitudes and the variable with which they grow have enabled physicists to determine the general mathematical *type* to which the relation must conform. Then, knowing what the type of the curve is, we find it to be relatively easy to determine the constants (such as the "$\frac{1}{4}$" of the equation $Y = \frac{1}{4}\sqrt{X}$) which makes the general equation applicable to a given specific case. This is done by using experimental results, such as those given in Table 9, to calculate the constants for the specific type of curve which has been determined upon.

Not all functional relations can be subjected to this type of logical analysis, however, and it is sometimes impossible to tell what sort of equation the results should really follow. In that case any mathematical curve "fitted" to the data has no more special meaning than the graphic curve drawn through the center of the observations; both are merely empirical descriptions of the relations, and both are limited in their interpretation to the range of the particular data upon which they are based. This fact will be discussed more fully later on.

It is apparent from the table that there are great variations in the distances which different cars or different drivers required to stop, even when traveling at the same speed. This is shown even more clearly when we make a dot chart of the data in just the same way as illustrated in Figure 3. The graphic comparison between speed

TABLE 10

RELATION BETWEEN SPEED OF AUTOMOBILE AND DISTANCE TO STOP AFTER SIGNAL, AS SHOWN BY 50 INDIVIDUAL OBSERVATIONS

Speed when signal is given	Distance traveled after signal before stopping*	Speed when signal is given	Distance traveled after signal before stopping*
Miles per hour	*Feet*	*Miles per hour*	*Feet*
4	2	19	46
7	4	24	93
17	50	14	26
14	36	12	28
12	20	9	10
11	28	10	34
20	48	15	20
15	54	24	70
17	40	25	85
13	34	20	64
15	26	19	36
19	68	13	26
10	26	10	18
18	56	7	22
22	66	16	40
18	84	14	60
8	16	20	52
4	10	24	120
12	14	24	92
20	56	17	32
23	54	13	34
18	76	11	17
12	24	13	46
16	32	14	80
18	42	20	32

* These observations were made before 4-wheel brakes were common.

and distance-to-stop, shown in Figure 4, reveals that there is only a general agreement between the different tests. There is certainly some relation between the two variables, but it is vague and uncertain in comparison with the relatively sharp and clear-cut relations shown in Figure 3.

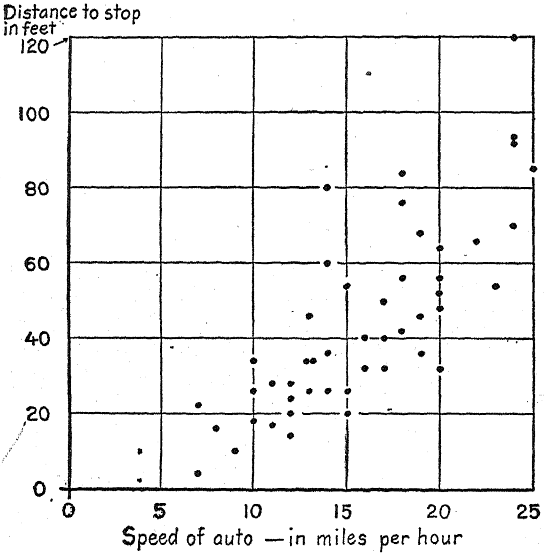


FIG. 4. Relation of speed of automobile to distance it takes to stop, as shown by individual observations.

M. Cee

There is no particular difficulty in understanding why the relation is not more definite. The data represent a great variety of different elements—cars with two-wheel brakes and cars with four-wheel brakes; cars with brakes in adjustment and cars with brakes well worn; cars nearly empty and cars heavily loaded; cars with balloon tires and cars with high-pressure tires. In addition, the drivers differ. Some are experienced drivers, some inexperienced; some strong and some unable to press the brakes fully down; some with almost instantaneous reaction to our signal to stop, some with faltering or lagging response; some bright and wide awake, others tired and unobservant; some calm and steady, others nervous and erratic. Finally the conditions of the tests might be different—some on concrete pavement, others on asphalt; some on up-grades, some downhill.

explaining variation

There are two different ways by which we might go about deciding exactly what these varying observations showed. (1) One way would be to divide up the data so that the effect of some of the different factors to isolate cases

mentioned would be removed from the results. Thus if we separated the observations into different groups according to the make of car, and then reported each of these groups according to the model or the year made, the relation between speed and distance for any single group would no longer be affected by differences in braking equipment so far as engineering design went. Most of the remaining factors, however, would still be present to affect the results, so that even within each subdivision the records would still show great diversity in the relation. Only if we continued the process of subdivision of our sample until we got down to successive observations of a single car operated by a single driver at the same place, would we be likely to get observations as consistent with each other as those in the previous physical and chemical illustrations. Differences in the promptness with which the driver responded to the signal, in the preciseness with which the speed at the moment of giving the signal was observed, and possibly in the force with which the driver applied his brakes, all might influence the result, so that even then the results might be less consistent—"the curve be less definitely defined"—than in a series of laboratory experiments where *all* the important outside variables could be definitely controlled and so prevented from affecting the results obtained.

Should the entire mass of observations be analyzed as suggested, that would give a great number of different sets of relations, each one showing how long it took a given car to stop when driven by a given driver, when traveling at different speeds. But this great number of different curves might not be suitable to answer our question. They might be so different from curve to curve that it might seem that there was no real general relation between speed and distance. A new car, with four-wheel brakes, driven by an experienced driver, might stop in its own length at the same speed at which an old car, with brakes nearly worn out, and driven by an inexpert driver, might require a hundred feet or more. Obviously neither one of these extremes would be typical of the general relation; but what would be typical? Even the less extreme cases might show great variations among themselves, so that it would be almost impossible to pick from the great diversity of curves one or a few that would serve as a basis of judgment for our problem.

2 A second way of going about it would be to try to determine some sort of *average* relation between speed and distance. In that case we should admit that there were great differences from the average in individual cases, yet should feel that the average would serve as a

general indication of what the relation was, even though we were aware it would not be true in every, or perhaps even in any, individual case. If we knew *nothing* about a car except the speed at which it was moving, that average relation, however, would serve to give us the best guess we could make as to how far it would take it to stop. Since we should have to make our speed limits the same for all passenger cars, that might give us the best basis of judgment as to how high it was safe to place it. Of course we should also need to know something about how much *more* than the average time exceptional cars or drivers might require and how far above the average any large proportion of them fell, so as to decide how much leeway to allow; but even so, the average relation would be the first interest and the point of departure in reaching our decision.

Where the relation between two variables is clear and reasonably sharply defined, as in the experimental case discussed, it is not difficult to determine the average relationship, since the relation for individual cases and the average relation for all cases are nearly identical. Where the relation is not so well defined, however, and where many other relations are involved in addition to the particular one which is being studied, it is by no means so easy to determine exactly what the true relationship is. A considerable body of statistical methods has therefore been developed to treat this particular problem. Since this problem pertains to the relation between variables, it has become known as the problem of co-relation, or "correlation." Just how statistical technique may be applied to the solution of the traffic problem which has just been presented will be considered in detail in the next chapter.

Summary. A statement of the change in one variable which accompanies specified changes in another is known as a statement of a *functional relation.* A functional relation may be stated either graphically by a curve or algebraically by a definite equation. Although functional relations may be readily determined from experimental conclusions where all influences except the one being studied are held constant, many problems cannot be studied by such methods. The statistical methods of *correlation analysis* may be used to study functional relations where experimental methods are not satisfactory.

CHAPTER 4

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN ANOTHER CHANGES: (1) BY THE USE OF AVERAGES

classification Data

The problem stated in the previous chapter was to determine how many feet automobiles traveling at a given speed require to stop. It involves determining the *average* extent to which one variable changes when another variable changes. Stated mathematically, the problem is to find the functional relation between speed and distance—the probable distance required to stop with any given initial speed. Of the many different ways of doing this, the simplest, and the one which would suggest itself most naturally, would be to classify ~~the records~~ into groups, placing all of one speed in one group, all of another speed in another group, making as many groups as there are different rates of speed recorded, and then *averaging* the different distances for all the cases in each group. This would then give an average distance to stop for each given rate of speed in the series of records. Table 11 shows this operation carried out. Table 11

Where there were only single observations, this fact has been indicated by placing the average—the single report—in parentheses.

The averages in the last column of Table 11 show quite specifically how the distance required to stop tends to increase with the speed a machine is traveling. The machines which were tested at 12 miles per hour stopped at an average distance of 21.5 feet, those at 15 miles per hour at 33.3 feet on the average, and those at 20 miles per hour at 50.4 feet. But the increase is not uniform. The cars at 10 miles per hour averaged a greater distance than those at either 11 or 12, and the cars at 19, a shorter distance than those at 18. from Table 11

Plotting

If the successive averages from Table 11 are plotted and connected by lines, both the general increasing tendency and the irregular change from group to group are easily seen. Figure 5 shows this comparison (see page 49).

Statement of the problem

Do these differences between the different group averages have any real significance? Is there any reason to think that this very jagged line is the *true* average relation between speed and distance? We can consider that from two points of view; the logic of the relation and the statistical basis of the differences. Logically the differences are quite nonsensical. If a given machine can stop in 22 feet when it is going 11 miles an hour, of course it can stop in at least the same distance when going 10 miles per hour, and probably something less.

Two tests in the zigzag
(1) Logically nonsensical

TABLE 11

Computation of Average Distance to Stop after Signal, for Different Initial Speeds

Speed when signal is given	Different distances noted for that speed*	Average distance for that speed
Miles per hour	*Feet*	*Feet*
4	2, 10	6.0
✓7	4, 22	13.0
✓8	16	(16)
✓9	10	(10)
10	26, 34, 18	26.0
11	28, 17	22.5
12	20, 24, 28, 14	21.5
13	34, 26, 34, 46	35.0
14	36, 26, 60, 80 ✓	50.5
15	54, 26, 20	33.3
16	32, 40	36.0
17	50, 40, 32	40.7
18	56, 84, 76, 42	64.5
19	68, 46, 36	50.0
20	48, 56, 64, 52, 32	50.4
22	66	(66)
23	54	(54)
24	93, 70, 120, 92	93.75
25	85	(85)

* Data taken from Table 10.

52
4

3 12

17|451

17
32
34
51
544

It certainly would not take 26 feet, as the table shows. ✓ Then from the statistical point of view the groups are entirely too small to show very definitely how far on the average it takes to stop at *any* one speed. Even the largest group, at 20 miles per hour, has only 5 cases, whereas we have seen in Chapter 2 that 10 to 25 cases may be required as a minimum to give an average of much reliability. Computing the standard error for the average from the 20-mile group of reports, it comes out 5.3 feet. With only 5 reports, however, Figure A (in Ap-

(2) Statistically insignificant

pendix 3) shows that we have to take a range of 1.1 times the standard error to make the observed value come within that range of the true value in 2 samples out of 3. We may say that the standard error of the average, taking this into account, is 5.83 feet.[1] The average for this group of records may therefore be written 50.4 ± 5.8 feet. When we say that the average distance required to stop when traveling 20 miles per hour (for all automobiles in town, say) is between 44.6 feet and 56.2 feet, we are likely to be wrong in 1 out of 3 such statements, on the average. With the average from the *largest* group showing as little reliability as this, it is quite clear that the zigzag variation from average to average has no real meaning. So few cases are included in each group that the averages are not statistically reliable to anything like the individual differences. All the irregular differences from group to group can therefore be accounted for by purely chance variations in sampling. It is quite possible that they are due solely to the small number of cases. As they have no statistical significance there is therefore no need to be worried about them.

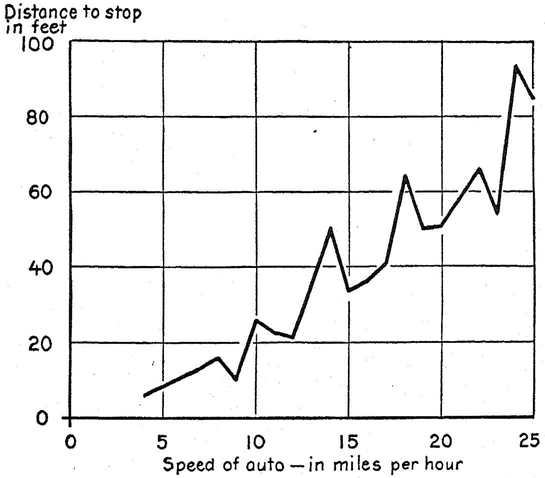


FIG. 5. Relation of speed of automobile to distance it takes to stop, as shown by averages of small groups.

Does that mean that in spite of the relationship we can *see* in

[1] The standard error is computed from the standard deviation of the five reports at 20 miles, using equation (7.1). This gives a value of 5.3. Figure A, in Appendix 3, shows that for five reports a range of 1.1 times the computed standard error must be taken to secure a reliability of .67 (or probability of .33 for the specified departure), so the final standard error is (5.3) (1.1), or 5.83.

Conclusion

Figure 5 that we can get no accurate statistical measurement of the relation? That is overstating the case a little; all that we have determined so far is that the line of averages, the irregular function shown in Figure 5, has but little statistical meaning, *just as it stands now.*

two alternatives (1) having a larger sample

We might be able to make the results more accurate by basing our averages on a larger number of reports. As we have seen previously, the more cases there are in a group the more reliable the average of that group is likely to be. One way of doing ~~that would~~ be to go out and get more records, so that we should have enough cases in each group to make the averages reliable within small enough limits to suit our needs. But that would be a long and expensive process. Isn't there some way we can find out something more just from the records we have?

(2) Re-grouping data

Another way of making the conclusions more stable would be by combining the records so as to give fewer groups, but with more cases in each group. So far we have been working with 19 different groups, one for each of the 19 different speeds measured. If instead we group them into a few groups—say four or five—we shall have considerably larger groups to work with.

or Resultant

Independent and dependent variables. The question might be asked whether the groups should be made on the basis of the rate of speed or of the distance to stop. (In preparing Table 11 we used the rate of speed without discussing the matter.) That comes back to the question of what we really want to find out. Do we want to know the *average speed at which machines were traveling* when it took them, say, 20 feet to stop; or do we want to know the *average distance* machines took to stop when they are traveling at a given speed? Obviously, the thing we are going to set is the speed limit, and we are merely interested in the distances to stop as one factor to guide us in deciding what the speed limit should be. We therefore want to know the effect of *speed* upon *average distance,* and not the reverse. For that reason we shall classify our records on the basis of speed, and then average together all the different distances for the cars traveling at that speed.

independent or causal

The same question is met with in nearly all problems where the relation between two variables is to be dealt with. It is always necessary to think over the problem carefully, and decide which variable we are going to regard as the independent or *causal* variable, and which one as the dependent, or *resultant.* Thus if we were relating variations in tobacco yields to applications of fertilizer, obviously the differences in fertilizer would be the cause and the differences in yield the result, so we would sort our records according to the differences in fertilizer. Other relations may not be so clear cut. If the size of stores were being related to profits, it might be as logical in some situations to consider that the more successful men were able to afford the largest stores as to consider that the larger stores returned the greater profits. Careful consideration of the facts in each given case is necessary to clarify exactly what is the particular relation involved.

As shown later (pages 113 to 121 and 450 to 451), it is frequently impossible to say which variable is the cause and which is the effect. All that can be definitely established is that the two vary together. Yet one may wish to regard one variable as the one whose values are given or known. It is then called the *independent variable* and plotted as the abscissa. The second variable will then be regarded as the one whose values are to be related to, or estimated from, the values of the known variable. It is then called the *dependent variable,* since it is treated as *depending upon* the given values of the independent variable. It is sometimes desirable in particular problems to consider first one ~~variable as the~~ independent variable and then the other one as independent.

TABLE 12

Average Relation between Speed of Car and Distance to Stop, as Shown by Records Thrown into Groups

Speed when signal is given*	Number of reports	Average speed	Average distance, to stop
Miles per hour		*Miles per hour*	*Feet*
Under 4.5	2	(4.0)	(6.0)
4.5 to 9.5	4	7.8	13.0
9.5 to 14.5	17	12.2	32.4
14.5 to 19.5	15	17.1	46.8
19.5 and over	12	22.2	69.3

* 4.5 to 9.5 means 4.5 and up to, but not including, 9.5.

Groups of larger size. To return to our automobile problem. Since the speeds varied up to 25 miles per hour, and we have 50 reports to deal with, we might try breaking them up into 5 groups and see what kind of averages that will give us. Using groups covering a range of 5 miles per hour each, we can group the records and determine the averages for the 5 groups thus formed, getting the results shown in Table 12.

from Table 11

In this group, belong the speeds 7, 8 & 9 (Table 11)

Sum of noted distances in these speeds is 4+22+16+10 = 52 divided by No. of distances (i.e., 4, in this case) the average [illegible] distance [illegible] this group

?

How to calculate average speed

rounded figures

1040008

These averages can then be plotted and connected by straight lines, just as were the averages in Figure 5. In constructing Figure 6, which shows this process, it is necessary to use the average speed as well as the average distance-to-stop in locating each point. This is because each of the average distances, as shown in Table 12, represents not one speed, but several different speeds thrown together. If we wish to compare the average distances, it seems most sensible to compare them on the basis of the average of the speeds which they represent. The circles in Figure 6 represent the several group averages plotted this way. The first one is located at the intersection of the lines for 4.0 miles per hour and 6.0 feet; the second at 7.8 miles per hour and 13.0 feet; and so on for the remainder.

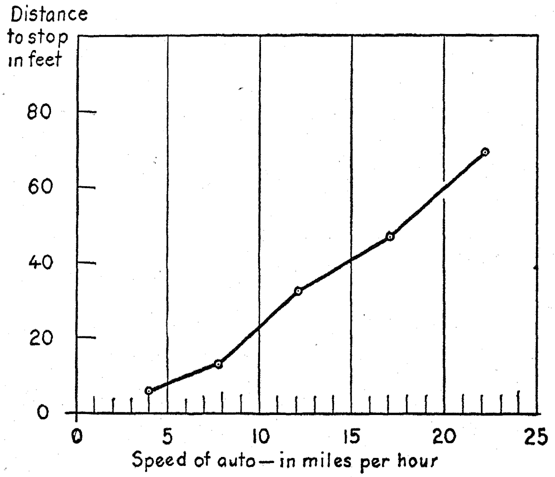


FIG. 6. Relation of speed of automobile to distance it takes to stop, as shown by averages of large groups.

In plotting the curve Cols. 3 & 4 of Table 12 have been used.

Results

(a)

When the group averages of Figure 6 are connected by straight lines the relation between speed and distance is shown much more satisfactorily than it was in Figure 5. The line in the new figure shows a continuous relation between speed and distance. It indicates that, when the averages are taken from groups large enough to eliminate the effect of individual cases, the higher the speed the greater the distance it takes to stop.

(b)

(1)

But on close examination even the relation shown in this last figure is not found fully satisfactory. If we compute the change in distance-to-stop for each change of 1 mile in speed, we find that the conclusions are somewhat erratic. Between the first two averages, the change in speed from 4.0 to 7.8 miles per hour, an increase of 3.8 miles per hour, is accompanied by a change in distance from 6.0 to 13.0 feet, or an increase of 7.0 feet. Between 4 and 7.8 miles per hour, therefore, the distance-to-stop apparently increases 1.8 feet for each increase of 1 mile per hour in the speed of the machine. Similar computations for all the other groups are shown in Table 13, carrying out just the same process.

see col 2 in Table 13

The results shown in Table 13 reveal that even the averages of Figure 6 are not altogether consistent. Between 4 and 8 miles per

~~(5 and 10)~~ rounded from 7.8

TABLE 13

Computation of Change in Distance for Each Change of One Mile in Speed, for Different Groups of Records

Speed when signal is given	Average speed	Average distance to stop	Increase in speed	Increase in distance	Increase in distance per 1 mile increase in speed
Miles per hour	*Miles per hour*	*Feet*	*Miles per hour*	*Feet*	*Feet*
Under 5	4.0 (=4)	6.0	✓	✓	
			3.8	7.0	1.8 ✓
5 to 10	7.8 (=8)	13.0			
			4.4	19.4	4.4 ✓
10 to 15	12.2 (=12)	32.4			
			4.9	14.4	2.9 ✓
15 to 20	17.1 (=17)	46.8			
			5.1	22.5	4.4 ✓
20 to 25	22.2 (=22)	69.3			

In effect these classes are same as in Table 12

hour they indicate that the distance-to-stop increases 1.8 feet for each increase of 1 mile in the speed of the machine; between 8 and 12 miles per hour the distance suddenly starts increasing 4.4 feet for each 1 mile per hour increase in the speed of the machine; then between 12 and 17 miles per hour the effect of further increase on the speed becomes less again, averaging only 2.9 feet increase in stopping distance for each increase of 1 mile per hour in speed; and then, finally, between 17 and 22 miles per hour changes again to 4.4 increase in feet to stop for each 1 mile increase in the speed of the auto.

rounded from 12.2

(11) This same variability in the rate of change can be seen directly from Figure 6 by noting the steepness of the several portions of the

line. Between 4 and 8 miles per hour, where there is the least average change in distance for each change in speed, the line has the least slope, that is, is the nearest horizontal. Between 8 and 12 miles, where the average distance to stop is much larger, the line tilts up abruptly; then between 12 and 17 miles per hour, where the average change in distance is less rapid, the line is flatter again, tilting up once more for the more rapid rate of change shown by the last group. It should be noted, too, that the slope of the line is almost exactly the same between the 7- and 12-mile averages, and the 17- and 22-mile averages, illustrating the fact that in both these intervals the increase in distance was the same for each mile-per-hour increase in speed. The irregular and zigzag character of the line in Figure 6 therefore shows the same vacillation in the group averages that the computations in Table 13 show. Simply by examining this chart closely it would have been possible to tell about this unsatisfactory character of the conclusions without taking the time to calculate out the exact rates.

Test of significance

Are the irregularities shown in Table 13 and Figure 6 of any significance statistically, or are they due simply to the possibilities of variation in using so small a sample, just as were the differences in Figure 5 and Table 11? Is it really true that an increase in speed has a larger effect upon the distance required to stop between 7 and 12 miles per hour than between 12 and 17?

Computing the standard error on average speed & distances

Reliability of group averages. The answer to these questions again involves a consideration of the statistical basis upon which our conclusions are based. These last results were calculated from the average speed and average distance for the several groups of records; obviously they can be no more reliable than are those averages themselves. In measuring the reliability of those averages by the methods we have already discussed, the thing to do is to compute the standard errors which will tell us about how much confidence we can have in each figure. That means that, by calculating these statistical constants, we can judge at least the *range within which* the true average may fall, in two samples out of three, provided the sample is a random sample.

The next step, therefore, is to calculate the standard error for each of the five averages of speed and the five averages of distance. The computation, which is exactly the same as that used before, based on equation (7.1), is shown in Table 14.

Comparing the several averages with their respective adjusted standard errors, as shown in the last column of Table 14, we find that there is not a great chance that if we made the same number of ob-

Conclusions

servations over again and used the same grouping, we should get averages different enough to change the location of the points materially. But with regard to the distance required to stop, the averages are much less reliable. If we collected enough records to determine

TABLE 14

COMPUTATION OF STANDARD ERRORS FOR THE AVERAGES SHOWN IN TABLE 12

Group	Number of cases, n	Standard deviation, σ	Computed standard error $\frac{\sigma}{\sqrt{n}}$	Range within which chances are $\frac{2}{3}$ that average will fall *	Average plus range for $\frac{2}{3}$ probability †
	For speed				
Miles per hour		*Miles per hour*	*Miles per hour*	*Miles per hour*	*Miles per hour*
Under 5	2	0	....	....	4.0 ± ?
5 to 10	4	0.83	0.48	0.58	7.8 ± 0.6
10 to 15	17	1.39	0.35	0.36	12.2 ± 0.4
15 to 20	15	1.41	0.38	0.40	17.1 ± 0.4
20 and over	12	1.95	0.59	0.62	22.2 ± 0.6
	For distance				
		Feet to stop	*Feet to stop*	*Feet to stop*	*Feet to stop*
Under 5	2	4.00	4.00	7.20	6.0 ± 7.2
5 to 10	4	6.71	3.87	4.68	13.0 ± 4.7
10 to 15	17	16.09	4.02	4.18	32.4 ± 4.2
15 to 20	15	17.62	4.71	4.90	46.8 ± 4.9
20 and over	12	23.25	7.00	7.35	69.3 ± 7.4

How to judge it to be statistically insignificant

* These values are obtained by adjusting the computed standard error to indicate the range for which the probability is only 0.33 that the true average lies outside. By interpolating in Figure A, Appendix 3, the necessary adjustments to be applied to the computed standard errors are found to be: for 2 observations, times 1.80; for 4, times 1.21; for 15 or 17, times 1.04; and for 12, times 1.05.

† In addition to the ranges shown here, there is a further margin of uncertainty due to the standard error of these estimated standard errors. It ranges from 71 per cent for the smallest group to 18 per cent for the largest.

the several averages quite accurately, there is one chance out of three that we might find that the true distance for the first group was practically nothing, or else more than 14 feet; or for the second group was less than 8 feet or more than 18 feet; and so on until for the last group it might be under 62 feet or over 77 feet.[2] With this wide pos-

probability

$1 - \frac{2}{3} = \frac{1}{3}$

see col 5 in Table 14

7.2

[2] If the standard errors of the estimated standard errors were also taken into account, the zones of uncertainty would be even wider.

sible variation in the true values, it is quite evident that the real facts have not yet been measured accurately enough to justify detailed computations of the differences in the slope of different portions of the line. By changing any one of the averages as much as has been indicated, the slope of the line would be very materially changed.

Plotting

Range within which true relation may fall. The extent to which reliance may be placed in the relationship between the two variables as shown by the 50 observations which we have to deal with may be judged from Figure 7. Here the actual averages have been plotted,

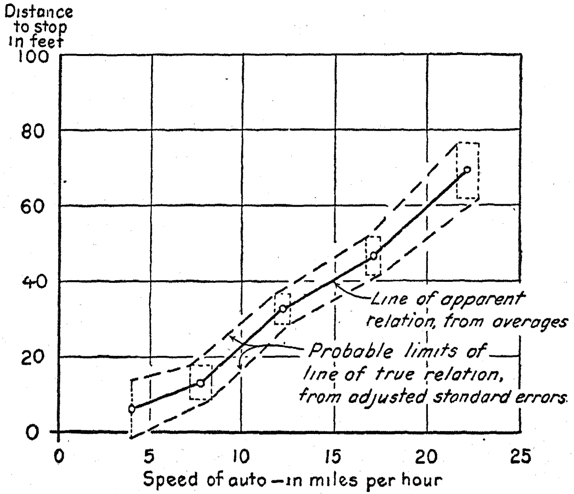
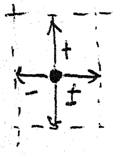


Fig. 7. Relation of speed of automobile to distance it takes to stop, as indicated by the range around group averages for which the probability is $\frac{2}{3}$ that the true average is included.

?



and lines drawn connecting them, just as before. But, in addition, rectangles have been drawn around each average to indicate the zone within which the true value would probably be found to lie if enough records were taken, using plus or minus the range for two chances out of three each way as the distance in laying off the rectangles from each average.[3] The corners of these rectangles have then been

[3] As the rectangles have been laid off with regard to both distance and speed, only in less than half the samples would the true values fall within the rectangles. In two out of three such samples the average speed will not differ from the true average speed by more than the stated amount. Similarly, in two out of three such samples the observed average distance will not differ from the true average by more than the extent calculated. Since $\frac{2}{3}$ times $\frac{2}{3}$ equals $\frac{4}{9}$, only in four samples out of nine, on the average, would it be likely that *both* observed speed and distance would fall within the calculated ranges from the true values *at the same time.*

Confidence limit = 66%

(col 5, Table 14)

connected by lines just as were the averages before. The probabilities now are that the line showing the true average relationship between speed and distance would run somewhere between these upper and lower boundaries, even though it might not be the particular irregular line of averages we have used so far.

Conclusion of the Range

Figure 7 indicates that there is really a rather wide zone within which the true relation might fall, even when we take the zone as indicated by statements which will be incorrect one time out of three. For example, it indicates that machines traveling 15 miles per hour would probably stop in 36 to 46 feet after the brakes were applied, whereas those traveling 20 miles an hour would probably stop in 52 to 68 feet. But this is still a pretty rough measure—would increasing the speed from 15 to 20 miles per hour increase the distance from 46 to 52 feet, only 6 feet; or would it increase it from 36 to 68 feet, 32 feet? Of and by themselves, the data do not tell us. We do not yet have any general statement of the relation between speed and distance.

The new problem?

Alternate Solution

We have seen how increasing the number of cases included in a single group increased the dependence which would be placed in that group. However, even by reducing our 50 cases to 5 groups we have not been able to get a consistent and satisfactory statement of the relation. Is it possible that by handling all the data as a single group we could get a better result? One way of doing this would be to average all the speeds and all the distances together. But that would only tell us what was the average distance to stop and the average speed. What we want to know is what distance is most likely to be required at any given speed, and the treatment just suggested would not give us that.

(a) average

There is one way, though, of determining the relation while considering all the records together. If we are willing to assume that an increase of one mile per hour in the rate of speed will increase the distance required to stop by exactly the same number of feet, no matter how rapidly or how slowly the machine is already moving, then we can determine this relation for all the data as a whole. On this basis a straight line can be used to represent the relation. All that we have to do is to determine a straight line which will come as near as possible to representing the relation as shown by all 50 individual observations.

(b) assumption

Straight Line Function

Summary. The change in one variable with changes in another may be approximately determined by grouping the records according to the independent variable and determining the corresponding averages for the dependent variable. Unless a very large number of observations is available, however, the functional relation shown by the successive averages will be irregular and inconsistent, owing solely to sampling variability. For that reason some method is needed for measuring the functional relation for the group of records as a whole. The simplest way in which this can be done is by assuming that the relation can be represented by a continuous straight line. Methods of determining such a line will be considered in the next chapter.

[illegible] variables

Note 1, Chapter 4. As already noted earlier in this chapter, it is always possible to reverse the dependent and the independent variables. Thus the data presented in Figure 3, on page 38, might have been plotted with time as the independent variable and with distance fallen as the dependent. A curve might then have been drawn in to show the average distance which a body can traverse for a given time of fall. Similarly, the data charted in Figure 4, on page 44, might have been charted with distance as the abscissa and speed as the ordinate. The data would then be in shape to consider the question, what is the average speed of cars which require a given specified distance to stop? The functions which express these relations are not exactly the reciprocal of the functions which express the reverse relation. That is, when

$$Y = f(X)$$

and

$$X = \phi(Y)$$

$$f(X) \neq \frac{1}{\phi Y}$$

The reasons for this will be considered subsequently.

CHAPTER 5

DETERMINING THE WAY ONE VARIABLE CHANGES WITH ANOTHER: (2) ACCORDING TO THE STRAIGHT-LINE FUNCTION

There are a good many ways by which a straight line can be determined to show the functional relation between the two variables, speed and distance. One way would be simply to place a ruler over the chart along the several group averages, or to stretch a black thread over them, and draw the line in by eye so as to fall as nearly as possible along them. Although no two persons would draw their lines exactly the same, still this method might give fairly satisfactory results where only a rough measure was wanted. In the present case, however, in view of the expensive field work necessary to collect the data, it would seem worth while to put as much clerical time on analyzing those we have as is needed to give the most accurate results. We shall therefore use the exact correlation method of determining the straight line.

The equation of a straight line. The determination of what this line will be consists in finding the *constants* for the *equation* of the line. Just as we have already seen (Chapter 3) that the curve showing the relation between the distance a body has to fall and the time it takes can be expressed by the relation,

$$Y = \tfrac{1}{4}\sqrt{X}$$

so any straight line can be expressed by the relation [1]

$$Y = a + bX \tag{8}$$

[1] Written this way, the equation is a perfectly general one which can be applied to the relation between *any* two variables, by calling one of them Y and the other one X. The symbol Y in the equation simply represents *the number of units* of the variable we designate as Y, whatever that may be, acres, dollars, pounds; and the symbol X likewise represents the number of units of the variable we designate as X. Thus if X is the number of rooms in each of a series of houses, X may be 4 for the first house, 7 for the next, 6 for the next, and so on. When we write X we then mean the number of rooms in each house, no matter how large or how small that number may be in any particular case. The particular number which X represents in any given case is said to be the value of X. Thus for a house of 5 rooms, we should say "the value of X is 5."

Figure 8 illustrates the meaning of a and b in this formula. When the value of X is 0, b times X is zero, and Y is equal to a. This constant, a, therefore, gives the height of the line (in terms of Y or vertical units) at the point where X is zero. This is indicated at the left edge of the chart.

From the same equation, every time X increases one unit, Y increases b times one unit, since Y is computed as a plus b times X. The difference of the height of the line (measured in Y units) between the point where X is 1 and where X is 2, is therefore b units of Y, just as indicated on the chart. And this continues to hold true for every unit change in X, whether from 1 to 2, or from 0 to 1, or from 99 to 100.

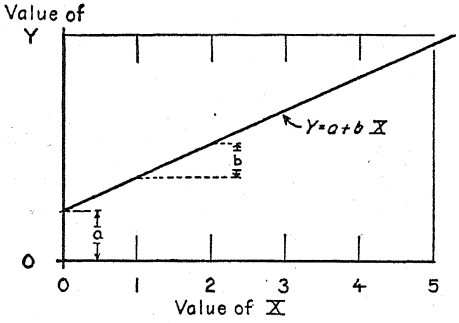


FIG. 8. Graph of the function $Y = a + bX$.

The meaning of these constants in the *equation of the straight line*, as equation (8) is known, may be illustrated more concretely by taking some actual values for the constants a and b, and seeing how the line would look then. If we take 3 for a, and 2 for b, the equation would then read:

$$Y = 3 + 2X$$

Figure 9 shows the line for which this is the equation. Thus if X is taken as zero, the value of Y is found to be

$$Y = 3 + (2 \text{ times } 0) = 3 + 0 = 3$$

And 3 is therefore the Y value corresponding to the X value, zero.

Similarly if X is taken as 10,

$$Y = 3 + (2 \text{ times } 10) = 3 + 20 = 23$$

And the Y value corresponding to the X value of 10 is therefore 23. All other values of Y which may be computed for values of X within the range shown in Figure 9 will similarly be found to lie exactly on the same line.

Figure 9 illustrates again the meaning of the constants a and b. When X is zero, the value of Y is three units above zero, as indicated, and for every unit increase in X (say from 5 to 6) the value of Y goes up 2 units. This is exactly the same thing as shown in Figure 8, except that there no definite values were assigned to a and b, whereas here they have been given exact numerical values.

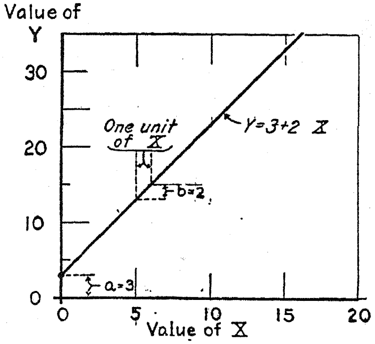


FIG. 9. Graph of the function $Y = 3 + 2X$.

To represent the general relation between the speed of an automobile and the distance it takes to stop, therefore, we can use this same kind of equation, letting X stand for the speed in miles per hour and Y stand for the distance-to-stop in feet.

Thus when we write the equation:

$$Y = a + bX$$

we shall be using that as shorthand for

$$\text{Feet to stop} = a + b\ (\text{speed in miles per hour})$$

But to give this equation definite meaning we must determine the numerical values for a and b, just as in our previous illustration we had to assume numerical values for these constants before the graph had any definite meaning for us.

The "observation equations." One way of finding what the values should be is by regarding each one of our original observations (Table 10) as an algebraic equation itself. Thus the first observation, 2 feet to stop at 4 miles per hour, would be written

$$2 = a + b\ (4)$$

putting the 2 feet in place of Y in the equation and the 4 miles in place of X.

Similarly the next observation, 4 feet to stop at 7 miles per hour, would be expressed

$$4 = a + b\ (7)$$

and so on right through to the last observation, 32 feet to stop at 20 miles per hour, which would be written—

$$32 = a + b\ (20)$$

Bringing all these different equations together would give a series looking like this:

$$\begin{aligned} 2 &= a + 4b \\ 4 &= a + 7b \\ 50 &= a + 17b \\ \cdot\ &\quad \cdot \quad \cdot\cdot \\ \cdot\ &\quad \cdot \quad \cdot\cdot \\ 80 &= a + 14b \\ 32 &= a + 20b \end{aligned}$$

(The middle equations are omitted here to save space.)

Since we had 50 original observations, we should have 50 different equations, each one containing the two unknown constants a and b.

Now by the rules of simple algebra, any *two* independent equations containing *two* unknown constants can be solved simultaneously to obtain the numerical values for those constants. One way to find the values of our unknown a and b would be to pick two of the equations representing our observations and solve them simultaneously. Suppose we take the first and the last ones; we shall then have:

$$\begin{aligned} a + 4b &= 2 \quad \text{— 1st equation} \\ a + 20b &= 32 \quad \text{— last equation} \end{aligned}$$

Solving these two equations simultaneously, we find the values

$$a = -5\tfrac{1}{2} \checkmark$$

$$b = 1\tfrac{7}{8} \checkmark$$

But in getting these values we have used only 2 out of the 50 observations. Should we have got the same result if we had used another pair? Suppose we take the second observation and next to the last—

Then

$$\begin{aligned} a + 7b &= 4 \\ a + 14b &= 80 \end{aligned}$$

These equations, solved simultaneously, give the values

$$a = -72$$

$$b = 10\tfrac{6}{7}$$

which are certainly far different from those secured before. Apparently the values secured by this method would depend upon the particular pair of observations selected, perhaps varying with each pair.

If we work out estimated values for Y for given values of X by these two solutions, we get estimates as follows:

According to the first result,

$$Y = -5.5 + 1.875X$$

when $X = 10,\ Y = 13.25;$ when $X = 20,\ Y = 32$

According to the second result,

$$Y = -72 + 10.86X$$

when $X = 10,\ Y = 36.6;$ when $X = 15,\ Y = 90.9$

If we should then plot the two calculated points for the first of these equations, and connect them by a straight line, we should find that that line also passes through the two dots which represent the two observations from which the values were calculated. Similarly, if we should plot the two computed points for the second equation, and pass a straight line through them, that also would pass through the two dots which represent the values from which it was calculated. Clearly, therefore, fitting a line to two observations is merely determining the line that passes through them. We could compute as many different lines as there are different pairs of observations *not* lying on the same line.

Fitting a straight line to two points, as we have done here, is simply equivalent to drawing a line to pass through those two points. This is evident in Figure 9A. Here the dot chart shown originally as Figure 4 has been replotted. The dots used in computing the above equations have been designated by crosses. The two lines computed have been plotted in. Quite clearly no *single* line could pass through all the different points. If we computed more lines by this process of using selected pairs of points, we should just get a larger variety of different lines.

Fitting the line by "least squares." If we are going to use a mathematically determined straight line at all, what we need is one which represents all 50 observations instead of any particular pair of them. No one line can *exactly* fit all 50 observations, for, as we have just seen, the line which would agree with the first and the last would not agree at all with the second and next to the last. What we shall have to find is some compromise line which will come as near as possible to agreeing with all the 50 observation equations, even though it does not *exactly* agree with any one. Mathematicians have worked out a method of obtaining such a line by the use of what is known as the "method of least squares." Although the process of determining the values of the constants a and b by this method is somewhat complicated, it takes all the observations into account, and gives each one of them an equal weight in the process. It is therefore of very great value in handling problems of this sort.

Statement of the problem

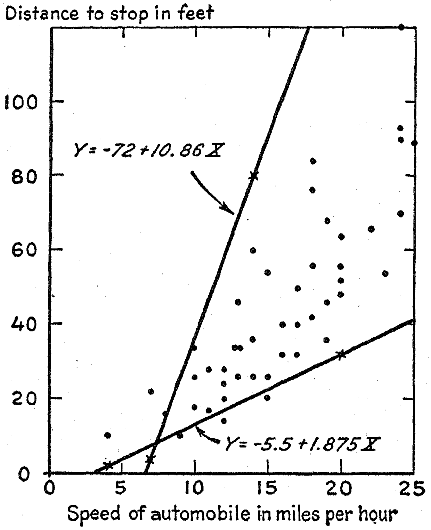


FIG. 9A. Data for automobile problem, and straight lines fitted to pairs of individual observations.

The equations upon which the process is based are derived by the use of calculus, and their derivation is given in Note 2, Appendix 2. The method itself, however, is very simple and can be used by anyone having a knowledge of simple algebra.

Computing the extensions. The individual observations are first listed as shown in Table 15. The speed in miles per hour is placed

TABLE 15

COMPUTATION OF VALUES FOR DETERMINATION OF LINE BY LEAST SQUARES

Speed in miles per hour, X	Distance to stop in feet, Y	X^2	XY
4	2	16	8
7	4	49	28
17	50	289	850
14	36	196	504
12	20	144	240
11	28	121	308
20	48	400	960
15	54	225	810
17	40	289	680
13	34	169	442
15	26	225	390
19	68	361	1292
10	26	100	260
18	56	324	1008
22	66	484	1452
18	84	324	1512
8	16	64	128
4	10	16	40
12	14	144	168
20	56	400	1120
23	54	529	1242
18	76	324	1368
12	24	144	288
16	32	256	512
18	42	324	756
19	46	361	874
24	93	576	2232
14	26	196	364
12	28	144	336
9	10	81	90
10	34	100	340
15	20	225	300
24	70	576	1680
25	85	625	2125
20	64	400	1280
19	36	361	684
13	26	169	338
10	18	100	180
7	22	49	154
16	40	256	640
14	60	196	840
20	52	400	1040
24	120	576	2880
24	92	576	2208
17	32	289	544
13	34	169	442
11	17	121	187
13	46	169	598
14	80	196	1120
20	32	400	640
Totals, 770 = ΣX	2,149 = ΣY	13,228 = $\Sigma(X^2)$	38,482 = $\Sigma(XY)$

under the heading "X," and the distance-to-stop in feet is placed under the heading "Y." Then each X item is squared, and entered in the column headed "X^2"; and each X item is multiplied by the accompanying Y item, and entered in the column headed "XY." Then all the items in each column are summed, giving the totals at the foot of each column. Just as before, in computing the standard deviation, we shall use the symbols "ΣX" to represent the sum of all the X items; "ΣY" to represent the sum of all the Y items; "$\Sigma(X^2)$" to represent the sum of all the X^2 items; and similarly, we shall use "$\Sigma(XY)$" to represent the sum of all the products in the XY column.

Solving the equations. Having obtained these values as indicated in Table 15, we can next proceed to find the values of a and b by the aid of the following formulas:

$$b = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - n(M_x)^2} \tag{9}$$

$$a = M_y - bM_x \tag{10}$$

see also notes below

In using these formulas the value of b is determined first, then it is used in the next formula to determine the value of a.[2]

$$M_x = \frac{\Sigma X}{n} = \frac{770}{50} = 15.4$$

$$M_y = \frac{\Sigma Y}{n} = \frac{2149}{50} = 42.98$$

Alternate way of derivation

[2] It should be noted that if both X and Y had been stated in terms of deviation from their mean values (just as was done when the standard deviation, σ, was computed in Table 6), they would have been denoted by the symbols small x and small y. If the product shown in the fourth column of Table 15 had then been obtained by multiplying together these two values, it would have been designated xy, and its sum, $\Sigma(xy)$. The correction factors used in the first part of the formula (9) just given are used simply to change the product sum of the original observations, $\Sigma(XY)$, to what it would have been if it had been computed from the deviations of the mean instead. That is to say,

$$\Sigma(XY) - nM_xM_y = \Sigma(xy) \tag{11}$$

Similarly, $\Sigma(X^2) - n(M_x)^2 = \Sigma(x^2)$

Hence $b = \Sigma(xy)/\Sigma(x^2)$ ✓

Equations (9) and (10) are only another way of stating the "normal equations," Using the values for ΣX, ΣY, $\Sigma(X^2)$, and ΣXY given in Table 15, in equations 9 and 10, we find the values of b and a to be:

$$b = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - n(M_x)^2} = \frac{38{,}482 - 50(15.4)(42.98)}{13{,}228 - 50(15.4)(15.4)} = \frac{5{,}387.4}{1{,}370} = 3.93$$

$$a = M_y - bM_x = 42.98 - (3.93)(15.4) = -17.54$$

The equation for the straight line, as thus determined by all the observations, is therefore

$$Y = -17.54 + 3.93X$$

(For an exercise, plot this line in on the dot chart shown in Figure 4, on page 44.)

This line is called the *line of best fit,* since it is the line which gives, for all the 50 observed values of X, values of Y which come as near as possible to agreeing with *all* the different Y values observed. While some equations, such as the two computed from 2 observations each, would come closer than would this one for some individual cases, they would be much farther off for other cases; this one comes closer to agreeing with all the cases than any other straight line.[3]

Estimating Y from X. We can see just how the equation for this line works by taking any given value for X we wish and working out what the estimated value for Y would be. That is, we can take

which can be solved simultaneously to give the values for a and b. These equations are

normal equations

$$na + (\Sigma X)\,b = \Sigma Y$$
$$(\Sigma X)a + (\Sigma X^2)b = \Sigma XY$$

These two equations can be solved simultaneously to get the values for a and b which will best fit all the equations, in the same way that the previous paired observations were put into simultaneous equations and solved simultaneously to get the values which would exactly fit the two observations.

Principle

The method by which this line is fitted rests upon the assumption that the scatter of the individual observations around the fitted line will approximate a normal distribution. If one or two observations are exceedingly erratic as compared to the others, so that the scatter of the observations around the line will be very skew, this method of fitting may be unsatisfactory.

why least squares

[3] The way in which this equation gives the best fit may be explained mathematically. If the differences between each of the actual observations and the estimated values given by this equation are computed, squared, and summed, that sum will be smaller than it would be if any other straight line were used. Since this method determines the line with the smallest possible squared deviations, the line is known as the "least-squares" line, and the method of computing it is known as the "method of least squares."

calculate a + b [illegible] from equations any initial speed we wish and compute from the equation what would be the most probable distance required to stop, on the basis of the straight-line relationship.

If 14 miles per hour is taken, X will be 14. Substituting this value in the equation gives the estimated value of Y.

Estimated value Y

$$\begin{aligned} Y &= -\,17.54 + 3.93\,(14) \\ &= -\,17.54 + 55.02 \\ &= 37.48 \end{aligned}$$

So the number of feet which would probably be required to stop, when traveling at 14 miles per hour, would be about 37.5 feet. Comparing this with the original observations, we see that the 4 cars recorded at this speed stopped in 36, 26, 60, and 80 feet, respectively. At 23 miles per hour the single car observed took 54 feet to stop. What estimate will the equation give for that speed? Let us see:

$$\begin{aligned} Y &= -\,17.54 + 3.93\,(23) \\ &= -\,17.54 + 90.39 \\ &= 72.85 \end{aligned}$$

This is much higher than the single observation. But referring to Figure 4 we see that that observation fell far below the general trend of the other observations. The straight-line equation, based on all the observations, thus seems to give a more reliable estimate of the distance which is most likely to be required to stop at any given speed than does any one individual observation.

But how far is it true that the straight line gives the most accurate estimate? Will it hold true for a speed of 1 mile per hour or for a speed of 50? Let us see.

For 1 mile per hour the equation becomes:

hypothetical

$$\begin{aligned} Y &= -\,17.54 + 3.93\,(1) \\ &= -\,17.54 + 3.93 \\ &= -\,13.61 \end{aligned}$$

nonsensical

For 50 miles per hour it gives:

Hypothetical

$$\begin{aligned} Y &= -\,17.54 + 3.93\,(50) \\ &= -\,17.54 + 196.5 \\ &= 178.96 \end{aligned}$$

Of these two results, only the latter sounds at all sensible. To say that a machine moving 1 mile per hour stops in *minus* 13.61 feet is saying that it stopped 13.61 feet *back* of where the brakes were applied, which is certainly nonsense. On the other hand, to say that a machine traveling 50 miles per hour would stop in about 179 feet after the brakes were applied might be quite reasonable—if we had any direct evidence for machines traveling at that speed. But that we do not have. All that we have are observations on 50 machines traveling at rates varying from 4 to 25 miles per hour. Since we have no observations for speeds below 4 miles per hour, we cannot expect our equation to be of any reliability below that point; and, since we have no observations of speeds above 25 miles per hour, we cannot be sure that our equation will give good estimates beyond that point. *Only within the range covered by the original observations can an estimating equation of this type be used.* [illegible]

Of the 50 observations, there were 6 below 10 miles per hour and only one above 24, so 43 out of the 50 were between 10 and 24 miles per hour. For that reason no great reliance can be put in the equation below 10 miles per hour and above 24 miles per hour. Only within those limits where the bulk of the observations fell can the equation really be trusted.[4] For that reason the final equation, showing the average relation between speed and distance for automobiles, should be written:

$$Y = -17.54 + 3.93\ (X), \text{ for values of } X \text{ between 10 and 24}$$

Then the application of the equation is limited to the range given, and there is no danger of its being used to give absurd values for speeds too low or untested values for speeds too high.

Now that the limits of the line have been considered, it may be well to compare it to the group averages used before, to see how this single line, based on all the observations, compares with the irregular line obtained when the observations were grouped. This can be done conveniently by drawing in the line on Figure 7, which showed not only the line of averages but also the limits within which those averages were probably correct. This comparison is shown in Figure 10. The straight line determined by the least-squares solution has been

[4] See pages 113 to 121 for a discussion of the type of problem in which a formula may be used to make estimates beyond the range covered by the data. See also Chapter 18 for formulas for estimating the standard errors for a and b.

drawn in solidly for the range of speed in which most of the observations fell and has been dotted in for the remainder of the range.[5]

Comparing the straight line with the group averages and the error limits within which they probably would fall, we see that the line does fall within those limits in every case but one, and in that case it just barely misses it. That shows that, so far as indicated by the number of observations we have on which to base the results, the straight line may serve as a more reliable indication of the general relation than does the irregular line of the group averages.

Comparison of group averages (Ch. ? p 56) to the line of best fit

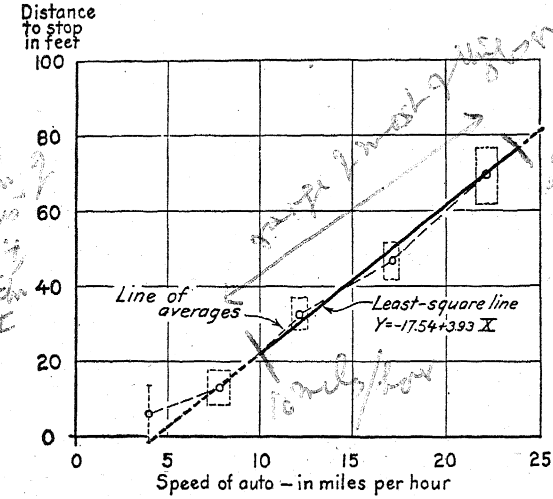


FIG. 10. Relation of speed of automobile to distance-to-stop as indicated by ranges around group averages and by least-squares straight line.

The estimated distance required to stop, for each speed considered, is shown by the corresponding ordinate of the line in Figure 10. The estimated values may also be obtained by substituting the X value in the equation, just as has been done for the observations at 14 miles and at 23 miles. Carrying out this computation gives the estimated values shown in Table 16. Subtracting the estimated distances from the actual distances gives the *residuals*, or the difference between the

[5] This line is drawn in according to the equation by determining the Y values for any two convenient values of X, and then drawing a straight line connecting them. Thus if the values at the end of the bulk of the observations, 10 and 24, are taken ~~for X,~~ the accompanying values for Y are found to be 21.8 and 76.8. These Y values are then plotted opposite 10 and 24 for X; a straight line drawn connecting them; and extended as a dotted line to cover the rest of the range.

two values. The symbol z is used in the table to designate these differences. The average of these differences, taken without regard to sign, is 11.6 feet; their standard deviation is 15.07 feet.[6]

TABLE 16

SPEED OF AUTO, DISTANCE TO STOP, AND DISTANCE ESTIMATED FROM SPEED BY LINEAR EQUATION

Miles per hour, X	Actual distance, Y	Estimated distance, Y'	Residual $(Y-Y')$, z	Miles per hour, X	Actual distance, Y	Estimated distan e, Y'	Residual $(Y-Y')$, z
4	2	−1.8	3.8	19	46	57.1	−11.1
7	4	10.0	−6.0	24	93	76.8	16.2
17	50	49.3	0.7	14	26	37.5	−11.5
11	36	37.5	−1.5	12	28	29.6	−1.6
12	20	29.6	−9.6	9	10	17.8	−7.8
11	28	25.7	2.3	10	34	21.8	12.2
20	48	61.1	−13.1	15	20	41.4	−21.4
15	54	41.4	12.6	24	70	76.8	−6.8
17	40	49.3	−9.3	25	85	80.7	4.3
13	34	33.6	0.4	20	64	61.1	2.9
15	26	41.4	−15.4	19	36	57.1	−21.1
19	68	57.1	10.9	13	26	33.6	−7.6
10	26	21.8	4.2	10	18	21.8	−3.8
18	56	53.2	2.8	7	22	10.0	12.0
22	66	68.9	−2.9	16	40	45.3	−5.3
18	84	53.2	30.8	14	60	37.5	22.5
8	16	13.9	2.1	20	52	61.1	−9.1
4	10	−1.8	11.8	24	120	76.8	43.2
12	14	29.6	−15.6	24	92	76.8	15.2
20	56	61.1	−5.1	17	32	49.3	−17.3
23	54	72.9	−18.9	13	34	33.6	0.4
18	76	53.2	22.8	11	17	25.7	−8.7
12	24	29.6	−5.6	13	46	33.6	12.4
16	32	45.3	−13.3	14	80	37.5	42.5
18	42	53.2	−11.2	20	32	61.1	−29.1

Interpreting the linear equation. Just what does *the line of least squares* tell us, now that we have decided it is a fairly accurate indicator of stopping distances—at least within the range 10 to 24 miles? We can answer that by trying to explain what the constants a and b

[6] The significance of this standard deviation of the residuals is explained on pages 129 and 494.

of the equation mean—the values −17.54 and 3.93, which we determined by least squares.

Meaning of constant (a)

The first of these constants, a, is merely an empirical value to place the height of the line. If observations available and the type of equation used were such that they could be expected to give a sensible value for the distance to stop when X was zero—that is, when the machine was not moving—then a would give that value, since when $X = 0$, $Y = a$. But, of course, when a machine is not moving, it does not take it any distance to stop, so in this case the a has no sensible interpretation *at that point*. But that is to be expected—as has been seen, the line as a whole has but little meaning below 10 miles per hour, and none at all below 4 miles; which was the lowest speed covered by the records. The constant a, therefore, has no meaning of and by itself in this particular example, but merely serves to place the height of the line as a whole for that range within which the line does have some meaning.

Constant (b)

The constant b, on the other hand, is always significant. It shows the difference in Y for every difference of one unit in X, on the average of all the observations, and within the range covered. In this particular problem, the value of 3.93 for b indicates that between 4 and 24 miles per hour each increase of one unit in X, that is to say, each increase of one mile per hour in speed, causes on the average an increase of 3.93 units in Y—that is, of 3.93 feet in the distance required to stop. This interpretation of b can always be made, and is one of the most significant results secured by determining the constants for the straight line. In comparison with the values shown in Table 13, ranging from 1.8 feet to 4.4 feet increase in stopping distance for each one mile increase in speed, this figure of 3.93 feet per mile increase in speed is seen as a sort of weighted average, averaging together all the different possible sorts of comparisons like those in Table 13.[7]

[7] The value determined for b, like the value previously determined for the mean yield of corn, is not the true value for all the cars in the city studied, but is only the estimate of that value as determined from the cars included in the sample. Just as the sample mean may vary from the true mean for the universe, so the b computed from the sample may vary from the true b for the universe. Likewise, the possible extent of that variation may be indicated by estimating its standard error. The increase in distance-to-stop for each additional mile in speed should be stated as

$$\boxed{3.93 \text{ feet} \pm (\text{standard error of } b)}$$

Pages 312 to 315 show how to calculate the standard error of b and explain its meaning more fully.

It should be noted that even though the straight line does fall within the standard error limits of most of the averages, as it does in this case, that by itself is no proof that the straight-line formula really expresses the true underlying relation between the speed of a machine and the distance that it takes it to stop in this example. It is a purely arbitrary method of describing relation, which apparently expresses the observed relation fairly well; but that is all. It is, after all, only an empirical expression of the relationship; and because it happens to agree fairly well is no proof that it expresses the true nature of the relation. In fact, there is as yet no proof that it is even the best empirical description of the observed relation that can be obtained; further tests, to be described in the next chapter, are necessary.

But whether or not the straight line is the best function in this particular example, it is a type of relation of very great importance and usefulness. It is one of the simplest functions to fit and to explain, and for that reason it is very widely used. The equations used in determining the constants of the equation (equations [9] and [10], page 66) are therefore of great importance. The student of analytical statistics should become thoroughly familiar with the methods of determining the constants of the equation and should understand thoroughly both the meaning and the limitations of this type of analysis.

Determining the constants for the linear equation for a given set of observations is called "'fitting' the equation to the data." Because the linear equation is one of the simplest of all equations to "fit," it is widely and frequently used. In many cases, no other possible relation is even considered. Actually, however, the linear equation is very limited in its logical meaning. By its very nature, it can represent only a situation where the change in the dependent variable, for a unit change in the independent variable, would be expected to be just the same regardless of how large or how small the independent variable was. This is a very precise and narrow relation. In many sets of relationship, the relation which theoretically would be expected would be a changing relationship as the value of the independent variable changed, instead of this unchanging relationship. Unless there is a good logical reason to expect the linear equation to represent truly the situation present, fitting a straight line can be regarded only as an empirical exercise, with no meaning to the constants obtained beyond the purely formal one of specifying the straight line that most nearly represents the data.

Fitting concept

many complicated cases

Summary. To express a functional relationship by a straight line, the constants may be determined arithmetically by the "method of least squares." Such a line gives the "line of best fit" under the assumptions of that method: a normal distribution of the observations around the line and the reduction of the squared residuals to a minimum. Estimates of the dependent variable may be made according to the linear function for any value of the independent variable. Only within the range which includes the bulk of the independent values does this estimate have meaning, however; and only then if the straight line gives a satisfactory expression of the observed relation, either empirically or logically.

Note 1, Chapter 5. Just as a straight line can be fitted to show the average distance-to-stop for each given rate of speed, so another straight line can be fitted if the variables are reversed. In that case the speed, miles per hour, could be regarded as the dependent or Y variable, and the distance-to-stop, feet, would be regarded as the independent or X variable. Working out the values of a and b for this reverse statement of the problem will be left as an exercise for the student. In line with the note to Chapter 4, it will be found that the value of this new b is not equal to $\frac{1}{b}$ as previously determined, but will differ slightly from it.

CHAPTER 6

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN ANOTHER CHANGES: (3) FOR CURVILINEAR FUNCTIONS

A straight-line equation is frequently a fairly good empirical statement of the relation between two variables even when the true relation is more complex than the straight line can portray. Yet it may be just as important to know the exact or approximate nature of the relationship as it is to have an empirical statement of it. For that reason it is necessary to consider other ways of expressing a relationship than the straight line.

In the automobile-stopping case we have been using as example, Figures 4 and 10 showed that the straight line agreed fairly well with the averages from the observations. Closer examination of the figures, however, reveals that for speeds below 10 miles per hour the actual stopping distance was usually greater than is indicated by the line; for speeds 10 to about 17 miles per hour the average stopping distance was about the same as indicated by the line; above 20 miles per hour the stopping distance was frequently much greater than is indicated by the straight line. These considerations rob the line of much of its usefulness for the purpose for which the study was started—to serve as a basis for establishing speed limits. The linear relation between speed and stopping distance is apparently not accurate above 20 miles per hour, tending to underestimate the distance required at higher speeds. Since that might be the very range within which it was desired to set the speed, the conclusions most needed for that particular purpose would be lacking.

[illegible]

The real difficulty involved is in the assumption that the straight-line function applies. We have assumed that an increase of one mile in the speed of the car increases the distance required to stop by the same number of feet, no matter how fast the car is already traveling. When we examine Figures 5 and 10 closely, we see that this is not correct; the line of averages slants up slowly at first, then tends to rise more steeply as the speed is increased, until it has the steepest slope at the highest speed. It is therefore incorrect to assume that

Assumption in St line function is [illegible] real we can express the relation by determining the average increase in stopping distance for an increase of one mile in the rate of speed; for *the increase in stopping distance is not the same regardless of the rate of speed, but tends to become greater as the rate of speed increases.* Only if our expression of the relation can express that fact too will it sum up all our observations with sufficient accuracy.

Statement of problem

What is needed is some general way of stating the relation between speed and distance, similar to the general relation expressed in the straight-line formula, yet expressing *a changing relationship* instead of the uniform linear relation shown by the straight line.

Fitting the curve

Different types of equations. In the same way that it is possible to represent relations mathematically by a straight line, it is possible to represent them by curves of various types. We have seen how the equation $Y = a + bX$ can be used to represent any straight line by determining the proper values to be assigned to the constants a and b. There is practically no limit to the different kinds of curves which can be similarly described by mathematical equations. The equations of a number of curves which are useful in statistical analysis of the relations between variables are:

$$Y = a + bX + cX^2 \qquad (a)$$

$$\log Y = a + bX \qquad (b)$$

$$\log Y = a + b \log X \qquad (c)$$

$$Y = a + b \log X \qquad (d)$$

$$Y = \frac{1}{a + bX} \qquad (e)$$

$$Y = a + bX + cX^2 + dX^3 \qquad (f)$$

hyperbola

$$Y = a + bX + c\left(\frac{1}{X}\right) \qquad (g)$$

Each of these equations can be used to represent a certain type of curve. Thus type (a) is the equation of a parabola. If we take certain values for the unknown constants a, b, and c, substitute them in the formula, work out the values of Y for various values of X, and plot them the same as we did before, we will see the sort of curve this equation can be used to express. Thus if we take 1 for a, 0.5 for b, and $-$ 0.1 for c, the equation will read:

$$Y = 1 + 0.5X - 0.1X^2$$

When the value of X is 0, Y will be 1, obviously. When X is 1, Y will be

$$\begin{aligned} Y &= 1 + 0.5\ (1) - 0.1\ (1^2) \\ &= 1.4 \end{aligned}$$

When X is 2, Y will be

$$\begin{aligned} Y &= 1 + 0.5\ (2) - 0.1\ (2^2) \\ &= 1 + 1 - 0.4 \\ &= 1.6 \end{aligned}$$

Similarly, when X is 3

$$\begin{aligned} Y &= 1 + 0.5\ (3) - 0.1\ (3^2) \\ &= 1 + 1.5 - 0.9 \\ &= 1.6 \end{aligned}$$

For X equal to 4

$$\begin{aligned} Y &= 1 + 0.5\ (4) - 0.1\ (4^2) \\ &= 1.4 \end{aligned}$$

and for $X = 5$

$$\begin{aligned} Y &= 1 + 0.5\ (5) - 0.1\ (5^2) \\ &= 1 \end{aligned}$$

and for $X = 6$

$$\begin{aligned} Y &= 1 + 0.5\ (6) - 0.1\ (6^2) \\ &= 0.4 \end{aligned}$$

Plotting each of these values on cross-section paper and drawing a smooth curve through the several points, we get the result shown in Figure 11 in the center of the top section. Examination of the figures above and of this chart discloses one characteristic of this type of curve—the curve is always symmetrical on both sides of the highest point—the point where it stops going up and starts to turn down (as half way between $X = 2$ and $X = 3$ in this case). The value of Y when $X = 2$ is the same as when $X = 3$. When $X = 1$ it is the same as when $X = 4$ and, for $X = 5$, Y is the same as when $X = 0$. As a result the curve could be cut into halves at the point of turning downward, one of which would be the reverse of the other. Besides this characteristic symmetry, this curve has another peculiarity—it has one, and only one, change from moving upward to moving down-

merits
(?)
symmetry

5

ward, no matter what values are assigned to a, b, and c, or how far it is carried out. For the equation shown, the curve reaches its highest point when $X = 2.5$. As shown in Figure 11, the curve continues downward on both sides of this point, no matter how large the positive or negative values of X become. Thus if $X = 100$,

$$\begin{aligned} Y &= 1 + 0.5\,(100) - 0.1\,(100^2) \\ &= 1 + 50 - 1000 \\ &= -\,949 \end{aligned}$$

If $X = -\,100$

$$\begin{aligned} Y &= 1 + 0.5\,(-100) - 0.1\,(-100^2) \\ &= 1 - 50 - 1000 \\ &= -\,1049 \end{aligned}$$

If the value of b were negative and of c were positive, the curve would then be concave from above instead of convex and would be symmetrical with respect to its lowest point.

Because of the characteristics mentioned, this type of curve is not very satisfactory to represent many types of relations. It does have great flexibility, in that many differently shaped curves can be represented by some particular segment of the parabola; but on the other hand the parabolic shape itself is so simple that many times the real relation between the variables cannot be represented by a parabola.

The characteristics of a number of other types of simple curves are also illustrated in Figure 11. In each case an equation of the type indicated has been assumed, and the values of Y corresponding to values of X have been computed as has just been done for the simple parabola. Then plotting these computed values gives the curves shown. Thus type (f), the cubic parabola, is seen to have one maximum point and one minimum point and one point of inflection (the point where the curve changes from concave from above to convex, or *vice versa*). No matter what values are assigned the constants in this equation, it can have only the single inflection and the two points of maxima and minima. Of course the particular data to be represented might fall anywhere along the entire course of the curve —if only a single change from positive to negative slope were required, the point of inflection in the cubic parabola might lie beyond the extremes of the data, and so not show at all when the fitted curve was plotted for the range covered by the data.

Figure 11 also illustrates curves of types (b) to (e), as well as some others not given special type designations. In each case where the log of Y is used in place of Y, it is evident that the previous curve has been modified as if by compressing the ordinates nearest zero and stretching out the ordinates farthest away from zero, stretching them more and more as they depart more and more from zero. This process transforms the straight lines of $Y = a + bX$ to a curve concave from above when $\log Y = a + bX$ is used instead; or, when log

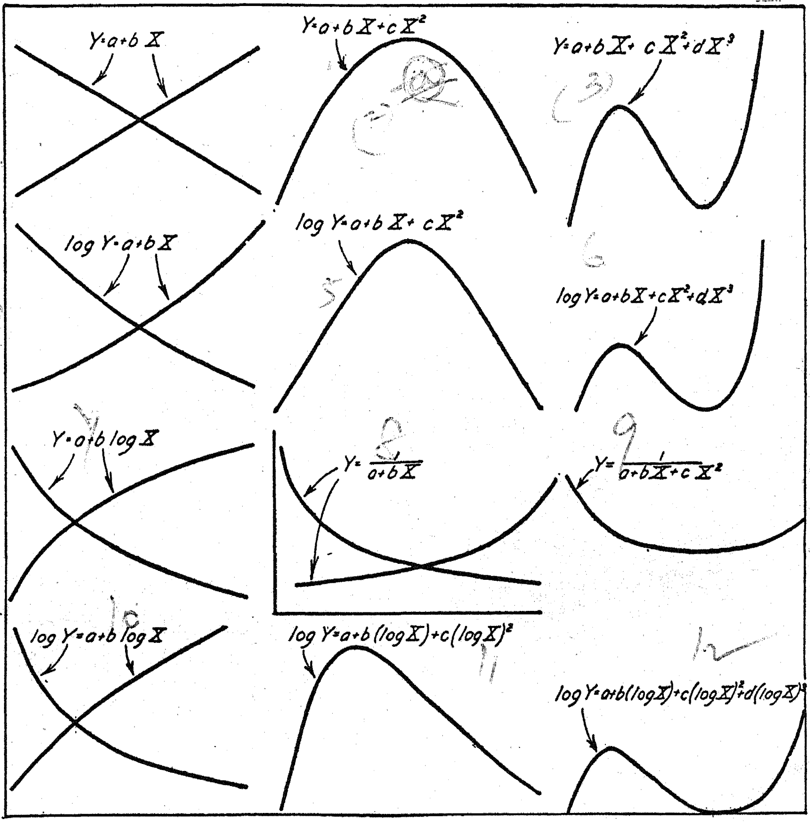


FIG. 11. Curves illustrating a number of different types of mathematical functions.

(11) $Y = a + bX + cX^2$ is substituted for $Y = a + bX + cX^2$, it lengthens out the top of the bend if b is positive, or flattens out the bottom of the dip if b is negative. Similar results are found with the cubic parabola.

Similarly, when $\log X$ is used in place of X, the previous curves are modified as if the abscissas were compressed near zero, and stretched out in the higher values. This changes the straight line (1)

(No. 7 in Fig 11)

Parabolas

of $Y = a + bX$ to a curve for $Y = a + b \log X$, convex from above when b is positive and concave from above when b is negative. The parabolas are similarly transformed, making the slopes different on each side of the bend in the simple parabola or on each side of the inflection in the cubic. The effect is to move the "hump" or "dip" in nearer to the zero abscissa and to stretch out the remainder of the curve (including the second bend, in the case of the cubic parabola).

When logarithms are used for both X and Y, the effect is to modify both sets of coordinates in the manner previously described. The curve $\log Y = a + b\ (\log X)$ may have either a concave or convex bend if b is positive, but is always concave from above if b is negative. Similar modifications are noted in the case of the simple parabola.

Main features

In any event it should be noted that the curves whose equations contain logarithms retain some of the same characteristics as those with similar equations without logarithms. Thus the linear equations (with only a and b) *never* change from a positive to a negative slope; (2) the simple parabola *always* has one such change, if carried out far enough; and the (3) cubic parabola always has two such changes. (4) In addition, it should be noted that a variable can be stated in terms of logarithms only if it has no negative values. Whereas the other functions can express negative values as readily as positive ones, the logarithmic curves always become asymptotic as they approach zero—that is, they tend to flatten out and to run almost parallel with the axis. This is because a logarithm cannot be obtained for a negative number. (5) No matter how small a logarithm becomes, the corresponding antilogarithm is still positive, even if only a very small decimal fraction.

Hyperbola

The hyperbola (type [e]) shown just below the center of Figure 11 also is peculiar in that it can become asymptotic as it approaches both the X axis and the Y axis, even if one or both of the variables are in negative values.[1] However, the values of X and Y which it ap-

[1] There are three types of simple hyperbolas which are frequently useful in curve fitting:

$Y = \dfrac{1}{a + bX}$ is an equilateral hyperbola, asymptotic to a line parallel to the X axis;

$Y = a + b\left(\dfrac{1}{X}\right)$ is an equilateral hyperbola asymptotic to a line parallel to the Y axis;

$\dfrac{1}{Y} = a + b\left(\dfrac{1}{X}\right)$ is an equilateral hyperbola asymptotic tc lines parallel to both axes.

proaches are not the zero values, as with the logarithmic curves, but special values which vary in each particular case and depend upon the value of the constants a and b in the equation. Still more complex curves of the same hyperbolic type may be obtained by including higher powers of X, such as

$$Y = \frac{1}{a + bX + cX^2}$$

[illegible] contradictions

Still other curves may be represented by hybrid equations, which combine two or more of the simple types described thus far. Thus type (g) is a compound of a simple linear equation and a simple hyperbola. This is sometimes useful to represent curves which cannot be represented by the simpler types. The choice of an equation to represent a particular set of data, however, depends upon logical analysis as well as upon the empirical ability of a given equation to represent the relation found. This matter is discussed at length subsequently on pages 113 to 125.

Imp

The equations discussed to this point all have one characteristic in common. They can all be fitted to the data by relatively elementary arithmetic operations, as will be shown subsequently. There are many other types of more complicated equations which cannot be fitted so readily. These can reproduce curves with recurrent or periodic oscillations, growth curves, and other complicated biological or physical phenomena. Discussion of the use and fitting of such complicated curves lies outside the scope of this book.[2]

Illustration

The inability of any one equation to represent many simple curves may be illustrated by taking a different example from the automobile-stopping case we have been considering previously. Table 17 shows a series of observations of two variables—the protein content of different samples of wheat, as determined by chemical analysis, and the proportion of "hard, dark, vitreous kernels" in each sample, as determined by visual examination with the naked eye. The relation here is quite different from the one we have been considering so far. There is no causal connection between these two variables in the sense of one's being caused by the other. Instead, they are merely two different ways of measuring the character of the wheat. It is a short, rapid process, however, to examine the samples by eye and determine

[2] For examples of such complicated curves and methods of fitting them, see Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics*, pp. 540-571, 441–462, New York, Henry Holt and Co., 1940.

problem

the percentage of hard, dark, vitreous kernels, whereas it is a long and expensive process to run a chemical test on each lot. For that reason it is of importance to know whether it is possible to estimate the protein content from the percentage of vitreous kernels, and, if so,

TABLE 17

PROTEIN CONTENT AND PROPORTION OF VITREOUS KERNELS FOR EACH OF A NUMBER OF SAMPLES OF WHEAT*

Sample number	Protein content	Proportion of vitreous kernels
	Per cent	*Per cent*
1	10.3	6
2	12.2	75
3	14.5	87
4	11.1	55
5	10.9	34
6	18.1	98
7	14.0	91
8	10.8	45
9	11.4	51
10	11.0	17
11	10.2	36
12	17.0	97
13	13.8	74
14	10.1	24
15	14.4	85
16	15.8	96
17	15.6	92
18	15.0	94
19	13.3	84
20	19.0	99

* These values are actual items, picked so as to show the relationship more clearly. Actually, the correlation is not so high as is shown by these selected cases.

how closely. So even though the vitreous kernels do not *cause* the differences in protein, we can still regard the proportion of vitreous kernels as the independent variable and the percentage of protein as the dependent variable. That means only that we are going to try to estimate the dependent (protein) from the independent (percentage

Good

of vitreous kernels) even though there is no direct cause-and-effect relation present.

The relation between the proportion of vitreous kernels and the per cent of protein may be seen more readily if a dot chart is made, showing the two variables for each of these individual observations. According to the previous discussion, we shall regard the proportion of kernels vitreous as X, the independent variable; and the percentage of protein as the dependent variable, Y. In preparing the dot chart, shown in Figure 12, we shall therefore plot the X values, or percentage of vitreous kernels, along the horizontal axis and the Y values, the proportion of protein, along the vertical axis.

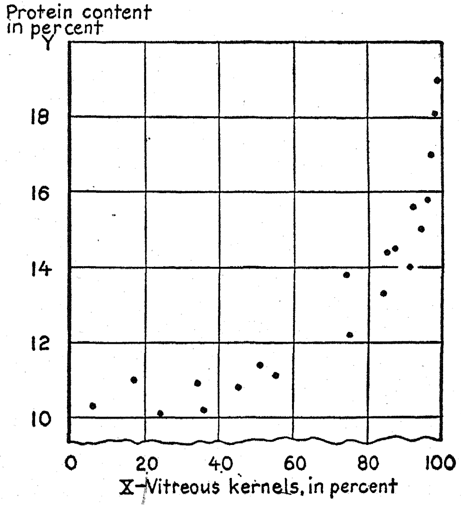


FIG. 12. Dot chart showing relation of proportion of vitreous kernels to protein content of wheat.

It is quite obvious from an inspection of the figure that a straight line would not do to represent the change in protein with change in vitreous kernels. Some type of curve is necessary. Let us see if the simple parabola is the proper type of curve.

"Fitting" a simple parabola. To represent the relationship between the two variables according to the formula

$$Y = a + bX + cX^2 \qquad (12)$$

we shall have to determine from the 20 observations the values to assign to the constants a, b, and c, just as before for the straight line we had to determine values for a and b. (Of course the a and b for the parabola will not be the same as the values for the straight line—unless c happens to be zero, which would make the equation for the parabola give a straight line instead.) The values for these constants are determined by constructing and solving the following equations:[3]

Also see below

$$\begin{rcases}(\Sigma x^2)b + (\Sigma xu)c = \Sigma xy \\ (\Sigma xu)b + (\Sigma u^2)c = \Sigma uy\end{rcases} \qquad (13)$$

and

$$a = M_y - b(M_x) - c(M_u) \qquad (14)$$

The values necessary in constructing equations (13) and (14) are derived as follows:

Use U to represent the X^2 values of equation (12).[4] foot-note

Then

$U = X^2$

$$\begin{rcases} M_x = \frac{\Sigma X}{n};\; M_u = \frac{\Sigma U}{n};\; M_y = \frac{\Sigma Y}{n} \\ \Sigma x^2 = \Sigma X^2 - nM_x^2 \\ \Sigma xu = \Sigma XU - nM_xM_u \\ \Sigma u^2 = \Sigma U^2 - nM_u^2 \\ \Sigma xy = \Sigma XY - nM_xM_y \\ \Sigma uy = \Sigma UY - nM_uM_y \end{rcases} \qquad (15)$$

After computing these values, the two equations (13) are solved simultaneously to obtain the values for b and c, and then these values are substituted in equation (14) to obtain the value for a.

Table 18, following, shows the form of computation in the first step to obtain these values for the data of Table 17.

[3] An alternative method is to solve the following three equations simultaneously. The clerical work is about the same in both methods.

$$\begin{aligned} na + (\Sigma X)b + (\Sigma U)c &= \Sigma Y \\ (\Sigma X)a + (\Sigma X^2)b + (\Sigma UX)c &= \Sigma XY \\ (\Sigma U)a + (\Sigma UX)b + (\Sigma U^2)c &= \Sigma YU \end{aligned}$$

These equations are derived by the process explained in Note 2, Appendix 2.

[4] If U is made equal to X^2 divided by some convenient number, say 1,000, the volume of necessary arithmetic can be materially reduced, without affecting the accuracy of the result. See Note 3, Appendix 2, for proof.

TABLE 18

COMPUTATION, FOR WHEAT PROBLEM, OF VALUES NEEDED TO DETERMINE CONSTANTS OF THE SIMPLE PARABOLA

Per cent vitreous kernels X	Per cent protein (minus 10)* Y	X^2 and U	XU	U^2	XY	UY
6	0.3	36	216	1,296	1.8	10.8
75	2.2	5,625	421,875	31,640,625	165.0	12,375.0
87	4.5	7,569	658,503	57,289,761	391.5	34,060.5
55	1.1	3,025	166,375	9,150,625	60.5	3,327.5
34	0.9	1,156	39,304	1,336,336	30.6	1,040.4
98	8.1	9,604	941,192	92,236,816	793.8	77,792.4
91	4.0	8,281	753,571	68,574,961	364.0	33,124.0
45	0.8	2,025	91,125	4,100,625	36.0	1,620.0
51	1.4	2,601	132,651	6,765,201	71.4	3,641.4
17	1.0	289	4,913	83,521	17.0	289.0
36	0.2	1,296	46,656	1,679,616	7.2	259.2
97	7.0	9,409	912,673	88,529,281	679.0	65,863.0
74	3.8	5,476	405,224	29,986,576	281.2	20,808.8
24	0.1	576	13,824	331,776	2.4	57.6
85	4.4	7,225	614,125	52,200,625	374.0	31,790.0
96	5.8	9,216	884,736	84,934,656	556.8	53,452.8
92	5.6	8,464	778,688	71,639,296	515.2	47,398.4
94	5.0	8,836	830,584	78,074,896	470.0	44,180.0
84	3.3	7,056	592,704	49,787,136	277.2	23,284.8
99	9.0	9,801	970,299	96,059,601	891.0	88,209.0
1,340	68.5	107,566	9,259,238	824,403,226	5,985.6	542,584.6

* To simplify the following calculations, 10.0 has been subtracted from each protein reading (See Note 3, Appendix 2.)

The values at the foot of the table give the values called for in equations (15). Substituting the values as computed for those shown symbolically, the arithmetic appears as follows:

$$M_x = \frac{\Sigma X}{n} = \frac{1,340}{20} = 67$$

$$M_y = \frac{\Sigma Y}{n} = \frac{68.5}{20} = 3.425$$

$$M_u = \frac{\Sigma U}{n} = \frac{107,566}{20} = 5,378.3$$

$$\Sigma X^2 - nM_x^2 = 107{,}566 - 20(67)^2 = 17{,}786$$

$$\Sigma XU - nM_xM_u = 9{,}259{,}238 - 20(67)(5{,}378.3) = 2{,}052{,}316$$

$$\Sigma U^2 - nM_u^2 = 824{,}403{,}226 - 20(5{,}378.3)^2 = 245{,}881{,}008$$

$$\Sigma XY - nM_xM_y = 5{,}985.6 - 20(67)(3.425) = 1{,}396.1$$

$$\Sigma UY - nM_uM_y = 542{,}584.6 - 20(5{,}378.3)(3.425) = 174{,}171.05$$

These calculations give the values needed in equations (13), which are to be solved simultaneously to obtain the values of b and c. Substituting the values just computed in the equations gives the two equations to be solved as follows:

$$\left.\begin{array}{ll}\text{(A)} & (\Sigma x^2)b + (\Sigma xu)c = \Sigma xy \\ \text{(B)} & (\Sigma xu)b + (\Sigma u^2)c = (\Sigma uy)\end{array}\right\} \left\{\begin{array}{r}17{,}786b + \quad 2{,}052{,}316c = \quad 1{,}396.1 \\ 2{,}052{,}316b + 245{,}881{,}008c = 174{,}171.05\end{array}\right.$$

The simplest way to solve these is by the Doolittle method, as indicated in Appendix I, page 464.

Solving the equations simultaneously gives $b = -0.0879$, $c = 0.001442$. These values are then substituted in equation (14) to obtain the value for a.

$$\begin{aligned} a &= M_y - b(M_x) - c(M_u) \\ &= 3.425 - (-0.0879)(67) + (0.001442)(5{,}378.3) \\ &= +1.56 \end{aligned}$$

With our values for a, b, and c, we can now write out the equation for the parabola, $Y = a + bX + cX^2$ (12), for this particular case as follows:

$$Y = 1.56 - 0.088X + 0.00144X^2$$

Since 10 was subtracted from the percentage of protein before calculating the equation,[5] to estimate the actual percentage 10 must be added back in, making the equation read

affected in the value of a constant

$$Y = 11.56 - 0.088X + 0.00144X^2$$

This then is the equation of the simple parabola which comes nearest to describing the relationships between ~~Y and X.~~ From it the percentage of protein in a given sample of wheat may be estimated from the percentage of hard, dark, vitreous kernels in that sample.

[5] See Note 3, Appendix 2, for proof that this does not affect the values obtained for $\Sigma(x^2)$, $\Sigma(xy)$, etc.

We can see how the estimates are made by working them out for some of the samples. If we take the values of X for the first five samples in Table 18—6, 75, 87, 55, and 34, for example—and substitute them in equation (I) above, we obtain estimated values for Y as follows:

When $X = 6$
$$Y = 11.56 - 0.088(6) + 0.00144(36) = 11.08$$

values of x to y

When $X = 75$
$$Y = 11.56 - 0.088(75) + 0.00144(5625) = 13.06$$

When $X = 87$
$$Y = 11.56 - 0.088(87) + 0.00144(7569) = 14.80$$

When $X = 55$
$$Y = 11.56 - 0.088(55) + 0.00144(3025) = 11.08$$

When $X = 34$
$$Y = 11.56 - 0.088(34) + 0.00144(1156) = 10.23$$

Substituting each of the values of X in the formula in turn in a similar manner, we obtain estimated values for Y as shown in Table 19. So as to distinguish between the actual values of Y, and the values for Y estimated from X according to the equation of the parabola, we shall designate the latter as Y' values.

interpolati

It is quite apparent from the table that the actual and the estimated values generally fall rather near each other, the estimates part of the time being too high and part of the time too low. We can get a better idea of the relation between the estimated and actual values by plotting both on a dot chart (Figure 13), similar to the way we did in Figure 12, using dots as before to represent the values of Y originally observed and crosses to represent the estimated values, Y'. Since the Y' values are all computed from the formula, the crosses all lie on a continuous smooth curve, which we can sketch in freehand, as indicated by the dotted line in the figure. Now if we want to estimate the protein for a sample with a proportion of vitreous kernels not included in our problem, say 65 for example, we can determine it either by substituting 65 for X in equation (I), and computing it out, or by reading from our smooth curve the Y value corresponding to an X value of 65. Of course this *graphic interpolation*, as it is called, will not be quite so exact as will the actual computation, but for many purposes the result will be sufficiently accurate.

Let us now examine Figure 13 and decide whether the formula for the parabola gives a satisfactory "fit" in this case—whether the estimated values do agree fairly well with the actual. We see at once that the curved line of the estimates does come closer to agreeing with the actual values than any straight line could. But on the other hand we see that the general shape of the parabolic curve and the general trend of the actual relationship is rather different. For low proportions of vitreous kernels, the estimated values are generally too low; for the highest proportions, they are also generally too low; whereas for proportions of vitreous kernels ranging from 70 to 95 per cent, the estimates are too high.

TABLE 19

Comparison, for Wheat Problem, of Actual Protein Content with Protein Content Estimated From Per Cent of Vitreous Kernels on Basis of the Simple Parabola

Per cent vitreous kernels, X	Per cent protein (minus 10), Y	Estimated per cent protein (minus 10), Y'	Difference between actual and estimated protein, $(Y - Y')$
6	0.3	1.08	−0.78
75	2.2	3.06	−0.86
87	4.5	4.80	−0.30
55	1.1	1.08	+0.02
34	0.9	0.23	+0.67
98	8.1	6.79	+1.31
91	4.0	5.50	−1.50
45	0.8	0.52	+0.28
51	1.4	0.83	+0.57
17	1.0	0.48	+0.52
36	0.2	0.26	−0.06
97	7.0	6.60	+0.40
74	3.8	2.95	+0.85
24	0.1	0.28	−0.18
85	4.4	4.51	−0.11
96	5.8	6.41	−0.61
92	5.6	5.68	−0.08
94	5.0	6.04	−1.04
84	3.3	4.35	−1.05
99	9.0	6.99	+2.01

curve to fit shows true relationship

Apparently the equation of the simple parabola is not adequate to describe this particular relationship. Especially for high proportions of vitreous kernels, the estimates are quite inaccurate. For 99 per cent vitreous, the parabola would estimate 17.0 per cent protein, whereas both samples over 97 per cent vitreous kernels had over 18 per cent protein. The failure of this curve to give a satisfactory "fit" is not due to any error in the computations but merely to the fact that this formula cannot give the proper-shaped curve to fit the relationship in this case. The mathematical properties of the equation itself are such that, no matter what constants are used for a, b, and c, it cannot come any closer to describing the true relation. The method just used in computing a, b, and c gives the *best* values for this case; any other three values substituted in the same formula would do even less well in "fitting" this particular set of observations.

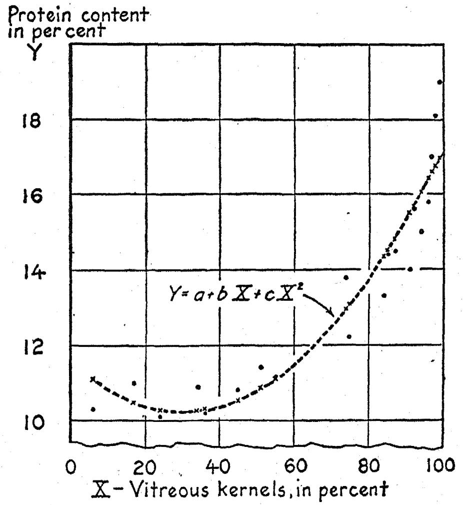


FIG. 13. Dot chart showing relation of vitreous kernels to protein content of wheat, and parabolic curve fitted to same.

"Fitting" a cubic parabola. The cubic parabola, type (f) of the equations on page 76, might be tried to see if it would describe this particular relationship more closely.

The equation of the cubic parabola,

$$Y = a + bX + cX^2 + dX^3 \tag{16}$$

has four constants a, b, c, and d to be computed. Here again, of course, a, b, and c will be different from those we have computed previously, unless the d value comes out zero. The values b, c, and d are computed by the simultaneous solution of the following three equations:[6]

Use U to represent the X^2 of equation (16) and V to represent the X^3.

$$\left.\begin{aligned}(\Sigma x^2)b + (\Sigma xu)c + (\Sigma xv)d &= \Sigma xy\\ (\Sigma xu)b + (\Sigma u^2)c + (\Sigma uv)d &= \Sigma uy\\ (\Sigma xv)b + (\Sigma uv)c + (\Sigma v^2)d &= \Sigma vy\end{aligned}\right\} \qquad (17)$$

The value for a is then computed from the following equation:

$$a = M_y - b(M_x) - c(M_u) - d(M_v) \qquad (18)$$

The values for Σx^2, Σxu, Σxy, Σu^2, and Σuy are computed as shown previously, equations (15). The additional values required in equation (17) are computed as follows:

$$\left.\begin{aligned}M_v &= \frac{\Sigma V}{n}\\ \Sigma uv &= \Sigma UV - nM_uM_v\\ \Sigma xv &= \Sigma XV - nM_xM_v\\ \Sigma v^2 &= \Sigma V^2 - nM_v^2\\ \Sigma vy &= \Sigma VY - nM_vM_y\end{aligned}\right\} \qquad (19)$$

$x^2 + x^3$
$= x^5$

It should be noted that among the values required to "fit" this cubic parabola, ~~that is~~, to determine the constants a, b, c, and d, are such values as ΣV^2 and ΣUV. Remembering that $V = X^3$, and $U = X^2$, we need to calculate X^5 and X^6. For $X = 10$, $X^6 = 1{,}000{,}000$, so for values of X such as those in Table 17, ranging from 6 to 99, it would ~~take a~~ tremendous volume of computation to compute the values required in equations (17), (18), and (19). This may be reduced by letting $U = X^2/100$, and $V = X^3/10{,}000$. The computa-

[6] The alternative method here involves the simultaneous solution of 4 equations, as follows:

$$\begin{aligned}na + (\Sigma X)b + (\Sigma U)c + (\Sigma V)d &= \Sigma Y\\ (\Sigma X)a + (\Sigma X^2)b + (\Sigma XU)c + (\Sigma XV)d &= \Sigma XY\\ (\Sigma U)a + \Sigma(UX)b + (\Sigma U^2)c + (\Sigma UV)d &= \Sigma UY\\ (\Sigma V)a + (\Sigma VX)b + (\Sigma UV)c + (\Sigma V^2)d &= \Sigma VY\end{aligned}$$

tion is not shown here in detail. It follows the general form of that given in Table 18; and the solution of the equations (17), starting in just as shown on page 200, may be most conveniently carried through by the method shown subsequently on page 464.

Even when the cubic parabola is "fitted" to the data given, however, it ~~does~~ not give a satisfactory "fit." Thus Figure 14 shows the cubic parabola fitted to the data, worked out as ~~just described~~. The values found gave the equation

$$Y = 0.35 + 0.0345X - 0.1397(X^2/100) + 0.1788(X^3/10{,}000)$$

or, clearing of fractions,[7]

$$Y = 0.35 + 0.0345X - 0.0014X^2 + 0.000018X^3$$

Adding in the 10 which was subtracted from Y before making the computations, the equation becomes

$$Y = 10.35 + 0.0345X - 0.0014X^2 + 0.000018X^3$$

In Figure 14, the original observations are represented by dots, the estimated values from the cubic parabola are represented by stars, and the curve of the simple parabola is also shown. A curve has been drawn through the stars to show the general shape of the cubic parabola.

The last curve comes much closer than the previous curve to describing the relationship which actually exists. Even so, however, it is not entirely satisfactory, for it gives estimates which are still too low at the very highest percentage of vitreous kernels. Except for this portion, and the downturn at the beginning, it seems quite satisfactory.

How to conclude this

There are still other types of curves, however, some of which might give better fits than the ones we have tried. For instance the fourth-order parabola,

Other curves

$$Y = a + bX + cX^2 + dX^3 + eX^4$$

can be fitted by an extension of the methods just described, as can parabolas with even more terms. Those are rarely useful, however, as the greater the number of terms, the greater the tendency becomes for the curve to "wiggle." In addition, the volume of arithmetic required becomes extremely burdensome—the computations for the fourth-order parabolas involving powers of X up to X^8.

Def

[7] See Note 3, Appendix 2, for proof of this step.

Good

Furthermore, there are only a limited number of observations, 20 in all. If a parabola were fitted with 20 constants, for example, it would simply twist and turn so as to pass through every observation. Since it would simply reproduce these 20 observations, it would be of no value at all in indicating the relation which probably holds true in the universe from which the observations in the sample are drawn. (See Chapters 18 and 22 for further discussion and mathematical measures of this question of the sampling significance of a fitted curve.)

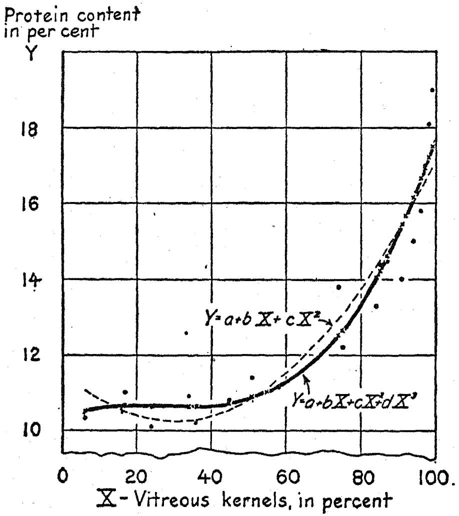


FIG. 14. Dot chart, with parabola and cubic parabola.

Fitting lines or parabolas to time series. In studying time series, it is sometimes desirable to fit a straight line or a curve to the successive observations as a means of determining the long-time trend. The techniques of time-series analysis lie outside the scope of this book, and therefore are not given especial consideration here.[8] Fitting a mathematical trend to a time series involves regarding the successive ~~months or~~ years as values of the *X*, or independent, variable. The fact that these values are regularly spaced, 1, 2, 3, 4, etc., and

[8] An excellent discussion of the methods and meaning of time-series analysis is given by Frederick C. Mills in his textbook, *Statistical Methods*, Chapters VII, VIII, and XI, revised edition, Henry Holt and Co., New York, 1938. See also Max Sasuly, *Trend Analysis of Statistics*, The Brookings Institution, Washington, 1934.

that the same succession reoccurs in many problems, makes possible special methods and special tables, which greatly reduce the labor of fitting the equations. This method of computation, known as *orthogonal polynomials,* should be used in determining lines or parabolic curves for such data.[9]

"Fitting" a logarithmic curve. Some of the other types of curves mentioned on page 76, particularly types b, c, and d, involving logarithms, and type e, using reciprocals, may be fitted with relatively little computation. The methods of fitting one of each of these types may be shown for the present case, even though they may fail to give any better fit than the curves which have already been computed. The three simple types of logarithmic curves, b, c, and d, may all be fitted by exactly the same method previously used in fitting a straight line, except that the logarithms of X, of Y, or of both together are employed where otherwise the values of the variables themselves are used. Comparison of the straight-line formula with the logarithmic formula indicates how this is done.

If we use $\overline{Y}$ to represent the logarithms of the Y values, and $\overline{X}$ to represent the logarithms of the X values, our equations will change as follows:

$$(b)\quad \log Y = a + bX, \text{ to } \overline{Y} = a + bX$$

$$(c)\quad \log Y = a + b \log X, \text{ to } \overline{Y} = a + b\overline{X}$$

$$(d)\quad Y = a + b \log X, \text{ to } Y = a + b\overline{X}$$

In each case it is evident that the new equation is identical in form with the simple straight-line equation,

$$Y = a + bX$$

and the same methods may therefore be used in determining the constants a and b as were used earlier in equations (8) to (11).

Some indication as to which one of the three logarithmic formulas will come nearest to fitting a given set of data can be obtained by converting both the X and Y values to logarithms, variables $\overline{X}$ and $\overline{Y}$, and then making dot charts of $\overline{Y}$ against X, of $\overline{Y}$ against $\overline{X}$, and of Y against $\overline{X}$. If one chart shows the dots falling in substantially a straight line

[9] For methods of fitting orthogonal polynomials, see Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics,* pp. 433–35, Prentice-Hall, Inc., New York, 1940, and R. A. Fisher, *Statistical Methods for Research Workers,* seventh edition, Oliver and Boyd, Edinburgh and London, 1938, pp. 148–155.

Sel [illegible]

the equation corresponding to that chart will give the most satisfactory fit.[10]

The first step in applying any one of the three logarithmic equations to the data of the wheat example is to work out the logarithms

TABLE 20

Variables in Wheat Problem and Logarithms of Values

Per cent protein	Per cent vitreous kernels	Logarithms of Variables:*	
		Protein	Vitreous kernels
Y	X	$\overline{Y}$	$\overline{X}$
10.3	6	1.013	0.778
12.2	75	1.086	1.875
14.5	87	1.161	1.940
11.1	55	1.045	1.740
10.9	34	1.037	1.531
18.1	98	1.258	1.991
14.0	91	1.146	1.959
10.8	45	1.033	1.653
11.4	51	1.057	1.708
11.0	17	1.041	1.230
10.2	36	1.009	1.556
17.0	97	1.230	1.987
13.8	74	1.140	1.869
10.1	24	1.004	1.380
14.4	85	1.158	1.929
15.8	96	1.199	1.982
15.6	92	1.193	1.964
15.0	94	1.176	1.973
13.3	84	1.124	1.924
19.0	99	1.279	1.996

* Logarithms to base 10.

and construct the three dot charts, to indicate which formula to use. The form of computation is shown in Table 20.

[10] This is strictly true only if the "goodness of fit" is measured in terms of the logarithms used.

Logarithms may also be used with parabola of higher orders, such as:

$$\text{Log } Y = a + bX + cX^2$$

Such involved curves will not be considered at length in this book, however.

See below

It should be noted that in working out the logarithms nothing can be added or subtracted from any of the variables (except for rounding off decimals).[11] In all the previous work the protein had been stated as protein in excess of 10 per cent, but now the original percentage figures are used once more. That is because logarithms deal with *relative* values, and the relation of 1 to 2 is quite different from the relation of 11 to 12. All the previous equations have dealt with absolute values or differences from the average; and the absolute difference between 1 and 2 is of course just the same as that between 11 and 12.

Reason why

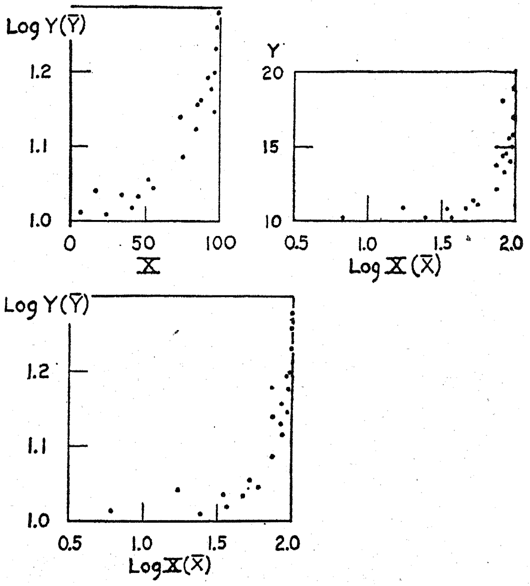


FIG. 15. Dot charts illustrating log $Y=f(X)$; $Y=f(\log X)$; log $Y=f(\log X)$.

Figure 15 gives the three dot charts in which the three different ways of combining the logarithmic and actual values are shown. None of the three gives a very close linear relation, but the one where $\bar{Y}$ and X are plotted seems to come nearest. The equation

$$\log Y = a + bX, \quad \text{or} \quad \bar{Y} = a + bX$$

to a straight line

will therefore be used.

[11] After the logarithms are once computed, however, they can be "coded" by subtracting a constant or by division, just as other variables have been treated formerly, with the same effect on the final constants obtained.

See above

The values necessary to determine a and b are as follows, using equations (9) and (10):

$$\checkmark\ \Sigma X\overline{Y},\quad M_x,\quad M_{\overline{y}},\quad \Sigma X^2$$

Table 21 shows in full the computation of these values from the original values of the two variables.

TABLE 21

COMPUTATION, FOR WHEAT PROBLEM, OF VALUES NEEDED TO DETERMINE CONSTANTS FOR LOGARITHMIC CURVE

Per cent protein Y	Per cent vitreous kernels X	Logarithms of Y $\overline{Y}$	Extensions	
			X^2	$X\overline{Y}$
10.3	6	1.013	36	6.078
12.2	75	1.086	5,625	81.450
14.5	87	1.161	7,569	101.007
11.1	55	1.045	3,025	57.475
10.9	34	1.037	1,156	35.258
18.1	98	1.258	9,604	123.284
14.0	91	1.146	8,281	104.286
10.8	45	1.033	2,025	46.485
11.4	51	1.057	2,601	53.907
11.0	17	1.041	289	17.697
10.2	36	1.009	1,296	36.324
17.0	97	1.230	9,409	119.310
13.8	74	1.140	5,476	84.360
10.1	24	1.004	576	24.096
14.4	85	1.158	7,225	98.430
15.8	96	1.199	9,216	115.104
15.6	92	1.193	8,464	109.756
15.0	94	1.176	8,836	110.544
13.3	84	1.124	7,056	94.416
19.0	99	1.279	9,801	126.621
Sums......	$\Sigma X = 1{,}340$	$\Sigma\overline{Y} = 22.389$	$\Sigma X^2 = 107{,}566$	$\Sigma X\overline{Y} = 1{,}545.888$

This computation gives the values necessary to compute a and b by formulas (9) and (10).

The averages of X and $\overline{Y}$ of course are:

$$M_x = \frac{\Sigma X}{n} = \frac{1,340}{20} = 67.0$$

$$M_{\bar{y}} = \frac{\Sigma \overline{Y}}{n} = \frac{22.389}{20} = 1.11945$$

Then

$$b = \frac{\Sigma X\overline{Y} - nM_xM_{\bar{y}}}{\Sigma X^2 - nM_x^2} = \frac{1,545.888 - 20(67)(1.11945)}{107,566 - 20(67)^2} = 0.002576$$

and

$$a = M_{\bar{y}} - b(M_x) = 1.11945 - (0.002576)(67) = 0.9469$$

In terms of the variable, the equation required is therefore

$$\overline{Y} = a + bX = 0.9469 + 0.002576X$$

or

$$\log Y = a + bX = 0.9469 + 0.002576X$$

The percentage of protein can now be estimated from the proportion of vitreous kernels observed for any sample of wheat, by substituting the percentage of vitreous kernels (the X values) in this equation and working it out. Thus for the first example, with 6 per cent of vitreous kernels, it would work out as follows:

Log Y, when x = 6

$$\log Y = a + bX = 0.9469 + 0.0026(6)$$

$$\log Y = 0.9624$$

Using a table of logarithms we find that the number corresponding to the logarithm 0.9624 (that is to say, its antilogarithm) is 9.17. The estimated proportion of protein is therefore 9.17 per cent.

Similarly if the proportion of vitreous kernels in the second sample, 75, is substituted in the equation, the work to calculate the estimated proportion of protein is:

Log Y, when x = 75

$$\log Y = a + bX = 0.9469 + 0.002576(75)$$

$$\log Y = 1.1401$$

$$\text{antilog } 1.1401 = 13.81$$

Imp

The estimated proportion of protein is therefore 13.81 per cent.

Table 22 shows this computation carried through for each of the 20 observations.

TABLE 22

COMPUTATION, FOR WHEAT PROBLEM, OF ESTIMATED PROTEIN CONTENT FROM PER CENT OF VITREOUS KERNELS ON THE BASIS OF A LOGARITHMIC CURVE

($\text{Log } Y = 0.9469 + 0.00258X$)

Col 5 —

Per cent vitreous kernels X	Estimated per cent protein		Actual per cent protein Y	Percentage errors in estimating protein proportion $100\left(\frac{Y}{Y'} - 1.00\right)$
	Estimated logarithm $\bar{Y}'$	Antilog of estimate Y'		
6	0.9624	9.2	10.3	+12.0
75	1.1401	13.8	12.2	−11.6
87	1.1710	14.8	14.5	− 2.0
55	1.0888	12.3	11.1	− 9.8
34	1.0345	10.8	10.9	+ 0.9
98	1.1993	15.8	18.1	+14.6
91	1.1813	15.2	14.0	− 7.9
45	1.0628	11.6	10.8	− 6.9
51	1.0783	12.0	11.4	− 5.0
17	0.9907	9.8	11.0	+12.2
36	1.0396	11.0	10.2	− 7.3
97	1.1968	15.7	17.0	+ 8.3
74	1.1375	13.7	13.8	+ 0.7
24	1.0087	10.2	10.1	− 1.0
85	1.1659	14.7	14.4	− 2.0
96	1.1942	15.6	15.8	+ 1.3
92	1.1839	15.3	15.6	+ 2.0
94	1.1890	15.5	15.0	− 3.2
84	1.1633	14.6	13.3	− 8.9
99	1.2019	15.9	19.0	+19.5

Inf — Note on Col. 5 in Table 22

It should be noted in this table that errors made in estimating the proportion of protein are stated as relative errors rather than absolute errors. That is done because the thing that is really estimated is the logarithm of the percentages of protein, or $\bar{Y}$, and the errors are really the differences between the actual logarithms and the estimated logarithms. If z is used to stand for the error, in this case z is really in terms of logarithms, that is:

$$z = \log Y - \text{estimated } \log Y, \text{ or } \bar{Y} - \bar{Y}'$$

or in terms of natural numbers:

$$\text{anti-log } z = \frac{\text{antilog } \overline{Y}}{\text{antilog } \overline{Y}'} = \frac{\text{actual } Y}{\text{estimated } Y}$$

See below footnote

Subtracting the constant 1.00 and multiplying by 100 changes this relative figure to the percentage which the observed value is above or below the estimate.[12]

Where log Y is taken as the dependent variable, as has been done here, fitting the equation by the methods just shown involves making the square of the *logarithmic* residuals around the line as small as possible. That means that instead of minimizing the sum of the *absolute* errors, squared, as heretofore, we now minimize the sum of the *percentage* errors, squared. In some cases it may be desired to use the logarithmic curve, yet to continue to minimize the absolute errors. Relatively simple methods are available to accomplish that result.[13]

[12] The reason for making this distinction will be seen later on, when the question of measuring the accuracy of the estimate is taken up.

[13] To fit the equation

$$\log Y = a + b(\log X)$$

under the conditions that the sum of the squares of the *absolute* departures of the estimated values, Y', from the actual values, Y, will be as small as possible, determine the values of a and b by solving the equations

$$\Sigma(Y^2)a + \Sigma(Y^2\overline{X})b = \Sigma Y^2\overline{Y}$$
$$\Sigma(Y^2\overline{X})a + \Sigma(Y^2\overline{X}^2)b = \Sigma Y^2\overline{X}\,\overline{Y}$$

where $\overline{Y} = \log Y$, and $\overline{X} = \log X$, as above.

To compute the several sums involved in these equations, the following form may be used:

X	Y	Y^2	$\overline{X}$	$\overline{Y}$	$Y^2\overline{X}$	$Y^2\overline{X}\,\overline{Y}$	$Y^2\overline{Y}$	$Y^2\overline{X}^2$
6	10.3	106.09	0.778	1.013	82.54	83.61	107.47	64.21
75	12.2	148.84	1.875	1.086	279.08	303.08	161.64	523.27
..								
..								
Sums	—	ΣY^2	—	—	$\Sigma Y^2\overline{X}$	$\Sigma Y^2\overline{X}\,\overline{Y}$	$\Sigma Y^2\overline{Y}$	$\Sigma Y^2\overline{X}^2$

The two simultaneous equations can be solved conveniently by the same procedure described in Appendix 1, page 464.

For the derivation of these equations, see W. Edwards Deming, *Some Notes on Least Squares*, pp. 136–141. U. S. Department of Agriculture Graduate School, Washington, 1938.

In Figures 16 and 17 the actual proportions of protein, shown as dots, are compared with the estimated values as worked out by the logarithmic relation. In the first of these figures the actual and estimated values are both stated in terms of the logarithms. It is quite apparent here that this equation assumes a straight-line relation between the proportion of vitreous kernels and the logarithms of the proportion of protein; since they were computed by a straight-line equation (log $Y = a + bX$) the estimated values all lie along the continuous straight line indicated. The next figure, however, compares the actual proportion of protein with the estimated, both stated in actual terms. Here the continuous curve which the logarithms produce in the estimated actual values is clearly shown. The relation between the proportion of vitreous kernels and the percentage of protein, as shown by this curve, does not agree with the actual relation as shown by the original observations even as closely as did the previous curves computed by means of parabolic equations.

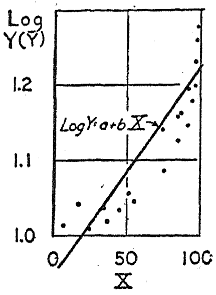


FIG. 16. Dot chart showing observations and fitted line for equation log $Y = a + bX$, in logarithms of Y.

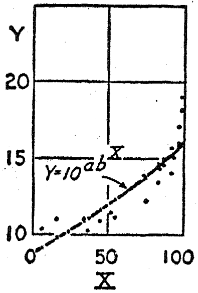


FIG. 17. Dot chart showing observations and fitted line for equation $Y = 10^{ab^X}$, in natural values of Y.

Before discussing other ways of expressing the curvilinear relation it might be well to discuss the procedure to determine the constants a and b if either of the other two forms of simple logarithmic equations were used.

If the equation $Y = a + b \log X$ is employed, the form $Y = a + b\overline{X}$ is used.

The values which must be computed are

$$M_y, \quad M_{\bar{x}}, \quad \Sigma Y\overline{X}, \quad \Sigma \overline{X}^2$$

and the constants are determined from the equations

$$b = \frac{\Sigma Y\overline{X} - nM_yM_{\bar{x}}}{\Sigma\overline{X}^2 - nM_{\bar{x}}^2}$$

$$a = M_y - bM_{\bar{x}}$$

Since the equation is in terms of Y itself, the estimated values, computed from the logarithms of X, will be directly in values of Y, and will not have to be converted to the antilogarithms.

6 If the equation $\log Y = a + b \log X$ is to be fitted, the form $\overline{Y} = a + b\overline{X}$ is used.

The values which will have to be computed are:

$$M_{\bar{y}}, \quad M_{\bar{x}}, \quad \Sigma\overline{Y}\overline{X}, \quad \Sigma\overline{X}^2,$$

and the constants are determined from the equations

$$b = \frac{\Sigma\overline{Y}\overline{X} - nM_{\bar{y}}M_{\bar{x}}}{\Sigma\overline{X}^2 - nM_{\bar{x}}^2}$$

$$a = M_{\bar{y}} - bM_{\bar{x}}$$

In this case the equation is in terms of $\overline{Y}$, the logarithms of Y, and the estimated values will therefore have to be converted from logarithms into natural numbers to show just what the relationship is, just as was done in the case that was worked out in detail earlier.

It is evident that no matter which one of the three logarithmic curves is employed, the arithmetic is exactly the same as in determining the simple straight line, with the exception of computing the logarithms and of substituting the appropriate logarithms where the actual values would otherwise be employed.

In cases where other modifications of the straight-line equation, such as type (e), are to be used, the process is to transform the equation to a linear form, then compute the constants just as before. Thus the type

$$✓ \quad Y = \frac{1}{a + bX}$$

Type e equation

can be converted to the form

$$\frac{1}{Y} = a + bX$$

or, letting $\frac{1}{Y} = Q$,

$$Q = a + bX$$

The computation can then be carried out in the usual way, and after the estimated values of Q, Q', are worked, converted back into Y values by the equation $Y' = \frac{1}{Q'}$.

Limitations of equations in describing relationships. Up to this point an expression of the relation between the proportion of vitreous kernels and the proportion of protein in each sample has been worked out on the basis of a number of different mathematical formulas. Each different equation has given a different curve. Some, such as the cubic parabola or the logarithmic curve, have given curves coming somewhere near to the relationship shown by the actual observations themselves; others, such as the simple straight line, have entirely failed to describe the relation. Yet the exact slope or shape of each curve was determined from the same set of observations; the constants of each curve were determined by "fitting" the same data. The diversity in the shape of the different curves is strikingly shown in Figure 18, where the several different curves are all drawn on one scale, and the original observations are shown as well. It is quite apparent that the differences in the shapes of the several curves are due solely to the particular form of equation used in computing them. There are certain types of relations which can be accurately represented by each of these equations. When it is "fitted" to data where that type of relation is really present, it can give a curve which accurately represents the true relation shown by the data. When, however, as in the present case, an attempt is made to represent a relation by an equation which does not truly express the nature of the relation, the resulting curve gives only a distorted representation

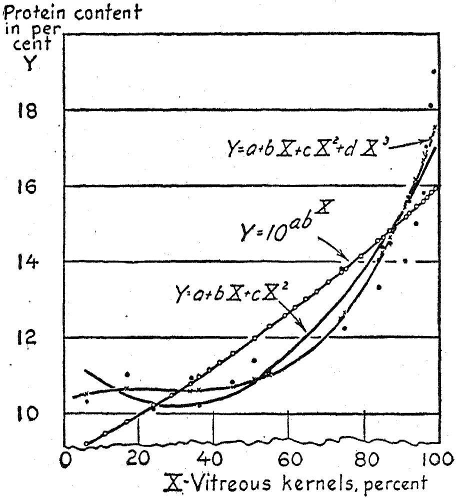


FIG. 18. Original observations, and several different types of fitted curves.

false equation

of the true relation—*it shows the relation only insofar as it is possible to do so within the limits of the particular equation used.*

So far there has been no attempt to show what there is in the "nature" of relations which may make them of the type to be represented accurately by one type of equation or by another. Instead, the purely empirical test of the way each one fits has been relied upon. If, as judged by the eye, the relation shown by the fitted curve *looked like* the relation shown by the original observations, we have said it gave a satisfactory fit; if it has not looked like it, we have said it did not give a satisfactory fit. And in this particular case, none of the computed curves has been really fully satisfactory—we can readily see that there might be some other smooth continuous curve which would come much closer to the actual observations than does any of the curves so far computed.

how to know the accuracy of the fit

Of course we might continue the process, using more and more complex equations, until finally we found one which did satisfactorily describe the relation. Or we might find that *no* ordinary mathematical expression would describe the relation. It might be that the underlying curve was so complex that it could not be represented in elementary algebraic terms. But even if we could describe the relation satisfactorily by some type of equation, the only advantage would be that then we would have some way of estimating values of the dependent variable (percentages of protein) from the independent variable (proportion of vitreous kernels) such as would agree reasonably well with the values actually observed. So long as the equation had been derived merely by the "cut-and-try" method described, it would have no meaning beyond serving as a simple device for estimating values of the one variable from known values of the other and would throw no particular light upon the real or inherent nature of the relation. For if we could find, by enough trying, one equation which would represent the relation satisfactorily, it might be that we could also find another. As a matter of fact, sometimes it is found that two different types of equations may each give exactly the identical curve when figured out.[14] Which one expresses the "true" nature of the relation? Merely because a given equation *can* reproduce a certain relation is no proof that it really "expresses" the nature of the relation. Something more must

advantage of the effort

[14] An example of this type may be seen in the bulletin, What makes the price of oats, by Hugh Killough, *U. S. Department Agriculture Bulletin* 1351, page 8. Here equations of two different types were found to yield almost identical curves, within the range covered by the observations studied.

be known than merely that it *can* express the relation. What that something is will be taken up in a later section.

Alternative method

If, however, it is not desired to determine what the "real nature" of the relationship is, but it is merely desired to express it sufficiently well so that values of one variable (such as protein content) can be estimated from known values of another (such as the proportion of vitreous kernels), it does not make any difference what type of equation is used, so long as it represents the observed relationship adequately. As a matter of fact, it is not really necessary to have an equation at all. If we have only a graph of the curve, or a table of values for one variable corresponding to values of another, from which we can construct a graph, that is all that is really necessary. For if we have a graph of the curve we can very readily estimate the value for one variable from corresponding known values for another by simply reading it from the curve. Thus in Figure 13 the curve for the equation

$$Y = a + bX + cX^2$$

is shown. If we wish to estimate the percentage of protein for a sample having, say 50 per cent of vitreous kernels, we need only to run up the line for $X = 50$ and note the value of Y corresponding to that point on the curve. In this case it is apparently about 10.8 per cent. Similarly, the estimates of the percentage of protein corresponding to any other percentage of vitreous kernels within the range covered by the curve may be read off directly from the curve. Further, by enlarging the chart and making the scale sufficiently detailed, we may read off the estimated values to any degree of accuracy that is desired —much more accurately, as a matter of fact, than our ability to determine the real relation usually justifies, as will be evident later on.

In many cases—perhaps in the great majority of cases—simply the working expression of the relation may be all that is either needed or desirable. The "true relation" between the variables may be so involved that a very complex mathematical expression would be required to represent it properly. Even simple types of physical relations may require rather complex curves to represent them. In many cases, too, the knowledge of the causes of the relation may be so undeveloped that there is no real basis for expressing the relationship mathematically. The relation between vitreous kernels and percentage of protein would be an example of this type—very complex details of chemical content and physical and biological structure are probably responsible, so complex as to be quite beyond satisfactory reduction to mathematical expression. Yet the original observations undeniably indicate that there is some sort of definite relation. For many practical purposes it may be entirely satisfactory merely to know what the relationship *is*, without bothering at all with what it really means. Even in scientific study that may frequently be satisfactory as a first step, since in many cases it is essential to know what are the facts before trying to work out the reasons *why* they are as they are.

When the expression of the relation is not to be used except as an empirical basis for estimating values of the dependent variable from the independent, or for showing just what the relationship *is*, the elaborate technique of determining the constants of a mathematical equation and working out the estimated values by the use of that equation becomes largely unnecessary. In many cases a curve can be determined with only a small fraction of the effort required in "fitting" a mathematical equation, yet it fits the data quite as well as any mathematical curve. In such cases the curve may afford quite as satisfactory a description of the relation and a basis for estimating one variable from the other as if elaborate computations had been made. This method is known as freehand smoothing.

Free-hand Curve

Expressing a curvilinear relation by a freehand curve. The process of determining a freehand curve may be very simply illustrated. In fact, it has already been suggested in much of the previous discussion. The very simplest way to do it would be to plot the original observations on coordinate paper, just as has been shown so many times before, and then draw a continuous smooth curve through them by eye in such a way as to pass approximately through the center of the observations all along its course. Where the nature of the relation is indicated quite as closely by the original observations as it is in the wheat problem which we have been discussing, this might yield quite a satisfactory expression of the relation. In other cases, however, the observations might be more widely scattered, and the underlying relation might be more difficult to determine, so that different persons, drawing in the curves freehand, might draw in rather different curves. Some method is therefore needed to give a greater degree of precision to the result, and to insure that the same data would yield substantially the same result even in the hands of different investigators.

Average method

This stability of result can be secured by a relatively minor extension of the methods already discussed in the first illustration of a two-variable relationship—the automobile-stopping problem. There

Average curve

it was found that by classifying the observations in appropriate groups, the general nature of the relation could be expressed by an irregular line connecting the several group averages. ~~All that~~ is needed is some method of deriving a continuous smooth curve from that irregular line. Smoothing out that irregular line, freehand, is a very evident and simple method. At the same time, starting with the irregular line of group averages gives a certain stability to the process and insures that different persons would draw in the curve with about the same position and shape.

TABLE 23

COMPUTATION OF AVERAGES TO USE IN FITTING FREEHAND CURVE, FOR WHEAT-PROTEIN PROBLEM

	Vitreous kernels below 25 per cent		Vitreous kernels 25 to 49 per cent		Vitreous kernels 50 to 74 per cent		Vitreous kernels 75 to 100 per cent	
	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein
	6	10.3	34	10.9	55	11.1	75	12.2
	17	11.0	45	10.8	51	11.4	87	14.5
	24	10.1	36	10.2	74	13.8	98	18.1
	....	......	....	......	....	......	91	14.0
	....	......	....	......	....	......	97	17.0
	....	......	....	......	....	......	85	14.4
	....	......	....	......	....	......	96	15.8
	....	......	....	......	....	......	92	15.6
	....	......	....	......	....	......	94	15.0
	....	......	....	......	....	......	84	13.3
	....	......	....	......	....	......	99	19.0
Totals....	47	31.4	115	31.9	180	36.3	998	168.9
No. cases.	3	....	3	....	3	....	11	....
Averages.	15.67	10.47	38.33	10.63	60.00	12.1	90.73	15.35

Smoothing the average curve

Applying the process to the wheat problem, the first step is to classify the data into appropriate groups according to the values of the independent variable, the proportion of vitreous kernels, and to determine the average percentage of vitreous kernels and of protein content for the observations falling into each group. The discussion of the automobile problem has shown that, for the differences in averages to be significant, it is necessary for the groups to be large enough so that the averages would not vary erratically from group to group. In some cases a little experimenting might be necessary to determine what this size would be. In the present case, inspection of the dot chart showing the original observations (Figure 12, page 83) indicates that a class interval of 25 per cent of vitreous kernels will give groups large enough to make the averages of protein content fairly stable from group to group.

The form of computation most convenient to obtain the group averages, using groups of the size suggested, is shown in Table 23.

The averages for the several groups are shown in Figure 19, indicated by hollow circles, whereas original observations are again shown by solid dots. A smooth continuous dashed curve has been drawn through the series of group averages, ignoring the individual observations and following only the general trend shown by the averages. This smooth curve comes quite near to representing the relation shown by the individual observations through most of its extent; but beyond 95 per cent of vitreous kernels it fails to follow the individual observations—through that portion of the range the protein content rises much faster than is indicated by the average for the whole range from 75 through 100 per cent vitreous kernels.

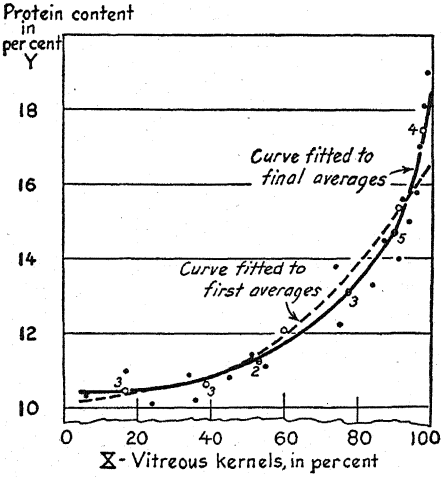


FIG. 19. Original observations and averages of protein content, and freehand curve.

Defects of the first average line

Because over half of all the observations fall in this upper portion of the range, it would seem reasonable to classify them into smaller

need for another average

Part 2 data re-classified into smaller groups

groups so as to give a better basis for determining this portion of the curve. Let us try splitting the observations above 50 into four groups, each with about the same number of observations—say 50 to 69, 70 to 84, 85 to 94, and 95 to 100. The computation of the new averages is shown in Table 24.

TABLE 24

Computation of Sub-averages for Last Groups in Wheat Problem, for Fitting Freehand Curve

	Vitreous kernels 50 to 69 per cent		Vitreous kernels 70 to 84 per cent		Vitreous kernels 85 to 94 per cent		Vitreous kernels 95 tŏ 100 per cent	
	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein	Per cent vitreous kernels	Per cent protein
	55	11.1	75	12.2	87	14.5	98	18.1
	51	11.4	74	13.8	91	14.0	97	17.0
	....	....	84	13.3	85	14.4	96	15.8
	....	....	....	....	92	15.6	99	19.0
	....	....	....	....	94	15.0	....	....
Totals....	106	22.5	233	39.3	449	73.5	390	69.9
No. cases.	2	....	3	....	5	....	4	....
Averages.	53	11.25	77.67	13.1	89.8	14.7	97.5	17.48

These new averages, together with the previous ones for the lower groups, are also plotted in Figure 19, and the number of cases that each represents is indicated next to it, to aid in judging what weight to assign to that average. Finally, a smooth continuous curve has been drawn in, to pass as near as possible to the different averages without making illogical twists or turns. As is evident in the figure, it has been possible to draw the line with no point of inflection in it, yet so that it passes quite near to all the group averages and approximately through the middle of the individual observations. Further, the general course of the line is sufficiently well defined by the several group averages so that if it were redrawn, either by the same person or another person, it could have only minor differences from the line actually shown. Making the chart over two or three times, and drawing a separate curve on each trial, then averaging the two or three curves together, is one method of reducing the variation due to individual judgment in drawing the curve.

Cautions in freehand fitting. In drawing in the freehand curve no attempt has been made to have the curve follow all the twists and turns of the irregular line of averages. As was shown previously with the automobile illustration, these irregular differences from group to group may very readily be due to chance fluctuations in sampling where the groups are small. Not unless the groups included a very much larger number of cases than these do here would one be justified in bending the curve because of the position of a single group average, and not even then unless there was some logical basis for a curve of that shape. In doubtful cases breaking up a particular group into smaller groups, as was just done in the wheat example, or reclassifying the observations into somewhat different groups, will help to determine whether or not the data positively indicate that an extra inflection is needed. It is also necessary to see if some single observation is responsible for the abnormality; if it is, it is better to disregard it and draw the curve without the extra twist.

The case of twisting a curve

In drawing in a freehand curve, it is desirable to place certain logical limitations on the shape of the curve rather than to have it be purely an empirical representation of the data. To do this, it is necessary to decide before the curve is drawn what those limitations should be. The limitations should be based upon a logical analysis of the relation under examination, in the light of all the information available to the investigator. In this case, for example, a consideration of the biological structure of the kernels, of the portions which run high in protein content, and of the appearance and size of those portions might lead one to the following conclusions:

Logical limitations in the wheat problem free hand curve

✓(*a*) An increase in the proportion of vitreous kernels might be associated with no change in the proportion of protein, or with an increase in the proportion, but never with a decrease in the proportion.

✓(*b*) The relation between vitreous kernels and protein should be a progressive one, consistently changing throughout the range of variation, rather than fluctuating back and forth.

✓(*c*) The maximum proportion of protein would be found with the largest proportion of vitreous kernels.

These three logical expectations might then be expressed in the following limitations to be placed on the shape of the curve to be drawn:

(1) The curve should have no negative slope throughout its length.

(2) The curve should have no points of inflection, but should change shape continuously and progressively.

✓(3) The maximum should be reached at the end of the curve.

These three logical limitations are all fulfilled by both the curves shown in Figure 19, yet they would exclude other types of curves which might be drawn. For example, they would rule out a curve with a hump or twist in it, or one which sloped down and then up.[15]

In some cases, examination of the data by the method of successive group averages, even after all the tests suggested above, will show the presence of a relation which cannot be expressed within the logical limitations imposed on the shape of the curve. In that case, the reasoning underlying the logical analysis should be reexamined, to see if some step requires restatement and if the limitations themselves should be changed. (For a further discussion of this interaction of induction and deduction, see pages 443 to 452 of Chapter 24.) For a curve to have real meaning, it must be consistent with a careful logical analysis, no matter whether the curve is obtained mathematically or freehand, or whether the logical limitations are expressed in a mathematical equation or in a set of limitations placed on the shape of the curve drawn by freehand fitting.[16]

which? the free hand curves

Interpreting the fitted curve. It is evident that the freehand curve comes closer to agreeing with all the original observations than did any of the mathematically determined curves. So far as can be judged by eye alone, it "fits" the relation actually observed quite satisfactorily. So far as giving a definite statement of the relation, and serving as a basis for estimating values of one variable from known values of the other, this freehand curve, obtained by the very simple process shown, is more satisfactory than any of the curves obtained by the mathematical computations.

The use of the freehand curve in estimating values of the dependent variable, percentage of protein, from known values of the independent variable, proportion of vitreous kernels, may be readily illustrated. Taking the first observation, with 6 per cent of vitreous kernels, and reading off the corresponding proportion of protein from the curve

[15] This use of logical analysis in stating the limitations on a freehand curve may be compared with the use of logic in deciding on the type of mathematical equation to employ. Note the subsequent section in this chapter on "The logical significance of mathematical functions."

[16] For a more detailed discussion of the pros and cons of freehand versus mathematical fitting, see W. Malenbaum and J. D. Black, The use of the short-cut graphic method of multiple correlation, *Quarterly Journal of Economics*, Vol. LII, November, 1937, and The use of the short-cut graphic method of multiple correlation: comment, by Louis Bean, and Further comment, by Mordecai Ezekiel, and Rejoinder and concluding remarks, by Malenbaum and Black, *Quarterly Journal of Economics*, February, 1940.

in Figure 19, we get 10.4 per cent as the estimated protein content. Similarly for the second observation, 75 per cent vitreous kernels, the curve indicates 12.9 per cent as the proportion of protein. Reading off the estimated protein for each of the 20 observations we get the estimates shown in Table 25.

Even though in using the freehand curve we do not have an equation stating the relation between X and Y, we still have a mathematical expression of the relation between them. For we can write

Comparing actual observed results with estimates on basis of freehand curve

TABLE 25

ACTUAL PER CENT OF PROTEIN AND PROPORTION ESTIMATED ON BASIS OF FREEHAND CURVE

Proportion of vitreous kernels X	Actual proportion of protein Y	Proportion of protein estimated from vitreous kernels $Y' = f(X)$	Difference between actual and estimate $Y - Y'$
6	10.3	10.4	−0.1
75	12.2	12.9	−0.7
87	14.5	14.5	0
55	11.1	11.4	−0.3
34	10.9	10.7	0.2
98	18.1	17.4	0.7
91	14.0	15.2	−1.2
45	10.8	11.1	−0.3
51	11.4	10.3	1.1
17	11.0	10.5	0.5
36	10.2	10.8	−0.6
97	17.0	17.0	0
74	13.8	12.8	1.0
24	10.1	10.6	−0.5
85	14.4	14.2	0.2
96	15.8	16.7	−0.9
92	15.6	15.5	0.1
94	15.0	15.9	−0.9
84	13.3	14.0	−0.7
99	19.0	18.0	1.0

$$Y' = f(X)$$

which simply means that the estimates~~, or Y'~~ values, are *a function of X*; that is, for every X value there is some corresponding Y'

Graphic vs Analytical Relation Statement

value. Of course, we can find what this corresponding value is only by reading it off the curve; yet that is enough. We have a graphic statement of the functional relation; if we had a definite formula to represent the curve, we would have an *analytical* statement of the relation as well.

Although we do not have a definite equation to represent the free-hand curve, it is still possible to state the relation shown by the curve other than in graphic form. This can be done by constructing a table showing, for whatever values of the independent variable may be selected, the corresponding estimated values of the dependent variable. Such a tabular statement of the relation may be more readily comprehended by readers not accustomed to graphic presentation. Further, it provides a basis for reconstructing the curve on any scale desired for the purpose of making further estimates. Table 26 illustrates this method of stating the relation.

Tabular Statement

TABLE 26

Per Cent of Protein Corresponding to Various Proportions of Vitreous Kernels in Samples of Wheat, as Indicated by 20 Observations

Proportion of vitreous kernels	Corresponding proportion of protein	Proportion of vitreous kernels	Corresponding proportion of protein
Per cent	*Per cent*	*Per cent*	*Per cent*
10	10.4	70	12.4
20	10.5	80	13.5
30	10.7	90	15.0
40	10.9	95	16.2
50	11.2	99	18.0
60	11.7		

In the range where the curve is rising most steeply the readings are taken more closely together, to provide for reproducing that portion of the curve more accurately. In addition, no readings are taken beyond the range covered by the original observations, nor are any shown for the extreme ends where the observations are few. This raises the whole question of how curves like this can serve as a basis for estimating when measurements are made of the independent variable, such as proportion of vitreous kernels, in cases other than those used in determining the relation. This problem will be taken up at the end of this chapter. But first the question of whether to use freehand or analytical curves will be discussed.

The logical significance of mathematical functions. There has been frequent reference previously to the question whether an equation did or did not express "the real nature" of a relationship, with little explicit attempt to explain exactly what that meant. To know when we are justified in using the simple freehand curve, and when we should go to the additional work of determining an equation for the curve, we must understand the logical bases for different types of equations, so that we can judge whether or not any particular type of curve can logically be expected to express the relation in any given set of observations.

The linear equation. Many relations are so simple that ordinarily we would not think of expressing them mathematically. Thus, if a train is traveling 45 miles an hour, the distance traveled is equal to the time multiplied by the speed. Using t for the time in hours, d for distance, and s for speed, the relation is obviously

$$d = st$$

This is a simple straight-line relation. Now, if, in addition, the train were a miles away from a given station at the beginning, after t hours of additional travel away from the station it would be D miles away, where

$$D = a + d = a + st$$

This is now expressed in the usual form for the straight-line equation, $Y = a + bX$. This equation is therefore the one to be used when it can logically be expected that each unit change in X causes a corresponding change in Y, regardless of the size of X. Thus in computing the distance the train has traveled we are assuming that it continued to travel at a definite rate, say 45 miles an hour, the whole way, and traveled the 200th mile just as fast as the first mile. Now if we were dealing with something where the change in Y was not the same for different values of X, the equation would no longer be satisfactory. For example, an airplane on a long-distance flight has to carry a heavy load of gasoline at the start and hence cannot attain full speed; the farther it goes the lighter its load becomes and the higher speed it can make. In such a case the straight-line formula would not be applicable, since the speed of the plane would increase with the distance it had gone. If the straight-line formula were used, it would indicate that it would take just as long to travel the first hundred miles as the last hundred, whereas actually it would take longer than that to travel the first hundred and less than that to travel the final hundred. Only an equation which included some value that properly took into account the change in speed with the change in distance could satisfactorily represent this relation.

1. ✓*The quadratic equation.* Another case in which the rate at which Y increases changes as the value of X increases is that of a weight falling to the ground. Since the attraction of the earth is for practical purposes a constant, it exercises a constant pull on a falling body. Thus, the farther a body falls, the faster it travels. It is just as if, in throwing a ball, a boy did not let go the ball for it to travel by its momentum but was able to keep shoving against it, adding more and more speed to the momentum it already had. Physicists express this relation by saying that the velocity with which an object falls is accelerated at a constant rate. This equation, therefore, is:

$$✓ V = gt$$

where g is a constant measuring the force of gravity, V is velocity in feet per second, and t is time in seconds.

With regard to the distance a body will fall in any given time, therefore, the case is much the same as with our airplane. The velocity, or speed, is increasing with every passing moment, and therefore the distance traveled in each succeeding second will be greater than the distance traveled in the previous second.

If we assume that the value of g in the equation is already known to be 32, the equation

$$V = gt$$

can then be written

$$V = 32t$$

We can then estimate the distance traversed by a falling body in each successive second by a process of approximation like this:

Let us figure that the average speed for each 2 seconds is the same as at the midpoint (which may not be exactly right) and then let us estimate the distance traversed in those 2 seconds by multiplying this average speed by the time. Then by adding all the distances together we can get an approximation of the total distance.

First we need to calculate the average speed for each period, using the last equation, $V = 32t$:

End of 1st second, speed = $32(1) = 32$ = average speed for 1st two seconds
End of 3d second, speed = $32(3) = 96$ = average speed for 2d two seconds
End of 5th second, speed = $32(5) = 160$ = average speed for 3d two seconds
End of 7th second, speed = $32(7) = 224$ = average speed for 4th two seconds
End of 9th second, speed = $32(9) = 288$ = average speed for 5th two seconds

Then we can estimate the distance traveled in each 2-second period, as follows:

Period	Average speed, feet per second	Distance in that period, feet
1st	32	64
2d	96	192
3d	160	320
4th	224	448
5th	288	576
Estimated total distance................		1600

Another estimate could be obtained by estimating the distance for each second separately, for there might be less error in assuming that the speed at the middle of each second would represent the average for that second. On this basis the problem would work out.

Speed at middle of 1st second = $32\ (\frac{1}{2}) = 16$; distance in that second = 16
Speed at middle of 2d second = $32(1\frac{1}{2}) = 48$; distance in that second = 48
Speed at middle of 3d second = $32(2\frac{1}{2}) = 80$; distance in that second = 80
Speed at middle of 4th second = $32(3\frac{1}{2}) = 112$; distance in that second = 112
Speed at middle of 5th second = $32(4\frac{1}{2}) = 144$; distance in that second = 144
Speed at middle of 6th second = $32(5\frac{1}{2}) = 176$; distance in that second = 176
Speed at middle of 7th second = $32(6\frac{1}{2}) = 208$; distance in that second = 208
Speed at middle of 8th second = $32(7\frac{1}{2}) = 240$; distance in that second = 240
Speed at middle of 9th second = $32(8\frac{1}{2}) = 272$; distance in that second = 272
Speed at middle of 10th second = $32(9\frac{1}{2}) = 304$; distance in that second = 304

In 10 seconds, total distance traversed........................... = 1,600

This comes out exactly the same as before. On reflection, it is evident that this is to be expected. Since the velocity increases at *a uniform rate for each moment of time,* the true average rate of speed for any period will be just half way between the speed at the beginning and at the end.[17] If we consider our 10 seconds as a whole, the velocity at the beginning is equal to

$$V = 32(t) = 32(0) = 0$$

that is, the initial velocity is zero; whereas the velocity at the end is

$$V = 32(t) = 32(10) = 320$$

The average speed for the period, therefore, is

$$\frac{0 + 320}{2} = 160$$

which is exactly the same as the speed at which the body is falling at the middle of the period, at the end of the fifth second, which is

$$V = 32(t) = 32(5) = 160$$

Computing the total distance traversed by multiplying the total time by this average speed, we have

$$d = (160)(10) = 1{,}600$$

giving exactly the same answer as our earlier computation.

The average speed during any period of t seconds is therefore $32t/2$. The total distance traversed in the t seconds can therefore be determined by multiplying the average speed, $32t/2$, by the total number of seconds, t. This gives

$$d = 32\left(\frac{t}{2}\right)t$$

or

$$\checkmark d = 32\frac{t^2}{2}$$

$$= 16t^2$$

So far, we have assumed that we know the acceleration, or rate of increase in velocity per second. Suppose instead we had not known it to begin with. How could we have found it out?

If we had used the symbol g to represent this value, we could have carried out all the previous calculations, except that we should have used "g" where instead we have used "32."

[17] This would not be true of all types of relations. If, for example, velocity increased at a *changing* rate, the smaller the units taken the more accurate would be the result.

Our last formula then would have been

when "g" is unknown.

$$d = g\frac{t^2}{2}$$

or

✓ $$d = \frac{g}{2}t^2$$

If we let $\frac{g}{2} = b$, the equation then would read

or, $$d = bt^2$$

We could readily determine the value for b by observing the distance a given body falls in 1 second, in 2 seconds, in 3 seconds, etc., and then working out the probable value for the constant, just as has been done before.

After we had made measurements of several distances d in the several periods t, we could determine b most readily for the straight-line equation by using T for t^2. Then

✓ $$d = bT$$

(which is the same form as $Y = a + bX$).

Since we may assume $a = 0$, it follows, from equation (10),

$$a = M_y - bM_x$$

that

$$0 = M_y - bM_x$$

Hence

$$bM_x = M_y$$

and

→ ? $$b = \frac{M_y}{M_x} = \frac{\Sigma Y}{\Sigma X}$$

or, in the terms of this particular example,

$$b = \frac{M_d}{M_T}$$

which gives a basis for determining g, the acceleration due to gravity in feet per second, simply by making observations of the time for bodies to fall varying distances.

Substituting an observation of 64 feet in 2 seconds in this equation gives $b = \frac{64}{4} = 16$; hence $g = 32$.

In this case it should be noted that the formula

$$d = \frac{g}{2}t^2$$

is derived on the assumption that the attraction of gravity is a constant, tending to increase velocity at a uniform rate per second, or other unit of time. Only if this assumption is correct can the equation be used. The equation is directly based upon this assumption; the reasoning used in deriving the equation also serves to explain what the constants obtained *really represent.* On the basis of this reasoning the equation determined is not a mere empirical expression of the relation between time falling and distance traversed. Instead, it is a fundamental measurement of *why that distance is what it is,* and relates it in a logical manner to the attraction of the earth.

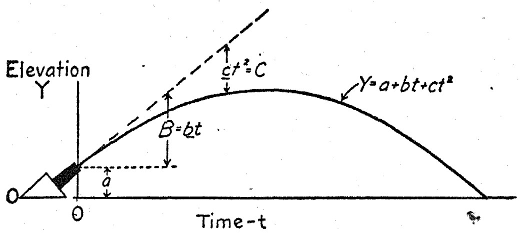


FIG. 20. The trajectory of a projectile, illustrating the equation $Y = a + bX + cX^2$.

Although it would be quite possible in this particular case to draw a freehand curve expressing the relation between time and distance, it would not be so satisfactory as the mathematical equation. The curve would merely state what the relation was; the equation, in addition, explains *why* it is, in the terms of a particular hypothesis.

2. *The parabolic equation.* Another physical case in which a definite relationship may be established logically, and then measured statistically, is the firing of a projectile from a gun.

Disregarding the resistance of the air, there are three elements which will determine the height the projectile will have reached at any given instant after it leaves the muzzle of the gun. The simplest of these elements is the height of the muzzle of the gun itself, represented by a in Figure 20. All the subsequent changes in elevation will obviously have to be added to that.

The second element is the rate at which the projectile is moving upward at the instant it leaves the muzzle. That is dependent, of course, on the angle at which the gun is elevated and the muzzle velocity. If the gun were elevated 1 per cent from the horizontal and the muzzle velocity were 1,000 feet per second, the projectile would leave the muzzle moving upward at the rate of 10 feet per second. If there were no resistance of the air, and if there were no force of gravity to pull the projectile off its course, its momentum would carry it on in this direction to infinity, as illustrated by the straight line in the picture. Here b represents the increase in elevation the projectile would attain for each additional second of flight, and a and bt the elevation it would attain if gravity did not influence it.

But gravity is at work too. As we have already seen, as soon as a body is released, the pull of gravity tends to move it downward at ever-increasing speed. Even if it is headed upward as when shot from a gun, the pull of gravity starts tending to pull it down. The diagram illustrates what happens, with C used to represent the distance the body would have fallen if it had no upward velocity. At first the gain in height from its upward momentum is more than enough to offset the tendency to lose height because of the pull of gravity, and the projectile moves upward along the curved course indicated. But finally the loss due to gravity becomes greater than the gain from its original upward momentum and the trajectory gradually turns downward, until the projectile finally comes to rest in the earth or on its target.

The height that the projectile reaches at any moment is the sum of these three components—the original height, the upward course, and the loss by gravity. Its height, then, can be expressed by adding together the three elements.

a remains the same, regardless of the time elapsed.

B, the height due solely to the original momentum, depends on the time, increasing as the time increases. If we let b represent the initial rate of gain in elevation per second of time, B can then be stated:

$$B = bt$$

Finally, C depends on the time elapsed, and, as we have just seen, varies with the square of time. With the same notation as in our falling-stone problem, but with C substituted for distance fallen

$$C = -\frac{g}{2}t^2 = ct^2$$

Adding these three elements together, we obtain the equation for the height of the projectile at any instant, letting H represent height in feet.

$$H = a + bt + ct^2$$

It will be seen that this equation is exactly identical in form with the equation for a parabola

$$Y = a + bX + cX^2$$

three elements decide the projectile height; hence three constants a, b, c,

Measurements of the height of the projectile at various given times after firing the charge, made for a given gun, firing the same charge at the same elevation of the gun, would give a series of X and Y values which could be used in computing the constants a, b, and c, even if all were unknown to start with.

If the equation were actually worked out, it would tell much more than merely the graph of the relation. For if the reasoning on which the several different constants were included in the equation was correct, then the equation would furnish a real explanation of why the projectile moved as it did, in terms of the laws of motion and of gravity upon which all such movements depend.

in social analysis, the math accuracy is lacking

Reasoning such as this, carried out to much greater lengths, has formed the basis for the scientific "laws" which have been discovered in physics and chemistry and expressed in definite equations. The methods for determining the constants in such equations, as presented earlier in the chapter, were devised to serve in determining such types of relations. But when the same methods are applied to biological, economic, educational, or other relationships in the natural or social sciences, their value is much more limited. Only rarely is there real basis for expecting a particular mathematical relationship such as can be expressed in a given type of equation. In many cases our knowledge of the reasons for the relationship are altogether too limited to enable us to say *why* the relationship is; and even where we can establish the reasons, they are frequently too complicated or too involved—or even too biological—to admit of mathematical treatment. If we express a given relation by a formula, merely on the basis that that formula seems to describe the observed relation satisfactorily, we do not have any greater knowledge of the relation than if we merely drew in a freehand curve. The equation is simply an empirical description of the relation; of and by itself, it offers no clue as to what the relation means.

3. *When to fit a mathematical equation.* From this discussion, the following tentative conclusion may be reached: Only when there is

some good logical basis for expecting a certain type of relation to hold should mathematical curves be employed in describing the relationship. When there is a logical basis for using a given formula, the constants of the equation serve as an explanation of the real nature of the relationship. In all other cases the mathematical curve has no more significance than the freehand curve; the latter may therefore be employed to describe the nature of the relation, and can be determined with much less expenditure of effort. That does not mean that a mathematical curve, based on adequate logical analysis, is of no additional value. If it can be shown that such a curve does fit the data, that may verify an hypothesis and so provide a "law" to state the nature of the relationship, which may be of far more value than the mere empirical statement of what the relationship is observed to be. If, however, there is no logical basis for anything except the empirical statement of the observed relation, the freehand curve is just as valuable as one fitted by aid of a mathematical equation.

v.g.

Where the logical expectations do not lead to a relation which can be formally expressed in a simple equation, they may, as has already been shown, still be sufficient to state a set of limiting conditions to be used in fitting a freehand curve.

4. *A mathematical equation used in an economic problem.* Economists sometimes use the hypothesis that for any one commodity there will tend to be a constant relation between the rate of change in the quantity consumers would buy and the rate of change in price. That is, if an increase of, say, 1 per cent in price would cause a 2 per cent decrease in consumption when prices were low, a similar increase of 1 per cent in price would still cause a decrease of 2 per cent in consumption even when prices were high and consumption was already low.

This economic hypothesis can be stated in definite mathematical terms quite as readily as the various physical hypotheses which have been mentioned; for it makes certain definite assumptions as to the precise way the two variables (price and consumption) are related.

If C is used for quantity consumed and P for price, the statement says that the relation

$$C = f(P)$$

that is, that the quantity consumed depends upon and varies with price, is a function of the type

$$C = kP^b$$

The reason for its being that type can be seen by stating the last equation in logarithmic form:

$$\log C = a + b \log P$$

This says now that a given change in the logarithm of P is always accompanied by a change of b times as much in the logarithm of C. Remembering that the same absolute change in the *logarithm* of a number always means a constant *percentage* change in its actual value, we can see that this equation states the economic hypothesis that a given proportional change in price is always accompanied, on the average, by a constant proportional change in consumption, no matter whether price was high or low to start with.

The practical application of the logarithmic demand equation may be illustrated by a concrete case. Table 27 shows the slaughter of hogs (under federal inspection) in the United States during the years 1922 to 1927 and the average price paid by packers during those years. If we assume that all the meat and other products from these hogs was consumed and ignore any possible shifts in the levels of demand during that period, we may ask whether the relation between the annual

TABLE 27

Slaughter of Hogs, and Average Price, and Computation of Logarithmic Curve

($\log C = a + b \log P$)

Year*	Weight of hogs slaughtered† (C)	Price of hogs‡ (P)	Logarithms of data		Extensions	
			Slaughter	Price	$\bar{C}\bar{P}$	$\bar{P}^2$
	Billion pounds	*Dollars per cwt.*	$\bar{C}$	$\bar{P}$		
1922–23	11.66	7.62	1.0667	0.8820	0.94083	0.77792
1923–24	11.83	7.61	1.0730	0.8814	0.94574	0.77687
1924–25	10.25	10.71	1.0107	1.0298	1.04082	1.06049
1925–26	9.66	12.16	0.9850	1.0849	1.06863	1.17701
1926–27	10.04	10.84	1.0017	1.0350	1.03676	1.07123
1927–28	10.99	9.20	1.0410	0.9638	1.00332	0.92891
Sums....	....	....	6.1781	5.8769	6.03610	5.79243

* From November to October, inclusive.

† Live weight of hogs slaughtered under federal inspection.

‡ Average costs to packers, at live weight. Adjusted for differences in price level, to 1928 level.

average price and the consumption of hog products in the United States during this period agrees with the hypothesis that a given proportional fall in price causes a constant proportional rise in consumption. We may at least roughly hold constant the effect of changes in price level by adjusting the price averages for concurrent changes in the level of wholesale prices.

Accordingly we "fit" the equation

$$\log C = a + b \log P$$

(where C = consumption, and P = price)

to the data by the methods previously discussed. The actual computations are all shown in Table 27.

$$\begin{aligned}
M_{\bar{c}} &= \frac{\Sigma \overline{C}}{n} = \frac{6.1781}{6} = 1.02968 \\
M_{\bar{p}} &= \frac{\Sigma \overline{P}}{n} = \frac{5.8769}{6} = 0.97948 \\
\Sigma(\overline{cp}) &= \Sigma(\overline{CP}) - nM_{\bar{c}}M_{\bar{p}} \\
&= 6.03610 - 6(1.02968)(0.97948) = - 0.01521 \\
\Sigma(\bar{p}^2) &= \Sigma(\overline{P}^2) - nM_{\bar{p}}^2 = 5.79243 - 6(0.97948)^2 = 0.03614 \\
b_{\overline{cp}} &= \frac{\Sigma(\overline{cp})}{\Sigma(\bar{p}^2)} = \frac{-0.01521}{0.03614} = - 0.42086 \\
a_{\overline{cp}} &= M_{\bar{c}} - bM_{\bar{p}} = 1.02968 - (- 0.42086)(0.97948) \\
\overline{C} &= a_{\overline{cp}} + b_{\overline{cp}}\overline{P} \\
&= 1.4419 - 0.42086\overline{P} \\
\log C &= 1.4419 - 0.4209 \log P
\end{aligned}$$

We may next test how well this equation describes the relationship by plotting both the original observations and the curve corresponding to the equation. Figure 21 shows this comparison in terms of the logarithmic values used in the computation and with the logarithmic values of the function (which, of course, is a straight line). It is seen that this straight line seems to fit the original values quite closely; they fall very close to it, above and below, in such a random fashion that no other type of curve seems necessary.

The comparison may also be made in terms of the original values, using the estimated values of the curve transformed from logarithms back to real numbers. Figure 21 shows the comparison of these values. Here again, the demand curve is seen to be a satisfactory "fit" to the actual data.[18]

The economic hypothesis as to the relation between price and consumption would therefore seem to be borne out so far as this particular illustration is concerned, and with the assumptions stated. The size of the constant, b, $-$ 0.42, indicates that anywhere along the curve a 1 per cent increase in the price of hogs is accompanied by approximately 0.4 per cent decrease in hog consumption, or *vice versa*.[19]

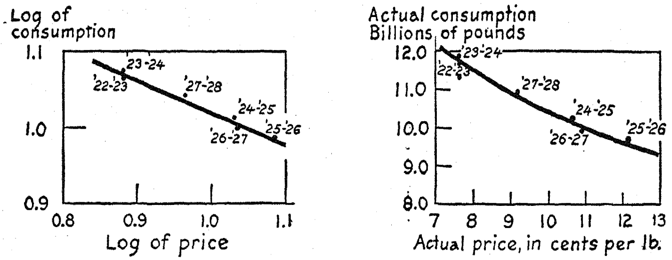


FIG. 21. The relation of consumption of hog products to hog prices, fitted by a logarithmic demand curve, both in logarithms of consumption and price and in natural numbers.

The wheat-protein example, on the other hand, illustrated a case where there was no logical basis for the use of any particular equation and where a freehand curve was therefore as satisfactory as any other type and gave a better fit than any of the analytical types which were tried. As has been stated, the great majority of the problems in the natural and social sciences are probably of this type, where

[18] Six observations, such as used in this case, are far too few to give stable or dependable results in price analysis or any other form of correlation. A curve from a sample of six observations is still less reliable than is an average from a sample of six observations. The close fit of the line to the observations in this case is partly due to the small number of observations utilized. The student can check this by recomputing this example including additional data for a longer period, say through 1937–1938, as given in *Agricultural Statistics*, p. 327, U. S. Department of Agriculture, 1939.

[19] In calculating this simple illustration, no attempt has been made to allow for the effect of changes in other factors which might also influence hog prices, such as the level of consumer buying power, the supplies or prices of other competing meat animals, or the changes in export demand. Chapter 23 discusses actual price analyses involving much more elaborate work than this shown here.

the relation can be measured even though the specific causes for it cannot be stated in mathematical language. Only where the relations can be explained on some logical basis which lends itself to mathematical statement is there justification for a large amount of work to "fit" a specific formula; and even then, if it is found that that particular formula does not give as good a "fit" as a simple freehand curve, there would be question as to whether the hypothesis was in agreement with the facts in that particular case.

Limitations in estimating one variable from known values of another. The methods shown so far provide a definite technique by which an investigator can determine the way in which the values of one variable differ as the values of another related variable differ. These same operations afford a basis for estimating values of the dependent variable from given values of the independent variable, for cases in addition to those from which the functional relation was determined. Whether such estimated values, for cases not included in the original study, can be expected to agree with the true values if they could be determined, depends upon two groups of considerations: (*a*) the descriptive significance of the curve and (*b*) its representative significance when it comes to applying to new observations.

These two groups of considerations apply (*a*) to exactly what a given curve means, with regard solely to the particular cases from which it was determined; and (*b*) the significance of the curve with regard both to the ability of those observations to represent the universe (whole group of facts) from which they were drawn and the ability of the curve to represent the true relations existing in that universe. This second group involves an extension of the points which were raised in the first chapter as to the reliability of an average; discussion of these questions will be deferred to Chapters 18 and 19.

Just as an average computed from a sample may differ more or less widely from the true average of the universe from which that sample was drawn, so a regression line or curve determined from a sample may differ more or less widely from the true regression in the universe. The following chapter discusses this problem, and Chapter 18 presents methods of estimating how far the regression line or curve from an individual sample may miss the true regression of the universe.

The representative significance of a curve depends upon the number of observations from which its shape was determined and how closely the curve as determined "fits" those observations. Since the number of observations usually differs along the different portions of a curve, it may be much more reliable in its central portions, where the bulk of observations occurs, than in the extreme portions where the number of observations may be much less. This may be especially marked in the case of complex curves fitted by mathematical means, where single extreme observations may have a material effect upon the shape of the end portions. In any event, only those portions of the curve where there are enough observations to make its shape and position definite should be regarded as statistically determined; the end portions, when dependent upon a few observations, should either not be used at all or else stated as very rough indications of the true curve.

It is particularly to be noted that determination of the line or curve of relationship gives no basis for estimating beyond the limits of the values of the independent variable actually observed. No matter whether a formula has been fitted or not, any attempt to make estimates beyond the range of the original data by "extrapolation," i.e., by extending the curve beyond the range of the observed data, gives a result that is not based on the statistical evidence. In case a formula has been used which has a good logical basis, extrapolation may give a result which it is logical to expect—but its reasonableness rests on the validity of the logic rather than on a statistical basis. The statistical analysis indicates only what the relations are within the range of the observations which are used in the analysis.

The "closeness" with which the line or curve fits the original data is another criterion of the reliance which can be placed in it. If the data all fall quite close to the line, that fact inspires more confidence in it than if they differ widely and erratically from it. But there are special statistical measures of just what this "closeness" is, and they will be given separate considerations in the next chapter.

As noted earlier, many more cases are required to determine a relation with any degree of dependability than were used in the hog-consumption example just considered. That example was given to illustrate the type of problem where a definite equation might be applied but not as an illustration of a real research problem.

Summary. In some functional relations, the change in the dependent variable with changes in the independent variable cannot be represented by a straight line. Such a relation may be represented by a curve showing the value of the dependent variable for each particular value of the independent variable. Curves may be fitted to given sets of observations either by use of mathematical functions, such as parabolas, logarithmic curves, and hyperbolas, or by various processes of freehand smoothing. When there is some logical basis for the selection of a particular equation, the equation and the corresponding curve may provide a definite logical measurement of the nature of the relationship., When no such logical basis can be developed, a curve fitted by a definite equation yields only an empirical statement of the relationship and may fail to show the true relation. In such cases a curve fitted freehand by graphic methods, and conforming to logical limitations on its shape, may be even more valuable as a description of the facts of the relationship than a definite equation and corresponding curve selected empirically.

In any event, estimates of the probable value of the dependent variable cannot be made with any degree of accuracy for values of the independent variable beyond the limits of the cases observed; and can be made most accurately only within the range where a considerable number of observations is available. It may be possible to extrapolate the curve if its equation is based on a logical analysis of the relation as well as on the cases observed; but in that case the logical analysis, and not the statistical examination, must bear the responsibility for the validity of the procedure.

Note 1, Chapter 6. The methods described in this chapter have been illustrated by determining the curve expressing the average change of percentage of protein with changes in percentage of vitreous kernels. In more general terms, that is, they have been limited to determining the relation

$$Y = f(X)$$

Exactly the same methods can be used to determine the reverse regression, which would show the average change in percentage of vitreous kernels with a given change in percentage of protein. Although this regression is not precisely the reciprocal of the other, it will usually be found that, where a curve rather than a straight line is necessary to represent one regression, a curve will similarly be needed for the other. regression. It will not necessarily be a curve of the same shape, however, or one that can be represented by the same equation

Note 2, Chapter 6. When an equation is used with the dependent variable stated as a logarithm, as types (*b*) and (*c*) on page 93, the further assumption is involved that the errors to be minimized vary proportionately with the size of the dependent variable. The standard error of estimate also must be stated as a percentage of the value estimated, rather than as a natural number. For an example of a problem where the range of error increases with the size of the dependent variable, and where a logarithmic equation would therefore be justified, see Figure 23, on page 154.

CHAPTER 7

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION

The methods developed up to this point may be used to estimate the values of one variable when the values of another are known or given. They also furnish an explicit statement of the average difference or change in the values of the estimated or dependent variable for each particular difference or change in the value of the known or independent variable. But that is not enough. In addition it is frequently desirable to answer three queries: ✓(1) How close can values of the dependent variable be estimated from the values of the independent variable? ✓(2) How *important* is the relation of the dependent variable to the independent variable? ✓(3) How far are the regression curve and these relations, as shown by the particular sample, likely to depart from the true values for the universe from which the sample was drawn? Special statistical devices, termed (1) the *standard error of estimate* and (2) the *coefficient* and *index of correlation*, have been developed to meet the need indicated by the first two questions. Error formulas and knowledge of the distributions of these coefficients, and standard errors for the regression line or curve, provide approximate answers for the third, under the assumption that the conditions of sampling are ideal (an assumption rarely valid even in experimental work).

Three separate queries & their answers.

The Closeness of Estimate—Standard Error of Estimate

Attention has previously been called to the fact that when some dependent variable, such as the distance required for an automobile to stop after the brake is applied or the protein content in wheat samples, is estimated from another variable, such as the speed at which the car is moving or the proportion of vitreous kernels in the sample, the estimated values in many cases will not be the same as the values of the dependent variable that were originally observed. These differences are obviously due to *residual* causes; that is, to variations in the dependent variable which were unrelated to changes in the par-

Residuals.

ticular independent variable used in the analysis. For that reason the differences between the estimated values and the actual values are termed residual differences or, more simply, *residuals.*

For linear relations. The meaning of the residuals and their use in determining the standard error of estimate and the coefficient and index of correlation can best be understood if illustrated by a concrete case. Such an illustration is given in Table 28. Here 18 observations of the number of days (X) that horses worked on different farms and the quantity of grain fed each horse (Y) have been fitted by a straight line to estimate the quantity of feed from the days of work. The estimated quantities, Y', and the residuals, z, or differences between the estimate and the actual, are also shown.

TABLE 28

Days Worked by Horses, Grain Fed per Horse, and Grain Estimated from Days of Work

Days worked X	Grain fed, in hundred weight Y	Estimated grain fed* Y'	Excess of actual over estimate z
107	49	48.0	1.0
70	28	40.9	−12.9
81	44	43.0	1.0
57	36	38.4	− 2.4
87	58	44.2	13.8
114	38	49.4	−11.4
73	49	41.5	7.5
74	53	41.7	11.3
42	33	35.5	− 2.5
90	45	44.8	0.2
100	59	46.7	12.3
59	39	38.8	0.2
86	38	44.0	− 6.0
89	41	44.6	− 3.6
98	42	46.3	− 4.3
95	45	45.7	− 0.7
76	39	42.1	− 3.1
98	46	46.3	− 0.3

* Computed by regression formula $Y = 27.43 + 0.1927X$.

The residuals vary from $+13.8$ to -12.9. If we wish to say how large they are on the average, we can ignore the plus and minus signs and compute the average deviation. For the 18 residuals in Table

28, the average deviation is 5.25, and the standard deviation is 7.13. If these residuals are grouped in a frequency distribution, they fall as shown in Table 29.

The standard deviation of z is different from the standard deviations previously computed. Instead of showing the standard deviation of grain fed from the mean quantity (that is, σ_y), it shows the standard deviation around a changing quantity, depending on the number of days worked. The σ_z is thus the standard deviation around the fitted line of relation, and may be indicated graphically on a correlation chart as a certain area above and below the fitted line. (Note Figure 22, page 151 of Chapter 8.)

TABLE 29

Frequency Distribution of Residuals in Estimating Grain Fed

Residual*	Number of times occurring	Residual*	Number of times occurring
−16 to −12	1	0 to + 4	4
−12 to − 8	1	+ 4 to + 8	1
− 8 to − 4	2	+ 8 to +12	1
− 4 to 0	6	+12 to +16	2

* As stated in Chapter 1, −12 to −16 means from −16 up to, but not including, −12; and so on for the other groups.

The standard deviation is 7.13, so we should expect two-thirds of the residuals to come between +7.13 and −7.13. Of the 18 cases, 12 came within this range of the line, or 67 per cent of all the cases. Similarly, only 5 per cent of the cases would be expected to fall outside the range $\pm 2\sigma$, or below −14.3 or above +14.3. Actually none come outside this range, which is close to the expected proportion for a normal distribution with this limited number of observations.

Where the same set of conditions prevails as those under which the original data were selected and only the independent variable is known, it may be desired to estimate the probable value of the dependent variable from the known value of the independent. Thus if the number of days that horses work on other farms in the same area is known, it may be desired to estimate the quantity of grain that will be needed to feed them. Or in a case where yield of cotton with various applications of irrigation water has been determined (note the example in the next chapter) it may be desired to estimate the most probable yield on other fields, solely from the amount of water applied. In case the estimates were to be made for new observations taken from the same "universe"—for example, on the same soil type, in the same area, and for the same year—as were the previous samples, a knowledge of the standard deviation of the residuals for original samples gives a basis for judging how closely the new estimates are likely to approximate the true, but unknown, yields for the new observations. Similarly in the feeding case it is evident that the errors of estimate will not often be greater than 14.3 hundred weight of grain, and usually will be less than 7.1 hundred weight.

Since the standard deviation of the residuals does thus serve to indicate the closeness with which new estimated values may be expected to approximate the true but unknown values, it has been named the *standard error of estimate.*[1]

The symbol S is used to denote the standard error of estimate. $S_{y \cdot x}$ indicates the standard error for estimates of Y made from a linear relation to X, by the equation $Y = a + bX$. Similarly, $S_{y \cdot f(x)}$ would indicate the standard error for estimates of Y made on the basis of a freehand curve relation to X, as indicated by the equation $Y = f(X)$.

The standard error of estimate is therefore defined by the two equations:

$$\left.\begin{aligned} S^2_{y \cdot x} &= \sigma^2_z = \frac{\Sigma z^2}{n} \\ S^2_{y \cdot f(x)} &= \sigma^2_{z''} = \frac{\Sigma (z'')^2}{n} \end{aligned}\right\} \qquad (20.1)$$

The standard error of estimate in estimating grain fed the horses from number of days worked, by the linear equation, is therefore 7.13 hundred weight.

For curvilinear relations. The calculation of the standard error where a curvilinear function is used to express the relation may also be illustrated by the horse-feeding data. From a freehand curve, fitted by methods already described, estimates of Y from the relation $Y = f\ (X)$ were obtained, as shown in Table 30.

The standard deviation of the new residuals is 6.85. This is then the standard error of estimate for estimates based on the curve.

The standard error of estimate of 6.85 from the curve, compared

[1] Chapter 19 gives more refined measures of the accuracy with which estimates may be made for new observations.

with that of 7.13 from the straight line, indicates that in both cases the amount of feed fed horses in a year can be estimated, for the cases included in the sample, from the number of days they work in a year with a standard error of between 675 and 725 pounds. It appears at this stage that the estimates made on the basis of the curvilinear relation are a little more reliable than those based on the linear relation.

TABLE 30

Days Worked by Horses, Grain Fed per Horse, and Grain Estimated from Days of Work, by Freehand Curve

Days worked X	Grain fed, in hundredweight Y	Estimated grain fed Y''	Excess of actual over estimate z''
107	49	46.5	2.5
70	28	41.4	−13.4
81	44	44.2	− 0.2
57	36	37.4	− 1.4
87	58	45.5	12.5
114	38	46.5	− 8.5
73	49	42.2	6.8
74	53	42.5	10.5
42	33	32.5	0.5
90	45	45.9	− 0.9
100	59	46.5	12.5
59	39	38.1	0.9
86	38	45.2	− 7.2
89	41	45.8	− 4.8
98	42	46.5	− 4.5
95	45	46.4	− 1.4
76	39	43.0	− 4.0
98	46	46.5	− 0.5

The standard error of estimate can also be used to indicate the probable reliability of a series of estimates of the values of the dependent variable for new observations when only the values of the independent variable are known, but only where it is definitely known that the new cases are drawn at random from exactly the same universe—the same set of conditions—as were the observations from which the relation was determined. In case they do not represent exactly the same conditions—as if, for example, they represent a different period of time [2]—then the standard error of estimate has meaning only with respect to the scatter of the residuals around the regression line *for the cases used in determining the relationship.* It measures (when adjusted) what the differences probably would have been in the universe from which the observations came but does not give more than a clue or a possible indication as to what the differences may be when the same relations are applied to data from new or different conditions.

Adjustment of standard error of estimate for the number of observations. The standard deviations of a series of samples drawn from any stable universe will vary from one to another, owing to statistical fluctuations. The same is true for the standard error of estimate computed for a fitted line. The standard deviations, or standard errors of estimate, not only vary but on the average also are somewhat smaller than the result that would be obtained from a large sample from the same universe. Because of this tendency of the standard error of estimate from the sample to understate the standard error in the universe, an adjustment is necessary. An unbiased estimate of the value of the standard error of estimate for the entire universe may be calculated from the standard error of estimate for the sample by the use of the following equations:

$$\overline{S}^2_{y\cdot x} = \frac{n\sigma_z^2}{n-2} = \frac{nS^2_{y\cdot x}}{n-2} \tag{21.1}$$

hence

$$\overline{S}^2_{y\cdot x} = \frac{\Sigma(z^2)}{n-2} = \sigma_z^2\left(\frac{n}{n-2}\right) \tag{21.2}$$

And for curvilinear functions

$$\overline{S}^2_{y\cdot f(x)} = \frac{n\sigma^2_{z''}}{n-m} = \frac{nS^2_{y\cdot f(x)}}{n-m} \tag{22.1}$$

hence

$$\overline{S}^2_{y\cdot f(x)} = \frac{\Sigma\,(z''^2)}{n-m} = \sigma^2_{z''}\frac{n}{n-m} \tag{22.2}$$

In these equations, $\overline{S}_{y\cdot x}$ is used to indicate the estimated standard error of estimate for the universe, just as $\overline{\sigma}$ was used (in Chapter 2) to indicate the estimated standard deviation in the universe from which the sample was drawn.

[2] See Chapter 2, page 15, for the other conditions assumed before error formulas apply exactly.

In equations (21.1) to (22.2), n stands for the number of observations. In equations (22.1) and (22.2), m stands for the number of constants in the regression equation, such as a, b, and c. In the case of a parabola of the second order (type a), m would be 3; for a cubic parabola (type f), it would be 4. Where a freehand curve has been used, it is necessary to estimate how many constants would be needed to represent the curve mathematically. (See pages 76 to 81 for the constants needed to represent various shapes of curves.)

The standard error of estimate in estimating grain fed the horses by the linear equation, after the standard deviation of the residuals is adjusted by equation (21.1), works out to be:

$$\bar{S}^2_{y \cdot x} = \frac{n\sigma_z^2}{n - 2}$$

$$= \frac{18(7.13^2)}{18 - 2} = 57.19$$

$$\bar{S}_{y \cdot x} = 7.56$$

The new value indicates that the errors in estimating grain from days worked, when the estimate is made for new observations drawn at random from the same universe, will run slightly larger than was indicated by the residuals for the cases included in the study, as tabulated in Table 29.

When the standard deviation for the curvilinear function is calculated by equation (22.1), a different result from that before appears. If it is assumed that the regression curve used could have been represented mathematically by an equation with three constants (such as a parabola) then the correction works out to be:

$$\bar{S}^2_{y \cdot f(x)} = \frac{n\sigma^2_{z''}}{n - m}$$

$$= \frac{18(6.85^2)}{18 - 3} = 56.31$$

$$\bar{S}_{y \cdot f(x)} = 7.50$$

The adjusted standard error of estimate for the curvilinear relation, 7.50, is barely smaller than that for the linear equation, 7.56. This indicates that when estimates are made for new observations from the same universe, the straight line is likely to give about as reliable results as is the regression curve. Not unless the adjusted standard error for the curve is materially smaller than for the straight line can the curvilinear regression be expected to improve the accuracy of estimate.[3]

Units of statement for standard error of estimate. The standard error of estimate is necessarily stated in exactly the same kind of units that the original dependent variable is stated in. Where the dependent variable is stated in feet, as in the automobile problem, the standard error of estimate will be in feet; where it is in percentage points, as in the wheat problem, the standard error will be in percentage points; and where it is in logarithms, as in Table 27, the standard error will be in logarithms. Thus in a case like that shown in Table 27, the standard error might be the logarithm 0.038. That means that the logarithm of the estimates is likely to agree with the logarithm of the true values to within ± 0.038, two-thirds of the time. With an estimated logarithm of 1.00, the logarithm of the true value would then be between 0.962 and 1.038, two-thirds of the time. In terms of anti-logarithms, this gives values of 9.16 and 10.91, or between 9.1 per cent above and 8.4 per cent below the value 10. Since a given logarithmic difference always means the same percentage difference, no matter how large or how small the base to which it is applied, when the standard error is thus stated in logarithms it indicates the range within which the estimates may be expected to be reliable, not as absolute quantities such as pounds of grain but as percentages. In terms of absolute differences, the estimate might be expected to be right within 100 pounds, no matter whether the quantity fed was estimated at 1,000 pounds or 4,000 pounds; whereas using logarithms, if the estimate was expected to be right within 100 pounds for an estimate of 4,000 pounds, it would be expected to be right within 25 pounds for an estimate of 1,000 pounds.

The *standard error of estimate* is thus computed from the standard deviation of the residuals for the cases on which the relation is based. It indicates the closeness with which values of the dependent variable may be estimated from values of the independent variable. Its exact interpretation differs with the particular units in which the values of the dependent variable are expressed.

[3] The values of $\bar{S}_{y \cdot x}$ are subject to errors of sampling, just as the values of σ_x are subject to errors of sampling. Accordingly, the values of $\bar{S}_{y \cdot x}$ must be regarded only as estimates of the true values, S_e, which prevail in the universe from which the sample is drawn. Also, it must be remembered that the adjustment, m, for the number of degrees of freedom removed, is only an approximate adjustment in the case of a freehand curve, and that this introduces a further limitation to the accuracy of $\bar{S}_{y \cdot f(x)}$.

The Relative Importance of the Relationship—Correlation

In certain problems it might be found that every bit of variation in one variable could be explained, or accounted for, by associated differences in the value of an accompanying variable. Thus all the variation in the volume of a cube can be explained by the corresponding difference in the length of one side. No other variable is needed to account for the volume of the cube. If we know what the length of the side is, we can compute accurately what the volume will be. All the variation in volume can therefore be said to be explained, or accounted for, by the known relation to the length of the side.

In most problems with which the statistician has to deal, however, all the variation cannot be explained by the relation to another variable, and residual variation is left over. As has just been pointed out, this residual variation can be measured and used as an indication as to the errors in estimate.

It is obvious that if no relation has been found, the independent variable considered does not explain any of the observed variation in the dependent variable, and so none of the variation can be explained as due to, or associated with, the independent variable. If, as in the case of the cube, the estimates all agree exactly with the actual values, there are no residual elements, and the variation is perfectly explained. But between these two extremes lie the cases of partial explanation, where a portion of the variation can be explained by the independent variable considered, and a portion cannot. In the automobile case, part of the variation in stopping distance, but not all, was associated with the speed; in the wheat case, part of the variation in protein content, but not all, could be estimated from variations in the proportion of vitreous kernels; and in the horse-feed case, part of the variation in feed fed, but not all, could be accounted for by variations in number of days worked. In many problems it is of interest to determine what proportion of the variation in the dependent variable can be explained by the particular independent variable considered, according to the relation observed.

Measurement of the relative importance of the relation between two variables calls for a different type of statistical constant than the standard error of estimate. The standard error of estimate simply indicates the size of the residuals without regard to the amount of variation in the dependent variable as first observed. If the standard error of estimate for a cotton-yield problem, for example, were 50 pounds, that would be the standard error no matter whether the yield of cotton in the original cases varied only between 200 and 400 pounds or between 50 and 1,200. If the yields varied only between 200 and 400 pounds, and the standard error was 50, practically all the variation in the original yields would still be left in the residuals; whereas if the yields varied between 200 and 1,200 and the standard error was 50, only a very small portion of the original variation would be left in the residuals. Yet the standard error of estimate would be of the same size in both cases.

What is needed to show the relative importance of the relationship is some measure which shows what *proportion* of the original variation has been accounted for. The amount of the variation in the series of estimated (Y') values shows *how much* variation has been accounted for. All that need be done is to compare that variation with the variation in the original series to determine what proportion of the variation has been explained.

The standard of deviation may be employed for the purpose of measuring the amount of variation. The actual values, Y, shown in Table 28, have a standard deviation of 7.92. The values estimated from the linear regression equation, Y', have a smaller standard deviation, 3.47. If we determine how large the latter is compared to the former, we get $\sigma_{y'}/\sigma_y = 3.47/7.92$, or 0.44. This is then a measure of the importance of relationship between the two variables—or the amount of *correlation*, as it is termed—according to the particular type of curve for which the relationship was determined.

Linear relations—coefficient of correlaion. Where the relationship between the two variables is found or assumed to be a straight line, the value of $\sigma_{y'}/\sigma_y$ is termed the *coefficient of correlation.* The symbol r is used to represent it. When values of Y are estimated from values of X according to a straight-line equation, then the proportion of the variation in Y which is so accounted for is indicated by the notation r_{yx}, which is read "the coefficient of correlation between Y and X."

The coefficient of correlation may therefore be defined

$$r_{yx} = \frac{\sigma_{y'}}{\sigma_y} \tag{23.1}$$

This formula gives values of r identical with those given by the more usual formula, equation (27), presented subsequently on page 148, as can be proved by simple algebra (see Note 3*a*, Appendix 2).

The method of computing the coefficient of correlation which has just been shown demonstrates that the coefficient is simply a measure of how large the variation in the estimated values is, in proportion to the variation in the original values. The coefficient of correlation thus measures the *proportion* of the variation in one variable which is associated with another variable, and therefore is a measure of the relative importance of the concomitance of variation in the two factors.

Curvilinear relations—index of correlation. In case the relation has been determined as a curvilinear function instead of a straight line, the ratio $\sigma_{y''}/\sigma_y$ is termed the *index* of *correlation,* and is represented by the symbol ρ_{yx}.

The index of correlation may therefore be approximately defined as

$$\rho_{yx} = \frac{\sigma_{y''}}{\sigma_y} \tag{23.2}$$

(A more exact value for the index of correlation is given in equation (29) on page 156.)

Computing the index of correlation for the horse-feed case, $\sigma_{y''}/\sigma_y = 3.86/7.92 = 0.49$. From this figure, it would appear that the correlation is definitely higher for the curve than for the straight line.[4]

Characteristics of the measures of correlation. It should be noted that in the case of straight-line relations, if the line has a positive slope, so that as X increases the values of Y' (the estimated values of Y) increase, the correlation is said to be *positive,* and a plus sign is affixed to the correlation coefficient. Similarly, if the line has a negative slope, so that as the values of X (the independent variable) are larger, the values of Y' (the estimated values for the dependent variable) become smaller, the correlation is said to be negative, and a minus sign is affixed to the correlation coefficient. The coefficient of correlation thus takes the same sign as the constant b of the corresponding linear equation. In the case of the correlation index, the curve may be positive in one portion and negative in another, so no sign is used, and reference to the curve is necessary to indicate the nature of the relationship.

In a case where the observed relation explains *all* the variation in the dependent variable, the estimated values will be identical with the actual values. The standard deviation of Y' will therefore be exactly as large as the standard deviation of Y, and the ratio $\sigma_{y'}/\sigma_y$ will equal 1.0. This is termed *perfect correlation,* and is indicated when $\rho = 1.0$, or when $r = +1.0$ or -1.0.

[4] In some statistical texts, r_{yx} is used to represent the correlation observed in a given sample, and ρ_{yx} is used to represent the true correlation existing in the universe from which that sample was drawn. The student should not confuse that use of the Greek rho, ρ, with the way it is used here.

At the other extreme of no relation, no variation can be accounted for by the particular independent variable considered, and the estimated values Y' are therefore all the same, being merely the average of Y. In that case the standard deviation of the estimated values is zero, and the ratio $\sigma_{y'}/\sigma_y = 0/\sigma_y = 0$. The case of complete absence of correlation, therefore, is indicated by values of 0 for either r or ρ.

The possible values of the coefficient of correlation therefore range from 0 to + 1.0 or to − 1.0; whereas the values for the index of correlation range from 0 to 1.0. Since most problems with which the investigator has to deal involve cases that are intermediate, where there is some but not perfect correlation, it is these intermediate cases which are of most importance. The precise significance of different values of r and ρ will next be considered.

Where both X and Y are assumed to be built up of simple elements of equal variability, all of which are present in Y but some of which are lacking in X, it can be proved mathematically that r^2 measures that proportion of all the elements in Y which are also present in X. For that reason in cases where the dependent variable is known to be causally related to the independent variable, r^2 may be called the *coefficient of determination.* It may be said to measure the percentage to which the variance in Y is determined by X, since it measures that proportion of all the elements of variance in Y which are also present in X.[5] The coefficient of determination, d_{xy}, may be defined by the equation

$$d_{xy} = r_{xy}^2 \tag{24.1}$$

Where some elements are present in each variable which occur in the other, the coefficient of determination is the product of these joint proportions. That is, if 2/3 of the elements in X are the same as 2/3 of the elements in Y, then the coefficient of determination will be equal to 4/9.

Although the coefficient of correlation was the earliest measure used, it can be seen that it may be misinterpreted. Thus if half the variance in Y is directly due to X, the coefficient of correlation would be 0.707 $(=\sqrt{1/2})$. Yet the coefficient of alienation[6] is also 0.707. If instead the coefficient of determination is used, when we know that that is 0.50, we know at once that the *coefficient of non-determination*[6] is also

[5] See Note 4, Appendix 2.

[6] See Note 5, Appendix 2, for a fuller definition of these new terms.

0.50; or if the determination is 0.60, the non-determination is 0.40. The coefficient of non-determination may be defined.

$$\cancel{d}_{xy} = 1 - r^2_{xy} \tag{24.2}$$

Since this is the most direct and unequivocal way of stating the proportion of the variance in the dependent factor which is associated with the independent factor, it may be used in preference to the other methods.

Where curvilinear relations have been used in determining the relationship, the term *index of determination* will be used to denote the value of ρ^2, thus retaining the same relation to the index of correlation that the coefficient of determination bears to r, the coefficient of correlation. The index of determination, $d_{y \cdot f(x)}$ may be defined

$$d_{y \cdot f(x)} = \rho^2_{yx} \tag{24.3}$$

When an expression is used such as "Forty per cent of the variance in yield is due to differences in rainfall," it will be understood that it is either the coefficient or the index of determination which is being stated.

Relation of the measures of correlation to the two regression lines. Attention has been called in several previous chapters to the fact that two regression lines can be fitted to any set of observations. These are denoted by the two coefficients b_{yx} and b_{xy} in the two equations

$$Y = a_{yx} + b_{yx} X$$

and

$$X = a_{xy} + b_{xy} Y$$

Although there are these two regression lines, there is only a single coefficient of correlation for any one set of observations. In fact, the coefficient of correlation has certain definite relations to the two lines. It indicates how closely the two lines approach one another. The higher the correlation, the closer the two lines come together; the lower the correlation, the farther they diverge. In perfect correlation ($r = \pm 1$) the two lines coincide. When there is no correlation ($r = 0$) the two lines will be at right angles to one another.

This relationship is so exact that the value of the correlation coefficient can be computed from the slopes of the two lines according to the equation

$$r_{yx} = \sqrt{b_{yx}\, b_{xy}} \tag{24.4}$$

It follows from this equation that when $r = 1$, $b_{yx} = \frac{1}{b_{xy}}$, and therefore the two regression lines will coincide.[7]

Although there can be only a single coefficient of correlation for a single set of observations, there can be two *indexes* of correlation. This follows from the fact that the curve which expresses the relation

$$Y = f(X)$$

may be a curve of quite a different type from that which expresses the relation

$$X = \phi(Y)$$

Accordingly, the index of correlation, ρ_{yx}, which measures the closeness of correlation according to the first curve, may be quite different from the index of correlation, ρ_{xy}, which measures the closeness according to the second curve. Only in the special case where all the observations lie precisely along the curve, so that $\rho = 1$, will the two indexes have the same value. In that case it will also hold true that the curves $Y = f(X)$ and $X = \phi(Y)$ will be identical with the coordinates reversed.

There is only one correlation coefficient, r, however. It measures the correlation according to both regression lines. Since $r = r_{yx} = r_{xy}$, either notation can be used interchangeably.

Adjustments for number of observations. Where the number of cases in the sample is not very large, both the coefficient and index of correlation require certain adjustments before the values calculated from the sample, as given by equations (23.1) and (23.2), can be used to indicate the values which are most probably true for the universe from which that sample was drawn. Without correction,

[7] This property of the two lines can be used to estimate graphically the closeness of correlation. When the two variables, X and Y, are stated in terms of unit standard deviation, X/σ_x and Y/σ_y, by dividing each observation by the standard deviation of the series, the coefficient of correlation will then be a precise mathematical function of the angle between the two lines. By stating the variables in this way, plotting them on a dot chart, and drawing in the two lines graphically, a fairly close approximation to the coefficient can be obtained.

the observed coefficient or index of correlation tends to exceed the true correlation.[8]

Denoting the adjusted constants as $\bar{r}_{yx}$ and $\bar{\rho}_{yx}$, the adjustment formulas are:

$$\bar{r}^2_{yx} = 1 - (1 - r^2_{yx}) \left(\frac{n-1}{n-2}\right) \tag{25}$$

$$\bar{\rho}^2_{yx} = 1 - (1 - \rho^2_{yx}) \left(\frac{n-1}{n-m}\right) \tag{26}$$

If the value to the right of the first "1 −" in equation (25) or (26) exceeds unity, 0 must be taken for the value $\bar{r}$ or $\bar{\rho}$.

In these equations, n and m have the same meaning as in equations (22.1) and (22.2), presented on page 133. The adjusted value $\bar{r}$ is the value which most probably exists in the universe, if the correlation is 0.80 or better. In half the samples, the value $\bar{r}$ will be as large as the true value; and in half, it will be smaller than the true value. If, however, the correlation is low, 0.60 or less, $\bar{r}$ is a somewhat more conservative estimate of the true correlation.

Applying the correction to the value of r_{yx} previously computed for the horse problem, the correlation of grain fed with number of days worked is found to be:

$$\bar{r}^2_{yx} = 1 - \frac{[1 - (0.44)^2]\,(18 - 1)}{18 - 2} = 0.1432$$

$$\bar{r}_{yx} = 0.38$$

The index of correlation is even more likely to be spuriously high when based on a small number of cases than is the coefficient of corre-

[8] The value of r calculated from a sample is derived from the standard deviation of the estimated values $\sigma_{y'}$ and the standard deviation of the dependent variable σ_y. It was noted in Chapter 2 that when standard deviations are computed from a small sample, they tend to be less than the true standard deviation of the universe, and this applies to σ_y. At the same time, $\sigma_{y'}$ is determined from a limited number of observations. It was already pointed out that a straight line would exactly fit any two observations with no residuals at all. When a straight line is fitted to ten observations, there are only eight "degrees of freedom" in determining the values a and b, as the "freedom" of two of these observations is used up in the determination. As a consequence of these conditions, the $\sigma_{y'}$ tends to be larger than it should be, and σ_y tends to be too small. Hence the quotient, $\sigma_{y'}/\sigma_y$ tends to be too large, on the average. Also, since $\sigma_{y'}$ tends to be too large, σ_z tends to be too small, and hence the observed standard error of estimate also needs correction, as provided in equations (21.1) to (22.2).

lation and is even more in need of the adjustment, indicated by equation (26).[9]

Computing the index of correlation for the horse-feed problem, with the corrections shown in equation (26):

$$\bar{\rho}^2_{yx} = 1 - \frac{(18 - 1)\left(1 - \dfrac{3.86^2}{7.92^2}\right)}{18 - 3} = 0.1389$$

$$\bar{\rho}_{yx} = 0.37$$

After adjusting, we find that in this case the index of correlation is almost the same as the coefficient, agreeing with the conclusion shown by the two standard errors of estimate. Just as with the standard errors, so it is with the correlation—not unless the index of correlation is still definitely higher than the coefficient, after they have been adjusted by formulas (25) and (26), can it be said that there is definite indication of curvilinear correlation rather than of linear.[10]

It should be noted that in any case the adjustment to r or ρ is small compared with its own standard error—that is, the value given by the sample may miss the true value in the universe by a margin much larger than the difference between the observed value and the adjusted value. Chapter 18 discusses methods of estimating the probable range of such departures of the observed correlation from the true. Even so, the average value from a series of samples always tends to have the bias mentioned, and it is worth eliminating this average bias as far as possible, even if the adjusted value from an individual sample is still subject to a considerable standard error of its own.

The reliability of the regression line or curve and of the measures of correlation. Chapter 2 shows how a series of samples drawn from the same universe would yield varying estimates of the true average in that universe. It also presented methods of estimating how far the

[9] The adjusted index of correlation $\bar{\rho}$ has the same interpretation as the adjusted coefficient of correlation—half of the samples will give values of $\bar{\rho}$ which will not exceed the true value of ρ in the universe from which the sample was drawn.

Just as the a and b of the linear equation eliminate two degrees of freedom, a curve representing three constants (or more) can be passed exactly through three observations (or more) and so may eliminate three (or more) degrees of freedom. There is therefore even more tendency for ρ to be spuriously high than for r, and the correction is even more needed.

[10] See Figure F of Appendix 3 for a graphic method of computing adjusted coefficients or indexes of correlation from the unadjusted values.

average from a single sample might miss the true average in the universe. In exactly the same way, if regression lines or curves are determined for a series of samples from the same universe, they will yield regressions which vary among themselves. Similarly, the coefficients or indexes of correlation and the standard errors of estimate will vary from sample to sample. Standard errors of each of these measures are available. They provide estimates of the range from the true values in the universe within which two-thirds of the values from such samples will fall and of the wider range within which larger proportions of the samples will fall. These measures of reliability for the sample results are much more complicated, both in computation and in interpretation, than the standard error of an average. Accordingly, their presentation is deferred to a later chapter (Chapter 18). In addition, the special problem of the reliability of an individual estimate for an individual new observation, from the results shown by a sample, is treated in a separate chapter (Chapter 19). The methods given in the present chapter and Chapter 8 are sufficient for determining the correlation and regression *as shown in the individual sample.* Before a student or research worker uses the results of the sample to draw more general conclusions as to the relations which hold true in other samples or in the universe as a whole, or before he makes estimates for new observations, he should master these later chapters and should apply the checks and limitations set forth there in stating his general conclusions or in making his estimates.

Summary. This chapter has pointed out that the closeness of relation between two variables may be measured either, by the absolute closeness with which values of one may be estimated from known values of the other or on the basis of the proportion of the variation in one which can be explained by, or estimated from, the accompanying values of the other. The absolute accuracy of estimate is measured by the standard error of estimate, which indicates the reliability of values of the dependent variable estimated from observed values of the independent value.

The relative closeness of the relation is best measured by the coefficient of determination, in the case of linear relationship, or by the index of determination, in the case of curvilinear relationship. These measures show the proportion of the variance in the dependent variable which is associated with differences in the other variable. In the case of variables causally related, they measure the proportion of the variance in one which can be said to be *due to* the other.

The best methods of computing the various measures of correlation will be shown in the next chapter; the methods used in this chapter are designed rather to show the significance of the measures themselves.

This chapter has also called attention to the fact that the measures of correlation obtained from a sample will vary from the true facts of the universe, has referred to later chapters where standard errors for estimating such variation are discussed, and has warned against drawing general conclusions or making new estimates from a single sample unless the precautions described in these subsequent chapters are observed.

CHAPTER 8

PRACTICAL METHODS FOR WORKING TWO-VARIABLE CORRELATION PROBLEMS

Terms to be used. The preceding discussion has developed the means by which values of one variable may be estimated from the values of another, according to the functional relation shown in a set of paired observations. Simple correlation involves only the means for making such estimates, and for measuring how closely those estimates conform to, and account for, the original variation in the variable which is being estimated, for the given set of observations.

The *regression line* is used, in statistical terminology, to designate the straight line used to estimate one variable from another by means of the equation

$$Y = a + bX$$

This equation is termed the *linear regression* equation; and the coefficient b, which shows how many units (or fractional parts) Y changes for each unit change in X, is termed the *coefficient* of *regression.*

Where a curvilinear function has been determined, either by the use of an equation or by graphic methods, the corresponding curve is similarly designated as the *regression curve.* Either the mathematical equation or, if none has been computed, the expression

$$Y = f(X)$$

where the symbol $f(X)$ stands for the relation shown by the graphic curve, is termed the *regression equation.*

The coefficient of correlation and the index of correlation have both been defined as the ratio of the standard deviation of the estimated values of Y to the standard deviation of the actual values, whereas the standard error of estimate has been defined as the standard deviation of the residuals from the estimates so made. In the case of linear relations, however, the coefficient of correlation and the standard error of estimate can both be computed directly from the same values as were employed in computing the constants of the regression equation. This will be illustrated by the practical example which follows.

Working out a linear correlation. As was illustrated in Chapter 5, pages 64 to 71, the values for a and b of the regression equation can be determined for any two variables, X and Y, between which it may be desired to determine the relation, by working out the values, M_x, M_y, ΣX^2 and $\Sigma(XY)$, and then substituting them in the appropriate equations. In order to compute directly the coefficient of correlation, r_{xy}, and the standard error of estimate, S_{yx}, it is necessary only to compute in addition the value ΣY^2 and substitute it in appropriate formulas. The data given in Table 31 illustrate the necessary operations.

TABLE 31

COMPUTING THE VALUES NEEDED TO DETERMINE LINEAR REGRESSION AND CORRELATION COEFFICIENTS

	Irrigation water applied per acre * (X)	Yield of Pima cotton per acre * (Y)	X^2	XY	Y^2
	Feet	*Units of ten pounds*			
	1.8	26	3.24	46.8	676
	1.9	37	3.61	70.3	1,369
	2.5	45	6.25	112.5	2,025
	1.4	16	1.96	22.4	256
	1.3	9	1.69	11.7	81
	2.1	44	4.41	92.4	1,936
	2.3	38	5.29	87.4	1,444
	1.5	28	2.25	42.0	784
	1.5	23	2.25	34.5	529
	1.2	18	1.44	21.6	324
	1.3	22	1.69	28.6	484
	1.8	18	3.24	32.4	324
	3.5	40	12.25	140.0	1,600
	3.5	65	12.25	227.5	4,225
Total.	27.6	429	61.82	970.1	16,057
Mean.	1.97	30.64			

* From James C. Muir and G. E. P. Smith, The use and duty of water in the Salt River Valley, *Agricultural Experiment Station Bulletin* 120, University of Arizona, 1927. All the plots were on the same type of soil, Maricopa sandy loam.

The computations shown in this table—squaring both X and Y, calculating the product XY, summing both X, Y, and the three columns of extensions, and dividing the first two sums by the number of cases to give the mean of X and Y—provide all the basic data necessary.[1] The values a and b for the regression equation may next be computed by substituting these extensions in equations (9) and (10), which were used previously in Chapter 5, page 66.

$$b_{yx} = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - n(M_x)^2} = \frac{970.1 - 14(1.97)(30.64)}{61.82 - 14(1.97^2)}$$

$$= \frac{125.050}{7.4874} = 16.701$$

$$a = M_y - bM_x = 30.64 - 16.701(1.97) = - 2.261$$

The *regression line*, $Y = a + bX$, therefore is for this case

$$Y = - 2.261 + 16.701X$$

The unadjusted coefficient of correlation, r_{xy}, may now be computed from the following new formula:

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{\sqrt{[\Sigma(X^2) - nM_x^2]\,[\Sigma(Y^2) - nM_y^2]}} \qquad (27)$$

$$= \frac{970.1 - 14(1.97)(30.64)}{\sqrt{[61.82 - 14(1.97)^2]\,[16{,}057 - 14(30.64)^2]}} = 0.847$$

It should be noticed that the numerator of this fraction is the same as that in the equation for b and that half of the denominator is the same, except that it is under the radical sign.

Comparison of equations (9) and (27) with equation (5) for the standard deviation

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{n} - M_x^2}$$

shows that they may be written more simply

$$b_{yx} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x^2} \quad \text{or} \quad = \frac{\Sigma(xy)}{n\sigma_x^2} \qquad (27.1)$$

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x\sigma_y} = \frac{\Sigma(xy)}{n\sigma_x\sigma_y} \qquad (27.2)$$

[1] Where the number of cases to be handled is large, various short cuts may be used to reduce the volume of computation required in computing the sums of extensions ΣX^2, ΣXY, and ΣY^2. The use of these short cuts is developed in Appendix 1, pages 455 to 463.

The second form, in each case, uses the notation $\Sigma(xy)$ for $\Sigma(XY) - n(M_xM_y)$ as discussed on page 66.[2] The forms shown in equations (9), (10), and (27), however, are the ones ordinarily used in actual computation, and should be kept clearly in mind.

Once r_{xy} has been computed, the value adjusted for the number of cases can then be obtained by equation (25).

$$\bar{r}^2_{xy} = 1 - (1 - r^2_{xy})\left(\frac{n-1}{n-2}\right)$$

For the present problem, that becomes

$$\bar{r}^2_{xy} = 1 - \frac{[1 - (0.847)^2]\,(14 - 1)}{14 - 2} = 0.6939$$

$$\bar{r}_{xy} = 0.833$$

Knowing $\bar{r}_{xy}$, we may next compute the standard error of estimate by the following equation:

$$\bar{S}_{yx} = \sqrt{\frac{\Sigma(Y^2) - n(M_y)^2}{n-1}(1 - \bar{r}^2_{xy})} \tag{28}$$

$$= \sqrt{\frac{16{,}057 - 14(30.64^2)}{13}[1 - (0.833)^2]}$$

$$= \sqrt{68.62} = 8.28$$

Since this equation includes $\bar{r}_{xy}$, already adjusted for the number of observations, no further adjustment is necessary. The standard error computed by equation (28) is identical with that obtained by equation (21.1), or (21.2), given in the previous chapter.

As noted earlier, though $r_{xy} = r_{yx}$, b_{xy} is *not* the same as b_{yx}. The former regression, showing the change in X for each unit change in Y (that is, regarding the dependent factor as the independent factor instead), is obtained by modifying equation (9) to the following form:[3]

$$b_{xy} = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(Y^2) - n(M_y)^2}$$

[2] The value of $\Sigma(xy)$ is sometimes called the *product moment*.

[3] When the correlation is perfect, so that $r_{xy} = 1$, the two regression coefficients will have the definite relation $b_{yx} = 1/b_{xy}$. Under these conditions the regression lines will be identical, no matter which variable is regarded as the independent variable and which as the dependent.

The new regression coefficient, b_{xy}, shows the average change in water applied with each additional unit (ten pounds) of cotton harvested. With the quantity of water subject to human control, as in this case, this relation appears to have little meaning. However, if it is desired to chart it on Figure 22 along with the other regression line, it can be charted according to the linear regression equation

$$X = a_{xy} + b_{xy}Y$$

The value of the new a can be computed by restating equation (10) in the form

$$a_{xy} = M_x - b_{xy}M_y$$

Equation (28) completes the computation of all the values needed[4] except the coefficient of determination, d_{xy}, which is simply r^2_{xy}. That is:

$$\bar{d}_{xy} = \bar{r}^2_{xy} = (0.833)^2 = 0.694$$

Interpreting the results of a linear correlation. The next step is to take the several constants which have been computed and see what they mean.

The coefficient of regression of *Y on X*, $b_{yx} = 16.70$, shows that on the average the acre yield of cotton increases 16.7 ten-pound units, or 167 pounds, for each additional acre-foot of water applied. The constant a shows that with no water applied, a yield of $-$ 2.26 ten-pound units, $-$ 22.6 pounds, or less than no cotton at all, might be expected. Since these results are based on observations extending from 1.2 acre-feet of water to 3.5, the relations shown by the regression line do not necessarily hold beyond those limits, and it is not certain what the yield would be when no water is applied. Extrapolating the regression line to that point is only a guess.

The regression equation

$$Y = -\ 2.26 + 16.7(X)$$

or

$$\text{Yield} = -\ 22.6 + 167\ (\text{feet of water})$$

then gives the yields of cotton estimated as most likely to be obtained from the quantity of water applied within the limits of 1.2 to 3.5 feet. Figure 22 shows how these estimated values, along the regression line, compare with the actual yields observed.

[4] Except also the calculation of measures of reliability, as explained in Chapters 18 and 19.

The standard error of estimate, 8.28 ten-pound units or 82.8 pounds, shows that the (adjusted) standard deviation of the differences between the actual and the estimated values is 82.8 pounds of cotton. Two lines have been drawn in Figure 22, at 82.8 pounds above and below the regression line. It will be seen that of the 14 cases, 9 fell between these two lines, or in the zone within one standard error on either side of the regression line.

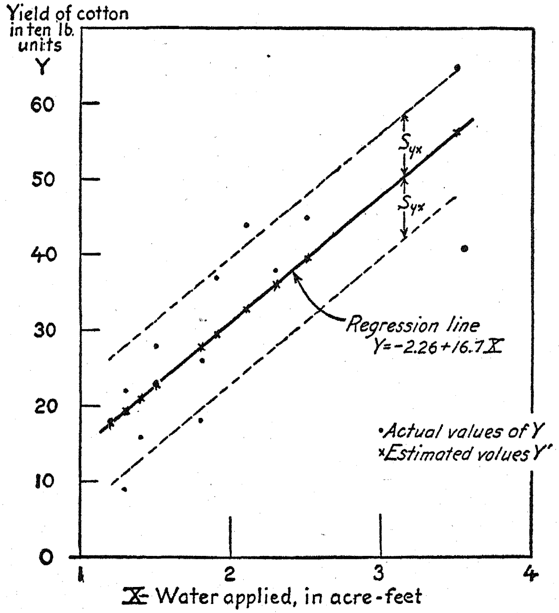


FIG. 22. Relation of yield of cotton to irrigation water applied; estimated yields from a linear regression and zone of probable yields indicated by the standard error of estimate.

The coefficient of correlation, $\bar{r}_{xy} = 0.83$, and the coefficient of determination, $d_{xy} = 0.69$, show that about 69 per cent of the variance in the yield of this crop in this area, on the farms from which these records were obtained, could be accounted for by the differences in the quantity of water used in irrigation. Since this leaves only 31 per cent of the variance to be accounted for by all other factors, it would appear that the quantity of water applied (or other factors associated with it) was the most important factor which was associated with the yield of cotton on these farms and on this type of soil.

The fact that 69 per cent of the variance in yield can be explained by corresponding differences in the quantity of water applied does not in itself mean that the differences in irrigation *caused* the differences in yield. For example, it might be possible that the quantity of water applied was regulated to conform to the fertility of the land and that the differences in yield were really due to the differences in fertility. The statistical measure merely tells how closely the variance in one variable was associated with variance in the other; whether that association is due to, or can be taken as evidence of, cause-and-effect relation is another matter, and is outside the scope of the statistical analysis. (For more extended discussion of this point, see the last two chapters of this book.)

Working out a curvilinear correlation. The next step is to consider whether the straight line is adequate to describe the way that the yield increases as more water is applied, or whether a curve had better be employed. (This step can be taken before any of the linear results are worked out, and, if a curve is decided on, the previous work can be skipped entirely, if desired.)

Before fitting the curve, we must consider what type of curve it is logical to expect. In most agricultural production problems, diminishing returns are experienced.[5] That is, the application of successive increments of fertilizer or other productive aid on the same areas will be expected to produce a smaller and smaller increase in the product. Also, it is known that if too much of some factors are applied, the result may be to produce a decline in output. The decline after the point of optimum application is reached may be gradual, or it may be sudden, owing to a toxic effect of too much of one substance upon the plant or animal. These considerations would lead us to expect a curve with the following characteristics:

1. It should rise steeply at first, and then less and less sharply until a maximum is reached.
2. It might show a decline after the maximum is reached, either gradual or sharp.
3. It would have only the single point of inflection (change of direction) at the optimum application.

These are the conditions we shall apply in fitting the curve.

Examining Figure 22 more closely, we see that, in the range up to 1.8 acre-feet of water, the actual yields lie below the regression line four times, and above four times; in the range from 1.9 to 3 acre-

[5] William J. Spillman, *The Law of Diminishing Returns,* World Book Co., Yonkers-on-the-Hudson, New York, and Chicago, 1924.

feet, the actual yields lie above in all four observations; and above 3 acre-feet the one yield below the line is much farther below than is the one above. These facts suggest that a curve convex from above, giving lower estimated yields than the straight line for the lowest and highest applications of water and higher estimated yields for the intermediate applications, would more accurately represent the relations in this case. (The number of observations is far too low to serve as a very accurate indication of the shape of the curve, but it will serve at least as a simple illustration of the way the whole problem may be worked through.)

The next step is to group the observations according to the value of *X* (the quantity of water) and average both *X* and *Y*, water and yield. In view of this small number of observations, rather large groups are taken; were more cases available, the groups might be made narrower.

TABLE 32

Computation of Group Averages to Indicate Regression Curve—Cotton Example

X (water) 1 to 1.4		*X* (water) 1.5 to 1.9		*X* (water) 2.0 to 2.9		*X* (water) 3.0 to 3.9	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
1.4	16	1.8	26	2.5	45	3.5	40
1.3	9	1.9	37	2.1	44	3.5	65
1.2	18	1.5	28	2.3	38		
1.3	22	1.5	23				
		1.8	18				
Sums... 5.2	65	8.5	132	6.9	127	7.0	105
Means.. 1.3	16.25	1.7	26.4	2.3	42.33	3.5	52.5

These averages are then plotted, as shown in Figure 23, an irregular line dotted in connecting them and as smooth a curve as possible which fulfills the stated conditions drawn in freehand through the averages and the broken line, just as discussed in pages 105 to 110, Chapter 6. This then gives the regression curve. It is seen to fit the data well, and yet to fulfill the logical conditions stated. The point of maximum yield, however, apparently lies beyond the limit of the observations.

Next the estimated yields for each different application of water are read off from this curve, and the difference between the actual and the estimated yields is determined. These residuals are then squared to determine their standard deviation. In case the linear correlation has not been previously worked, the yields, or Y values, are also squared as shown, so as to determine their standard deviation, and so give the basis for measuring the amount of correlation.

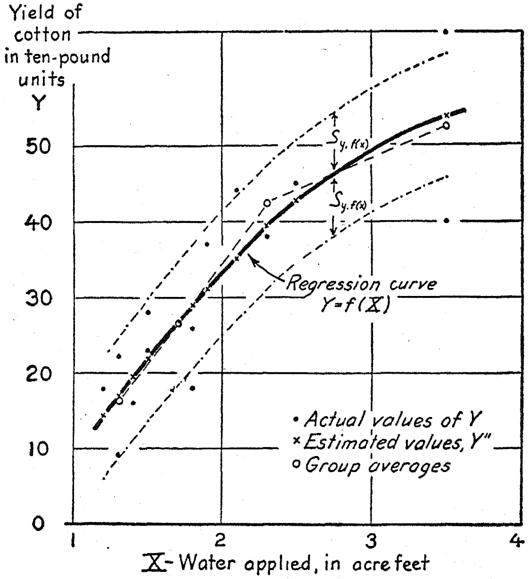


FIG. 23. Relation of yield of cotton to irrigation water applied; estimated yields from a curvilinear regression; and zone of probable yields as indicated by the standard error of estimate.

The sum of the Y'' values is slightly smaller than the sum of the Y values, and the mean of the z'' values is therefore not exactly zero, but 0.264. That indicates that the curve shown in Figure 23 should be shifted up 0.264 unit, or 2.64 pounds, to make the estimated and actual averages agree.[6] Representing this curve by $f(X)$,

[6] In problems with many observations, the sum of the Y values and of the Y'' values may be determined separately for the several different portions of the curve, to see if its position should be shifted in one portion and not in another. This process cannot be carried too far, however, for if the divisions are made too small the effect will be to make the curve pass through each successive group average, without smoothing out the irregularities into a continuous function.

the regression equation for the curvilinear correlation may therefore be written:

$$Y = k + f(X)$$
$$Y = 2.64 + f(X)$$

TABLE 33

Computation of Residuals and Standard Deviation for Curvilinear Regression—Cotton Example

Water per acre, X	Yield, in ten-pound units, Y	Yield estimated from X, in ten-pound units, Y''	$Y-Y''$, (z'')	$(z'')^2$	Y^2
1.8	26	29.0	− 3.0	9.00	676
1.9	37	31.0	6.0	36.00	1,369
2.5	45	42.8	2.2	4.84	2,025
1.4	16	19.2	− 3.2	10.24	256
1.3	9	16.8	− 7.8	60.84	81
2.1	44	35.2	8.8	77.44	1,936
2.3	38	39.5	− 1.5	2.25	1,444
1.5	28	21.9	6.1	37.21	784
1.5	23	21.9	1.1	1.21	529
1.2	18	14.2	3.8	14.44	324
1.3	22	16.8	5.2	27.04	484
1.8	18	29.0	−11.0	121.00	324
3.5	40	54.0	−14.0	196.00	1,600
3.5	65	54.0	11.0	121.00	4,225
Sums......	429	425.3	+ 3.7	718.51	16,057

The values at the foot of Table 33 now give the constants necessary to measure the closeness of the correlation. First the standard deviations of Y and of z'' are computed, using the formula

$$\sigma_y = \sqrt{\frac{\Sigma Y^2 - n(M_y^2)}{n}} = 14.44$$

$$\sigma_{z''} = \sqrt{\frac{\Sigma(z'')^2 - n(M_{z''}^2)}{n}} = \sqrt{\frac{718.51 - 14(0.264^2)}{14}} = 7.16$$

Then, by equation (22.2),

$$\overline{S}^2_{y\cdot f(x)} = \sigma^2_{z''}\left(\frac{n}{n-m}\right) = (7.16^2)\left(\frac{14}{14-3}\right) = 65.23$$

$$\overline{S}_{y\cdot f(x)} = 8.07$$

Here 3 is used for the value of m, since it is judged that a parabolic equation of type (a), with 3 constants, would be adequate to reproduce the freehand curve.

The standard error of estimate for the graphic regression curve is thus 8.07 ten-pound units, or 80.7 pounds. This is 2.1 pounds smaller than the corresponding value in the case of the linear correlation, indicating how much more closely the curve fits the data than does the straight line, even after allowing for its greater flexibility. In Figure 23 two dotted lines have been drawn in, each 80.7 pounds away from the regression curve, indicating the zone of estimate within which approximately two-thirds of the cases fall (10 out of 14 in this instance) and within which two-thirds of the actual yields may be expected to fall if new estimates of yield are made from the water applied for additional cases drawn from the same universe. (Note also the discussion, in Chapters 18 and 19, of the reliability of such estimates.)

The index of correlation, $\bar{\rho}_{yx}$, may next be computed by substituting the two standard deviations in formula (29):

$$\bar{\rho}^2_{yx} = 1 - \left(\frac{\sigma^2_{z''}}{\sigma^2_y}\right)\left(\frac{n-1}{n-m}\right) \tag{29}$$

This formula includes the corrections for the number of variables and constants. It should always be used in calculating the index of correlation where the curve has been determined freehand, as in this case, since it gives a more accurate measure of the correlation than does equation (23.2), shown previously.

Where the equation of the curve has been determined by mathematical means, the standard error of estimate and the index of correlation may be computed without working out the estimates and residuals for each of the individual cases. These methods will be described subsequently.[7]

In the example given, the index of correlation works out

$$\bar{\rho}^2_{yx} = 1 - \left[\frac{(7.16)^2}{(14.44)^2}\right]\left[\frac{14-1}{14-3}\right] = 1 - 0.2905 = 0.7095$$

$$\bar{\rho}_{yx} = 0.842$$

Since the index of determination is simply $\bar{\rho}^2_{yx}$, it is 71.0 per cent. Comparing these results with those obtained by linear correlation the index of determination of 71.0 per cent compares with the coefficient of

[7] See page 412, Chapter 22.

determination of 69.4 per cent. Apparently taking into account the curvilinear nature of the relations has increased the proportion of the variance in yield accounted for by differences in water application by 1.6 per cent of the total variance in the yield.[8] (Only the measures of determination can be directly compared in this way. If the coefficient of correlation, 0.833, were subtracted from the index of correlation, 0.842, that would give an incorrect idea of the importance of taking account of the curvilinear nature of the relation.)

Interpreting the results of curvilinear correlation. The index of determination and the accompanying standard error of estimate have been interpreted for the curve in much the same manner as were the coefficient of determination and the standard error of estimate for the straight line. In the case of the regression curve itself, however, a somewhat different method of presentation may be best, since a mathematical equation expressing the relation has not been computed.

TABLE 34

YIELD OF PIMA COTTON, WITH DIFFERENT APPLICATIONS OF IRRIGATION WATER, ON MARICOPA SANDY LOAM SOILS IN THE SALT RIVER VALLEY, ARIZONA, IN 1913, 1914, AND 1915

Irrigation water applied	Average yield of cotton lint
Acre-feet	*Pounds per acre*
1.25	156
1.50	222
1.75	283
2.00	335
2.25	385
2.50	431

The regression curve just worked out for the cotton problem, for example, may be presented either as a curve showing graphically the yield to be expected for various applications of water, as is illustrated in Figure 23, or as a table showing the same thing, as in Table 34. In both instances the constant which has been determined from the average of z'' is added to the values read from the curve in Figure 23, $f(X)$, so as to give the final estimates which would be made by taking into account this slight shift in the position of the curve.

[8] See Chapter 18, page 319, for tests as to whether this difference is large enough to be significant.

Similar presentation could be given the regression line in cases of linear correlation, if desired, but then the chart would show only a straight line and the table would show exactly the same changes in the dependent variable for each successive uniform change in the independent variable. In preparing the table, the relation is shown only for that range of water application within which the bulk of the observations fall. Similarly, only this range should be shown by the solid line in the chart; a dotted line might be used to indicate the relations beyond that up to the extremes observed. Neither the regression line nor curve should, ordinarily, be carried beyond the limits of the observations on which it was based. Also, before general conclusions are drawn as to the application of the results to cases other than those included in the sample (as, in this instance, to other fields in the same area), the standard errors set forth in Chapters 18 and 19 should be calculated and included in the interpretation.

Summary. This chapter has illustrated the way in which correlation analysis may be applied to a specific problem, the manner in which linear and curvilinear regressions may be determined most simply, and the way in which they may be interpreted. In addition, the simplest manner of computing the standard error of estimate and the coefficient and the index of correlation have been illustrated, and their significance has been briefly discussed.

CHAPTER 9

THREE MEASURES OF CORRELATION—THE MEANING AND USE FOR EACH

So many different statistical coefficients have been introduced in the discussion of correlation that there may be some confusion among them as to the meaning and use of the different coefficients. Particularly in linear correlation, there are three constants which summarize nearly all that a correlation analysis reveals.

First, the standard error of estimate shows how nearly the estimated values agree with the values actually observed for the variable being estimated. This coefficient is stated in the same units as the original dependent variable, and its size can be compared directly with those values.

Second, the coefficient of determination (r^2) shows what proportion of the variance in the values of the dependent variable can be explained by, or estimated from, the concomitant variation in the values of the independent variable.[1] Since this coefficient is a ratio, it is a "pure number"; that is, it is an arbitrary mathematical measure, whose values fall within a certain limited range, and it can be compared only with other constants like itself, derived from similar problems.

Finally, the coefficient of regression measures the slope of the regression line; that is, it shows the average number of units increase or decrease in the dependent variable which occur with each increase of a specified unit in the independent variable. Its exact size thus depends not only on the relation between the variables but also on the units in which each is stated. It can be reduced to another form, however, by stating each of the variables in units of their own individual standard deviation. In this form it has been termed β or the "beta" coefficient[2] The relation between beta and the coefficient

[1] These statements are all subject to the error limitations set forth later, in Chapters 18 and 19.

[2] See Truman Kelley, *Statistical Method,* p. 282, The Macmillan Co., New York, 1924.

of regression may be indicated by stating the regression equation in both ways:

$$Y = a + b_{yx}X$$

$$\frac{Y}{\sigma_y} = a' + \beta_{yx}\left(\frac{X}{\sigma_x}\right)$$

$$\beta_{yx} = b_{yx}\left(\frac{\sigma_x}{\sigma_y}\right)$$

$$a' = \frac{M_y}{\sigma_y} + \beta_{yx}\frac{M_x}{\sigma_x}$$

Stated in this way, β for the cotton-yield problem is 0.845. That is, for each increase of one standard deviation (0.73 acre-foot of water) in X, the yield of cotton increased 0.845 of one standard deviation. Since the standard deviation of Y was 144.3 pounds, that is equal to 121.9 pounds of cotton for each 0.73 acre-foot of water. This is at the rate of 167 pounds of cotton for each foot of water, which is the same thing as was shown by the coefficient of regression. However, for comparisons between problems where the standard deviations are much different, the "beta" coefficient may have value. It is evident that in simple correlation the value of beta is the same as that of r.

Relation of the different coefficients to each other. Even though each of the three coefficients measures certain aspects of the relation between variables, it does not follow that all three coefficients will vary together, or that a problem which shows a high coefficient of determination will also show a high regression coefficient or a low standard error of estimate. That is because they measure different aspects of the relation.

The particular usefulness of each of the three different groups of correlation measures is illustrated in Figure 24, which shows three sets of simple relationships, with hypothetical data.

Here the regression coefficient is smaller in A than B. In A an additional inch of rain causes an average increase of 2.5 bushels in yield, as compared with an increase of 3.1 bushels in B. But in case A, a considerable part of the variation in yield is apparently due to rainfall, as shown by the high correlation ($r = 0.83$) and the small size of the standard error of estimate (2.2 bushels); whereas in case B, factors other than rainfall apparently cause most of the differences in yield, as indicated by the lower correlation ($r = 0.71$) and the larger standard error of estimate (3.8 bushels). In terms of determination apparently about 69 per cent of the differences in yield are related to differences in rainfall in the first case, and only about 50 per cent in the second.

In comparison with *A* and *B*, case *C* has much less variable yields, ranging from only about 8 bushels to 12 bushels, compared with a range of 8 to 21 in case *A* and 0 to 20 in case *B*. Only a small part (22 per cent) of the variation in yields is associated with rainfall differences, as indicated by the low correlation (0.47). An increase of 1 inch in rainfall apparently causes only 0.5 bushel increase in yield. Yet in spite of this low relation, it is possible to estimate yields more accurately, given the rainfall, in this case than in either of the other two, as is shown by the standard error of estimate of 1.15 bushels as compared to 2.2 bushels for *A* and 3.1 for *B*. The original variation in yields is so slight in case *C* that even the small relation shown to rainfall is enough to make it possible to estimate yields more accurately than in either of the other cases.[3]

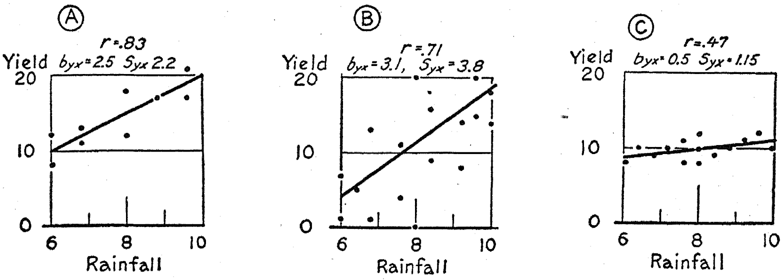


FIG. 24. Hypothetical sets of data, illustrating three types of correlation coefficients.

These three cases illustrate the relative place of each of the three types of correlation measure. Case *B* shows the greatest change in yield for a given change in rainfall (the regression measure); case *A* shows the highest proportion of differences in yields accounted for by rainfall (the correlation or determination measure); and case *C* shows the greatest accuracy of estimate (the error of estimate meas-

[3] In calculating the measures for these illustrative cases, the corrections for numbers of cases have been ignored, as they would not have affected the particular points these examples were set up to illustrate.

ure). Which of these measures should have most attention in a particular investigation depends upon the phase of the investigation which is most important: the *amount* of change (regression); the *proportionate* importance (correlation); or the *accuracy* of *estimate* (standard error). All have their place, and none should be entirely overlooked or ignored.

CHAPTER 10

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (1) BY SUCCESSIVE ELIMINATION

The Problem of Multiple Relations

The relations studied up to this point have all been of the type where the differences in one variable were considered as due to, or associated with, the differences in one other variable. But in many types of problems the differences in one variable may be due to a number of other variables, all acting at the same time. Thus the differences in the yield of corn from year to year are the combined result of differences in rainfall, temperature, winds, and sunshine, month by month or even week by week through the growing season. The premiums or discounts at which different lots of wheat sell on the same day vary with the protein content, the weight per bushel, the amount of dockage or foreign matter, and the moisture content. The speed with which a motorist will react to a dangerous situation may vary with his keenness of sight, his speed of nervous reaction, his intelligence, and his familiarity with such situations. The price at which sugar sells at wholesale may depend upon the production of that season, the carryover from the previous season, the general level of prices, and the prosperity of consumers. The weight of a child will vary with its age, height, and sex. The volume of a given weight of gas varies with the temperature and the barometric pressure.

The physicist and the biologist use laboratory methods to deal with problems of compound or multiple relationship. Under laboratory conditions all the variables except the one whose effect is being studied may be held constant, and the effect determined of differences in the one remaining varying factor upon the dependent variable, while effects of differences in the other variables are thus eliminated. In the case of a gas, for example, the temperature may be held constant while the volume at different barometric pressures is determined experimentally, and then the pressure held constant while the volume at different temperatures is determined. For many of the problems with which the statistician has to deal, however, such laboratory controls cannot be used. Rainfall and temperature and sunshine vary constantly, and only their combined effect upon crop yields can be noted. Economic conditions are constantly shifting, and only the total result of all the factors in the existing situation can be measured at any time. And so on through many other types of multiple relations similar to those mentioned—the statistician has to deal with facts arising from the complex world about him, and frequently has but little opportunity to utilize laboratory checks or artificial controls.

Theoretical example. Where a dependent variable is influenced not only by a single independent variable, as in the relation of Y to X, but also by two or more independent variables, we can represent the relation symbolically by the equation

$$X_1 = a + b_2X_2 + b_3X_3 + \ldots b_nX_n \tag{29.1}$$

Here X_1 represents the dependent variable, and X_2, X_3, . . . X_n represent the several independent variables.

The meaning of the several constants in this equation and the way in which it may be interpreted geometrically can be shown by making up a simple example.

Let us assume that in a new irrigation project the farms are all alike in quality of land and kinds of buildings and that the price at which each one is sold to the settlers is computed as follows:

Buildings, $1,000 per farm
Irrigated land, $100 per acre
Range (non-irrigated) land, $20 per acre.

Using X_1 to represent the selling price per farm in dollars, X_2 to represent the number of acres of irrigated land in each farm, and X_3 to represent the number of acres of range land, we can state the method of computing the selling price in the single equation

$$X_1 = 1{,}000 + 100X_2 + 20X_3$$

The relations stated in this equation may be represented graphically as shown in Figure 24.1. The representation is broken up into halves. The first half shows the relation of farm value to irrigated land for farms that have no range land; the second shows the relation of farm value to range land for farms that have no irrigated land. This figure is constructed exactly the same as was Figure 9 on page 61. Thus in the upper section of Figure 24.1, each change of 1 unit in X_2, as, for example, from 3 to 4, adds 1 unit of b_2, or \$100, to the farm value. Similarly, in the lower section of Figure 24.1, each change of 1 unit in X_3, as, for example, from 5 to 6, adds 1 unit of b_3, or \$20, to the farm value. In each case, as for zero acres, the line begins with the value of a, \$1,000, to cover the value of the buildings.

The equation just shown (29.1) is called the *multiple regression equation.* The term *multiple* is added to indicate that it explains X_1 in terms of two or more independent variables, $X_2, X_3 \ldots X_n$. The

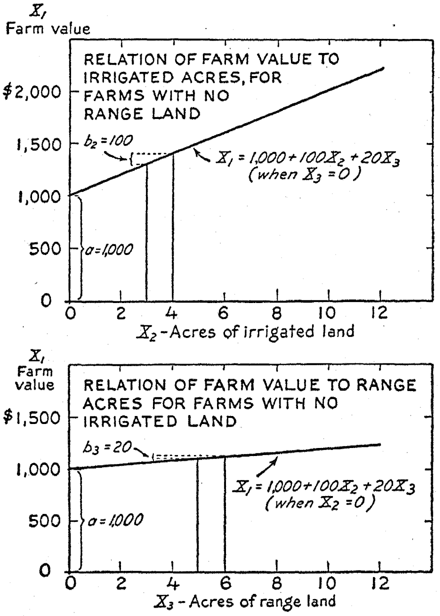


FIG. 24.1. Graph of the function $Y = 1{,}000 + 100X_2 + 20X_3$.

coefficients b_2 and b_3 are termed *net regression coefficients.* The term *net* is added to indicate that they show the relation of X_1 to X_2 and X_3, respectively, excluding, or *net of,* the associated influences of the other independent variable or variables. In contradistinction, the regression coefficient b_{yx} of equation (8),

$$Y = a + b_{yx}X$$

may be termed the *gross regression coefficient.* The term *gross* is added here to indicate that it shows the apparent, or gross, relation between Y and X without considering whether that relation is due to X alone, or to other independent variables associated with X.

The difference between the net and gross regression coefficients may be further shown by a simple arithmetic illustration, based on the farm-value formula just discussed.

Let us take a dozen assumed irrigated farms and calculate from the pricing equation what their selling prices should be. In setting up these illustrative farms, let us assume further that in general the farms with large irrigated areas had small range areas and those with little irrigated land had larger amounts of range land. Under these conditions the computation works out as follows:

TABLE 34.1

COMPUTATION OF ESTIMATED SELLING PRICE, WITH $X_1 = 1{,}000 + 100X_2 + 20X_3$

Observation number	X_2 (1)	X_3 (2)	$100(X_2)$ (3)	$20(X_3)$ (4)	Calculated values of X_1 (3)+(4)+1,000
1	8	5	800	100	1,900
2	4	5	400	100	1,500
3	3	10	300	200	1,500
4	7	8	700	160	1,860
5	7	10	700	200	1,900
6	8	15	800	300	2,100
7	6	12	600	240	1,840
8	1	15	100	300	1,400
9	4	17	400	340	1,740
10	2	22	200	440	1,640
11	4	20	400	400	1,800
12	5	13	500	260	1,760

The apparent relation of the values of X_1, as just computed, to X_2 and X_3 may be shown by preparing dot charts of the X_1 to X_2 relation and the X_1 to X_3 relation. These dot charts are shown in Figure 24.2.

Examining this figure, we find that X_1 is fairly closely related to X_2 but that it has no definite relationship to X_3. We could calculate the regression lines for each of the two relationships shown. The regression coefficient, b_{12}, for the first comparison, would show the average change in X_1 with unit changes in X_2. The regression coefficient, b_{13}, for the second comparison, would show the average change in X_1 with unit changes in X_3. The latter coefficient would come very close to zero, to judge visually from the chart. Both these would be *gross regression coefficients,* measuring only the apparent relation between X_1 and each of the other variables. We know in this case that the values of X_1 are completely determined by the values of X_2 and X_3. If we could hold constant, or eliminate, the true effect of X_2 on X_1, we should find that the relation of the corrected values of X_1 to X_3 was just as close as to X_2. In spite of the fact that the gross regression, b_{13}, appears to be zero, the net regression, b_3, is really 20.

By using the known net regression of X_1 on X_2, we can correct the X_1 values to eliminate that part of their variation which is due to X_2 and then relate the remaining fluctuation to X_3. Let us do that by subtracting b_2X_2 from X_1. This process is shown in Table 34.2.

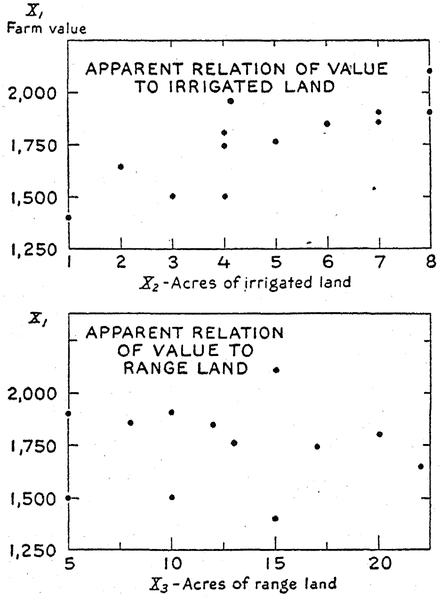


FIG. 24.2. The apparent relation of farm value to acres of irrigated land and to range land reveals little of the underlying net relationship.

We can now plot the values of X_1, corrected for X_2, $X_1 - b_2X_2$, as shown in the sixth column, against the X_3 value, as shown in the third column. The resulting dot chart is shown in Figure 24.3.

This figure now shows the underlying relation between X_1 and X_3, with all the dots falling exactly on one straight line. If we now draw in the regression line and calculate its slope, we shall find it is exactly the same as the line for b_2 which was illustrated in the lower section of Figure 24.1. Figure 24.3 illustrates the *net* regression of X_1 on X_3, as contrasted to the *gross* regression which was represented by the lower section of Figure 24.2. If X_1 were similarly corrected for X_3 and the values $X_1 - b_3X_3$ were plotted against X_2, the net regression of X_1 on X_2 would similarly be shown. (This step is left for the student to perform.)

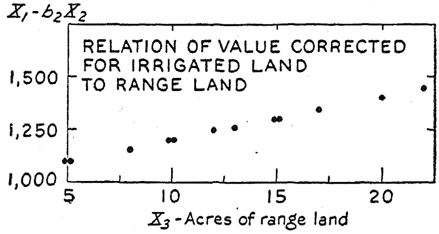


FIG. 24.3. After the net influence of irrigated land has been removed, the underlying relation of farm value to acres of range land is very clear.

If we had not known the underlying relationships as given in this case to start with, but merely had the series of observations of X_1, X_2, and X_3 shown in Table 34.1 and Figure 24.2, would it be possible to work out from those observations the underlying, or *net*, relationships? That is the problem which next will be explored. This time we shall use a series where we do not know the relationship, and see how we

TABLE 34.2

CORRECTION OF COMPUTED X_1 FOR CONTRIBUTION OF X_2

Observation number (1)	X_2 (2)	X_3 (3)	X_1 (4)	b_2X_2 ($100X_2$) (5)	$X_1 - b_2X_2$ (6)
1	8	5	1,900	800	1,100
2	4	5	1,500	400	1,100
3	3	10	1,500	300	1,200
4	7	8	1,860	700	1,160
5	7	10	1,900	700	1,200
6	8	15	2,100	800	1,300
7	6	12	1,840	600	1,240
8	1	15	1,400	100	1,300
9	4	17	1,740	400	1,340
10	2	22	1,640	200	1,440
11	4	20	1,800	400	1,400
12	5	13	1,760	500	1,260

can proceed to work it out. Also, as in most practical cases, we shall use an example where all the causes of variation are not known and where we must deal with independent variables which explain only a part of the variation in the dependent variable.

Practical example. The problem of multiple relations is illustrated by the data in Table 35. These represent 20 farms in one area, with varying crop acreages, dairy cows, and incomes. To determine from these records what income may be expected, on the average, with a given size of farm and with a given number of cows, it is necessary to estimate the effect of differences in the number of acres on income and also the effect of differences in the number of cows on income.

TABLE 35

ACRES, NUMBER OF COWS, AND INCOMES, FOR 20 FARMS

Record no.	Size of farm	Size of dairy	Income
	Number of acres	*Number of cows*	*Dollars per year*
1	60	18	960
2	220	0	830
3	180	14	1,260
4	80	6	610
5	120	1	590
6	100	9	900
7	170	6	820
8	110	12	880
9	160	7	860
10	230	2	760
11	70	17	1,020
12	120	15	1,080
13	240	7	960
14	160	0	700
15	90	12	800
16	110	16	1,130
17	220	2	760
18	110	6	740
19	160	12	980
20	80	15	800

From these data it would seem that both the size of the farm and the size of the dairy herd influence farm income, to judge from dot charts showing the relation of income to acres (Figure 25) and of income to number of cows (Figure 26). It appears from these charts that there may be a slight tendency for the farms with the larger acreage in crops to have larger incomes and a rather marked tendency for the farms with the larger number of cows to have larger incomes.

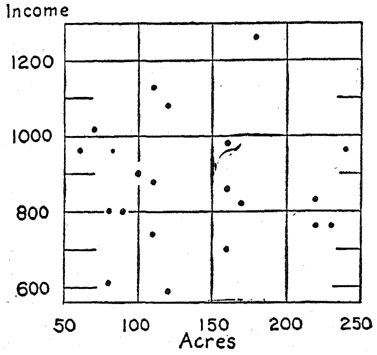


FIG. 25. Correlation chart of acres and income on individual farms.

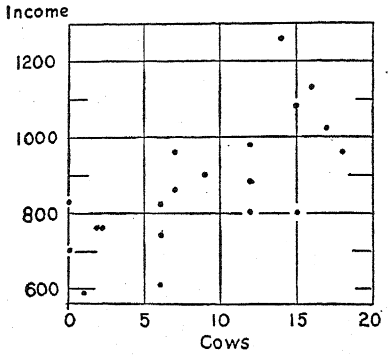


FIG. 26. Correlation chart of number of cows and income on individual farms.

Analysis by simple averages not adequate. The simple comparison alone, however, is not sufficient to tell exactly how incomes change with acres and with number of cows. That is because there is a marked relation between the size of the farms and the number of cows, as is illustrated in Figure 27. There is a definite tendency for the larger farms to have smaller dairy herds. As a result, the difference in incomes in Figure 25, which appeared to be due directly to differences in acreages, may be due in part to the differences in the sizes of the dairy herds on the farms with different acreages in crops. If we make groups of farms of 50 to 99 acres, 100 to 150 acres, and so on, and average the acres, cows, and income for each group, as is shown in Table 36, we find a marked difference in the number of cows from group to group, as well as in the number of acres and in the incomes.

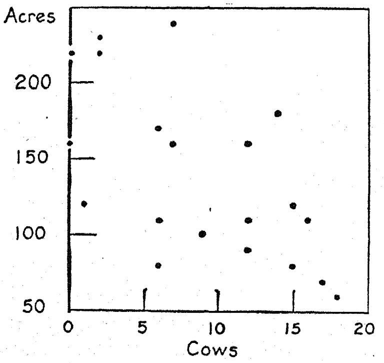


FIG. 27. Correlation chart of number of cows and number of acres on individual farms.

TABLE 36

AVERAGE NUMBER OF COWS AND INCOME, FOR FARMS OF DIFFERENT SIZES

Size group	Number of farms	Average size	Average size of dairy	Average income
		Number of acres	*Number of cows*	*Number of dollars*
50–99 acres	5	76	13.6	838
100–149 acres	6	111	9.8	887
150–199 acres	5	166	7.8	924
200–249 acres	4	228	2.8	828

The farms of 50 to 99 acres, with an average size of 76 acres, have incomes which average $838; the farms of 150 to 199 acres, with an average size of 166 acres, show incomes which average $924. Is this difference in income due to the difference in size? Before this can be definitely answered we must consider that the two groups also differ in the average number of cows, with 13.6 in the first group and only 7.8 in the second. So far, there is nothing to indicate whether the difference in income is due to the difference in the size of the farms or in the number of cows; we have shown that both vary from group to group, and that is all.

If, on the other hand, we should attempt to determine how far income varied with differences in the number of cows by classifying the records with respect to the number of cows, and averaging incomes, we should secure the result shown in Table 37.

TABLE 37

AVERAGE ACRES AND INCOME, FOR FARMS WITH DIFFERENT NUMBERS OF COWS

Size of herd	Number of farms	Average size of dairy	Average size of farms	Average income
		Number of cows	*Number of acres*	*Number of dollars*
Under 5 cows	5	1.0	190	728
5–9 cows	6	6.8	143	815
10–14 cows	4	12.5	135	980
15 cows and over	5	16.2	88	998

Even though the income is higher on the farms with more cows, Table 37 does not indicate how much of that can be credited to the cows and how much to other factors. It is evident from the table that as the number of cows goes up, the number of acres goes down; are the differences in income associated with changes in number of cows, in number of acres, or in part with both?

Eliminating the approximate influence of one variable. What we need to know is how far income varies with size of farm, as between farms with the same number of cows; and how far income varies with the number of cows, as between farms of the same size as to acres. One way of determining this would be to adjust the income on each farm to eliminate the differences due to (or associated with) the number of cows, and then compare the adjusted incomes with the size of the farm to determine the effect of size on income. To start this process the effect of the number of cows upon incomes is needed. We can secure an approximate measure of this by determining the straight-

TABLE 38

ADJUSTING FARM INCOMES FOR DIFFERENCES IN NUMBER OF COWS

Size of farm	Size of dairy	Income	Income assumed due to cows	Income adjusted to no-cow basis
Number of acres	*Number of cows*	*Dollars*	*Number of dollars*	*Number of dollars*
60	18	960	362	598
220	0	830	0	830
180	14	1,260	282	978
80	6	610	121	489
120	1	590	20	570
100	9	900	181	719
170	6	820	121	699
110	12	880	241	639
160	7	860	141	719
230	2	760	40	720
70	17	1,020	342	678
120	15	1,080	302	778
240	7	960	141	819
160	0	700	0	700
90	12	800	241	559
110	16	1,130	322	808
220	2	760	40	720
110	6	740	121	619
160	12	980	241	739
80	15	800	302	498

line equation for estimating incomes from cows—approximate only, since the differences in the size of the farms are ignored at this point.

Determining the straight-line relation according to Chapter 5, we find that the relation between cows and income is given by the equation:

$$\text{Income} = \$694 + 20.11 \text{ (number of cows)}$$

According to this equation, farms with no cows averaged about $694 income, and these incomes increased $20.11 for each cow added, on the average. Knowing this relation, we can adjust the incomes on the several farms by deducting that part of the income which would be assumed due to the cows, according to this average relation.

Table 38 illustrates the process of adjusting the incomes to a no-cow basis, by subtracting out this approximate effect of cows on incomes. The next step is to see what the relation is between the acres in the farm and these adjusted incomes. Plotting both on a dot chart, Figure 28 shows this relation graphically. Comparing this figure with Figure 25, where the relation between the acres and the unadjusted incomes was plotted, we see that the relation is much closer and more definite for the adjusted incomes than for the unadjusted incomes. This is only natural; now that the marked relation of number of cows to income has been removed, even if only approximately, the underlying relation of size to income can be more clearly seen.

It is evident from Figure 28 that size has a more marked effect upon income than appeared in Figure 25, where the effect of cows was mixed in also. As was pointed out earlier, the fact that cows and acres were correlated meant that the effects of differences in cows were mixed in with the effects of differences in acres. Now that the effect of cows has been at least roughly removed, the change in incomes with changes in acres can be more accurately determined.

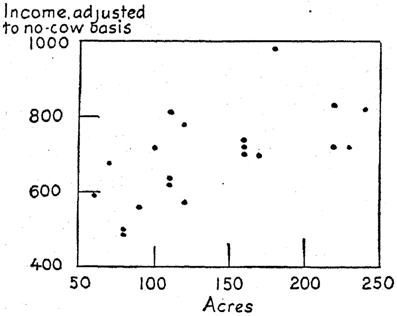


FIG. 28. Relation of income, adjusted for number of cows, to number of acres.

Fitting straight lines to the relations shown in Figures 25 and 28, to determine the average change in income with changes in acres, we obtain regression equations as follows:

$$\text{Income} = \$868.74 + \text{(number of acres) } \$0.0234$$

$$\text{Income, effect of cows removed,} = \$508.51 + \text{(number of acres) } \$1.33$$

It is evident that the determination of the effect of acres upon income without making some allowance for the effect of the correlated variable, number of cows, in this case would have seriously underestimated the effect of acres upon income. Such a determination would have shown only $0.02 increase in income for each acre increase in size, whereas the later determination shows $1.33 increase in income for each acre increase in size.

The relation now shown between income and acres illustrates the extent to which one variable may really influence a second, even though its influence is concealed by the presence of a third variable. From Figure 25, which indicates that there is practically no correlation between acres and income, one might conclude that differences in income were not at all associated with differences in acreage; yet when the variation in income associated with cows is removed, even by the rough method shown, a very definite relation of income to size is found. For that reason one cannot conclude that, because two variables have no correlation, they are not associated with each other; the lack of correlation may be due to the compensating influence of one or more other variables, concealing the hidden relation.

Eliminating the approximate influence of both variables. We now have two equations, one showing the effect of cows upon income and the other the effect of acres:

(A) Income = \$694 + (number of cows) \$20.11

(B) Income, effect of cows removed,
= \$508.51 + (number of acres) \$1.33

These two equations can be combined into a single equation by taking that part of the first one which shows the increase in income for each cow and adding it to the second one. This gives an equation which includes allowances for both factors, as follows:

(C) Income = \$508.51 + (number of acres) \$1.33
+ (number of cows) \$20.11

The last equation gives a basis for indicating the effect of both acres and cows on income and for computing the income that might be expected, on the average, with a farm of a given size and with a given number of cows. For example, for a farm of 120 acres and 15 cows, the expected income would work out as follows:

$$\text{Income} = \$508.51 + (120)\ \$1.33 + (15)\ \$20.11$$
$$= \$508.51 + \$159.60 + \$301.65 = \$970$$

If 5 cows were added, making it 120 acres and 20 cows, the estimated income would be:

$$\begin{aligned}\text{Income} &= \$508.51 + (120)\ \$1.33 + (20)\ \$20.11 \\ &= \$1070\end{aligned}$$

Or if 50 acres were added, making 170 acres and 15 cows, the income would be estimated:

$$\text{Income} = \$508.51 + (170)\ \$1.33 + (15)\ \$20.11$$

TABLE 39

ACTUAL INCOME AND INCOME ESTIMATED FROM NUMBER OF ACRES AND COWS

Acres	Cows	Computation of estimated income:		Estimated income (A) + (C) +$508.51	Actual income	Actual income minus estimated income
		Estimate for acres $1.33 (acres) (A)	Estimate for cows $20.11 (cows) (C)			
60	18	$ 80	$362	$ 950.5	$ 960	$ 9.5
220	0	293	0	801.5	830	28.5
180	14	239	282	1,029.5	1,260	230.5
80	6	106	121	735.5	610	−125.5
120	1	160	20	688.5	590	− 98.5
100	9	133	181	822.5	900	77.5
170	6	226	121	855.5	820	− 35.5
110	12	146	241	895.5	880	− 15.5
160	7	213	141	862.5	860	− 2.5
230	2	306	40	854.5	760	− 94.5
70	17	93	342	943.5	1,020	76.5
120	15	160	302	970.5	1,080	109.5
240	7	319	141	968.5	960	− 8.5
160	0	213	0	721.5	700	− 21.5
90	12	120	241	869.5	800	− 69.5
110	16	146	322	976.5	1,130	153.5
220	2	293	40	841.5	760	− 81.5
110	6	146	121	775.5	740	− 35.5
160	12	213	241	962.5	980	17.5
80	15	106	302	916.5	800	−116.5

Equation (C) can be used as illustrated, to work out what income might be expected, on the average, for each of the farms shown in Table 39. The estimated income can then be compared with the actual income and the difference, if any, determined.

As is illustrated in Table 39, the estimated incomes vary somewhat from the actual. This is just another way of saying that all the differences in income cannot be accounted for by the effect of differences in acres and in cows, according to the relations summarized in equation (C). This failure of the estimated values to agree exactly with the original values is seen graphically in Figure 28 by the fact that all the dots do not lie exactly along the regression line. Subtracting the estimated values from the actual values gives the residual differences of the actual income above or below the income estimated from the two factors, acres and cows.

Correcting results by successive elimination. It may now be recalled that, even though the incomes were adjusted to eliminate the effects of cows upon income before determining the relation between income and acres, the determination of the relation between income and cows was made without making any allowance for the concurrent effect of acres. Since we now have an approximate measure of the effect of acres determined while eliminating to some extent the effect of cows, we can use that new measure, equation (B), to adjust the incomes for the effect of the acres and then get a more accurate measure of the true effect of cows alone upon incomes. This process is shown in Table 40. Here estimates of income are worked out by equation (B) on the basis of acres, showing what the incomes might be expected to average if all the farms had no cows. The difference between these estimates and the actual incomes may then be considered to be the part due to cows alone, while eliminating the effect of differences in the numbers of acres. On the first farm, for example, equation (B) indicates that with no cows the income for 60 acres should be \$588. Subtracting this from the \$960 actually received leaves \$372 as the income apparently accompanying the 18 cows.

The adjusted incomes may then be plotted on a dot chart with the number of cows as the other variable, as shown in Figure 29. Comparing this figure with Figure 26, where the number of cows was plotted against income without first making any adjustment in the original incomes, we easily see how much closer the relation is after making the adjustment. Further, it is evident that cows have a greater effect upon income than was indicated by the earlier comparison. Computing the straight-line relationship for Figure 29 gives the equation:

(D) Income, adjusted to constant acres,

$$= -\$68.77 + (\text{number of cows})\ \$27.88$$

By this last computation (equation [D]), each increase of one cow causes an average increase in income of $27.88, whereas according to the earlier comparison (equation [A]), each increase of one cow caused an average increase in income of only $20.11. The second value is larger than the first, again showing the necessity of making allowances for the effect of one factor before the true value of the other can be properly measured.

Now that we have a new measure of the effect of cows, we might go on to adjust incomes for cows by this new measure and then get a revised value for the effect of acres upon incomes on a no-cow basis, in place of the relation shown in equation (B). This possibility of further correction will be referred to later. But before that we will make some experiments with the new equation (D).

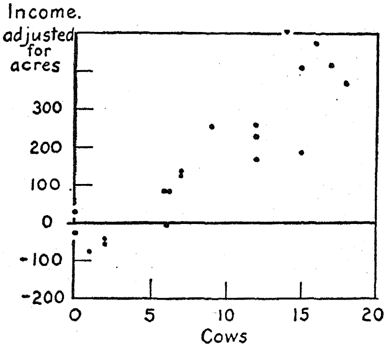


FIG. 29. Relation of income, adjusted for number of acres, to number of cows.

We now have equations for the relation of incomes, adjusted for the other factors, to the remaining factors. These two equations, (B) and (D), are:

(B) Income, effect of cows removed,
= $508.51 + (number of acres) $1.33

(D) Income, adjusted to constant acres,
= – $68.77 + (number of cows) $27.88

These two equations may be combined to give a revised equation to indicate the effect of both cows and acres upon incomes, equation (E).

(E) Income = $439.74 + (number of acres) $1.33
+ (number of cows) $27.88

Equation (E) is exactly the same as the previous equation (C) except that the revised effect of cows is included, and the constant term has also been changed owing to changing the allowance for cows.

In exactly the same way that equation (C) could be used to work out the estimated income for any given combination of cows and acres, equation (E) can be also used. Thus for 120 acres and 15 cows, it would give

$$\begin{aligned}\text{Estimated income} &= \$439.7 + (120)\ \$1.33 + (15)\ \$27.88\\ &= \$439.7 + \$159.6 + \$418.2 = \$1{,}018\end{aligned}$$

TABLE 40

Adjusting Farm Incomes for Differences in Number of Acres

Size of farm	Size of dairy	Income	Income estimated for acres, with no cows	Income with effects of acreage differences eliminated *
Number of acres	*Number of cows*	*Dollars*	*Number of dollars*	*Number of dollars*
60	18	960	588	372
220	0	830	801	29
180	14	1,260	748	512
80	6	610	615	− 5
120	1	590	669	− 79
100	9	900	642	258
170	6	820	735	85
110	12	880	655	225
160	7	860	722	138
230	2	760	815	− 55
70	17	1,020	602	418
120	15	1,080	669	411
240	7	960	828	132
160	0	700	722	− 22
90	12	800	629	171
110	16	1,130	655	475
220	2	760	802	− 42
110	6	740	655	85
160	12	980	722	258
80	15	800	615	185

* Where the actual income is below that expected for a farm of that size with no cows, the deficit is indicated by the minus sign.

The result, $1,018, is $48 higher than the $970 worked out by equation (C). This higher estimate is due to the fact that equation (E) makes a larger allowance for the effect of each cow, and 15 is more than the average number of cows. If less than the average number of cows were used, equation (E) would give a lower estimate than equation (C).

Working out the estimated incomes for each of the original observations according to equation (E), we obtain results as shown in Table 41.

TABLE 41

ACTUAL INCOME AND INCOME ESTIMATED FROM NUMBER OF ACRES AND NUMBER OF COWS, REVISED RELATIONS

Acres	Cows	Computation of estimated income		Estimated income, (A) + (B) +$439.7	Actual income	Actual income minus estimated income
		Estimate for acres $1.33 (acres) (A)	Estimate for cows $27.88 (cows) (B)			
60	18	$ 80	$502	$1,021.7	$ 960	−$ 61.7
220	0	293	0	732.7	830	97.3
180	14	239	390	1,068.7	1,260	191.3
80	6	106	167	712.7	610	−102.7
120	1	160	28	627.7	590	− 37.7
100	9	133	251	823.7	900	76.3
170	6	226	167	832.7	820	− 12.7
110	12	146	335	920.7	880	− 40.7
160	7	213	195	847.7	860	12.3
230	2	306	56	801.7	760	− 41.7
70	17	93	474	1,006.7	1,020	13.3
120	15	160	418	1,017.7	1,080	62.3
240	7	319	195	953.7	960	6.3
160	0	213	0	652.7	700	47.3
90	12	120	335	894.7	800	− 94.7
110	16	146	446	1,031.7	1,130	98.3
220	2	293	56	788.7	760	− 28.7
110	6	146	167	752.7	740	− 12.7
160	12	213	335	987.7	980	− 7.7
80	15	106	418	963.7	800	−163.7

Comparing the residuals, or differences between the actual and estimated income, obtained by means of this new equation with those obtained using the equation in its first form (shown in Table 39), we see that in more than half the cases they are smaller with the revised form. A more definite comparison can be made by computing the standard deviation of the residuals in each case. The standard deviation of the residuals shown in Table 39, using equation (C), is \$90.29, whereas the standard deviation of the residuals shown in Table 41, using equation (E), is but \$78.70. It is apparent from this that the revised equation, determined after the effects of the other variables had been eliminated, gives more accurate estimates of income than does the original equation in which the effects of the other variables had not been so fully eliminated.

It was suggested previously that the last corrected values for the relation of cows to income gave a new basis for correcting income so as to measure more accurately the relation of acres to income. This in turn would give a new basis for measuring the effect of cows, and so on, until a final stable value had been reached. So long as a new correction would result in a further change in the computed effect of either variable, the new values would give a better basis for estimating income than did the previous values. Only when the point was reached where no further change need be made in the effect of either variable could it be said that the relation of each variable to income had been quite correctly measured while allowing for the influence of the other factor, and that might involve a large number of successive corrections.

This method of allowing for the effect of other factors so as to determine the true relation of each one to the dependent factor (as income, in this case), by first correcting for one, and then for another, is known as the method of successive elimination. This method can be used where there are three or more independent factors related to (or accompanying variations in) a dependent (or resultant) factor just as it was used here for two factors, except that then the dependent needs to be corrected in turn to eliminate the effects of all the other independent factors except the particular one whose effect is being measured. But although it is possible to measure the relations by this method, it would be a very slow and laborious process. A shorter mathematical method which gives the same result by more direct processes is available instead. This method, known as the method of multiple correlation, is presented in detail in Chapter 12.

Summary. This chapter has shown that when two related factors both affect a third factor it is difficult to measure the effect of either factor upon the third without the result being affected by both causal factors. Allowing for this duplication by eliminating the effects of each factor in turn (successive elimination) can gradually determine the true effect of each, but the method is long and laborious.

CHAPTER 11

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (2) BY CROSS-CLASSIFICATION AND AVERAGES

We have previously seen (Chapter 4) how the relation between two variables can be studied by means of averages. An extension of the same method can be used for problems where two or more variables affect a third variable, such as that discussed in the last chapter.

Analysis by averages where there are two independent variables involves classifying the records first by one variable, then breaking each of the resulting groups into several smaller groups according to the values of the second variable. If a third independent variable were to be considered, these groups would be broken up into still smaller groups, according to the values of the third variable. Then the values of the dependent variable, as well as each of the independent variables, would be averaged for each subgroup. This process is known as subclassification or cross-classification.

Cross-classification for three variables. In the problem presented in the last chapter, there were two independent variables—number of cows and number of acres. The records would therefore need to be classified into groups both according to the number of cows and the number of acres on each farm. Since there is such a small number of records the groups should not be made too small. Let us take three groups for cows; less than 6, 6 to 11, and 12 and over; and four groups for the size of farm; from 50 to 99 acres, from 100 to 149, from 150 to 199, and 200 acres and over. This will give us twelve possible groups in all. The records may be classified into these twelve groups and totals and averages computed for each, as shown in detail in Table 42.

It is apparent that none of these groups has a sufficient number of farms represented to make the averages particularly significant; yet even at that a certain regularity in the averages can be observed. In each column the average income increases as the size of farm increases, though there is but little difference in the average number of cows from group to group; similarly across each line of averages the income increases as the number of cows increases, though there is but little difference in the average size of farm from group to

TABLE 42

Cross-classification of Reports According to Size of Farm and Size of Dairy Herd

Size of farm	Size of dairy herd								
	Under 6 cows			6 to 11 cows			12 cows and over		
	Acres	Cows	Income	Acres	Cows	Income	Acres	Cows	Income
	Number	*Number*	*Dollars*	*Number*	*Number*	*Dollars*	*Number*	*Number*	*Dollars*
50 to 99 acres	......	......	......	80	6	610	60	18	960
	......	......	......	......	......	......	70	17	1,020
	......	......	......	......	......	......	90	12	800
	......	......	......	......	......	......	80	15	800
Total.......	......	......	......	......	......	......	300	62	3,580
Average.....	......	......	......	80	6	610	75	15.5	895
100 to 149 acres	120	1	590	100	9	900	110	12	880
	......	......	......	110	6	740	120	15	1,080
	......	......	......	......	......	......	110	16	1,130
Total.......	......	......	......	210	15	1,640	340	43	3,090
Average.....	120	1	590	105	7.5	820	113	14.3	1,030
150 to 199 acres	160	0	700	170	6	820	180	14	1,260
	......	......	......	160	7	860	160	12	980
Total.......	......	......	......	330	13	1,680	340	26	2,240
Average.....	160	0	700	165	6.5	840	170	13	1,120
200 acres and over	220	0	830	240	7	960			
	230	2	760						
	220	2	760						
Total.......	670	4	2,350						
Average.....	223	1.3	783	240	7	960			

group. These relations may be more clearly seen in Figures 30 and 31, where the average incomes from Table 42 are charted, first for differences in the number of cows with farms of similar sizes, and then for differences in the number of acres, with farms of similar numbers of cows.

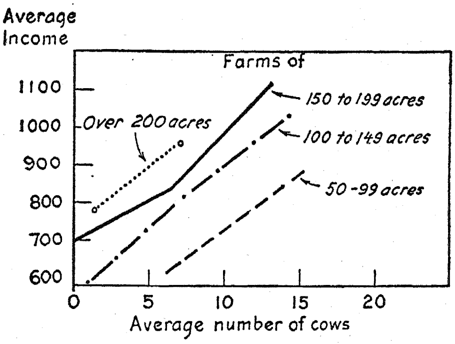


FIG. 30. Difference in average income with difference in number of cows, for farms grouped by size of farm.

Both figures show the tendency for income to increase with an increase in the independent variable, when the effect of the other variable is held fairly constant by the grouping process. In Figure 30 the lines show about the same general slope for each of the four groups, though there are some irregularities. Figure 31 similarly shows about the same general change in income with a given change in the size of the farm, no matter what is the number of cows; but here the irregularities from group to group are even more striking.

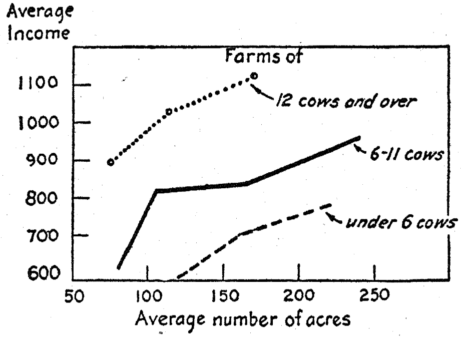


FIG. 31. Difference in average income with difference in number of acres, for farms grouped by numbers of cows.

In Chapter 4 it was shown that such irregularities from group to group might readily be due to random errors of sampling. In the present case, the number of items in each group is so small that it would be hardly worth while to compute the standard error for each average. Even if there were many more cases in each group than are available here, differences as large as those shown might be due simply to random differences in sampling and therefore have no real meaning as indicating differences prevailing in the universe from which the sample was selected.

Although the averages obtained by the process of subsorting may be considered to show the general effect of changes in one variable, such as cows, upon income, with the effect of the other variable, such as acres, removed, they cannot be considered to show the specific effect of specific differences. For example, much more evidence would be needed to prove that, between 75 and 100 acres, a change of 1 acre has much greater effect upon income on farms with 6 to 11 cows than on farms with 12 cows or more, even though the lines in Figure 31 would appear to indicate this. All that is really proved is that on farms of both numbers of cows there is a tendency for income to increase with an increase in the number of acres.

TABLE 43

Difference in Average Income for Farms of Different Sizes and With Different Sizes of Dairy Herd

Size of farm	Under 6 cows in herd		6 to 11 cows in herd		12 cows or over in herd	
	Size of group	Average income	Size of group	Average income	Size of group	Average income
	Number of farms	*Dollars*	*Number of farms*	*Dollars*	*Number of farms*	*Dollars*
50 to 99 acres...	....	....	1	610	4	895
100 to 149 acres...	1	590	2	820	3	1,030
150 to 199 acres...	1	700	2	840	2	1,120
200 to 249 acres...	3	783	1	960	....	....

The averages obtained by the process shown in Table 42 may be summarized for publication in a form similar to Table 43. The number of cases represented in each average is included to prevent the reader from placing an undue amount of confidence in an average based on a small number of observations. In addition, each should be followed by ± its own standard error.

The very small number of cases included in each of the groups is strikingly brought out in Table 43. Even if there were five times as many farms to deal with—100 in all—if they were distributed in the same manner, the largest group would have only 20 cases, and all the rest would have 15 or less, which, under ordinary conditions, would be hardly enough for really significant averages.

Average differences between matched sub-groups. After the observations have been grouped and averaged as shown in Table 43, average differences in the dependent variable (as here, dollars of income), with given differences in each independent variable, can be roughly determined while holding constant the other independent variable or variables. This involves determining the average differences between the averages for the dependent variable for matched groups. The computations are shown in Tables 43.1 and 43.2.

TABLE 43.1

CHANGE IN AVERAGE INCOME BETWEEN GROUPS MATCHED FOR SIZE OF FARM

Size of farm	*A* Under 6 cows	*B* 6 to 11 cows	*C* Increase (*B*−*A*)	*D* Over 12 cows	*E* Increase (*D*−*B*)
Acres	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
50–99	....	610	....	895	285
100–149	590	820	230	1,030	210
150–199	700	840	140	1,120	280
200–249	783	960	177	....	....
Average change with cows.	....	....	182	....	258

From these results it appears that increasing the number of cows from under 6 to between 6 and 11, without changing the size of farm, was accompanied by an average increase of $182. Increasing the cows further to over 12 cows was accompanied by a further increase of income of $258. Similarly, increasing the size of farm from under 99 acres to 100-149 acres, without changing the number of cows, was accompanied by an increase of $173 in income. A further increase to 150-199 acres was accompanied by a further average increase of $73 in income, and to 200-249 acres, by $102 more income. (In this discussion "increase" in size or cows has been used to designate differences between results for farms of different sizes or with different number of cows.) These rough measurements of differences in the dependent variable with differences in one independent variable, while holding a second independent constant by subsorting, may be compared with results obtained by the more exact methods set forth in subsequent chapters.[1]

This same method may be applied to get the average difference between matched subgroups, where two or more other independent variables are held constant by the grouping.

TABLE 43.2

Change in Average Income between Groups Matched for Number of Cows

Number of cows	*A* 50–99 acres	*B* 100–149 acres	*C* Increase ($B-A$)	*D* 150–199 acres	*E* Increase ($D-B$)	*F* 200–249 acres	*G* Increase ($F-D$)
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
Under 6	......	590	......	700	110	783	83
6 to 11	610	820	210	840	20	960	120
12 or over	895	1,030	135	1,120	90	......	......
Average change with acres	......	......	173	......	73	......	102

Limitation of cross-classification for many variables. This small problem illustrates one fundamental difficulty with the method of subclassification and averaging—the large number of cases required for conclusive results. Though there are only two independent variables involved, and the records are classified into only three groups one way and four the other, apparently 100 cases or more would be required for really significant results. If it had been desired to subclassify the records according to two more additional variables—say number of men employed and number of hogs kept—that would have greatly increased the number of records necessary. If each of the

[1] In computing Tables 43.1 and 43.2, no attention was paid to weighting the results according to the number of cases falling in each group, or to the sampling reliability of each average. For a discussion of the first of these points, and for possible methods of dealing with it, see F. A. Harper, Analyzing data for relationships, *Cornell University Agricultural Experiment Station Memoir* 231, June, 1940.

TABLE 44

FORM FOR SHOWING DIFFERENCES IN AVERAGE INCOME FOR FARMS CLASSIFIED BY ACRES, MEN EMPLOYED, COWS, AND HOGS

Area and number of hogs	1 man		2 men		3 men	
	Size *	Average income	Size *	Average income	Size *	Average income
	Under 6 cows					
Farms of 50 to 99 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						
Farms of 100 to 149 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						
	6 to 11 cows					
Farms of 50 to 99 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						
Farms of 100 to 149 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						
	12 cows and over					
Farms of 50 to 99 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						
Farms of 100 to 149 acres:						
Under 20 hogs........						
20–39 hogs............						
40 hogs and over......						

Etc.

* Number of reports in group.

groups already shown had been further divided into 1-man, 2-man, and 3-or-more-man farms, and each of these sub-groups had been further divided into farms with less than 20 hogs, 20 to 39 hogs, and 40 or more hogs, that would have increased the number of possible groups from 12 to 108. Where over 100 records would have been needed in the first case to give results at all reliable, probably a thousand or more records would be needed with this further classification. Although such large numbers of records are available in some types of work, as in census tabulations, they are rarely obtainable in most economic or social-science studies, and for that reason treatment of a large number of variables by the method of detailed sub-classification has but limited application in this field.

The way in which a fourfold classification, such as that described in the preceding paragraph, might be presented is indicated by the form in Table 44, even though it would only occasionally be used.

In addition to the large number of cases required to obtain reliable results, the method of sub-classification and averaging has further shortcomings; it provides no measure of how *important* the relation shown is as a cause of variation in the factor being studied, or of how closely that factor may be estimated from the others on the basis of the relations shown. Thus Table 43 shows that, on the average, certain differences in the number of cows and in the number of acres were accompanied by certain differences in the average income. By itself, however, it did not give any indication of how closely the income could be estimated if the number of acres or the number of cows were known; nor did it indicate the proportion of the variance in income which can be explained by concurrent differences in size of farm and size of dairy. For these reasons, as well as because of the large number of cases necessary to obtain reliable conclusions, the method of sub-classification and averaging does not determine the relationships where many variables are involved so satisfactorily as do other methods, which will be considered in subsequent chapters.

Significance of differences in group averages. When the data are classified as shown above, the results may be tested to determine whether the differences found between successive group averages are significant, or whether they might have occurred by chance. One method for testing this is to compute the standard error for each group average and to consider these standard errors in judging whether or not the differences are significant.[2] A second method of judging the significance of the differences is by determining whether the variation between the averages of the columns or cells is or is not significant, as compared to the variation between the individual items which fall in each column or cell. Relatively simple methods, set forth in standard textbooks,[3] are available for this "analysis of variance." Since these methods relate only to the *significance* of the observed differences, and not to the functional nature of the relations which underlie those differences, they are not presented here.

Summary. The relation of one variable to several others may be approximately determined by detailed cross-classification. Very large numbers of records are required to make the averages accurate, however, since the number of groups increases rapidly with additional variables. Further, the averages by themselves give no indication of the closeness of correlation.

[2] Formulas for the standard errors of the difference between two group averages are given by G. Udny Yule and M. G. Kendall in their *Introduction to the Theory of Statistics* (eleventh edition), pp. 387–88, C. Griffin and Co., Ltd., London, 1937.

[3] Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics*, pp. 351–59, Prentice-Hall, Inc., New York, 1939.

R. A. Fisher, *Statistical Methods for Research Workers* (seventh edition), Chapter VIII, Oliver and Boyd, London and Edinburgh, 1938.

G. W. Snedecor, *Statistical Methods Applied to Experiments in Agriculture and Biology*, Chapters 10, 11, Iowa State College Press, Ames, Iowa, 1937.

CHAPTER 12

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE VARIABLES CHANGE: (3) BY USING A LINEAR REGRESSION EQUATION

In Chapter 10 it was shown that an equation could be arrived at to express the average relation between income, acres, and cows, as follows:

Equation (E)

$$\text{Income} = 439.74 + 1.33 \text{ (number of acres)} + 27.88 \text{ (number of cows)}$$

If we designate the three series of variable quantities, income, acres, and cows, by the symbol X with different subscripts, using X_1 to represent dollars of income, X_2 to represent number of acres, and X_3 to represent the number of cows, we can rewrite the equation in the form

$$X_1 = 439.74 + 1.33X_2 + 27.88X_3$$

If now we use the symbol a to represent the constant quantity 439.74; b_2 to represent 1.33, the amount which X_1 increases for each increase of one unit in X_2 (one acre); and b_3 to represent 27.88, the amount which X_1 increases for each increase of one unit in X_3 (one cow); the equation appears as

$$X_1 = a + b_2X_2 + b_3X_3 \qquad (30)$$

Comparing this equation with the regression equation for the straight-line relation between two variables

$$Y = a + bX$$

we see that the two equations are just alike, except for the difference in the symbols used to represent the different variables and for our having added the expression for an additional variable. In equation (30), X_1, the variable which is being estimated, is termed the *dependent* variable, since its estimated value depends upon those of the other variable or variables; and X_2 and X_3 are termed *independent* variables, since their values are taken just as observed, independent of any of the conditions of the problem. Since there is more than one independent variable concerned, the equation is said to be a multiple estimating equation, or a *multiple linear regression equation.*

Chapter 10 showed that the values of the constants a, b_2, and b_3, which in the particular problem considered indicate what the average income would be for a farm and dairy of any given size, could be worked out by a cut-and-try method which gradually approached nearer and nearer to the right values. It is evident, however, that for any particular criterion of "rightness" only one set of values for these constants can be exactly right. If the criterion of "rightness" is taken as that which will make the standard deviation of the residuals, when income is estimated from the other two variables, as small as possible, the values of a, b_2, and b_3 which will give this result can be determined once and for all by a direct mathematical process. Determining these values so as to give the "best" equation for estimating X_1 on the basis of linear relations to X_2 and X_3 is the first step in the method of *linear multiple correlation.*

Determining a regression equation for two independent variables. The best values for a, b_2 and b_3 in the multiple regression equation (30), can be worked out by an extension of the same process used in working out the values for the estimating equation when only one independent variable was considered. Just as before, the value of the b constants will be determined first, equation (31), and then the a values will be worked out from them:[1]

$$\left.\begin{aligned}\Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 &= \Sigma(x_1x_2)\\ \Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 &= \Sigma(x_1x_3)\end{aligned}\right\} \tag{31}$$

$$a = M_1 - b_2M_2 - b_3M_3 \tag{32}$$

Here, just as in Chapter 5, the symbol M represents the mean value of each variable, and the subscript indicates the particular variable.

Similarly, the symbols $\Sigma(x_2x_3)$, $\Sigma(x_1x_2)$, and $\Sigma(x_1x_3)$ represent the sums of the products of the variables, corrected to adjust them to deviations from the mean; that is, $\Sigma(x_1x_2) = \Sigma[(X_1 - M_1)(X_2 - M_2)]$. Likewise the symbols $\Sigma(x_2^2)$, etc., represent the sums of the squares of the variables, also adjusted to deviations from the mean.

[1] See Note 6, Appendix 2, for the derivations of these equations. They are the normal equations for two independent variables, corresponding to the normal equations for one independent variable given on page 67, in the footnote.

Using the two basic formulas

$$\Sigma(x_1x_2) = \Sigma(X_1X_2) - nM_1M_2 \qquad (11)$$

and

$$\Sigma(x_2^2) = \Sigma(X_2^2) - n(M_2^2)$$

the other values shown in equation (31) may be worked out as follows:

$$\Sigma(x_1x_3) = \Sigma(X_1X_3) - nM_1M_3$$

$$\Sigma(x_2x_3) = \Sigma(X_2X_3) - nM_2M_3$$

$$\Sigma(x_3^2) = \Sigma(X_3^2) - n(M_3^2)$$

Computing the extensions. Inspection of these equations shows that there are eight arithmetic values which must be computed from the original data to work out the values to substitute in equations (31) and (32). These are ΣX_1, ΣX_2, ΣX_3, $\Sigma(X_2^2)$, $\Sigma(X_3^2)$, $\Sigma(X_1X_2)$, $\Sigma(X_1X_3)$, and $\Sigma(X_2X_3)$. The actual work of computing these values for the farm-income data originally presented in Table 35 is shown in Table 45. [The value $\Sigma(X_1^2)$ is not needed in solving equations (31) or (32); but, as it will be needed later, it is also worked out here for convenience in calculation.]

After we have multiplied through all the extensions shown in this table, and added each of the columns, our next step is to compute the values M_2, M_3, and M_1, by dividing the sums of each of the first three columns by the number of cases. The correction values for each of the products is then computed and entered below the value from which it is to be subtracted. Thus the value below the sum of the fourth column, $\Sigma(X_2^2)$, is its correction factor, $n(M_2^2)$. This is equal to $20(13.95)^2$, or 3892.05, which is the value entered. Similarly, the value below the sum of the fifth column, $\Sigma(X_2X_3)$, is its correction factor $n(M_2M_3)$, or 20(8.85)(13.95), which equals 2469.15. All the other correction factors are similarly worked out and entered. Then subtracting each correction factor from the value above it gives the values all ready for equations (31). Thus the value at the foot of column 4 is the value for $\Sigma(x_2^2)$; and so on. When these values are substituted in the appropriate spaces of equations (31), they become

$$\left.\begin{array}{ll}\text{(I)} & \Sigma(x_2^2)b_2+\Sigma(x_2x_3)b_3=\Sigma x_1x_2\\ \text{(II)} & \Sigma(x_2x_3)b_2+\Sigma(x_3^2)b_3=\Sigma x_1x_3\end{array}\right\} = \left\{\begin{array}{l}606.95\,b_2-394.15\,b_3=14.20\\ -394.15\,b_2+676.55\,b_3=1360.60\end{array}\right.$$

Solving the equations. The next step is to solve the two algebraic equations simultaneously to determine the values for b_2 and b_3.

The simplest way to carry this through is by the Doolittle method. The first equation is divided through by the coefficient of b_2, with the sign changed, giving the first derived equation (I′):

$$\text{(I)} \qquad 606.95\, b_2 - 394.15\, b_3 = 14.20$$

$$\text{(I}'\text{)} \qquad -b_2 + 0.64939\, b_3 = -0.02340$$

TABLE 45

COMPUTATION OF VALUES TO DETERMINE MULTIPLE REGRESSION EQUATION TO ESTIMATE ONE VARIABLE FROM TWO OTHERS

	1	2	3	4	5	6	7	8	9
	Number of acres*	Number of cows	Number of dollars income*						
	X_2	X_3	X_1	X_2^2	X_2X_3	X_1X_2	X_3^2	X_1X_3	X_1^2
	6	18	96	36	108	576	324	1,728	9,216
	22	0	83	484	0	1,826	0	0	6,889
	18	14	126	324	252	2,268	196	1,764	15,876
	8	6	61	64	48	488	36	366	3,721
	12	1	59	144	12	708	1	59	3,481
	10	9	90	100	90	900	81	810	8,100
	17	6	82	289	102	1,394	36	492	6,724
	11	12	88	121	132	968	144	1,056	7,744
	16	7	86	256	112	1,376	49	602	7,396
	23	2	76	529	46	1,748	4	152	5,776
	7	17	102	49	119	714	289	1,734	10,404
	12	15	108	144	180	1,296	225	1,620	11,664
	24	7	96	576	168	2,304	49	672	9,216
	16	0	70	256	0	1,120	0	0	4,900
	9	12	80	81	108	720	144	960	6,400
	11	16	113	121	176	1,243	256	1,808	12,769
	22	2	76	484	44	1,672	4	152	5,776
	11	6	74	121	66	814	36	444	5,476
	16	12	98	256	192	1,568	144	1,176	9,604
	8	15	80	64	120	640	225	1,200	6,400
Sums...	279	177	1,744	4,499	2,075	24,343	2,243	16,795	157,532
Means..	13.95	8.85	87.2						
Correction item....................				3,892.05	2,469.15	24,328.80	1,566.45	15,434.40	152,076.80
Corrected sums....................				606.95	− 394.15	14.20	676.55	1,360.60	5,455.20

* In these computations, X_2 and X_1 have been divided by 10. (See Note 3, Appendix 2.)

Then equation (II) is entered, and under it is written equation (I) multiplied by the coefficient of b_3 in equation (I′) (0.64939). The sum of these two equations is then taken, eliminating the values in b_2:

$$\begin{array}{lrrcr}
 & \text{(II)} & -394.15\,b_2 + 676.55\,b_3 & = & 1360.60 \\
(0.64939) & \text{(I)} & +394.15\,b_2 - 255.96\,b_3 & = & 9.22 \\
 & (\Sigma\text{II}) & 420.59\,b_3 & = & 1369.82 \\
 & (\text{II}') & b_3 & = & 3.25690
\end{array}$$

As indicated above, this step gives the value of b_3. This is then substituted in equation (I′) and the value of b_2 determined:

$$-\,b_2 + 0.64939(3.25690) \quad = -\,0.02340$$

$$b_2 = 0.02340 + 2.11500 = 2.13840$$

The values of b_2 and b_3 being thus obtained, the next step is to substitute them, together with the other values required, in equation (32) to work out the value for a:

$$a = M_1 - b_2M_2 - b_3M_3$$

$$a = 87.2 - (2.1384)(13.95) - (3.2569)(8.85)$$

$$= 87.2 - 29.83 - 28.82 = 28.55$$

Estimating X_1 from X_2 and X_3. Having computed the values for a, b_2, and b_3, we can now write out our regression equation (30), with the *best* values, as determined by the mathematical calculation:

$$\left(\frac{X_1}{10}\right) = 28.55 + 2.1384\left(\frac{X_2}{10}\right) + 3.2569X_3$$

$$X_1 = 285.5 + 2.1384X_2 + 32.569X_3$$

Comparing this equation with the last one obtained in Chapter 10, (page 178), we see that the mathematical determination has changed the \$1.33 allowed for the effect of each acre (b_2) to \$2.14, and increased the \$27.88 allowed for the effect of each cow (b_3) to \$32.57. Just what effect this has on the accuracy of the equation as a basis for estimating income from cows and acres may be judged by working out an estimated income for each of the 20 cases according to these last results, and then comparing the estimated values with the original values, just as was done before with the equations worked out by the approximation method. The necessary computation is shown in Table 46.

The operations that have been performed in this table may be mathematically stated as follows:

First, an estimated value of income, X_1, has been worked out by substituting in equation (30) the values for X_2 and X_3 given by each successive observation. Using the symbol X'_1 to represent this estimated value of X_1 it may be defined

$$X'_1 = a + b_2X_2 + b_3X_3 \tag{33}$$

Each estimated income has next been subtracted from the corresponding actual income. With the symbol z used to represent the *residual,* the amount by which the actual value exceeds or falls below the estimated value, it may be defined

$$z = X_1 - X'_1 \tag{34}$$

The residual z has exactly the same meaning when the estimated values of the dependent variable are based upon two or more variables, using multiple correlation, as it had previously when the estimate was based on a single variable, with simple correlation.

The accuracy of the last estimating equation, derived by an exact mathematical process, can now be compared with the accuracy of previous equations, obtained by a cut-and-try process. Computing the standard deviation of the residuals shown in this last table and comparing it with the standard deviations of the residuals worked out in Tables 39 and 41 of Chapter 10, we find the comparison to be:

Standard deviations of residuals using various straight-line equations:

First approximation equation,	$\sigma_z = 90.29$
Second approximation equation,	$\sigma_z = 78.70$
Mathematically determined equation,	$\sigma_z = 70.48$

The equation determined mathematically gives a closer estimate of the actual incomes from which it was derived than do either of the two previous equations. This will always hold true. The mathematically determined equation gives once and for all the estimates of X_1 which will make σ_z the smallest that can be obtained, assuming linear relations. The best that could be done by the approximation method would be to obtain the same conclusions as would be obtained by the other method. The successive steps in Chapter 10 have shown how difficult it is to do this when the several independent variables are correlated with each other, and so tend to vary with one another. The mathematical method for determining the estimating equation, as illustrated in this Chapter (or some alternative form of computation involving the same principle), has therefore been practically universally adopted as the standard way of determining the precise way in which one variable is related to, or may be estimated from, two or more variables related among themselves, if only straight-line relations are to be assumed.

TABLE 46

ACTUAL INCOME AND INCOME ESTIMATED FROM NUMBER OF ACRES AND COWS, ON BASIS OF MATHEMATICALLY DETERMINED RELATIONS

Acres, X_2	Cows, X_3	Computation of estimated incomes			Estimated income, X_1'	Actual income, X_1	Actual minus estimated income, X_1-X_1' z
		Estimated for acres, b_2X_2	Estimated for cows, b_3X_3	Constant, a			
60	18	128	586	286	1,000	960	− 40
220	0	470	....	286	756	830	74
180	14	385	456	286	1,127	1,260	133
80	6	171	195	286	652	610	−42
120	1	257	33	286	576	590	14
100	9	214	293	286	793	900	107
170	6	363	195	286	844	820	−24
110	12	235	391	286	912	880	−32
160	7	342	228	286	856	860	4
230	2	492	65	286	843	760	−83
70	17	150	554	286	990	1,020	30
120	15	257	489	286	1,032	1,080	48
240	7	513	228	286	1,027	960	−67
160	0	342	....	286	628	700	72
90	12	192	391	286	869	800	−69
110	16	235	521	286	1,042	1,130	88
220	2	470	65	286	821	760	−61
110	6	235	195	286	716	740	24
160	12	342	391	286	1,019	980	−39
80	15	171	489	286	946	800	−146

Nomenclature in multiple linear correlation. When the constants of the estimating equation are determined by the exact mathematical process, the equation is called a *multiple regression equation,* and the constants b_2 and b_3, which show, in this case, the average increase in income (X_1) for unit increases in acres (X_2), and cows (X_3), are termed *net regression coefficients*. The constant b_2 is termed "the net regression of X_1 on X_2, holding X_3 constant," and b_3 is termed "the net regression of X_1 on X_3, holding X_2 constant." All that that means for b_2, for example, is "the average change observed in X_1 with unit changes in X_2, determined while simultaneously eliminating from X_1 any variation accompanying (hence temporarily assumed due to) changes in X_3."[2]

In order that the mathematical notation for the net regression coefficients may show quite clearly which independent variables were held constant when a particular coefficient was determined, the subscripts under the b are sometimes more elaborate, showing first the dependent variable, then the independent variable whose effect is stated, then a period followed by the independent variables which were held constant in the process. Thus the b_2 we have been using would be written $b_{12.3}$. The whole regression equation would appear

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \tag{35}$$

This notation serves to distinguish these net regression coefficients from those which would be obtained if additional independent variables were included. Thus if a third independent variable, say X_4, were also considered, the equation would read

$$X_1 = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{36}$$

For still another variable it would be

$$X_1 = a_{1.2345} + b_{12.345}X_2 + b_{13.245}X_3 + b_{14.235}X_4 + b_{15.234}X_5 \tag{37}$$

The notation for a is changed as well as for each of the b's; $a_{1.234}$ will probably be a different value from $a_{1.23}$, just as $b_{12.34}$ is likely to be somewhat different from $b_{12.3}$. This is to be expected; if some other factor, such as the number of men working on each farm, were taken into account as well as the number of acres and the number of cows, the average increase in income per additional acre, with both the number of cows and the number of men held constant, might be quite different from what it would be with only the number of cows held constant. In the last case, any increase in income owing to more men being at work on the larger number of acres would be ascribed to the acres and not to the men, whereas in the former this element would be removed from the increase attributed to the acres.

[2] The term *partial regression coefficient* is used by some authors in place of *net regression coefficient.*

Determining a regression equation for three independent variables. Solely to illustrate the method, we may take the number of men on each of these 20 farms as given in Table 47 and work out an estimating equation considering men as well as acres and cows. (In actual practice, 20 observations are usually too few to determine, with any degree of reliability, the net relation of one variable to 3 independent variables. This problem is used here solely to illustrate the process.)

With the number of men designated as X_4, the unknown constants to be determined are those given in equation (36); $a_{1.234}$, $b_{12.34}$ $b_{13.24}$, and $b_{14.23}$. They can be obtained by the solution of the following set of equations.

$$\left.\begin{array}{llll} \Sigma(x_2^2)b_{12.34} & + \Sigma(x_2x_3)b_{13.24} & + \Sigma(x_2x_4)b_{14.23} & = \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_{12.34} & + \Sigma(x_3^2)b_{13.24} & + \Sigma(x_3x_4)b_{14.23} & = \Sigma(x_1x_3) \\ \Sigma(x_2x_4)b_{12.34} & + \Sigma(x_3x_4)b_{13.24} & + \Sigma(x_4^2)b_{14.23} & = \Sigma(x_1x_4) \end{array}\right\} \qquad (38)$$

$$a_{1.234} = M_1 - b_{12.34}M_2 - b_{13.24}M_3 - b_{14.23}M_4 \qquad (39)$$

Computing the extensions. All except 4 of the arithmetic values for equation (38) which need to be calculated from the original data have been worked out previously. Only the values which involve X_4, and its mean, are additional. The new values needed are therefore M_4, $\Sigma(x_1x_4)$, $\Sigma(x_2x_4)$, $\Sigma(x_3x_4)$, and $\Sigma(x_4^2)$. The computation of these values is shown in Table 47.

All the calculations, including correcting for the means at the end, are carried out just as in Table 45. The figures at the foot of each column provide the remaining values necessary to write out equations (38) in full. For convenience in writing these equations, we shall again use the abridged notation of b_2 for $b_{12.34}$, b_3 for $b_{13.24}$, etc., remembering, however, that b_2 here is a different constant from b_2 previously.

$$\left.\begin{array}{ll} \text{(I)} & \Sigma(x_2^2)b_2+\Sigma(x_2x_3)b_3 \\ & \qquad +\Sigma(x_2x_4)b_4=\Sigma(x_1x_2) \\ \text{(II)} & \Sigma(x_2x_3)b_2+\Sigma(x_3^2)b_3 \\ & \qquad +\Sigma(x_3x_4)b_4=\Sigma(x_1x_3) \\ \text{(III)} & \Sigma(x_2x_4)b_2+\Sigma(x_3x_4)b_3 \\ & \qquad +\Sigma(x_4^2)b_4 \ =\Sigma(x_1x_4) \end{array}\right\} = \left\{\begin{array}{l} 606.95b_2-394.15b_3 \\ \qquad +63.20b_4=14.20 \\ -394.15b_2+676.55b_3 \\ \qquad +11.60b_4=1360.60 \\ 63.20b_2+11.60b_3 \\ \qquad +17.20b_4=193.20 \end{array}\right.$$

Solving the equations. The three equations are now to be solved simultaneously to determine the values for b_2, b_3, and b_4. This can be done by the usual algebraic processes, but the peculiar symmetrical character of the equations, which the attentive reader has probably already noticed, makes it possible to use a much shorter method. Since the saving in clerical labor by the use of this method is quite significant, it will be shown in full.

TABLE 47

COMPUTATION OF ADDITIONAL VALUES TO DETERMINE MULTIPLE REGRESSION EQUATION, ADDING A THIRD INDEPENDENT FACTOR

Item number	Number of acres, X_2*	Number of cows, X_3	Number of men, X_4	Number dollars income, X_1*	X_2X_4	X_3X_4	X_1X_4	X_4^2
1	6	18	2	96	12	36	192	4
2	22	0	3	83	66	0	249	9
3	18	14	4	126	72	56	504	16
4	8	6	1	61	8	6	61	1
5	12	1	1	59	12	1	59	1
6	10	9	1	90	10	9	90	1
7	17	6	3	82	51	18	246	9
8	11	12	2	88	22	24	176	4
9	16	7	2	86	32	14	172	4
10	23	2	3	76	69	6	228	9
11	7	17	2	102	14	34	204	4
12	12	15	3	108	36	45	324	9
13	24	7	4	96	96	28	384	16
14	16	0	2	70	32	0	140	4
15	9	12	1	80	9	12	80	1
16	11	16	3	113	33	48	339	9
17	22	2	2	76	44	4	152	4
18	11	6	1	74	11	6	74	1
19	16	12	2	98	32	24	196	4
20	8	15	2	80	16	30	160	4
Sums.....	279	177	44	1744	677	401	4030	114.00
Means....	13.95	8.85	2.2	87.2				
Correction items........					613.80	389.40	3836.80	96.80
Corrected sums........					63.20	11.60	193.20	17.20

* Coded by dividing by 10.

The first step is to set down the first equation (I) and divide it through by the coefficient of the first term, Σx_2^2, *with the sign changed*, or -606.95 in this case. The resulting derived equation (I′) is set down just below it:

$$\text{(I)} \qquad 606.95b_2 - 394.15b_3 + 63.20b_4 = 14.20$$

$$\text{(I}'\text{)} \qquad -b_2 + 0.64939b_3 - 0.10413b_4 = -0.02340$$

The next step is to set down the second equation (II). The first equation (I) is then multiplied by the coefficient of the *second term* in the derived equation (I′), which is +0.64939 in this case, and the products set down just below equation (II). These two equations are added, giving the sum equation (Σ_2), which cancels out the first term, as shown below. The sum equation is then divided by the coefficient of its first term, with the sign changed, giving the second derived equation (II′). The second portion of the work now appears as follows:

$$\begin{array}{lrcl}
\text{(II)} & -394.15b_2 + 676.55b_3 + 11.60b_4 & = & 1360.60 \\
(0.64939)\ \text{(I)} & 394.15b_2 - 255.96b_3 + 41.04b_4 & = & 9.22 \\
\hline
(\Sigma_2) & 420.59b_3 + 52.64b_4 & = & 1369.82 \\
\text{(II}') & -b_3 - 0.12516b_4 & = & -\ 3.25690
\end{array}$$

The final step in the process of elimination is to write down equation (III), multiply the first equation (I) by the coefficient of the *third* term of the first derived equation (I′), which is −0.10413 in this case, and set the products down below equation (III); multiply the sum equation (Σ_2) by the corresponding coefficient (the second term) from the second derived equation (II′), −0.12516; and set these products down below the previous equation. Equation (III) and the two new equations are then added, giving an equation (Σ_3), from which values in both b_2 and b_3 have been eliminated. This equation is then divided by the coefficient of its first term, with the sign changed, −4.03 in this case, and the resulting new derived equation entered as equation (III′). (A method of checking each step in these computations is shown in Appendix 1, Methods of Computation, page 464.) All the computations to this point are:

$$\begin{array}{r|r|r|lcr}
\text{(I)} & 606.95b_2 & -\ 394.15b_3 & +\ 63.20b_4 & = & 14.20 \\
\text{(I}') & -b_2 & +\ 0.64939b_3 & -\ 0.10413b_4 & =- & 0.02340 \\
\text{(II)} & -394.15b_2 & +\ 676.55b_3 & +\ 11.60b_4 & = & 1360.60 \\
(0.64939)\ \text{(I)} & 394.15b_2 & -\ 255.96b_3 & +\ 41.04b_4 & = & 9.22 \\
\hline
(\Sigma_2) & & 420.59b_3 & +\ 52.64b_4 & = & 1369.82 \\
\text{(II}') & & -\ b_3 & -\ 0.12516b_4 & =- & 3.25690 \\
\text{(III)} & 63.20b_2 & +\ 11.60b_3 & +\ 17.20b_4 & = & 193.20 \\
(-0.10413)\ \text{(I)} & -\ 63.20b_2 & +\ 41.04b_3 & -\ 6.58b_4 & =- & 1.48 \\
(-0.12516)\ (\Sigma_2) & & -\ 52.64b_3 & -\ 6.59b_4 & =- & 171.45 \\
\hline
(\Sigma_3) & & & 4.03b_4 & = & 20.27 \\
\text{(III}') & & & -\ b_4 & =- & 5.02978
\end{array}$$

It is now very easy to compute the values of b_2, b_3, and b_4 from the three derived equations. From equation (III′), $b_4 = 5.02978$.

Substituting this value in equation (II′), which may be transposed to read

$$b_3 = 3.25690 - 0.12516b_4$$

we find

$$b_3 = 3.25690 - (0.12516)(5.02978)$$

$$= 3.25690 - 0.62953 = 2.62737$$

Then, transposing equation (I′), we find

$$b_2 = 0.02340 + 0.64939b_3 - 0.10413b_4,$$

and substituting the values for b_3 and b_4,

$$b_2 = 0.02340 + (1.70619) - (0.52375),$$

we find

$$b_2 = 1.20584$$

The values of b_2, b_3, and b_4, just computed, may next be verified by substituting them in the last equation (III). *Equations* (I) *or* (II) *should not be used for this verification, since they will not provide a complete check.* Equation (III)

$$63.20b_2 + 11.60b_3 + 17.20b_4 = 193.20$$

becomes, when the newly calculated values are substituted,

$$(63.20)(1.20584) + (11.60)(2.62737) + (17.20)(5.02978) = 193.20;$$

this works out to

$$76.21 + 30.48 + 86.51 = 193.20$$

or

$$193.20 = 193.20$$

This proves the accuracy of all the previous work.

The work just summarized is all that is needed to solve these three simultaneous equations. In view of the way the terms cancel out during the second and subsequent steps of the process, the work can be still further simplified by omitting all entries to the left of the solid line which has been drawn in through the last set of entries.

Having calculated the values of the three b's, we can calculate a very readily.

$$\begin{aligned} a &= M_1 - b_2M_2 - b_3M_3 - b_4M_4 \\ &= 87.2 - (1.20584)(13.95) - (2.62737)(8.85) - (5.02978)(2.20) \\ &= 36.06 \end{aligned}$$

The regression equation for the three variables is therefore

$$\left(\frac{X_1}{10}\right) = 36.06 + 1.20584\left(\frac{X_2}{10}\right) + 2.62737X_3 + 5.02978X_4$$

If we clear the fractions, the equation becomes

$$X_1 = 360.60 + 1.20584X_2 + 26.2737X_3 + 50.2978X_4$$

Using this equation, we may work out values of X_1 and of z just as we did previously. (This will be left as an exercise for the student. Is σ_z for the new estimates larger or smaller than for the previous estimates? Why should it be?)

Interpreting net regression coefficients. It should be noted that though the value of 1.20584 for $b_{12.34}$, just determined, compares with the value of 2.13840, for $b_{12.3}$, determined previously, they do not measure exactly the same thing. The coefficient $b_{12.34}$ shows the average increase in income for each acre increase in size of farm, with both the number of *cows* and the number of *men* remaining unchanged. The coefficient $b_{12.3}$ shows the average increase in income for each increase of one acre in size, with the number of *cows* remaining unchanged, but without making any allowance for differences in the number of men. Apparently a considerable portion of the differences in income which on the earlier analysis would have been ascribed to the additional acreage is shown by this more complete analysis really to have been associated with the larger labor force on the greater acreages, rather than to the greater acreages themselves. This result illustrates one property of net regression coefficients in common with all other correlation results. They ascribe to any particular independent variable not only the variation in the dependent variable which is directly due to that independent variable but also the variation which is due to such other independent variables correlated with it as have not been separately considered in the study. In the same way that acres, taken alone, included part of the effect due to cows, the effect of acres eliminating cows still included part of the effect due to men; and even the effect of acres holding constant the effect of both cows and men may still include variation due to other correlated variables, such, for example, as fertility of the land. These considerations illustrate the extreme care which is necessary in examination of the data and the theoretical analysis of the problem before deciding on the variables to be correlated and the caution which must be employed in interpreting the results.

Determining the regression equation for any number of independent variables. The same mathematical principle which has been used to determine the constants for regression equations involving one, two, or three independent variables can be extended to problems involving any number of variables it may be desired to employ.

For four independent variables the equations are:

$$\left.\begin{aligned}
&\Sigma(x_2^2)b_{12.345} + \Sigma(x_2x_3)b_{13.245} + \Sigma(x_2x_4)b_{14.235} \\
&\qquad + \Sigma(x_2x_5)b_{15.234} = \Sigma(x_1x_2) \\
&\Sigma(x_2x_3)b_{12.345} + \Sigma(x_3^2)b_{13.245} + \Sigma(x_3x_4)b_{14.235} \\
&\qquad + \Sigma(x_3x_5)b_{15.234} = \Sigma(x_1x_3) \\
&\Sigma(x_2x_4)b_{12.345} + \Sigma(x_3x_4)b_{13.245} + \Sigma(x_4^2)b_{14.235} \\
&\qquad + \Sigma(x_4x_5)b_{15.234} = \Sigma(x_1x_4) \\
&\Sigma(x_2x_5)b_{12.345} + \Sigma(x_3x_5)b_{13.245} + \Sigma(x_4x_5)b_{14.235} \\
&\qquad + \Sigma(x_5^2)b_{15.234} = \Sigma(x_1x_5)
\end{aligned}\right\} \quad (40)$$

$$a_{1.2345} = M_1 - b_{12.345}M_2 - b_{13.245}M_3 - b_{14.235}M_4 - b_{15.234}M_5 \quad (41)$$

When this set of equations is compared with equation (38) for three independent variables, it is evident that adding the additional variable, X_5, has made it necessary to add the additional equation, in which X_5 appears in each of the product terms, and also to add an additional term to each of the previous equations, the additional term including a product summation [such as $\Sigma(x_2x_5)$ and $\Sigma(x_3x_5)$] in which X_5 appears, and also the net regression coefficient $b_{15.234}$. The equation to compute a has also been extended by adding the term "$- b_{15.234}M_5$." In the same way the equations to be solved to determine the constants for any number of variables can be built up, if it is remembered that for each variable added a new term must be added to each of the previous equations and a new equation must be added, each term added including the new variable in some way.

The products which must be computed for any given set of variables, and the equations which will need to be solved, may be worked out readily by the use of the following scheme:

Write out the required regression equation (in terms of deviations from the mean), as, for example, for six variables:

$$b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 = x_1$$

Multiply each term through by the coefficient of the first unknown (that is, by x_2) and sum. This gives the first of the required equations:

$$\Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 + \Sigma(x_2x_4)b_4 + \Sigma(x_2x_5)b_5 + \Sigma(x_2x_6)b_6 = \Sigma(x_2x_1)$$

Then multiply through by the coefficient of the second unknown (x_3) and sum. The second equation is, therefore,

$$\Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 + \Sigma(x_3x_4)b_4 + \Sigma(x_3x_5)b_5 + \Sigma(x_3x_6)b_6 = \Sigma(x_3x_1)$$

The same process is carried out for the coefficient of each unknown in turn, giving five equations to be solved simultaneously to determine the values for the five unknowns. Setting up these equations may be reduced to a tabular form, as follows:

TABLE 48

FORM FOR WORKING OUT THE EQUATIONS TO DERIVE NET REGRESSION CONSTANTS

Independent variables	Independent variables (in deviations from means)							Dependent variable x_1
	x_2	x_3	x_4	x_5	x_6	x_7	x_8	
x_2	$\Sigma(x_2^2)b_2$	$\Sigma(x_2x_3)b_3$	$\Sigma(x_2x_4)b_4$					$=\Sigma(x_1x_2)$
x_3	$\Sigma(x_2x_3)b_2$	$\Sigma(x_3^2)b_3$	$\Sigma(x_3x_4)b_4$					$=\Sigma(x_1x_3)$
x_4	$\Sigma(x_2x_4)b_2$	$\Sigma(x_3x_4)b_3$	$\Sigma(x_4^2)b_4$					$=\Sigma(x_1x_4)$
x_5	$\Sigma(x_2x_5)b_2$	$\Sigma(x_3x_5)b_3$	$\Sigma(x_4x_5)b_4$					$=\Sigma(x_1x_5)$
x_6	$\Sigma(x_2x_6)b_2$	$\Sigma(x_3x_6)b_3$	$\Sigma(x_4x_6)b_4$					$=\Sigma(x_1x_6)$
x_7	$\Sigma(x_2x_7)b_2$	$\Sigma(x_3x_7)b_3$	$\Sigma(x_4x_7)b_4$					$=\Sigma(x_1x_7)$
x_8	$\Sigma(x_2x_8)b_2$	$\Sigma(x_3x_8)b_3$	$\Sigma(x_4x_8)b_4$					$=\Sigma(x_1x_8)$

The variables to be considered are listed at the head of columns from the left to right, ending with the dependent variable at the right. Then the independent variables are entered down the beginning of the lines at the left in the same order. The cells of the table are then filled by multiplying the variable at the head of the column by the variable at the end of the line. These products indicate the values to be computed (by equations [11] and [15]), to give the arithmetic values for the equations. The "b" terms represent, of course, the net regression coefficients for the particular number of variables concerned; that is, b_2 would be $b_{12.3}$ for two independent variables,

$b_{12.34}$ from three independent variables, and so on. The illustration is carried out to seven independent variables, but the scheme can be extended to as many as it is desired to consider.

The equation to compute a is simply the value of the mean of the dependent variable, minus the product of the mean of each independent variable multiplied by the coefficient for the net regression of the dependent variable on that independent variable.

As a matter of practical procedure, it is seldom that a problem is so complicated or that enough observations are available so that significant results for each variable will be obtained using ten or more variables; and, ordinarily, analyses involving not more than five variables are all that will yield stable results. To illustrate some of the details of the procedure necessary where a large number of variables must be considered, various methods to simplify the necessary calculations in carrying through a problem involving a large number of observations are presented in Methods of Computation, Appendix 1.

Interpreting the multiple regression equation. The same limitations apply in interpreting regression coefficients worked out with the effect of one or more variables held constant as when only two variables are considered. Thus for the data shown in Table 47: there were no observations with more than 18 cows, or 4 men, and none below 60 acres or above 240 acres. For that reason, there is no basis for using the regression equation to estimate income beyond those limits. Furthermore, for the extreme ranges where only a few observations were available—for example, less than 80 acres—the relations could not be expected to hold as well as where there were more observations upon which to base the conclusions. In Chapter 18 a more definite basis for determining the probable accuracy of such estimates is discussed. For the present the caution may be restated, that the results may be expected to hold true only within the range covered by the bulk of the observations upon which they were based.[3]

The meaning of the regression equation

$$X_1 = 360.60 + 1.21X_2 + 26.27X_3 + 50.30X_4$$

may be made clearer, in publishing correlation results, by working out the estimated values for a representative variety of conditions. Such a

[3] Even within the limits of the range of observations there may be combinations of values of independent variables which are not represented by the data, either exactly or even approximately. Estimates for such combinations will have less reliability than for those combinations which are represented. For a fuller discussion of this source of unreliability, see Chapter 19.

statement of the conclusions covered by the previous regression equation would be as follows:

TABLE 49

AVERAGE INCOME ON FARMS WITH VARYING NUMBERS OF ACRES, COWS, AND MEN
(As indicated by correlation analysis)

Labor force	100 acres			160 acres		
	0 cows	8 cows	16 cows	0 cows	8 cows	16 cows
	Dollars	*Dollars*	*Dollars*	*Dollars*	*Dollars*	*Dollars*
1 man......	532	742	952	*	*	*
2 men......	*	792	1,003	655	865	*
3 men......	*	*	1,053	705	915	1,125

* Omitted because of absence of observations representing this combination of factors.

It should be noted in Table 49 that, according to these results, increasing the number of men from 1 to 2, or from 2 to 3, will add $50 to income, no matter whether the farm has 100 acres and 8 cows, or 160 acres and 16 cows. Similarly, adding 8 more cows is indicated as having the same effect on income, no matter how large the farm is or how many men are employed. But that this conclusion has been reached is no proof that it is really true of the universe represented by the original data. Instead, such a conclusion is inherent in the linear equation (35, 36, or 37) which has been used. That equation necessarily assumes that an increase of one unit in any one independent variable will always be accompanied by an equal change in the dependent variable. Only insofar as the actual facts agree with that assumption can they be represented by a linear equation. Subsequent chapters (particularly 14 and 21) take up methods of analysis which may be employed when this type of relation is not true, and the linear equation is therefore unable to express the facts adequately.

Net regression coefficients, computed from a sample, may vary more or less widely from the true values for the universe from which that sample is drawn. Tests to indicate the reliability of such sample results are given in Chapter 18. They should always be calculated and considered before generalizing from such sample results.

Summary. This chapter has presented mathematical methods for determining the constants of a linear regression equation, so that changes in one variable may be estimated from changes in two or more independent variables. Equations so determined afford a more exact basis for making such estimates than do linear equations obtained by any other method. Furthermore, the multiple regression equation serves to sum up all the evidence of a large number of observations in a single statement which expresses in condensed form the extent to which differences in the dependent variable tend to be associated with differences in each of the other variables, as shown by the sample.

CHAPTER 13

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION FOR LINEAR MULTIPLE CORRELATION

Standard error of estimate. After working out equations by which values of one variable may be estimated from those for two or more independent variables, it is frequently desirable to have some measure of how closely such estimates agree with the actual values and of how closely the variation in the dependent variable is associated with the variation in the several independent variables. Attention has been called in the preceding chapters to the computation of the residuals, z, when the value of a variable is estimated from that of several others. Where the estimate is based on several independent variables the standard deviation of these residuals serves as a measure of the closeness with which the original values may be estimated or reproduced just as well as where the estimate is based on a single variable. Continuing the same terminology as before, this standard deviation is still called the "standard error of estimate." Thus for the regression equation for estimating income from known numbers of acres, cows, and men, the standard error of estimate is designated $S_{1.234}$. The subscripts "$_{1.234}$" indicate that that is the standard error for variable X_1 when estimated from the independent variables X_2, X_3, and X_4.

Where the size of the sample is small in proportion to the number of variables involved, the standard deviation of the residuals for the cases included in the sample tends to have a downward bias. That is, it tends to be smaller than the standard error which would be observed if the same constant were computed from large samples drawn from the same universe.

For that reason it is necessary to adjust the observed standard deviation of the residuals, σ_z, before it will give an unbiased estimate of the value of the standard error of estimate in the universe. This adjustment is:

$$\bar{S}^2_{1.234} = \frac{n\sigma^2_{z_{1.234}}}{n - m} \qquad (42)$$

where n = number of sets of observations in the sample,
m = number of constants in the regression equation, including a and the b's.

(Where the adjusted value for $\bar{S}^2_{1.234}$ exceeds the value of σ_1^2, the latter value should be used for the standard error.)

The standard errors for the equations obtained when one, two, and three independent variables were considered in the farm-income study in Chapter 12 may be summarized as follows:

Independent variables	Observed σ_z	n	m	Adjusted standard error
X_2	165.15*	20	2	$\bar{S}_{1.2} = 165.15$
X_2, X_3	70.48	20	3	$\bar{S}_{1.23} = 76.45$
X_2, X_3, X_4	66.77	20	4	$\bar{S}_{1.234} = 74.65$

* This value has not been shown previously. It is calculated from the data of Chapter 12.

(In this case the correlation between X_1 and X_2 is practically zero, so $\sigma_z = \sigma_1$. Under the rule given above, $\bar{S}_{1.2} = \sigma_1$.) The values tabulated in the last column illustrate the increase in the reliability of estimate as additional variables are taken into account.

So far, the standard errors of estimate (except for simple or two-variable correlation) have been determined by actually working out all the estimated values, subtracting to get the individual residuals, z, and then determining their standard deviation. For linear multiple regression equations, however, a much simpler process can be used. To compute the standard deviation of the residuals by this process, all that is required in addition to the values which have been used in computing the b's is the value, $\Sigma(x_1^2)$. The formula is as follows:

$$\bar{S}^2_{1.234\ldots n} = \frac{\left\{\begin{array}{l}\Sigma(x_1^2) - [b_{12.34\ldots n}(\Sigma x_1x_2) + b_{13.24\ldots n}(\Sigma x_1x_3) \\ \qquad + \ldots + b_{1n.23\ldots(n-1)}(\Sigma x_1x_n)]\end{array}\right\}}{n - m} \qquad (43)$$

Substituting the values for the regression equation computed with two independent variables, pages 193 and 194, the equation becomes

$$\bar{S}^2_{1.23} = \frac{\Sigma(x_1^2) - [b_{12.3}(\Sigma x_1x_2) + b_{13.2}(\Sigma x_1x_3)]}{n - 3}$$

In terms of coded values for X_1,

$$\frac{\bar{S}^2_{1.23}}{10^2} = \frac{5{,}455.20 - (2.1384)(14.20) - (3.2569)(1{,}360.60)}{20 - 3}$$

$$\frac{\bar{S}_{1.23}}{10} = \sqrt{\frac{993.50}{17}} = 7.645; \ \bar{S}_{1.23} = 76.45$$

The result is seen to be identical with the value computed (after adjustment) by the lengthy process illustrated in Table 46, on page 196, of working out all the individual estimates, computing their standard deviation, and then adjusting by equation (42).

Multiple correlation. The standard error of estimate for a multiple regression equation, just as with simple correlation, measures the *closeness* with which the estimated values agree with the original values. The standard error, however, offers no measure of the *proportion* of the variation in the dependent factor which can be explained by, or is associated with, variation in the independent factor or factors. For example, in one area the farm income might be twice as variable as in another. If two or three independent factors such as those discussed came as near accounting for all the variation in incomes in one area as in the other, the standard errors of estimate would be the same in both cases. There was originally more variance in income in the one case than in the other; therefore with the same amount left unaccounted for the independent factors would have been associated with a larger proportion of the original variance, in the case where it was largest to begin with, and would have been relatively more important in that case. In simple correlation, the *relative* importance of the independent factor was measured by the ratio of the standard deviation of the estimated values to the standard deviation of the actual values, and the name *coefficient of correlation* was given to this ratio. In exactly similar manner, when the estimates are based on several variables, instead of on one, the relative importance of all those variables combined may be measured by dividing the standard deviation of the estimated values by that of the original values. This ratio is named the *coefficient of multiple correlation,* since it measures the combined importance of the several independent factors as a means of explaining the differences in the dependent factor.

If we use $X_{1(234)}$ to designate the estimates of X_1 made from variables X_2, X_3, and X_4, and use $R_{1.234}$, to represent the unadjusted *coefficient of multiple correlation,* the coefficient may be defined:

$$X_{1(234)} = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \qquad (44)$$

$$R_{1.234} = \frac{\sigma_{1(234)}}{\sigma_1} \qquad (45)$$

The same short formula which has been shown for computing the standard error of estimate may be employed to facilitate the computa·

tion of the coefficient of multiple correlation, using only values already involved in equation (43). The equation for computing the coefficient of correlation by this method is: [1]

$$R^2_{1.234\,\ldots\,n} = \frac{\left\{\begin{array}{c} b_{12.34\,\ldots\,n}(\Sigma x_1x_2) + b_{13.24\,\ldots\,n}(\Sigma x_1x_3) \\ + \ldots + b_{1n.23\,\ldots\,(n-1)}(\Sigma x_1x_n) \end{array}\right\}}{\Sigma(x_1^2)} \tag{46}$$

There is a tendency for the multiple correlation shown by the sample to be in excess of the correlation existing in the universe from which the sample was drawn, especially where the number of observations is small, or the number of variables large. For that reason the coefficient $R_{1.23\,\ldots\,n}$, computed as shown in equation (46), has to be adjusted before it will give $\bar{R}_{1.23\,\ldots\,n}$, the unbiased estimate of the correlation most probably existing in the whole universe. The adjustment is:

$$\bar{R}^2_{1.234\,\ldots\,n} = 1 - (1 - R^2_{1.234\,\ldots\,n})\left(\frac{n-1}{n-m}\right) \tag{47}$$

m and n have the same meaning for this equation as in equation (42).

If the value for $\bar{R}^2$ comes out a minus quantity, use 0 for $\bar{R}^2$.

The square of the coefficient of multiple correlation, $\bar{R}^2$, may be termed the *coefficient of multiple determination.*

The same relations hold between the coefficient of multiple correlation and the standard error of estimate in the case of multiple correlation as in the case of simple correlation. For that reason, one of these measures may be computed from the other, whichever is determined first, according to the following equations:

$$\bar{R}^2_{1.234\,\ldots\,n} = 1 - \left(\frac{\bar{S}^2_{1.234\,\ldots\,n}}{\sigma_1^2}\right)\left(\frac{n-1}{n}\right) \tag{48}$$

$$\bar{S}^2_{1.234\,\ldots\,n} = \sigma_1^2\,(1 - \bar{R}^2_{1.234\,\ldots\,n})\left(\frac{n}{n-1}\right) \tag{49}$$

Using equation (48) to compute the values of $\bar{R}$ from the values of $\bar{S}$ previously computed, the multiple coefficients for the three regression equations previously worked out may be stated in the following different ways:

[1] This may be computed most conveniently by following the form shown on pages 467 and 469.

Dependent variable	Independent variable(s)	$\bar{S}$ Standard error of estimate	$\bar{R}$ Coefficient of multiple correlation	$\bar{R}^2$ Coefficient of multiple determination
X_1(income)	X_2(acres)	165.15	0*	0
X_1(income)	X_2(acres); X_3(cows)	76.45	0.892	0.796
X_1(income)	X_2(acres); X_3(cows); X_4(men)	74.65	0.898	0.806

* The value shown here should be that of $\bar{r}_{12}$. In this case it happens to be zero.

It is evident that the correlation increases as the standard error decreases. Here the residual variation in each case is being compared with the same original standard deviation, so that that necessarily follows. Where different studies are being compared, however, such as two samples with widely different original deviations in the dependent variable, the standard error of estimate would not necessarily decrease as the correlation increased, since the former is an *absolute* measure whereas the latter is a *relative* measure.[2]

It is evident from the figures just shown that the coefficient of multiple correlation, if incorrectly interpreted, makes the relationship seem closer than does the coefficient of multiple determination ($\bar{R}^2$). It cannot be demonstrated that the coefficient of multiple determination will measure in all cases that proportion of the variance in the dependent factor which is associated with the independent factors. Yet it is sufficiently true so that, if such a statement is to be made as "seventy-five per cent of the variance in income was associated with (or related to) variances in numbers of acres farmed, or cows milked, and men hired," it is more accurate to use the coefficient of multiple determination than to use the coefficient of multiple correlation. The latter would overstate the case. This principle holds true both for simple correlation ($\bar{r}$) and multiple correlation ($\bar{R}$): the square of the coefficient indicates the proportion of the variance in the dependent variables which has been mathematically accounted for; whereas one minus the square of the coefficient indicates the proportion which has not been accounted for.[3]

[2] This point is of considerable significance in certain types of economic problems, particularly in time-series analysis. For example, taking the first differences of a series of values frequently tends to make the deviations much larger than by taking deviations from trend. A study which gives a higher coefficient of correlation for first differences than for deviations from trend may still yield the less accurate estimate, as measured by the standard error of estimate.

[3] See Note 7, Appendix 2.

The coefficient of multiple correlation, $R_{1.234\ldots n}$, may also be defined as the simple correlation between the actual X_1 values and the $X_{1(234)}$ values estimated from the several independent factors. This interpretation illustrates the way it sums up the combined relation of the dependent variable to the several independent variables.

(For the most convenient methods of calculating the various measures discussed in this chapter, see Appendix 1, pages 459 to 478.)

Measuring the separate effect of individual variables. In addition to the measures of the importance of all of the independent variables combined, it is sometimes desirable to have measures of the importance of each of the individual variables taken separately, while simultaneously allowing for the variation associated with remaining independent variables. There are two different types of these measures: *the coefficient of partial correlation* and the *"beta" coefficient.*[4]

Partial correlation. Coefficients of *partial correlation* serve to determine the correlation between the dependent factor and each of the several independent factors, while eliminating any (linear) tendency of the remaining independent factors to obscure the relation. Thus in the problem where income was correlated with numbers of acres, cows, and men, the partial correlation of income with acres, while holding constant cows and men, indicates what the average correlation would probably be between acres and income in samples of farms in which all the farms in each sample had the same number of cows and the same number of men.

If the data we have just been discussing were classified into groups which had the same number of cows and men in each group, and the correlation of the income and acres for the farms in each group was calculated separately, that would give a series of values for the correlation between acres and income for series of groups in each of which there was no variation in cows or men. If a weighted average of this

[4] Discussion of the coefficient of part correlation (which was covered on pages 182 and 183 of the first edition of this book) has been dropped from this edition. It is defined by the formula

$${}_{12}\bar{r}^2_{34} = \frac{b^2_{12.34}\sigma^2_2}{b^2_{12.34}\sigma^2_2 + \sigma^2_1(1 - \bar{R}^2_{1.234})} \qquad (51)$$

Little practical use has been found for this coefficient, except that it does provide a maximum value for the coefficient of partial correlation. Although its formal interpretation was correct as given previously, it seems to provide insufficient information to justify its detailed presentation. However, its derivation is still given in Note 9, Appendix 2, as before.

series of correlations was then calculated,[5] it would correspond to the partial correlation of income with acres, while holding cows and men constant ($r_{12.34}$). A similar interpretation can be made for the other two partial correlation coefficients. Even in problems (such as the present one) where the number of observations is not sufficient to permit of many such subgroups being formed, the partial correlation coefficient indicates about what such an average correlation in selected subgroups would be, if computed from a larger sample drawn from the same universe.

Any group of independent variables may serve to explain some, but not all, of the variation in a dependent variable. If an additional independent variable is added, it may account for part of the variation left unexplained by the factors previously considered. The coefficient of partial correlation may be defined as a measure of the extent to which that part of the variation in the dependent variable which was *not* explained by the other independent factors can be explained by the addition of the new factor. For example, in the farm-income problem, considering only acres and cows, the correlation was $\bar{R}_{1.23} = 0.892$. When acres, cows, and men were considered, the correlation was $R_{1.234} = 0.898$. Squaring both values shows that, whereas the two variables explain 79.6 per cent of the variance in income, the three variables explain 80.6 per cent. Whereas 20.4 per cent of the variance is left to be explained when the two variables are considered, only 19.4 per cent is left to be explained when three are considered. Adding the additional variable has increased the variance which can be explained by the difference between these two figures, or 1.0 per cent (20.4 − 19.4 per cent). If the importance of this increase is determined by comparing it to the variance left unexplained before the new variable was added, we find that $\frac{1.0}{20.4}$, or 4.90 per cent of the variance left unexplained by acres and cows, has now been found to have been associated with differences in numbers of men. Taking its square root gives the coefficient of partial correlation, 0.221.

The coefficient is designated $\bar{r}_{14.23}$, since it shows the partial correlation between X_1 and X_4, after X_2 and X_3 had been taken into account. As is indicated in the discussion, it may be computed by the formula[6]

$$\bar{r}^2_{14.23} = \frac{(1 - \bar{R}^2_{1.23}) - (1 - \bar{R}^2_{1.234})}{1 - \bar{R}^2_{1.23}}$$

[5] The calculation of the average of a series of correlation coefficients would involve the use of Fisher's z-transformation.

[6] This is different from the formula customarily given. See Note 7, Appendix 2, for its derivation.

For purposes of computation, this formula may be simplified to

$$\bar{r}^2_{14.23} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.23}} \tag{50}$$

If it is desired to compute coefficients of partial correlation for the other independent variables, acres and cows, the corresponding formulas are[7]

$$\bar{r}^2_{13.24} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.24}}$$

$$\bar{r}^2_{12.34} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.34}}$$

It should be noticed that, although the numerator of the fraction is the same in each case, the denominator is different. This is a peculiarity of coefficients of partial correlation—they measure the importance of each of the several variables by determining how much it reduces the variation *after all the other variables except it are taken into account.*

If we work out the new multiple correlations necessary,[8] $\bar{R}_{1.24}$ and $\bar{R}_{1.34}$, and substitute them in the equations given just above, the whole set of coefficients of partial correlation and partial determination for the farm-income problem works out as follows:

$$\bar{r}^2_{13.24} = 1 - \frac{1 - 0.806}{1 - 0.458} = 0.642$$

$$\bar{r}^2_{12.34} = 1 - \frac{1 - 0.806}{1 - 0.791} = 0.072$$

[7] Equation (50) and these following equations will give values for the partial regression coefficients, which will differ slightly from those computed by the classical equations used by Yule, and then adjusted by equation (47). In view of the definition of the adjusted partial correlation coefficient just given, however, it is believed that this method of computation directly from the adjusted values, $\bar{R}_{1.234}$ and $\bar{R}_{1.23}$, is sufficiently accurate for all practical purposes.

[8] The two new coefficients of multiple correlation are obtained by rearranging the arithmetic values previously computed so as to give the necessary regression coefficients, and then determining the value of $\bar{R}$ by equations (46) and (47). The two new sets of equations are:

To determine $R_{1.24}$

$$(\Sigma x_2^2)b_{12.4} + (\Sigma x_2x_4)b_{14.2} = (\Sigma x_1x_2)$$
$$(\Sigma x_2x_4)b_{12.4} + (\Sigma x_4^2)b_{14.2} = (\Sigma x_1x_4)$$

Similarly for $R_{1.34}$

$$(\Sigma x_3^2)b_{13.4} + (\Sigma x_3x_4)b_{14.3} = (\Sigma x_1x_3)$$
$$(\Sigma x_3x_4)b_{13.4} + (\Sigma x_4^2)b_{14.3} = (\Sigma x_1x_4)$$

Relative Importance of Individual Factors Affecting Income, as Indicated by Coefficients of Partial Correlation

Factors already considered	Factor added	Coefficient of partial correlation ($\bar{r}_{12.34}$, etc.)	Reduction in unexplained variance ($\bar{r}^2_{12.34}$, etc.)
Cows (X_3), men (X_4)	Acres (X_2)	0.27	0.072
Acres (X_2), men (X_4)	Cows (X_3)	0.80	0.642
Acres (X_2), cows (X_3)	Men (X_4)	0.22	0.049

When income was correlated with acres alone, there was no correlation at all. (Before adjusting for the number of observations, $r_{12} = 0.01$.) Yet the partial correlation of income with acres, while holding constant the variation associated with cows and men, has just been seen to be 0.27. Although this is not high, it is certainly more than no correlation at all. Furthermore, even though the correlation of income with cows alone is 0.64, the correlation with both acres and cows is 0.89.

On the surface of the data there appears to be no relation between acres and income, since the positive relation of acres to income is hidden. Acres are negatively correlated with cows to a sufficient extent so that the decreased income with decreased number of cows offsets the increases with more acres. Only when the number of cows is allowed for can the influence of acres be seen.

It is evident that a mere surface examination of a set of data cannot reveal which independent factors are important and which are unimportant. A variable which shows no correlation with the dependent variable may yet show significant correlation after the relation to other variables has been allowed for.

Investigators sometimes think they are doing "research" when they study the relation of a given variable, say the price of a commodity, to a number of other factors, discard all those factors that show no correlation with price, and select out for further study by multiple correlation the factors that show the highest simple correlation with the price. As the preceding discussion shows, that procedure may result in discarding factors which would show a truly important relation to price after the effect of other associated factors had been allowed for. A careful, logical examination of the problem, the selection of the factors to be considered on the basis of these qualitative considerations, and then preliminary examination of all the intercorrelations among the selected independent factors will provide more trustworthy results. (See Chapter 24 for a more detailed discussion of the places of qualitative and quantitative analysis in such studies.)

The test whether a given independent variable may really be related to the dependent variable, even if it shows no apparent correlation, is whether that independent variable is correlated with other independent variables, which in turn are correlated with the dependent. Thus in the example just discussed, although acres showed no correlation with income, they did show significant correlation with cows. If acres had had no correlation with either income, cows, or men, it would have been impossible for acres to have correlation with income even after the relation to cows and men was allowed for.

"Beta" coefficients. The importance of individual variables may also be compared by their net regression coefficients. The size of the regression coefficients, however, varies with the units in which each variable is stated. They may be made more comparable by expressing each variable in terms of its own standard deviation, using the "beta" coefficients mentioned in Chapter 9. In terms of betas, the regression equation for four variables would be

$$\frac{X_1}{\sigma_1} = \beta_{12.34}\frac{X_2}{\sigma_2} + \beta_{13.24}\frac{X_3}{\sigma_3} + \beta_{14.23}\frac{X_4}{\sigma_4} + a'$$

Hence the partial betas may be defined

$$\beta_{12.34} = b_{12.34}\frac{\sigma_2}{\sigma_1} \tag{52}$$

For the problem we have been considering, the betas may be calculated very readily:

$$\beta_{12.34} = b_{12.34}\frac{\sigma_2}{\sigma_1} = 1.2058\left(\frac{5.51}{16.52}\right) = 0.402$$

$$\beta_{13.24} = b_{13.24}\frac{\sigma_3}{\sigma_1} = 2.6274\left(\frac{5.82}{16.52}\right) = 0.926$$

$$\beta_{14.23} = b_{14.23}\frac{\sigma_4}{\sigma_1} = 5.0298\left(\frac{0.927}{16.52}\right) = 0.282$$

If the relative importance of each of the different factors, as judged by the two different types of individual measurement, is compared, the relations are:

Relative Importance of Individual Factors Affecting Income, as Indicated by Two Different Coefficients

Independent factor	Factors held constant	Coefficients of partial correlation ($\bar{r}_{12.34}$)	Beta coefficients $\beta_{12.34}$
Acres (X_2)..............	Cows (X_3), men (X_4)	0.27	0.402
Cows (X_3)..............	Acres (X_2), men (X_4)	0.80	0.926
Men (X_4)...............	Acres (X_2), cows (X_3)	0.22	0.282

It is evident from this comparison that, although the exact values differ for the two sets of measures, the rank of the three variables in order of importance is the same and the relative sizes are comparable.[9] This does not always hold true, owing to the mathematical differences in the meaning of the two sets.

Besides the coefficients which have been discussed, which measure either the total relative importance of all the independent variables or the importance of each one separately, it is sometimes desirable to measure the correlation between one variable and a group of others, after eliminating from the dependent variable that part of its variation imputed (by the analysis) to a single one of the independent variables. The problem may be stated as follows:

Where $R_{1.234}$ measures the relation between X_1 and X_2, X_3, X_4, according to the regression equation (36), the problem stated is to determine the correlation between $(X_1 - b_{12.34}X_2)$ and the two remaining independent variables, according to the equation

$$(X_1 - b_{12.34}X_2) = a_{1.234} + b_{13.24}X_3 + b_{14.23}X_4$$

This could be determined by actually carrying out the operations indicated, but it can be much more readily computed by use of the formula [10]

$$\left.\begin{array}{c}\text{Multiple correlation}\\ \text{squared of}\\ (X_1 - b_{12.34}X_2)\\ \text{with } X_3 \text{ and } X_4\end{array}\right\} = 1 - \frac{\sigma_1^2(1 - R_{1.234}^2)}{\sigma_1^2 - 2b_{12.34}(\Sigma x_1x_2/n) + b_{12.34}^2\sigma_2^2} \qquad (53)$$

[9] One other type of measure of individual importance, the coefficient of separate determination, is discussed in Note 11, Appendix 2.

[10] See Note 12, Appendix 2, for derivation of this equation.

An illustration of the type of problem to which this method may be applied can be drawn from the field of price analysis. If X_2 in the case illustrated above were an index of price level, X_1 the price of some commodity, and X_3 and X_4 other factors affecting price, such as production and storage stocks, it might be desired to determine not only how closely the price of the commodity was related to all the factors, including the price index, but also how closely it was related to the remaining factors after the variations in price found to be associated with changes in price level were removed from it. Formula (53) would enable this determination to be made.

Reliability of results from a sample. All the coefficients presented in this chapter are subject to fluctuations of sampling just as are simpler coefficients. A later chapter (Chapter 18) discusses the extent of these fluctuations with various sizes of samples and gives methods of estimating how far the coefficients from a given random sample may miss the true values of the coefficient in the universe from which the sample was drawn.

Summary. This chapter has shown that the accuracy of a regression equation for estimating one variable from two or more others may be measured by the standard error of estimate. The extent to which variation in the dependent variable is associated with the variation in the several independent variables may be measured by the coefficient of multiple correlation, or, with respect to variance, by the coefficient of multiple determination. The relative importance of each of the independent variables may be measured (*a*) by the coefficient of partial correlation, relative to the variation remaining after the effects of the other variables have first been removed, or (*b*) by the beta coefficients, which reduce the net regression coefficients to a comparable basis. Finally, a method is provided for measuring the proportion of the variation in the dependent variable which is explainable by a group of independent variables, after eliminating from the dependent variable that portion of its variability which has been found to be associated with another independent factor.

CHAPTER 14

DETERMINING THE WAY ONE VARIABLE CHANGES WHEN TWO OR MORE OTHER VARIABLES CHANGE: (4) USING CURVILINEAR REGRESSIONS

The discussion of multiple correlation to this point has been limited to linear relationships—relations where the change in the dependent variable accompanying changes in each independent variable was assumed to be of exactly the same amount, no matter how large or how small the independent variable became. Thus in the farm income example, it was assumed that each additional cow would be accompanied by the same increase in income, no matter whether it was the first, the tenth, or the thirtieth. Similarly, each additional acre in crops or each additional man employed was assumed to be accompanied by an identical contribution to the income, no matter how large or how small the business already was. It is quite evident that such an analysis makes no provision for there being an optimum size of operation for given circumstances or for differences in the contributions of different numbers of units. In this particular case, it assumes that there is no such thing as the principle of diminishing returns. Such an analysis might therefore fail entirely to reveal the proper size of productive unit, or the number of each of the several elements to be employed to yield maximum returns.

In many other types of problems for which multiple correlation analysis might be used, limitation of the analysis to linear relations would seriously restrict its value or prevent its use altogether. In dealing with the effect of weather upon crop yields, several variable weather factors are usually concerned. There may be an optimum point for growth, with respect to both temperature and precipitation, with values either above or below the optimum tending to produce lower yields. Linear regressions are obviously unfitted to express such relations. In problems such as these, and many others which might be enumerated, determination of the exact curvilinear relation between independent and dependent variable, while simultaneously eliminating the effect of other factors which also affect the dependent variable, is the most important feature in the investigation. Unless the curve itself can be determined, the other conclusions are of little value.

The problem in its simplest outlines may be stated as follows: Given a series of paired observations of the values of a dependent variable X_1 and two or more independent variables X_2, X_3, X_4, etc., required to find the change in X_1 accompanying the changes in X_2, X_3, and X_4, in turn, while holding the remaining independent factors constant, so that for any given values of X_2, X_3, and X_4, etc., values may be estimated for X_1, according to the regression equation

$$X_1 = a' + f_2(X_2) + f_3(X_3) + f_4(X_4) + \text{etc.} \tag{54}$$

The expression "$f_2(X_2)$" is used here simply as a perfectly general term meaning any regular change in X_1 with given changes in X_2, whether describable by a straight line or a curve. The equation is read "X_1 is a *function* of X_2 plus a *function* of X_3," etc.

The several partial (or "net") regression curves may be determined either by the use of definite mathematical expressions, one for each independent variable, with the constants all determined simultaneously just as in linear multiple correlation; or by a method known as "successive graphic approximation," which involves no prior assumptions as to the shapes of the curves.

Multiple Regression Curves Mathematically Determined

In using definite mathematical functions, it is necessary to express the curvilinear relations by simple mathematical curves of some type, so that the constants for the curves may be determined by methods similar to those already presented. If simple parabolas were used, involving only the first and second powers of each independent variable, equation (54) could be expressed

$$X_1 = a + b_2X_2 + b_{2'}(X_2^2) + b_3X_3 + b_{3'}(X_3^2) + b_4X_4 + b_{4'}(X_4^2) \tag{55}$$

However, this type of parabola is not very flexible, and in practice it fits but very few actual curves. If the more flexible cubic parabola were employed, involving the first, second, and third powers of each independent variable, the equation would be

$$\begin{aligned} X_1 = a + b_2(X_2) + b_{2'}(X_2^2) + b_{2''}(X_2^3) + b_3(X_3) + b_{3'}(X_3^2) \\ + b_{3''}(X_3^3) + b_4(X_4) + b_{4'}(X_4^2) + b_{4''}(X_4^3) \end{aligned} \tag{56}$$

This last equation for three independent variables involves 10 constants and increases the error in their determination accordingly, and the clerical labor of dealing with the squared and cubed values would be large (unless they were coded). Even then, it offers no guarantee that the curves for each function would truly represent the real relationship. The curves corresponding to the three functions in equation (54) would be:

$$f_2(X_2) = b_2X_2 + b_{2'}(X_2^2) + b_{2''}(X_2^3)$$

$$f_3(X_3) = b_3X_3 + b_3'(X_3^2) + b_{3''}(X_3^3)$$

$$f_4(X_4) = b_4X_4 + b_{4'}(X_4^2) + b_{4''}(X_4^3)$$

Whether or not these curves would actually be a good fit to the true functions could not be told beforehand, for the problem is not to find the curves expressing the relation between X_1 and each of the other variables according to the apparent relation but according to the underlying relation, which may become apparent only when the differences in X_1 associated with differences in the other factors have been eliminated. Each of the independent factors may be correlated with the other independent factors to a greater or less degree. Thus in the problem which follows, correlating X_2 with X_3, $r = +0.07$; X_2 with X_4, 0.00; and X_3 with X_4, -0.67. The last correlation is sufficient to tend to obscure the relations. When we make a dot chart showing the apparent relation between X_1 and X_3, we cannot tell how much of the observed differences in X_1 are due to the differences in X_4 associated with the differences in X_3. For that reason we cannot be sure what type of curve would truly represent the differences in X_1 with differences in X_3 after allowances had been made for these other factors. Even though the apparent relation might indicate that a straight line or some type of parabola would fit, there would be no guarantee that this would truly represent the net functional relationship. The successive approximation method, which makes no rigid assumption as to the type of curve, is therefore to be preferred.[1]

Multiple Regression Curves by Successive Approximations

The general method of determining partial regression curves by the successive approximation method may be outlined as follows:

The conditions to be imposed on the shape of each curve, in view of the logical nature of the relations, are first thought through and stated. This procedure, for each curve, is similar to that described on page 109 of Chapter 6.

[1] The determination of multiple regression curves by fitting definite mathematical equations is dealt with at more length in Chapter 22, on pages 396 to 401.

The linear partial regressions are next computed. Then the dependent variable is adjusted for the deviations from the means of all independent variables except one, and a correlation chart, or dot chart, is constructed between these adjusted values and that independent variable. This provides the basis for drawing in the first approximation curve for the net regression of the dependent variable on that independent variable, within the limitations of the conditions stated. The dependent variable is then corrected for all except the next independent variable, the corrected values plotted against the values of that variable, and the first approximation curve determined with respect to that variable. This process is carried out for each independent variable in turn, yielding a complete set of first approximations to the net regression curves. These curves are then used as a basis for correcting the dependent factor for the approximate curvilinear effect of all independent variables except one, leaving out each in turn; and second approximation curves are determined by plotting these corrected values against the values of each independent variable in turn. New corrections are made from these curves, and the process is continued until no further change in the several regression curves is indicated.

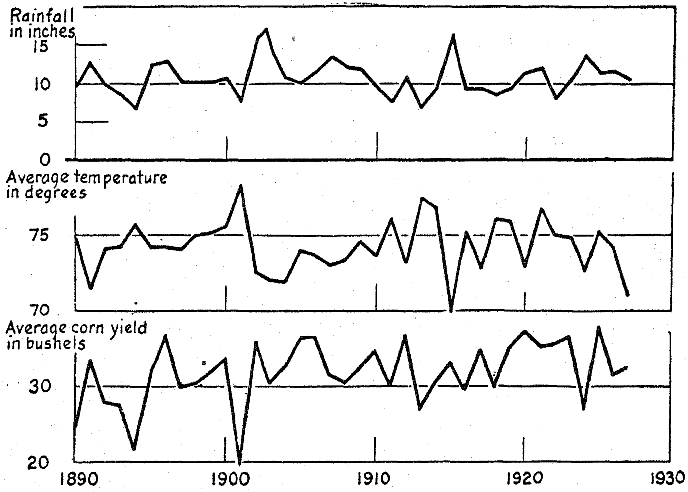


FIG. 32. Rainfall, temperature, and corn yields in the Corn Belt, 1890 to 1927.

The process of determining net curvilinear regressions by the successive graphic approximation method may be illustrated by the data shown in Table 50. These data show, for a period of 38 years, the average rainfall during June, July, and August, for nine weather stations scattered through the Corn Belt. This precipitation has been designated as variable X_3. The average temperature during the same months, at the same stations, has been designated as X_4. The average yield of corn per acre, in the six leading Corn Belt states, is shown as X_1—the variable whose fluctuations are to be explained, so far as possible, by the other factors.

It is evident from the table that there has been a marked upward trend in corn yield during this period, although there has not been a similar trend in rainfall or temperature. Plotting each one of the three factors, X_3, X_4, and X_1 as shown in Figure 32, we notice, however, that there have been marked though irregular long-time cycles in rainfall and temperature during the period. To a certain extent the upward swing in yields has agreed with the high point of the rainfall cycles, particularly from 1919 to 1921. It is not safe, therefore, to fit a long-time trend to yield and to assume that in removing that trend we are merely taking out the effects of such factors as better varieties, improved methods of tillage, or concentration of acreage in the more fertile sections. Since there is some association between rainfall and time, at least over considerable periods, in eliminating all the variation associated with time we might be eliminating a part of the variation which really reflected differences in rainfall. Accordingly we may make time itself one of the factors in the multiple correlation and ascribe to time only that part of the long-time change in yields which is not associated with differences in rainfall or in temperature. Each year, numbered from 0 up, is therefore included as one of the factors in the multiple correlation[2] and is designated as variable X_2.

Before starting the statistical process, we must state the conditions to be observed in fitting a curve to each function. For rainfall, the considerations are quite similar to those discussed in Chapter 8 for irrigation water applied, so we shall use the same conditions as stated there (page 152).

For temperature, the range of possible relations might be wider. There may be certain temperatures to which the plant does not respond and then certain higher temperatures which produce a marked response. Again, if the temperature is too high, a marked reduction in yield

[2] Note the parallel treatment of changes in time as an independent factor in R. A. Fisher, *Statistical Methods for Research Workers*, second edition, p. 174.

TABLE 50

YIELD OF CORN, RAINFALL, AND TEMPERATURE IN SIX LEADING STATES; AND YIELD ESTIMATED BY LINEAR REGRESSIONS ON THREE FACTORS *

Year	Time, X_2	Rainfall, in inches, X_3	Temperature, in degrees, X_4	Yield, in bushels, X_1	Estimated yield, X'_1	Difference, $X_1 - X'_1$ z
1890	0	9.6	74.8	24.5	28.4	−3.9
1891	1	12.9	71.5	33.7	31.6	2.1
1892	2	9.9	74.2	27.9	29.1	−1.2
1893	3	8.7	74.3	27.5	28.5	−1.0
1894	4	6.8	75.8	21.7	27.0	−5.3
1895	5	12.5	74.1	31.9	30.9	1.0
1896	6	13.0	74.1	36.8	31.4	5.4
1897	7	10.1	74.0	29.9	30.0	−0.1
1898	8	10.1	75.0	30.2	29.7	0.5
1899	9	10.1	75.2	32.0	29.8	2.2
1900	10	10.8	75.7	34.0	30.1	3.9
1901	11	7.8	78.4	19.4	27.5	−8.1
1902	12	16.2	72.6	36.0	34.6	1.4
1903	13	14.1	72.0	30.2	33.8	−3.6
1904	14	10.6	71.9	32.4	32.1	0.3
1905	15	10.0	74.0	36.4	31.1	5.3
1906	16	11.5	73.7	36.9	32.2	4.7
1907	17	13.6	73.0	31.5	33.7	−2.2
1908	18	12.1	73.3	30.5	32.9	−2.4
1909	19	12.0	74.6	32.3	32.5	−0.2
1910	20	9.3	73.6	34.9	31.6	3.3
1911	21	7.7	76.2	30.1	29.8	0.3
1912	22	11.0	73.2	36.9	33.0	3.9
1913	23	6.9	77.6	26.8	29.1	−2.3
1914	24	9.5	76.9	30.5	31.0	−0.5
1915	25	16.5	69.9	33.3	37.7	−4.4
1916	26	9.3	75.3	29.7	31.8	−2.1
1917	27	9.4	72.8	35.0	33.0	2.0
1918	28	8.7	76.2	29.9	31.4	−1.5
1919	29	9.5	76.0	35.2	32.1	3.1
1920	30	11.6	72.9	38.3	34.6	3.7
1921	31	12.1	76.9	35.2	33.4	1.8
1922	32	8.0	75.0	35.5	32.1	3.4
1923	33	10.7	74.8	36.7	33.8	2.9
1924	34	13.9	72.6	26.8	36.5	−9.7
1925	35	11.3	75.3	38.0	34.2	3.8
1926	36	11.6	74.1	31.7	35.0	−3.3
1927	37	10.4	71.0	32.6	35.7	−3.1

* Data from E. G. Misner, Studies of the Relation of Weather to the Production and Price of Farm Products, I. Corn. Mimeographed publication, Cornell University, March, 1928. The six states are Iowa, Illinois, Nebraska, Missouri, Indiana, and Ohio.

might be produced.[3] These considerations lead to the following conditions for the temperature curve:

1. It might rise none at all or slowly in the lower range, then more steeply, then taper off until a maximum is reached.
2. It might decline after the maximum, gradually or sharply, but would have only one maximum.
3. It might have two points of inflection, one where it started to rise rapidly, the second where it starts to rise less rapidly.

With respect to the third curve, that for trend, there is no *a priori* reason to expect any given shape during the period concerned, except that there be no sudden changes from year to year. Accordingly, the only condition imposed is that the trend have a smooth, gradual change, with no sharp inflections.

As a preliminary step before starting to determine the net regression curves, we may examine the apparent relation of yield to rainfall, before the other factors (temperature and time) are taken into account.

The apparent relation between rainfall (X_3) and yield (X_1) is indicated in Figure 33, by a dot chart of the relation, with the average yield indicated for each group of years of similar rainfall. The broken line connecting these averages indicates that there is a marked curvilinear relation, the lower increases in rainfall being accompanied by much greater increases in yield than the higher increases. Fitting a straight regression line to these two variables, the relation is found to be

$$X_1 = 23.55 + 0.776X_3$$

This line is accordingly drawn in on the chart, cutting across the curve indicated by the line of group averages.

Although Figure 33 shows yields to be definitely associated with differences in rainfall, it must be noted that rainfall is significantly correlated with X_4, temperature, the correlation being $r_{34} = -0.67$, and is also slightly correlated with time. To some extent, then, the changes in yield shown in the figure to be associated with differences in rainfall may really be due to concomitant differences in the other two

[3] More elaborate investigations, experimental and statistical, have shown that the effect of both temperature and rainfall vary at different times of the season, and especially at certain critical times in the growth of the plant, such as at tasseling. Also, the particular combination of moisture and heat may be important. These possibilities will be referred to subsequently, in connection with more refined and elaborate methods of analysis.

factors. The extent to which these other two factors may have influenced the relations can be judged by determining the multiple correlation of X_1 with all three factors, and then noting how the regression of X_1 on X_3 alone (b_{13}), which has just been shown plotted in the figure, compares with the net regression of X_1 on X_3 ($b_{13.24}$) determined while simultaneously holding constant the linear effects of X_2 and X_4. The first step toward determining the net regression curve, therefore, is to determine the multiple regression equation and the coefficient of multiple correlation, according to the methods outlined in Chapters 12 and 13.

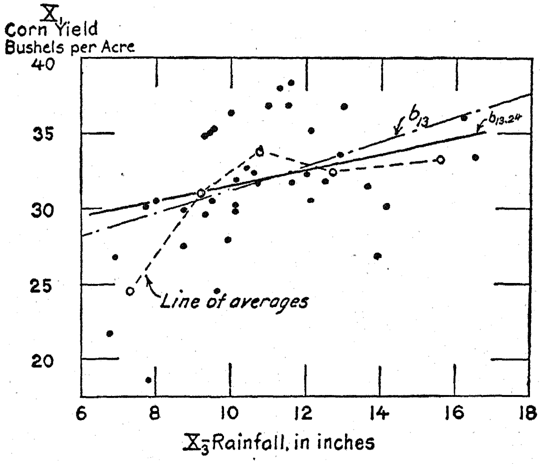


FIG. 33. Apparent relation of corn yields to rainfall (with simple and net regression lines).

The regression equation works out to be

$$X_1 = 53.505 + 0.146X_2 + 0.537X_3 - 0.405X_4$$

and the multiple correlation, adjusted for the number of observations and constants, $\bar{R}_{1.234}$, is 0.49.[4]

[4] Using units of years of time, inches of rainfall in tenths, and degrees of temperature in tenths, and corn yields in tenth bushels, we find the normal equations for the data of Table 50 to be:

$$\begin{aligned} 4{,}569.50b_{12.34} + 248.00b_{13.24} - 8.50b_{14.23} &= 6{,}813.00 \\ 248.00b_{12.34} + 18{,}989.06b_{13.24} - 10{,}279.41b_{14.23} &= 14{,}726.97 \\ -8.50b_{12.34} - 10{,}279.41b_{13.24} + 12{,}408.86b_{14.23} &= -8{,}442.64 \end{aligned}$$

$n\sigma_1^2 = 70{,}455.03$; $\sigma_1 = 43.0$; or 4.3 bushels.

This result shows that when the net linear influence of trend and of temperature is allowed for, yield increases on the average only 0.54 bushel for each increase of one inch in rainfall, whereas, before these other factors were taken into account, yield appeared to increase 0.78 bushel with each additional inch of rainfall. The difference between the simple regression and the net regression may be shown by plotting the latter as well in Figure 33.[5] It is then quite apparent how different are the relations as shown by the two lines.

Considering the effect of the other factors reduces the linear regression of X_1 on X_3 by nearly $\frac{1}{3}$. If other factors have so much effect on the average linear relation, they may have an even greater effect on the shape of the curve. The net regression line in Figure 33 shows the average change in the values of X_1 with different values of X_3, after the differences in X_2 and X_4 are taken into account. The average yield for different groups according to rainfall, connected by the broken line, shows definitely that the simple regression line is but a poor indication of the underlying relation between X_1 and X_3. The net (or partial) regression line may be an equally poor indication of the relation with the other factors held constant. What is needed is some way of seeing the differences in the *individual* values of X_1 for different values of X_3, after the variation due to X_2 and X_4 has been eliminated. It is impossible to do this entirely, for we have as yet no measure of the *curvilinear* relation of X_1 to X_2 or X_3. But we do have our net regression coefficients, which measure the linear regression of X_1 on these other factors, and by using them we can eliminate from X_1 that part of its variation associated with the linear effects of X_2 and X_4, and then see if that gives us any clearer picture of the curvilinear relation between X_1 and X_3.

Determining the "first approximation" net regression curves. Having determined the linear multiple regression equation, we next

[5] The net regression line, showing the change in yield with changes in rainfall while holding constant time and temperature, may be computed from the multiple regression equation by substituting the average values for time and for temperature for X_2 and X_4, and then working out the new constant. For the data given in Table 50, the averages are:

$$M_2 = 18.500;\ M_3 = 10.784;\ M_4 = 74.276;\ M_1 = 31.916$$

If we substitute the means of X_2 and X_4 for their values in the multiple regression equation, that equation becomes:

$$X_1 = 53.505 + (0.146)(18.500) + 0.537X_3 - (0.405)(74.276) = 26.124 + 0.537X_3$$

The net regression line in Figure 33 is therefore drawn in from this last equation.

calculate the estimated value of X_1 for each one of the 38 observations, by substituting the corresponding values of X_2, X_3, and X_4 in the equation. Each of the estimated values (X_1') is then subtracted from the actual value (X_1), giving the residual values (z), as also shown in Table 50.

The next step is to construct a scatter diagram to show the relation between variations in X_3 and the variation in X_1 after that associated with X_2 and X_4 has been eliminated. To do that, the net regression line for X_1 on X_3 is plotted on Figure 34, just as it had been on Figure 33.[6]

The residuals for each observation, from Table 50, are then plotted on the chart, with their X_3 value for abscissa and with the value of z as ordinate *from the net regression line as zero base.* For the first observation, $X_3 = 9.6$ and $z = -3.9$. The ordinate of the point on the net regression line corresponding to $X_3 = 9.6$ is 31.3, and the dot for this observation is correspondingly plotted 3.9 lower than that, at 27.4. For the second observation, $X_3 = 12.9$ and $z = +2.1$. The ordinate of the point on the regression line corresponding to $X_3 = 12.9$ is 33.1; so the dot for this observation is plotted at 33.1 + 2.1, or 35.2. After the corresponding operation has been carried out for all the observations, the figure appears as shown in Figure 34.[7]

If Figure 34 is compared with Figure 33, it is readily seen that the scatter of the dots has been reduced. This will always be true when the other variables show any significant relation to the dependent factor; that is, when $\bar{R}_{1.234}$ exceeds $\bar{r}_{13}$. The scatter is reduced because

[6] To plot the line, all that is necessary is to take the equation of the line to be used (see previous footnote)

$$X_1 = 26.124 + 0.537X_3$$

and substitute any two convenient values for X_3, say 6 and 16.

$$\text{For } X_3 = 6,\ X_1 = 26.124 + (0.537)(6) = 29.35$$

$$\text{For } X_3 = 16,\ X_1 = 26.124 + (0.537)(16) = 34.71$$

With these two sets of coordinates, the line is then drawn in with a straight edge through the points indicated.

[7] The simplest way of plotting the individual observations is to use a scale, which can be slid along the regression line as zero. The values of z are then plotted directly as vertical deviations from the points on the regression line corresponding to the particular values of the independent variable considered, as X_3 in the present case.

that part of the variation in X_1 which can be expressed as net linear functions of X_2 and X_4 has now been eliminated.[8]

Consideration of Figure 34 can be facilitated by computing the means of the ordinates corresponding to the values of X_3 falling within convenient intervals. These can be obtained by simply averaging together the z values for each selected group of values of X_3 and plotting those averages as deviations from the regression line, just as the individual deviations were plotted previously. The necessary averages are as shown in Table 51.

TABLE 51

AVERAGE VALUES OF z, FOR CORRESPONDING X_3 VALUES

X_3 values	Number of cases	Average of X_3	Average of z
Under 8.0	4	7.30	−3.85
8.0– 9.9	10	9.19	+0.16
10.0–10.9	8	10.35	+1.49
11.0–11.9	5	11.40	+2.56
12.0–13.9	8	12.76	−0.52
14.0 and over	3	15.60	−2.20

These averages, when plotted the same as the individual observations and connected by a broken line, give the irregular line also shown in Figure 34. Comparing this line with the similar one in Figure 33,

[8] This can be readily proved. Each point on the net regression line was obtained by the formula:

$$\text{(A)} \qquad X_1 = a_{1.234} + b_{12.34}M_2 + b_{13.24}X_3 + b_{14.23}M_4$$

To these values have been added the residuals, z. These residuals equal $X_1 - X_1'$, and therefore for each observation are equal to

$$\text{(B)} \qquad X_1 - a_{1.234} - b_{12.34}X_2 - b_{13.24}X_3 - b_{14\cdot 23}X_4$$

The ordinate of each dot in Figure 34 is the ordinate of the regression line plus z, and is therefore equal to the sum of the two equations, (A) and (B). If we use π to represent these ordinates, they are therefore equal to

$$\begin{aligned}\pi &= a_{1.234} + b_{12.34}M_2 + b_{13.24}X_3 + b_{14.23}M_4 + X_1 - a_{1.234} - b_{12.34}X_2 \\ &\qquad - b_{13.24}X_3 - b_{14.23}X_4 \\ \pi &= X_1 - b_{12.34}(X_2 - M_2) - b_{14.23}(X_4 - M_4) \\ \pi &= X_1 - b_{12.34}x_2 - b_{14.23}x_4\end{aligned}$$

The adjusted values shown on Figure 34 are therefore simply the values of X_1 less net linear corrections for deviations in X_2 and X_4 from their mean values.

on page 227, we see that though the lines are in general similar there are some marked differences. The average for the second group ($X_3 =$ 8.0–9.9) is now above the straight net regression line, whereas previously it was below it. Likewise the average for $X_3 = 14$ and over is now slightly below the average for $X_3 = 12.0$ to 13.9, whereas before it was a little above it. Also, the difference between the first two averages is not so large as it appeared before. Apparently part of the previous deviations reflected other independent factors.

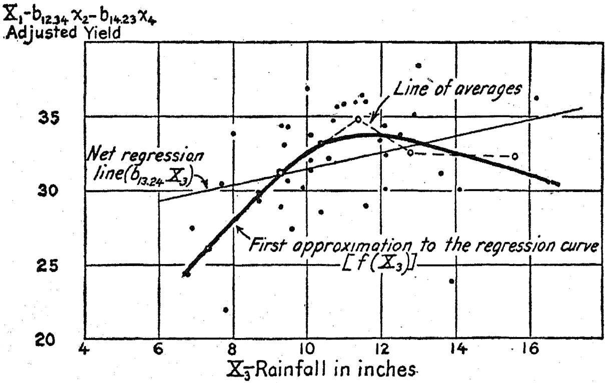


FIG. 34. Rainfall and yield of corn adjusted to average temperature and year, and first approximation curve fitted to the averages. [The notation $f(X_3)$ on the figure corresponds to $f_3(X_3)$.]

It is quite evident that a regression curve is indicated, rising sharply to a maximum yield between 10 and 12 inches of rain, then declining gradually for higher rainfalls. Such a curve is accordingly drawn in freehand, passing as near to the several group averages as is consistent with a continuous smooth curve, and yet conforming to the limiting conditions as to its shape. This curve is the first approximation to the curvilinear function.

$$X_1 = f_3(X_3)$$

which was required to be determined while simultaneously taking into account the curvilinear effects of X_2 and X_4 on X_1. It is only a first approximation because it has been determined while allowing for only the net *linear* effects of the other two variables. If their *curvilinear* effect were determined and allowed for, that might change somewhat the shape of this curve.

The next step is to determine similar first approximations to the curvilinear relation between X_1 and X_2, and between X_1 and X_4, with the net linear effects of the other variables eliminated just as has been done for X_3. It is not necessary to plot the apparent relation between X_1 and X_2 or X_1 and X_4. This was done in the case of X_3 (Figure 33) solely to illustrate the difference between taking the apparent relations and taking the net relations after the linear influence of the other factors had been allowed for (Figure 34). Instead, we may proceed at once to determine the net relations for X_1 to X_2. Figure 35 shows this step.

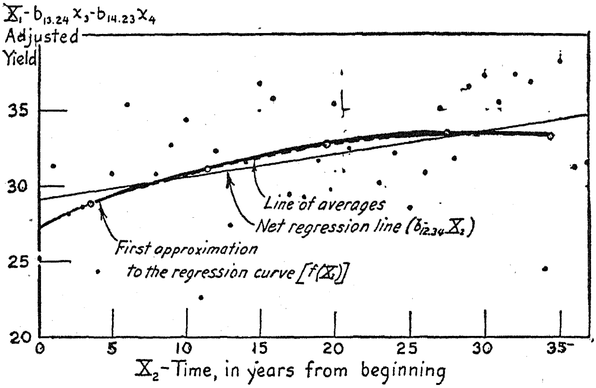


FIG. 35. Time and yield of corn adjusted to average temperature and rainfall, and first approximation curve fitted to the averages. [The notation $f(X_2)$ on the figure corresponds to $f_2(X_2)$.]

This figure is constructed exactly as was Figure 34, by the following steps: (1) Plot the net regression line.[9] (2) Plot in the individual residuals, z, as deviations from that line.[10] (3) Average the residuals grouped according to X_2, plot the group averages, and connect them by

[9] The regression equation, for mean values of X_3 and X_4, becomes

$$
\begin{aligned}
X_1 &= 53.505 + 0.146X_2 + 0.537(M_3) - 0.405(M_4) \\
&= 53.505 + 0.146X_2 + (0.537)(10.784) - (0.405)(74.276) \\
&= 29.214 + 0.146X_2
\end{aligned}
$$

This equation is then the equation to which the net regression line in Figure 35 is drawn. Substituting the values $X_2 = 0$ and $X_2 = 20$ in the equation, values for X_1 of 29.214 and 32.13 are obtained, giving the coordinate points for drawing in the line.

[10] For the first observation, $X_2 = 0$ and $z = -3.9$. The point on the regression line corresponding to $X_2 = 0$ has an ordinate of 29.2. The dot for this observation is accordingly plotted at 29.2 − 3.9, or 25.3. For the next observation, $X_2 = 1$ and $z = 2.1$. The corresponding ordinate on the regression line is 29.4, so the dot is plotted at 29.4 + 2.1, or 31.5. The dot for each observation is plotted in turn in the same way, with a sliding graphic scale to place the dots above or below the regression line.

a broken line. (4) Draw in a smooth curve through the line of averages, if a curve is indicated, conforming to the limiting conditions stated for this curve.

After the first two steps have been carried out, just as described for Figure 34, grouping and averaging the residuals with respect to X_2 give the averages shown in Table 52.

TABLE 52

AVERAGE VALUES OF z FOR CORRESPONDING X_2 VALUES

X_2 values	Number of cases	Average of X_2	Average of z
0- 7	8	3.5	−0.38
8–15	8	11.5	+0.24
16–23	8	19.5	+0.64
24–31	8	27.5	+0.26
32–37	6	34.5	−1.00

The average residuals shown in the table are then plotted in above and below the regression line in Figure 35 and connected by a broken line. This line of averages indicates that corn yield (for years of similar rainfall and temperature) rose rapidly during the earlier years, then more and more gradually, until during the last ten years it tended to remain about on the same level. A smooth continuous curve is therefore drawn through the averages, completing step (4) and giving the first approximation to the curvilinear net regression of X_1 on X_2, $f_2(X_2)$.

The same operations are then carried out for X_4 as shown in Figure 36. After drawing in the net regression line,[11] and plotting in the individual observations,[12] we group the residuals on X_4 and average, with the results shown in Table 53.

[11] The net regression line for X_1 and X_4 may be determined by an alternative method to that used before. On such charts as Figures 34, 35 or 36, the net regression line will always pass through the mean of the two variables. For Figure 36, therefore, X_1 will have its mean value, 31.92, when X_4 has its mean value, 74.28. From the net regression coefficient, $b_{14.23}$, it is evident that each unit increase in X_4 is accompanied by -0.405 unit increase in X_1. If X_4 is increased from 74.28 to 78.28, or 4 units, X_1 will change by $(-0.405)(4)$, or -1.62. For $X_4 = 78.28$, X_1 will therefore be $31.92 - 1.62$, or 30.30. This gives the two sets of points necessary to locate the line; when $X_4 = 74.28$, $X_1 = 31.92$; and when $X_4 = 78.28$, $X_1 = 30.30$.

[12] The individual residuals are plotted in the same way as indicated in the other two cases; the residual -3.9 for $X_4 = 74.8$ is plotted 3.9 units below the corresponding point on the regression line, and similarly for the other observations.

TABLE 53

AVERAGE VALUES OF z FOR CORRESPONDING X_4 VALUES

X_4 values	Number of cases	Average of X_4	Average of z
Under 72.0	4	71.08	−1.28
72.0–72.9	5	72.58	−1.24
73.0–73.9	5	73.36	+1.46
74.0–74.9	10	74.30	+0.49
75.0–75.9	7	75.33	+0.91
76.0–76.9	5	76.44	+0.64
77.0 and over	2	78.00	−5.20
76.0 and over	7	76.89	−1.03

The last group, on the first grouping, has but two cases, so the last two groups are combined, giving the averages shown in the last line. The fact that both the items above 77 degrees are low, also evident in Figure 36, would give a little more reliability to the average based on only two items; but it is generally unsafe to give such an extreme bend to the end of a regression curve as this would call for, on the basis of so few observations. The larger grouping will therefore be used in this case, leaving the subsequent approximations to determine whether the more extreme bend is justified.

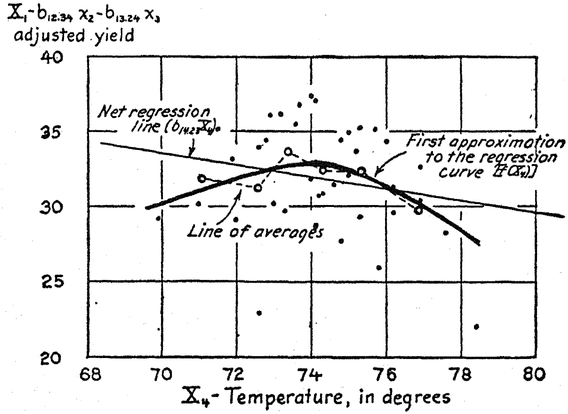


FIG. 36. Temperature and yield of corn adjusted to average rainfall and year, and first approximation curve fitted to the averages. [The notation $f(X_4)$ on the figure corresponds to $f_4(X_4)$.]

The line of averages in Figure 36 indicates that yields may tend to rise as temperature increases up to between 73 and 75 degrees, and then to fall as the temperature goes still higher. A smooth curve is therefore drawn in, averaging out the irregularities shown in the broken line of the group averages and conforming to the limiting conditions stated on page 226. It does not make much difference if these first approximation curves are not drawn in in exactly the right position or shape, as the subsequent operations will tend to correct them to the proper shape if the original one is incorrect. It is for that reason that fairly accurate results can be secured by this graphic process, even though the true shape of the curves is not known at the beginning.

Estimating X_1 *from the first approximation curves.* We have now arrived at first approximations to the net regression curves for X_1, against each of the three factors. It must be remembered that in making the adjustments on X_1 to arrive at these curves, only the net *linear* effects of the other independent variables have been eliminated. Now that we have at least an approximate measure of the curvilinear relations of X_1 to the independent variables, making adjustments to eliminate these approximate curvilinear effects may enable us to determine more accurately the true curvilinear relation to each variable.

The first step in the next stage of the process is to work out estimated values of X_1 based on the curvilinear relations. To do this we may designate the relation between X_1 and X_2 shown by the curve in Figure 35 as $f'_2(X_2)$; the relation between X_1 and X_3 shown in Figure 34 as $f'_3(X_3)$; and the relation between X_1 and X_4 shown in Figure 36 as $f'_4(X_4)$. The estimates of X_1 may then be worked out by the regression equation

$$X''_1 = a'_{1.234} + f'_2(X_2) + f'_3(X_3) + f'_4(X_4) \tag{57}$$

The symbol X''_1 is used to designate this second set of estimates, just as X'_1 was used to designate the first set, worked out from the linear regression equation. The constant $a'_{1.234}$ is different from the constant $a_{12.34}$ used in equation (36); its value is given by the formula

$$a'_{1.234} = M_1 - \frac{\Sigma[f'_2(X_2) + f'_3(X_3) + f'_4(X_4)]}{n} \tag{58}$$

To work out $a'_{1.234}$ according to equation (58), it is first necessary to work out the value $f'_2(X_2) + f'_3(X_3) + f'_4(X_4)$ for each set of observations. For the first observation, for example, $X_2 = 0$, $X_3 = 9.6$, and $X_4 = 74.8$. From $f'_2(X_2)$, given in Figure 35, the curve reading (or ordinate) corresponding to a value of 0 for X_2 is 27.3. For $f'_3(X_3)$, Figure 34, the ordinate of the curve corresponding to $X_3 = 9.6$ is 31.7. For $f'_4(X_4)$, Figure 36, the curve ordinate corresponding to $X_4 = 74.8$ is 32.5. The value $[f'_2(X_2) + f'_3(X_3) + f'_4(X_4)]$ for the first observation is therefore [27.3 + 31.7 + 32.5], or 91.5. The sum of these values for each observation is the value required in equation (58).

Before continuing the process of reading each value from the charts for the remaining observations, it should be noted that, since many observations of each variable have the same values, the same point would be read from each chart many times. The process of working out the computations can be much simplified by reading each required value from each chart once for all and recording it so that it can be used each time. Since each chart indicates each individual observation for each independent variable, only those points for which there are observations need be recorded. Carrying out this process, we may record the functional relations as shown in Tables 54, 55, and 56, which show the readings from Figures 35, 34, and 36, respectively.[13]

TABLE 54

Values of X_1 Corresponding to Given Values of X_2, from the First Approximation Curve

X_2	$f'_2(X_2)$	X_2	$f'_2(X_2)$	X_2	$f'_2(X_2)$	X_2	$f'_2(X_2)$
0	27.3	10	30.8	20	32.8	29	33.4
1	27.8	11	31.0	21	33.0	30	33.5
2	28.2	12	31.3	22	33.1	31	33.5
3	28.6	13	31.5	23	33.1	32	33.5
4	29.0	14	31.7	24	33.2	33	33.5
5	29.4	15	31.9	25	33.2	34	33.5
6	29.7	16	32.1	26	33.3	35	33.5
7	30.0	17	32.3	27	33.3	36	33.5
8	30.3	18	32.5	28	33.4	37	33.5
9	30.6	19	32.6				

[13] In entering these values it is not worth while reading further than the first decimal, for the line is not drawn more accurately than to within 0.1 or 0.2. The accuracy depends, of course, on the scale; but it is not worth using very large charts to secure spuriously high accuracy, when the standard error of any particular point on the curve is probably several units and when the curve is only a first approximation, subject to subsequent modification.

The values to determine $a'_{1.234}$ may now be worked out in orderly manner, as shown in Table 57, in the fourth to the seventh columns.

TABLE 55

VALUES OF X_1 CORRESPONDING TO GIVEN VALUES OF X_3, FROM THE FIRST APPROXIMATION CURVE

X_3	$f'_3(X_3)$	X_3	$f'_3(X_3)$	X_3	$f'_3(X_3)$	X_3	$f'_3(X_3)$
6.8	24.6	9.5	31.5	10.8	33.4	12.9	33.3
6.9	25.0	9.6	31.7	11.0	33.5	13.0	33.2
7.7	27.1	9.9	32.4	11.3	33.6	13.6	32.9
7.8	27.4	10.0	32.5	11.5	33.7	13.9	32.7
8.0	27.9	10.1	32.6	11.6	33.7	14.1	32.5
8.7	29.7	10.4	33.1	12.0	33.7	16.2	31.0
9.3	31.0	10.6	33.3	12.1	33.6	16.5	30.8
9.4	31.2	10.7	33.4	12.5	33.5		

TABLE 56

VALUES OF X_1 CORRESPONDING TO GIVEN VALUES OF X_4, FROM THE FIRST APPROXIMATION CURVE

X_4	$f'_4(X_4)$	X_4	$f'_4(X_4)$	X_4	$f'_4(X_4)$	X_4	$f'_4(X_4)$
69.9	30.2	73.0	32.5	74.2	32.8	75.7	31.6
71.0	31.0	73.2	32.6	74.3	32.7	75.8	31.5
71.5	31.4	73.3	32.6	74.6	32.6	76.0	31.3
71.9	31.7	73.6	32.7	74.8	32.5	76.2	31.0
72.0	31.8	73.7	32.7	75.0	32.3	76.9	30.1
72.6	32.2	74.0	32.8	75.2	32.1	77.6	29.0
72.8	32.3	74.1	32.8	75.3	32.0	78.4	27.6
72.9	32.4						

This computation gives us the sum of the respective functional values for the 38 observations. Substituting this sum and the number of observations in equation (58), we find the required constant to be

$$a'_{1.234} = 31.916 - \frac{3621.9}{38} = -63.397$$

Since the functional values for our regression equation are only expressed to one decimal point, we shall use -63.4 for $a'_{1.234}$, which will result in the estimated values being 0.003 unit too low, on the average.

It is now possible to complete the process of computing X_1'', the estimated value of X_1, using the first approximation curves, according to equation (57), and the constant which has just been computed. When equations (57) and (58) are compared, it is evident that, except

TABLE 57

Computation of Functional Values Corresponding to Independent Variables, of the Estimated Value of X_1, and the New Residual, for Each Observation

1	2	3	4	5	6	7	8	9	10
X_2	X_3	X_4	$f_2'(X_2)$	$f_3'(X_3)$	$f_4'(X_4)$	$f_2'(X_2)$ $+f_3'(X_3)$ $+f_4'(X_4)$	$\Sigma(f)+a'$ $=X_1''$	X_1	X_1-X_1'' z''
0	9.6	74.8	27.3	31.7	32.5	91.5	28.1	24.5	−3.6
1	12.9	71.5	27.8	33.3	31.4	92.5	29.1	33.7	4.6
2	9.9	74.2	28.2	32.4	32.8	93.4	30.0	27.9	−2.1
3	8.7	74.3	28.6	29.7	32.7	91.0	27.6	27.5	−0.1
4	6.8	75.8	29.0	24.6	31.5	85.1	21.7	21.7	0
5	12.5	74.1	29.4	33.5	32.8	95.7	32.3	31.9	−0.4
6	13.0	74.1	29.7	33.2	32.8	95.7	32.3	36.8	4.5
7	10.1	74.0	30.0	32.6	32.8	95.4	32.0	29.9	−2.1
8	10.1	75.0	30.3	32.6	32.3	95.2	31.8	30.2	−1.6
9	10.1	75.2	30.6	32.6	32.1	95.3	31.9	32.0	0.1
10	10.8	75.7	30.8	33.4	31.6	95.8	32.4	34.0	1.6
11	7.8	78.4	31.0	27.4	27.6	86.0	22.6	19.4	−3.2
12	16.2	72.6	31.3	31.0	32.2	94.5	31.1	36.0	4.9
13	14.1	72.0	31.5	32.5	31.8	95.8	32.4	30.2	−2.2
14	10.6	71.9	31.7	33.3	31.7	96.7	33.3	32.4	−0.9
15	10.0	74.0	31.9	32.5	32.8	97.2	33.8	36.4	2.6
16	11.5	73.7	32.1	33.7	32.7	98.5	35.1	36.9	1.8
17	13.6	73.0	32.3	32.9	32.5	97.7	34.3	31.5	−2.8
18	12.1	73.3	32.5	33.6	32.6	98.7	35.3	30.5	−4.8
19	12.0	74.6	32.6	33.7	32.6	98.9	35.5	32.3	−3.2
20	9.3	73.6	32.8	31.0	32.7	96.5	33.1	34.9	1.8
21	7.7	76.2	33.0	27.1	31.0	91.1	27.7	30.1	2.4
22	11.0	73.2	33.1	33.5	32.6	99.2	35.8	36.9	1.1
23	6.9	77.6	33.1	25.0	29.0	87.1	23.7	26.8	3.1
24	9.5	76.9	33.2	31.5	30.1	94.8	31.4	30.5	−0.9
25	16.5	69.9	33.2	30.8	30.2	94.2	30.8	33.3	2.5
26	9.3	75.3	33.3	31.0	32.0	96.3	32.9	29.7	−3.2
27	9.4	72.8	33.3	31.2	32.3	96.8	33.4	35.0	1.6
28	8.7	76.2	33.4	29.7	31.0	94.1	30.7	29.9	−0.8
29	9.5	76.0	33.4	31.5	31.3	96.2	32.8	35.2	2.4
30	11.6	72.9	33.5	33.7	32.4	99.6	36.2	38.3	2.1
31	12.1	76.9	33.5	33.6	30.1	97.2	33.8	35.2	1.4
32	8.0	75.0	33.5	27.9	32.3	93.7	30.3	35.5	5.2
33	10.7	74.8	33.5	33.4	32.5	99.4	36.0	36.7	0.7
34	13.9	72.6	33.5	32.7	32.2	98.4	35.0	26.8	−8.2
35	11.3	75.3	33.5	33.6	32.0	99.1	35.7	38.0	2.3
36	11.6	74.1	33.5	33.7	32.8	100.0	36.6	31.7	−4.9
37	10.4	71.0	33.5	33.1	31.0	97.6	34.2	32.6	−1.6
Totals...	....	....	1,208.4	1,204.2	1,209.3	3,621.9			

for the constant term, X_1'' is equal to the values that have just been computed in the seventh column of Table 57. Accordingly, all that is necessary is to subtract 63.4 from each of those values. This step is shown also in Table 57, in the eighth column.

The column headed X_1'' shows the estimated values obtained by this process. The next step is to see whether the new estimates come any nearer to reproducing the observed values of X_1 than did the first set of estimates, based on the linear regression equation. We therefore compute a new set of residuals, z'', by subtracting the new estimates from the actual values of X_1. This step, also, is shown in Table 57.

$$z'' = X_1 - X_1'' \tag{59}$$

If the individual residuals shown are compared with the residuals obtained by the linear regression, as computed in Table 50, it will be seen that in general the new residuals are smaller than the previous ones, though the reverse is true in many cases. There are 23 cases in which the new residual is smaller, and 15 in which it is larger than the original residual. A more accurate comparison can be obtained by comparing the standard deviations of the residuals for the two sets. For the linear correlation, the standard deviation of the residuals was 3.6 bushels, whereas the standard deviations of the new residuals is 3.0 bushels. Apparently the new estimates do come nearer to the observed values, on the average, than did the first set of estimates. (See also Note on page 258.)

Determining the second approximation regression curves. The regression curves used in constructing the estimate X_1'' were only the first approximations to the true curvilinear relations, since they were determined by eliminating only the linear effects of the other independent factors. Now that the residuals obtained by the use of the first approximation curves have been computed, however, we can determine whether any change in the shape of the several curves is necessary.

To do this we construct Figure 37 by drawing in the regression curve from Figure 35, using the same scale as before. Use of Table 54 makes it easier to reproduce the curve. Next we plot each of the last residuals as a deviation just as before, except that now the residuals are plotted as deviations from the regression curve, instead of from the regression line, at the point corresponding to the independent variable X_2. Thus the first observation, with $X_2 = 0$, has $z'' = -3.6$. The point on the curve corresponding to $X_2 = 0$ is 27.3; so the dot has for ordinate 27.3 − 3.6, or 23.7. The values for next observation are $X_2 = 1$ and

$z'' = 4.6$. The corresponding value of $f_2'(X_2)$ is 27.8, so the ordinate for the dot is 27.8 + 4.6, or 32.4. The coordinates for this dot are therefore 1 and 32.4. The remaining observations are plotted in the same manner, shortening the process by scaling the value for z'' directly above or below from the corresponding point on the regression curve.

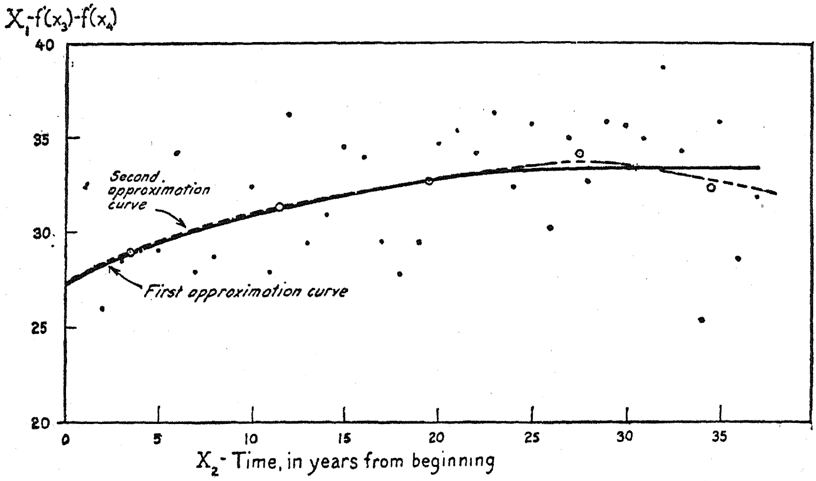


FIG. 37. Time, and yield of corn adjusted to average temperature and rainfall on basis of first approximation curves; and second approximation to $f_2(X_2)$.

With the dots all plotted, it is evident that the scatter is too great to indicate definitely changes which may be needed in the curve, if any, simply from the dots alone. Accordingly the residuals are averaged in groups, employing the same grouping as before (Table 52), which eliminates the need of averaging the corresponding X_2 values over again. The new averages work out as shown in Table 58.

TABLE 58

AVERAGE VALUES OF z'', FOR CORRESPONDING X_2 VALUES

X_2 values	Number of cases	Average of X_2	Average of z''
0– 7	8	3.5	+0.10
8–15	8	11.5	+0.16
16–23	8	19.5	−0.08
24–31	8	27.5	+0.64
31–37	6	34.5	−1.08

The averages are next plotted as deviations from the first approximation curve. They indicate that a slight raise in the lower part of the curve may be needed, and a downward bend toward the end. It appears that now that the influence of rainfall and temperature on yield have been more accurately allowed for, the upward trend with time is slightly less than it seemed before in the early years; and the trend seems to have turned downward toward the end of the series—the exact year or extent of the turn is indeterminate. A new curve is therefore drawn in in Figure 37, and, as it happens, a smooth, continuous curve can be drawn exactly through each of the first three group averages, but not having the extreme bend indicated by the last two group averages.

The same process may now be applied to X_3, to see if any change need be made in the first regression curve for the change in X_1 with changes in that variable. This process is carried out as shown in Figure 38, the first approximation curve being drawn in just as before, using the data given in Table 55.

Instead of plotting the individual residuals for each observation, as was just done with respect to X_2, we may proceed at once to compute the average residuals for each of the groups of values of X_3, since it is sufficiently apparent from Figure 37 that the scatter of the individual observations is still too great to serve as a guide in correcting the first approximation curves. Averaging the residuals gives the averages shown in Table 59.

TABLE 59

Average Values of z'', for Corresponding X_3 Values

X_3 values	Number of cases	Average of X_3	Average of z''	Average of X_3	Average of z''
Under 8.0	4	7.30	+0.58		
8.0– 9.9	10	9.19	+0.03		
10.0–10.9	8	10.35	−0.15	10.75	+0.09
11.0–11.9	5	11.40	+0.48		
12.0–13.9	8	12.76	−1.11	13 53	−0.34
14.0 and over	3	15.60	+1.73		

Again the averages are somewhat irregular when plotted, so the last four groups are reduced to two, and the new averages plotted and indicated separately. The number of observations represented by each of the first set of averages is indicated next to it, so that averages based on a small number of observations will not be given undue weight in drawing in the curve. It might be desirable in some cases, also, to try

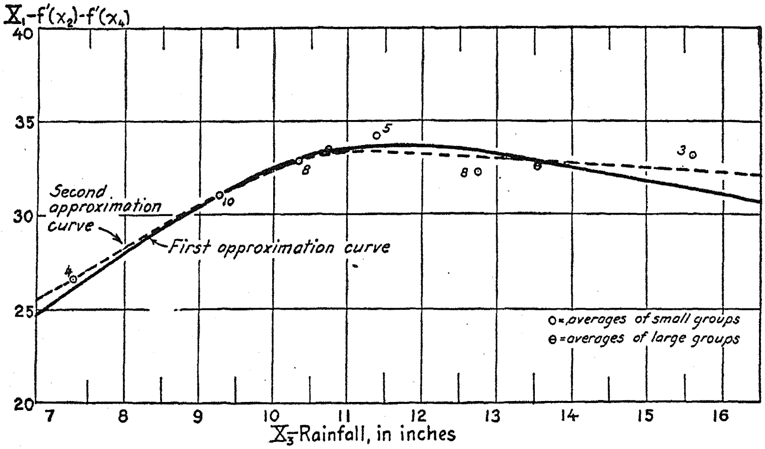


FIG. 38. Rainfall, and yield of corn adjusted to average temperature and time on the basis of first approximation curves; and second approximation to $f_3(X_3)$.

regrouping the cases into different groups—say from 8.5 to 9.4, 9.5 to 10.4, etc.—and see if that would change at all the indications as to the

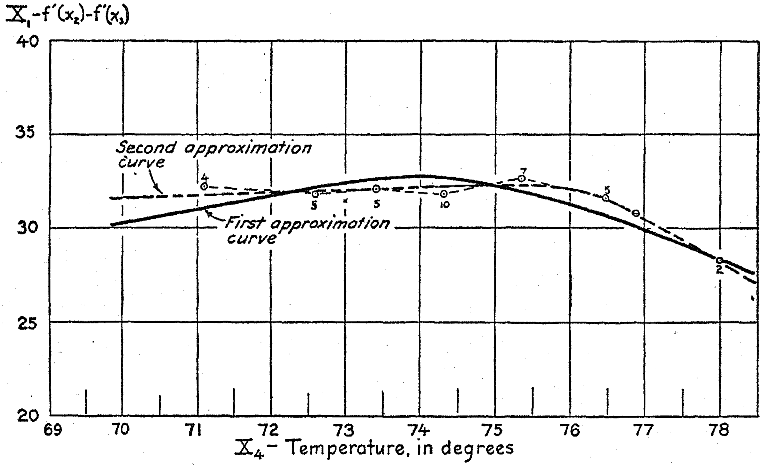


FIG. 39. Temperature, and yield of corn adjusted to average rainfall and time on the basis of first approximation curves; and second approximation to $f_4(X_4)$.

shifts needed in the first curve. Working that out in this case, the changes needed are still found to be about the same as shown by the group averages in Figure 38, though somewhat less regular, owing to the smaller size of groups. A new curve is then drawn in freehand, as indicated by the group averages, rising somewhat higher than formerly at both ends, and not rising quite so high in the central portion as before.

Turning to the relation between X_1 and X_4, the first approximation curve for $f_4^{'}(X_4)$ is reproduced in Figure 39, using the values given in Table 56. The next step is to average the values of z'' for corresponding values of X_4. Using the same groupings used in Table 53, we arrive at the averages shown in the following table:

TABLE 60

AVERAGE VALUES OF z'', FOR CORRESPONDING X_4 VALUES

X_4 values	Number of cases	Average of X_4	Average of z''
Under 72.0	4	71.08	+1.15
72.0–72.9	5	72.58	−0.36
73.0–73.9	5	73.36	−0.58
74.0–74.9	10	74.30	−0.86
75.0–75.9	7	75.33	+0.63
76.0–76.9	5	76.44	+0.90
77.0 and over	2	78.00	−0.05
76.0 and over	7	76.89	+0.63

Plotting these new averages, and connecting them by a broken line, we see that the relation of yield to temperature may be quite different from the way it appeared on the first approximation. Apparently the highest yields are obtained around 75 to 76 degrees, instead of at 74 degrees; higher temperatures appear to reduce the yield markedly, but lower temperatures have only a slight influence on the yield. These indications are all within the theoretical limitations on the shape of the curve, as stated on page 226. The new curve, drawn in freehand so as to pass as nearly through these new averages as possible and still maintain a smooth continuous shape, with only a single maximum, expresses these relations.

Estimating X_1 from the second approximation curves. Now that the second approximation curves have been determined for each variable, we can proceed to estimate values of X_1 on the basis of the revised curves, to see whether the new curves enable us to estimate X_1 any more accurately than the first set of curves did. To facilitate the process we first construct tables for $f_2''(X_2)$, $f_3''(X_3)$, and $f_4''(X_4)$, showing the readings for the functions from the revised curves.

TABLE 61

VALUES OF X_1 CORRESPONDING TO GIVEN VALUES OF X_2, FROM THE SECOND APPROXIMATION CURVE

X_2	$f_2''(X_2)$	X_2	$f_2''(X_2)$	X_2	$f_2''(X_2)$	X_2	$f_2''(X_2)$
0	27.4	10	31.0	20	32.7	29	33.6
1	27.9	11	31.2	21	33.0	30	33.5
2	28.4	12	31.4	22	33.2	31	33.4
3	28.8	13	31.6	23	33.3	32	33.2
4	29.2	14	31.8	24	33.4	33	33.0
5	29.5	15	32.0	25	33.5	34	32.8
6	29.8	16	32.1	26	33.6	35	32.6
7	30.2	17	32.3	27	33.7	36	32.4
8	30.4	18	32.5	28	33.7	37	32.2
9	30.7	19	32.6				

To simplify the calculations, 20 is subtracted from each of the functional values in making subsequent entries. The computations to determine the estimated values are then carried out as shown in detail in Table 64 following, just as for Table 57. In practical computation these entries, for the second approximation curves, would be made on the same sheet as were the entries in Table 57 for the first approxima-

TABLE 62

VALUES OF X_1 CORRESPONDING TO GIVEN VALUES OF X_3, FROM THE SECOND APPROXIMATION CURVE

X_3	$f_3''(X_3)$	X_3	$f_3''(X_3)$	X_3	$f_3''(X_3)$	X_3	$f_3''(X_3)$
6.8	25.5	9.5	31.5	10.8	33.3	12.9	33.0
6.9	25.7	9.6	31.7	11.0	33.4	13.0	33.0
7.7	27.5	9.9	32.2	11.3	33.4	13.6	32.8
7.8	27.8	10.0	32.3	11.5	33.3	13.9	32.7
8.0	28.2	10.1	32.5	11.6	33.3	14.1	32.7
8.7	29.9	10.4	32.9	12.0	33.2	16.2	32.2
9.3	31.1	10.6	33.1	12.1	33.2	16.5	32.1
9.4	31.3	10.7	33.2	12.5	33.1		

tion curves, thus eliminating the work of entering the values of X_1, X_2, X_3, and X_4 over again.

Table 64 is worked out just as was Table 57. Thus the data for the first observation show values of 0, 9.6, and 74.8 for X_2, X_3, and X_4, respectively. Looking up the corresponding values in Tables 61, 62, and 63 gives values of 27.4, 31.7, and 32.3, for the three functional values. Subtracting 20 from each value, to reduce the subsequent clerical work, we enter 7.4, 11.7, and 12.3 in the functional columns.

TABLE 63

Values of X_1 Corresponding to Given Values of X_4, from the Second Approximation Curve

X_4	$f_4''(X_4)$	X_4	$f_4''(X_4)$	X_4	$f_4''(X_4)$	X_4	$f_4''(X_4)$
69.9	31.6	73.0	32.0	74.2	32.2	75.7	32.2
71.0	31.7	73.2	32.0	74.3	32.2	75.8	32.2
71.5	31.8	73.3	32.0	74.6	32.2	76.0	32.1
71.9	31.8	73.6	32.1	74.8	32.3	76.2	32.0
72.0	31.8	73.7	32.1	75.0	32.3	76.9	30.7
72.6	31.9	74.0	32.2	75.2	32.3	77.6	29.1
72.8	32.0	74.1	32.2	75.3	32.3	78.4	27.3
72.9	32.0						

The three functional values are then added, and the sum entered in the seventh column. The entries for the functional readings are completed as shown, and the sum computed for each observation. Then the average of the seventh column is determined, giving the value 35.30. As the average of X_1 is 31.916, the value of the new constant, $a''_{1.234}$, is found by equation (58) to be

$$\begin{aligned} a''_{1.234} &= 31.916 - 35.300 \\ &= -3.384 \end{aligned}$$

Accordingly, 3.4 is subtracted from each of the values in column 7 to give the estimated value of X_1, X_1''', which is then entered in the eighth column of Table 64.

The final step in computing the table is to subtract each of the estimated values, X_1''', from the actual value X_1, giving the residuals z''', which appear in the last column.

Comparing the new residuals, z''', with the previous ones, z'', given in Table 58, we find that the size of the residuals has been increased in just about as many cases as it has been decreased. But when we compute the standard deviation of the new residuals, we find that the

TABLE 64

Computation of Functional Values, from the Second Approximation Curves, Corresponding to Independent Variables for Each Observation, and Computation of Estimated Value for X_1 and of New Residuals

Independent variables			Corresponding functional values *			$f_2''(X_2)$ $+f_3''(X_3)$ $+f_4''(X_4)$	$(7)-a$ $=X_1'''$	Dependent variable X_1	X_1-X_1''' z'''
X_2	X_3	X_4	$f_2''(X_2)$	$f_3''(X_3)$	$f_4''(X_4)$				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	9.6	74.8	7.4	11.7	12.3	31.4	28.0	24.5	−3.5
1	12.9	71.5	7.9	13.0	11.8	32.7	29.3	33.7	4.4
2	9.9	74.2	8.4	12.2	12.2	32.8	29.4	27.9	−1.5
3	8.7	74.3	8.8	9.9	12.2	30.9	27.5	27.5	0
4	6.8	75.8	9.2	5.5	12.2	26.9	23.5	21.7	−1.8
5	12.5	74.1	9.5	13.1	12.2	34.8	31.4	31.9	0.5
6	13.0	74.1	9.8	13.0	12.2	35.0	31.6	36.8	5.2
7	10.1	74.0	10.2	12.5	12.2	34.9	31.5	29.9	−1.6
8	10.1	75.0	10.4	12.5	12.3	35.2	31.8	30.2	−1.6
9	10.1	75.2	10.7	12.5	12.3	35.5	32.1	32.0	−0.1
10	10.8	75.7	11.0	13.3	12.2	36.5	33.1	34.0	0.9
11	7.8	78.4	11.2	7.8	7.3	26.3	22.9	19.4	−3.5
12	16.2	72.6	11.4	12.2	11.9	35.5	32.1	36.0	3.9
13	14.1	72.0	11.6	12.7	11.8	36.1	32.7	30.2	−2.5
14	10.6	71.9	11.8	13.1	11.8	36.7	33.3	32.4	−0.9
15	10.0	74.0	12.0	12.3	12.2	36.5	33.1	36.4	3.3
16	11.5	73.7	12.1	13.3	12.1	37.5	34.1	36.9	2.8
17	13.6	73.0	12.3	12.8	12.0	37.1	33.7	31.5	−2.2
18	12.1	73.3	12.5	13.2	12.0	37.7	34.3	30.5	−3.8
19	12.0	74.6	12.6	13.2	12.2	38.0	34.6	32.3	−2.3
20	9.3	73.6	12.7	11.1	12.1	35.9	32.5	34.9	2.4
21	7.7	76.2	13.0	7.5	12.0	32.5	29.1	30.1	1.0
22	11.0	73.2	13.2	13.4	12.0	38.6	35.2	36.9	1.7
23	6.9	77.6	13.3	5.7	9.1	28.1	24.7	26.8	2.1
24	9.5	76.9	13.4	11.5	10.7	35.6	32.2	30.5	−1.7
25	16.5	69.9	13.5	12.1	11.6	37.2	33.8	33.3	−0.5
26	9.3	75.3	13.6	11.1	12.3	37.0	33.6	29.7	−3.9
27	9.4	72.8	13.7	11.3	12.0	37.0	33.6	35.0	1.4
28	8.7	76.2	13.7	9.9	12.0	35.6	32.2	29.9	−2.3
29	9.5	76.0	13.6	11.5	12.1	37.2	33.8	35.2	1.4
30	11.6	72.9	13.5	13.3	12.0	38.8	35.4	38.3	2.9
31	12.1	76.9	13.4	13.2	10.7	37.3	33.9	35.2	1.3
32	8.0	75.0	13.2	8.2	12.3	33.7	30.3	35.5	5.2
33	10.7	74.8	13.0	13.2	12.3	38.5	35.1	36.7	1.6
34	13.9	72.6	12.8	12.7	11.9	37.4	34.0	26.8	−7.2
35	11.3	75.3	12.6	13.4	12.3	38.3	34.9	38.0	3.1
36	11.6	74.1	12.4	13.3	12.2	37.9	34.5	31.7	−2.8
37	10.4	71.0	12.2	12.9	11.7	36.8	33.4	32.6	−0.8
Totals...	....	....	447.6	445.1	448.7	1341.4			

* Less 20.0 for each functional reading.

standard deviation of z''' is 2.80 bushels, or slightly smaller than the standard deviation of z'', 3.0 bushels. (See Note on page 258.)

Correcting the curves by further successive approximations. The process ordinarily would be carried through one or more additional approximations by repeating the steps shown. Thus the last residuals, z''', when averaged and plotted with respect to the second set of approximation curves, would indicate whether any further modifications were needed in the curves; if any were made, new readings would be made from the new curves, new estimates of X_1 obtained from them, and another set of residuals determined. So long as the standard deviation of each new set of residuals is smaller than that of the previous set (and no more complicated curves were drawn in, which would require more constants to represent them), the approximation curves may be regarded as approaching closer and closer to the underlying true curves. When, however, the curves have been determined as closely as is possible from the given data, the standard deviation of the residuals will show no further decrease and may even increase slightly. In such case the set of curves showing the lowest standard deviation of residuals (and yet conforming to the hypothetical limitations) may be regarded as the final curves determined by the process.[14]

We can make a check on the slope and amplitude of the final curves by the method of least squares, using the supplementary methods set forth in pages 401 to 403 of Chapter 22. Or if it is desired to have a mathematical expression of the several curves, equations may be selected capable of representing the several curves whose shape has been determined by the graphic successive approximation process, fitting the mathematical curves according to the methods presented briefly earlier in this chapter, on pages 221 and 222, and described in more detail in the first section of Chapter 22.

Stating the final conclusions. After the final shape of the several net regression curves has been determined, it still remains to state those curves in such shape that their meaning is perfectly clear. The several functions may be stated to show the value of the dependent factor associated with given values of the particular independent factor when values of other independent factors are held at their mean. There are two alternative ways of stating the associated values: (1) as actual values and (2) as deviations from the mean values.

[14] In very exact work, the effect upon the residuals of modifications in each curve separately might be tested after this point, to insure that each individual regression curve had been fitted to the data with the greatest degree of accuracy.

To state the associated values as actual values, we may use the following procedure:

First, the mean of all the values read from the final curve is determined. For $f_2(X_2)$, this mean may be designated $M_{f(2)}$. The values from the curve are read off for selected intervals of X_2. Then the estimated values of X_1 for each of these values of X_2 (with values of X_3, X_4, etc., at their means) are determined by subtracting the mean of the curve readings from each of these actual readings and adding to the result the mean of X_1. That is, if we use $X_1 = F_2(X_2)$ to designate these values of X_1, estimated from the net curvilinear relation to X_2, we can define them by the equation

$$X_1' = F_2(X_2) = f_2(X_2) - M_{f(2)} + M_1 \qquad (60)$$

If, however, the expected values of X_1 for given values of X_2 are to be stated merely as deviations from the mean values, those deviations may be determined by subtracting from each curve reading the mean of all the curve readings. If we use $F_2(x_2)$ to designate these expected deviations from the mean values, we may define them by the equation

$$x_1' = F_2(x_2) = f_2(X_2) - M_{f(2)} \qquad (61)$$

It is evident, from equations (60) and (61), that

$$F_2(X_2) = F_2(x_2) + M_1$$

In the actual statement of the results of a correlation study, it is frequently desirable to state the relation of the dependent factor to the most important independent factor according to equation (60), and to state the relation for the remaining independent factors according to equation (61). When that is done, the estimated values of X_1, based on all the independent factors, may be readily computed by taking the estimate from the most important factor, and then adding to or subtracting from that the corrections to take account of the departures of other factors from their means. Using X_1' to designate this final estimate of the value X_1, and taking X_3 as the most important factor, we make the estimate by the equation

$$X_1' = F_2(x_2) + F_3(X_3) + F_4(x_4) + \cdots + F_n(x_n) \qquad (62)$$

The process of working out these final statements of the net curvilinear regression lines may be illustrated by the data of the corn-yield problem. Since the rainfall (X_3) was apparently the most important factor, that may be taken as the one for which the regression is to be stated according to equation (60). If we regard the second approximation curve shown in Figure 38 and Table 62 as the final curve, then Table 64 gives the readings from this curve for each of the individual observations.

The mean of the readings of $f_3(X_3)$ is next computed from the values of Table 64. The sum of the 38 $f''(X_3)$ readings is 445.1, so

$$M_{f(X_3)} = \frac{445.1}{38} = 11.71$$

The mean value of X_1 is $M_1 = 31.92$. From equation (60),

$$F_3(X_3) = f_3(X_3) - M_{f(3)} + M_1$$

which is

$$F_3(X_3) = f_3(X_3) - 11.71 + 31.92$$
$$= f_3(X_3) + 20.21$$

All that is necessary, therefore, is to add the new constant, 20.2, to the values read from the curve. This process is shown in Table 65.

TABLE 65

Computation of Average Yield of Corn With Varying Rainfall, Holding Trend in Yield and Influence of Temperature Constant

Inches of rainfall, X_3	Readings from final curve,* $f_3''(X_3)$	Constant, $M_1 - M_{f(3)}$	Average yield, $F_3(X_3)$
7	6.0	20.2	26.2
8	8.2	20.2	28.4
9	10.5	20.2	30.7
10	12.3	20.2	32.5
11	13.4	20.2	33.6
12	13.2	20.2	33.4
13	13.0	20.2	33.2
14	12.7	20.2	32.9
15	12.5	20.2	32.7
16	12.3	20.2	32.5

* Curve readings minus 20, just as entered in TaLle 64.

The computation for $F_4(x_4)$ follows the same form as that for $F_3(X_3)$, save that equation (61) is used instead, and hence the mean of X_1 is not involved. First the mean of all the readings for $f_4(X_4)$, as shown in Table 64, is computed, giving the value of 11.81. The values for $F_4(x_4)$ are therefore given by the equation

$$\begin{aligned} F_4(x_4) &= f_4''(X_4) - M_{f(X_4)} \\ &= f_4''(X_4) - 11.81 \end{aligned}$$

These values are worked out in Table 66.

TABLE 66

COMPUTATION OF DEVIATION OF CORN YIELDS FROM YIELDS OTHERWISE EXPECTED, BECAUSE OF DIFFERENCES IN TEMPERATURE FOR SEASON

Average temperature, X_4	Readings from final curve,* $f_4''(X_4)$	Constant, $M_{f(4)}$	Correction to expected yield, $F_4(x_4)$
70.0	11.6	−11.8	−0.2
71.0	11.7	−11.8	−0.1
72.0	11.8	−11.8	0
73.0	12.0	−11.8	0.2
74.0	12.2	−11.8	0.4
75.0	12.3	−11.8	0.5
76.0	12.1	−11.8	0.3
77.0	10.5	−11.8	−1.3
78.0	8.3	−11.8	−3.5

* Curve readings minus 20, just as entered in Table 64.

The net correction in the estimated yield to allow for the influence of trend can be obtained by carrying through a similar computation for $F_2(x_2)$. The readings for $f_2''(X_2)$ sum to 447.6, so $M_{f(2)} = 11.78$. The values of $F_2(x_2)$ are then given by the equation

$$F_2(x_2) = f_2''(X_2) - 11.78$$

This computation is carried out in Table 67.

The conclusions of the study can then be stated as shown in the last column of each of the last three tables, free from all the previous details.

The relations for each of the variables can also be combined to show the expected or estimated yield for various combinations of the independent factors. Thus for the present case, it might be desired to combine the findings into a table showing the expected or probable yield for any given combination of rainfall and temperature, with the 1927 trend of yield. These values can be obtained by taking the trend correction for 1927, +0.4, and combining it with the estimated influence of various quantities of rain and degrees of temperature. These estimates would then be defined by the equation

TABLE 67

COMPUTATION OF DEVIATION OF CORN YIELDS FROM THOSE OTHERWISE EXPECTED, BECAUSE OF NET TREND IN YIELDS

Number of year, X_2	Date	Readings from final curve,* $f_2''(X_2)$	Constant, $M_{f(2)}$	Correction to expected yield, $F_2(x_2)$
0	1890	7.4	−11.8	−4.4
5	1895	9.5	−11.8	−2.3
10	1900	11.0	−11.8	−0.8
15	1905	12.0	−11.8	0.2
20	1910	12.7	−11.8	0.9
25	1915	13.5	−11.8	1.7
30	1920	13.5	−11.8	1.7
35	1925	12.6	−11.8	0.8

* Curve readings minus 20.

$$X_1' = F_2(x_2) + F_3(X_3) + F_4(x_4)$$

$$= 0.4 + F_3(X_3) + F_4(x_4)$$

Combining the readings for $F_3(X_3)$ from Table 65 with those for $F_4(x_4)$ from Table 66, and adding in the correction for $F_2(x_2)$ as just stated, we obtain estimated yields as shown in Table 68.[15]

In preparing a table such as Table 68, we should not enter values for combinations of the several factors which were not represented in the data on which the relations were based. Examination of a dot chart of the relation between rainfall and temperature, for the data included in the analysis, shows that no combinations of rainfall below 9 inches and temperature below 74° appeared in the record, and no cases of temperature above 78° with rainfall above 9 inches occurred. Accordingly, these combinations, and other combinations which were not represented, are left blank in the table, as shown. (A more exact

[15] Table 68 may be compared with the results secured by cross-classifying and averaging the same data, by the methods of Chapter 11.

method for measuring the representativeness of the relations is referred to in Chapter 19, on page 349.)

By combining a table such as Table 68 with a statement of the extent to which yields averaged higher or lower than those shown at different times through the period, all the conclusions from the study can be presented in simple form, easy to understand.

TABLE 68

Estimated Yield of Corn, in Bushels per Acre, With Varying Rainfall and Temperature Conditions, for 1927

Inches of rainfall *	Average temperature †				
	70°	72°	74°	76°	78°
7	‡	‡	27.0	26.9	23.1
9	30.9	31.1	31.5	31.4	‡
11	33.8	34.0	34.4	34.3	‡
13	33.4	33.6	34.0	‡	‡
15	32.9	33.1	‡	‡	‡

* Total for June, July, and August; average for 9 Corn-Belt stations.
† Average for June, July, and August, at same 9 stations.
‡ This combination of factors was not represented in the observations analyzed.

The final results of curvilinear correlation studies, after being simplified to the form shown in Tables 65 to 67, or in Table 68, may also be expressed graphically for final publication. Thus all three relations might be combined into a single figure, such as shown in Figure 40, to present in relatively simple form the final conclusions reached by the statistical analysis.[16]

It might be noted at this point that Table 68 is much more than merely a table of average yields for various rainfall and temperature groups. There were only 38 observations to begin with, and only 14 of those were under 74 degrees temperature. If these 14 observations had been grouped according to year and rainfall, and the average yield determined for each class, only the roughest sort of groups could have been made, and even then the averages would have had but little reliability. As the result of the correlation study, however, all 38 observations have been drawn on to determine the relations. The table shows the yield most likely to be received with any of 16 different

[16] A three-dimensional chart illustrating Table 68 is shown on page 373.

combinations of rainfall and temperature, for the trend in 1927. Other estimates could be shown for a large number of other combinations. Furthermore, it is known that estimates made from such tables agreed with the actual yields to within 2.8 bushels in about two-thirds of the original cases. The reliability of these estimated yields is thus greater than it would be for any average of a few cases alone. This example illustrates the ability of correlation analysis both to bring out of a series of observations relations which are not observable on the surface and to provide a basis for estimating the probable effect on the dependent factor of new combinations of the independent factors.

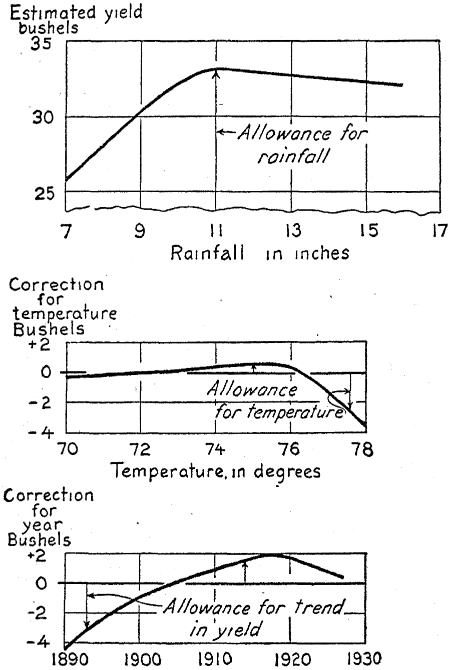


FIG. 40. Relation of yield of corn to rainfall, temperature, and time.

In this particular case the final shapes of the regression curves showing the net differences in yield with differences in rainfall and time are not greatly different from those indicated by simple correlation. In some cases, however, the final shape of the curves may be markedly different from the apparent shape before the variation associated with other factors has been eliminated. Thus the final shape of the curve showing the net differences in yield with differences in temperature, after allowing for the influence of rainfall and time, is quite different from what might have been expected from the original observations, as is illustrated in Figure 41. The curvilinear net regression is also quite different from the linear net regression, indicating that 74 to 76 degrees is the optimum temperature, whereas the straight line indicated that the lower the temperature, the higher the yield. With multiple correlation, as with simple correlation, the determination of the regression curves makes the results much more definite, adequate, and usable than does merely the determination of the linear regressions.

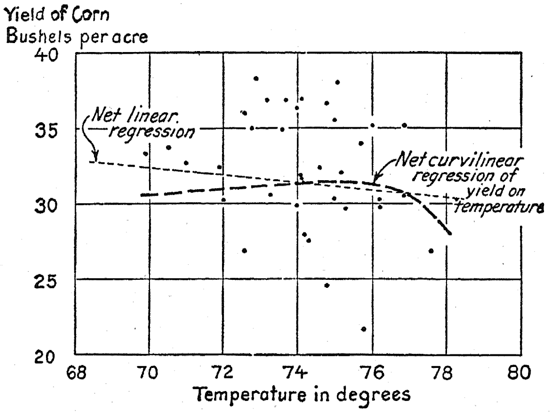


FIG. 41. Comparison of apparent relation of corn yields to temperature with net relation after eliminating influence of rainfall and of trend in yield.

Limitations on the use of the results. It should be noted that the results of the corn-yield analysis apply only to the same area from which the data were drawn and to the period which they covered. Thus they provide no basis for estimating corn yields in other sections, and their use in estimating yields in other periods—as in subsequent years—is attended by increasing risk due to the necessity of extrapolating the trend regression. Although this may give fair results for a year or two, as has been illustrated, it may tend to become increasingly inexact. For example, it may be that the trend of yield did not really turn downward about 1920, but only flattened out—additional years of observations will be needed really to tell which is correct.

Other multiple curvilinear correlation studies illustrate other limitations to the application of the results secured. Thus in a study of the price of eggs in New York City, records were secured during a period of a few days on the retail sales price of each of a number of dozens of eggs, and of the size, color, and quality of the eggs. (The data are given in the problem in Chapter 17.) By determining the net regression of price upon each of the factors, using the method illustrated, the net change in egg prices with changes in each of these factors can be determined. But it is readily apparent that size, quality, and color are not the only factors which might cause egg prices to vary. Prices change from one time of year to another, because of changes in seasonal demand, in supplies on the market, and in response to other factors as well. Prices also vary from place to place on the same day, and even at different stages in the marketing process in the same city on the same day—between sales at wholesale and retail, for instance. When we say that the results of the egg-price study enabled us to estimate egg prices to within five cents two-thirds of the time, it must be remembered that the statement holds true only for *the same universe from which the original samples were selected.* In this case the samples were all selected from sales at retail, in the New York metropolitan area, in a particular period. The results therefore apply only to the reasons for variations in egg prices between particular stores, in that particular city, in that particular period. They might indicate the effect of similar differences in quality or weight on prices from store to store in the same city at other times of the year, or in other cities; but we could not be certain of that from this material alone. Other studies, covering those other "universes," would be needed to prove or disprove that supposition; for the conclusions, of and by themselves, offer no statistical evidence except for their own particular "universe." For that reason, each of the final tables should indicate clearly the conditions to which its conclusions apply and thus definitely limit the statistical statement of the results to the particular conditions which they really represent.

A test in actual forecasting of yield. The two preceding paragraphs stand exactly as they were written in 1929. Now that this book is being revised (in 1941) the regressions based on the period from 1890 to 1927 can be given a severe test, by using them to estimate the yields during the subsequent 12 years. The necessary data for this estimate are given in Table 68A.

Estimates of yield for each of these years, according to the final curvilinear regressions shown in Tables 65 to 67 and Figure 40, are given in Table 68B, together with the residuals.

The new years included years of weather conditions more extreme than any experienced in the base years. It was, therefore, necessary to extrapolate the earlier curves in making the estimates. This was done by extending them with the same slope or curve as in the adjacent portions of the curve determined from the earlier data.

TABLE 68A

YIELD OF CORN, RAINFALL, AND TEMPERATURES IN SIX LEADING STATES, 1928 to 1939

Year	Time X_2	Rainfall in inches X_3	Temperature in degrees X_4	Actual yield in bushels X_1
1928	38	15.1	72.8	33.4
1929	39	10.6	73.4	31.5
1930	40	6.4	76.4	25.8
1931	41	10.4	76.9	32.7
1932	42	13.5	76.0	35.4
1933	43	7.2	77.3	29.4
1934	44	7.5	80.0	18.9
1935	45	9.6	76.2	31.7
1936	46	4.9	80.0	18.5
1937	47	10.1	76.6	36.4
1938	48	12.6	76.3	35.9
1939	49	11.7	75.8	41.1

Source: Computed from June, July, and August records for nine weather stations in Corn Belt states. Stations averaged include Kansas City, St. Louis, Toledo, Omaha, Peoria, Cincinnati, Topeka, Indianapolis, and the Iowa state average, as in the original study.

It is evident that the regressions gave fairly good estimates for the first few years of extrapolation, but thereafter gave increasingly large underestimates of the yield. It would appear that the introduction of hybrid seed corn, the possible improvement of cultivation with better machinery, the increase of soil fertility and the restriction of corn to the better fields with acreage-limitation and soil-conservation programs after 1933, and other factors, all combined to produce a new "universe," in which the corn yield to be expected for a given combination of weather became progressively higher than it had been in earlier years. Also, extremes of weather not previously experienced (such as the combination of an average temperature of 80° with a rainfall of 4.9 inches in 1936), which lay far outside the previous observations, apparently produced results somewhat different from those in the years analyzed. Even so, the estimates for the years of extreme conditions (1934 and 1936) were not extremely in error, as contrasted to other years in the last five. The doubts as to the correctness of the trend, as expressed in 1929, have been clearly confirmed by the subsequent data.

These actual results of extrapolation of a regression formula indicate the way that the conditions of a universe may shift and show the need of recalculating forecasting formulas for time series every year or two, to make sure that they are still applicable.

TABLE 68B

YIELD ESTIMATED BY CURVILINEAR REGRESSIONS ON THREE FACTORS, 1928 to 1939

Year	$F_2(x_2)$	$F_3(X_3)$	$F_4(x_4)$	X_1''	X_1	z'' $X_1 - X_1''$
1928	0.2	32.7	0.2	33.1	33.4	0.3
1929	0	33.4	0.3	33.7	31.5	−2.2
1930	−0.2	24.8	−0.3	24.3	25.8	1.5
1931	−0.4	33.2	−1.1	31.7	32.7	1.0
1932	−0.6	33.0	0.3	32.7	35.4	2.7
1933	−0.8	26.7	−1.9	24.0	29.4	5.4
1934	−1.0	27.4	−9.0	17.4	18.9	1.5
1935	−1.2	31.8	0	30.6	31.7	1.1
1936	−1.4	21.0	−9.0	10.6	18.5	7.9
1937	−1.6	32.7	−0.5	30.6	36.4	5.8
1938	−1.8	33.3	−0.1	31.4	35.9	4.5
1939	−2.0	33.5	0.4	31.9	41.1	9.2

The residuals for the first six years have a root-mean-square error $\left(= \sqrt{\frac{\Sigma(z'')^2}{n}} \right)$ of 2.7 bushels. This compares well with the standard deviation of 2.8 for the estimates for the 38 years included in the study. The next six years, however, had a root-mean-square error of 5.8 bushels. Since these latter errors were all in the same direction, the shift in the trend would appear to be primarily responsible for this increased unreliability.

(For an exercise in curve fitting by this method, the student can fit a set of regressions to the data for the whole period 1890 to 1939. Also, it would be valuable to fit separate regressions for the periods 1890 to 1920, and 1910 to 1940, and compare the two sets of results. Do they show a significant change in the relation of yields to the three factors?)

Reliability of Regression Curves

The regression curves show the net relation between the dependent variable and each independent variable, with the net variation associated with the other independent variables held constant, for the particular observations included in the sample. If another sample were drawn from the same universe, and similar net regression curves were determined, they would vary somewhat from the curves determined from the first sample. The lower the multiple correlation in the universe, or the smaller the sample, the larger would be this variation between successive samples. Methods have been developed for estimating the proportion of such samples which will give regression results falling within given ranges of the true regressions prevailing in the universe. (See Chapter 18, pages 327 to 340.) In publishing regression results, as shown in Tables 65 to 68, or in presenting charts of the regression results, such as shown in Figure 40, the reliability range of the regressions should be indicated, as shown subsequently. Even if the regressions (as in the example here) are determined from a time series, and so are based upon *all* the evidence for that portion of the constantly evolving universe, the reliability limits may still be used as an indication of possible significance, in view of the closeness with which the relations can be determined. (For a more extended discussion of the meaning of sampling errors with respect to time, see Chapter 19, pages 349 to 356.)

Summary. In this chapter methods of determining curvilinear multiple regressions have been discussed. These show the extent to which changes in the dependent variable are associated with changes in each particular independent variable, while simultaneously removing that part of the variation in the dependent variable which is associated (linearly or curvilinearly) with other independent variables. A method of determining the curves by successive graphic approximations is presented step by step. Since this method does not involve making definite assumptions as to the final shape of the curves, it is to be preferred to more mathematical methods, presented in a subsequent chapter, unless there is a logical basis for the choice of specific functions. Methods of simplifying the conclusions for popular statement are illustrated, and the universe to which they are applicable is briefly considered.

Correction Note.—On pages 239 and 247 the standard deviations of the residuals, σ_z, are used to determine whether the new regression curves show any gain in closeness of fit over the previous regressions. These comparisons can be made most accurately by using the *standard errors of estimate,* adjusted for n and m as explained on pages 208 and 261 (eqs. 42 and 65). The successive approximation process should be continued only until the adjusted standard error of estimate shows no further reduction.

CHAPTER 15

MEASURING ACCURACY OF ESTIMATE AND DEGREE OF CORRELATION FOR CURVILINEAR MULTIPLE CORRELATION

In presenting linear multiple correlation it was pointed out that coefficients could be computed to show (1) how closely estimated values of the dependent variable, based on the linear regression equation, could be expected to agree with the actual values; and (2) what proportion of the total observed variation in the dependent factor could be explained or accounted for by its relation to the independent factors considered. These coefficients were, respectively, the standard error of estimate and the coefficient of multiple correlation. Exactly parallel coefficients can be computed to show the significance of curvilinear multiple correlation, employing curvilinear net regressions such as those discussed in Chapter 14. The term "standard error of estimate" is again used to indicate the measure of the probable accuracy of estimated values of the dependent factor. In measuring the proportion of variation explained we will follow the usage in simple curvilinear correlation, and use the term "index" to denote the fact that curvilinear regressions have been employed. The proportion of variation accounted for is therefore shown by the "index of multiple correlation."

Standard error of estimate. In working through the various steps in determining the net regression curves by the method of successive approximations, in Chapter 14, the estimated values were subtracted from the actual values for each observation, and the resulting residual values, z'', z''', etc., were obtained. The standard deviations of these residuals were used as an indication of the accuracy of estimate for each set of curves. Where a very large number of observations is employed, such standard deviations of the residuals may be regarded as an indication of the extent to which estimated values of the dependent variable made from new sets of observed values drawn from the same universe may be expected to agree with the actual value of the dependent variable. Thus if we use $S_{1.f(2.3,4)}$ to designate the standard error of estimates of X_1, made on the basis of curvilinear relations to X_2, X_3, and X_4, and $z_{1.f(2,3,4)}$ to represent the residuals obtained using the final curvilinear regressions to estimate the dependent factor, the standard error may be defined by the equation

$$S_{1.f(2,3,4)} = \sigma_{z_{1.f(2,3,4)}} \tag{63}$$

If the standard error of estimate for the final regression curves for the egg-price problem mentioned in the previous chapter were 5 cents, that would mean that, if other purchases of eggs had been made in the same territory on the same day, it would have been possible to estimate the price to be paid for each dozen from their physical characteristics, to an accuracy indicated by that standard error. Two-thirds of the estimated values would probably have fallen within a range of 5 cents of the prices actually charged.

With the corn-yield problem, the standard deviation of the residuals from the last set of curves was 2.8 bushels. In this case no other "sample" can be drawn from the same "universe" except those included in the problem, for the universe was restricted to the years studied, 1890 to 1927. Extrapolating the trend line, however, it is fairly safe to say that estimates made for the same region for subsequent years can be expected to have a standard deviation of at least 2.8 bushels. If the trend used did not prove correct for subsequent years, the errors might be considerably larger.[1]

The relation shown in equation (63) holds exactly true only where there are a very large number of cases included in the sample dealt with. Where the sample is no larger than is usually available to the research worker, there is a tendency for the standard deviation of z to be somewhat smaller than the standard error which would be found in a very large sample drawn from the same universe. The smaller the number of observations, the larger the number of independent variables included, and the more complex the curves employed, the greater will be the tendency for the observed standard deviation to underestimate the true standard error. This may be illustrated by results from an experimental study of the stability of multiple curvilinear correlation results. In this case a universe of known correlation was employed, and successive samples were drawn of various sizes, repeating each drawing a number of times for the samples of each size. The curvilinear regressions were then determined for each sample separately by the successive approximation method, and

[1] This statement, written a decade ago, may be compared with the actual extrapolations made subsequently, as shown on pages 255 to 257 of the previous chapter.

the standard deviations were worked out for the residuals in each case. The entire analysis was then repeated, employing a universe of a higher correlation. The central values of these standard deviations of the residuals, for the samples of each size, were:

Number of observations	Observed standard deviation of z *	
	Universe 1	Universe 2
30	1.95	1.53
50	2.18	1.64
100	2.21	1.72
Entire universe	2.40	1.80

* These values are the median values observed.

It is quite evident from these results that the samples tended to give standard deviations smaller than that which actually was true for the universe as a whole and, further, that the smaller the sample employed, the greater the overestimate of the reliability of the estimated values.

It is therefore necessary to adjust the observed σ_z to give $\bar{S}_{1.f(2,3,\text{etc.})}$, which is an unbiased estimate of $S_{1.f(2,3,\text{etc.})}$ for the universe from which the sample was drawn. This adjustment is given in the following equation:

$$\bar{S}^2_{1.f(2,3,\text{etc.})} = \frac{\sigma^2_{z_{1.f(2,3,\text{etc.})}}}{1 - m/n} \qquad (64)$$

or

$$\bar{S}^2_{1.f(2,3,4,\text{etc.})} = \frac{n\sigma_z^2}{n - m} = \frac{\Sigma(z^2)}{n - m} \qquad (65)$$

Where n = number of observations in the sample
and m = number of constants represented (either mathematically or graphically) in the regression equation

It will be seen that equation (65) is exactly similar to equation (42) for the standard error of estimate in linear multiple correlation problems. For curvilinear problems, however, the value m has a somewhat different meaning. Thus in the experimental results just discussed, three independent factors were involved, so the regression equation was of the form

$$X_1 = a + f_2(X_2) + f_3(X_3) + f_4(X_4)$$

The corresponding linear regression equation would involve only four constants, so m would be equal to 4. For curvilinear regressions, however, at least two constants would be necessary to represent each regression curve, and possibly more. In the experimental study each curve had only one bend, either upward or downward. It was judged, however, that the curves could not be represented by second-order parabolas, since their shapes did not follow the smooth symmetrical curve which that type of function is capable of describing. Instead, it was judged that a third-order parabola would be necessary to give a fairly satisfactory fit to each regression curve. The conclusion was therefore reached that three constants would be necessary for a mathematical representation of each regression curve. On that basis the entire regression equation would represent approximately ten constants, three for each of the three curves, and one for the value a. (See pages 76 to 81 for other types of curves.)

Using 10 for m in equation (64), we may work out the value of $\bar{S}_{1.f(234)}$ for the smallest sample shown in the statement on page 261 as follows:

$$\bar{S}^2_{1.f(234)} = \frac{\sigma_z^2}{1 - \frac{m}{n}} = \frac{(1.95)^2}{1 - \frac{10}{30}} = \frac{3.80}{0.667}$$

$$= 5.70$$

$$\bar{S}_{1.f(234)} = 2.39$$

It is evident that this corrected value is much closer to the true value for the entire universe, 2.40, than was the original standard deviation of z.

Carrying the same adjustment through for the other values shown on page 261, we obtain standard errors of estimate as shown in the following statement.

Number of observations n	Value used for m	Universe 1		Universe 2	
		Observed σ_z	Calculated $\bar{S}_{1.f(234)}$	Observed σ_z	Calculated $\bar{S}_{1.f(234)}$
30	10	1.95	2.39	1.53	1.87
50	10	2.18	2.43	1.64	1.83
100	10	2.21	2.33	1.72	1.82
Entire universe	....	2.40	....	1.80	....

The superior accuracy of the adjusted values is evident throughout this table—in each case they come much nearer to agreeing with the true value for the universe than do the unadjusted values.

Using equation (64) to obtain the standard error of estimate for the corn-yield problem, we find it necessary first to decide on the value to use for m. That problem also employed three independent variables, just as did the experimental study, and the final σ_z was 2.80 bushels. Although none of the three regression curves has more than one bend, none of them is of the symmetrical shape that can be described by the parabola; instead, at least a cubic parabola would be required to represent the curves for $f_3(X_3)$ and $f_4(X_4)$, whereas probably a quartic parabola, involving four constants, would be required to represent $f_2(X_2)$ with its final shape, or three constants with its first form. The final regression equation for corn yields might therefore be assumed to represent one constant for a, four for $f_2(X_2)$, three for $f_3(X_3)$, and three for $f_4(X_4)$, or a total of eleven in all. When this value and the number of cases are inserted, in formula (64), it becomes

$$\bar{S}^2_{1.f(234)} = \frac{\sigma_z^2}{1 - \frac{m}{n}} = \frac{(2.80)^2}{1 - \frac{11}{38}} = 11.03$$

$$\bar{S}_{1.f(234)} = 3.32$$

Although the standard deviation of the observed residuals was only 2.8 bushels, this standard error of estimate indicates that, in using the results in making estimates for other years, the accuracy is likely to be less, even though the trend line is correctly extended. Instead of the estimated values probably coming within 2.8 bushels of the actual values in 68 per cent of the cases, they are likely to come so close in only about 58 per cent of the estimates, and an error of 3.5 bushels would have to be allowed to take in 68 per cent of the cases. In this particular problem, with 3 regression curves determined from 38 observations, the correction embodied in equation (64) is important. If the same set of conclusions had been obtained from 20 observations, with the same standard deviation of the residuals, applying the correction formula would have increased the standard error of estimate to above 4.1 bushels, illustrating again the tendency of a small sample to exaggerate the accuracy of estimate.[2]

[2] As is indicated later (Chapter 19, pages 341 to 347), each individual estimate for a new observation has its own standard error. Those standard errors are all larger than the standard error of estimate from the sample. The interpretation given above for the use of the standard error of estimate therefore understates the standard error for new observations.

Index of multiple correlation. The coefficient of multiple correlation, it will be remembered, indicated the proportion of the total variation in the dependent factor which could be accounted for on the basis of the linear relations to the several independent factors. In exactly the same way the proportion of variation which can be accounted for on the basis of the curvilinear relations to the several independent factors is termed the "index of multiple correlation," and is designated by the term P, that is, capital *rho.* Following the definition, and using X_1'' to indicate values of X_1 estimated from the other factors on the basis of the net curvilinear regressions, we may define the index of multiple correlation roughly by the equation

$$\mathrm{P} = \frac{\sigma_{X_1''}}{\sigma_{X_1}}$$

It is more accurately computed, however, by making use of the standard deviations of the residuals. Using z'' to represent $X_1 - X_1''$, then

$$\mathrm{P}^2 = 1 - \frac{\sigma^2_{z''}}{\sigma^2_1} \qquad (66.1)$$

With small samples $\sigma_{z''}$ tends to be smaller than the actual standard error of estimate in the universe as a whole. For that reason, the index of correlation, as computed by the formula just given, tends to exceed the correlation that actually obtains in the universe from which the observations are drawn. Data from the experiment mentioned earlier illustrate this point. The following tabulation shows the modal index of multiple correlation for the samples of each size, in comparison with the true index of correlation for the entire universe.

Number of observations in sample	Observed index of multiple correlation in samples drawn from same universe
30	0.77
50	0.71
100	0.68
Entire universe	0.62

In every case the observed correlation exceeds the true correlation in the universe, and the smaller the size of the sample, the larger the difference. It is therefore necessary to apply to the index of multiple correlation the same type of adjustment which was applied in obtaining the standard error of estimate, if unbiased estimates of the population value are to be obtained. This may be done either by substituting the adjusted standard error of estimate for the observed standard deviation of the residuals in the equation to determine P, or by making the adjustment directly in the equation itself. The following formulas show both methods.

$$\overline{\mathrm{P}}^2_{1.234} = 1 - \left[\left(\frac{\overline{S}^2_{1.f(2,3,4)}}{\sigma_1^2}\right)\left(\frac{n-1}{n}\right)\right] \tag{66.2}$$

$$\left.\begin{aligned}\overline{\mathrm{P}}^2_{1.234} &= 1 - \left[\left(\frac{\sigma^2_{z_{1.f(2,3,4)}}}{\sigma_1^2}\right)\left(\frac{n-1}{n-m}\right)\right]\\ &= 1 - \left[\left(\frac{\Sigma(z^2_{1.f(2,3,4)})}{\Sigma(x_1^2)}\right)\left(\frac{n-1}{n-m}\right)\right]\end{aligned}\right\} \tag{66.3}$$

The adjusted indexes of multiple correlation work out for the experimental data as shown in the following statement:

Number of observations, n	Value used for m	Crude, P	Adjusted, $\overline{\mathrm{P}}$
30	10	0.77	0.64
50	10	0.71	0.63
100	10	0.68	0.65
Entire universe	....	0.62	

Here again the adjusted values are found to be in much better agreement with the true value for the entire universe than are the crude values. For that reason equations (66.2) or (66.3) should always be employed in calculating the index of multiple correlation.

Unless the index of multiple correlation, as calculated with the adjustment, is larger than the coefficient of multiple correlation, with its comparable adjustment by equation (47), there is no statistical evidence of significant curvilinearity in the regression lines. Unless the standard error for the curves is lower even after adjustment, any reduction in the unadjusted standard deviation of $z_{1f(2,\,3,\,4)}$, as compared with σ_z from the linear regression, would be merely a fictitious improvement in accuracy. If we take additional variables into account, or use up more degrees of freedom by employing more constants in the curves, we obtain a certain amount of spurious increase in the apparent correlation. Correcting for n and m removes this spurious effect.

Once the index of multiple correlation has been computed by equations (66.2) or (66.3), the square of its value may be employed to represent the total determination, i.e., to measure the proportion of the total variance in X_1 which can be accounted for on the basis of the curvilinear relations to the several independent factors. To maintain the same terminology, this may be termed the *index* of total determination, to distinguish it from the *coefficient* of total determination, which applies to linear multiple correlation.

The computation of the index of multiple correlation may now be illustrated from the data of the corn-yield problem.[3] In that study the original standard deviation of the yields was 4.30 bushels, the standard error of estimate by linear multiple correlation, 3.87 bushels, and the coefficient of multiple correlation, after adjusting for the number of cases, 0.49. The standard error of estimate for the final regression curves, as worked out on page 263, was 3.46 bushels. Computing the index of multiple correlation by equation (66.2), we have

$$\begin{aligned}\overline{\mathrm{P}}^2_{1.234} &= 1 - \frac{\overline{S}^2_{1.f(234)}}{\sigma_1^2}\left(\frac{n-1}{n}\right)\\ &= 1 - \frac{(3.46)^2}{(4.30)^2}\left(\frac{37}{38}\right)\\ &= 0.369\\ \overline{\mathrm{P}}_{1.234} &= 0.61\end{aligned}$$

The index of multiple correlation is therefore 0.61, as compared with the coefficient of multiple correlation of 0.49. The total determination, whch was 24 per cent for the linear relation, has been raised to 37 per cent for the curvilinear. The increase indicates that the linear relations did not express all the effect of the three independent variables, and that taking the curvilinearity of the regressions into account has added significantly to the importance of the factors considered. With such a low determination, however, it is evident that there are other perhaps more important factors not yet taken into account.

Measuring the net curvilinear importance of individual factors. No method has been devised as yet to determine the portion of the index of total determination which can be ascribed to each of the several independent factors, solely from the methods used in obtain-

[3] See pages 225 and 227.

ing the several regression curves themselves. The final slope and shape of the curves may be tested, however, by correlating the curve readings for each observation with the original values of the dependent factor, so as to obtain the partial regression coefficients indicated in equation (89), and explained in Chapter 22.

$$X_1 = a' + b_{12'.3'4'}[f_2(X_2)] + b_{13'.2'4'}[f_3(X_3)] + b_{14'.2'3'}[f_4(X_4)]$$

If that is done, the coefficient of multiple correlation, $R_{1.2'3'4'}$, measures the total correlation with respect to the several curvilinear functions (including the final adjustments) and is therefore the index of multiple correlation, $\text{P}_{1.234}$. It is, however, still subject to the same adjustment for number of constants as are indexes of multiple correlation computed in other ways, and should therefore be corrected as follows:

$$\overline{\text{P}}^2_{1.234} = 1 - (1 - R^2_{1.2'3'4'})\frac{n-1}{n-m} \qquad (67)$$

Indexes of partial correlation can be determined with respect to the curvilinear regressions of the several independent variables, as shown in equation (89), in exactly the same way that the parallel coefficients of partial correlation are obtained. Since the curvilinear transformation relates solely to the net regression of X_1 on each of the independent variables, the meaning of the partial indexes with respect to the separate variables is open to some doubt.

Summary. For curvilinear multiple regression equations it is possible to obtain standard errors of estimate, indexes of multiple correlation, and indexes of partial correlation, which serve the same purpose that the comparable coefficients serve for linear multiple regressions. Owing to the extent to which the process of fitting the curves may exaggerate the significance of the results, it is even more important to adjust the several measures with respect to the number of observations and numbers of constants involved than it is with linear multiple correlation.

CHAPTER 16

SHORT-CUT METHODS OF DETERMINING NET REGRESSION LINES AND CURVES

In problems where the correlation is fairly high, the number of variables is not too large, and the number of observations is relatively small (say not over 50 to 100 cases), net regression lines and curves may be determined by a combination of inspection and graphic approximation which takes only a fraction of the time required by the methods previously presented in detail.[1] This graphic method is very speedy, and in the hands of a careful worker can yield results almost as accurate as those obtained by the longer methods previously set forth. It must be used, however, with the same regard to the meaning of correlation results, to the care in selection of material, and to the consistency of results with those logically expected as the other methods. It is subject to even more severe limitations with respect to the sampling variability of the results obtained from successive samples than are the other methods. For these reasons the student should first become thoroughly acquainted with the preceding methods, and their meaning and limitations, and then use this method only as a more rapid procedure for obtaining substantially the same results.

The general basis of the short-cut method is to select, by inspection, several individual observations for which the values of one or more independent variables are constant, and then note the change in the dependent variable for given changes in the remaining independent variable. This process is repeated for additional groups of observations for which the other independent variable or variables are constant (or practically so) but at a different level than for the first group. The relation between the dependent variable and the remaining independent variable, as indicated by a series of such groups, approaches the *net* regression line or curve, since the cases have been selected so as largely to cancel out the variation associated with other

[1] L. H. Bean, Applications of a simplified method of graphic curvilinear correlation, mimeographed preliminary report, U. S. Bureau of Agricultural Economics, April, 1929; and A simplified method of graphic curvilinear correlation, *Journal of the American Statistical Association,* Vol. XXIV, pp. 386–397, December, 1929.

independent variables. A first approximation line or curve is then drawn in by eye, and the residuals from this curve, measured graphically, are used to determine the regression for the next variable, cases again being selected so as to eliminate the influence of other independent variables. The final fit of the several lines or curves is tested by the same successive approximation process employed in Chapters 10 and 14, or by a shorter graphic equivalent of it. Since the initial lines or curves approach much more closely to the final net regressions, and since graphic transfers of residuals are substituted for curve reading and computation of the z's, the process is much shorter and fewer steps are required.

Linear net regressions. The short-cut method for linear regressions may be illustrated by the same farm-income problem utilized in Chapters 10, 11, and 12. The first step is to number each one of the observations as listed in the first four columns of Table 47, page 199, so that they may be distinguished from one another.

Preliminary examination of inter-relationships. The next step is to make dot charts of the intercorrelations of the *independent variables,* to see how they are related. Since there are three independent variables, X_2, X_3, and X_4, there are three sets of such intercorrelations—X_2 with X_3, X_2 with X_4, and X_3 with X_4. Dot charts for these combinations are shown in Figure 42. In entering these charts, we identify each observation by its own number, for future reference.

Examination of Figure 42 shows a moderate negative correlation between cows and acres and men and acres, and a slight positive correlation between cows and men. If charts such as these showed practically perfect correlation between any two independent variables—all the dots clustering closely together along a line or curve—that would be a warning that those two variables were so closely interrelated that it would be difficult or impossible to untangle the separate effects of each, regardless of what method was used. In such a case, one of the independent variables should be dropped, and the regressions found for the other variable should be stated as the relation of the dependent variable to the values of the independent variable retained *and the associated values of the independent variable which was excluded.* In this case, the intercorrelations are all low enough so that it will not be difficult to separate out the effects of each one.[2]

[2] Intercorrelation among the independent variables that is high but not perfect reduces the speed with which the successive approximations converge toward the best values, those which would be found by least squares. In such cases many more approximations may be required to get the best simultaneous fit.

The next step is to chart the values of the four variables for each observation in succession and connect them by lines just as if they were entries in a time series, as shown in Figure 43. (Classifying the records in order with respect to one of the independent factors before taking this step might be advisable.)

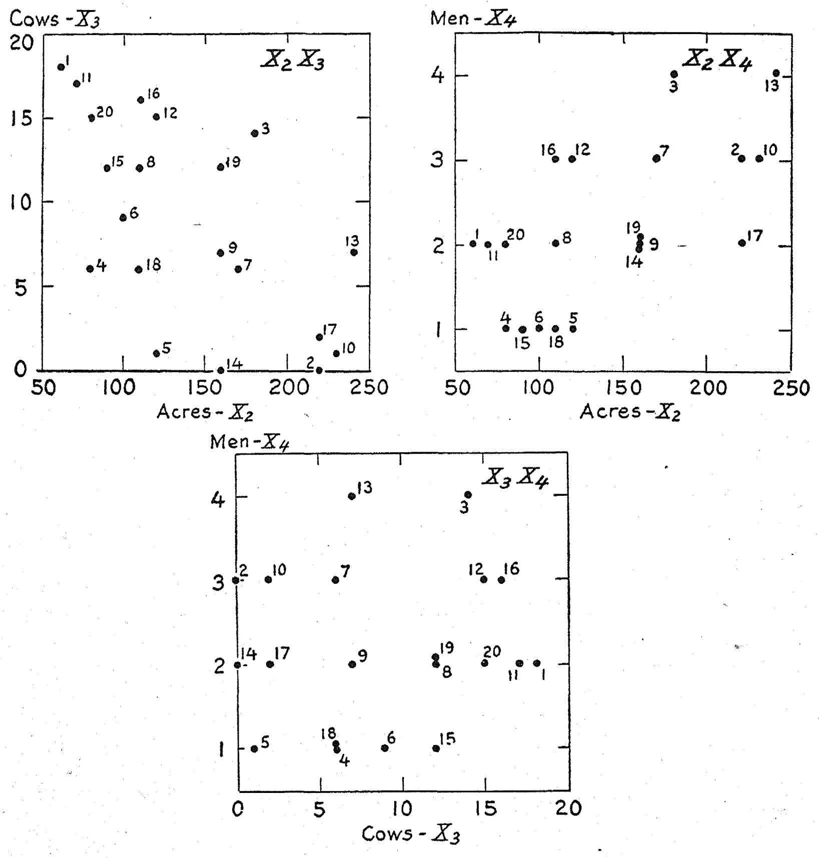


FIG. 42. Dot charts showing the intercorrelations of the independent variables, X_2, X_3, and X_4.

Comparing the different lines in Figure 43, we see that variation in incomes appears to be more closely associated with variations in cows than with either of the other factors. (Dot charts of X_1 with X_2, X_1 with X_3, and X_1 with X_4, might be used instead to reach this conclusion.) The relation of X_1 to X_3, number of cows, *for constant numbers of acres and men,* will therefore be examined first.

Determination of first approximation regression lines. From Figure 42 we note that of the farms with the largest numbers of acres, both farms 2 and 10 have 3 men employed, whereas farms 13 and 17 have 4 and 2, respectively. Accordingly we plot the cows and incomes for these farms on a new dot chart as shown in Figure 44, indicating the number of the farm represented by each dot, and using solid dots. The placing of these dots does not seem to indicate any marked relation of income to the number of men; we therefore draw in a straight line freehand, to fit approximately the change in income with changes in cows, as shown by these four observations. (The values may be taken from Table 69, page 277.)

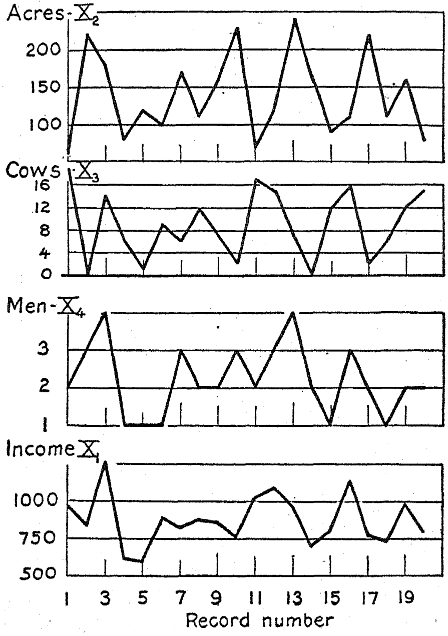


FIG. 43. Acres, cows, men, and income, on 20 farms.

Turning to the small farms, on the X_2X_4 section of Figure 42, we note that farms 6, 15, and 18, each with between 90 and 110 acres, have 1 man apiece; and farms 8, 11, and 20, with 70 to 110 acres, have 2 men apiece. Plotting the corresponding observations as hollow dots on Figure 44, again we have little evidence of any influence of the differences in number of men. The other small farms, 4, 5, and 16, are accordingly plotted, and a line, estimated graphically to pass through the nine observations as well as possible, is drawn in as shown.

Finally, it is noted that farms 7, 9, 14, and 19 all have 160 to 170 acres, so these are plotted on Figure 44 as crosses, to distinguish them. The differences in the number of men are ignored at this step, since they have been found to have little apparent relation to the income in the previous cases, and a line is drawn through these last cases, as indicated.

Comparing the three lines, we see that all have about the same slope, so a single line is drawn in to pass through the intersection of the averages of cows and of income, with a slope averaging the slope of the other three lines. This last line is the first approximation to the net regression of income on cows, with acres and men constant. The dots for the remaining farms, numbers 1, 3, and 12, are then plotted in, with the numbers to indicate their identity.

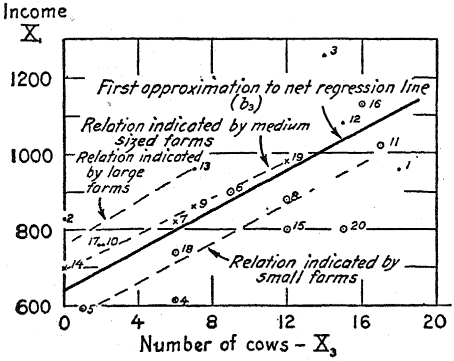


FIG. 44. Income plotted against cows, on specified farms, and first approximation to net linear regression on cows.

For the next step, a blank chart is prepared, as shown in Figure 45, to show the relation between acres, X_2, and the departures of income, X_1, from that expected on the basis of the approximate regression on number of cows. This chart is completed by scaling off the vertical departure of each .observation in Figure 44 from the approximation line, and then plotting that departure in Figure 45 as a departure from the zero line, with the number of acres for the same observation as abscissa.[3] The identity of the observation represented by each dot is again shown by its number. Here, to aid in identifying observations according to the other independent variable,

[3] For a convenient and speedy method of scaling off and transferring these departures graphically, see pages 479 to 485.

solid dots have been used for farms with 1 man, circled dots for farms with 2, and crosses for farms with 3. The 2 farms with 4 men are also shown as solid dots. The relation of acres to income is now clearly evident (in fact, were this not a discussion of linear correlation, fitting a curve would seem to be justified). It is next noted that farms 4, 5, 6, 15, and 18 have but 1 man apiece. Accordingly a line is dotted in to pass as near the dots for these farms as possible. Farms 2, 7, 10, 12, and 16 have 3 men each, so a line is fitted to them graphically, as indicated. Farms 1, 8, 9, 11, 14, 17, 19, and 20 have 2 men each, so they are designated by enclosing each of them with a circle, and a line is fitted freehand to them. All these lines are of somewhat the same slope, so a final line is drawn in by eye, averaging the slope of the other lines and intersecting the zero line at the abscissa corresponding to the average number of acres. This line is the first approximation to the regression of income on acres determined while holding constant the approximate effects of both cows and number of men.

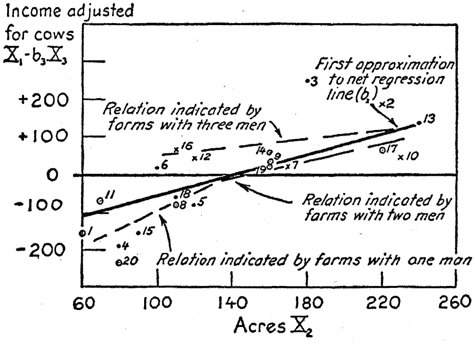


FIG. 45. Income adjusted for cows (by first approximate regression), plotted against acres on specified farms, and first approximation to net linear regression on acres.

The next step is to prepare a chart for number of men and adjusted income, as shown in Figure 46. The deviations of the individual observations from the approximate regression line in Figure 45 are measured graphically, and plotted in as deviations from the zero line in Figure 46, with the number of men for each observation as abscissa. The placing of these dots indicates a tendency for income to increase with number of men. The average adjusted income for each number of men is determined by inspection, and indicated by the small circles. Then a straight line is fitted by eye so as to intersect the zero line at the average of X_4, and fit these averages as well as possible.

Determination of second approximation net regression lines. The next step is to check the slope of the previous approximate net regression lines, to see if any changes are needed, now that the effect of other factors has been more accurately allowed for. To do this, the line from Figure 44 is drawn in on Figure 47. The deviations of each of the observations in Figure 46 are then scaled off graphically, and plotted in Figure 47 as vertical deviations from the line, with the number of cows, X_3, as abscissa. The plotting of these deviations indicates that a slightly steeper line might fit better, since it is found that, although in the range 0 to 2 cows, 2 dots fall below the line whereas 3 fall above, in the range 14 to 18, 4 out of 6 dots fall above the line, and in the range 6 to 8 cows, 3 of the 5 observations fall below the line. Accordingly a revised line is drawn in free hand, passing

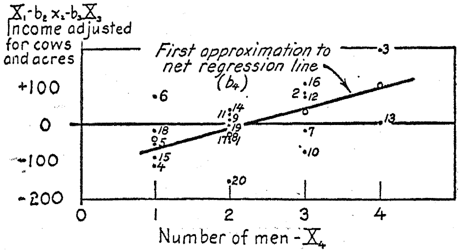


FIG. 46. Income adjusted for cows and acres (by first approximate regressions), plotted against number of men on specified farms, and first approximation to net linear regression on men.

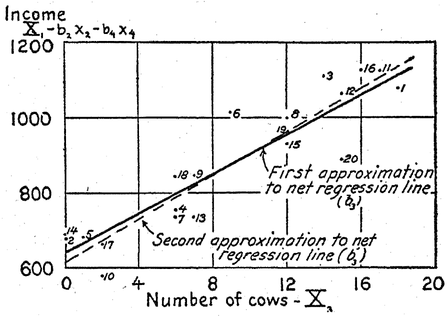


FIG. 47. Income adjusted for acres and men (by first approximate regressions), plotted against cows, and first and second approximations to net regressions on cows.

through the intersection of the averages of cows and income as before, and fitting the new dots as well as possible. The first line for the regression of income on acres is then checked in the same manner, by plotting the deviations from the new line in Figure 47 as deviations from the first approximate regression on acres (Figure 45). This process, carried out by graphic plotting just as before, is shown in Figure 48.

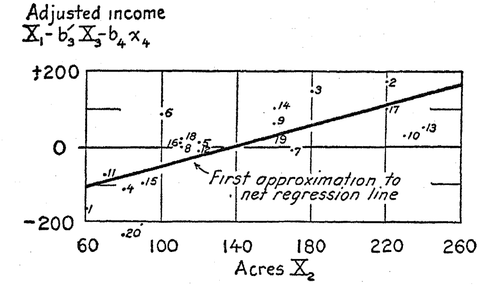


FIG. 48. Income adjusted for cows (by second approximate regression) and men (by first approximation), plotted against acres.

The distribution of the dots in Figure 48 shows that the observations are so nearly evenly balanced about the line now that no further change in the line is necessary. It is evident that a curve would fit better than the straight line, but for the present we are considering linear relations only.

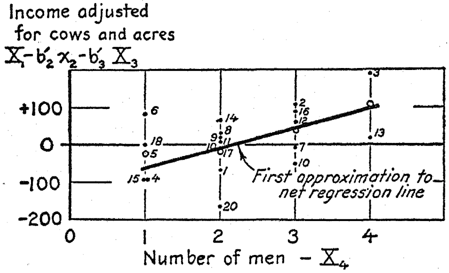


FIG. 49. Income adjusted for cows and acres (by second approximate regressions), plotted against men.

Since no change has been made in the regression for X_2, all that remains is to check the first line for the regression on X_4, using the deviations from the line in either Figure 47 or in Figure 48. Plotting these deviations graphically as before, above or below a line with the same slope as in Figure 46, gives the result shown in Figure 49. Since this figure shows no significant change from Figure 46, the line is left unchanged, and the lines on Figures 47, 48, and 49 are accepted as giving the approximate values for $b_{13.24}$, $b_{12.34}$, and $b_{14.23}$, respectively. If the increases in income per unit change are calculated from these lines they come out 29.2 dollars per cow, 1.34 dollars per acre, and 52.7 dollars per man, as contrasted to the exact values of 26.3, 1.21, and 50.3, worked out in Chapter 12. Although the values are not identical, they are quite close—so close, probably, that the differences between them have no statistical significance in view of the small number of observations on which they are based. (If a larger number of succesive approximations were used, and the average residuals were computed at each step as a guide to the new lines, the final values would come even closer to the exact values.)

Estimating values of dependent variable. The estimated income may now be worked out for each farm, either by taking readings directly from each curve or by substituting the approximate values found for the regression coefficients in equation (39) to determine a, and then working out the estimates mathematically. In either case, the correlation and standard error could be computed only by working out the estimated values, calculating the residuals and their standard deviation and substituting those in equations (42) and (48). The process of computing the estimates by using values read directly from the figures is shown in Table 69.

Calculating standard error of estimate and multiple correlation. The standard deviation of the z's computed in Table 69 is 69.06. By substituting this value in equations (42) and (48), the standard error of estimate and the multiple correlation work out as follows:

$$\bar{S}^2_{1.234} = \frac{n\sigma_z^2}{n - m} = \frac{20(4{,}632)}{16} = 5{,}790$$

$$\bar{S}_{1.234} = 76.09$$

$$\bar{R}^2_{1.234} = 1 - \frac{\bar{S}^2_{1.234}}{\sigma_1^2}\left(\frac{n-1}{n}\right) = 1 - \frac{5{,}790}{27{,}276}\left(\frac{19}{20}\right) = 0.798$$

$$\bar{R}_{1.234} = 0.893.$$

The new standard error of $76.09 compares with that of $74.65 obtained by the regular least-squares method, and the multiple correlation of 0.893 by the approximation method compares with the value 0.898 obtained by the more exact method. As indicated by these slightly lower coefficients, the approximation method is not quite so precise, yet for most practical purposes the results are nearly the same.[4]

The short-cut method applied to curvilinear regressions. The greatest usefulness of the short-cut method is in determining net curvilinear regressions. Since the method of successive graphic approximations presented in Chapter 14 also depends on the convergence of successive approximate curves, the short-cut method secures results which are exactly as reliable, at a great saving of time.

TABLE 69

CALCULATION OF ESTIMATED INCOME FROM LINEAR REGRESSIONS DETERMINED BY APPROXIMATION METHOD

Number	X_2 Acres	X_3 Cows	X_4 Men	X_1 Income	$f_2(X)_2$	$f_3(X_3)$	$f_4(X_4)$	X_1'	z
1	60	18	2	960	−106	1,134	−11	1,017	− 57
2	220	0	3	830	+110	612	+42	764	66
3	180	14	4	1,260	+ 56	1,022	+94	1,172	88
4	80	6	1	610	− 80	789	−62	647	− 37
5	120	1	1	590	− 26	641	−62	553	37
6	100	9	1	900	− 52	876	−62	762	138
7	170	6	3	820	+ 43	789	+42	874	− 54
8	110	12	2	880	− 39	964	−11	914	− 34
9	160	7	2	860	+ 29	818	−11	836	24
10	230	2	3	760	+123	670	+42	835	− 75
11	70	17	2	1,020	− 93	1,110	−11	1,006	14
12	120	15	3	1,080	− 26	1,051	+42	1,057	23
13	240	7	4	960	+136	818	+94	1,048	− 88
14	160	0	2	700	+ 29	612	−11	630	70
15	90	12	1	800	− 66	964	−62	836	− 36
16	110	16	3	1,130	− 39	1,080	+42	1,083	47
17	220	2	2	760	+110	670	−11	769	− 9
18	110	6	1	740	− 39	789	−62	688	52
19	160	12	2	980	+ 29	964	−11	982	− 2
20	80	15	2	800	− 80	1,051	−11	960	−160

[4] In fact, the differences between the values obtained by exact solution and those obtained by the approximation method are no larger than might readily occur by chance if the mathematical analysis were repeated on a second sample of the same size, to judge from the standard errors of the three regression coefficients, when computed by the methods explained in Chapter 18.

The procedure will be illustrated by a problem of four variables. The same method may be applied to larger or smaller problems equally well.

The data to be considered are:

TABLE 69A

DATA FOR SHORT-CUT METHOD OF DETERMINING REGRESSION CURVES*

Year X_4	Cost per ton of finished steel X_1	Proportion of capacity operated X_2	Average hourly earnings X_3
	Dollars per ton	*Per cent*	*Cents per hour*
1920	72.3	88.3	77.5
1921	78.5	47.5	60.2
1922	57.9	71.3	58.5
1923	63.0	88.3	67.0
1924	63.7	69.0	70.8
1925	62.9	78.4	70.3
1926	60.3	88.0	70.8
1927	59.6	78.9	71.3
1928	55.2	83.4	71.8
1929	51.5	89.2	72.5
1930	58.6	65.6	73.2
1931	65.6	38.0	70.8
1932	81.4	18.3	61.0
1933	65.0	28.7	59.0
1934	64.6	31.2	70.0
1935	65.4	38.8	73.0
1936	61.1	59.3	74.0
1937	65.6	71.2	86.0

[*] The data are calculated from regular published reports of the U. S. Steel Corporation. See Kathryn H. Wylie and Mordecai Ezekiel, The cost curve for steel production, *Journal of Political Economy*, Vol. XLVIII, pp. 777–821, December, 1940.

Data for 1938 and 1939 are also available, but we shall disregard them until the analysis is completed, and then use them for checking the results.

Logical relation of the variables. These data are from a study of the relation of volume of steel output to cost per ton. The qualitative examination of the problem (see discussion in publication cited in the footnote to Table 69A) indicated that changes in wage rates might be expected to have a relative, or multiplying, effect upon the cost for a given output, so that the relation might best be examined in terms of:

$$\log X_1 = f_2(X_2) + f_3(X_3)$$

Also, the qualitative examination revealed that major changes in technical methods of production, especially the beginning of the substitution of continuous-strip mills for hand mills, had taken place during the period under consideration, and that these improvements in technology might need to be included, either directly as a labor-efficiency factor or, indirectly, as a trend factor.

To simplify this illustrative presentation, the data will be used in absolute values, instead of using the logarithms. The charts will be examined for indications of multiplying relationship, however, since (as is shown in detail on page 296) this graphic method can also be used to spot the presence of such non-additive relations.

Conditions on the curves to be drawn. Before proceeding to the statistical steps in the examination of these data, the types of curves logically expected and the resulting conditions to be placed upon the shapes of the curves to be obtained must also be considered. Without going into the underlying technical reasons (presented more fully in the original study), let us assume that the following conditions will be imposed:

On the net relation of cost to capacity:

1. The curve may fall, at a declining rate, until a minimum is reached, and may then increase gradually after that minimum is passed. No points of inflection are expected.

On the net relation of cost to wages:

2. The curve will rise steadily, possibly at an increasing rate with higher wages, but otherwise will be fairly uniform—that is, will be either a straight line or a shallow curve concave from above. There should be no inflections.

On the net relation of cost to the time elements (efficiency, etc.):

3. The curve will tend to decline, perhaps slowly at first and then more and more rapidly as new techniques are introduced. There might also be irregular changes reflecting the changes in general price level (and in various purchased materials and services other than labor) during the period under examination, especially in the early 1920's and after 1929. (Note how this trend factor lumps together labor efficiency, price levels, and perhaps other factors, each of which might be given separate consideration in a more elaborate investigation.)

Preliminary examination of inter-relationships among the independent variables. As before, the inter-relationships of the several independent variables (including time for the trend factor) must be examined before the short-cut approximations can be begun. These are presented in Figure 50, the years being used to designate the observations. After the dots were located, the successive years were connected by a light line, making it possible to consider the relations of X_4 (time) to X_3 and X_2, as well as of X_2 to X_3, all on this one chart. (This same method could be used even in non-time-series data by first classifying the data on the ascending values of one independent variable. Successive observations, by number, would then indicate increasing values for that variable.)

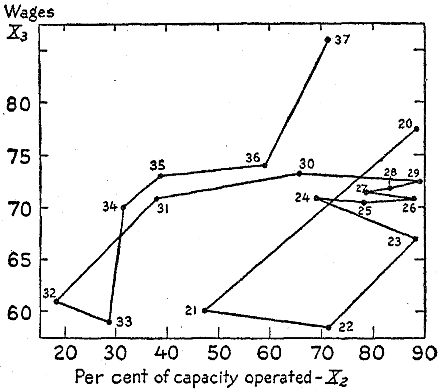


FIG. 50. Wages and per cent of capacity operated, with successive observations connected to indicate shift in the X_2X_3 relationship with time.

Examining first the location of the dots in Figure 50, without regard to their sequence, a moderate intercorrelation between wages (X_3) and rate of operations (X_2) is evident. No low values of X_2 are found, except together with low values of X_3. In the higher ranges of X_2 the values of X_3 fan out more, varying from quite low to quite high. Apparently there is enough independence in the occurrence of the two variables to permit of fairly good separation of their effects.

When examined with regard to time, however, the independence is not so good. The low wages at high output all occurred in one period—1921 to 1923. The marked positive correlation of wages and operations from 1930 to 1937 is also a correlation with time, both generally declining from 1930 to 1933, and both rising from 1933 to 1937. Since this was the period when technological changes were greatest, it may be difficult to disentangle the time or trend elements here, reflecting these technological changes, from the effects of the associated advances in output and in wages. We shall have to be on guard for this as we proceed with the analysis.

Looking for groups of observations which hold the other factor constant, we note on Figure 50 that there were a considerable number of years when wages [5] fell between 70 and 75 cents per hour. These observations for these years may be used to hold wages substantially constant, while the data are examined for the apparent effects of operation rate and time.

Determination of first approximation curve for first independent variable. The observations for the years with wages of 70 to 75 cents are accordingly plotted on Figure 51 with percentage capacity operated (X_2) as the abscissa and cost per ton (X_1) as the ordinate.[6] After the dots are plotted, successive observations (when they occur in this group) are connected by light dotted lines. This enables us to examine the relation of cost to operation rate and time while holding wages constant.

These observations indicate at once a marked negative correlation between operation rate and cost. The data from 1924 to 1929 suggest a rapid fall in cost for a given rate, especially from 1927 to 1929. Apparently there was some further decline from 1931 to 1934, but the data for 1935 to 1936 fall almost precisely on those for 1930 to 1931. (However, examination of Figure 50 shows that wages were slightly higher in this latter period, which might obscure the trend factor at this point.) No curve is indicated as yet. Accordingly, a line is drawn in lightly, as indicated, to show the relation of cost to operation rate for these observations, with the trend factor also considered.[7]

[5] "Wage rates per hour" is quite a different thing from "average earnings per hour employed," since the latter is a weighted figure reflecting all changes in the composition of the labor force. The latter is the figure used here (note Table 69A), since an average wage-rate figure was not available. For brevity, however, the term "wages" will be used here to describe the data, even though that is not the technically correct designation.

[6] Great care should be exercised in plotting these values, as their exact location becomes the basis for all the successive graphic transfers. Chart paper of adequate size to separate the dots should be used.

[7] By drawing this line parallel to the lines connecting successive years, all trend is eliminated except the one-year change. If the line were tilted slightly steeper than the line connecting successive years, that would provide an approximate correction for the year-to-year change, also. With the uncertainty of trend effects after 1931, however, that was not done here, but was left for subsequent approximations to clarify.

The observations for years of very low wage rates—1921, 1922, 1932, and 1933—are next plotted, and consecutive years again connected by dotted lines. Both show exaggerated drops in costs with increases in output. Only 1933 shows a cost lower than might be expected from the observations previously plotted. If 1932 were also to show a cost below the usual relation, the regression curve would have to swing up sharply, so as to pass above it. The high value for 1921 may be ignored for the moment, as possibly reflecting the high price levels at the end of the first World War inflation.

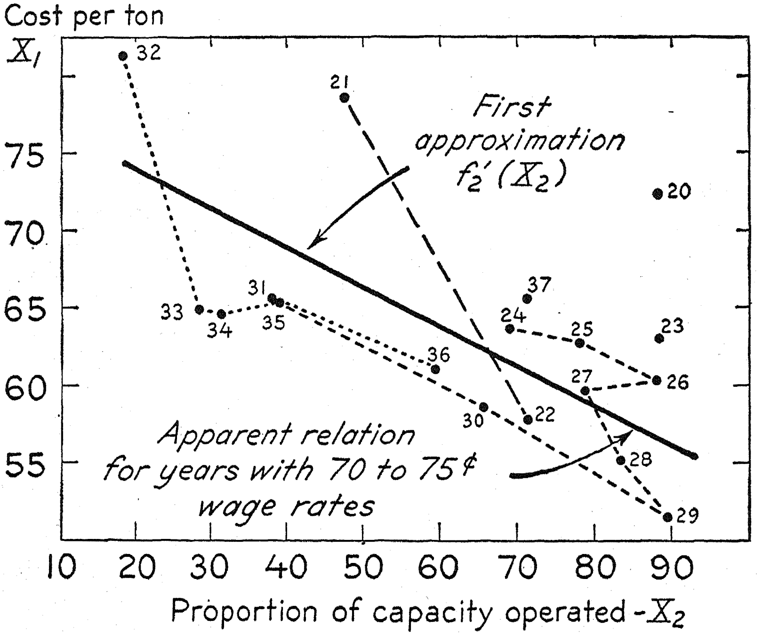


FIG. 51. Cost per ton and per cent of capacity operated, and first approximation to $f_2(X_2)$.

The two years of high wages—1920 and 1937—and the one remaining year of moderately low wages, 1923, are next plotted. The dot for 1937 falls above the other observations, and that for 1920 much higher still, apparently confirming the unusual (trend?) factors affecting the position of the 1921 observation. Similarly 1923 is fairly high, despite its moderate wage rate, as compared to subsequent years.

The evidence as to wage rates, to this point, sums up as follows: 1920 to 1923 all show relatively high costs (with the exception of

1922). Apparently trend elements outweighed the effects (if any) of the low wages in 1921 and 1923. With low rates, 1933 shows quite a low cost for the low rate of output, whereas 1932, with somewhat higher wage rate, shows a much higher cost. Apparently the fall in output to near zero increases cost very greatly per unit. On the basis of these considerations, a curve could be drawn in as the first approximation, extending the previous line but bending it up to pass well above 1932, with its low wage rate. With only one or two observations to support that bend at this stage, it seems best to be more conservative until the other factors have been more definitely allowed for, and until the evidence for a curve (if any) is more clearly established (even though a curve of declining costs was expected.)

Accordingly the straight line previously drawn in lightly is extended and used as the first approximation toward the net regression, $f_2'(X_2)$. (If a curve had been clearly indicated by the examination of the data as described above, it would have been drawn in at this point, thus starting the successive approximations from a curve instead of from a straight line.)

Determination of first approximation curve for second independent variable. The next step is to examine the relation of costs, as now approximately corrected for the relation to operation rate by $f_2'(X_2)$, to wages and time. Accordingly, the vertical departures of the dots on Figure 51 from the line of $f_2'(X_2)$ are scaled off, and are plotted in Figure 52.[8] The departures are plotted as ordinates, with the values of X_3, wages, as abscissas. If the fourth variable, X_4, were not a time series, or not arranged in order, it would be necessary to group these observations according to its value, also, as was done in plotting Figure 51. Since the numbers of the successive years indicate the successive values of X_4, that is not necessary. After the dots are all plotted, the successive years are connected by a light dotted line, to aid in separating the trend influences from that of wages.

If the dotted line to the successive years is followed, it is apparent that there was a general downward trend in the adjusted costs. The years 1920 and 1921 appear on one level, the years 1922 to 1927 on a lower level, and the years from 1928 on (with the exception of 1932) on a still lower level. In each of these groups of years there is a positive relation between adjusted costs and wages, as indicated by the light lines drawn through each group. Only the last group has any

[8] As with the linear short-cut method, the job of making these readings and transfers can be made swifter and more accurate by using the technique outlined on pages 479 to 485.

indication of a curve. Even there, the curve depends entirely on the position of the two extreme observations, one at each end. Here, however, the lower portion of this curve parallels, almost exactly, the lines indicating the apparent positions for the two other groups, which in turn lie mainly on the left half of the lower group of observations. Furthermore, the shape of the curve—shallowly concave—is consistent with that logically expected. Accordingly, a shallow curve passing through the center of the observations is drawn in, approximately paralleling the apparent lines and curve representing the relations for the three groups. The succeeding successive approximations will show

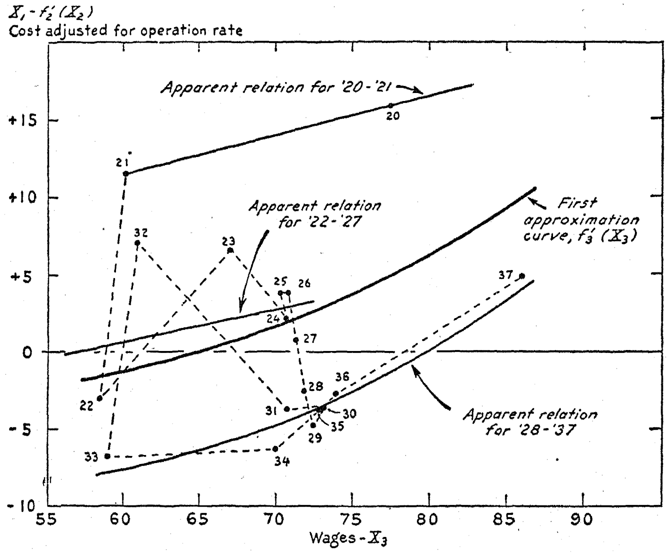


FIG. 52. Wages and cost per ton adjusted to average operation rate on the basis of the first approximation, and first approximation to $f_3(X_3)$.

whether this curve is justified or whether a straight line should be substituted.

Determination of first approximation curve for third independent variable. The next step is to examine the relation of costs, now approximately adjusted for both wages and operation rate, to time. Accordingly, the vertical departures of the dots on Figure 52 from the curve $f_3'(X_3)$ are scaled off, and are plotted in Figure 53. Again the departures are plotted as ordinates, with this time the values of X_4 as abscissas. Since this is the last independent variable to be considered, it is not necessary to group the observations with respect to any other variable but all can be plotted and examined as a whole.

Figure 53 shows the resulting chart. Connecting the successive years makes it easier to study the type of trend present.[9]

Except for the single wide departure in 1932, Figure 53 indicates a definite downward trend from the beginning, tapering off about 1930 and running flat or gradually rising thereafter. Taking midpoints between each pair of observations (indicated by the crosses) helps to locate the approximate level of this trend. The one extreme departure, 1932, is disregarded in the process. Its position in Figure 51, at the extreme end of the line, meant that its adjustment for X_2 was in doubt. A smooth curve is then drawn in, declining to about 1930, and running flat thereafter. The rising trend indicated by the observations for 1936 and 1937 is left for subsequent approximations to confirm. In general it is unwise to give an extra "twist" to a regression curve simply on the evidence of one or two observations.

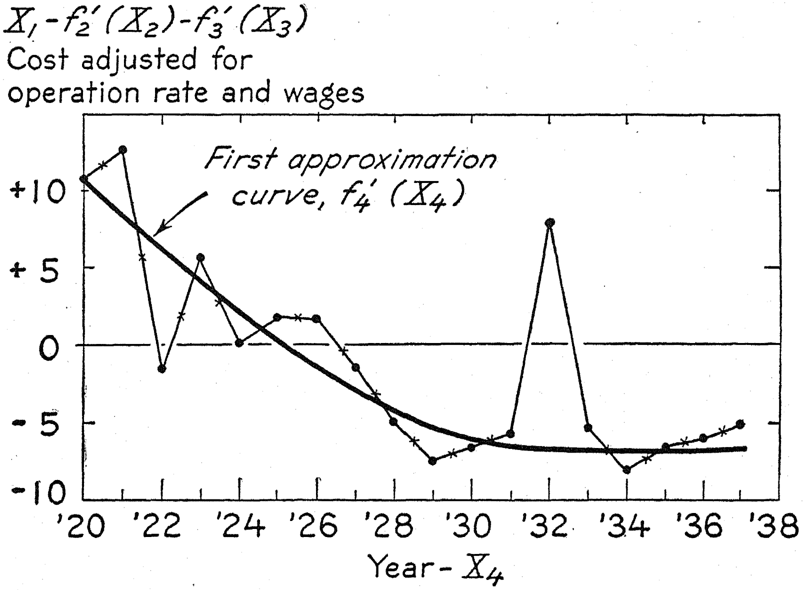


FIG. 53. Time and cost per ton adjusted to average operation rate and wages, on the basis of the first approximation curves, and first approximation to $f_4(X_4)$.

[9] If joint functions are suspected (see Chapter 21) the data might again be grouped for values of X_2 and X_3, in plotting Figure 53. If these groups showed varying relations to X_4, even after the approximate relations to X_2 and X_3 had now been eliminated, that would indicate the presence of a joint relation. Note Figure 57, and the discussion on pages 296 to 299 of this chapter.

Determination of second approximation curve for first independent variable. We now have determined first approximation lines or curves to the net regressions of X_1 on X_2, X_3, and X_4. The departures of the dots on Figure 53 from the regression line $f'_4(X_4)$ are the residuals, z'', from this first set of curves. The remaining steps involve the graphic transfer of these residuals to each curve in turn, the correction of each curve on the basis of the fit of the new residuals, and in turn the transfer of the newly corrected residuals to the next curve, and so on until no further change is indicated in any of the curves. Ordinarily the residuals from Figure 53 would be plotted back on the original curve for X_2, Figure 51. To show the process clearly, however, the dots and the first approximation curve for $f'_2(X_2)$, from Figure 51, are reproduced again as Figure 54.

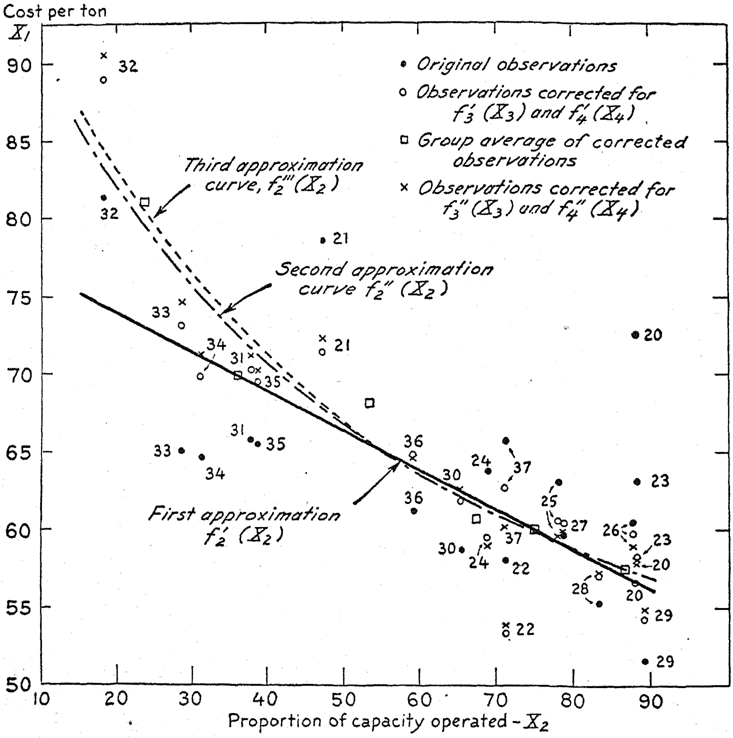


FIG. 54. Per cent of capacity operated, and cost per ton unadjusted and adjusted to average values of other variables, and second and third approximations to $f_2(X_2)$.

The vertical departures of the dots on Figure 53 from the approximation curve, $f'_4(X_4)$, are then plotted on Figure 54 as departures above and below the regression line, $f'_2(X_2)$, with the corresponding values of X_2 as abscissas. To prevent confusion with the original values shown as solid dots, the corrected values are indicated as hollow dots.

It is at once apparent, on inspection of Figure 54, after the corrected values are all plotted in, that the new values show much less scatter than the original values. Closer inspection reveals that every one of the adjusted observations below 60 per cent of capacity falls *above* the first approximation line, with a single exception. In the range from 60 per cent to 80 per cent, three cases fall below the first approximation line (two widely) and three slightly above, indicating in this range that the new line should be lower than before. The five observations above 80 per cent fall two below, two about the same distance above, and one right on the line, indicating that the position of the line here is about correct. These departures confirm the suggestion previously given by the 1932 value in Figure 51 that the regression should be a curve, concave from above. This accords, also, with the logical conditions originally imposed on this relation. Accordingly such a curve is drawn in freehand, passing as near as possible through the averages of the adjusted values in each successive group. (To facilitate drawing the curve, the average of the residuals in successive

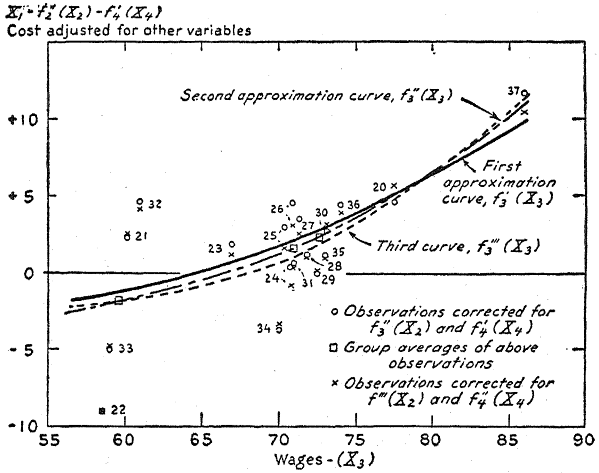


FIG. 55. Wages, and cost per ton adjusted to average values of all other variables, and second and third approximations to $f_3(X_3)$.

ranges of 10 to 15 units of X_2 are estimated graphically and drawn in as hollow squares.)

Determination of second approximation curve for second independent variable. The vertical departures of the adjusted values (the hollow dots) above or below the second approximation curve, $f_2''(X_2)$, are next scaled off graphically and plotted as ordinates from the values of the $f_3'(X_3)$ curve, as zero, with the corresponding X_3 values as abscissas. This is generally done on the original X_1X_3 chart (Figure 52). For clarity, however, the curve of Figure 52 is here reproduced on Figure 55, and the departures from Figure 54 are transferred to this new chart. The four observations around 60 for X_3 average definitely below the line; both the next group up to 72.5 and the next group 72.5 up to 75 average slightly below, whereas the single observation above 85 falls above the line. These averages are indicated by squares on Figure 55.[10] The single high observation at the end alone would not be enough to indicate a change in the curve, but it is consistent with the group averages, which indicate the need for a slightly steeper curve than the original one. Accordingly this new curve is drawn in, approximately through the group averages, but still conforming to the conditions stated on page 279. To this point none of the relations, as indicated by the data, has differed sufficiently from the shapes logically expected to require any reconsideration of the logical analysis from which the conditions limiting the shapes to be drawn were derived.

Determination of second approximation curve for third independent variable. The same process is used in determining the second approximation for the next variable. The vertical departures of the dots on Figure 55 above or below the second approximation curve, $f_3''(X_3)$, shown as a dashed line, are scaled off and plotted as departures from the $f_4'(X_4)$ curve, with the corresponding X_4 values as abscissas. Again a new chart is prepared, Figure 56, with $f_4'(X_4)$ reproduced, although the original chart, Figure 53 (on page 285), is still clear enough so that these new values could readily have been plotted upon it. Again, as the observations are equally spaced in time, a continuous light line is drawn in, connecting the successive observations.

If the curve were any ordinary function—anything except a trend allowance for a number of unrepresented factors—there would be little evidence, from the dots in Figure 56, for any further change in the fitted curve. Since it is a trend allowance, however, and was ex-

[10] These averages have been estimated graphically, by the technique explained on page 485.

pected to be irregular on logical grounds (note the conditions stated on page 279), more flexibility may be in order. Comparing Figure 56 with Figure 53, we see that the observations have been changed only slightly by the further adjustments for $f_2(X_2)$ and $f_3(X_3)$. The individual observations on both charts show a pronounced fall from 1920 to 1924, a flattening out then for three or four years, then another fall to 1929. Between 1923 and 1927, Figure 56 shows that 4 out of 5

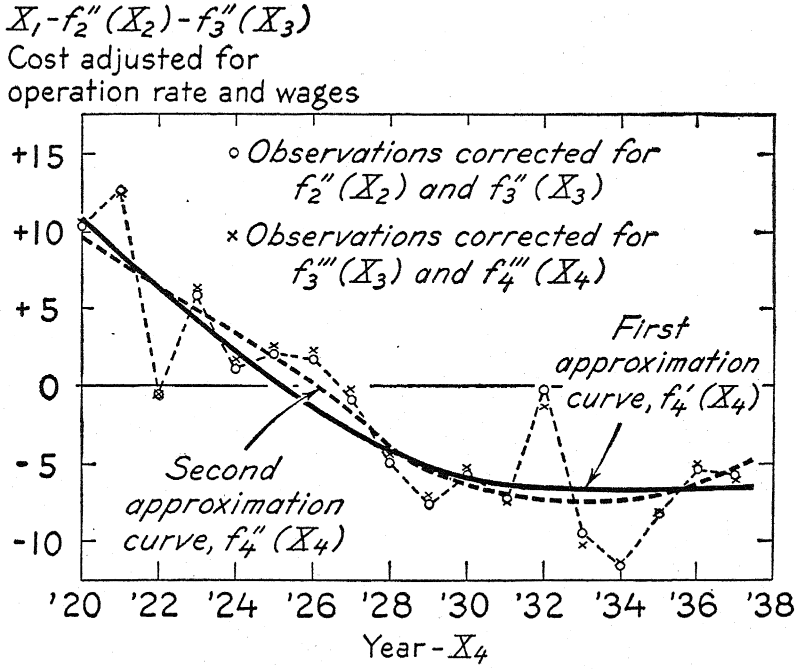


FIG. 56. Time, and cost per ton adjusted to average operation rate and wages on basis of second approximation curves; and second approximation to $f_4(X_4)$.

observations fall above the $f_4'(X_4)$ line, whereas, between 1928 and 1935, 6 out of the 8 observations fall below the line. These departures indicate that some changes in the first curve are justified. It is apparent that these changes would not be inconsistent with the possible composite effects of price-level changes and a general downward trend in production efficiency. The sharp fall from 1920 to 1924, however, largely reflects the two high observations for 1920 and 1921, offset somewhat by a very low observation in 1922. Accordingly, the trend may be interpreted as moderately downward from 1920 to 1926, more sharply downward to about 1929, then gradually tapering off to a low about 1933 or 1934, and rising gradually thereafter. A more flexible trend is therefore drawn in according to these general changes but not following single observations to the extremes of their departures.[11]

Determination of third approximation curves. The same process as before is now repeated, plotting the departures from $f_4''(X_4)$ around the $f_2''(X_2)$ curve, with X_2 values as abscissas. This time the new departures shown on Figure 56 are plotted back on the previous chart, Figure 54. Crosses are used for the new departures, to distinguish them from the previous values shown as hollow dots. To prevent confusing the chart, the observation (year) number is not shown with the cross, except where there are two or more observations with about the same X_2 value.

Examining the location of these new crosses on Figure 54, we notice that, for every observation with a value below 50 for X_2, the cross is one to one and one-half units (of X_1) higher than the corresponding dot. For values of X_2 above 50, however, the crosses fall alternately above and below the corresponding dots, with the averages of the crosses hitting just about the curve. This pattern indicates that the $f_2''(X_2)$ curve should be raised somewhat below 50, to be still steeper. Accordingly, a new curve is drawn in, changed as indicated, to pass as near as possible through the group averages of the crosses (as graphically estimated) and yet conform with the logical limitations on its shape.

The vertical departures of the crosses from the new curve, $f_2'''(X_2)$, are then carried forward to Figure 54, as departures from $f_3''(X_3)$. Again crosses are used to represent the new values.

Inspection of Figure 55, after the crosses are inserted, discloses a different situation from that in the previous chart. In the left portion of Figure 55, for values of X_3 below 65, the crosses fall very close to the corresponding dots, with no change for the average. In the right-hand portion, for values of X_3 above 75, the crosses also fall above and below the corresponding dot. Between 65 and 75, however, a number of the crosses fall a considerable distance below the corresponding dot, so that out of the twelve observations in this range, six crosses fall slightly above the f'' line and six fall a considerable distance below. This pattern indicates that the f'' curve should be made more sharply concave, without changing the elevation of either

[11] Only in rare instances would a curve with this much flexibility be justified. In this particular case its use is in line both with the theoretical analysis and the resulting conditions imposed on the shape of the curve.

end. A new curve is therefore drawn in to correct this, through the group averages of the crosses. (To prevent confusion, these averages are not shown on Figure 55.) The sharp lift in the last portion of this curve is dependent only upon the two observations, 1920 and 1937. However, the shape of this part of the curve is consistent with the logical limitations and with the other observations. Except for these two observations, a straight line would fit the crosses almost as well as the curve. The evidence for the existence of a curve, or for its exact shape, is thus very uncertain, as the data are distributed here.[12]

If the f''' curves are compared with the f'' curves on both Figure 54 and Figure 55, it is evident that we have determined the shape of these curves about as well as we can with the data at hand. Even with the material change in the trend by using the much more flexible curve of $f_4''(X_4)$, the differences between the f'' curves and the f''' curves for X_2 and X_3 are insignificant. However, to complete the process we carry the final residuals, the departures of the crosses on Figure 55 from the $f_3'''(X_3)$ curve, over to Figure 56, as departures from the trend line $f_4''(X_4)$.

There is no improvement in the average closeness of the crosses to the trend line, $f_4''(X_4)$, as a result of the slight changes in f_2 and f_3. The general characteristics of the trend, as fitted by the previous flexible curve, remain the same. From 1923 to 1930, every cross falls slightly above the corresponding dot, suggesting the possibility of a slightly better fit if the trend was raised a little in this portion. The single high value in 1932 continues to stand out, alone and unexplained. It seems hard to justify it on any trend basis. We could eliminate the wide departure for 1932 by twisting the lower end of $f_2(X_2)$ up sharply to pass through this single observation. In the absence of confirmatory evidence from another such low year for percentage of capacity operated, this would be a risky assumption.

Although it would be possible to modify the trend further, as suggested in the preceding paragraph, it seems best to let it stand unchanged. In view of the slight changes in the f_2 and f_3 curves in the last approximation, we end the successive approximation process at this point, feeling we have carried the process about to the point of diminishing returns in increased accuracy.

It should be noted, in Figures 54, 55, and 56, that the final curves at the end of the approximation process differ significantly from the

[12] See page 338 of Chapter 18 for the sampling reliability of the portion of a curve determined by such extreme observations, where the theory of random sampling may be properly applied.

first approximations only in the case of $f_2(X_2)$. Almost the same flexible trend of $f''_4\ (X_4)$ could have been drawn in the first approximation on Figure 53. The closeness with which $f'_3\ (X_3)$, $f'_4\ (X_4)$, and $f''_2\ (X_2)$ approximate the final curves is an indication of the great power of the graphic method in making a rapid approach to the underlying relations. The routine of comparing selected observations for which the values of the other independent variables are constant, or almost so, and judging the net relations from these selected comparisons provides a much closer initial approximation to the final curves than does the initial assumption of linear net regressions, used as the starting point in the successive approximation process presented in Chapter 14.

(For an exercise, the student might take the example which has just been analyzed and determine the net regression curves by the method of Chapter 14, using the same limitations on the shape of the curves as used here. That will enable him to compare the relative speed and effectiveness of the two methods in approaching the final curves.)

As already noted the intercorrelations among X_2, X_3, and X_4 were only moderate in this case. In a problem where the intercorrelations among the independent variables were quite high, the improvement in the fit of the several regression curves as a result of the successive approximation process might be more marked than it was in the example just completed. In such a case the convergence toward the curves of best fit will be slower than where the intercorrelations are low, and a larger number of successive approximations will be required to determine the final curves.

If, after several approximations have been made, the new curves start swinging up and down over curves previously determined, the approximation has probably been carried far enough. Especially where the intercorrelations for two independent variables are very high, a rise in the slope of one curve will cause a fall in the slope of the other. In such a case the exact position of each of the two curves is indeterminate, and the zone within which the last two or three approximations vary will indicate something of the uncertainty as to the exact shape or location of each curve. As will be shown later (Chapter 18), the reliability of *any* net regression line or curve varies inversely with the extent to which the particular independent variable is correlated with the other independent variables. Where two variables are so closely correlated that the relation to the dependent variable may be ascribed to either independent variable or parceled out between the two, their individual effect is indeterminate. Only by securing a large enough sample can the true influence of each be judged. When a large enough sample cannot be secured, that is the inherent fault of the data and not of the method employed. When used with due regard to the logical significance of the curves obtained, any one of the several methods will tend to give results which are substantially the same—that is, which lie within the range of possible accuracy imposed by the facts of the particular sample.

Determining standard error of estimate and the index of multiple correlation. The standard error of estimate may now be determined by first computing the value of $\sigma_{z'''}$. This can be done most simply by scaling off, on Figure 56, the departures of the last adjusted values (the crosses) from the final trend curve. These departures are the z''''s. Any errors which have been made in any of the successive graphic transfers will accumulate in these residuals. A more exact check can be made by reading off the estimated values for each observation from the final curves and adding them up to calculate the estimated X_1''' and z''', according to the same method used in Chapter 14. The z''' values as computed in this manner should agree closely with the z''''s scaled from the final approximation chart. These calculations are shown in Table 69B.

Column 10 of Table 69B gives the residuals as scaled off from the last approximation curve on Figure 56. Column 9 gives the residuals as computed in the usual way from the several curve readings. It is evident that the two columns agree very closely, the largest difference being only 0.4. This is an indication of the degree of accuracy maintained in the successive graphic transfers. In this case graph paper 8 by 10 inches was used in preparing the charts for Figures 51 to 56, and each of the transfers was double-checked. If higher accuracy in the mechanical process is desired, a still larger scale could be employed.

Taking the residuals in Column 9 as the most accurate, we may now calculate their standard deviation (around their own mean). It works out at 2.88. This compares with a standard deviation for X_1 of 7.19.

Before computing $\overline{S}_{1.f(2,3,4)}$ and $\overline{P}_{1.234}$, we need the values for n and m. A simple parabola or hyperbola with two constants would probably represent $f_2'''(X_2)$ and $f_3'''(X_3)$. However, $f_4''(X_4)$ with its two inflections would probably require at least three constants. In addition, there is an a constant, represented by the mean of the z''''s. Altogether, then, it would probably take eight constants to fit mathematical curves to the regression functions graphically determined. Accordingly, $n = 18$ and $m = 8$. With these values, we can now compute $\overline{S}$ and $\overline{\text{P}}$ by equations (65) and (66.2).

$$\overline{S}^2_{1.f(2,3,4)} = \frac{n\sigma^2_{z''}}{n-m} = \frac{18(2.88^2)}{18-8} = 14.9299$$

$$\overline{S}_{1.f(2,3,4)} = 3.86$$

$$\overline{\text{P}}^2_{1.234} = 1 - \frac{\overline{S}^2_{1.f(2,3,4)}}{\sigma^2_1}\left(\frac{n-1}{n}\right) = 1 - \frac{14.9299}{(7.19)^2}\left(\frac{17}{18}\right)$$

$$= .7272$$

$$\overline{\text{P}}_{1.234} = 0.85$$

TABLE 69B

CALCULATION OF ESTIMATED X_1 FROM FINAL REGRESSION CURVES

Year X_4 (1)	X_2 (2)	X_3 (3)	$f_2'''(X_2)$ (4)	$f_3'''(X_3)$ (5)	$f_4''(X_4)$ (6)	$\Sigma(f_2+f_3+f_4) = X_1'''$ (7)	X_1 (8)	z''' (8–7) (9)	z''' * (10)
1920	88.3	77.5	57.1	4.9	9.7	71.7	72.3	0.6	0.9
1921	47.5	60.2	67.8	−1.8	8.1	74.1	78.5	4.4	4.4
1922	71.3	58.5	60.5	−2.1	6.5	64.9	57.9	−7.0	−7.0
1923	88.3	67.0	57.1	−0.3	4.9	61.7	63.0	1.3	1.5
1924	69.0	70.8	61.0	1.0	3.4	65.4	63.7	−1.7	−1.8
1925	78.4	70.3	59.1	0.8	1.9	61.8	62.9	1.1	0.8
1926	88.0	70.8	57.2	1.0	0.3	58.5	60.3	1.8	2.1
1927	78.9	71.3	59.0	1.2	−1.6	58.6	59.6	1.0	1.3
1928	83.4	71.8	58.1	1.4	−3.7	55.8	55.2	−0.6	−0.5
1929	89.2	72.5	57.0	1.8	−5.4	53.4	51.5	−1.9	−1.7
1930	65.6	73.2	61.9	2.2	−6.3	57.8	58.6	0.8	1.0
1931	38.0	70.8	72.2	1.0	−6.9	66.3	65.6	−0.7	−0.7
1932	18.3	61.0	84.6	−1.7	−7.3	75.6	81.4	5.8	5.9
1933	28.7	59.0	77.3	−2.0	−7.5	67.8	65.0	−2.8	−2.8
1934	31.2	70.0	75.8	0.7	−7.4	69.1	64.6	−4.5	−4.1
1935	38.8	73.0	71.7	2.0	−7.0	66.7	65.4	−1.3	−1.1
1936	59.3	74.0	63.5	2.6	−6.4	59.7	61.1	1.4	1.3
1937	71.2	86.0	60.5	11.0	−5.4	66.1	65.6	−0.5	−0.8

* These are the values of z''' scaled off from Figure 56.

The multiple correlation 0.85 is still close, even after the adjustment for the number of observations and constants. The standard error of estimate works out at $3.86 per ton. This indicates that if it were possible to measure this same relationship between other factors and costs from a very large sample drawn from the same universe, the errors in estimating steel costs for the observations in that large sample would *probably* have a standard deviation of $3.86.[13]

[13] See pages 341 to 356 of Chapter 19 for the errors of individual forecasts and for the application of error formulas to time series.

Estimating cost for new observations. We can now use the data for 1938 and 1939, which we have disregarded to this point, to work out estimates for those years from the regression curves, by the same process shown in Table 69B. The values are:

Year	X_2	X_3	$f_2''(X_2)$	$f_3'''(X_3)$	$f_4''(X_4)$	X_1'''	X_1	z'''
1938	36.2	90.0	73.0	14.5	−4.3	83.2	80.5	−2.7
1939	60.7	89.7	63.1	14.2	−3.0	74.3	76.0	1.7

Just as in the similar example in Chapter 14, it is necessary to extrapolate two of the regression curves beyond the base data in making this estimate for subsequent years. In spite of the additional possibility of error which this introduces, both of the new estimates show residuals no larger than $\overline{S}_{1.f(2,3,4)}$. This indicates that the changes in steel costs during these next two years were in general related to the same factors as during earlier years and to about the same degree. (The student can check this conclusion by adding these two new observations to the original data, and re-analyzing the resulting sample of twenty observations.) If the trend or other factors were extrapolated much further, or if a sudden change in the conditions surrounding the industry were to occur, much larger errors of estimation might be experienced.

Restating short-cut results for publication. The same methods described on pages 247 to 254 of Chapter 14 can be used with curves obtained by the short-cut process, to prepare them for publication. There is a shorter method, however, which takes advantage of the fact that the curves obtained by the short-cut method are already in terms of a net value of X_1, for one variable, plus adjustments to that value for the other variables. All that is necessary is to determine the average value of the final z's and use this average as the a constant. (In the illustrative example just given, this average was only 0.08, and consequently was ignored.) Then the final functions are determined as follows (for the final curves of the illustrative problem)

$$F_2(X_2) = a + f_2'''(X_2)$$

$$F_3(x_3) = f_3'''(x_3)$$

$$F_4(x_4) = f_4''(x_4)$$

It is evident that, except for the slight adjustment of adding a to the first curve, these curves are the same as the final curves shown on Figures 54, 55, and 56.

Identifying "joint" relations by the short-cut process. In some problems the relation between the variables is such that the independent variable cannot be explained fully by a regression equation which *adds* the regression of X_1 on variable X_2, to that on X_3, etc. Instead, in such cases the relation is so complex that the net change in X_1 with given changes in X_2 will vary with the associated values of X_3 or other variables. This type of relationship, designated "joint correlation," is discussed subsequently (Chapter 21). Where such correlation is present, it will show up in the process of examining the subgroups of observations in the first steps of the short-cut process.

The following empirical data will serve to illustrate the occurrence of joint correlation:[14]

Observation Number	X_1	X_2	X_3	X_4
1	216	9	4	6
2	160	10	8	2
3	140	2	7	10
4	264	4	11	6
5	30	5	2	3
6	56	7	1	8
7	5	1	5	1
8	16	2	2	4
9	70	2	5	7
10	126	7	6	3
11	180	10	3	6
12	280	5	7	8
13	120	3	4	10
14	25	1	5	5
15	224	4	8	7
16	120	6	0	2

The number of cases here is so small that it is difficult to eliminate the effects of X_3 and X_4, to determine the first approximation to the X_1X_2 relation. An approximate grouping can be made, however, by classifying the observations into three groups, as follows:

One, those with X_3 and X_4 both *larger* than their respective means.

Two, those with X_3 and X_4 both *smaller* than their respective means.

[14] From Wilfred Malenbaum and John D. Black, The use of the short-cut graphic method of multiple correlation, *Quarterly Journal of Economics,* Vol. LII, p. 97, November, 1937.

Three, those with X_3 and X_4 one above and one below their respective means.

This gives groupings with four observations (3, 4, 12, and 15) in the first group, four (5, 7, 8, and 14) in the second, and eight (1, 2, 6, 9, 10, 11, 13, and 16) in the third. Plotting each of these groups of observations, and drawing an approximate line through each, gives the results shown in Figure 57.

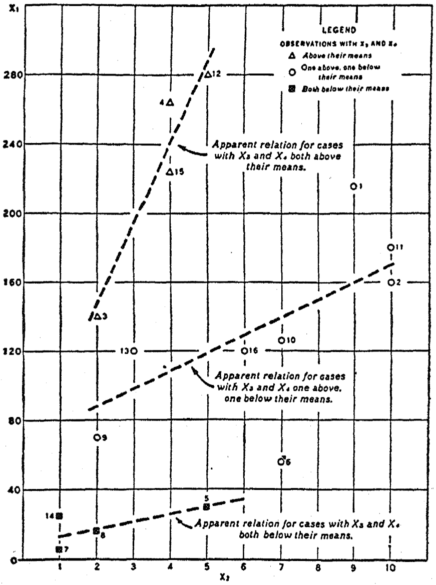


FIG. 57. Relation of X_1 to X_2, with observations classified on X_3 and X_4. When natural numbers are used, the net regression of X_1 on X_2 appears to shift with the accompanying values of X_3 and X_4.

This figure differs from those we have examined previously (such as Figure 44 on page 272 or Figure 52 on page 284) in that the relations as shown by the several subgroups do not parallel one another at relatively constant distances, but instead diverge sharply. It appears, therefore, that the relation of X_1 to X_2 depends not only on the value of X_2 but also on the *associated values of X_3 and X_4*.

In this particular case the progressive nature of the relations shown on Figure 57 might lead us to suspect that the relation, instead of being an additive one, is a multiplying one. If that is the case, though it could not be represented adequately by an equation of the type:

$$X_1 = f_2(X_2) + f_3(X_3) + f_4(X_4)$$

it still might be represented by:

$$X_1 = [\phi_2(X_2)]\,[\phi_3(X_3)]\,[\phi_4(X_4)]$$

If that is the case, it can be determined by using the relation:

$$\log X_1 = f_2\,(\log X_2) + f_3\,(\log X_3) + f_4\,(\log X_4)$$

We can test whether this is likely to give a satisfactory fit by replotting Figure 57 on double logarithmic paper, or by plotting it on ordinary paper, substituting the logarithms of X_1 and X_2 for the natural values. Let us do the latter.

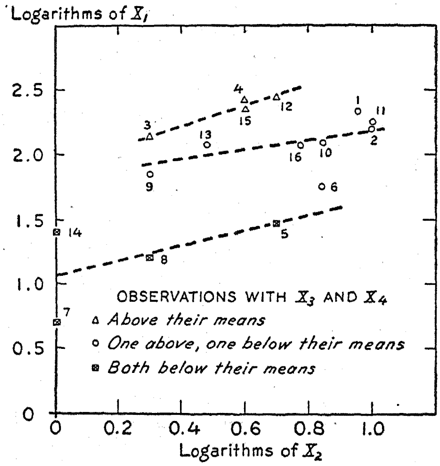


FIG. 58. When the logarithms of the data shown in Figure 57 are used, the net regression of X_1 on X_2 is found to be about the same, regardless of the accompanying values of X_3 and X_4.

When that is done, the relations appear as shown in Figure 58. The three lines, fitted roughly to the three sets of observations, now appear more nearly parallel. In particular, the line of the upper group, which in Figure 57 made almost a 60° angle with the line for the lower group, is almost perfectly parallel to it in Figure 58. Apparently in this example the problem can be handled satisfactorily by the usual short-cut procedures, merely by transforming the variables from natural numbers to logarithms.

Where this transformation, or other simple transformations, do not serve to make the successive sub-groups show approximately parallel relations, the methods of Chapter 21 must be employed instead.

Application of the short-cut method to large samples. The short-cut method might be applied to samples too large to plot the individual observations separately, by using a modification of the process of subgrouping and averaging illustrated in Chapter 11. The averages from Table 42, plotted in Figures 30 and 31, indicated quite well the final slope of the net regression lines. That was because the influence of the other independent variable had been largely held constant by the process of subclassifying. In the same way the lines of averages from subgroups would tend to indicate the regression curves in problems where curves were needed. With a sufficient number of observations, the first approximation to each of the net regression curves might be obtained from charts of subaverages similar to Figures 30 and 31 on page 183. These several first approximation curves could then be made the basis for working out estimated values of X_1 and residuals. The process of successive approximations could then be continued exactly as illustrated in Chapter 14. Since the first approximation curves would approach fairly near to the true net regressions, the number of approximations required to obtain the same closeness of fit would usually be less than by the earlier method.

Combination of short-cut procedures and mathematical procedures. Both the short-cut method of this chapter and the longer successive-approximation method of Chapter 14 depend on graphic methods in arriving at the curves of best fit. Where especially high accuracy is desired, the final slope of the several curves can be checked by least squares, according to the methods set forth in Chapter 22 on pages 401 to 403.

Some investigators prefer to use the short-cut method to determine the approximate shapes of each of the several net regression curves, and then to fit mathematical net regressions capable of representing those several shapes. The technique for fitting these mathematical curves to several variables is also set forth in Chapter 22 on pages 396 to 401. If there is a logical basis to support the curves employed, there is some value to this procedure. If the equations are simply selected empirically, however, the mathematical curves have no more meaning than the graphic ones, for the reasons already discussed fully in Chapter 6. It is true that any one fitting the same set of mathematical curves to the same data by the same method will get exactly the same result, to the fifth decimal place in the values of the constants, if desired. Curves obtained by different investigators by either graphic process, on the contrary, may vary slightly from one to another. But the identical constants obtained by the least-squares fit have only a fictitious accuracy, as compared with their standard errors, or with the zone of uncertainty within which the function can be determined from the given set of observations. Multiple regression curves are significant only with respect to this zone, rather than to the exact line (as explained fully in Chapter 18). With proper care in analyzing the data for interrelationships and in carrying through the successive approximations, as explained in Chapter 14 and in this chapter, either graphic method will ordinarily give results about as significant, within their error zone, as results obtained by the more laborious methods of fitting mathematical curves by extensive arithmetic calculations.

Summary. Under certain conditions first approximations to multiple regression lines or curves may be obtained directly from the original observations by a graphic process based on the comparison of individual observations, considering several variables simultaneously. This process eliminates the necessity of computing linear regressions by arithmetical means. Further, it substitutes graphic measurements for arithmetic calculations in correcting these curves to their final shape by successive approximations. It requires the researcher to examine his data more thoroughly and so to exercise thought and care in working out the relations and in interpreting their significance. Carefully used, it materially reduces the time required in determining multiple regression curves.

Note 1, Chapter 16. In view of the extensive discussions which have occurred concerning the validity of the short-cut method, certain key articles on this point are listed here.

WAITE, WARREN C., Some characteristics of the graphic method of correlation, *Jour. Amer. Stat. Assoc.*, Vol. XXVII, pp. 68–70, March, 1932.

EZEKIEL, MORDECAI, Further remarks on the graphic method of correlation, *Jour. Amer. Stat. Assoc.*, Vol. XXVII, pp. 183–185, June, 1932.

MALENBAUM, W., and J. D. BLACK, The use of the short-cut graphic method of multiple correlation, *Quart. Jour. Econ.*, Vol. LII, pp. 66–112, November, 1937.

BEAN, L. H., and MORDECAI EZEKIEL, The use of the short-cut graphic method of multiple correlation, Comment, and Further comment, *Quart. Jour. Econ.*, Vol. LV, pp. 318–346, February, 1940.

WELLMAN, H. R., Application and uses of the graphic method of multiple correlation, *Jour. Farm Econ.*, Vol. XXIII, pp. 311–316, February, 1941.

WAITE, WARREN C., Place of, and limitations to, the method, *Jour. Farm Econ.*, Vol. XXIII, pp. 317–322, February, 1941.

WORKING, E. J., and GEOFFREY SHEPHERD, Notes on the place of the graphic method of correlation analysis, *Jour. Farm Econ.*, Vol. XXIII, pp. 322–323.

FOOTE, RICHARD J., and J. RUSSELL IVES, The relationship of the method of graphic correlation to least squares, U. S. Department of Agriculture, Bureau of Agricultural Economics, mimeographed report, December, 1940.

These discussions, especially the report by Foote and Ives, and an address by Meyer A. Girshick at the same meeting, as summarized in the February, 1941, *Journal of Farm Economics*, have provided definite proof of the meaning of the graphic method. They have shown that in linear multiple correlation the graphic method gives results which tend to approach the lines secured by a least-squares solution, even if the first approximations are purely arbitrary guesses. Further, they have shown that the speed of convergence depends on the intercorrelation among the independent variables. The higher their intercorrelation, the slower tends to be the speed of the convergence.

The discussion and procedures in this chapter, as now revised, take into account these recent examinations of the meaning of the short-cut graphic method, and incorporate the most useful and significant suggestions to the student which have come out of them.

Note 2, Chapter 16. The comments made in the note on page 258 apply to Chapter 16 as well. If the standard error of estimate is calculated (as shown on pages 293 and 294) as each new set of approximation curves is completed, it will show whether the gain in closeness of fit is sufficient to offset any additional flexibility introduced in the curves. The validity of this test, however, depends upon the user's skill in estimating the value of m to employ.

CHAPTER 17

MEASURING THE WAY A DEPENDENT VARIABLE CHANGES WITH CHANGES IN A NON-QUANTITATIVE INDEPENDENT FACTOR

It is frequently desirable to determine the change in one variable associated with changes in an independent factor which varies in such a way that it cannot be measured quantitatively. Thus if the significance of various factors affecting farm values is to be determined, one may wish to include type of road as one of the factors, since a farm on a concrete road should be expected to be worth more than one on a dirt road, other factors being the same. Yet the designations, concrete, brick, macadam, gravel, and dirt, cannot be considered in the correlation analysis in the way that the numbers measuring variable factors are treated.

Where no other factors are involved, a non-quantitative factor may be treated by sorting with respect to that factor, and averaging the dependent variable. Thus if only type of road is being considered, the average value per acre of farms fronting on each type of road may be taken as the measure of the influence of roads on value. If, however, several other factors must be considered at the same time, such as value of improvements, productivity of the soil, distance from town, etc., and if there is any relation between differences in these factors and differences in road type (as in general there will tend to be), the influence of road type must be measured by some application of multiple correlation methods. Fortunately the methods of multiple curvilinear correlation, as presented in Chapters 14, 15, and 16, can be extended to treat non-quantitative factors as well, and thus provide the answer to the difficulty.

Eliminating the influence of other variables. The method of determining regressions for non-quantitative variables may be illustrated by the data shown in Table 70. These data are from a study of the relation of various quality factors to the price of eggs sold at retail.[1] The factors shown in Table 70 are X_2, an index of the interior quality

[1] Original data collected by C. B. Howe. See reference 42 of Chapter 23.

TABLE 70

DATA FOR EGG PROBLEM, WITH A NON-QUANTITATIVE INDEPENDENT VARIABLE

Independent variables				Dependent variable, X_1	z'''	$f(X_5)$	z''''
X_2	X_3	X_4	X_5 *				
21	23	4	C	35	− 7.3	+0.6	− 7.9
35	24	12	C	45	− 8.4	+0.6	− 9.0
26	23	12	B	55	3.4	+0.9	2.5
27	24	12	B	55	3.3	+0.9	2.4
31	22	12	A	50	− 1.8	−2.5	0.7
35	24	12	C	44	− 9.4	+0.6	−10.0
28	23	12	C	60	8.2	+0.6	7.6
41	23	12	B	50	− 4.8	+0.9	− 5.7
28	26	2	C	45	− 1.6	+0.6	− 2.2
24	23	11	B	52	4.6	+0.9	3.7
28	20	12	C	45	− 5.5	+0.6	− 6.1
49	24	12	C	55	− 3.6	+0.6	− 4.2
30	24	12	C	55	2.4	+0.6	1.8
48	23	12	B	60	1.9	+0.9	1.0
19	22	9	C	45	1.8	+0.6	1.2
22	23	3	A	45	1.7	−2.5	4.2
33	25	12	C	60	6.6	+0.6	6.0
26	24	12	C	59	6.9	+0.6	6.3
35	23	12	B	55	2.1	+0.9	1.2
20	23	12	B	50	− 0.9	+0.9	− 1.8
25	25	12	B	55	2.6	+0.9	1.7
46	24	12	B	60	2.5	+0.9	1.6
30	26	1	B	45	− 3.2	+0.9	− 4.1
24	24	12	B	55	3.1	+0.9	2.2
48	23	12	B	60	1.9	+0.9	1.0
17	22	12	C	55	4.8	+0.6	4.2
18	22	12	A	45	− 5.3	−2.5	− 2.8
41	24	12	C	55	− 0.3	+0.6	− 0.9
30	25	12	C	67	14.0	+0.6	13.4
19	24	2	B	53	8.3	+0.9	7.4
47	24	0	B	55	0.9	+0.9	0.0
32	24	12	B	55	2.2	+0.9	1.3
26	24	12	B	49	− 3.1	+0.9	− 4.0
38	24	12	A	42	−12.2	−2.5	− 9.7
29	23	12	B	42	− 9.9	+0.9	−10.8
24	24	0	A	45	− 2.9	−2.5	− 0.4
37	25	12	A	40	−14.3	−2.5	−11.8
36	23	12	A	48	− 5.1	−2.5	− 2.6

* A designates "sold without carton," B "sold in carton but unbranded," and C "sold in carton with brand name."

TABLE 70—*Continued*

Independent variables				Dependent variable, X_1	z'''	$f(X_5)$	z''''
X_2	X_3	X_4	X_5 *				
10	23	0	B	47	1.2	+0.9	0.3
35	24	12	C	59	5.6	+0.6	5.0
22	22	12	B	52	1.2	+0.9	0.3
29	21	12	B	55	4.0	+0.9	3.1
16	23	0	B	40	−6.5	+0.9	−7.4
6	22	3	B	40	−1.0	+0.9	−1.9
31	23	12	B	55	2.8	+0.9	1.9
26	23	12	B	55	3.4	+0.9	2.5
36	21	12	B	60	7.8	+0.9	6.9
39	22	12	B	55	1.4	+0.9	0.5
42	23	12	B	60	4.8	+0.9	3.9
36	24	12	C	60	6.4	+0.6	5.8
47	22	12	B	60	2.8	+0.9	1.9
27	24	12	C	55	2.8	+0.6	2.2
31	22	12	A	50	−1.8	−2.5	0.7
26	22	11	A	40	−7.2	−2.5	−4.7
45	23	12	A	60	3.5	−2.5	6.0
18	25	12	C	45	−6.6	+0.6	−7.2
35	24	12	C	50	−3.4	+0.6	−4.0
21	23	12	C	55	4.0	+0.6	3.4
44	23	12	A	60	3.9	−2.5	6.4
48	24	12	A	55	−3.6	−2.5	−1.1
33	24	12	A	55	2.0	−2.5	4.5
47	24	12	C	55	−3.1	+0.6	−3.7
16	22	5	A	45	3.9	−2.5	6.4
32	25	0	B	50	0.8	+0.9	−0.1
45	25	12	B	55	−2.4	+0.9	−3.3
46	23	12	B	57	0.0	+0.9	−0.9
32	24	12	C	55	2.2	+0.6	1.6
16	23	1	C	41	−4.2	+0.6	−4.8
30	25	1	C	50	2.3	+0.6	1.7
24	22	0	A	42	−5.0	−2.5	−2.5
44	24	11	B	50	−2.6	+0.9	−3.5
25	22	12	B	49	−2.1	+0.9	−3.0
16	23	0	A	45	−1.5	−2.5	1.0
31	24	8	A	48	3.2	−2.5	5.7

* A designates "sold without carton," B "sold in carton but unbranded," and C "sold in carton with brand name."

of the eggs in each dozen; X_3, the weight of each dozen in ounces; X_4, the number of white eggs in each dozen; X_5, the type of carton the eggs were sold in; and X_1, the price of eggs per dozen, in cents. Net curvilinear regressions have been determined for the three quantitative factors by the successive approximation method, and estimated prices have been worked out by the regression equation

$$X_1' = a' + f_2(X_2) + f_3(X_3) + f_4(X_4)$$

The residuals, z''', obtained by subtracting these estimated prices from the observed prices, X_1, are shown in the table. The values in the last two columns are explained later.

Determining the net influence of the new variable. The first step in determining the net regression of X_1 on X_5 is to group the residuals from the previous curves, z''', according to the new factor X_5, and determine the average for each group. This gives results as follows:

Value of X_5	Average of z'''
A—no carton	−2.5
B—carton	+0.9
C—carton and brand name	+0.6

These results show that, after making allowances for the size, color, and quality of the eggs, those with unmarked cartons sold 3.4 cents above those sold in bulk, on the average, but those with branded cartons sold only 3.1 cents above eggs in bulk. These results cannot be accepted as the final effect of package on price without first raising the question whether the curves previously determined to show the influence of the other factors might be changed somewhat were the type of package taken into account. Whether this will be true or not depends upon whether there is any correlation between the new factor and the factors previously considered, or whether they are quite independent of each other. This can be determined by sorting the other factors according to the values of X_5, and determining their averages for each group. The results are:

Value of X_5	Averages of other independent variables			Number of cases
	X_2	X_3	X_4	
A—no carton	30.6	23.1	8.6	17
B—carton	31.6	23.2	9.6	33
C—carton and brand	29.9	23.8	10.2	24

There does seem to be some correlation between X_5 and the other variables. Apparently the eggs sold in unmarked cartons are, on the average, of the best quality and of medium size; the eggs sold in cartons under brand names are of larger size, but are not of such high quality, on the average; whereas those sold in bulk average medium in quality but low in size.[2] Accordingly, the curves previ-

[2] The exact correlation between X_5 and X_2, X_3, and X_4 can be computed by estimating each of the other variables from the values of X_5, using the averages of X_2, X_3, and X_4 for each group of X_5 as the estimated values of X_2, X_3, and X_4, for the cases falling in each group. The residuals between the estimated and actual values, and their standard deviation, can then be computed for each of the three variables. Then the indexes of correlation can be computed in the usual way. When computed this way by using group averages instead of a continuous function, the special name *correlation ratio* is given to the correlation, and the symbol η is used to designate it. This value may be more rapidly computed by the following formula (using Y to represent the dependent variable, and X the independent variable, just as with simple correlation in Chapters 5 to 7):

$$\eta_{yx} = \sqrt{\frac{\Sigma[n_0(M_0)^2] - n(M_y)^2}{n\sigma_y^2}} \tag{68}$$

Here η_{yx} is the correlation ratio for Y values estimated from group averages when sorted on X; $n_0(M_0)^2$ is the number of cases in each group times the square of the average value of Y for that group, $\Sigma[n_0(M_0)^2]$ is the sum of all such values, σ_y is the standard deviation of the variable being estimated, and n is the number of all the observations $(=\Sigma n_0)$.

The process may be illustrated by calculating η_{25}, the correlation ratio between X_2 and X_5, from the data above:

X_5	M_0 Mean X_2	n_0 Number of cases	$(M_0)^2n_0$
A	30.6	17	15,918.12
B	31.6	33	32,952.48
C	29.9	24	21,456.24
Σ	....	74	70,326.84

$$\eta_{25}^2 = \frac{\Sigma[n_0(M_0)^2] - n(M_2)^2}{n\sigma_2^2} = \frac{70,326.84 - 70,285.61}{7,505.76} = .005493, \quad \eta_{25} = 0.074$$

The value as calculated is subject to the same correction equation (26) as the correlation index, with m = number of groups.

So adjusted, η_{25} shrinks to 0, showing no real correlation.

This same measure of correlation by group averages can be applied to quantitative variables as well as to non-quantitative ones, but in that case it has less significance than the index of correlation, which relates to a continuous function instead of an irregular line of averages.

ously determined for the change in price with differences in size and in quality may have included some portion of the effect really associated with cartons instead. Now that at least an approximate measure has been obtained of the influence of carton on price, the previous curves may be modified by taking this factor also into account.

Taking account of the non-quantitative variable in estimating X_1 and z. The first steps in the procedure of allowing for the extent to which prices varied with the carton are shown in Table 70. In the column headed $f(X_5)$ the approximate influence of differences in carton on price are entered, the averages found in the tabulation on page 305 being used. Since these values would be added to the previous estimated values of X_1 to obtain the new estimates, they may instead be subtracted from the previous residuals (z''') to obtain the revised residuals. The last column shows these new values for z''''. Before using these new values to see if any changes are necessary in the other regression curves we may first determine how much the standard error of estimate has been reduced by taking X_5 into account. This could be determined directly by computing the standard deviation of the new z'''' values; but a much shorter method is available, using the same principle employed in footnote 2. By the use of this method, the $\sigma_{z''''}$ may be computed from the $\sigma_{z'''}$ by the formula

$$\sigma^2_{z''''} = \frac{n\sigma^2_{z'''} - [\Sigma(n_0 M_0^2) - n(M_{z'''})^2]}{n}$$

The necessary computations are:

X_5	$M_{z'''}$	Number of cases	nM_0	$n(M_0)^2$
A	−2.5	17	−42.5	106.25
B	0.9	33	29.7	26.73
C	0.6	24	14.4	8.64
		Sums........	1.6	141.62

$$M_{z'''} = \frac{1.6}{74} = 0.0216$$

So

$$\sigma^2_{z''''} = \frac{74(5.06)^2 - (141.62 - 0.04)}{74} = 23.69$$

$$\sigma_{z''''} = 4.87$$

Computing the standard error for estimates based on X_5 and the other variables, we must recognize that the value of m has been increased by three by the introduction of the new factor; so, whereas m was assumed to equal 8 previously, it now equals 11. Adjusting the values of 5.06 for $\sigma_{z'''}$ and 4.87 for $\sigma_{z''''}$ by equation (65), we find $\bar{S}_{1.f(2,3,4)} = 5.36$, and $\bar{S}_{1.f(2,3,4,5)} = 5.27$. Apparently the introduction of X_5 as a factor has had as yet but slight effect on the accuracy with which egg prices might be estimated.

Making further successive approximation corrections. It is still possible, however, that the regressions for the other factors might be modified now that X_5 has been at least approximately allowed for. Consequently the values of z'''' are classified according to the values of X_2, X_3, and X_4, and the averages computed for each group. The averages given in Tables 71, 72, and 73 are secured. The averages in Table 71 suggest that the curve for $f_2(X_2)$ might be modified slightly, so as to rise more steeply in the portion up to $X_2 = 40$ and less steeply thereafter. Table 72 does not indicate any consistent relation between X_3 and z'''', so no further change in $f_3(X_3)$ is indicated. Table 73 indicates that the curve for $f_4(X_4)$ might also be altered slightly, so as to have a somewhat steeper slope.

TABLE 71

AVERAGE VALUES OF z'''' FOR CORRESPONDING X_2 VALUES

X_2 values	Number of cases	Average of X_2	Average of z''''
0–14	2	8.0	−0.9
15–19	9	17.2	−0.2
20–29	23	25.1	+0.1
30–39	24	33.5	+0.3
40–49	16	45.5	−0.1

If $f_2(X_2)$ and $f_4(X_4)$ were modified as suggested, a new estimated value of X_1 might then be worked out, using these new curves and the previous curve for $f_3(X_3)$, and using the values for $f_5(X_5)$ already entered in Table 70. The new z's based on these new estimates might then be classified with respect to X_5, to determine if any change need be made in the values for $f_5(X_5)$ worked out on page 305. If any material change were found necessary in X_5, the residuals might be corrected accordingly, and then averaged with respect to X_2, X_3, and X_4, to see if any further changes would be needed in their values.

This process of successive approximation should be continued until no further significant change was indicated in any of the curves, or until the $\bar{S}_{1.f(2,3,4,5)}$ showed no further reduction.

TABLE 72

AVERAGE VALUES OF z'''' FOR CORRESPONDING X_3 VALUES

X_3 values	Number of cases	Average of z''''
20	1	−6.1
21	2	5.0
22	13	0.1
23	23	0.2
24	25	−0.1
25	8	0
26	2	−3.2

In view of the fact that none of the averages of z'''' shown in Tables 71 to 73 are so large but what they might very readily have occurred by chance, it does not seem worth while, in this problem, to carry out the additional steps just outlined. In a problem where the non-quantitative factor is an important one, however, and where it is significantly correlated with the other independent variables, the determination of the net function for that factor should be carried through a sufficient number of approximations to measure the final net effect of each factor as accurately as possible.

TABLE 73

AVERAGE VALUES OF z'''' FOR CORRESPONDING X_4 VALUES

X_4 values	Number of cases	Average of X_4	Average of z''''
0	7	0	−1.3
1– 2	5	1.4	−0.4
3– 5	4	3.8	+0.2
8–11	5	10.0	+0.5
12	53	12.0	+0.2

Taking the preliminary results shown on page 305 as the final measure of the influence of type of container on price, we may then conclude that eggs sold in an unmarked carton brought, on the average, 3.4 cents more per dozen than eggs of the same quality, size, and color sold in bulk, and 0.3 cent more than eggs sold in a carton with a brand name. (This last result might reflect the experience of consumers with branded eggs of poor quality, as indicated in the tabulation on page 305, which might tend to make them sell at a discount even when they were of equal quality.) The significance of the relation may be measured by the slight reduction in the standard error of estimate previously noted, or else by the increase in the index of multiple correlation. Computing the indexes of multiple correlation corresponding to the standard errors of estimate before and after the type of carton is allowed for, by equation (66.2), we find them to be $\overline{P}_{1.234} = 0.59$; $\overline{P}_{1.2345} = 0.62$. The corresponding indexes of determination, 35 and 38 per cent, indicate that taking into consideration the differences in the carton has increased the proportion of egg prices which can be explained by 3 per cent of the original variance, even after due allowance is made for the additional constants the process introduces into the estimating equation.

It should be noted that the first approximation to the regression on non-quantitative factors can be made directly from the first set of residuals, computed from the linear multiple regression equation, instead of waiting until after approximate regression curves are determined for the other factors. In case a non-quantitative factor is a very important one, so that ignoring it in determining the net linear regressions may seriously impair their accuracy, it may be roughly included by designating successive groups by a numerical code which approximates the expected influence of the variable. Then if the true influence is of a different order from the expected influence, that fact will show up when the first approximation curves are worked out. (For the non-quantitative factor the averages of residuals must be interpreted as discrete points for each class, however, rather than as a continuous function.) Thus for the egg problem it might have been tentatively assumed that eggs in branded cartons would sell above eggs in unbranded cartons, and both would sell well above eggs in bulk. The bulk eggs could then have been designated by 1; the unbranded cartons by 3; and branded cartons by 4. The net linear regression would have been positive; but the analysis of the residuals would have revealed that the eggs in branded cartons really averaged lower in price (other factors equal) than the eggs in unbranded cartons, so the final conclusion would probably be much the same as the one just determined.

Summary. Where an independent factor is not a continuous variable, but may be classified into two or more groups, the regression of a dependent factor may be determined with respect to each group, while holding other factors constant by the usual multiple correlation process. Standard errors and indexes of correlation may be worked out to include the effects of non-quantitative independent factors equally as well as for continuously variable factors.

CHAPTER 18

DETERMINING THE RELIABILITY OF CORRELATION CONCLUSIONS

Early in this book it was pointed out that when any statistical measure, such as an average, is determined from a sample selected from a universe under study, the true value of that measure in the universe might be different from the value shown by the sample. Methods were discussed which enable one to estimate how far the average from such a sample may vary from the true average, for a stated proportion of such samples. Such estimates enable one to judge how much confidence may be placed in an average calculated from a given sample.

Simple Correlation

Regression coefficients. Correlation constants determined from finite samples are just as subject to variation as are other statistical constants. Thus in an experiment 5 samples of 30 observations each were drawn at random from the same universe. The true value of b_{yx} for the universe was 0.152. The regression of Y on X was determined separately for each sample. The values for b_{yx} which were secured from the 5 samples varied from -0.136 to $+0.449$, as shown in Table 74. When 5 samples of 50 observations each were drawn, and

TABLE 74

VALUES OF b_{yx} SECURED IN SUCCESSIVE SAMPLES DRAWN FROM THE SAME UNIVERSE, WITH DIFFERENT NUMBERS OF OBSERVATIONS

	30 observations	50 observations	100 observations
	0.292	0.175	0.113
	0.012	−0.297	0.120
	−0.136	0.144	0.303
	−0.022	0.130	0.197
	0.449	0.167	0.132
True value	0.152	0.152	0.152

the regressions computed for each, the range was reduced to -0.297 to $+0.175$; but the variation between samples was still large. Even when 100 observations were included in each sample, the regressions were by no means identical, though the range was reduced still more.

It is evident that the observed values of b_{yx} fell both above and below the true value for the universe from which the samples were being selected.[1] It is also evident that the smaller the number of observations, the larger the variation in the results between different samples and the greater the possibility of a serious difference between the true value and that indicated by the sample. The amount of variation likely to be present in regressions determined from random samples of any specified size may be estimated by the equation

$$\text{Standard error of } b_{yx} = \frac{\bar{S}_{y.x}}{\sigma_x\sqrt{n}} \tag{69}$$

Since this constant is computed from the adjusted value, $\bar{S}_{y.x}$, no further adjustment is required.

If only one of the samples in Table 74 had been obtained—say the first one with 50 observations—the observed value for b_{yx} would have been $+0.175$. The standard error of estimate for this sample was 2.46, and the σ_x was 2.44. Computing the standard error of b_{yx} for this sample by means of equation (69),

$$\sigma_b = \frac{2.46}{2.44\sqrt{50}} = \frac{2.46}{17.25} = 0.143$$

the value of b_{yx}, as determined from this single sample, may therefore be stated to be 0.175 ± 0.143.

The standard error of the regression coefficient is interpreted exactly the same as the standard error of the average was interpreted in Chapter 2. In two samples out of three, on the average, the observed regression will miss the true regression by not more than one standard error calculated from the sample. Therefore, if in this case we say that the true regression lies between $0.175 - 0.143$ and $0.175 + 0.143$, or between 0.032 and 0.318, we are making a statement of a type which, if made for a succession of such samples, will be wrong one time out of three, on the average. Similarly, if we said that the true regression

[1] In some textbooks, b_{yx} would be used to represent the regression as determined from the sample and β_{yx} would be used to represent the true value of the corresponding regression in the universe from which the sample was drawn. In this notation, in Table 74, the value for $\beta_{yx} = 0.152$. In consulting textbooks using this notation, we should not confuse this use of the β with the special definition given it in Chapter 13, equation (52).

probably lies between -0.111 and 0.461, i.e., within a range of twice the standard error from the observed value, we are making a statement of a kind which, if made for a series of samples, will be wrong in one sample out of twenty, on the average.

It happens, in this particular case, that four out of five of the observed regressions (for samples of 50) fall within one σ_b of the regression from the first sample.[2] It also happens that the true value also falls within that range. This will not always be true, however. For example, if the sample had happened to give the same results as the third sample of 30 observations, with $b_{yx} = -0.136$, the case might have been different. For that sample, the values of the other constants were such as to make $\sigma_b = 0.109$. The value of b_{yx} as indicated by this sample, therefore, -0.136 ± 0.109, is such that the observed value lies 2.6 times its own standard error from the true value, 0.152. Although a departure as large as this would ordinarily be expected to occur only once out of every 100 samples on the average (0.009), still it *may* happen with any particular sample.[3] For that reason, if very great accuracy is desired, a range of three times the standard error may be used as the criterion. There is but one chance out of nearly 400 (0.0027) that a given random sample will yield a constant such as a regression coefficient which will fall more than three times its own standard error away from the true value for the universe.

These probabilities apply only in case there are thirty or more degrees of freedom (n-m) in the sample. As was pointed out in Chapter 2, if the number of degrees of freedom is less than thirty, the probabilities of falling outside of any given range of the true value are increased, as shown in Table A on page 23. In using this table for regression coefficients, subtract 1 from the number of cases in the sample before looking the probability up in the table.[4]

Thus if a value of $b_{yx} = 0.50 \pm 0.12$ were found from a random sample of 11 cases, the reliability of the observed regression could be judged from the column headed 10 in Table A. That column indicates

[2] A more precise way of stating this comparison would be to show a series of regressions frcm samples drawn from the same universe, such as those listed in Table 74, with each sample regression followed by ± its own standard error. If that were done, it would then be found that, in two samples out of three, on the average, the value $b_{yx} + \sigma_b$ would overlap the true value of b_{yx} for the universe.

[3] Probability tables, such as that given in Table A of Chapter 2, or shown graphically in Figure A, page 505, list these odds for various multiples of the σ.

[4] That is because two constants (a and b) have been determined simultaneously in the process of getting b, whereas the table is stated for arithmetic means, which represent the determination of only a single constant. (See page 22, footnote 7.)

that, with samples of this size, 34 out of each 100 samples, on the average, would miss the regression in the universe by as much as 0.12 (1 σ_b); about 8 out of each 100 would miss by as much as 0.24 (2 σ_b); and 15 samples out of each 1,000 would miss by as much as 3 σ_b, or 0.36. Thus in this case, if we say that the true value probably lies between 0.14 and 0.86, we are making a statement of the sort which is likely to be wrong only once or twice out of each hundred such statements—if the sample was drawn under such conditions that the formulas of simple sampling hold true.

It should be noted from equation (69) that the standard error of the regression coefficient varies inversely with the square root of the number of observations. The effect of this is illustrated in Table 74. The variation of the regression coefficients obtained from samples of 100 observations is only about half as great as the variation of the regression coefficients from samples of 30.

Regression line. Not only may the observed *slope* of the regression line vary from the true slope, but the elevation of the line, as observed from a sample, may vary from the true elevation. Formula (69) has already indicated a way of determining the standard error of the regression coefficient, and so of estimating the probable range within which the true slope lies. The height of the regression line is most accurately determined for the mean estimated value, $M_{y'}$, of the dependent factor, corresponding to the observed mean value of X, the independent factor. If we define the mean as

$$M_{y'} = a_{yx} + b_{yx}M_x$$

we may find its standard error by the formula

$$\sigma_{M_{y'}} = \frac{\bar{S}_{y.x}}{\sqrt{n}} \tag{70}$$

The standard error of the whole regression line may now be determined from equations (69) and (70). We may illustrate by data from the cotton-yield problem used as an example in Chapter 8, on page 147. With 14 observations, the values were $b_{yx} = 16.70$, $a_{yx} = -2.261$, $M_x = 1.97$, $\bar{S}_{yx} = 8.28$, $\sigma_x = 0.73$, $M_y = M_{y'} = 30.64$, $\sigma_y = 14.43$.

$$M_{y'} = -2.261 + (16.70)(1.97) = 30.64$$

$$\sigma_{M_{y'}} = \frac{8.28}{\sqrt{14}} = 2.21$$

$$\sigma_{b_{yx}} = \frac{8.28}{0.73\sqrt{14}} = 3.03$$

Since the estimated value, Y' equals $M_{y'} + b(x)$, the standard error of the estimate for any value of x will be composed of the sum of the standard errors of $M_{y'}$ and of $b(x)$. Standard errors are standard deviations; hence they can be summed only by adding their squares (as demonstrated in Appendix 2, Note 1). The standard error of Y', for any particular value of x, is therefore given by the equation [5]

$$\sigma_{y'} = \sqrt{\sigma^2_{M_{y'}} + (\sigma_{b_{yx}}x)^2} \tag{70.1}$$

By using this relation, the calculation of the standard error of Y', for selected values of X, is shown in the following tabulation:

Selected values of X	Departures from mean x	Calculation of $\sigma_{y'}$				
		$\sigma_{b_{yx}}x$ $= (3.03x)$	$(\sigma_{b_{yx}}x)^2$	$\sigma^2 M_{y'}$ $= (2.21)^2$	$\sigma^2_{y'}$ $(\sigma_b x)^2$ $+\sigma^2 M_{y'}]$	$\sigma_{y'}$
0.97	−1.00	−3.030	9.1809	4.8841	14.0650	3.75
1.47	−0.50	−1.515	2.2952	4.8841	7.1793	2.68
1.97	0	0	0	4.8841	4.8841	2.21
2.47	0.50	1.515	2.2952	4.8841	7.1793	2.68
2.97	1.00	3.030	9.1809	4.8841	14.0650	3.75
3.47	1.50	4.545	20.6570	4.8841	25.5411	5.05
3.97	2.00	6.060	36.7236	4.8841	41.6077	6.45

There are 14 cases; subtracting the one extra constant involved in correlation determinations gives 13 as the number of observations with which to judge from Table A the significance of these standard errors. Taking values midway between those for 10 and for 16 cases, we find that the statement that the true values of b_{yx} and of $M_{y'}$ do not differ from the observed values by more than the calculated standard errors will be wrong for 34 out of each 100 such statements, on the average. Similarly, the statement that they do not differ by more than twice the calculated standard errors will be wrong for 7 out of 100 such statements, on the average. The chances are therefore 93 out of 100 that the true regression line would fall within twice the standard errors just calculated. Plotting $2\sigma_{y'}$ above and below the corresponding values of Y', given by the regression line, shows this range. These limits are

[5] Holbrook Working and Harold Hotelling, Applications of the theory of error to the interpretation of trends, *Journal of the American Statistical Association Papers and Proceedings*, xxiv, pp. 73–85, March supplement, 1929.

plotted in Figure 59, together with the original observations and the regression line. The limits within which the line probably fell could be shown in a similar manner for any other desired limit of probability. It is now clear why great caution must be exercised in extending even a linear regression' line beyond the range of the data from which it is derived. As is evident in the figure, the true position of the line becomes very uncertain as the limits of the data are approached, and increases rapidly beyond them.

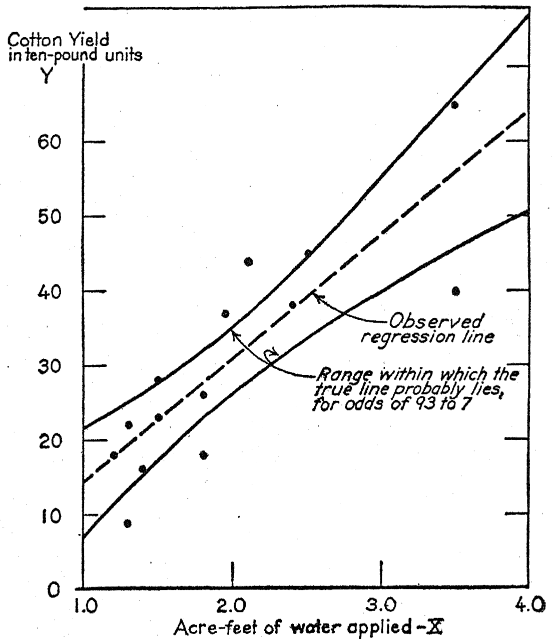


Fig. 59. Linear regression of cotton yield on irrigation water applied, and range within which the true relation probably lies.

In many correlation problems, the regression line is the most important result of the study. The confidence that can be placed in the line determined from a random sample is no greater than is indicated by the probable error of its slope, or the standard error zone of its position. Accordingly, the final statement of the regression coefficient or regression line should always indicate clearly the standard error or probable error zone, and should also state the number of observations on which the conclusions are based. This will serve to caution the reader of the extent to which the values may vary from the true value simply due to chance fluctuations of sampling, and so caution him not to attach more importance to them than their significance justifies.

Correlation coefficients. In exactly the same way that regression coefficients will vary from sample to sample, all other statistical constants tend to vary. Regression coefficients from random samples tend to be normally distributed around the true value, so that the probability of a given departure from the true value occurring may be judged from the normal curve;[6] but that is not equally true of correlation coefficients. If the number of observations in the sample is exceedingly large, so that fairly stable results are secured, the distribution of the observed correlations will tend to be nearly normal, so that the standard error may be estimated by the formula[7]

$$\text{Standard error of } r_{yx} = \frac{1 - r^2}{\sqrt{n - 2}} \tag{71}$$

This equation applies only when n is large, say 100 or more. To test the significance of correlation coefficients obtained from small samples, Fisher has developed the equation

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \tag{71.1}$$

The value t is used to judge the probability of the occurrence of such a correlation purely by chance, in exactly the same way that the number of times an average is times its standard error is used to judge the probability of the significance of the average. Thus if a correlation of 0.60 is secured with a sample of 21 cases, $t = 3.26$. Looking up this value in Table A on page 23, or Figure A of Appendix 3, using 20 for n,[8] we find that only in one sample out of 200 random samples, on the average, would a value this large or larger be obtained from a universe with no correlation present. If, however, a correlation of 0.60 had been secured with only 7 cases, t would equal

[6] The normal curve is the basis for the probability data given in the last column of Table A of Chapter 2.

[7] Equation (71) holds precisely true only when the value used for r is the true correlation in the universe, rather than the value observed in the sample. This limitation does not apply to equation (71.1).

[8] Just as with regression coefficients, 1 less than the number of cases should be taken for n when Table A is used to judge the significance of a correlation coefficient. The unadjusted correlation, r, should be used in all tests of significance, *not* the adjusted value $\bar{r}$.

1.68. Figure A indicates that, with this value of t, the chances of getting a correlation this large or larger from random samples drawn from a universe with no true correlation would be almost 0.16. This means that out of 100 such samples obtained from a universe in which the true correlation was zero, 16, on the average, would show a correlation as high as 0.60.[9]

Although this method may be used in conjunction with Table A to determine whether or not the correlations computed from small samples are any valid indication of a correlation in excess of zero, it cannot be used to determine the significance of the difference in correlation between two samples or to determine whether or not the correlation in a given sample exceeds any specific value. In the first illustration, for example, where $r = +0.60$, one might wish to know the probability that the true correlation in the universe exceeds $+0.20$. Owing to the skewed distribution of values of r when computed from small samples, this cannot be determined by a simple sampling formula. R. A. Fisher has devised a method, however, of so transforming observed values of r as to give them a normal distribution, and then solving such problems as this from the transformed values. For methods of dealing with this phase of sampling, the reader is referred to his presentation of the method in Statistical Methods for Research Workers, seventh edition, pages 202 to 211.

Certain of Fisher's methods to determine the reliability of observed correlations may be put into more simple form for general use, as shown in Figure B in Appendix 3. This figure is based upon the idea that, although we cannot state the true correlation existing in the universe from the correlation shown in a given sample, we can estimate a minimum value for the true correlation, with a given chance of being wrong. Figure B has been calculated, by Fisher's methods, to show such probable minimum correlations in the universe, with the probability that the statements based on the figure will be wrong for 1 sample out of 20, on the average. The results have been plotted for different sizes of sample and observed correlations. Thus if a random sample of 20 gives an observed correlation of 0.70, the figure shows at a glance that we can say that the true correlation is greater than 0.44, with the expectation that such statements will be wrong only once in twenty times, on the average. Similarly, for an observed correlation of 0.55 with a sample of 35 cases, reading from the line

[9] See R. A. Fisher, *Statistical Methods for Research Workers*, seventh edition, Oliver and Boyd, London and Edinburgh, 1938, pages 197 to 202, for a fuller discussion of the use of t in judging the reliability of correlation coefficients.

for observed correlation = 0.55, and interpolating between $n = 30$ and $n = 40$, gives 0.32, which means that we can say that the true correlation is greater than 0.32, with the same degree of confidence. The figure can be used in a similar manner for any other size of sample up to 100, and any observed correlation.

Figure B deserves close study, for it tells a great deal about the sampling reliability, or, rather, unreliability, of correlation coefficients. The bottom line, for example, shows that, when samples are drawn from a universe where the true correlation is zero, 1 sample out of 20 will show a correlation as high as ± 0.60, on the average, with samples of 10 cases; as high as $\pm$ 0.49, with samples of 15 cases; and as high as $\pm$ 0.35, even with samples of 30 cases. Similarly, if the samples are drawn from a universe where the true correlation is 0.50, 1 sample out of 20, on the average, will show a correlation as high as 0.81, with samples of 10; as high as 0.73, with samples of 20; and as high as 0.69, with samples of 30. Many other similar comparisons can be made readily. For example, if the true correlation is 0.80 and samples of 10 cases are used, 5 per cent of the samples will show correlations as high as 0.93. These facts do not take into account the tendency of many students to examine a number of possible independent variables and to select for more detailed study those which show the highest correlation with the independent factor. If that is done, the possible minimum correlation in the universe, corresponding to the correlation observed in the sample so selected, will be even lower than would be estimated from Figure B.

Correlation indexes. The reliability of indexes of (curvilinear) correlation, ρ, determined from very large samples, may be judged by the use of the following equation:

$$\text{Standard error of index of correlation} = \frac{1 - \rho^2}{\sqrt{n - m}} \tag{72}$$

In using Table A to test the significance of such correlations for small samples by the t method, we must deduct 1 less than the number of constants necessary to represent the regression line mathematically [the value m of equation (26) minus 1] from the number of cases before using Table A. Thus if a correlation index computed for a cubic parabola fitted to 7 observations were to be judged, its reliability would be determined by using the column headed 4. Since 4 constants would be represented in the regression equation, $7 - (m - 1) = 4$. If the computation gives $t = 2.8$, Table A (or Figure A) indicates that 7 out of 100 such samples, on the average, would give a correlation as high or higher than the observed correlation, even if there were no true correlation in the universe.

Empirical studies of the sampling variability of indexes of correlation indicate that they tend to be skewed in their distribution, just as do coefficients of correlation; therefore parallel special methods must be employed in judging their significance. Figures C, D, and E, on pages 507 to 509, prepared to apply to multiple correlation coefficients in the same way that Figure B applies to simple correlation coefficients, may be used tentatively to judge the reliability of indexes of correlation, until more exact measures have been developed. Where $m = 4$, Figure C (for $R_{1.234}$) may be used; where $m = 6$, Figure D (for $R_{1.23456}$); and where $m = 8$, Figure E (for $R_{1.2345678}$). These figures, also, are based upon methods developed by R. A. Fisher.

Multiple Correlation

Coefficients of multiple correlation and net regression. Correlation constants derived from multiple regression studies are even more subject to chance variation than are those from simpler analyses. In a random sample of 30 cases drawn from a known universe, for example, the following values were obtained:

$$R_{1.234} = 0.538; \; b_{12.34} = 0.583; \; b_{13.24} = 0.366; \; b_{14.23} = 0.949$$

By drawing 15 more random samples of 30 cases each, 10 of 50 cases, and 5 of 100 cases, values were secured for R and the b's as shown in the following statement and in that on the next page.

Distribution of Values for Multiple Correlation Coefficients for Repeated Samples Drawn from the Same Universe (True Value 0.563)

Range of values	30 observations	50 observations	100 observations
0.300–0.399	4	1	
0.400–0.499	3	5	
0.500–0.599	4	1	4
0.600–0.699	4	3	1
0.700–0.799	1		

From these two tables we can see how the variation decreases as the number of cases increase, and can also see what values the constants from the several samples tend to center around, and so estimate the approximate true value. But some definite idea of the range within which this true value probably would lie could have been obtained by computing the standard error of each constant by means of the formulas:[10]

$$\left.\begin{array}{l}\text{Standard error for a coefficient cf mul-}\\\quad\text{tiple correlation } R_{1.234\,\ldots\,n}\end{array}\right\} = \frac{1 - R^2_{1.234\,\ldots\,n}}{\sqrt{n - m}} \qquad (73)$$

$$\left.\begin{array}{l}\text{Standard error for a coefficient of par-}\\\quad\text{tial regression } b_{12.34\,\ldots\,n}\end{array}\right\} = \sqrt{\frac{\bar{S}^2_{1.234\,\ldots\,n}}{n\sigma^2_2(1 - R^2_{2.34\,\ldots\,n})}} \qquad (74)$$

Reliability of multiple correlation coefficient. The standard error for the value of R, 0.538, given for our first sample works out to be

Distribution of Values for Net Regression Coefficients for Repeated Samples Drawn from the Same Universe

Range of values	30 observations	50 observations	100 observations	True value
Values for $b_{12.34}$:				
−0.79 to −0.60.......	1			
−0.59 to −0.40.......	0			
−0.39 to −0.20.......	1			
−0.19 to −0.00.......	0	2		
0 to 0.19.......	2	1	1	
0.20 to 0.39.......	6	4	4	+0.320
0.40 to 0.59.......	4	3		
0.60 to 0.79.......	3			
Values for $b_{13.24}$:				
−0.19 to − 0.......	2			
0 to 0.19.......	3			
0.20 to 0.39.......	5	6	2	+0.377
0.40 to 0.59.......	2	2	2	
0.60 to 0.79.......	1	1	1	
0.80 to 0.99.......	2	1		
1.00 to 1.10.......	1			
Values for $b_{14.23}$:				
0 to 0.19.......				
0.20 to 0.39.......		1		
0.40 to 0.59.......	1	1		
0.60 to 0.79.......	8	2	2	
0.80 to 0.99.......	3	4	3	+0.824
1.00 to 1.19.......	4	1		
1.20 to 1.39.......		1		
1.40 to 1.59.......				

[10] The standard errors of the several net regression coefficients can be determined at the same time that the regression coefficients are determined, and as part of the same set of computations. See Appendix 1, "Methods of Computation."

0.139. If we ignore the fact that the distribution of R, just as of r, is not normal, we may interpret that roughly by saying that in 2 out of 3 such samples, on the average, the true value of R in the universe will be within the range $R \pm \sigma_R$, or between 0.40 and 0.68. As it happens for this particular sample, the true value, 0.563, does lie within this range. Ten of the 16 samples, or 63 per cent, gave values falling within 0.139 of the true value, so the computed standard error is not so misleading in this particular case. For still smaller samples, or for higher correlations, the standard error computed by equation (73) would be less reliable. For such cases we would use instead the equation:

$$t = \frac{R_{1.234}\sqrt{n - m}}{\sqrt{1 - R^2_{1.234}}} \tag{74.05}$$

This equation is used together with Table A to judge whether there is real evidence that the true correlation exceeds zero, just as equation (71.1) is used in the case of the correlation coefficient and index. In using Table A, $m - 1$ must be subtracted from the number of observations. (This also applies in using Table A for coefficients of net regression.) For more exact interpretations, Fisher's transformation method, previously referred to, may be utilized.[11]

For small samples, the reliability of coefficients of multiple correlation varies not only with the correlation and the size of sample, but also with the number of independent variables. Fisher has developed an exact method for judging the probable significance of observed coefficients of multiple correlation.[12] Figures C, D, and E on pages 507 to 509 provide a simple method of applying his conclusions for multiple correlation coefficients, in the same way that Figure B provides for simple correlation coefficients. For problems involving 3, 5, and 7 independent factors, respectively, these figures show the approximate minimum true correlation that probably exists in the universe with any size of sample up to 100, and for any observed correlation, with the probability that the statements based on the figure will be right for 19 samples out of 20, on the average. Thus if, with 30 observations, a correlation of $R_{1.23456} = 0.80$ should be obtained, we can say that the true correlation (from Figure D) is at least 0.58. Similarly, if

[11] See pages 469 to 474 in Appendix 1, "Methods of Computation," for the most effective method of computing the standard errors of net regression coefficients, according to equation (74).

[12] R. A. Fisher, The general sampling distribution of the multiple correlation coefficient, *Proceedings of the Royal Society*, A, Vol. 121, pp. 654–673, 1928.

for 50 observations a correlation of $R_{1.234} = 0.62$ were obtained, Figure C gives 0.42 as the probable minimum correlation in the sample. These conclusions, of course, apply only if the conditions of random sampling are fulfilled. Problems with 2, 4, or 6 independent variables may be considered by interpolating between the corresponding values given for 1, 3, 5, or 7 independent variables.

Considering the problem mentioned above, where a sample of 30 observations showed $R_{1.234} = 0.538$, Figure C gives a value of 0.16 as the probable minimum correlation. From the single sample we could then say that the true correlation is probably at least 0.16 in the universe from which the sample was drawn, with one chance in twenty of being wrong.

Figures C, D, and E show the possibilities of getting high correlations from a random sample, even when there is little or no correlation in the universe from which that sample was drawn. Thus for three independent variables, Figure C shows that, if samples of 15 observations are used, in 1 sample out of 20, $R_{1.234}$ will be as large as 0.69, even if the correlation in the universe is zero, and as large as 0.78, even if the true correlation in the universe is only 0.40. Similarly, if there are 7 independent variables, Figure E shows that, if samples of 20 cases are used, in 1 sample out of 20, on the average, $R_{1.2345678}$ will be as high as 0.79, with zero correlation in the universe; 0.85, with 0.50 in the universe; and 0.91, with 0.70 in the universe. Even with samples as large as 100 cases, $R_{1.2345678}$ in 5 per cent of the samples will be as high as 0.37 for samples drawn from a universe with zero correlation, and as high as 0.57 for samples drawn from a universe with 0.40 as the true correlation. Figure D gives similar probabilities for 5 independent variables. Many other combinations of size of sample, true correlation in the universe, and observed correlation for 5 per cent of the samples are given in these figures.

If the several independent variables in the multiple correlation had been selected by considering a large number of possible independent variables, and by retaining only those which showed the highest gross or net correlation with X_1, there is a much larger possibility of the correlation in the sample exceeding the true correlation in the universe by a wide margin. In fact, it is almost certain to be erroneously high. If error calculations are to be used to judge the sampling significance of the correlations or regressions observed, the variables must be selected purely on logical or deductive grounds (as discussed at length in Chapter 24), rather than on any such basis of empirical selection of those which show the apparent closest relation.

It should always be remembered that, if the choice is purely empirical, the next following period might readily reverse the order of apparent importance of the several variables.

Reliability of net regression coefficients. Turning to the meaning of the regression coefficients, we may illustrate the case with one constant, $b_{12.34}$. The value given by the original sample was 0.538. For that sample $\sigma_2 = 2.53$, $\bar{S}_{1.234} = 2.81$, and $R_{2.34} = 0.708$. If these values are substituted in equation (74), the standard error works out to be 0.287. The observed regression may therefore be stated to be 0.538 ± 0.287. This indicates that we can say that the true regression probably lies between 0.251 and 0.825, with the expectation that such statements will be right two times out of three, on the average; or we can say that it lies between -0.036 and 1.112, with the expectation that such statements will be wrong only one time out of twenty (0.045). Actually, the true value in this case was 0.320, or within the first range. It may be noted that 11 of the 16 samples showed regression coefficients for $b_{12.34}$ within 0.287 of the true value, and all but one fell within 0.574 of the true value. Again this illustrates how the variability of constants which tend to be normally distributed may be estimated by appropriate error formulas, and hence how the reliance to be placed in conclusions from a given sample may be judged.

From equation (74) it is evident that the reliability of a net regression coefficient varies directly with the multiple correlation of the dependent factor with the other factors, but inversely with the multiple correlation of the particular independent factor with the other independents. The more closely a particular independent factor can be estimated from the other independent factors present, the less accurately can the net relation of the dependent factor to it be determined.

The qualifications the use of this error formula throws around regression results may be illustrated in a problem where the theory of sampling is fairly applicable, namely, the relation between the feed a herd of cows receives and the resulting milk production. Table 75 shows these results for two different studies.

This table illustrates two points: first, that the regression results are not very accurate even with a multiple correlation of 0.80; and, second, that the reliability of the regressions varies from variable to variable, being much greater in some cases than in others. It is obvious that some of the regressions would have no statistical significance at all, whereas others would indicate the probable relations within a fairly close range of accuracy.

Thus for the percentage of lime, with the P.E. = 67.4 per cent of the regression, there is 1 chance out of 2 that the true net regression varies from that observed in this sample by two-thirds of the observed value, and 1 chance out of 6 that the true net regression is of opposite sign from that observed. With the total digestible nutrients, on the other hand, with the probable error only 12 per cent of the observed value, there is but little chance that the observed value differs from the true regression by more than 30 per cent, and very little chance that it differs as much as 40 per cent.

If the regression equation is to be used solely as a basis for making new estimates of the value of the dependent factor to be expected for given values of the independent factors, then the accuracy of the several regression coefficients does not make such a great difference. Any deficiency in one may be compensated for by an excess in another. (This does not hold true, however, if estimates are made for extreme values of variables whose regressions are subject to large errors. See Chapter 19 on this point.) But if the major interest is not in the total estimate, but in the changes in the dependent factor with changes in each particular independent factor, then the reliability of each particular regression coefficient becomes of real importance. In the illustration cited, for example, it would not do to know merely that the milk production per cow varied both with protein content and with lime, if it was desired to know how much to allow for protein and how much for lime in compounding a ration. Instead, the probable errors indicate that the influence of protein (as represented in the "nutritive" ratio) has been fairly accurately measured, whereas the influence of lime has not been accurately measured at all. Not much confidence therefore can be placed in the conclusions as to this latter factor.

In any correlation study where the results are based upon a sample of observations drawn at random from a known universe, and where any importance is to be attached to the values found for the several regression coefficients, it is essential that the standard errors of each of those coefficients be determined and considered. As is illustrated in the examples just discussed, a sample may have a very significant multiple correlation and yet yield regression coefficients for some variables which are almost entirely the result of chance fluctuation, and therefore of little or no significance. This may occur even with moderately large samples, such as the sample of 95 cases in the first example just considered. Computation, presentation, and discussion of the standard errors of the regression coefficients are therefore vital parts of any such multiple correlation study.[13]

TABLE 75

Probable Errors of Partial Regression Coefficients, in Per Cent of the Value of the Coefficient *

Item	Wisconsin study	Minnesota study
Number of observations	95	77
Number of variables	10	8
Multiple correlation, adjusted for number of variables	0.805±0.039	0.862±0.034

Probable Error of Regression Coefficients †

Independent variable	Per cent	Per cent
Total digestible nutrients	12.0	11.5
Nutritive ratio	12.4	9.5
Per cent of protein "good"	28.3	
Per cent of lime	67.4	
Per cent summer feeding	17.5	
Per cent silage	21.5	13.7
Fat test of milk	10.6	3.7
Per cent fall freshening	18.5	11.8
Value per cow	26.8	
Age of cows	....	17.9
Per cent grain in ratio	....	20.2

* Mordecai Ezekiel, The application of the theory of error to multiple and curvilinear correlation, *Journal of the American Statistical Association,* Vol. XXIV, No. 165 A, March, 1929, Supplement, p. 103.

† The coefficients are for the net regression of milk production on the factors stated. P.E. = 0.6745 of the standard error.

Multiple curvilinear correlation. All the formulas for multiple regression constants cited up to this point have been derived for

[13] For illustrations of ways of presenting net regression coefficients, together with their standard errors, see M. J. B. Ezekiel, P. E. McNall, and F. B. Morrison, Practices responsible for variations in physical requirements and economic costs of milk production on Wisconsin dairy farms, *Agricultural Experiment Station of Wisconsin Research Bulletin* 79, August, 1927, pp. 21–23; and Kathryn H. Wylie and Mordecai Ezekiel, The cost curve for steel production, *Journal of Political Economy,* Vol. XLVIII, pp. 792–93, December, 1940.

linear correlation use. For curvilinear multiple correlation results, however, no measures of the probable error have yet been devised for the freehand process by logical and mathematical deduction. Experiments mentioned previously were initiated to provide at least some empirical measures of reliability. The results indicate that the *index* of multiple correlation must be corrected for the number of constants involved or assumed just as much as the coefficient of multiple correlation, as has already been illustrated.

The reliance to be placed on regression curves requires separate treatment. Where those curves are determined by fitting mathematical functions, the probable accuracy with which the true relation is expressed by the mathematical curve may be judged by error formulas which have been worked out mathematically by an extension of the same methods upon which those previously presented were based. For regression curves determined by the successive approximation process or by the graphic approximation process, no such mathematical treatment is possible. Experimental study of the reliability of regression curves determined by successive approximations, however, has thrown some light on the reliability of such curves and made it possible to state the following general principles:

First, the reliability of regression curves appears to vary inversely with the standard error of estimate for the entire sample.

Second, the reliability of any point on a regression curve appears to vary directly with the square root of the number of observations on which that portion of the curve was based.

Third, the reliability of any point on a regression curve, when stated as the difference between the value of the function at that point and the value of the function at the point corresponding to the mean of the independent variable, appears to vary inversely with the square root of the distance the selected point is from the mean of the independent variable, measured in units of the standard deviation of the independent variable.

All these points apply equally to simple regression curves and net regression curves, computed while holding the influence of other factors constant. For net regression curves, one further point is involved, the extent to which one independent factor tends to vary with the other independent factors, which may be stated: Fourth, the reliability of points on a net (or partial) regression curve appears to vary inversely with the multiple curvilinear correlation of the particular independent factor with the other independent factors.

The following formulas give a rough approximation to the standard error of net regression curves. These formulas express the four points just mentioned. In experimental work, these formulas, when computed from the results shown by individual samples have, on the average, successfully indicated the range within which the true regression curves lie 17 times out of 20 (using a range of twice the computed standard error). The proportion of very large errors, up to 5 or more times the computed standard errors, has been larger than would be expected from a normal distribution of errors. These preliminary formulas may leave out some essential element in occasional cases, or the results of graphic freehand curve fitting may show errors in exceptional samples out of proportion to those ordinarily made.[14]

The formulas are:

$$\sigma_{f(X)-f(X_M)} = \sqrt{\frac{\overline{S}^2_{y.f(x)}ux}{\sigma_x^2 n_u}} \tag{74.1}$$

$$\sigma_{f_{12.34}(X_2)-f_{12.34}(X_{M2})} = \sqrt{\frac{\overline{S}^2_{1.f(2,3,4)}ux_2}{\sigma_2^2 n_u(1-\overline{\text{P}}^2_{2.34})}} \tag{74.2}$$

Since several new symbols are introduced in these two equations to cover the points which have been enumerated, they will first be defined.

The symbols have exactly the same meaning for both equations, except for the additional term $(1 - \overline{\text{P}}^2_{2.34})$ in equation (74.2) for regression curves determined by multiple correlation. The standard errors of estimate, $\overline{S}_{y.f(x)}$, and $\overline{S}_{1.f(2,3,4)}$ have the same meaning as defined in equations (21.1) to (22.2) and (64) or (65); σ_x and σ_2 are the usual standard deviations of the independent variable. The new terms have the following meanings:

$f(X_M)$ means the reading from the regression curve $f(X)$ for the point where X is equal to M_x.

$f_{12.34}(X_{M_2})$ means the reading from the net regression curve $f_{12.34}(X_2)$ for the point where $X_2 = M_2$.

n_u represents the number of observations falling within some selected group interval of X or X_2, with X_a, the point for which the accuracy of the curve is to be determined, at the center. This interval

[14] The derivation of these formulas is given in Mordecai Ezekiel, The sampling variability of linear and curvilinear regressions, *Annals of Mathematical Statistics*, September, 1930.

must be taken large enough to include the observations which were taken into account in determining the shape of that portion of the curve, yet small enough not to take in observations whose values did not enter into the determination of that part of the curve.

u designates the range over which n_u is taken, stated in units of the independent variable X or X_2, and using the same units as those in which the standard deviation, σ_2, is stated. Thus if the standard deviation is in units of pounds or dollars, u is also stated in pounds or dollars.

The other term, x or x_2, has the same meaning as used previously—the deviation of the independent variable from the mean of that variable, stated in the same terms as the standard deviation is stated in. Thus for the point along the curve where the independent variable has the value X_a, $x = X_a - M_x$. There is this difference from the usual usage, however,—in equations (74.1) and (74.2) x and x_2 are to be taken as positive numbers without regard to sign.

The several steps in working out the reliability of a regression curve, and the meaning of the results, may be illustrated by applying these equations to one of the curves previously determined.

The reliability of the regression curve worked out for cotton yields in Chapter 8 may be tested by equation (74.1). The curve obtained (Figure 23), on page 154, shows that with an average application of water, 1.97 feet, a yield of 328 pounds of cotton would probably be obtained, whereas with an application of 1.4 feet of water, a yield of 195 pounds would probably be obtained. Apparently reducing the application of water 0.57 foot would reduce the yield of cotton by 133 pounds. How accurate is this last conclusion?

Picking out the values necessary to compute the standard error according to equation (74.1), we have $\overline{S}_{y.f(x)} = 80.7$ pounds, $\sigma_x = 0.73$ foot, and $x = 0.57$ foot. Noting that the average yield in the groups of 1 to 1.4 feet of water and 1.5 to 1.9 feet of water both had some influence in determining the position of the curve at 1.4 feet, we may let the interval for which we take the number of cases extend half way into the upper group, or to 1.7, and an equal distance below 1.4, or to 1.1. The number of items for n_u, then, will include all the cases having 1.05 or more feet of water applied, and less than 1.75 feet applied, a range of 0.70 foot. The number of cases falling in this range (Table 31 on page 000) is found to be 6; so $n_u = 6$ and $u = 0.70$. Substituting these several values in equation (74.1), we find the arithmetic to be as follows:

Standard error of decrease of 133 pounds

$$=\sqrt{\frac{\overline{S}^2_{y\cdot f(x)}ux}{\sigma_x^2 n_u}}=\sqrt{\frac{(80.7)^2(0.70)(0.57)}{(0.73)^2(6)}}=\sqrt{\frac{(6523)(0.70)(0.57)}{(0.5348)(6)}}=\sqrt{811.11}$$

Standard error = 28.5

The difference of 133 pounds, therefore, has a standard error of 29 pounds. The statement that the reduction of 0.57 foot in water applied reduces yields by 133 pounds, therefore, really means that the reduction is probably between 104 and 162 pounds, but there is at least 1 chance in 20 that it is as little as 77 pounds, or as much as 189 pounds. That is, if we make the statement that the true value for all the farms in the universe lies between 77 and 189 pounds, we should be wrong in at least 1 out of 20 such statements, on the average. (Table A need not be considered in computing these chances, unless the total number of observations in the problem, $n - m$, is less than 30. Then $n - m$ should be used to find the column to determine the probabilities from, rather than n_u. In this case, with $n - m = 11$, the conclusions are not quite so reliable as these statements indicate.)

It should be noted that the estimated range of error would not have been changed very greatly if a different interval had been used for u. Had the range from 1.25 to 1.45 been taken instead, u would have been 0.30 and n_u would have been 5. If these values are substituted in equation (74.1) instead of the ones used previously, the standard error works out as somewhat lower than before. The greater the total number of observations, the less effect a change in u will have on the computed error—it is only the small number of cases and very irregular distribution that causes as considerable a difference as in this particular case—and even so, the indicated reliability is still of the same order.

To compute the range of error for the entire curve, we may pick out a number of selected points—say at each 0.2 foot of water—and work out the error for the reading at each of those points. The process may be shortened by noting that, in equation (74.1), the values of the several terms remain unchanged for every point along the curve, with the exception of u, n_u, and x; and that, if the same range is taken for u at each point, only the two other values are changed. Accordingly, equation (74.1) may be restated as follows:

$$\sigma_{f(X)-f(X_M)} = \sqrt{k\frac{x}{n_u}}; \quad \text{where} \quad k = \frac{\overline{S}^2_{y\cdot f(x)}u}{\sigma_x^2} \qquad (74.11)$$

Since the value of k is the same for every point along the curve, it can be worked out once for all; then all that needs to be computed at each point is $\frac{x}{n_u}$, the product $k\frac{x}{n_u}$, and the square root of the product.

The work of applying this process to the cotton-yield curve may be shown in tabular form. First the value of k must be worked out. If we continue to use the same range for u as before, 0.70 foot of water, the computation is:

$$k = \frac{\bar{S}^2_{y\cdot f(x)}u}{\sigma^2_x} = \frac{(80.7)^2(0.70)}{(0.73)^2} = 8{,}538$$

The next step is to enter X for each selected point, compute the value of x, and determine the value of n_u and of x/n_u; then multiply by k and extract the square root. These several steps are shown in Table 75A.

TABLE 75A

Computing the Standard Errors for Points Along a Regression Curve

X	x	n_u	$\frac{x}{n_u}$	(Error)^2 $k\left(\frac{x}{n_u}\right)$	Error $\sqrt{k\frac{x}{n_u}}$
1.2	0.77	6	0.12833	1095.71	33.1
1.4	0.57	6	0.09500	811.11	28.7
1.6	0.37	8	0.04625	384.89	19.6
1.8	0.17	6	0.02833	241.91	15.5
2.0	0.03	5	0.00600	51.23	7.1
2.2	0.23	4	0.05750	490.94	22.1
2.4	0.43	3	0.14333	1223.78	35.0
3.5	1.53	2	0.76500	6531.57	80.8

The values of n_u are determined just as in the single case before, by taking all the cases falling within 0.35 foot above and 0.35 foot below the value of X selected. Thus for $X = 1.2$, there are 6 cases between 0.85 and under 1.55; whereas for $X = 2.4$, there are but 3 cases between 2.05 and below 2.75. The series of values in the n_u column add to considerably more than the total number of observations, since the range taken is such that there is considerable overlapping. This does not affect the final errors computed, however, since the unit selected for u tends to have no effect on the size of the computed error.

The importance of the errors shown in the last column of Table 75A may be judged by comparing them to the values to which they apply—the difference between the estimated cotton yield for the several values of X and the yield estimated for the mean value of X. Table 75B shows this comparison.

TABLE 75B

Deviation of Points on a Regression Curve from the Value for the Mean, and the Standard Errors

X	$f(X)$	$f(X)-f(X_M)$	Standard errors *	2.25 (standard error) †
1.2	145	−183	±33	±75
1.4	195	−133	±29	±65
1.6	248	− 80	±20	±44
1.8	293	− 35	±16	±35
1.9	328	0		
2.0	335	7	± 7	±16
2.2	376	48	±22	±50
2.4	414	86	±35	±79
3.5	543	215	±81	±182

* From Table 75A.

† For a case where $n - m = 11$, range from the true value within which 95 per cent of the sample values will fall, on the average.

The standard errors will have to be interpreted with respect to the total number of observations, adjusted by m. For this problem, $m = 3$, so Table A should be entered with 12. Interpolating, we find that for such samples a departure of more than one standard error from the true value is likely to occur 34 times out of 100, and a departure of more than twice the standard error is likely to occur about 8 times out of 100 (as compared to less than 5 times for a very large sample). To estimate the extent of the true differences in yield lying beyond the observed differences which will be exceeded only in one sample out of 20, on the average, it is necessary to add about ± 2.25 times the standard error to the observed differences. This value is accordingly entered in the final column of Table 75B.[15]

[15] In view of the rather rough approximation to the true standard error of the curves given by these formulas, this use of Table A may be a refinement which is hardly justified. As indicated before, 2 or 3 samples out of 20, on the average, may show departures exceeding twice the standard error, as calculated by this method.

Just as with a regression line, the value of Y corresponding to the mean value of X is not exactly certain. No special sampling study has been made of regression curves in this connection. It would hardly be correct to apply equation (70) directly to this, as the central value for a freehand regression curve is not as definitely determined as for a straight regression line. Accordingly, the errors may be interpreted only with respect to differences from the mean value, rather than with respect to actual values.

The regression curve with its computed standard error is shown in Fig. 60, together with the wider range to reduce the probability of error to 0.05. This chart indicates the interpretation which may be given to the computed errors. The inner borders indicate the range within which the relation probably lies, with the regression curve from 1 sample out of 3, on the average, differing from the true curve by more than the range shown; whereas the outer borders indicate the range within which the true curve probably falls, with at least 1 sample out of 20, on the average, giving a regression curve which differs from the true curve by more than the range shown.

It is now quite evident why the table showing the relation according to the regression curve, as set forth in Chapter 8, page 154, was not carried beyond 2.5 feet of water. In fact, it might be just as well not to carry it beyond 2.25 feet, to judge from Figure 60, for at 2.5 feet there is at least 1 chance out of 20 that the true difference in yield above the yield for the average water application differs from that estimated by nearly the difference shown in the estimate. The wide range of possible error for the true position of this curve reflects both the small number of observations upon which it is based and the relatively low correlation shown by those observations. Even so, the computed range of error indicates what degree of reliance can be placed in the findings under these limiting conditions, and so makes the results of the analysis of more value than if we had no knowledge of their probable stability.

It is evident from the illustration that certain portions of a regression curve may be much less accurately determined than certain other portions. It is not merely the total number of observations in the sample, but the way they are scattered or bunched along the curve which is fitted, which affects the reliability of the various portions of the regression curve.

The process of working out the standard error for a net or partial regression curve is exactly the same as that just illustrated, except that equation (74.2) is used instead of (74.1). The computation may be broken into two steps just as illustrated for simple correlation, as follows:

$$\sigma_{f_{12.34}(X_2)-f_{12.34}(X_{M_2})} = \sqrt{k'\frac{x_2}{n_u}}, \quad \text{where} \quad k' = \frac{\bar{S}^2_{1.f(2,3,4)}u}{\sigma_2^2(1-\bar{\mathrm{P}}^2_{2.34})} \qquad (74.21)$$

The multiple curvilinear intercorrelation of each independent variable with the remaining independent variables can be determined fairly rapidly by the use of the short-cut graphic method. Thus for the

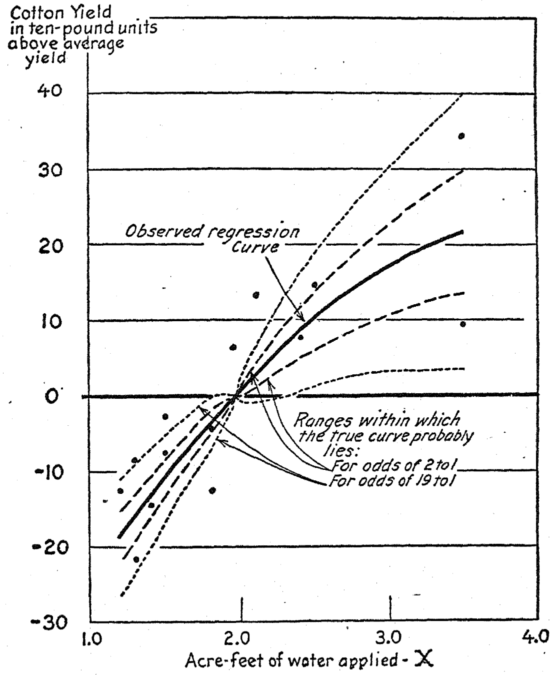


FIG. 60. Curvilinear regression of cotton yield on irrigation water applied, and range within which the true relation probably lies.

second problem used in Chapter 16, computation of the standard error zone for the several independent variables involves determination of the index of multiple correlation, $\bar{\mathrm{P}}$, for each of the three supplementary regression relations, as follows:

(A) $\bar{\mathrm{P}}_{2.34}$ from $X_2 = f_{23.4}(X_3) + f_{24.3}(X_4)$

(B) $\bar{\mathrm{P}}_{3.24}$ from $X_3 = f_{32.4}(X_2) + f_{34.2}(X_4)$

(C) $\bar{\mathrm{P}}_{4.23}$ from $X_4 = f_{42.3}(X_2) + f_{43.2}(X_3)$

In these three regression equations, to prevent confusion the same notation has been used for the subscripts to the "f's" in designating the several net regression curves as is used ordinarily in distinguishing the several net regression coefficients.

After the three sets of regression curves, (A), (B), and (C), are determined by the short-cut process, the final residuals for each may be read off from the final charts, the values of σ_2 and z_2''', σ_3 and z_3''', σ_4 and z_4''' determined, and the values of $\overline{\text{P}}_{2.34}$, $\overline{\text{P}}_{3.24}$, and $\overline{\text{P}}_{4.23}$ computed from these by equation (66.3), just as was illustrated in Chapter 16. With these three values, and the standard deviations of all the variables, we can then compute the standard error zone for each net regression curve by equation (74.21), carrying through for each variable in turn computations similar to those just indicated.

TABLE 75C

Computing the Standard Errors for Points along the Net Regression Curve $f_{12.34}(X_2)$

X_2	x_2	n_u	$\dfrac{x_2}{n_u}$	(Error)^2 $k'\left(\dfrac{x}{n_u}\right)$	Error $\sqrt{k'\dfrac{x}{n_u}}$
25	37.97	3	12.66	24.41	4.9
35	27.97	4	6.99	13.48	3.7
45	17.97	3	5.99	11.55	3.4
55	7.97	2	3.98	7.67	2.8
65	2.03	5	.41	.79	0.9
75	12.03	7	1.72	3.32	1.8
85	22.03	7	3.15	6.07	2.5
95	32.03	4	8.01	15.44	3.9

Carrying out these curvilinear correlation analyses, we obtain values of $\overline{\text{P}}^2_{2.34} = 0.6986$ and $\overline{\text{P}}^2_{3.24} = 0.4809$. The standard deviations are also computed, giving $\sigma_2^2 = 513.76$ and $\sigma_3^2 = 44.43$. The $M_2 = 62.97$, and $M_3 = 69.87$. The values of $\overline{S}^2$ and σ_1^2, page 294, are 14.93 and 51.70, respectively.

Using equation (74.21), we next calculate the value of k'. This involves deciding on the value of u to use. For X_2, where the observations run from 18.3 to 88.3, an interval of 20 seems appropriate, beginning at 15. Accordingly, k' becomes

$$k' = \frac{\overline{S}^2_{1.f(2,3,4)}u}{\sigma_2^2(1 - \overline{\text{P}}^2_{2.34})} = \frac{(14.93)(20)}{(513.76)(1 - 0.6986)} = 1.928$$

Table 75C shows the work set up in the same form as in Table 75A. The values in the n_u column are taken from Table 69A of Chapter 16, by calculating the frequency of X_2 in each 20-unit range around the X_2 values stated. Thus, for the group with $X_2 = 35$, there are four observations in the range from 25 to 45, and 4 is therefore the n_u value for this group. The next group, $X_2 = 45$, includes the range 35 to 55, with three observations. The fact that some observations are counted twice makes no difference, as that is allowed for by the inclusion of the u value in equations (74.2) and (74.21).

Similarly, for the regression $f_{13.24}(X_3)$, k' becomes

$$k' = \frac{\overline{S}^2_{1.f(2,3,4)}u}{\sigma_3^2(1 - \overline{P}^2_{3.24})} = \frac{(14.93)(10)}{(44.43)(1 - 0.4809)} = 6.473$$

Here, with X_3 varying from 59 to 86, units of 10 are used for u. The computation of the errors is as follows:

TABLE 75D

Computing the Standard Errors for Points along the Net Regression Curve $f_{13.24}(X_3)$

X_3	x_3	n_u	$\frac{x_2}{n_u}$	(Error)^2 $k'\frac{x}{n_u}$	Error $\sqrt{k'\frac{x}{n_u}}$
60	9.87	4	2.47	15.99	4.0
65	4.87	3	1.62	10.49	3.2
70	0.13	12	.01	.06	0.3
75	5.13	12	.43	2.78	1.7
80	10.13	1	10.13	65.57	8.1
85	15.13	1	15.13	97.94	9.9

The values for n_u are likewise obtained from Table 69A, taking the frequencies of X_3 in each 10-unit range around the X_3 values selected.

The standard errors as computed in Tables 75C and 75D could next be compared with the values to which they apply, by working out the departures of the net regression curves from their means just as was shown in Table 75B. That step will be omitted here. Instead, the errors are plotted graphically as ± departures from their respective net regression curves, as shown in Figure 61. In this case, with $n = 18$, and $m = 8$ (note page 293 of Chapter 16), we enter Table A (or Figure A) with 11 to find the significance of the departure. That gives us approximately 0.34. Accordingly, we conclude that in one sample out of three, on the average, the net regressions would miss the true regressions in the universe by larger amounts than those indicated in Figure 61 for this particular problem.

When we compare the zones of standard error in Figure 61 with the distribution of the original observations as shown in Figures 51 and 52 of Chapter 16, we see that the values of the independent variable for which the regression is fairly accurately determined are the values where the bulk of the observations fall. Thus, in the case of X_2, capacity operated, the observations are thickly clustered from 65 to 90, and thinly spread over the rest of the range. In consequence, the zone of standard error is narrowest in this region where relatively more observations were available to determine the slope of the curve. Similarly, for X_3, wage rates, the bulk of the observations fell between 70 and 75, with a thin scatter below 70 and with only one observation above 80. This distribution also is faithfully reflected in the standard errors, with a very wide error zone about 75, indicating that little is known of the slope of the regression curve in that range. The error equations (74.2) and (74.21) have this especial property of indicating

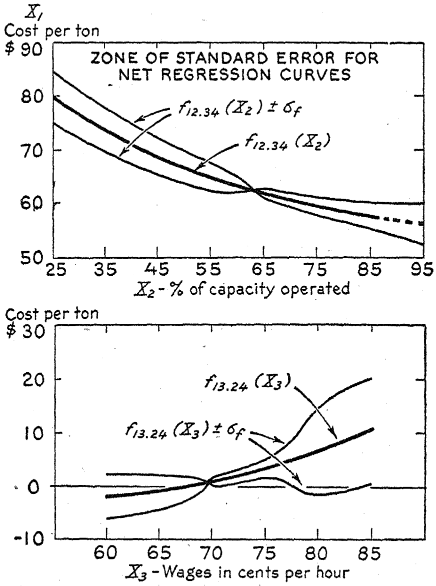


FIG. 61. Curvilinear net regressions of steel costs per ton on operation rate and wage rates, and standard error range for the net regressions.

the accuracy of determination of the various *portions* of each regression curve, in view of the adequacy with which that *portion* of the curve is represented by the distribution of the observations in the sample.

It will also be noted, in Figure 61, that if a horizontal line were passed through the mean of each curve (the point of zero error) it would fall entirely within the standard error zone for most of its length for $f_{13.24}(X_3)$, but would fall largely outside the zone of error for $f_{12.34}(X_2)$. That means that there is no certainty that there was any net relation between costs (X_1) and wages (X_3), whereas there is definite indication of a net relation between costs (X_1) and capacity operated (X_2). This result is secured even though the correlation of X_2 with X_3 and X_4 is materially higher than the correlation of X_3 with X_2 and X_4. The net relation found between X_1 and X_3 could readily have occurred by chance; there is much less possibility that the observed net relation between X_1 and X_2 could have been a chance result of the particular rates of operation which occurred during the years under study. The errors for $f_4(X_4)$ are not calculated, since there seems little real meaning in a standard error for a trend regression.

In a problem drawn from a time series, the meaning of error computations is less certain than in samples drawn at random from a true universe. Even in time series, however, the error computations may serve as some indication of the closeness within which the available data can locate the underlying relationships. (See also the next chapter for the significance of error formulas in time series.)

In all problems based on random sampling, where any generalizations as to the relations in the universe are to be based on the shape or slope of the final curves obtained by the graphic method, the error zones should be computed and should be given due consideration when the data are presented. This is just as important for curvilinear regressions as is the use of standard error values for linear regression coefficients.

Regression curves fitted mathematically. Where regression curves are obtained by fitting definite mathematical equations to the data, the standard error of the curve may be judged by the same methods previously presented for determining the probable errors of net regression coefficients. Thus if a parabola of the formula

$$X_1 = a + bX_2 + b'X_2^2$$

is determined, the standard errors of b and b' may be determined by equation (74), treating X_2 and X_2^2 as two independent variables.

The range within which the true curve probably lies may then be worked out just as has been illustrated for a linear regression. Similarly, if net regression curves are determined by fitting several mathematical equations simultaneously (as presented in detail in Chapter 22), an extension of this same method may be used to judge the reliability of each of the net regression curves so obtained.[16]

Summary. Coefficients of correlation and of regression, indexes of correlation, and regression curves, when determined from a limited sample, may depart more or less widely from the true value for the universe from which that sample was drawn. This chapter presents methods by which the possible extent of that variability may be judged. These methods represent an extension of the methods presented in Chapter 2 for judging the reliability of averages.

The methods of this chapter apply only when the correlations are determined from samples so selected as to comply with all the assumptions of random sampling. Where the samples are selected by other methods, the results may be of greater or of less reliability than if random sampling had been employed. Furthermore, in many types of problems, such as in time series, the observations can hardly be regarded as samples drawn from a universe. In such cases, statistical measures of reliability have a less precise meaning, but may still be valuable as a caution on the use of the results.

[16] See Henry Schultz, The standard error of a forecast from a curve, *Journal of the American Statistical Association,* pp. 139–185, June, 1930.

CHAPTER 19

THE RELIABILITY OF AN INDIVIDUAL FORECAST AND OF TIME-SERIES ANALYSES

The preceding chapter has indicated the kind of variability from sample to sample that may be expected in determining statistical constants, such as regression and correlation coefficients, and in determining regression lines and curves. It has provided means of estimating, from the values obtained from a single sample, various indications of how far and how frequently the results from successive samples of the same size are likely to vary from the true values in the universe from which the samples are drawn.

Reliability of an Individual Forecast

The practical statistician frequently has to deal with a quite different problem. Having taken a given sample, and having determined from that sample how the selected dependent variable is related to one or more independent variables, he then has the problem of drawing new observations of the same independent variable(s) from the same universe, and of estimating from those new values the most probable value of the dependent variable for the new cases. In series involving time relations, this becomes the problem of forecasting. In the corn-yield problem of Chapter 14, for example, it is possible to forecast the ultimate yield for the season as soon as the rainfall and temperature during the growing season are recorded. In problems involving crop production and price, it is possible to forecast the average price for the season as soon as the average crop is known. In these two cases, involving successive observations in time, theories of simple sampling do not apply rigorously, since the observations are not drawn fully at random. (See discussion of the time series problem later in this chapter.)

In other problems, however, sampling theory may be fully applicable. In a sample of children drawn at random from the school population of a given city, certain relations may be determined between their age and height and their weight. From these relations, how closely can we expect to estimate the weight of a new child, selected at random from the same population? In problems such as this, we are concerned with the possible difference between the estimated value, X'_1, and the actual value, X_1, for new observations drawn from the same universe as the sample. Heretofore we have calculated standard errors for the regression coefficient and line and standard errors of estimate for the observed errors in estimating X or X_1, in the sample under study. Also we have used methods of adjusting the standard error of estimate to obtain the most probable variation from the true regression line in the parent universe. The present problem, however, involves the accuracy of estimates made from the line or curve *obtained from the sample,* in the light of the possible sampling errors *of that line,* as compared to the true line, plus the possible range of errors of the estimates around the true line. What we need, therefore, is a means of combining the standard error of the regression line, σ_b, with the standard error of estimate, $\bar{S}_{1.23 \ldots n}$.

Simple correlation. For a simple two-variable correlation, the square of the standard error of a single estimate is given by the equation[1]

$$\sigma^2_{Y'-Y} = \sigma^2_{M_{y'}} + (\sigma_{b_{yx}}x)^2 + \bar{S}^2_{yx} \tag{75}$$

Applying this equation to the illustration used previously, on page 316, we can tabulate the calculation of various values as follows:

Selected values of X	Departures from mean, x	Calculation of $\sigma_{y-y'}$			
		$\sigma^2_{y'}$	$\bar{S}^2_{yx}$	$\sigma^2_{y-y'} =$ (3) + (4)	$\sigma_{y-y'}$
(1)	(2)	(3)	(4)	(5)	
0.97	−1.00	14.0650	67.6784	81.7434	9.0412
1.47	−0.50	7.1793	67.6784	74.8577	8.6520
1.97	0	4.8841	67.6784	72.5625	8.5183
2.47	0.50	7.1793	67.6784	74.8577	8.6520
2.97	1.00	14.0650	67.6784	81.7434	9.0412
3.47	1.50	25.541	67.6784	93.2194	9.6550
3.97	2.00	41.6077	67.6784	109.2861	10.4539

The last column gives the standard errors of estimate for values of Y' estimated from new values of X drawn from the same universe. It is apparent from these values that standard errors for individual

[1] The derivation of this equation is given in Note 13, Appendix 2.

forecasts near the mean of X are but little larger than $\bar{S}_{yx}$. Thus the standard error for the forecast of 26.8 for Y' when $X = 1.47$ is only $\sigma_{y-y'} = 8.65$, as compared with $\bar{S}_{yx} = 8.28$. The further the observed value of X departs from the mean, the larger the uncertainty of the individual forecast. Thus when $X = 3.97$, $\sigma_{y'-y} = 10.4539$. We can state this uncertainty of the estimate more simply by expressing the relation as follows:

$$\text{When} \quad X = 1.47, \quad \overline{\overline{Y}} = 26.8 \pm 8.65$$
$$\text{When} \quad X = 3.97, \quad \overline{\overline{Y}} = 67.6 \pm 10.45$$

Here we have introduced a new symbol, $\overline{\overline{Y}}$, to designate the probable range within which the true value will lie, for two estimates out of three on the average.

These standard errors of individual forecasts are interpreted in the same way as any other standard error, as indicating (for various selected multiples of the standard error) the proportion of a succession of such forecasts which will show departures from the true values of stated sizes. Thus, in the problem illustrated on pages 147 and 151, when yields are estimated for new plots with 3.97 feet of water applied, two out of three new observations, on the average, should show yields falling within 10.45 ten-pound units of the estimated yield. Table A, in Chapter 2, should be used in interpreting this standard error in exactly the same way it has been used before.

If we wish to know the ranges within which the actual value will agree with the forecasted values except for a specified proportion of the estimates, say 5 out of 100, we can determine those ranges by computing each one of them from the formula

$$Y = Y' \pm t\,\sigma_{y'-y} \tag{76}$$

The value to be used for t is obtained from Table A, page 23, or Figure A, in Appendix 3, by selecting the value which gives 0.05 as the proportion of cases. (In Table A, t corresponds to the values in the left-hand column; in Figure A, to the abscissas, shown across the bottom.) For example, assume that we were estimating the probable yield of cotton for a new plot where 2.97 acre-feet of water had been applied. The estimate, Y', is 51.9 ten-pound units. What is the true yield likely to be? Equation (70.21) then becomes

$$Y = 51.9 \pm (9.04)t$$

The regression line (page 148) was determined from a sample of 14 observations. The straight line involves two constants, so the n for

Table A, Chapter 2, $= 14 - 2 + 1 = 13$. Interpolating between the lines for $n = 12$ and $n = 16$ in Figure A (page 505) on the ordinate corresponding to 0.05, we find the corresponding abscissa gives t as 2.16. Upon substitution, the equation becomes

$$Y = 51.9 \pm (9.04)(2.16)$$

$$Y = 51.9 \pm 19.5$$

Accordingly, we estimate that the true yield will lie between 32.4 and 71.4 ten-pound units, or between 324 and 714 pounds, knowing that we are likely to be wrong only in one out of twenty such estimates, on the average.

Multiple correlation. The equation for the standard error of an individual forecast made from a multiple regression equation is similar to that given for simple correlation, with the addition of expressions for the additional variables, as follows:

$$\sigma^2_{x'_{1.234}-x_1} = \bar{S}^2_{1.234}\left[1 + \frac{1}{n} + c_{22}x_2^2 + c_{33}x_3^2 \right.$$
$$\left. + c_{44}x_4^2 + 2c_{23}x_2x_3 + 2c_{24}x_2x_4 + 2c_{34}x_3x_4\right] \quad (77)$$

In this equation x_2, x_3, and x_4 are the values of the independent variables for which the forecast is made, stated as departures from the respective means M_2, M_3, and M_4, as calculated in the original sample from which the regression equation was calculated. The c values for equation (77) are obtained by the simultaneous solution of the following equations:

$$\left.\begin{aligned}(\Sigma x_2^2)c_{22} + (\Sigma x_2x_3)c_{23} + (\Sigma x_2x_4)c_{24} &= 1\\ (\Sigma x_2x_3)c_{22} + (\Sigma x_3^2)c_{23} + (\Sigma x_3x_4)c_{24} &= 0\\ (\Sigma x_2x_4)c_{22} + (\Sigma x_3x_4)c_{23} + (\Sigma x_4^2)c_{24} &= 0\end{aligned}\right\} \quad (78)$$

$$\left.\begin{aligned}(\Sigma x_2^2)c_{32} + (\Sigma x_2x_3)c_{33} + (\Sigma x_2x_4)c_{34} &= 0\\ (\Sigma x_2x_3)c_{32} + (\Sigma x_3^2)c_{33} + (\Sigma x_3x_4)c_{34} &= 1\\ (\Sigma x_2x_4)c_{32} + (\Sigma x_3x_4)c_{33} + (\Sigma x_4^2)c_{34} &= 0\end{aligned}\right\} \quad (79)$$

$$\left.\begin{aligned}(\Sigma x_2^2)c_{42} + (\Sigma x_2x_3)c_{43} + (\Sigma x_2x_4)c_{44} &= 0\\ (\Sigma x_2x_3)c_{42} + (\Sigma x_3^2)c_{43} + (\Sigma x_3x_4)c_{44} &= 0\\ (\Sigma x_2x_4)c_{42} + (\Sigma x_3x_4)c_{43} + (\Sigma x_4^2)c_{44} &= 1\end{aligned}\right\} \quad (80)$$

If these equations are solved, it will be found that $c_{23} = c_{32}$, $c_{24} = c_{42}$, and $c_{34} = {}_{43}$. Also, it is evident that the coefficients of the equations to the left hand of the equality signs are identical in all three sets of equations and are also identical with those of the normal equations (38), used in determining the net regression coefficients. These two facts make it possible to compute the values of all the c's at the same time that the b's are computed, with only a relatively slight additional amount of work. This process is given in detail in the appendix, "Methods of Computation," pages 469 to 474.

The c's for a large number of independent variables are obtained by an expansion of equations (78) to (80), setting up as many sets of simultaneous solutions as there are independent variables and placing the 1 on the right-hand side of the equations opposite the variable whose (Σx_n^2) occurs with the c_{nn}'s, just as for the second set of equations (79) above; 1 occurs to the right of the equation where $(\Sigma x_3^2)c_{33}$ occurs as one of the items on the left of the equality sign.

The standard error of the individual forecast, according to equation (77), will differ for each combination of values of the various independent variables shown in the new observation. If the values of these several independent variables all fall at about their mean values, $\sigma_{x-x'}$ will be only slightly larger than $\bar{S}_{1.234}$. If they fall far from it, or even if one independent variable falls far from it, where the standard error of the net regression for that variable is very large, the standard error of the estimate will be correspondingly large.

For n variables, the general formula for the square of the standard error of the individual estimate is given symbolically by

$$\sigma^2_{x'_{1.23\ldots n}-x_1} = \bar{S}^2_{1.23\ldots n}\left[1 + \frac{1}{n} + (c_2x_2 + c_3x_3 + \ldots + c_nx_n)^2\right] \quad (81)$$

In expanding equation (81) for any number of variables, it must be interpreted by the special condition that $c_2c_2 = c_{22}$, $c_2c_n = c_{2n}$, etc.

The standard errors of individual estimates made from multiple regression equations, according to equations (77) or (81), can be interpreted in exactly the same way as given above for standard errors of individual estimates from simple regression equations, from equation (75).

Curvilinear correlation. Where a simple or multiple curvilinear relation is determined by fitting mathematical regression equations, the standard error of individual estimates can be computed by an extension of equation (77). Thus if a cubic parabola has been fitted using

$$Y = a + b_2X + b_3X^2 + b_4X^3$$

we can compute this equation most readily by writing it in the form

$$Y = a + b_2X + b_3U + b_4V$$

where $$U = X^2 \quad \text{and} \quad V = X^3$$

The standard error of an individual estimate is then given by the equation

$$\sigma_{y'_{xuv}-y} = \overline{S}^2_{y.xuv}\left[1 + \frac{1}{n} + c_{xx}x^2 + c_{uu}u^2 + c_{vv}v^2 \right.$$
$$\left. + 2c_{xu}xu + 2c_{xv}xv + 2c_{uv}uv\right]$$

Similar expansions are available for mathematical regression equations for two or more variables.[2]

Where the regression curve has been determined graphically, the standard error for simple correlation is as follows:

$$\sigma^2_{y'_{y \cdot f(x)}-y} = \overline{S}^2_{y \cdot f(x)}\left(1 + \frac{1}{n}\right) + [\text{standard error of } f(X) - f(X_M)]^2 \quad (82)$$

In equation (82), the last term of the equation is that determined by equation (74.1), for the particular value of X for which Y is to be estimated.

In the case of multiple curvilinear correlation, with the regressions graphically determined, no precise equation has yet been developed to give the standard error of individual estimates. A roughly approximate value may, however, be calculated as follows:

$$\sigma^2_{x'_{1 \cdot f(234)}-x} = \overline{S}^2_{1.f(234)}\left(1 + \frac{1}{n}\right) + \sigma^2_{f_2(X_2)} + \sigma^2_{f_3(X_3)} + \sigma^2_{f_4(X_4)} \quad (83)$$

where the second, third, and fourth terms are the standard errors of the curvilinear regressions for the particular values of X_2, X_3, and X_4, which are represented in the estimate of X_1, calculated according to equations (74.2) or (74.21).

[2] Henry Schultz, The standard error of a forecast from a curve, *Journal of the American Statistical Association*, pp. 139–185, June, 1930.

Equation (83) gives only a rough approximation to the true standard error because it excludes the terms which provide for the cross-products between the different independent variables. Where the intercorrelations ($\overline{P}_{2.34}$, etc.) between the independent variables are low—say, 0.50 or lower—this will probably not affect the calculated error very much. Where the intercorrelations are quite high, this estimated value may overestimate or underestimate the true error by a considerable margin.

The Applicability of a Regression Equation to an Extrapolation beyond the Observed Range

We have already seen examples, in Chapters 14 and 16, of how estimates might sometimes need to be made for new observations which lie beyond the range included in the original sample. We have also seen the possibility of exceptionally large errors of estimate when the formulas or curves are extrapolated in this way beyond the observed range. A rough rule-of-thumb has been given that estimates beyond the observed range should never be made, or, if they must be made, should be regarded as exceptionally hazardous. This present section will explore further the meaning of the statement "beyond the range of observation."

Where only two variables are concerned, there is no question as to the range covered in the original observations. Thus if we consider the data plotted in Figure 23, on page 154, it is apparent at once that the independent variable, *X*, covers the range from 1.2 to 3.5. Any new values of *X* smaller or larger than those values would be beyond the observed range.

Where two or more independent variables are concerned, the situation is more complex. Thus the data of the example plotted on page 170, in Figures 25 and 26, show that the acres range from 60 to 240, and the cows range from 0 to 18. Suppose a new observation were drawn from the same universe, with 225 acres and 17 cows. Would that observation be within the original range? At first it might seem that it would, since the number of acres falls within the original acreage range, and the number of cows within the original range for cows

Multiple correlation, however, is concerned not merely with the relation of the dependent variable to each independent variable separately, but with the composite relation to all the independent variables together. Is the *combination* of 17 cows and 225 acres, whose effect was represented, either exactly or approximately within the original observations? This combination involves the joint values for X_2 and X_3, which were represented in the original observations. These are shown plotted on Fig. 27, page 170. It is evident from this figure that the new combination lies well outside the observed joint distribution of cows and acres.

The original sample had some farms of between 200 and 250 acres, but none of them had more than 6 cows. It also had some farms of 15 or more cows, but none of them had more than 120 acres. The single original case that came anywhere near the new observation was a farm with 14 cows and 180 acres. Even this one case is quite different from the new observation with 17 cows and 225 acres. Since the new observation lies well outside the *joint distribution* or combination of values represented in the original sample, any estimate made for it from a regression equation based on that sample is subject to an extra degree of hazard, beyond that given by the error formulas discussed in the preceding portion of this chapter. Those formulas give accurate values of the probable error of individual estimates only within the range represented by the original sample. Extrapolation of the regression equation or curves beyond that range, or combination of values, represents an extension into unknown fields, where sudden changes in the nature of the relations might conceivably occur. *A priori* knowledge of the relations, based on technical facts and theories, or on other evidence, may justify extrapolations of the curves. Any assumption that the errors of such extrapolations can be calculated from the error formula derived from the sample depends on a continuation of the observed relations into the unsampled range of values. Such an assumption can be justified only by other information, independent of the observed values and the constants calculated from them. Estimates of error for such extrapolations are only as reliable as the assumptions on which the extrapolations are based.

Where there are three or more independent variables, it is still more difficult to determine whether a given new combination of values lies outside the joint distribution of the three or more variables in the original sample. In many cases this can be determined by careful checking of the new observation against such dot charts as those in Figure 42, on page 270 of Chapter 16. Thus, suppose a new observation were drawn with 2 cows, 100 acres, and 4 men. Would this be within the range of the original observations?

Careful inspection of the charts on page 270, and of the data on page 199, reveals that, although the combination of 2 cows and 100 acres is well within the observed joint distribution for those two variables, no such combination occurred with 4 men, or even with 3 men. The nearest values are one observation (No. 7) of 3 men with 6 cows and 170 acres and one other (No. 12) of 3 men with 15 cows and 120 acres. The new observation, of 4 men with 2 cows and 100 acres, would apparently involve much more human labor, to care for that many cows and acres, than was represented in the original observations, and therefore lies far outside the joint distribution represented in the sample for the three values. It is quite possible that that much labor would represent a wasteful use, so that the additional men would be more likely to reduce the farm income rather than increase it. An estimate of income for this new farm, based on the relations shown in the sample for quite different farms, might therefore be very sadly in error.

The rough process of comparing the new observation with the values of the independent variable for the original observations, as illustrated above, may serve reasonably well for determining whether the new observation is or is not represented in the original sample. Methods are available for computing the exact probability of the new observation being drawn from the distribution represented in the original observations.[3] Carrying through such calculations ordinarily would seem to involve an amount of labor out of proportion to the value of the information obtained. For very exact work, or for estimates of very great importance, however, it might be worth working them out. This would be true especially where the new observation happened to fall at about the edge of the distribution zone of the previous observations, so that it was uncertain whether or not it would be safe to estimate the dependent variable from the relations previously observed.

The Use of Error Formulas with Time Series

Many of the problems that are important in economics and other social sciences involve measurements in time. Even in crop forecasting and in some other problems involving biological reactions, time series must be analyzed.

All the error formulas presented in this chapter and in the preceding chapter, as well as those in Chapter 2, are based upon the theory

[3] The article by Waugh and Been, cited in full at the end of this chapter, gives formulas for calculating this probability. This article also considers the standard error of the individual estimate and gives error formulas similar to those presented earlier in this chapter.

of simple sampling. That theory assumes that each observation in a sample is selected purely at random from all the items in the original universe. It also assumes that successive samples are selected in such a way that value found in one sample have no relation or connection with the values found in the next sample.[4] If the successive months or years in a time series are regarded as successive observations, the first assumption obviously may not hold true. Each successive item of a linear trend line is perfectly correlated with each preceding item. Each price of a given commodity on succeeding days or months may show some relationship to prices in the preceding period. If the correlation between each item of a series and each item of the same series following it in time is calculated by the usual methods, the resulting correlation coefficient is termed the *coefficient of serial correlation.* In time series, almost every variable will show serial correlations that differ significantly from zero. That fact has been urged as a reason why the theory of errors cannot be used at all with such data. It also has been urged as a reason for not even using ordinary correlation techniques with time series, unless special devices, such as successive first differences, are used to eliminate the serial correlations.[5]

Time series also differ from the situation assumed in simple sampling in their lack of constancy of the universe. The formulas of simple sampling assume that there is a large or infinite universe of similar events, from which the sample is drawn at random. Such a universe might be, for example, the number of dots turned up at each throw by throwing a pair of dice a large number of times. They also assume that new observations or new samples will be obtained by drawing in exactly the same way from exactly the same universe, as by making additional throws with the same set of dice under exactly the same conditions. Precise probability forecasts can be made from the original sample concerning the proportions of new samples or observations that will show certain characteristics under these ideal and highly simplified conditions.

When any phenomenon is sampled at successive intervals of time the "universe" being studied can never be precisely the same. Even

[4] Note the way this assumption comes into the derivation of the equation for the standard error of the arithmetic average, in Note 1, Appendix 2. See also Richard von Mises, *Probability, Statistics and Truth,* The Macmillan Co., New York, 1939.

[5] See Alexander Sturges, Price analysis as a guide in marketing control: the use of correlation in price analysis, *Journal of Farm Economics,* Vol. XIX, pp. 699–706, August, 1937.

successive astronomical observations differ, even if in imperceptible degrees, because of the loss of matter radiated from the various stars. Surveying measurements in successive years may differ because of slight geological shifting of the earth's surface, or because of erosion or other changes in the soil surface. Normal crop or livestock yields, as seen earlier, may change because of improvements in the biological make-up of the seed or in the strains of stock so that what would be normal yields for certain weather or feed at one time become subnormal yields at another. The "population" of corn plants or of cows is not static—it changes constantly as one generation passes away and new ones come upon the scene. Human populations, too, change constantly by birth, growth, and death, so that what is the normal average height or weight in one year is different in another. The habits and ideas of living men may change much faster than the people themselves change. Ordinarily, perhaps, those habits and ideas change slowly—most people will react to the idea of "socialism," for example, one day in much the same way that they reacted to the same idea a week or a month earlier. But sometimes, under the force of social pressures, world-sweeping events, or economic or other catastrophes, ideas change swiftly and dramatically. Many Iowa farmers who, in 1928, were born-and-bred Republicans of the most conservative type, were threatening to hang judges to prevent foreclosure sales four years later, in 1932.

To the extent that changes in the universe follow a steady rate of progression, they can sometimes be allowed for by trend factors (as shown earlier) or by progressive shifts in the regressions themselves. So long as the composition of the universe—in correlation analyses, the character of the relations or reactions which are under study in a given set of circumstances—remains substantially constant from period to period, with at most only well-defined patterns of change, the changing character of the universe can thus be allowed for, at least in part. In such cases, forecasts of future changes depend upon a continuation of the same rate or degree of change. It is never possible to be certain, however, that a new event may not make a sudden change or break in the trend—as the declaration of war in September, 1939, produced a sudden change in prices and markets.

But what of the independence of successive observations? Does the fact of serial correlations mean that correlation cannot be applied successfully, and that regression relations found in past cases will never work out equally well in practice?

We have seen already in several cases (notably in Chapters 14 and 16) that forecasts worked out by extrapolating an earlier formula to subsequent years have given results which agreed remarkably well with the standard error of estimate. Had we calculated the wider standard error of individual forecasts (equation [77] or [82]) for these cases, the agreement of the actual errors with the expected range of error would have been even better. This agreement is contrary to what we would have expected, on the basis of the theories set forth above. Is it merely a lucky accident, or does it indicate that the sampling equations have a wider applicability than their basic assumptions would lead us to expect?

This problem is one of the greatest unsolved questions in the whole field of modern statistical methods. It is one where the widest possible range of judgments may be found among the experts who should be able to agree upon the answer. Without presuming to give a final answer to the question, we may advance certain considerations to explain why correlation results with time series may be more reliable than some critics have believed they could be:

In the discussion to this point, we have assumed that after we knew the values for, say, 1940, the values for 1941 constituted the next observation. Also, we have tacitly assumed that since 1941 will have only a single set of values—say of rainfall, temperature, and corn yields—we are not selecting a new observation at random, but are drawing a unique and predetermined set of values.

Let us examine this a little closer. So far as the trend value is concerned, that is so. The trend reading for 1941 (as for variable X_4 in the problem of Chapter 14) is bound to be one unit larger than for 1940, and the estimated contribution of trend to the yield is therefore expected to be exactly in line with that of preceding years. As explained above, this is inherent in our assumption of a gradual yet continuous shift in the composition of the universe. But what of the values of the other variables? Are they predetermined?

Rainfall in a given season, so far as meteorologists have been able to explain it, is the final result of a large number of accidental circumstances, mainly unpredictable very far in advance. Efforts to forecast the rainfall from that of the preceding season or seasons have yielded unsatisfactory results. There seems to be something of an irregular periodicity in weather over considerable periods of time, but the unpredictable year-to-year fluctuations around that irregular trend are of much greater magnitude than the trend itself. Much the same is true for temperature. So far as the rainfall and temperature are concerned, then, the values encountered in a given year may be regarded as pretty completely random "drawings" out of nature's grab-bag of all the possible weather combinations that might occur that year. The yield of corn, in turn, represents a similar "drawing" out of all the possible yields that might accompany that weather combination, for a year when the gradually shifting elements in the universe were of the magnitude as measured (more or less accurately on its extrapolation into the new territory) by the trend. Explained in this way, all the observations may be regarded as reasonably "random" sampling from the observations that otherwise might have been secured for the given year. And if the drawings for that year, as well as the drawings for each of the other years, have no particular relation with what other observations might have been drawn each year if the forces of nature had nodded another way instead, we may feel reassured that the successive observations were really random—always excepting the factor of progressive change measured, more or less accurately, by the trend element. (It must be remembered, also, that this is not a trend in yield or a trend in rainfall but rather a trend for the yield secured under constant conditions of rainfall and temperature. The trend is itself a net regression, measured while eliminating the variation in yield associated with changes in the other variables.) For the meteorological problem, then, we may feel that we can calculate standard errors with reasonable propriety, even though all the data are time series. Also, we see that we can construct a reasonably satisfactory explanation for why the data behave *as if* the theory of sampling did apply, in the sense of the forecast showing about the expected range of error.

But what of time series where the data are economic, and not meteorological? How about the steel-cost problem of Chapter 16? In that problem, steel costs, wage rates, and percentage of capacity operated are all parts of a progressing and evolving economic system. It was obvious from the data that wage rates changed progressively and in most cases slowly, with relatively little change from one year to another. Also, it was evident (as revealed eventually in the trend factor) that costs changed gradually and progressively with relation to such factors as price level which also changed with time and which were not otherwise explicitly recognized in the analysis.

Once the trend allowance has been made to take care of the changing nature of the universe (including, in this case, rough and imperfect allowance both for technology changes and price-level changes), the ordering of the data in time has no influence on the final correlation result. If the costs adjusted for the (net) trend were taken as the dependent factor, the observations could be jumbled in any sequence selected and the net regressions on wages and percentage capacity would still be the same. Although wages cannot be regarded as random events for any year, the cost rate can be regarded as a random selection from all the possible cost rates that might accompany that particular wage rate and operation rate. It is precisely by revealing the distribution of cost rates for given values of the other factors that the net regressions are determined. So it seems that from this point of view we can say that the cost (adjusted for net trend) which accompanies a given combination of values of each of the other factors is a random sample from all the costs that might accompany that combination—and that the error formulas may therefore be used.

There are two limitations, however, on the conclusions concerning the possible independence of events, even in economic time series. (1) In the steel-cost problem, there was some indication of a lag in effect. For example, an increase in wage rate one year (because of accounting difficulties or even because of length of process in the production of finished products) might not show up fully in per-unit costs until the next year, and vice versa. To the extent that values of the dependent factor one year reflect not current year values but previous year values of the independent factors, that might introduce a systematic error in the regressions. This error would be particularly serious if the serial correlation was high for the independent factor involved, such as the wage rate. Only if such a lagging effect was specifically allowed for and measured by introducing the wage rate of the preceding year as an additional independent variable affecting the cost of the given year, could such erroneous results in the regressions be definitely eliminated.

(2) In the previous examples we have been using illustrations—crop years or production periods—where there is a natural break between the successive intervals. Suppose we were studying corn prices, though, and took weekly prices, weekly national corn supply available, and weekly values of other factors. Over four years we should have 208 separate sets of values. Should we be correct in saying we had 208 independent observations of the effect of these several variables on corn prices, in the same way that we could say that observations of corn yields for 25 years gave us 25 independent observations of the factors influencing yield? Obviously, we should not be correct. Instead of having observations of 208 different phenomena, we have a number of repeated observations of what are essentially the same phenomena. To calculate sampling errors at all, or even to judge what *n*

to use in applying the various corrections to our constants, we must take our successive observations of time-series phenomena at sufficiently long intervals so that we sample essentially different phenomena at each successive observation. With crop years or other natural breaks in the process, it is easy to pick such appropriate breaks. Where the process is a continuous one, such as the production of steel, it is more difficult to select appropriate intervals for making the "cuts" to provide independent cross-sections of the continually changing phenomena. If the variables represent the total of output or production over given periods, these successive "cuts" must be uniformly spaced in time. If the variables are rates in time, however, that would not be essential, and the observations might be spaced at varying intervals so as to catch particular values of independent variables. Thus in the steel-cost problem of Chapter 16, we might have selected the observations from the months in which operation rate (percentage of capacity) made new highs or lows, and used those as the basis for our analysis. (For the reasons set forth in Chapter 20, such selection could be applied only to independent factors. If applied to the dependent one, it would seriously bias the results.) The guiding principle must be to select the observations in such a way as to make each successive observation a new observation of a new set of the variable phenomena, except for such continuously progressive elements as can be appropriately eliminated by the simultaneous fitting of a trend or trends.[6]

The preceding discussion suggests some ways in which economic time series can be examined to see if conditions are present which would prevent the theory of sampling from applying, or can be selected so as to make the use of the theory reasonable. If they are found or can be made reasonably free from such conditions, forecasts based on such analyses might be expected to follow reasonably well the error limits given by the formulas based on sampling theory. That may explain why forecasts from economic time series, such as those shown for the steel-cost problem of Chapter 16, may in fact agree quite well with what would be expected if the error formulas did apply.

Where there is clear indication of lagging effects from period to period which cannot be specifically allowed for, or where the serial correlations in the data are so high as to make the several observations

[6] A rough method for judging the length of intervals necessary to obtain such independent values is given by L. R. Hafstad, On the Bartels technique for time-series analysis and its relation to the analysis of variance, *Journal of the American Statistical Association*, Vol. 35, July, 1940. Pages 347 to 353 are especially germane to the discussion here.

not really independent observations at all, then the sampling formulas simply do not apply, because the assumptions on which they are based are not fulfilled in the given problem. In such cases the error formulas may still be calculated, in the hope that they will indicate the minimum possible reliability of the results instead of the maximum possible unreliability. But whether they will do even that correctly is not yet known.

In closing this discussion of the time-series problem, one word may be added on practical procedure. That is on the possibility of testing the actual forecasting efficiency of an analysis by saving the last two or three observations as test values on which to try out the adequacy of the regression equation derived from the earlier observations. This technique, illustrated in several of the examples analyzed in previous chapters, has been found a useful precaution in practical research work. But sometimes the test values will indicate a sudden change in the level of the trend, or will suggest a change in one of the other regressions. In such cases it will usually be better to recalculate the data, using all the information down to the latest year, rather than to base the real forecast for the next year to come on an extrapolation already found to be in doubt. For, as illustrated in several of the problems and as demonstrated in the first section of this chapter, the longer the extent of extrapolation beyond the base data, the greater the possibility of error. And that applies just as definitely when a net trend regression is being extrapolated as any other regression curve. So save a few final values for test cases—but do not hesitate to add them to the sample analyzed, if that is found necessary to account satisfactorily for the most recent relations.

Practical Procedures in Judging the Reliability of Forecasts

Up to this point this chapter has presented mathematical procedures for judging the confidence that can be placed in an individual forecast, solely in the light of the information given by the individual sample from which the estimating (regression) equation was derived. In actual practice, the statistician has usually only this single set of sample data to judge from. Usually (and almost universally in time series) he cannot draw a new sample and contrast the results of the second sample with the first. He may, however, have other prior information to help guide him in making and interpreting his forecast. Or he may be able to draw a series of samples, each one throwing light on a different aspect of the same problem. Each one of those samples may give results subject to a wide margin of random error. Yet if the results of the several different approaches are all consistent with one another, the whole set together will provide a more dependable basis for a forecast than is indicated by the calculated standard errors for any one taken separately. Other relevant information may be of a quite non-quantitative nature, yet it may serve to help guide the analyst in making his forecast.

Any forecast is hazardous, for the future can never be perfectly known. Yet life always consists in making plans for the future. Every business man, every farmer, every consumer is constantly making judgments as to the future, and making commitments or taking action based upon those judgments. The success or failure of the actions in reaching the ends sought often depends in large measure upon the accuracy of those estimates. In every-day life such estimates are usually based on hunches, waves of opinion, the most recent happenings, rule-of-thumb analyses, or even blind guess-work. If the statistician is to serve society, it must be on this action front, where his analysis of past relations will help provide a surer guide into the unseen over the horizon. Many statisticians hesitate to make forecasts, for they know how little statistical dependability they can place in them. They fear to risk their reputations upon forecasting an uncertain event. They need not be so hesitant. To be useful, all that is required is for their forecasts to be more reliable, on the average, than the forecasts on which such judgments have been based in the past. Many business forecasting services have been accurate only to 55 per cent, yet have kept in business on the gain over the 50-50 odds of the completely uninformed guess. In events characterized by waves of emotions or by common response of many individuals to the same stimuli, as in the business cycle or the hog cycle, the accuracy of the uninformed guesses—and actions—of producers may have averaged only 30 to 40 per cent right. Even a forecast which is very fallible when judged by its mathematical or statistical significance may yet yield a greatly improved guide to human actions. If the statistician will base his forecast on all the information at his command, quantitative and non-quantitative, and will guard his forecast by some statement as to its range of dependability, he can both aid judgments and protect his reputation. In the end, he must be willing for that reputation to rest upon the average accuracy of a long series of estimates rather than upon the lucky calling of any one individual event. And the more the technical operations on the statistical side can be reinforced by the knowledge, theories, experience, and judgments of the researcher as practical agronomist, sociologist, economist, meteorologist, or other technician, the more valuable the statistical operations will become as a basis for an informed and useful projection from the events of the past into the still-malleable future.[7]

Summary. In this chapter we have discussed the problem of the accuracy of individual estimates of the dependent variable for new observations drawn from the same universe as the original sample and have presented methods of estimating the probable range of error for such estimates. We have considered the question of whether the new observations of independent factors represent portions of the distribution of the same factors present in the sample or whether they include new and therefore untested values or combinations of the same variables, with resulting unpredictable effects upon the estimates of the dependent variable. Finally, we have discussed the question of the applicability of error theory to time series, shown how in many cases it may still be applied, and indicated some rough tests to judge whether or not the time series is or is not of a character that will make error calculations completely inapplicable. Finally, we have given some hints for the handling of time-series correlations in such a way as to minimize the errors when extrapolating the regressions for purposes of practical forecasting, and have made some suggestions for combining tests of significance with other prior information as a basis for making and judging forecasts.

REFERENCES

Waugh, Frederick V., and Been, Richard O., Some observations about the validity of multiple regressions, *Statistical Journal of the College of the City of New York*, Vol. 1, No. 1, pp. 6–14, January, 1939.

Schultz, Henry, The standard error of a forecast from a curve, *Journal of the American Statistical Association*, pp. 139–185, June, 1930.

[7] Compare this discussion with the concluding section of Chapter 2, pages 30 to 32.

CHAPTER 20

INFLUENCE OF SELECTION OF SAMPLE AND ACCURACY OF OBSERVATIONS ON CORRELATION RESULTS

Selection of Sample

Methods of determining linear and curvilinear regressions, together with appropriate measures of their significance and accuracy, have been set forth in previous chapters. These methods do not yield results representative of the universe from which the sample observations have been drawn, however, if that sample is not truly representative of the particular relation being determined. There are various ways in which the sample may fail to represent the universe, and the resulting extent to which the correlation constants will be biased will vary both with the character of the unrepresentativeness and with the individual coefficients. Each type of abnormality must therefore be treated separately.

The samples may be selected from the universe in such a way as to exclude all the observations falling beyond a certain value of a given variable, thus ruling out values either at one or at both extremes, or perhaps ruling out middle values and selecting only extreme ones. This may be done for either the dependent variable or the independent variable or variables, or for both together. Such a selection of observations produces certain specific effects upon the correlation constants. Under some conditions it may be very desirable to select the observations in this way, if only the resulting aberrations in the correlation constants are recognized and allowed for.

A second and somewhat more difficult type of problem to deal with arises when there are errors of measurement in obtaining the values of one or more of the variables—such errors as might arise, for example, in estimating the total production of corn in the United States within a given year, or in working out, from a farmer's memory, what was the income on his farm the previous year. Here again the effect on the correlation constants will depend upon whether the errors are random or biased and upon which variable or variables are affected by the errors. A separate discussion must therefore be given each case.

The clearest way of indicating the effect of these various departures from truly representative sampling may be by first stating the general principles involved, and then illustrating the way those principles work out by concrete illustrations. Except where specially stated otherwise, the discussion will apply solely to linear relations. The effects for curvilinear relations are in general analogous to those to be discussed.

Selection of sample with respect to values of independent variable. If the sample is selected with respect to values of the independent variable, that will not tend to affect the slope of the regression line but will affect the value of the coefficient of correlation. If the selection is such that extreme values are rejected but intermediate ones are left in, the correlation will be lowered below that prevailing in the universe; if intermediate values are rejected and only extreme ones are used, the correlation will be raised above that prevailing in the universe. If the values of both variables are normally distributed, the standard error of estimate will tend to remain the same, regardless of the selection.

These principles may be illustrated by the set of hypothetical data shown in Table 76.

TABLE 76

Correlation Table, Showing Hypothetical Frequencies at Specified Values

Values of Y	Values of X				
	0	1	2	3	4
0	....	....	1	1	1
1	....	1	2	2	1
2	1	2	4	2	1
3	1	2	2	1	
4	1	1	1		

For the data shown, $r = -0.47$, $\sigma_x = \sigma_y = 1.134$, and $b_{yx} = -0.50$. If now the values had been selected with references to X, so as to exclude values below 1 or above 3, the number of observations would have been reduced from 28 to 22. For this restricted set of observations, $\bar{r} = -0.26$, $\sigma_x = 0.739$, $\sigma_y = 1.09$, but $b_{yx} = -0.50$. Computing the standard error of estimate, we arrive at $\overline{S}_{y.x} = 1.02$ for the first case and 1.07 for the second. It is quite apparent that the correlation has been lowered by the restriction in the selection of the values of X; but the regression of y on x has not been changed at all, and the standard deviation of the residuals has been only slightly changed.

If now the selection is such that only extreme values of X are taken, say below 1 and above 3, the number of observations is reduced to 6.

Computing the results for those values, we have $\sigma_x = 2.00$, $\sigma_y = 1.29$, $\bar{r} = -0.71$, but $b_{yx} = -0.50$! Also $\bar{S}_{y.x} = 1.00$.

Bringing the three sets of results together for comparison, we have the following tabulation.

With X used as *independent* variable:

	σ_x	σ_y	$\bar{r}_{xy}$	$\bar{S}_{yx}$	b_{yx}
All cases.	1.13	1.13	−0.47	1.02	−0.50
Extreme values of X excluded.	0.74	1.09	−0.26	1.07	−0.50
Only extreme values of X used. . . .	2.00	1.29	−0.71	1.00	−0.50

These three examples thus illustrate the principles stated before: that selection with respect to the independent factor does not tend to change the regression or the standard deviation of the residuals but does affect the correlation, lowering the correlation if it has lowered the dispersion of the independent factor and raising the correlation if it has increased the dispersion of that factor.

Selection of sample with respect to values of dependent factor. Selection with respect to values of the dependent factor is more serious, in that it affects all the constants. According as the effect is to raise or to lower the standard deviation of the dependent factor, such selection tends to raise or lower both the regression coefficient and the coefficient of correlation from the value for the universe and likewise to raise or lower, respectively, the standard error of estimate.

These principles may be illustrated from the three examples just used, by regarding X as the dependent factor and Y as the independent factor and noting the influence of the selection with regard to the dependent factor, X, upon the regression of X on Y, b_{xy}. For the first case, with all values left in, $b_{xy} = -0.50$ and $\bar{S}_{x.y} = 1.02$. For the second case, however, with extreme values of X left out, b_{xy} drops to -0.23 and $\bar{S}_{x.y}$ becomes 0.73. For the third case, with only extreme values of X included, b_{xy} increases to -1.20 and $\bar{S}_{x.y}$ becomes 1.55. Bringing these three sets together yields the following comparison.

With X used as *dependent* variable:

	σ_x	σ_y	r_{xy}	$\bar{S}_{xy}$	b_{xy}
All cases.	1.13	1.13	−0.47	1.02	−0.50
Extreme values of X excluded.	0.74	1.09	−0.26	0.73	−0.23
Only extreme values of X included. .	2.00	1.29	−0.71	1.55	−1.20

These results indicate the extent to which selection with regard to the dependent factor may completely destroy the significance of all the results.

Selection of samples with reference to values of both variables. Selection of cases with reference to values of both independent and dependent variables has an even greater effect upon the conclusions than the two cases discussed because selection of extreme values tends to exaggerate the correlation and regression, and of central values to lower both, to even greater extent than where the selection is with respect to the dependent factor alone.

If, in the data of Table 76, only those cases are selected in which values of X below 2 are associated with values of Y above 2, and in which values of X above 2 are associated with values of Y below 2, the observations are reduced to ten cases, as follows:

Values of X	Values of Y	Number of cases	Values of X	Values of Y	Number of cases
0	3	1	3	0	1
0	4	1	3	1	2
1	3	2	4	0	1
1	4	1	4	1	1

For these values, $\sigma_x = 1.48$, $\sigma_y = 1.48$, $\bar{r}_{xy} = -0.90$, $b_{yx} = -0.91$, and $\bar{S}_{yx} = 0.68$.

It is evident that such selection raises both the correlation and the regression above the true value for the universe. This is to be expected, for this selection is equivalent to picking out the pairs of values which *do* show correlation with each other. Restricting the selection to paired values of above 1 for both variables, and below 3 for both variables, likewise would be picking out cases so as to eliminate all correlation. Such selection obviously destroys the value of the results.

Conclusions with reference to selection of data. If an investigator is interested only in the regression line and not in the degree of correlation, and if the regression is truly linear, selection of data with reference to the independent factor (or factors) will not tend to change the slope of the regression line (or lines). Under those conditions selection of extreme cases of the independent factor may yield a reliable indication of the regression with much fewer observations than if the cases were selected at random. This principle is frequently applied in experimental or laboratory work, but is equally applicable in other types of investigations.

If the regressions are curvilinear, however, special selection of either extreme or central items of the independent variables forestalls the determination of the nature of the function, since curvilinear regressions can be determined only for the ranges of the independent factor within which observations have been secured. For such regressions, therefore, the nature of the function may be more accurately determinded if the independent items are selected so as to be spread fairly uniformly through the whole range of values, thus affording a sufficient number of observations for accurate determination of the nature of the relation throughout the whole range. Selection purely at random frequently provides more observations than are needed for certain portions of the curve, and provides so thin a scattering of observations at other portions as to make its true position and shape quite indeterminate, as has been illustrated previously. Even if curvilinearity is only suspected, such a uniform distribution of values for the independent variable provides an improved basis for determining whether or not the regression is truly linear, as compared with an equal number of observations selected at random. At the same time, where the dependent factor is normally distributed, selection with reference to the independent factor does not tend to change the standard error of estimate.

If the primary interest, however, is not in the nature of the relations and in determining how closely values of the dependent factor may be estimated (regressions and standard error), but instead is in determining what proportion of the original variation in the dependent factor can be accounted for on the basis of the relations determined (correlation and determination), then anything other than random selection with reference to any factor will give estimates of the closeness of the correlation which either over- or underestimate the true correlation in the universe from which the sample is drawn. For most accurate results in such problems, the distribution of the dependent factor in the sample should be an accurate representation of the distribution in the universe from which the observations were drawn, and the only selection which would be justified would be aimed at securing such a sample.

Since the correlation coefficient or index, and the parallel measures of determination, are of significance only with respect to the standard deviation of the observed values of the dependent factor, it follows that when the dependent factor has such an abnormal distribution that its standard deviation is of little value as a descriptive statistic, the measures of correlation also tend to be of little value. For any series which actually yielded such an extreme distribution as the dichotomous values used in the third case of those just illustrated, measures of correlation would have little significance except their formal mathematical definition. Yet the regressions and standard errors of estimate would tend to retain all their usual value and significance, so long as no selection had been made with reference to values of the dependent variable. In such a case, attempting to select the values of the dependent factor so as to make the series more nearly normal might seriously bias the regression results.

Accuracy of Observations

The data with which the statistician has to deal are frequently subject to errors of observation. If corn yields are being studied in relation to fertilizer applications, for example, farmers may be able to estimate the yield per acre on a given tract only to within 5 or 10 bushels of the true yield. If livestock prices are being studied, the market reporter may not be able to get his daily average nearer than within 10 or 25 cents per 100 pounds of the true average of all the sales for the day. Or if educational ratings are being studied, the instructor may not be able to grade the test papers nearer than to within 5 or 10 per cent of the grade each really deserves. All these illustrations are akin to the difficulties of the surveyor, who finds he cannot measure his angles more accurately than within a certain number of seconds; or of the astronomer, who finds his repeated observations disagree from each other by fractions of a second. But the errors of measurement are ordinarily tremendously greater in biological, economic, or social investigations than in physical observations; and for that reason statisticians must be particularly careful to use their data in such a way as to minimize the influence, upon their conclusions, of the errors which may be present.

Errors of observation may be such that they are not correlated with the value being observed, and hence tend to fall equally above and below the true values throughout the range of the variable; or else they may be such that they are correlated with the variable, tending usually to make the observed value fall above the true value in the upper part of the range, and below the true value in the lower part; or *vice versa*.

In correlation problems, there are two sets of true values involved, those for the dependent and independent variables; and there may also be two sets of errors, one tending to cause the observed values for the dependent variable to differ from the true values, and the other affecting the independent variable. The extent to which such errors, if present, modify or impair the results of correlation analysis depends both upon the type of the errors and the variables which they affect.

If the errors affect only values of the dependent factor, and if they are not correlated with the true values, their presence tends to lower the correlation and to increase the standard error of estimate, but does not tend to change the slope of the regression line from the true slope for the universe. If, however, uncorrelated errors are in the independent factor, that not only tends to lower the correlation and increase the standard error of estimate, but also tends to decrease the regression below the true value. Both of these cases may be illustrated from the same set of data used before.

Errors in the dependent variable. The data used in Table 76 may be modified by assuming some random error influences Y, making one-third of the values 1 unit higher, one-third 1 unit lower, and leaving one-third unchanged. With these changes, the data appear as follows:

TABLE 77

Correlation Table, Showing Hypothetical Frequencies at Specified Values, with Random Errors in Y

Values of Y	Values of X				
	0	1	2	3	4
−1	....	....	1	....	1
0	....	1	....	2	....
1	1	....	3	....	1
2	....	2	2	3	....
3	1	2	3	1	1
4	....	....	1	....	....
5	1	1	....	....	....

For these data, $\bar{r}_{yx} = -0.33$, $b_{yx} = -0.50$, and $\bar{S}_{yx} = 1.46$. The introduction of the random error into Y has lowered the correlation from that of -0.47 for the original values and increased the standard error of estimate; but it has had no significant effect upon the regression of Y on X, the new value -0.50 being identical with the value of -0.50 for the original data in Table 76.

Error in the independent variable. If, however, X is regarded as the dependent factor and Y as the independent, the regression coefficient for the new values, $b_{xy} = -0.28$, is found to be much reduced from that of -0.50 for the original values. Introducing even random errors into the observations of the independent factor markedly reduces the observed regressions below the true value.

The errors considered to this point have all been random errors. If, instead, the errors are correlated with either of the factors, their presence would obscure the true relationship and bias any correlation constants which might be computed, tending to make them either too high or two low, depending on the inter-relations between the errors and the variables.

Errors in both variables. If random errors are associated with both variables simultaneously, their effects are a blending of those just illustrated, tending to reduce both the closeness of correlation and the regression below the true values. For example, if random errors of the same magnitude are introduced into X as well as Y of Table 76, the values appear as follows:

TABLE 78

Correlation Table, Showing Hypothetical Frequencies at Specified Values, with Random Errors in Both X and Y

Values of Y	Values of X						
	−1	0	1	2	3	4	5
−1	....	....	....	....	2		
0	....	....	1	1	1		
1	1	....	1	1	1	1	
2	....	1	2	2	1	1	
3	....	1	2	2	1	1	1
4	....	....	....	....	1		
5	....	1	1				

With these changes, the correlation is reduced to practically 0, the standard error increased to 1.524, and the regression of Y on X changed to -0.179. The comparison of these constants with those for the original data in Table 76 illustrates the extent to which the presence of random errors in the observed values of the variables may reduce the accuracy and effectiveness of correlation analysis.

Dealing with errors in both variables. The methods of computing the regression line considered to this point are methods which take one variable as given, or independent, and the other variable as based upon it, or dependent. If it is known that all the errors of observation are random and are in one variable, and none are in the other, the effect of those errors may best be eliminated by considering the one with no errors as independent and the other as dependent. As has just been demonstrated, the regression line then obtained will be practically identical with that which would be obtained if no random errors at all were present.

In some cases it may be known that both variables are subject to random error, yet it may be desired to obtain a regression line which most accurately expresses the relation between the two. That can be done by a special method, which fits the line on the condition that the sum of the squares of the departures of each observation *perpendicular* to the fitted line shall be made a minimum (in contrast to the usual condition that the sum of the squares of the *vertical* departures from the fitted line shall be made a minimum, with the dependent variable plotted as the ordinate.) This special method involves an entirely different procedure for fitting the line, and is not given here. It has the disadvantage that it does not give a basis for estimating values for either variable from known values of the other, nor does it give a basis for measuring the closeness of the correlation between the two. It is referred to here merely to call attention to the fact that methods are available for determining the regression when both variables are known to be subject to random errors.[1]

Errors of observation in multiple correlations. The points which have been illustrated here for simple correlation are equally true for multiple correlation, both with respect to the influence of selection of sample and of the effect of errors of observation. The influence of errors of observation in multiple correlation problems may be illustrated by a case based on actual economic data.

Over the 17 years from 1907 through 1923, the monthly price of lambs shows a very high correlation with the price of wool and the price of dressed lamb. When X_2 is used for the price of wool, in cents per pound, X_3 for prices of dressed lamb, in cents per pound, and X_1 for prices of live lambs, in cents per pound, multiple correlation gives, for the 204 observations, $R_{1.23} = 0.991$ and $x_1 = 0.144x_2 + 0.354x_3$.

To test what effect random errors would have had on this correla-

[1] Abraham Wald, Fitting of straight lines if both variables are subject to error, *Annals of Mathematical Statistics*, Vol. XI, No. 3, pp. 284–299, September, 1940.

tion, two dice were thrown 204 times, giving random values from 2 to 12. These values were then added to the successive values of the dependent, and a similar set of 204 values to the successive observations of one independent factor, to see what effect that would have on the results. In the following tabulation the notation "$X + e$" is used to designate the variables to whose values these "random errors" had been added.

EFFECT OF INTRODUCING RANDOM ERRORS ON CORRELATION RESULTS

Independent variables	Dependent variable	Multiple correlation	Regression equation
X_2 and X_3	X_1	0.991	$0.144x_2 + 0.354x_3$
X_2 and X_3	$X_1 + e$	0.821	$0.112x_2 + 0.424x_3$
X_2 and $X_3 + e$	X_1	0.953	$0.163x_2 + 0.277x_3$
X_2 and $X_3 + e$	$X_1 + e$	0.804	$0.152x_2 + 0.306x_3$

These results illustrate the principles just set forth. The introduction of random errors into the dependent variable (X_1) reduces the correlation, but does not greatly change the size of the two regression coefficients. It would appear, especially from the amount of the reduction in net regression on X_2, that the errors in this case may not have been completely randomly distributed and uncorrelated with X_1, X_2, and X_3, even though determined by throws of dice.

But the second modification, where the error is introduced into the independent variable X_3 instead, is much more striking. The correlation is not reduced so much as in the first case, and the regression of X_1 on X_2 is changed only slightly from the original value—and increased as it happens. The net regression of X_1 on $X_3 + e$, however, is only three-fourths as large as was the net regression of X_1 on X_3, in spite of the fact that the error introduced was only enough to raise the standard deviation of X_3 from 6.14 to 6.64.

The final case, with errors introduced into both X_1 and X_3, shows the lowest correlation of any, as would be expected. The net regression of $X_1 + e$ on X_2 is but little different from what the regression of X_1 on X_2 was, whereas the net regression of $X_1 + e$ on $X_3 + e$, though larger than what the regression of X_1 on $X_3 + e$ was, is still definitely lower than the regression of X_1 on X_3. The regression equation in this last case, where $X_1 + e$ is the dependent, is not greatly different from what it was in the preceding case with X_1 as the dependent, in spite of the fact that one of the independent variables—X_3—had a significant random error of observation in its values both times.

These cases illustrate the extent to which random errors may confuse the true relations, if they are allowed to creep into the observations. Just how great an effect upon the results such random errors will have depends upon the magnitude of the errors, the original variations in the variables, and the closeness of the inter-correlation. Although equations can be derived to show how great a reduction in correlation errors of a given magnitude will produce, they are of little practical use in economic work, since it is usually difficult enough to determine whether there are errors of observation or not, much less to determine what magnitude they have.[2] In using reports or estimates of prices or commodity production or supply, we know that the data are nearly always subject to more or less error. The same is true of many other economic data—errors of observation of greater or less magnitude are nearly always present. It may be of some slight reassurance to know that observational errors even as large as those introduced in the example just considered still modify the regression results as little as these have been seen to do.

The practical significance of the principles which are stated here is that, if there is known to be a large but random error in observing some variable, that variable may still be used as the dependent variable in a correlation study without making the regressions or estimating equation very far wrong, if determined with a large number of cases; but, on the other hand, any use of that variable as an independent variable will be certain to yield results which understate the actual relations.

[2] In the problem given, the significant values determining the effect of the errors are:

$$\sigma_1 = 3.96 \qquad \sigma_{1+e} = 4.74$$
$$\sigma_3 = 6.14 \qquad \sigma_{3+e} = 6.64$$

If the errors are in the dependent variable alone, the relations between the true and the apparent correlation are indicated by the equation:

$$R^2_{(1+e).23\ldots n} = \frac{(R^2_{1.23\ldots n})\sigma_1^2}{\sigma_1^2 + \sigma_e^2} = R^2_{1.23\ldots n}\frac{1}{1 + \frac{\sigma_e^2}{\sigma_1^2}}$$

This gives what the new correlation would be if the errors were truly random, so that the new regression equation came out as identical with the old. In the problem given, this gives an expected value for R of 0.827 as compared to the 0.821 actually obtained.

In cases where the errors are biased, they tend to make the results of correlation analysis more or less in error, quite regardless of the variables to which they apply. If the errors tend either to magnify or to minimize the differences which actually exist, they will have a parallel effect on the regression coefficients if they apply to the dependent variables and an inverse effect if they apply to an independent variable. There are so many different types of bias, however, that no more definite statement of the effects can be laid down.

Random errors have the same type of effect in the case of curvilinear correlation that they do in linear correlation, since if they are truly random they will tend to be balanced out along all the portions of the regression curve alike, if in the dependent variable; or tend to confuse the relations along the curve, if in the independent variable; and so reduce the differences observed.

Biased errors, on the contrary, may happen to be concentrated along certain portions of the range, and hence have a much more marked effect at one point than at another. Although this might seriously disturb the significance of the curve, it probably would have an equally disastrous effect on the reliability of the straight line. About the only real difference between linearity and curvilinearity with regard to errors is that random errors in the dependent variable could be "balanced out" in the case of a straight-line regression with a somewhat smaller number of observations than would be necessary to secure valid results for a curvilinear regression.

Where, with random errors in the dependent factor, there are not enough cases available to "balance them out," the effect of the errors is to throw a varying amount of error into the conclusions, the exact amount of the error depending on how closely the errors approach being canceled out. The illustrative case, where with over 200 observations the regressions were still changed somewhat, probably indicates what may be obtained by a combination of slight departures from true "randomness" in the errors with a sample not quite large enough to eliminate entirely all the resulting instability. This may be nearer to what would usually happen in practice than the theoretical complete elimination of the errors in the dependent variable.

Summary. Modification of the observations from the true conditions, either by selection of the sample or by the presence of errors of observation, tends to alter the value of the coefficient of correlation. If the regression line or curve is of primary interest, however, its accuracy of determination may be increased by suitable selection of observations with respect to independent factors. Similarly, random errors of observation may not influence the regressions, if the factor they affect can be treated as the dependent factor and if enough observations are available to balance out the errors. These points hold true for multiple correlation problems as well as for 2-variable problems.

CHAPTER 21

MEASURING THE RELATION BETWEEN ONE VARIABLE AND TWO OR MORE OTHERS OPERATING JOINTLY

In working out the change in one variable with changes in other variables up to this point we have assumed that the relation of the dependent factor to each independent factor did not change, no matter what combination of other independent factors was present. In the case of the yield of corn, for example, as worked out in Chapter 14, we assumed that the effect of a given change in rainfall upon the yield was the same, no matter what was the temperature for the season. The significance of this assumption may be shown by combining the estimate for rainfall with the estimate for temperature, and plotting the combined influence of the two variables. In Table 68 (on page 252) we already have this combined influence worked out, so all we have to do is to plot it. Figure 62 shows the resulting figure. In reading this figure it should be noted that the inches of rainfall are read along the right-hand edge of the bottom of the cube, the degrees of temperature along the left-hand edge, and the yield along the vertical edge. The yield for any combination of temperature and rainfall is then shown by the distance the upper surface of the solid figure is above the point of intersection of the corresponding values in the base plane.[1]

Inspecting Figure 62, we can now see what is meant by saying that the changes in yield are assumed to be the same for each change in rainfall, no matter what the temperature. As shown in the figure, the maximum yield with a temperature of 70° is obtained at about 12 inches of rain—and that is also the rainfall which produces a maximum yield with a temperature of 72°, 74°, or 78°. Each curve has just

[1] The way this figure is made may be thought of as follows: Suppose we drew a series of charts of the estimated differences in yield with differences in rainfall, with one chart for an average temperature of 70°, one for 72°, one for 74°, etc. Then if we cut these charts off at the yield line, and arrange them one back of the other, at even distances, we have a figure looking much like Figure 62. The lines sloping across the surface from left to right represent what would be the tops of this series of charts. (In this figure the estimates are charted for all combinations of the two variables, even for some not represented in the sample and not shown in Table 68.)

exactly the same shape, and the only difference is their elevation above the base. On looking at it the other way, we find that the same is true of temperature. With 9 inches of rainfall the maximum yield is obtained at about 75° temperature, and the maximum is also at 75° with other levels of rainfall. This relation necessarily follows the assumptions made in measuring it. Figure 62 merely shows the estimate we get by the use of equation (54):

$$X_1 = a + f_2(X_2) + f_3(X_3)$$

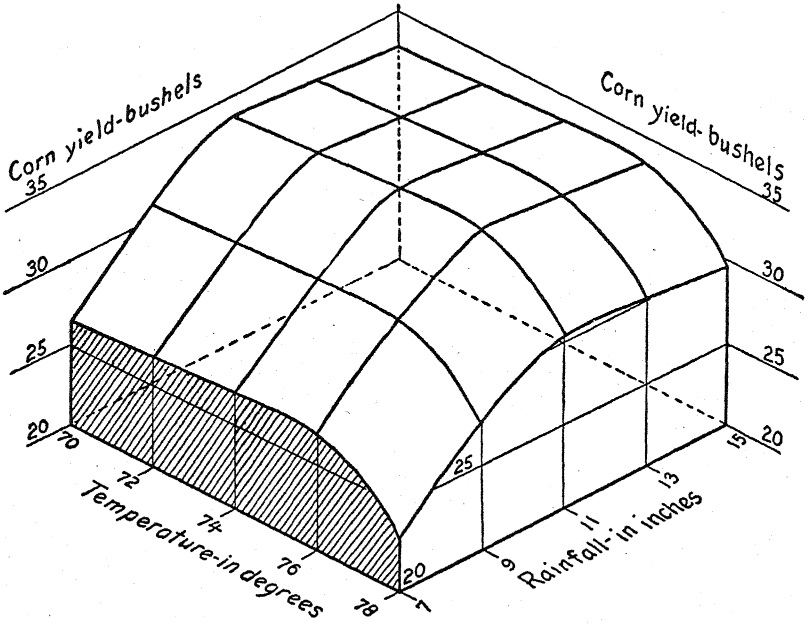


FIG. 62. Probable yield of corn for various specific combinations of rainfall and temperature, from multiple curvilinear correlation.

In working out these estimates we simply add together the estimated value for X_2 and the estimated value for X_3. It does not make any difference what the value of X_3 is, the changes in X_1 assumed to accompany particular changes in X_2 are the same—and that is what the figure shows.

Only a little reflection is needed to indicate that Figure 62 may not tell the whole truth of the relation of yield to rainfall and temperature. It is quite possible that the crop can use more rain in a hot season than in a cool one, so that the rainfall which will produce the maximum crop may be higher in a season of high average temperature than in a season of low temperature. If that is really the case, equation (54) is unable to express the relationship, for, as just pointed out, that equation assumes that the change in yield with rainfall is the same, no matter what the temperature.

An extreme illustration of a changing relationship is shown in Figure 63. This figure, which is based on actuarial investigations,[2] shows the differences in mortality among men from the usual rate, for differences in weight at different ages. Taking the 22-year line, for example, we see that men who are much over normal weight have a much higher mortality than normal for that age. Then as the weight is less the mortality is less, until at normal weight there is only normal mortality. But as the weight drops still more, the mortality increases again, until below 80 per cent of the normal weight the mortality is more than 20 per cent in excess of normal.

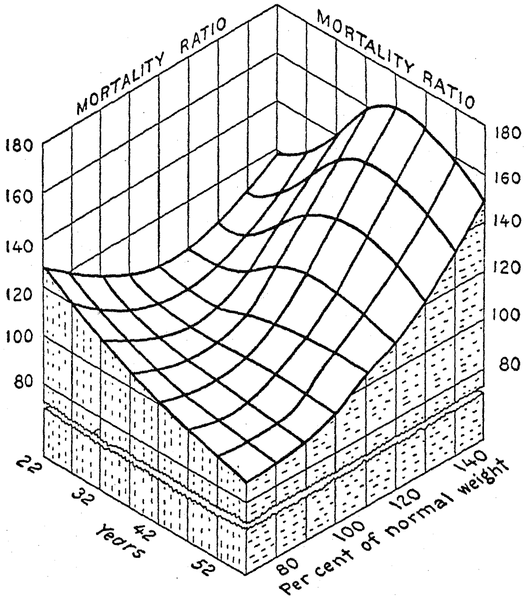


FIG. 63. Differences in mortality with differences in weight, for men of various ages. (Each in percentage of average mortality for that age.) Illustration taken from an article by Andrew Court.

The relation is somewhat different for 52-year-old men, however. For them the mortality is also higher for those who are above normal

[2] *Medico-Actuarial Investigations,* Vol. II, p. 24, 1913.

in weight and decreases as normal weight is reached. But as the weight falls below normal the mortality continues to decrease, until for men who are only 70 per cent of normal weight, the mortality is more than 15 per cent *below* the normal for that age. For ages intermediate between these two, the change is also intermediate—as is shown in the chart, 27 years is similar to 22, but not so marked, and the line for 47 years is similar to that for 52. At 42 years, there is apparently little difference in mortality anywhere between 70 per cent of normal weight and 100 per cent.

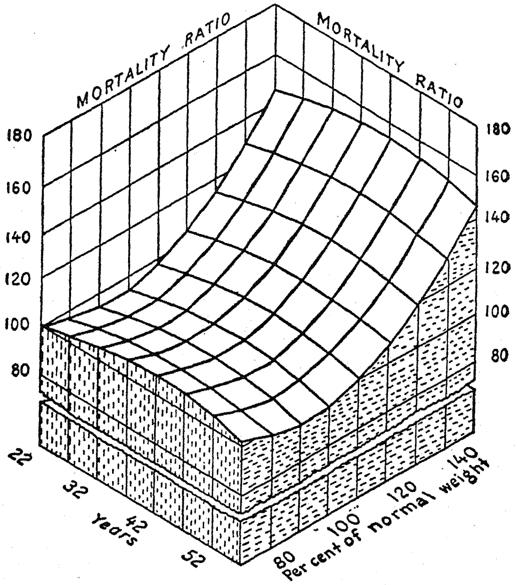


FIG. 64. Relation shown in previous figure, represented by equation

$$X_1 = f_2(X_2) + f_3(X_3).$$

Figure 63 illustrates a situation which the previous methods of analysis would be quite incapable of dealing with adequately. Were equation (54) used to represent this relation, the higher mortality with lower weights for young men would tend to balance out the lower mortality for the older men at the same weight. In fact, the erroneous conclusion might be reached that the age does not affect the mortality at all. Figure 64 shows the results of an attempt to represent this relation by the methods previously discussed. It is quite obvious that the results fall far short of the relations as shown by Figure 63.

Use of "joint functions" to show combined effects. What is needed in both the corn-yield problem and the mortality problem is some way of determining what the yield, in the one case, or the mortality, in the other, is most likely to be for any given *combination* of the two independent variables. That is quite different from asking for the separate effect of each one. Obviously, a small change in one independent factor will be expected to be accompanied by only a small change in the dependent, so that all the estimated yields (or mortalities) will be expected to lie along a continuous surface like that shown in Figure 62 or 63; but the surface will be free to warp or change its shape in different portions like the surface shown in Figure 63, instead of being held rigidly to the same shape in each dimension, like the surfaces in Figure 62 or 64. Mathematically, such a changing relation between one variable and two or more others is known as a *joint functional relation,* and may be indicated by the equation:

$$X_1 = f(X_2,\ X_3) \tag{84}$$

This is read simply that "X_1 is a joint function of X_2 and X_3." That means only that, for any combination of values of X_2 and X_3, there will be some particular value of X_1. Equation (84) is therefore capable of representing either a relation such as that shown in Figure 62, or the more complex relation shown in Figure 63.

The problem of determining the extent to which corn yield varies with the joint effect of temperature and rainfall may be said to be one of determining the functional relation of yield to the two other factors, according to the relation shown in equation (84).

Determining a joint function for two independent variables. Where only two independent variables are concerned, the joint functional relation may be determined quite simply, if a large enough number of observations is available.

The process may be illustrated by data from a different problem, shown in Table 79. The observations are from a field study of haystack dimensions in the Great Plains area. Farmers in this area ordinarily sell their hay unbaled and in the stack. It is therefore necessary to estimate the quantity of hay each stack contains. Two measurements, which can be made readily with only a rope, are usually employed—the perimeter around the base of the stack and the "over," or the distance from the ground on one side of the stack over the center to the ground on the other. The observations shown in Table 79 are all for round stacks. These stacks vary in height and shape to some extent, however, so their volume cannot be computed from the basal circumference by any simple mathematical rule, such as for the volume of a hemisphere. The volumes shown in the table are computed from careful surveying measurements of all the dimensions of each stack—much more exact measurements than a farmer would be able to make in practice. The problem is to establish the average volume for specified circumferences and "overs," so the farmers may be able to use these two measurements, and also to determine how much confidence can be placed in estimates of volume based on these two factors.

The volume will tend to be some function of the basal area times the height. The basal area is a function of the square of the basal circumference; the "over" is a function of both the basal diameter and the height—but attempts to separate the two have been unsuccessful. It is obvious, however, that any attempt to represent the relation by a regression equation of the type

$$\text{volume} = f\,(\text{circumference}) + f\,(\text{“ over ”})$$

will be unsatisfactory because of the multiplying nature of the relations, that is,

$$\text{volume} = f\,(\text{circumference})\ (\text{over})$$

Such a relationship may be approached by use of the relation

$$\log_{\text{volume}} = f\,(\log_{\text{circumference}}) + f\,(\log_{\text{over}})$$

Attempts to determine the relationship by this equation, however, have not been fully successful. The shape of the stacks apparently shifts with changes in size.

The haystack problem is evidently one where the relation may best be expressed by a joint function such as

$$\text{volume} = f\,(\text{circumference, over})$$

Such a relation could be determined directly from the data by the methods which will presently be described. It is evident that the correlation surface would have a marked upward slope as the two dimensions increased together, even if the usual volume formulas applied. The work for this particular problem may be somewhat simplified by first stating each variable as a logarithm and then determining the joint relation according to the equation

$$\log_{\text{volume}} = f\,(\log_{\text{circumference}}, \log_{\text{over}})$$

TABLE 79

DATA TAKEN FROM NEBRASKA ROUND STACKS MEASURED IN 1927 AND 1928†

Volume, in cubic feet	Circumference, in feet	"Over," in feet	X_2*	X_3*	X_1*	X_1'	z
2853.00	69.0	37.00	0.139	0.168	0.455	0.478	−0.023
2702.00	65.0	36.50	0.113	0.162	0.432	0.450	−0.018
3099.00	73.0	38.50	0.163	0.185	0.491	0.447	0.044
1306.00	62.5	26.50	0.096	0.023	0.116	0.143	−0.027
2294.00	70.0	35.00	0.145	0.144	0.361	0.436	−0.075
2725.00	68.0	36.50	0.133	0.162	0.435	0.421	0.014
3309.00	71.0	39.25	0.151	0.194	0.520	0.557	−0.037
2790.00	64.0	36.75	0.106	0.165	0.446	0.450	−0.004
2756.00	62.0	38.50	0.092	0.185	0.440	0.478	−0.038
5237.92	80.0	43.00	0.203	0.233	0.719	0.705	0.014
3149.82	67.0	37.60	0.126	0.175	0.498	0.490	0.008
5498.46	79.0	44.60	0.198	0.249	0.740	0.739	0.001
3397.83	66.0	38.00	0.120	0.180	0.531	0.541	−0.010
3007.56	62.0	36.80	0.092	0.166	0.478	0.486	−0.008
4574.29	79.0	41.10	0.198	0.214	0.660	0.596	0.064
6228.59	73.0	48.00	0.163	0.281	0.794	0.780	0.014
2318.64	63.0	30.20	0.099	0.080	0.365	0.265	0.100
3176.71	68.0	37.75	0.133	0.177	0.502	0.502	0
2352.31	70.0	32.50	0.145	0.112	0.371	0.363	0.008
2174.44	69.0	31.62	0.139	0.100	0.337	0.333	0.004
2694.72	73.0	34.50	0.163	0.138	0.431	0.433	−0.002
3333.53	70.0	37.25	0.145	0.171	0.523	0.500	0.023
4328.92	78.5	40.00	0.195	0.202	0.636	0.617	0.019
2115.04	67.0	31.25	0.126	0.095	0.325	0.317	0.008
2489.08	66.5	33.75	0.123	0.128	0.396	0.388	0.008
2296.65	64.5	32.38	0.110	0.110	0.361	0.338	0.023
3117.21	65.5	37.58	0.116	0.175	0.494	0.480	0.014
4088.36	74.0	40.33	0.169	0.206	0.612	0.602	0.010
4180.88	72.0	40.50	0.157	0.207	0.621	0.594	0.027
2318.19	63.0	33.00	0.099	0.119	0.365	0.346	0.019
1946.90	58.0	31.00	0.063	0.091	0.289	0.255	0.034
2479.89	61.0	36.50	0.086	0.162	0.394	0.423	−0.029
3174.80	73.0	37.00	0.163	0.168	0.502	0.506	−0.004
2151.54	64.0	33.00	0.106	0.119	0.333	0.353	−0.020
3475.68	73.0	39.50	0.163	0.197	0.541	0.576	−0.035
4393.08	71.0	42.00	0.151	0.223	0.643	0.624	0.019
2819.50	69.0	35.00	0.139	0.144	0.450	0.432	0.018
3703.49	70.0	38.50	0.145	0.185	0.569	0.530	0.039
2742.81	72.5	34.50	0.160	0.138	0.438	0.430	0.008

* $X_2 = \log_{10}$ (circumference) − 1.700, stated to three decimal places.

$X_3 = \log_{10}$ ("over") − 1.4, stated to three decimal places.

$X_1 = \log_{10}$ (volume) − 3.0, stated to three decimal places.

† Acknowledgment is due W. H. Hosterman, of the Bureau of Agricultural Economics, U. S. Department of Agriculture, for the use of these data.

TABLE 79—*Continued*

Volume, in cubic feet	Circumference, in feet	"Over," in feet	X_2*	X_3*	X_1*	X_1'	z
3002.40	66.0	35.50	0.120	0.150	0.477	0.430	0.047
1854.19	69.0	30.50	0.139	0.084	0.268	0.297	−0.029
1982.07	62.0	31.00	0.092	0.091	0.297	0.288	0.009
2470.86	65.0	33.50	0.113	0.125	0.393	0.373	0.020
1203.15	60.1	26.25	0.079	0.019	0.080	0.117	−0.037
2843.84	71.0	36.00	0.151	0.156	0.454	0.469	−0.015
2636.25	66.0	36.00	0.120	0.156	0.421	0.443	−0.022
1998.39	65.0	32.00	0.113	0.105	0.301	0.330	−0.029
2005.03	64.0	32.00	0.106	0.105	0.302	0.323	−0.021
2568.76	66.0	35.00	0.120	0.144	0.410	0.418	−0.008
2161.18	65.0	32.50	0.113	0.112	0.335	0.345	−0.010
2112.20	67.0	32.00	0.126	0.105	0.325	0.333	−0.008
3009.33	65.0	38.00	0.113	0.180	0.478	0.438	0.040
1992.24	63.0	31.00	0.099	0.091	0.299	0.288	0.011
2746.98	70.0	34.00	0.145	0.131	0.439	0.407	0.032
2238.27	64.0	35.00	0.106	0.144	0.350	0.406	−0.056
1747.47	67.0	30.00	0.126	0.077	0.242	0.280	−0.038
2863.91	67.0	36.00	0.126	0.156	0.457	0.448	0.009
3593.47	72.0	39.00	0.157	0.191	0.555	0.555	0
2435.48	62.0	35.00	0.092	0.144	0.387	0.443	−0.056
2430.18	63.0	34.00	0.099	0.131	0.386	0.362	0.024
2590.07	67.0	35.00	0.126	0.144	0.413	0.423	−0.010
3577.68	70.0	41.00	0.145	0.213	0.554	0.596	−0.042
3299.24	73.0	40.00	0.163	0.202	0.518	0.598	−0.080
1986.14	64.0	32.50	0.106	0.112	0.298	0.338	−0.040
3109.04	68.0	38.00	0.133	0.180	0.493	0.508	−0.015
2821.56	71.0	37.00	0.151	0.168	0.450	0.498	−0.048
2932.24	67.0	38.00	0.126	0.180	0.467	0.501	−0.034
3304.63	69.0	38.00	0.139	0.180	0.519	0.514	0.005
2565.46	72.0	35.00	0.157	0.144	0.409	0.450	−0.041
4509.93	74.0	41.33	0.169	0.216	0.654	0.627	0.027
4804.01	81.0	42.00	0.208	0.223	0.682	0.683	−0.001
4241.80	75.0	40.75	0.175	0.210	0.627	0.619	0.008
4516.10	69.2	43.25	0.140	0.236	0.655	0.643	0.012
5011.62	77.5	43.10	0.189	0.234	0.700	0.691	0.009
2110.73	65.0	31.50	0.113	0.098	0.324	0.316	0.008
2775.70	76.0	34.60	0.181	0.139	0.443	0.448	−0.005
3927.90	72.0	39.00	0.157	0.191	0.594	0.555	0.039
4212.77	80.0	41.50	0.203	0.218	0.624	0.663	−0.039
3562.64	78.5	38.50	0.195	0.185	0.552	0.575	−0.023

* $X_2 = \log_{10}$ (circumference) − 1.700, stated to three decimal places.
$X_3 = \log_{10}$ ("over") − 1.4, stated to three decimal places.
$X_1 = \log_{10}$ (volume) − 3.0, stated to three decimal places.

TABLE 79—*Continued*

Volume, in cubic feet	Circumference, in feet	"Over," in feet	X_2*	X_3*	X_1*	X_1'	z
2853.96	75.0	35.50	0.175	0.150	0.455	0.461	−0.006
3294.38	69.0	38.00	0.139	0.180	0.518	0.514	0.004
1689.54	63.0	30.50	0.099	0.084	0.228	0.274	−0.046
2228.84	62.0	33.00	0.092	0.119	0.348	0.341	0.007
2362.61	64.0	34.00	0.106	0.131	0.373	0.379	−0.006
3088.28	68.0	38.50	0.133	0.185	0.490	0.520	−0.030
3820.79	70.0	40.00	0.145	0.202	0.582	0.570	0.012
3126.64	63.0	36.90	0.099	0.167	0.495	0.447	0.048
3624.75	71.0	38.45	0.151	0.185	0.559	0.536	0.023
3023.97	73.0	36.50	0.163	0.162	0.480	0.493	−0.013
6045.42	79.0	47.00	0.198	0.272	0.781	0.798	−0.017
3100.11	64.0	37.00	0.106	0.168	0.491	0.457	0.034
3378.07	70.0	38.00	0.145	0.180	0.529	0.519	0.010
3040.29	77.0	35.00	0.186	0.144	0.483	0.464	0.019
2252.16	65.0	32.50	0.113	0.112	0.353	0.345	0.008
3552.61	76.0	37.00	0.181	0.168	0.551	0.481	0.070
2635.90	66.0	34.50	0.120	0.138	0.421	0.405	0.016
3201.41	71.0	35.50	0.151	0.150	0.505	0.455	0.050
2590.21	69.0	35.00	0.139	0.144	0.413	0.432	−0.019
3743.55	76.0	38.25	0.181	0.183	0.573	0.558	0.015
3858.03	73.0	39.50	0.163	0.197	0.586	0.576	0.010
3829.44	74.0	39.75	0.169	0.199	0.583	0.586	−0.003
2556.44	66.0	33.00	0.120	0.119	0.408	0.365	0.043
3119.07	69.0	36.00	0.139	0.156	0.494	0.460	0.034
2122.38	65.5	32.00	0.116	0.105	0.327	0.332	−0.005
2921.92	69.0	36.00	0.139	0.156	0.466	0.460	0.006
2936.35	72.5	34.50	0.160	0.138	0.468	0.430	0.038
2427.66	76.0	33.00	0.181	0.119	0.385	0.399	−0.014
2069.38	65.0	31.50	0.113	0.098	0.316	0.315	0.001
1899.54	72.0	30.00	0.157	0.077	0.279	0.285	−0.006
4289.28	78.5	40.50	0.195	0.207	0.632	0.629	0.003
2407.39	67.5	32.50	0.129	0.112	0.381	0.358	0.023
3097.99	66.0	35.50	0.120	0.150	0.491	0.430	0.061
3893.67	75.5	39.25	0.178	0.194	0.590	0.582	0.008
2238.66	68.0	31.75	0.133	0.102	0.350	0.336	0.014
2314.79	64.0	33.10	0.106	0.120	0.364	0.356	0.008
2667.07	66.0	34.70	0.120	0.140	0.426	0.409	0.017
2582.07	68.0	33.50	0.133	0.125	0.412	0.388	0.024
3426.50	75.0	37.00	0.175	0.168	0.535	0.516	0.019
2307.34	60.0	33.40	0.078	0.124	0.363	0.336	0.027
3960.41	76.0	39.30	0.181	0.194	0.598	0.585	0.013

* $X_2 = \log_{10}$ (circumference) − 1.700, stated to three decimal places.
$X_3 = \log_{10}$ ("over") − 1.4, stated to three decimal places.
$X_1 = \log_{10}$ (volume) − 3.0, stated to three decimal places.

The logarithms (to base 10) are accordingly also shown in Table 79, and are designated as X_2, X_3, and X_1. (To facilitate the subsequent computations, 1.7 has been subtracted from the logarithm for circumference, 1.4 from the logarithm for "over," and 3.0 from the logarithm for volume.)

Subgrouping and averaging the observations. The first step in the process of determining the joint functional relation is to classify the observations according to X_2, and subclassify according to X_3, and determine the averages of X_1, X_2, and X_3 for each group. Since there are only 120 observations, it would not be worth while to make too many groups. Four groups each way would give 16 subgroups, and 5 each way would give 25. If the cases were uniformly distributed through 25 subgroups, that would make less than 5 cases to a group, which is rather thin for a satisfactory average (though it might be sufficient in this particular problem, where the correlation is much higher than in many problems which must be dealt with.) However, the cases will not necessarily be distributed uniformly through all the groups, so it will be best if we try the fivefold classification and see how the cases fall.

TABLE 80

NUMBER OF HAYSTACK OBSERVATIONS, CLASSIFIED ACCORDING TO X_2 AND X_3
(Logarithms of Circumference and "Over")

X_3 values	X_2 values				
	Under 0.090	0.090–0.119	0.120–0.149	0.150–0.179	0.180 and over
Under 0.100	2	7	3	1	
0.100–0.139	1	14	10	3	2
0.140–0.179	1	8	17	8	2
0.180–0.219	......	2	10	14	7
0.220 and over	......	......	1	2	5

There is a marked correlation between X_2 and X_3, so a few groups have 10 or more reports, whereas 15 out of the 25 have under 5. Preliminary examination of the data indicates that a unit change in X_3 is generally accompanied by a larger change in X_1 than is a unit change in X_2. Accordingly we may decide to halve the groups in the central portion of the range of X_3, making the class intervals with respect to that variable under 0.100, 0.100 – 0.119, 0.120 – 0.139, 0.140 – 0.159, 0.160 − 0.179, 0.180 − 0.199, 0.200 − 0.219, and 0.220 and over. With 5 classes for X_2, this will give a 40-group classification—but with many of the "cells" vacant. Averaging X_2, X_3, and X_1 for each of the resulting groups gives means as shown in Table 81.

Plotting the subgroup averages and drawing first approximation curves. Inspection of the averages of X_2 down each column in Table 81 shows that most of the variation in that factor has been eliminated, except in the upper subgroups of X_3, above $X_3 = 0.200$, where the averages tend to fall above the mean of the range. There is a more marked tendency for the averages of X_3 to rise across the rows from left to right. Accordingly the groups classified with respect to X_2 will be studied first, to determine the changes in X_1 with changes in X_3, X_2 being held (approximately) constant at various values. This may be done by plotting separately the average difference in X_1 with differences in X_3, for each column. Figure 65 shows these averages, with the

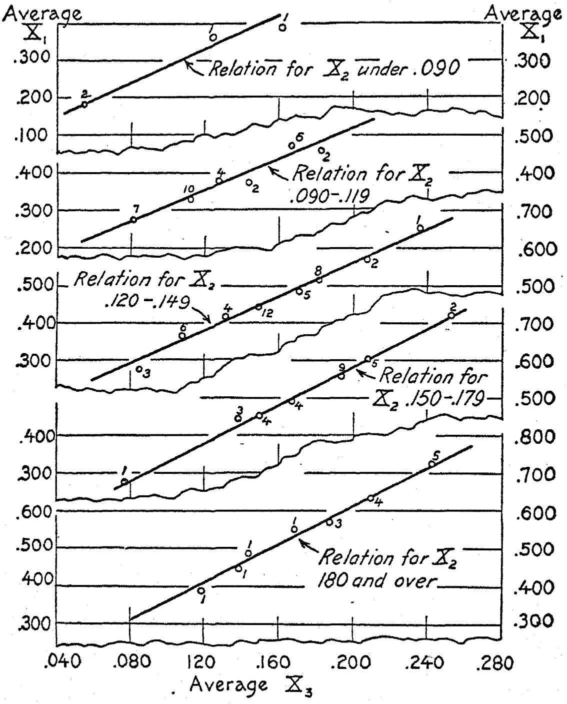


FIG. 65. Differences in X_1, with differences in X_3 for specified values of X_2, and first approximate curves.

TABLE 81

HAYSTACK DATA: AVERAGE X_2, X_3, AND X_1, FOR OBSERVATIONS CLASSIFIED BY X_2 AND X_3

X_3 values	Number of cases	X_2 under 0.090		
		Mean X_2	Mean X_3	Mean X_1
Under 0.100	2	0.071	0.055	0.185
0.100–0.119				
0.120–0.139	1	0.078	0.124	0.363
0.140–0.159				
0.160–0.179	1	0.086	0.162	0.394
		X_2 0.090–0.119		
Under 0.100	7	0.102	0.081	0.278
0.100–0.119	10	0.107	0.112	0.332
0.120–0.139	4	0.106	0.127	0.379
0.140–0.159	2	0.099	0.144	0.369
0.160–0.179	6	0.105	0.167	0.473
0.180–0.199	2	0.103	0.183	0.459
		X_2 0.120–0.149		
Under 0.100	3	0.130	0.085	0.278
0.100–0.119	6	0.132	0.108	0.362
0.120–0.139	4	0.130	0.131	0.417
0.140–0.159	12	0.129	0.149	0.440
0.160–0.179	5	0.135	0.171	0.483
0.180–0.199	8	0.135	0.181	0.515
0.200–0.219	2	0.145	0.208	0.568
0.220 and over	1	0.140	0.236	0.655
		X_2 0.150–0.179		
Under 0.100	1	0.157	0.077	0.279
0.100–0.119				
0.120–0.139	3	0.161	0.138	0.446
0.140–0.159	4	0.159	0.150	0.456
0.160–0.179	4	0.163	0.167	0.492
0.180–0.199	9	0.161	0.193	0.558
0.200–0.219	5	0.167	0.208	0.606
0.220 and over	2	0.157	0.252	0.719
		X_2 0.180 and over		
0.100–0.119	1	0.181	0.119	0.385
0.120–0.139	1	0.181	0.139	0.443
0.140–0.159	1	0.186	0.144	0.483
0.160–0.179	1	0.181	0.168	0.551
0.180–0.199	3	0.186	0.187	0.574
0.200–0.219	4	0.198	0.210	0.638
0.220 and over	5	0.199	0.242	0.724

number of observations represented by each one indicated. Apparently the relation, for each group of averages, tends to be linear, so straight lines are drawn in by eye as first approximations to the final relation. It should be noticed, however, that these lines are not all of the same slope but tend to slope more steeply as X_2 increases. In some problems curves instead of straight lines would be indicated by these group averages. In such a case, separate curves would be fitted freehand to each set of averages. In drawing such curves it is desirable to keep them as nearly of the same shape as the data will permit, and to change the shape only gradually from one to the next.

Obtaining a second approximation to the joint surface. After the first approximation lines or curves have been drawn along the X_3 axis, the next step is to smooth along the X_2 axis. To do this the values of X_1 according to the first approximation curves are read off at intervals on X_3, corresponding to the central values of the groups of X_3 in Table 81. These values are shown in Table 82.

TABLE 82

ESTIMATED VALUES OF X_1 FOR SPECIFIED VALUES OF X_2 AND X_3, FROM FIRST APPROXIMATION LINES

X_3	X_2				
	Under 0.090	0.090–0.119	0.120–0.149	0.150–0.179	0.180 and over
0.060	0.195	0.233	0.247	....	....
0.110	0.307	0.334	0.359	0.354	0.382
0.130	0.352	0.374	0.403	0.404	0.433
0.150	0.397	0.415	0.448	0.454	0.484
0.170	0.442	0.455	0.492	0.505	0.535
0.190	....	0.496	0.537	0.555	0.586
0.210	....	....	0.583	0.605	0.637
0.240	....	....	0.650	0.681	0.713

The readings shown in Table 82 may next be smoothed along the X_2 axis by plotting the estimated value of X_1 for the specified values of X_3, with varying values of X_2. This process is shown in Figure 66. The values used as the coordinates for X_2 are the corresponding average values of X_2 for each subgroup in Table 81. Thus in plotting the values of X_1 for $X_3 = 0.060$—the first line in the table above—the values for X_2 are the averages from the first lines in Table 81—0.071, 0.102, 0.130, etc. By using these actual averages allowance is made for cases such as those noted previously, where not even the process of subgrouping has completely removed the influence of the other independent variable.

The averages plotted in Figure 66 show a slight but consistent curvilinear relation of X_1 to X_2, with a gradually increasing slope for the higher values. In two cases straight lines would fit as well as curves, but, since all the remaining groups show consistent slight curves, they are drawn in here as well. The averages, smoothed freehand along the X_2 axis, give a second approximation to the joint functional relation.

Making the final smoothing of the approximation curve. As a final check, values from the curves in Figure 66 may be read off for stated values of X_2, and smoothed again with reference to X_3. Since the variation in the averages of X_2 in each column of Table 81, and in the averages of X_3 in each row, have now been allowed for by the methods used in constructing Figures 65 and 66, the new readings may be taken for any convenient interval on X_2. The values 0.070, 0.080, 0.100, 0.120, 0.140, 0.160, 0.180, and 0.200 may be used, giving convenient values for subsequent interpolations. Reading off the corresponding X_1 values from the curves in Figure 66, just shown before in Table 82, and plotting with the X_3 values as abscissas, gives the results shown in Figure 67.

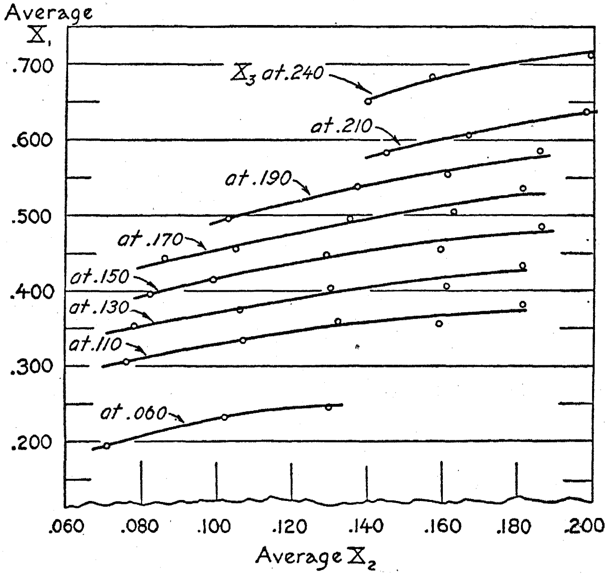


FIG. 66. Differences in X_1 with differences in X_2 for specified values of X_3, read from smoothed curves in Fig. 65.

After the X_1 readings from Figure 66 are plotted, as indicated by the hollow circles in Figure 67, they are smoothed with reference to the X_3 axis. It is found again that straight lines serve to describe the relation, and these are accordingly drawn in by eye, with some consideration of adjacent lines where otherwise the line would be out of agreement, as for $X_2 = 0.200$. These lines, showing the estimated values of X_1 for specified differences in X_2 and X_3, may be taken as defining the functional relation between the three variables. This figure indicates very clearly the "warping" of the regression surface. The increase in X_1 per unit increase in X_3 is much greater for large values of X_2 than for small values. That fact could not have been expressed in any regression equation of the form $X_1 = f_2(X_2) + f_3(X_3)$. The shape of the "correlation surface" may be seen in Figure 68, where the final lines from Figure 67 have been combined into a three-dimensional diagram.

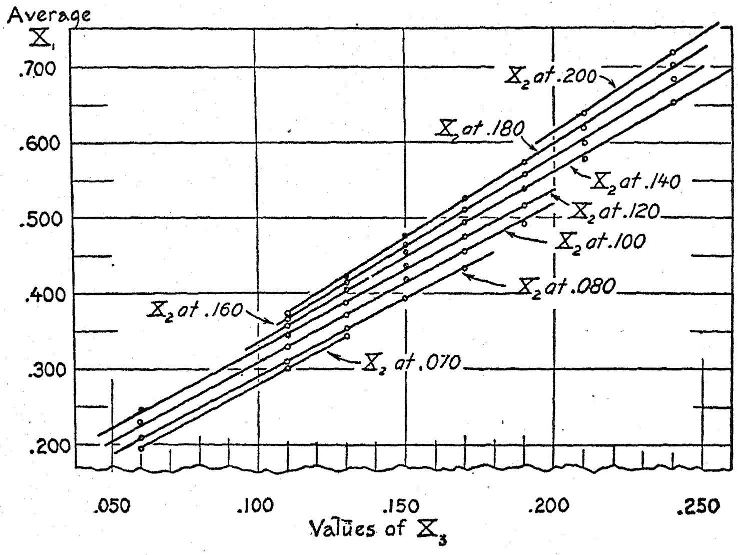


FIG. 67. Differences in X_1 with differences in X_3, from second smoothed curves.

Estimating X_1 from the joint function. Estimates of X_1 for any combination of values of X_2 and X_3 may be made directly from Figures 67 or 68 by making the necessary interpolations. The process may be more conveniently carried out by making a "contour chart," which shows the differences in X_1, for different combinations of X_2 and X_3, by a series of lines passing through combinations of the other two variables which will produce equal values of X_1. Thus if a series of planes were passed through the cube in Figure 68, parallel to the base plane, at $X_1 = 0.300$, 0.400, 0.500, etc., they would cut the top surface of the solid in the intersections indicated by the dotted lines. Then if one were to look straight down upon the top of the solid, these dotted lines would appear as shown in Figure 69. Other lines have been drawn in between these to indicate 0.050 differences in X_1.[3]

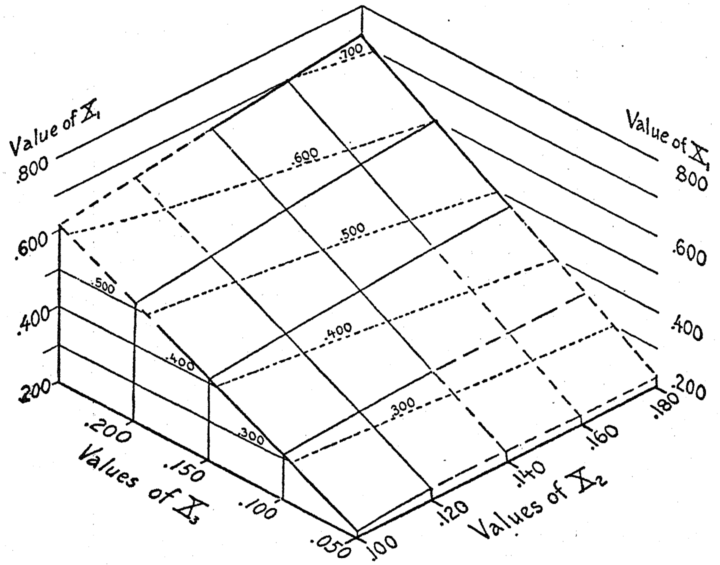


FIG. 68. Probable value of X_1 for specific combinations of X_2 and X_3, from smoothed surface.

Determining the standard error of estimate and index of multiple correlation. Estimated yields for each observation may now be worked out by the use of Figure 67 or 69, interpolating for the distances between the adjacent lines where an observation falls off the line. Table 79 also shows these estimated values (X'_1), and the difference between the actual and the estimated. The standard deviation of the original

[3] Figure 69 may be most readily drawn from Figure 67. By noting the value of X_3 necessary to produce a value of $X_1 = 0.300$, for each X_2 line, a series of values for X_2 and X_3 may be located, through which the contour for $X_1 = 0.300$ is drawn. The values for $X_1 = 0.400$ are then noted, giving the location of the 0.400 contour, and so on.

values of X_1 is 0.1265, whereas the standard deviation of the residuals computed in Table 79 is 0.0295. Apparently the regression surface accounts for almost all the variation in volume. The accuracy with which estimates of X_1 may be made from X_2 and X_3 may be determined by adjusting σ_z in the usual way.

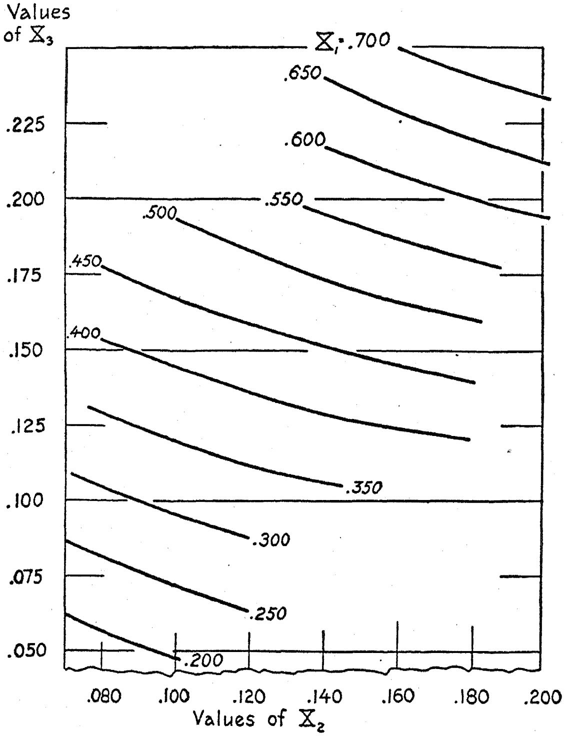


FIG. 69. Probable value of X_1 for specific combinations of X_2 and X_3, shown by contours.

For the type of surface shown, the relations might be quite closely represented by an equation of the type [4]

$$X_1 = a + b_2X_2 + (b_3 + b_3'X_2)X_3$$

This equation expresses the relation shown in Figure 68: for a constant value of X_2, the regression of X_1 on X_3 is linear; but as X_2 changes, this regression also changes at a uniform rate. The equation given has

[4] See page 408 for a further discussion of the possibilities of this type of formula.

4 constants, so in adjusting σ_z to determine the standard error of estimate, $m = 4$. Hence

$$\begin{aligned}\bar{S}^2_{x_1[f(x_2x_3)]} &= \sigma_z^2\left(\frac{n}{n-m}\right)\\ &= (0.0295)^2\left(\frac{120}{116}\right)\\ \bar{S}_{x_1[f(x_2x_3)]} &= 0.0300\end{aligned}$$

Similarly, the index of multiple correlation for X_1 as a joint function of X_2 and X_3 may be computed in the usual manner:

$$\begin{aligned}\mathrm{P}^2 &= 1 - \frac{\sigma_z^2}{\sigma_1^2} = 1 - \frac{(0.0295)^2}{(0.1265)^2}\\ &= 1 - \frac{0.000870}{0.01600}\\ &= 0.9456,\end{aligned}$$

and adjusting for the number of constants,

$$\begin{aligned}\bar{\mathrm{P}}^2 &= 1 - (1 - \mathrm{P}^2)\left(\frac{n-1}{n-m}\right)\\ &= 1 - (1 - 0.9456)\left(\frac{119}{116}\right)\\ &= 0.9442\\ \bar{\mathrm{P}} &= 0.972\end{aligned}$$

It is evident that the volume of a round haystack may be very closely estimated from the rough farm measurements of circumference and "over." The standard error of estimate, 0.0300, indicates that the logarithm of volume can be estimated to ± 0.0300 of the true logarithm for two-thirds of the observations and to ± 0.0600 of the true logarithm for 95 per cent of them. Taking the antilogarithms of 0.0300 and $\bar{1}.9700$, and of 0.0600 and $\bar{1}.9400$, we find that that means the volume can be estimated to between 107.2 per cent and 93.3 per cent, or between 114.9 per cent and 87.1 per cent, of the true values, respectively, for the proportions stated.

Stating the conclusions shown by the joint function. After the joint relation of one variable to two others has been determined by the method sketched, the final regression surface, as expressed in Figures 67, 68, or 69, may be restated in simpler terms by preparing tables showing the average or expected values of X_1 for stated combinations of X_2 and X_3. In this particular problem, where the surface was determined with respect to logarithmic values, that involves determining the logarithms of X_2 and X_3 for the selected values, reading off from the charts the corresponding estimated value for the logarithm of X_1, and finding its antilogarithm. Carrying out this process, we obtain the values shown in Table 83.

TABLE 83

Average Volume of Round Haystacks for Different Combinations of Circumference and "Over"

Circumference	"Over," in feet				
	30	34	38	42	46
	Cubic feet	*Cubic feet*	*Cubic feet*	*Cubic feet*	*Cubic feet*
60 feet	1,730	2,244			
65 feet	1,871	2,432	3,097		
70 feet	1,928	2,553	3,319	4,150	
75 feet	....	2,655	3,524	4,467	5,623
80 feet	....	....	....	4,710	6,026

Determining joint influence of two independent variables, holding other independent variables constant. In many cases it may be desirable to allow for the joint influence of two variables while simultaneously eliminating or holding constant the effect of one or more additional independent variables. In the corn problem it would be desirable to determine the relation of yield to rainfall and temperature jointly, while simultaneously allowing for the upward tendency in yield during the period studied. This may be done by determining the relation according to the equation

$$X_1 = f_{2,3}(X_2, X_3) + f_4(X_4) \tag{85}$$

This relation may be worked out by combining the method just shown for determining a joint function for two independent variables with the method of successive approximation for handling many variables, as discussed in Chapters 14 and 16. The essential steps are (1) to determine the curvilinear changes in X_1 with changes in X_2, X_3, and X_4, according to the simpler equation,

$$X_1 = a + f_2(X_2) + f_3(X_3) + f_4(X_4)$$

and then (2) to compute the residuals for each observation, using these curves, and subclassify the residuals according to the two variables for which the presence of a joint function is to be tested. If these averages of residuals indicate any significant warping of the surface, (3) they are next smoothed by the method presented following Table 81. The residuals may then (4) be adjusted to take account of this joint relation in addition to the individual curvilinear relations previously allowed for, and their standard deviation computed. If the variance has been significantly reduced, the residuals may then (5) be averaged with respect to the remaining independent factor, to see if the curve for that factor will be changed now that the joint relation to the other factors has been allowed for. If it is changed, the residuals are recomputed to see if any further change need be made in the joint function and the process continued until the final shape of the curve and joint surface is determined.

Measuring correlation with respect to joint functions. The correlation may be measured with respect to joint functions just as before it was measured with respect to curvilinear regressions. The standard error of the residuals, adjusted for the estimated number of constants, indicates the standard error of estimate; and this adjusted standard error, substituted in equation (66.2), gives the index of multiple correlation and of multiple determination. But since the *combined* influence of X_2 and X_3 is being determined, it is not possible to compute coefficients of partial correlation, or other measures of individual importance, for the variables which are being considered jointly. It would be possible to work out what portion of the variance in X_1 was accounted for by X_2 and X_3 together and how much by X_4, but that would be all.[5] It is something of a guess how many constants should be allowed for in computing the correlation and the standard error. It will be higher than for the individual curves for $f_2(X_2)$ and $f_3(X_3)$, in general. If the joint relation merely involves a gradual regular shifting of the slope of the curves across the surface, one additional constant would be enough; if it involves a shifting at an increasing rate, two might be assumed; and if it involves several changes in shift, even more might be needed.

Determining joint influence of three or more independent variables. The methods just described might be used to determine several joint

[5] This would involve determining, by least squares, the equation

$$X_1 = a + b_2[f_{2,3}(X_2, X_3)] + b_3[f_4(X_4)]$$

The separate determination with respect to b_2 would then indicate the determination by X_2 and X_3 combined. See Note 11, Appendix 2.

relations at the same time, each relation involving two independent variables. Thus if X_2 = rain in July, X_3 = temperature in July, X_4 = rain in August, X_5 = temperature in August, and X_6 = time, the yield of corn might be explained by a set of relations represented by the equation

$$X_1 = f_{2,3}(X_2, X_3) + f_{4,5}(X_4, X_5) + f_6(X_6) \tag{86}$$

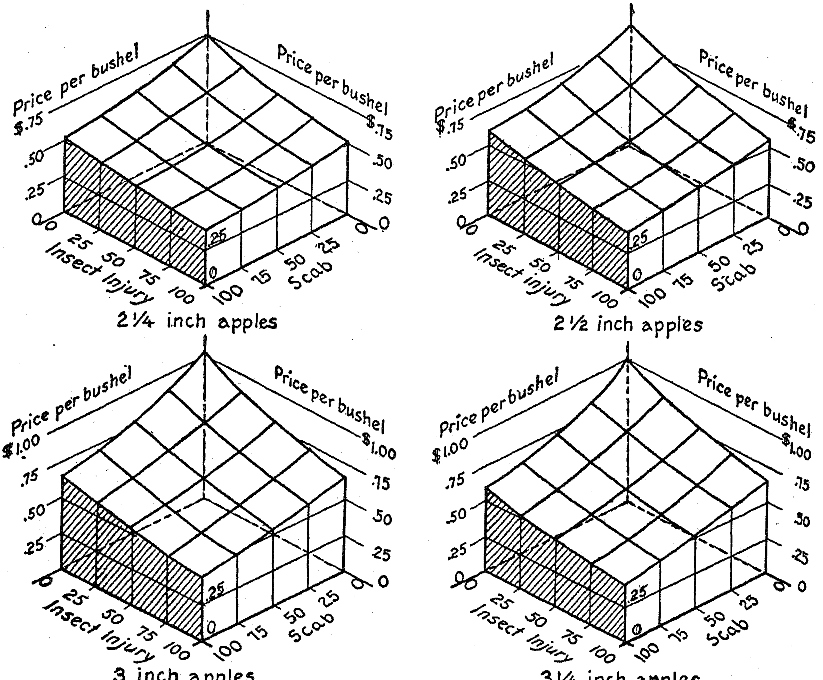
The functions would be determined by first getting the net regression curves for each factor separately, then the joint curves for $f_{2,\,3}(X_2, X_3)$ and $f_{4,\,5}\ (X_4, X_5)$ by classifying the residuals by the method just described, and then determining the final shapes by successive approximations. But there will be some cases where even so flexible a relation as represented in equation (86) will not be sufficient really to represent the relations. For example, yield might depend jointly on rainfall, temperature, and length of growing season, and a change in any one factor might cause differences in the effects of others as well. Such a relation would be represented by such equations as

$$X_1 = f(X_2, X_3, X_4, \dots X_n) \tag{87}$$

To determine the shape of such a function for even three independent variables would require a large number of observations, since a threefold subclassification would be needed. If only 4 classes were used for each variable, 64 subclasses would be possible. Not unless there were sufficient observations so that say 3 to 5 might fall in each class, on the average, could such a relation be determined with any degree of accuracy, unless the correlation was very high indeed. If the joint correlation were perfect, one case to a subclass would be sufficient to indicate the nature of the function.

With three independent variables, successive smoothing in three dimensions would be involved. Where an adequate number of observations was available, the process might be simplified by dividing the observations into several groups according to one variable, determining the functional relation to the other two independent variables separately for each group, and then smoothing the results for the different groups together to determine the change in joint function with changes in the first variable.

Figure 70 illustrates some results of this sort, for a four-dimensional joint function. These results were obtained from an analysis of 190 observations of sales of individual lots of apples. The records were first separated into those for each of the 5 sizes of apples, and the joint functional relation of price to amount of insect injury and amount of scab determined separately for each size. These results were then smoothed between apples of different sizes, to make the "surface" of the imaginary four-dimensional solid diagram show a gradual continuous change over every dimension.[6]



2 1/4 inch apples

2 1/2 inch apples

3 inch apples

3 1/4 inch apples

FIG. 70. Average price of apples of given sizes, for various combinations of amount of insect injury and amount of scab.

While, of course, it is not possible to draw a single diagram expressing the four-dimensional relationship

$$X_1 = f(X_2, X_3, X_4)$$

[6] This is done by reading expected prices for 0 scab, 0 insect injury, for apples of each size, and smoothing that series; reading for 0 scab, 20 per cent insect injury, and smoothing that series; and so on until every portion of the surface has been smoothed with respect to the third independent variable. The smoothed values could then be read off and smoothed again in other dimensions, until the final continuous function was obtained. This illustration is from an analysis supplied by Frederick V. Waugh. For a more elaborate study of the same type, see John R. Raeburn, Joint correlation applied to the quality and price of McIntosh apples, Cornell University Agricultural Experiment Station *Memoir* 220, March, 1939.

the relation may be visualized by a composite diagram, as illustrated in Figure 70. This figure in particular illustrates the significant relations brought out by the joint functional treatment. Thus it is seen that large apples, with neither scab nor insect injury, sold for a material premium over perfect apples of small size; but that if the apples were badly damaged it did not make much difference what size they were. This may be stated another way—the presence of defects reduced the price of large apples much more than the price of small ones. The figure also shows that the presence of either defect alone reduced the value of apples of any size materially, whereas the presence of both defects together reduced the price only slightly more. Thus for 3-inch apples, apples with 0 scab and 0 insect injury sold for \$1.05; those with 0 scab but 100 per cent insect injury, for \$0.66; those with 100 per cent scab and 0 insect injury, for \$0.65; and those with 100 per cent scab and 100 per cent insect injury for \$0.42. Increasing the insect injury from 0 to 100 per cent reduced the price 39 cents for apples with no scab, and only 23 cents for those with 100 per cent scab. Likewise for apples with neither scab nor insect injury 3-inch apples sold for \$1.05, and 2¼-inch ones for 75 cents; whereas for apples of these two sizes with 100 per cent of both injuries, the prices were 42 cents and 39 cents, respectively. These comparisons show what a difference the recognition of joint relations may make in research conclusions, and how important may be the resulting differences in the statement of relations.

Theoretically there is no limit to the number of variables which could be considered jointly. The only practical limitation is the number of observations available. Where it is possible to determine the joint relation, that affords by far the most satisfactory statement of the relationship, since then the real relation is not obscured by the assumptions hidden in the regression equation used. When a limited number of observations precludes a full recognition of joint relations, the use of mathematical transformations such as logarithms, logical grouping of the variables to determine the combinations of variables for which joint relations are most likely to obtain, and trial study of the residuals will serve to make the final regression equation set forth the true nature of the relations as closely as possible from the limited evidence available.

Just as the accuracy of net regression curves depends largely on the number of observations along various portions of the curve and the standard error of estimate, so the reliability of a joint regression surface (such as that shown in Figure 68) would depend on the standard error of estimate and the number of cases falling within the selected portion of the area. Where the joint regression surface is determined mathematically, its reliability can be estimated by an extension of the same equations presented in Chapters 18 and 19. Methods of estimating the standard errors of a surface determined graphically have not yet been developed.

Summary. This chapter has developed means by which the relation between one variable and two others operating jointly may be determined, either where no other variables are concerned or where one or more additional independent variables are taken into account. Methods are also discussed for measuring the influence of three or more independent variables operating jointly; but the increased number of observations necessary for such determinations restricts the field of usefulness of this type of analysis.

REFERENCES

Ezekiel, Mordecai. The determination of curvilinear regression "surfaces" in the presence of other variables, *Jour. Amer. Stat. Assoc.*, Vol. XXI, pp. 310–320. September, 1926.

CHAPTER 22

SUPPLEMENTARY METHODS FOR DETERMINING CURVILINEAR AND JOINT RELATIONS

Chapters 14, 16, and 21 have set forth means by which curvilinear regressions may be determined for functions either of the simpler type

$$X_1 = f_2(X_2) + f_3(X_3) + f_4(X_4) + \ldots + f_n(X_n)$$

or of the more complex joint type

$$X_1 = f(X_2, X_3, \ldots, X_n)$$

In each case the methods were purely empirical and depended on a combination of freehand smoothing with successive approximation to the best curve as the influence of other factors was gradually eliminated. In addition to the methods which have been presented, there are other techniques which have been suggested for considering even more complex relations. On the other hand, if a specific mathematical function is assumed, the curves may be determined by a more rigid process, using the principle of "least squares." This chapter presents some of these further methods, both for multiple curvilinear regressions of the simpler forms and for joint functional relations.

Determining net regression curves by mathematical functions. After the shape of the several net regression curves has been determined by the successive approximation method, a definite mathematical statement of the several functions may be obtained by an extension of the method presented on pages 221 and 222. The freehand curves would provide the basis for selecting functions which would fit the net shape of the regression curves fairly well, giving at least this empirical criterion as to what function to use. Applying this method to the egg problem mentioned on page 302, for example, the final curves indicate that a straight line is probably adequate to describe the net regression of price on X_3, that a cubic parabola would probably be required to describe the net regression on X_2, and that a second-degree parabola might be sufficient to fit the net regression on X_4. Accordingly, the equation

$$X_1 = a + b_2X_2 + b_{2'}(X_2^2) + b_{2''}(X_2^3) + b_3X_3 + b_4X_4 + b_{4'}(X_4^2)$$

might be fitted to the data. After the values for the seven constants were determined by the usual method for linear correlation, the closeness with which the several mathematical curves fitted the net regressions could be judged by computing the residuals from the new regression equation, and then plotting them as deviations from the several net regression curves, exactly as the residuals from the linear regressions were plotted in Figures 34, 35, and 36. If modifications in the fitted curves were found necessary, they could be determined by the approximation process again.

Where the original relations indicate a marked curvilinear relation, as in Figures 33 and 41, the mathematical curves may be fitted right at the start, just as described above, and these curves used as the basis for subsequent corrections by the approximation method. Whether determining net linear regression, as illustrated in the corn-yield problem, or determining net curvilinear regressions, as just suggested, will prove the most expeditious way of beginning the successive approximation method will depend on the circumstances in individual problems. Thus if one regression is known to be markedly curvilinear, while the others are substantially linear, taking that curvilinearity into account in the equation may bring the linear regressions for all the other variables much closer to their final form, and so reduce the number of steps necessary in the successive approximation determinations.

Determining the curves by least squares. The process of determining net regression curves by the use of a definite mathematical equation may be illustrated for the following data:

X_2	X_3	X_1	X_2	X_3	X_1
1	3	8	0	2	7
2	2	10	4	5	9
4	7	8	3	3	10
9	8	9	1	2	9
5	5	10	6	5	10
2	3	9	1	2	9
2	2	9	2	2	10
7	14	7	4	14	6
9	8	9	1	2	9
2	4	8	10	11	8

Preliminary examination by the graphic method indicates that the net regressions of X_1 on both X_2 and X_3 may be approximately represented by parabolas. Accordingly, a net regression curve may be assumed of the type

$$X_1 = a + b_2X_2 + b_2'(X_2^2) + b_3X_3 + b_3'(X_3^2)$$

The arithmetic required to determine the five constants can be reduced by "coding" the squared values. Let $U = X_-^2/10$, and $V = X_3^2/10$. The regression equation then may be written

$$X_1 = a + b_2X_2 + b_uU + b_3X_3 + b_vV$$

The normal equations to determine the four constants are next obtained in exactly the same manner as described in Chapter 12 for a multiple correlation involving four variables. The resulting normal equations are:

$$(\Sigma x_2^2)b_2 \quad + (\Sigma x_2u)b_u + (\Sigma x_2x_3)b_3 + (\Sigma x_2v)b_v = \Sigma x_1x_2$$

$$(\Sigma x_2u)b_2 \ + (\Sigma u^2)b_u \ \ + \Sigma(x_3u)b_3 \ + (\Sigma uv)b_v \ = \Sigma x_1u$$

$$(\Sigma x_2x_3)b_2 + (\Sigma ux_3b_u \ \ + \Sigma(x_3^2)b_3 \quad + (\Sigma x_3v)b_v = \Sigma x_1x_3$$

$$(\Sigma x_2v)b_2 \ + (\Sigma uv)b_u \ \ + \Sigma(x_3v)b_3 \ + (\Sigma v^2)b_v \ = \Sigma x_1v$$

Carrying out the required computations, the equations are found to be:

$$171.750b_2 + 170.625b_u + 165.000b_3 + 207.600b_v = - \ \ 2.50$$

$$170.625b_2 + 181.165b_u + 153.540b_3 + 192.316b_v = - \ \ 5.31$$

$$165.000b_2 + 153.540b_u + 295.200b_3 + 441.480b_v = - \ 50.80$$

$$207.600b_2 + 192.316b_u + 441.480b_3 + 696.072b_v = - \ 86.52$$

The $(\Sigma x_1^2) = 24.20$.

Solving these equations by the usual method and computing a from equation (41) by restating it

$$a_{1.2u3v} = M_1 - b_2M_2 - b_uM_u - b_3M_3 - b_vMv$$

we find the regression equation to be

$$X_1 = 9.411 + 1.2709\,X_2 - 0.7337\,U - 0.9957\,X_3 + 0.3309V$$

The net regressions of X_1 on X_2 and X_3 are now shown by the two parabolic equations:

$$X_1 = \ \ 5.596 + 1.2709\,X_2 - 0.7337\,X_2^2$$

and

$$X_1 = 12.515 - 0.9957\,X_3 + 0.03309\,X_3^2$$

The graph of these two curves is shown in Figure 71.

The (unadjusted) multiple correlation of X_1 with X_2, U, X_3 and V is 0.968. Since there are five constants in the regression equation, and 20 observations, this gives (by equation [47]) an adjusted correlation of 0.963. This is then the index of multiple correlation of X_1 with X_2 and X_3, according to the parabolic regressions. The standard error of estimate, similarly adjusted, is found to be 0.101.

It should be noted that where net curvilinear regressions are found by this method, the number of constants assumed in the regression equation is definitely known, and there can be no question as to the exact correction to apply to the computed correlation and standard error, or as to the probable significance of the observed correlation. On the other hand, the shape of the regression curve obtained is conditioned by the type of curve assumed; except where there is some logical basis for assuming a particular type of relation, the selection of the formula is still a purely empirical process. If the formula selected does not fit the data well, the resulting curves may fail to reveal the true relations.

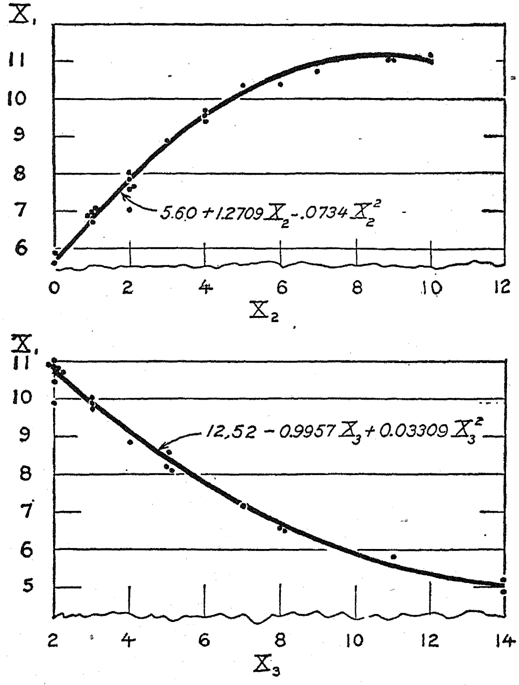


FIG. 71. Parabolic regression curves, fitted simultaneously, and net residuals.

Testing the fit of the curves graphically. The extent to which mathematical net regressions fail to fit the data adequately may be investigated, in any particular problem, by the same graphic methods set forth in Chapter 14. To make this check, after the regression equation is determined for the particular curves selected, estimated values of X_1 are calculated from the equation. The residual differences between X_1 and these estimated values are then computed. These residuals are then plotted as departures from the mathematical net regression curves, in the same manner that the residuals from linear regressions were plotted as departures from the linear regressions in Chapter 14. Carrying this out for the problem illustrated, we obtain these results:

X_2	X_3	X_1	X_1'	z	X_2	X_3	X_1	X_1'	z
1	3	8	7.9	0.1	0	2	7	6.7	0.3
2	2	10	9.8	0.2	4	5	9	9.2	−0.2
4	7	8	8.0	0.0	3	3	10	9.9	0.1
9	8	9	9.1	−0.1	1	2	9	8.8	0.2
5	5	10	9.8	0.2	6	5	10	10.2	−0.2
2	3	9	9.0	0.0	1	2	9	8.8	0.2
2	2	9	9.8	−0.8	2	2	10	9.8	0.2
7	14	7	7.2	−0.2	4	14	6	5.9	0.1
9	8	9	9.1	−0.1	1	2	9	8.8	0.2
2	4	8	8.2	−0.2	10	11	8	7.8	0.2

The residuals obtained above are then plotted as departures from the parabolic net regressions, as also shown in Figure 71. It is evident in this case that the parabolic regressions represent the relations quite well, with the departures in general evenly distributed on both sides of each curve throughout their length. Only in the curve for $X_1 = f_2(X_2)$ is there any indication of failure to obtain a good fit. Here most of the individual observations lie slightly above the curve for values of X_2 below 2. Above $X_2 = 8$ the individual observations do not agree with the downward turn of the parabola. Using a third-degree parabola for $f_2(X_2)$, which would mean adding a new term, $b_2''\ (X_2^3)$, to the regression equation, would produce a better fit for this function.

Where the graphic check on the adequacy of a mathematical net regression curve indicated that the functional relation was such that it could not be readily represented by a higher-order parabola or other simple mathematical expression, a freehand curve might be fitted to the residuals instead, and the final shape of the curves determined by successive approximations, just as described in Chapter 14. The determination of parabolic net regressions may thus be substituted for the determination of linear net regressions as the first step in the successive approximation method of obtaining net regression curves.

Any other type of mathematical function, the parameters of which can be expressed in the first degree, can be used to determine net regressions by the method of least squares. Besides still higher powers of X_2, such transformations as $10/X_2^2$, $100/X_2^3$, $\log X_2$, and $1/\log X_2$ may be employed as independent variables, either in place of the previous independent variables or as an addition to the simple statement of them.

Supplementary methods of determining the final shape of net regression curves. After the shapes of the several regression curves have been determined by the method of successive approximations, it is sometimes desirable to use the method of linear correlation to determine whether any further adjustment should be made in the *slope* of the several curves, to give the closest possible estimate of the X_1 values. There are two alternative ways of doing this, yielding slightly different types of corrections. The first and simplest method is to correlate the final residuals, z'''', with the values of the several independent variables. That is, a new multiple correlation is run to determine the regression equation

$$z'''' = a_{z.234} + b_{z2.34}X_2 + b_{z3.24}X_3 + b_{z4.23}X_4 \qquad (88)$$

If significant values are obtained for any of the b's, they indicate that the corresponding regression curve should be rotated counterclockwise, if the net regression coefficient is positive; or clockwise if the coefficient is negative. The final values of the several functions will then equal the readings for the curves as previously determined, plus the additional linear correction. That is, if the final curvilinear multiple regression equation is to be

$$X_1 = k + f_2(X_2) + f_3(X_3) + f_4(X_4)$$

the several terms will be:

$$\begin{aligned} k &= a''''_{1.234} + a_{z.234} \\ f_2(X_2) &= f_2'''(X_2) + b_{z2.34}X_2 \\ f_3(X_3) &= f_3'''(X_3) + b_{z3.24}X_3 \\ f_4(X_4) &= f_4'''(X_4) + b_{z4.23}X_4 \end{aligned}$$

Since the intercorrelations between X_2, X_3, and X_4 have already been computed in determining the original linear net regressions, much of the work required in determining the constants for equation (88) has already been performed, and the additional computation involved is not very heavy.

A somewhat different type of correction is obtained by determining the regression equation

$$X_1 = a_{1.2'3'4'} + b_{12'.3'4'}[f_2'''(X_2)] + b_{13'.2'4'}[f_3'''(X_3)] \\ + b_{14'.2'3'}[f_4'''(X_4)] \qquad (89)$$

To compute the new constants required in equation (89), the functional readings corresponding to the independent variables are correlated with the original values of the dependent variable. Thus, if the values in Table 64, page 246, had been obtained from the final curves determined by the successive approximation process, the values read from the curves, shown in the fourth, fifth, and sixth columns, would have been substituted for the original independent variables in running the multiple correlation with X_1. If X_2', X_3', etc., are used to represent these transformed values, the data to be correlated for the first four sets of observations would be:

X_2'	X_3'	X_4'	X_1	X_2'	X_3'	X_4'	X_1
7.4	11.7	12.3	24.5	8.4	12.2	12.2	27.9
7.9	13.0	11.8	33.7	8.8	9.9	12.2	27.5

If the net regression coefficients come out 1.0, in equation (89), that indicates that no change need be made in the curves. If any b comes out other than unity, however, the values read from the corresponding curve should be adjusted as indicated by the regression results. The adjustment may be worked out as follows:

In the same way that $f_2'''(X_2)$ was used to indicate the values read from the final set of approximation curves, let $f_2'''(x_2)$ represent the deviations of those readings for each variable from the average of all the readings for the particular variable. That is, for each observation

$$f_2'''(x_2) = f_2'''(X_2) - M_{f'''(X_2)}$$

The regression equation (89) may then be restated

$$x_1 = b_{12'.3'4'}[f_2'''(x_2)] + b_{13'.2'4'}[f_3'''(x_3)] + b_{14'.2'3'}[f_4'''(x_4)]$$

and the corrected functions will be as follows:

$$f_2(x_2) = b_{12'.3'4'}\,[f_2'''(x_2)]$$

$$f_3(x_3) = b_{13'.2'4'}\,[f_3'''(x_3)]$$

$$f_4(x_4) = b_{14'.2'3'}\,[f_4'''(x_4)]$$

The difference between the two types of corrections is illustrated in Figure 72. Here the final curve for $f_4(X_4)$, from the corn-yield problem, has been plotted, and in addition it is shown as if a correction $+ 0.5x_4$ had been worked out by the first method, equation (88), or a correction $1.5[F_4(x_4)]$ had been determined, by the second method, equation (89). It is evident that the first correction rotates the curve, so as to make its upward slope greater throughout, and its downward slope less; whereas the second correction merely expands the curve, making all the high values higher and all the low values lower, no matter where they fall with respect to X_4. This is typical of the effect of these two types of corrections when applied to a curve of the type shown here. For curves which do not depart so far from a straight line and which either rise or fall through their entire length, the difference between the two types of correction is less marked, as may readily be determined by experiment. For a straight line the correction given by the two methods will tend to be identical.

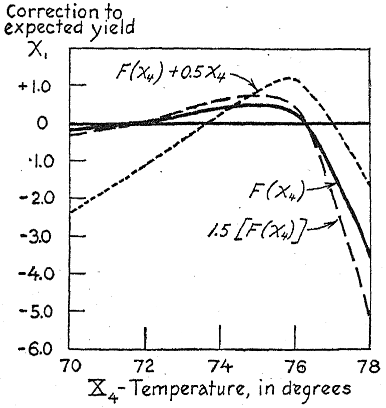


FIG. 72. Two types of corrections to net regression curves.

The Bruce adjustment. Besides the two methods shown of adjusting the final curves by linear correlation (the method of least squares), there is a somewhat different adjustment of the final readings, termed the Bruce method after its originator.[1] This method consists essentially of (1) constructing a dot chart showing the relation between the original values of X_1 and the values, X_1''', estimated from the final set of curves (even after corrections such as those just mentioned have been applied); (2) drawing in a curve showing average values of X_1 for corresponding values of X_1'''; and (3) using this

[1] See list of references at end of this chapter.

curve as the basis for making the final estimate. This method thus consists in finding the function θ for the equation.

$$X_1'''' = \theta(X_1''') \qquad (90)$$

$$X_1'''' = \theta[a + f_2'''(X_2) + f_3'''(X_3) + f_4'''(X_4)] \qquad (91)$$

The function θ corresponds to the curve just described.

The curve for θ can be determined only after all the other f's are worked out. The average value of X_1 is determined for each group of estimated values, X''', treating X_1''' as if it were a single independent variable. Plotting the average values of both variables against each other, just as in Figure 23, a curve may be fitted freehand if it is indicated, to show the change in X_1 with changes in X_1'''. This curve is then the function θ for equation (90).

This function may then be used to work out new estimates, X_1'''', which should have still higher correlation with X_1 than the previous estimates.

The Bruce adjustment is likely to prove of value in certain types of joint functions. Thus in the egg-price problem, it might be that when all the values of several factors, each of which by itself tended to lower the price, occurred in combination, the resulting price would be, on the average, even lower than the sum of the effects of each of the variables would indicate. On the other hand, it might be that when values of several factors, each of which would raise the price, occurred together, the price would not go quite as high as the sum of the probable effects of the several factors would indicate. The Bruce adjustment thus makes it possible to determine one type of joint relation without the considerable extra work described in Chapter 21 for determining joint functions in general.

Determining joint relations by contours. The method for determining joint relations presented in the last chapter is essentially one of subclassification and then two-way smoothing of the resulting averages, by successive smoothing for each of the two (or more) independent variables. A somewhat different method has been worked out by which a three-variable surface may be smoothed directly in both independent dimensions at the same time. The Waugh method is based on determining contours directly, instead of indirectly as described in the last chapter. In using this method, the averages of subgroups (either of original observations, as in Table 81, or of residuals) are plotted directly on a two-variable diagram, with one independent variable as ordinate and the other independent variable as abscissa, and with the group averages used for the two independent variables. The average of the dependent variable (or residual) is then written in next to the dot which designates the subgroup. Figure 73 shows such a chart for the averages of Table 81. The next step is to connect averages of equal values by a continuous line, or, if none are the same, to run in contour lines which will enclose averages within the same limits. Thus the lines on the chart have been drawn so as to separate off the groups with X_1 under 0.300, between 0.300 and 0.400, from 0.400 to 0.500, etc. Once the shape and direction of these contours are determined, they may then be redrawn so as to keep a similar shape or a continuously changing shape, and an even or a regularly changing interval, across the whole surface. It is evident that Figure 73 is quite similar to Figure 69, determined by the other method.

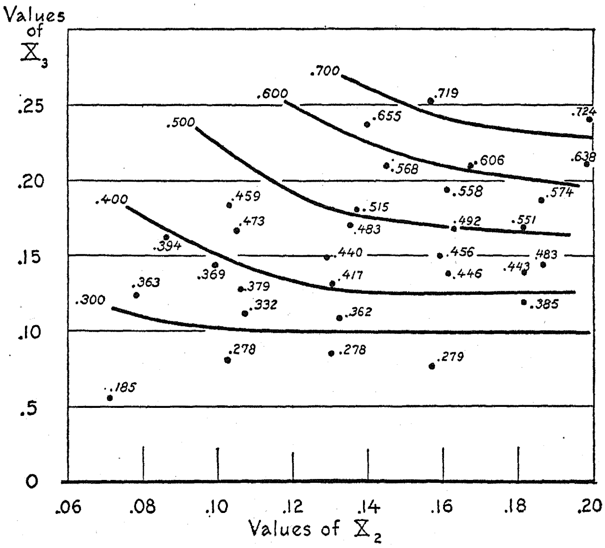


FIG. 73. Average values of X_1 for various combinations of X_2 and X_3, and contours fitted directly to the data.

Where the correlation is high, so that the individual observations define the regression surface rather closely, the Waugh method may be used directly with the individual observations, plotting each observation in the same way that the group averages were plotted in Figure 73. The following data illustrate this use of the method.

The data from Table 84 are plotted in Figure 74, with the yield adjusted for trend used as the dependent factor. Drawing in contours so as to separate years of similar yields, we find that a very peculiar type of surface is indicated—one that changes elevation very rapidly between the combination of high early rainfall and low late rainfall, and high early rainfall and high late rainfall. When these results are used to forecast the yield in 1928 (which year, it will be noted, was not plotted or used in determining the contours) a yield of about 175 bushels is indicated. This is only in fair agreement with the final yield of 219 bushels, determined several months after the climatic data were available to give the forecast stated.

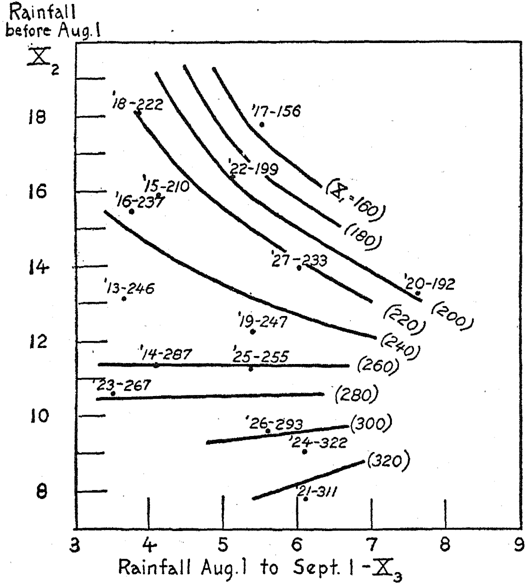


FIG. 74. Yield of potatoes for years of specified rainfall before August 1 and after August 1, and contours fitted directly to the data.

Reading off the estimated values for each year shown, the estimated adjusted yields as shown in the next to the last column of Table 84 are obtained. The standard deviation of the residuals, shown in the next column, is 10.6 bushels, whereas the σ of the deviations from trend is 63.0. If five constants are assumed to be necessary to represent the surface mathematically, the standard error of estimate would be 13.0 bushels and the index of correlation for the surface indicated by the contours would be 0.98. If it is assumed that the trend line could be fairly accurately projected, the standard error of estimate indicates that an error as great as that in 1928 would be likely to occur only very rarely.[2] The fact of high correlation and of low standard error could be judged directly from the closeness with which the contours fit the individual observations, in just the same way that closeness of the observations to the regression line indicates high correlation in the case of simple correlation.

TABLE 84

WEATHER CONDITIONS AND YIELD OF POTATOES IN MAINE

Year	Rainfall to August 1 (July doubled)	Rainfall August 1 to September 15	Yield	Adjustment for trend *	Yield adjusted for trend	Estimated yield	Residual
	Inches X_2	*Inches* X_3	*Bushels* X_1	*Bushels*	*Bushels* X_1	*Bushels* $f(X_2, X_3)$	z
1913	13.17	3.66	220	+26	246	248	− 2
1914	11.33	4.08	260	+27	287	260	27
1915	15.96	4.12	179	+31	210	229	−19
1916	15.46	3.77	204	+33	237	236	1
1917	17.77	5.53	125	+31	156	155	1
1918	18.09	3.87	200	+22	222	220	2
1919	12.25	5.41	230	+17	247	248	− 1
1920	13.29	7.62	177	+15	192	196	− 4
1921	7.82	6.11	298	+13	311	323	−12
1922	16.40	5.12	187	+12	199	197	2
1923	10.61	3.51	258	+ 9	267	278	−11
1924	9.10	6.13	315	+ 7	322	308	14
1925	11.30	5.38	250	+ 5	255	262	− 7
1926	9.60	5.60	290	+ 3	293	297	− 4
1927	13.98	6.02	232	+ 1	233	226	7
1928	15.45	6.45	220	− 1	219		

* Simultaneously determined while allowing for trend. See F. V. Waugh, Methods of forecasting New England potato yields, U. S. Department of Agriculture, Bureau of Agricultural Economics, Mimeographed Report, February, 1929.

Determining joint functions by definite mathematical functions. In exactly the same way that definite equations can be determined by the method of least squares to represent curvilinear net regressions, certain types of joint functional surfaces can be represented

[2] If the standard error of this particular estimate could be calculated along the lines indicated in Chapter 19, the error might not appear so unusual.

by definite equations. The simplest type is that shown by the haystack volume problem in Chapter 21, where the regression of X_1 on X_3 is substantially linear for any given value of X_2 but where the slope of the regression $b_{13.2}$ changes as the values of X_2 change. If it is assumed that the slope of $b_{13.2}$ changes at a constant rate with changes in X_2, this assumption may be expressed in the relation

$$X_1 = a + b(c + dX_2)X_3$$

Multiplied out, it becomes

$$X_1 = a + bcX_3 + bdX_2X_3$$

which may be stated

$$X_1 = a + eX_3 + g(X_2X_3) \tag{92}$$

The values of a, e, and g may then be determined by the usual methods of linear multiple correlation, with X_3 and the values of the product (X_2X_3) used as the independent factors.

If it is assumed that X_1 varies with X_2, other than through its influence on $b_{13.2}$, an additional term may be added to the equation, making it

$$X_1 = a + eX_3 + g(X_2X_3) + hX_2 \tag{93}$$

Determining the values of the four constants of equation (93) from the haystack data, and working out estimated values of X_1 for specific combinations of values of X_2 and X_3, we shall arrive at the same joint functional surface as was determined by the graphic method presented in Chapter 21.

We may extend the same method to n independent variables, assuming similar linear net regressions for X_1 on each independent factor, with the other independent factors constant at any given values and with these net regressions changing their slope progressively and uniformly as the other independent factors change. For three independent factors (four-dimensional space) the regression equation would be

$$X_1 = a + b_2X_2 + b_3X_3 + b_4X_4 + c_2(X_2X_3)$$
$$+ c_3(X_2X_4) + c_4(X_3X_4)$$

Determination of the seven constants would thus make possible a definite mathematical representation of a very complex set of relationships.

If it is assumed (1) that the regression of X_2 on X_1, for any given value of X_3, is a curve and (2) that the slope of this curve changes *at a changing rate* as X_3 changes, this assumption may be stated

$$X_1 = a + f\,[a + \theta(X_3)]X_2$$

This equation may be approximately represented by the following form:

$$X_1 = a + f_2(X_2) + f_{2,\,3}(X_2X_3) + f_3(X_3) \qquad (94)$$

Using X_2, X_3, and the product (X_2X_3) as the independent factors, we may determine the shape of the three functions by any of the methods presented previously. Then working out estimated values of X_1 for various combinations of X_2 and X_3, we can determine very warped curvilinear surfaces for $f(X_2,\,X_3)$. This last method is extremely flexible, and can be used to determine a wide variety of joint functional relations. It, too, may be generalized for n variables, with increasing numbers of observations. For three independent variables it would be

$$X_1 = a + f_2(X_2) + f_3(X_3) + f_4(X_4) + f_{2,3}(X_2X_3) + f_{2,4}(X_2X_4) + f_{3,4}(X_3X_4)$$

Although these methods do not reduce greatly the number of observations required to determine joint functions, they do make it possible to apply the systematic procedure developed in Chapter 14 and to judge more accurately the number of constants represented in the regression surface; and they enable the methods of Chapters 18 and 19 to be applied in judging the reliability of the conclusions.

The Court method. The Waugh method is essentially a way of simplifying the smoothing of the surface, while still leaving it primarily a graphic freehand process. Another method, developed by Andrew Court, reduces the determination of joint functions to a more definite process, similar to the determination of the usual regression curves. This method depends upon a mathematical rotation of the surface of cubes such as those shown in Figures 62 to 64, so that, instead of averaging the values only when viewed with respect to the rectangular axes X_2 and X_3, we may also average them with respect to axes cutting across the surface at an angle. Though similar to the mathematical method just described, this method is applicable to a somewhat different type of surface.

The characteristic feature of the method is the use of composite functions which represent two or more independent variables. Thus the regression surface in Figure 70, for apples of one size, might be expressed by the equation

$$X_1 = f_2(X_2) + f_3(X_3) + f_{2+3}(X_2 + X_3) \tag{95}$$

The effect of the introduction of the new composite element $(X_2 + X_3)$ may be explained by working out what the values of this composite variable will be for various combinations of X_2 and X_3. The following statement shows this in detail.

Value of Composite Variable $(X_2 + X_3)$, for Various Values of X_2 and X_3

X_3 values	X_2 values					
	0	20	40	60	80	100
0	0	20	40	60	80	100
20	20	40	60	80	100	120
40	40	60	80	100	120	140
60	60	80	100	120	140	160
80	80	100	120	140	160	180
100	100	120	140	160	180	200

It will be seen that the composite values $(X_2 + X_3)$ run diagonally across the surface. Thus the value 100 occurs with $X_3 = 100, X_2 = 0$; with $X_2 = 50, X_3 = 50$; and with $X_2 = 0, X_3 = 100$. If the surface shown in Figure 70 were to be described by equation (95), the curve for $f_{2+3}(X_2 + X_3)$ would rise gradually from 0 to 100, then rise more and more sharply as it approached 200.

One advantage of the Court method is that it makes it possible to estimate with much greater accuracy the number of constants represented by the surface. Thus in Figure 70 each curve might reasonably be represented by a second-degree parabola, so seven constants may be assumed for the entire relation in equation (95). If desired, we could write the equation in terms of parabolas, as follows:

$$\begin{aligned} X_1 = a + b_2X_2 + b_2'(X_2^2) + b_3X_3 + b_3'(X_3^2) \\ + b_4(X_2 + X_3) + b_4'(X_2 + X_3)^2 \end{aligned}$$

Stated in this form, the shape of the surface could be determined by a least-squares solution, giving exactly determinable shapes for each of the three functions, and a definite measure of the reliability of the results. However, unless the mathematical curves happened to be about right to represent the real relations, the final functions might not express the relationship so closely as would the freehand curves, determined by successive approximations.

Where two independent variables which are to be considered jointly are not of the same degree of variability, a 45-degree rotation of the surface, such as that shown in the tabular statement of $(X_2 + X_3)$, could still be secured by making the composite variable equal to $\frac{X_2}{\sigma_2} + \frac{X_3}{\sigma_3}$. If a negative rotation were desired, that could be obtained by using the value $\frac{X_2}{\sigma_2} - \frac{X_3}{\sigma_3}$. Further, if it were desired to rotate the surface either less or more than 45 degrees, that could be done by dividing one variable or the other by a suitable constant. Thus the form $\frac{X_2}{\sigma_2} + 2\frac{X_3}{\sigma_3}$ would rotate the surface about 67 degrees.

The general statement for the Court solution, for a two-variable joint function, is:

$$X_1 = f_2(X_2) + f_3(X_3) + f_{2+3}\left(\frac{X_2}{a} + \frac{X_3}{b}\right) + f_{2-3}\left(\frac{X_2}{a} - \frac{X_3}{c}\right) \quad (96)$$

Even a more complex form than equation (96) could be employed, by using combination functions of several different degrees of rotation in the same equation. Using such a combination with simple parabolas for each function, Court has successfully fitted the regression surface shown in Figure 63, illustrating the flexibility of the method. It is evident, however, that much judgment is necessary in selecting the way the combination variable or variables are to be stated in equation (96), both with respect to whether the rotation is to be positive, or negative, or both, and the extent of the rotation to be used. In the apple-price problem, it was known that the statement of equation (95) would fit quite well, because of the prior knowledge of the relations expressed in Figure 70. Where such information as to the shape of the function is not known *a priori,* considerable testing of different methods of statement and examination of group averages and profile charts like Figure 65 would be necessary to decide upon a form of statement which would yield adequate results.

The Court method may be extended to n-dimension joint functions, and it has very great flexibility for this purpose. The number of possible combination variables becomes increasingly great as the number of variables increases, however, so that stable results cannot be secured by this method either, unless a sufficiently large number of observations is available to define the relations for each of the subclasses which are obtained by successive sorting on each independent variable. Thus using only 45-degree rotations, we should find the full Court equation for three independent variables to be

$$\left.\begin{aligned} X_1 &= f_2(X_2) + f_3(X_3) + f_4(X_4) \\ &\quad + f_{2+3+4}\left(\frac{X_2}{\sigma_2} + \frac{X_3}{\sigma_3} + \frac{X_4}{\sigma_4}\right) + f_{2+3-4}\left(\frac{X_2}{\sigma_2} + \frac{X_3}{\sigma_3} - \frac{X_4}{\sigma_4}\right) \\ &\quad + f_{2-3+4}\left(\frac{X_2}{\sigma_2} - \frac{X_3}{\sigma_3} + \frac{X_4}{\sigma_4}\right) + f_{2-3-4}\left(\frac{X_2}{\sigma_2} - \frac{X_3}{\sigma_3} - \frac{X_4}{\sigma_4}\right) \end{aligned}\right\} \quad (97)$$

Even if each function were represented by only two constants, equation (97) would involve fourteen constants. The similar forms for four and five independent variables become increasingly complex It is probable that this method can be used only occasionally, where a very large number of observations can be obtained. For such problems, however—as for the apple-price example, where 190 observations were available—the use of equations such as (96) or (97) might reduce somewhat the factor of individual judgment and enable the researcher to determine joint relations in n-dimensional space with more facility than by graphic methods which essentially involve considering individual dimensions in succession, or at most two at the same time.

Measures of correlation for mathematically determined regressions. Where the curvilinear net regressions or regression surfaces have been determined by the use of mathematical functions such as those indicated in equations (55) and (56), then the several measures of closeness of fit can be obtained from the computations employed in determining the values of the several b's by the usual linear multiple correlation methods. For example, if equation (56) involving cubic parabolas for each variable has been employed, the regression equation is

$$\begin{aligned} X_1 = a + b_2X_2 + b_{2'}(X_2)^2 + b_{2''}(X_2)^3 \\ + b_3X_3 + b_{3'}(X_3)^2 + b_{3''}(X_3)^3 + \text{(etc.)} \end{aligned}$$

In that case the *coefficient* of multiple correlation with respect to the independent variables X_2, X_2^2, X_2^3, X_3, X_3^2, X_3^3, etc., becomes the *index* of multiple correlation with respect to the variables X_2, X_3, etc. The necessary adjustments because of the number of constants represented in the regression equation, as indicated by equation (67), still have to be made, of course. With the regression curves determined mathematically, there is no question of the value of m to be used. For the equation shown above, with only two independent variables, X_2 and X_3, m is 7. With this limitation, the index of multiple correlation for mathematically determined regressions may be defined by the equation

$$R_{1.2,\,2^2,\,2^3,\,3,\,3^2,\,3^3,\,\ldots\,n,\,n^2,\,n^3} = \text{P}_{1.23\,\ldots\,n} \tag{98}$$

Indexes of partial correlation could be worked out by parallel recombination of the elements involved in determining the constants of equation (56), but the steps necessary would become exceedingly complicated, and therefore are not set forth here.

The standard error of estimate in using a mathematically determined curvilinear regression equation is the same as the standard error of the multiple correlation results, with the appropriate correction for the number of constants. When the index of multiple correlation has been determined, with the proper adjustments, the standard error of estimate may readily be obtained by the formula

$$\bar{S}^2_{1.f(23\,\ldots\,n)} = \sigma_1^2\,(1 - \bar{\text{P}}^2_{1.23\,\ldots\,n})\left(\frac{n}{n-1}\right) \tag{99}$$

This operation is necessarily identical with that employed in computing the standard error in linear multiple correlation, using the adjusted coefficient of multiple correlation.

Differential regressions. The relation of rainfall or temperature to a growing crop can be measured more effectively if the distribution of rainfall or temperature through the entire season is considered, instead of breaking up the records into a series of arbitrary periods as in various illustrations to this point. Many years ago R. A. Fisher developed a method of fitting a continuous differential regression curve to the rainfall through the season, showing the changing effect of each inch of rainfall at different times in the growing period of the plant. (Note discussion on page 419 and reference 18 at the end of Chapter 23.) This technique has recently been extended to make it possible to obtain such a differential regression curve for one independent variable, such as rainfall distribution, while simultaneously making allowance for the effect of other independent variables, such as evaporation, and determining the differential regression on the second independent factor. These methods appear to be particularly valuable in agronomic and meteorological problems, but they may also be found of value in other applications. They are fully presented and discussed in the paper by Davis and Pallesen, listed at the end of this chapter.

Summary. Simple curvilinear regressions, determined by the successive approximation process, may be subjected to a final correction by mathematical means; or mathematical curves for each function may be fitted simultaneously; or certain types of joint relations may be represented by the use of a composite function, θ, which may be determined rather readily.

The smoothing of two-variable joint functions may be facilitated by the use of contours (the Waugh method) drawn freehand either from the subgroup averages, or, in the case of high correlation, from the original observations. Other methods employ combination variables composed of simple linear functions of two or more independent variables to rotate or warp the joint surface and so determine its shape other than at right angles to the axes of the independent variables. By using several such combination variables, and determining regression curves for them by successive approximations, we may represent very complex joint functional surfaces quite closely. These methods may be extended to joint functions of n variables, but they become increasingly complex and require an increasingly large number of observations. Even so, these methods reduce somewhat the element of human judgment involved in the determination of joint functions and simplify the steps involved to more nearly a routine process which can be expected to give identical results from the same data in the hands of different investigators.

It is possible to obtain standard errors of estimate and indexes of multiple correlation, which serve the same purpose for mathematically determined curvilinear multiple regression equations that the comparable coefficients serve for linear multiple regressions. Owing to the larger number of constants to be determined, it is even more important than it is with linear multiple correlation to adjust the several measures with respect to the number of observations and number of constants involved if we are to obtain unbiased estimates of the corresponding values in the universe from which the sample was drawn.

REFERENCES

Bruce, Donald. On possible modifications in the Ezekiel method of curvilinear multiple correlation. Typewritten manuscript, filed in the Library, Bureau of Agricultural Economics, U. S. Dept. of Agr., 19 pp.

Waugh, Frederick V. The use of isotropic lines in determining regression surfaces. *Jour. Amer. Stat. Assoc.*, p. 144, June, 1929.

Court, Andrew T. Measuring joint causation. *Jour. Amer. Stat. Assoc.*, Vol. XXV, pp. 245–254, September, 1930.

Davis, Floyd E., and Pallesen, J. E. Effect of the amount and distribution of rainfall and evaporation during the growing season on yields of corn and spring wheat. *Jour. Agr. Rersearch,* Vol. 6, No. 1, pp. 1–24, Washington. Jan. 1, 1940.

CHAPTER 23

TYPES OF PROBLEMS TO WHICH CORRELATION ANALYSIS HAS BEEN APPLIED

In the preceding chapters many different practical problems have been used to illustrate the kinds of correlation analysis and the actual steps in working out the results. It may now be worth while to turn attention to specific research problems to which these methods have been applied in the past. This will indicate the type of logical analysis which must be made before the statistical technique can be applied and show something of the kind of conclusions which may be reached by the use of these techniques.

Land values. One of the first comprehensive studies involving extensive correlation analysis was a study of land values by Haas (1).[1] In this study the sales prices of a number of different farms were obtained, and also supplementary facts about the farms, such as distance from town, value of buildings, proportion of crop land, fertility of the soil, and type of road on which the farm fronted. Changes in land values over the period were first eliminated, and the adjusted acre prices related to the other factors by linear correlation. Sortings of the residual values were used to determine the regressions for some of the less important factors. It was found that the differences in value per acre had a multiple correlation of $R = 0.81$ with the factors mentioned and that acre values could be estimated from the independent factors with a standard error of $19 per acre. As the assessor's valuations of these same farms showed a much larger error, as compared with the actual sales values, it was suggested that the impartial regression equation be substituted for the less reliable human judgment in assessing individual farms for taxation purposes.

In a later study of the same type (2) the value of the farm dwelling and the value of the barns were considered as separate variables, and curvilinear regressions were determined. It was found in this study that the contribution of the farm dwelling to the farm value was a joint function of the value of the dwelling and the size of the farm, an expensive dwelling adding more to the value of a large farm than

[1] The numbers in parentheses refer to references at the end of this chapter.

to the vaiue of a small one. Road type was one of the factors considered. Three classes of roads were used, and the method explained in Chapter 17 was employed to determine the net difference in farm value per acre with differences in the type of road. Preliminary work in this study, with the farm value stated on a per-acre basis, gave a linear correlation of $R = 0.98$. It was discovered, however, that this high correlation was due almost entirely to the presence of a few very small farms, which showed values of farms per acre and values of buildings per acre both running into the thousands of dollars. When these farms were excluded, the linear multiple correlation dropped to $R = 0.64$, indicating the spurious correlation obtained by dividing by the common factor, number of acres. In the final correlation, with curvilinear relations and joint functions being used, a multiple correlation of $\text{P} = 0.77$ was obtained. As 368 observations were used, more complex methods could be employed for this analysis than would be feasible in most cases.

Physical relations between input and output. Another type of problem to which multiple correlation has been applied is determining the physical relation between the number of input (or cost) elements applied in some production process and the resulting output or yield. This problem is particularly important in agricultural research, where many of the combinations of conditions which occur in practical farming cannot be reproduced or studied under experimental conditions, and where the number of variables is so great as to make the use of fully controlled experiments both lengthy and costly.

In one of these studies (3) the gain in weight of beef steers on feed was related to the quantities of corn, hay, and high-protein feeds fed per day, to the number of days on feed, and to the initial weight of the animals. Curvilinear regressions were determined for all factors, protein being the only one to indicate a true linear relation. The curves showed marked diminishing returns per unit of feed as added amounts of corn or of hay were fed per day. The younger the animals, and the shorter the time they were on feed, the smaller the amount of feed necessary to produce a given amount of gain. There were 67 observations for this study, each representing a different bunch of cattle. The multiple correlation was $\text{P} = 0.78$.[2]

[2] This study is of particular interest to the author, as it was in studying this particular problem that the successive approximation method of determining net regression curves was first worked out, and in this problem that regression curves were first determined by this method while holding the influence of other variables constant.

The same analysis of physical relations has been applied to the production of milk by dairy cows. In most of these studies (4, 5, 6, 7) the feeds used and milk produced have been worked out on a herd-average basis, the record for each herd constituting one observation. In one study, however, the records were available by individual cows, and the conclusions secured from those records agreed quite well with those obtained from the other analyses (8).

The total quantity of digestible nutrients in the feed, the proportion of protein in the feed, the proportion of butterfat in the milk, and the proportion of the herd freshening in the fall, all have been found to be important variables influencing the production of milk. Variables of less importance, but of some effect in some localities, have been the proportion of nutrients derived from silage, the proportion of feed fed while on pasture during the summer season, the age and weight of the cows, and their quality as indicated by their value per head. The breed of cow was considered in several studies, but was found to have only a negligible influence on production after other factors were allowed for. In spite of the fact that no measures have been found satisfactory for the nutrients the cows obtain from pastures, multiple correlations ranging up to 0.90 have been obtained in these studies, indicating how much the average production of a herd is dependent upon the physical conditions and practices.

Similar correlation studies of the influence of physical input upon output have been made in the case of potatoes (9, 10), cotton (11), and other crops. In the study of potatoes, yield was found to vary with the amount of seed used, the quantity of manure and fertilizer applied, and the depth of plowing. The regression for the latter factor was particularly interesting, in that it was convex from above, indicating that maximum yields were secured with a certain depth of plowing and that plowing either deeper or shallower decreased the yield.

In the study of cotton, the quantities of mixed fertilizer and of nitrate of soda used were considered as separate variables; the quantity of calcium arsenate applied was considered, and also the fertility of the land as indicated by the yield of other crops, notably corn. The results in this problem raise two interesting points which illustrate some of the logical problems which come up in correlation analyses. The arsenate influences the yield through killing the boll weevil. In the year studied there was a heavy weevil damage on untreated fields, and the applications of poison increased the yield very materially. But these results indicate nothing of how much influence poison would have on yield in years when weevil damage was lighter. It would be necessary to repeat the study over several years with varying weevil damage, and then relate the differences in the effectiveness of poison to differences in the climatic factors which affect the weevil infestation, before it would be possible to judge in any particular year whether or not it would pay to use poison that year—and the prices both of poison and of cotton would enter into the final consideration.

The inclusion of yields of other crops as a factor in the multiple correlation raises another interesting logical point. The net regressions show that, with other factors remaining the same, farms with high yields of other crops also tend to have high yields of cotton. This might be interpreted as indicating that farms with high yields of other crops also have high native fertility, and that in eliminating this factor the results as to the effect of using the other factors have been made more dependable. But it may be that the high yields of other crops are partly due to high fertilization, either during the same year or in previous years. In eliminating the increased cotton yield associated with high yields of other crops, then, we might really be eliminating part of the result of high fertilization of cotton. The simultaneous determination of the relations by the method of multiple correlation tends to allow for these inter-relations, if they exist, but whether it does so completely in any given case may still be queried. In the particular case cited, collection of additional information as to fertilizer applied on each field in previous years would give a more positive answer to the question of how much was native fertility and how much was the result of previous treatment, and so give a real solution to the logical dilemma.

Weather conditions and crop yields. Another type of complex physical relationships which has been satisfactorily treated by multiple correlation is the relation of weather factors to crop yields. The yield problem of Chapter 14, taken from the work of Misner (12), and the potato-yield problem of Chapter 21, from the work of Waugh (13), have already been discussed at length. Other problems of the same sort were early studies of the relation of rainfall in July and August to the size of the Illinois corn crop (14); studies of the influence of rainfall and temperature during the growing season on cotton yields (15); studies of the influence of precipitation, temperature, and relative humidity on spring-wheat yields (16); and many others which might be mentioned. One interesting study related the weather during the winter to the yield of cotton in the South. This study showed that extreme cold tended to exterminate the boll weevil and so increase the yield of cotton (17). In spite of the fact that the correlation was practically perfect during the brief period of six years for which the study was made, the author did not believe that he had explained all the causes of variations in the yield of cotton, and modestly refrained from concluding that he had a perfect forecaster of cotton yields. This was fortunate, as, after giving an excellent forecast of yield for one year, cotton yields the second year were diametrically opposite to the expected yields—with a departure of many times the standard error of the previous years. This case is interesting as indicating the limited meaning of computed standard errors in time series, and as further indicating that a result which is not sensible logically cannot be trusted as the sole basis for forecasting, no matter how high the correlation. As it was this same investigator who had previously worked out the influence of weather conditions during the growing season on the yield of cotton in individual states, he was forewarned as to the significance of his apparently perfect forecaster, and he was duly cautious in interpreting its meaning. The result was certainly important, however, as indicating that weather factors prior to planting time may be related to subsequent yield.

A somewhat different approach to the crop-yield problem has been taken by R. A. Fisher (18). In studying wheat yields at Rothamsted, he pointed out that it really made little difference to the growth of the crop whether a given rain occurred on April 30 or May 1; yet if the rainfall were studied by monthly totals the assumed effect might be quite different. Furthermore, if weekly periods were considered for all the different factors, the number of different constants in the regression equation might readily exceed the number of observations. He therefore devised a method of determining the differential relation of rainfall and yield, so as to determine the rate of change in yield with the rate of change in rainfall at any season of the year. The differential equation required a sufficiently small number of constants so that it could be accurately determined from the observations at hand. The resulting smooth curve for the change in yield with changes in rainfall showed that the maximum effect was in fall and in spring, with less effect during the winter. With rainfall through the year the only weather element considered, correlations ranging from 0.32 to 0.63 were obtained for various test plots. Although this method does not take into account joint effects of climate at different seasons (as did the potato-yield problem used in Chapter 22), and the method of analysis is more complicated mathematically than any of those presented in this book, the suggestion of determining differential regression equations may open up new possibilities of accurate and complete analysis. (See page 413.)

Relation of physical characteristics of samples to chemical characteristics. A quite different application of correlation analysis has been in determining the extent to which the chemical properties of a given sample were related to, or could be estimated from, observable physical properties. The estimation of the protein content of wheat from the proportion of vitreous kernels, used as an illustration in Chapter 6, is taken from a much more comprehensive series of studies (19) in which the weight, the percentage of vitreous kernels, and the region of the country from which the wheat came were all found to have significant influences. In addition, it was found that the relations changed slightly from year to year, so that further work remains to be done to determine the influence of differences in climatic factors on the relation between physical and chemical properties.

A somewhat different study, but also within the same general field as the last one mentioned, related the volume of bread a given quantity of flour would produce to the gluten content of the wheat and of the flour (20). Correlation was also used to determine the extent to which the digestible composition of different cuts of meat could be judged from the visible proportion of fat (21). Studies such as these illustrate how statistical methods may be used to generalize from the results of many tests, even where the tests themselves were carried out under the carefully controlled conditions of exact scientific experiment.

A somewhat different application cf statistical methods to the interpretation of data secured from exact scientific measurements is in the astronomical problem of the relation of the brightness, intensity, and distance of the stars. Careful investigations in this field (22) have leaned heavily upon correlation analysis for their final conclusions.

These last two types of problems deal with purely physical relations, which remain the same, or at worst change only gradually, over a series of years. The idea of a statistical universe which is being sampled may therefore have some application, though it is sometimes a limited one. But in the next type of problem, though the universe is stable at any one time, it may change radically from year to year, so that conclusions for one year may not be at all applicable to those of succeeding years.

Relation of farm organization to farm income. The question of what organization will produce the best returns for the farms in a given locality is one that has been given extensive statistical investigation, by correlation means and otherwise. In particular, studies of farm income in Pennsylvania (23), Iowa (24), and Virginia (25), to mention specific cases, have made extensive use of multiple correlation analysis. Such factors have been considered as the size of the farm, the acreage in each of the principal crops, the size of the important livestock enterprises, the efficiency of crop production and livestock production, and the capital invested. In general it has been found that about half the variation in earnings from farm to farm in the same year can be explained by such objective measures of their organization and management as just mentioned. The multiple correlations with income range up to a maximum of about 0.75 to 0.80. In addition, it has been found that the size of the dominant enterprise and the efficiency with which it is conducted are usually the most important factors affecting returns. Thus on Iowa hog farms (24) the yield of corn, the number of brood sows, and the efficiency of hog production are dominant factors; on Virginia tobacco farms (25), the acreage in tobacco, the yield of tobacco per acre, and the quality of tobacco; and on Pennsylvania dairy farms (23), the number of dairy cows and the efficiency of the dairy enterprise.

Beyond these broad generalizations, however, the results of detailed statistical studies of this type are distinctly limited. In the first place, the results hold true only for the particular year in which the records were collected. Differences in yields from one year to another and changes in the prices of each product and of each cost factor modify both the physical and the economic situation, so that many allowances must be made before the results can be applied in another year. Even if satisfactory adjustments can be made, there is still another limitation. Each individual farm is a different entity, and the organization which produces the best results *on the average* will not necessarily be the best for any one individual farm. If it were possible to observe one farm under one hundred different types of organization and operation, and record the resulting profit secured under each one, it would then be possible to judge from the analysis of those records what type of organization would yield the maximum returns for that farm under the same price conditions. But with the records of different farms representing not like entities but entities more or less unlike, the conclusions are not so applicable. Only on the assumption that the observations are drawn from a homogenous universe of similar conditions can the results of statistical studies of farm organization be interpreted to give the best organization for any one farm—and the areas where this assumption is justified are probably very few.

Relation of economic conditions to market price for a commodity. All the problems discussed to this point have been such that a certain universe might be specified, even though that universe would be likely to change more or less with the passage of time. The problem of prices, though, is of an entirely different character, for there only a single observation can be drawn for any given length of period, and the next period is essentially in a different universe. Even so, however, there is enough continuity to the way that individual persons react in the aggregate, and enough similarity between successive years, so that fairly stable results can sometimes be secured, and, where the change in reaction is continuous and progressive, that change itself can be made one variable in the analysis.

Annual prices. The simplest price studies are those which relate the market price for a commodity to the supply for a marketing year. The early work on this line by Moore (26) indicated the general relation of supply to price for corn, hay, oats, potatoes, and cotton. The influence of changing conditions were eliminated mainly by the use of first differences, so the resulting curves were not susceptible of logical economic interpretation. More recent work on potatoes (27, 28, 29), oats (30), and cotton (31, 32) has recognized the influence of price levels, trends in demand, carryover from previous years, and the prices of competing products as factors influencing price along with supply; and multiple correlation or alternative techniques are used to take into account the influence of the different variables. With relatively short periods on which to base the analyses, exceedingly high correlations have been secured in many cases—frequently above 0.95, even after adjusting for the number of constants. When forecasts have been made ahead, however, they have met with variable success, the forecasts in some years working out practically as well as in the period on which the analysis was based and in other cases missing by wide margins, sometimes by many times the standard error of estimate.

These extreme errors in forecasting seem to be due to the element of fortuitousness in economic events. Thus a factor which has been fairly constant for a number of years, and hence has shown little influence on price, may suddenly become very important—and upset a forecast based on the years in which it was unimportant. In addition to such universally upsetting changes as the outbreak of the World War, other illustrations of such sporadic and unforecasted events are the sudden decrease in the foreign demand for American hog products in the spring and summer of 1927 and the increasing competition of Indian cotton with American cotton in 1928 and 1929.[3]

Monthly prices. When monthly prices are considered, more elaborate statistical studies have been possible, with a larger number of individual observations. Although questions may be raised as to how closely the successive monthly prices of a staple commodity are really independent of each other, there is no question but that conditions are constantly changing, so that there are some elements of independence between successive observations. Of the earlier studies of monthly prices, a study of cotton prices by Smith (33) is of particular economic interest for its separation of the influence of actual and of prospective supply on price, and of the shifting of the regression curves for these factors through the season—determined as joint functions of the month and of the variable. A study of hog prices (34) developed both an empirical forecaster of prices and an economic interpretation of the influence of market receipts, storage stocks, competing products, and business conditions, on prices. The correlations were relatively low, however, and subsequent analyses have materially modified many of the conclusions. The monthly forecasts of hog prices based on this study were not as accurate as were forecasts which took a broader range of elements into consideration (35). Studies of monthly hog prices in Germany by Hanau (36) along the same line yielded reasonable results and gave forecasts which worked well in practice. A study of monthly prices of dressed lamb (37) which gave a correlation of 0.98 for the seventeen years studied is noteworthy in that the same formula served to estimate monthly prices (from current supply data) for three years afterwards, with almost the same accuracy as during the period studied. This analysis considered monthly per capita supplies, price level, competing products, and business activity as independent factors, and determined trend and seasonal variation while simultaneously eliminating the influence of other factors.

The studies mentioned include only a small portion of the statistical studies of price which have been made, but indicate some of the many ways in which statistical analysis has been used in this field—

[3] During the great economic depression after 1929, on the contrary, many price-analysis correlations continued to give fairly reliable forecasts, despite the great increase in the amplitude of fluctuation in industrial activity and consumer buying power.

and also some of the dangers and pitfalls that beset the investigator. Price analysis is the last place to apply statistical methods without thorough logical and economic analysis of the particular problem.

Weekly or daily prices. For very perishable products, where supplies and prices may fluctuate widely even from day to day, price studies may deal with the average prices for a week or even for an individual day. Representative statistical studies of this type are those of watermelons by Hedden and Cherniack (38) and of peaches by Kantor (39). In both these studies it was necessary to take into account a regular variation in demand from day to day of the week. Other factors influencing demand were also considered, and it was found that temperature had a marked influence on the price that would be paid for a given supply of watermelons. These short-period studies both related to an individual large market—New York City. Similar studies have been made for other markets and other products.

Relation of characteristics of different lots of a commodity to prices at which they sell. All the price studies which have just been discussed treated the reasons for the change in prices from time to time, for lots of the commodity of uniform or of average quality, and at the same stage of the marketing process. As has been pointed out, only one observation can be drawn from each successive universe. A type of study which presents different statistical problems is that of determining why different lots of the same commodity, sold within a given period and at the same stage of the marketing process, should sell for different prices. In this case there is a true universe—all the sales of the specified kind taking place within the specified period —and as large a sample as is desired can be secured, up to the limits of the universe. The studies of land prices previously mentioned are one example of this type of analysis; and the study of the relation of the price of apples to size, insect injury, and scab, used as an illustration in Chapter 23, is another example.

One of the most interesting studies of this type related the prices of different lots of asparagus to the length of green color in the stalk, the number of stalks in the bunch, and the uniformity of the stalks (40). The results of this study, presented very effectively in pictorial style (as reproduced in Fig. 75), have had a marked influence on the practices by the producers who supply the Boston market and have led to further experimental investigation as to how to produce asparagus with the desirable qualities (41). Similar studies have been made of the influence of size, color, interior quality, and type of carton on the prices received for eggs sold at retail, for both the New York and Philadelphia markets (42) and for the Wilmington market (43).

One logical point which cannot be overlooked in studies of the effect of quality on price, however, is that the premiums paid for high-quality lots may vary from time to time with differences in the relative supply of products of the different qualities. That is to say, though the conclusions as to the effect of quality upon price do apply

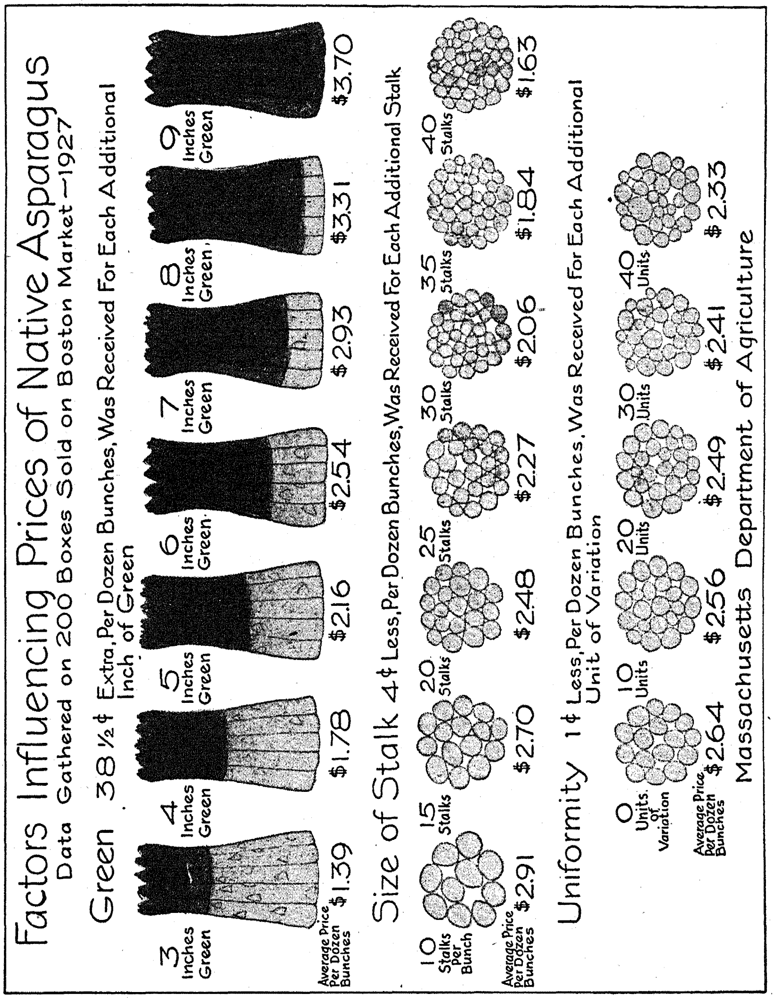


FIG. 75. A pictorial presentation of conclusions reached by a multiple correlation study. (From Frederick V. Waugh.)

in the universe from which the observations were drawn—with certain conditions as to the supply of the different sizes and qualities—they may not apply in a different universe in which the circumstances have changed. Other studies have therefore attempted to determine not only how the prices vary for different qualities under the set of supply conditions at one particular time but also how the premium or discounts varied from time to time with differences in the supplies of each quality. Thus studies of the influence of protein content, weight per bushel, dockage, and grade on the prices received for different cars of wheat (44) have shown that in crop years when high-protein wheat is very scarce, a wheat of high protein commands a marked premium, and that factor is much more important than weight; whereas, in years when high-protein wheat is more plentiful but much of the wheat is underweight, the weight factor becomes relatively more important, with the protein premium becoming of much less significance. With records of more than a thousand cars per year for several years, the changes in premiums were determined from month to month, by using two joint functions, one for month and protein content and the other for month and weight per bushel (44).

The effect of varying supplies of different sizes and varieties has also been studied in the case of peach prices (39). Here it was found that the premium for competing varieties changed with the supply of each, a variety which sold at a premium when only a small portion of the total supply, selling at a discount when it exceeded a certain percentage of the supply. The premium for peaches of large size, however, tended to persist in spite of increased supplies, though it was reduced somewhat when the proportion of large-sized peaches increased.

These last two groups of studies illustrate the way in which changing universes (in time series) may yet be brought within the purview of statistical analysis, and conclusions may be reached which will be of value in new sets of circumstances. If the complex of conditions changes from time to time because of factors such as differences in supply of different sizes, qualities, or varieties, or recurring differences in demand from day to day through the week, or from month to month through the year, which factors can be objectively taken account of and their influence measured with respect to the dependent factor or with respect to the influence of other variables on the dependent factor (joint relationships), then the fact that the circumstances are changing ceases to be a "bug-a-boo," because the reasons for the changes may be determined and allowed for. Just how far the conclusions from such analyses will hold under changed conditions depends upon how adequately the real causes of the changes from time to time have been determined, and how much unaccountable dynamic or evolutionary change there has been and may be. But even so, this approach seems the hopeful one in treating the baffling problem of changing conditions in time series; and it may yet be possible to apply laws of sampling and to make statistical forecasts for these cases with the same confidence that they can be made for stable universes.

Other price studies. Other types of price-analysis studies which may be mentioned briefly are those of differences in prices between different points in space or of different points in the marketing process. Correlation analysis has been applied to the first of these problems (45) in studying the relative influence of changes in freight rates, location of supplies, and price level on the margin between potato prices in Minneapolis and New York. Some studies of marketing costs (46, 47) have indicated the influence of size of creamery, distance of haul, and methods of operation on creamery costs and hence on prices received by farmers for their cream; but the general subject of the relation of prices of the same product at different points in the marketing process has not otherwise been investigated, except in the most general way (48).

Another variety of price study is in determining the influence of prices on the quantity of a product moved into consumption. In making studies of this sort for milk (49), it has been found that season of the year, day of the week, holidays, changing food habits, and income of the consumer have more influence on consumption than do price changes; but after these are eliminated a slight but significant change in consumption with change in price may be found. With cotton, on the contrary (32), it was found that price alone, with an upward trend in demand, almost completely determined the quantity consumed throughout the world; whereas the quantity consumed in the United States was also influenced by the general level of industrial activity (50). A similar relation for consumption of hog products to price was shown as an illustration in Chapter 6, Table 27. A parallel type of study indicates the effect of price on the quantity of cotton carried over at the end of the season or withheld or used by the producers. Thus it has been found that in years of low potato prices producers feed or waste much larger quantities, whereas when the prices fall below certain points much of the supply is left in the ground undug (51).

In all these price studies it must be recognized that logically price does not of and by itself determine consumption, carryover, and waste, nor does supply alone determine price. Instead where competition is effective there is a continuous dynamic balance of all the factors, which has been aptly described by the great economist Alfred Marshall as the closing of a pair of shears, where neither blade alone does the cutting. When, however, the relations have been analyzed step by step in the various ways which have been described, the different relations may then be pieced together in a harmonious whole which is logically consistent and which gives concrete statement to the economic hypotheses concerned (52).

Relation of changes in production to prices and other factors. Another type of problem in which prices are involved, but only as independent factors, is studying the influence of price changes on changes in production. The distinctive characteristic of these studies is that the prices in one period must be related to production in some subsequent period or periods, the length of lag depending on the technological length of the production process and on the time it takes producers to respond to changes in prices. One of the first studies of this type related cotton acreage to prices for the previous season (53). Subsequent experience showed, however, that continued high prices for two seasons might have a different influence than for a single season alone (54). Studies of hogs showed that it took eighteen months for differences in prices to be reflected in market receipts (34). The price of corn was found of equal importance with the price of hogs in causing changes in hog production. Hog production has also been studied by different type-of-farming areas, and this detailed study has shown marked differences in the responses to prices in different areas, depending on the position the hog enterprise occupied in the farming system. The weather conditions during the farrowing season in the spring and the relation of corn prices to hog prices during several critical periods in the production process were found to be important factors (55). The production of milk has likewise been found to respond to changes in the relation of the price of the milk to the costs of feedstuff (56). There is a short-time effect which is due to changes in the intensity of feeding and a long-time effect which is due to changes in the number of cows (57). The acreage of potatoes reflects prices for two years preceding, as well as prices for the year before. The responses for potatoes are quite parallel in different areas, though there are some important differences reflecting differences in the position the enterprise occupies in the farming system (54). In the case of some minor crops, the prices for the major crop of the region have as much influence on subsequent acreages of these competing crops as do the prices of the minor crops themselves. Thus sweet-potato acreage is influenced by cotton prices, and flax acreage by wheat prices. In other cases, yields or per-acre returns for preceding years must be considered, as well as prices alone. The general price level of competitive products or of all commodities has also usually been considered in judging the significance of a particular price.

In most of these studies of production responses, it has been found necessary to state the subsequent acreage or production as a percentage of, or as an absolute increase or decrease from, the acreage or production of the preceding year or production period. Stating the relation in this way recognizes the fact that the farmer or other producer must plan the next year's operations not with reference to any hypothetical normal or average but with reference to the actual production situation of the current year. Often, as is illustrated in Figure 76, a very high price will not call forth any larger increase in production in the following year than will a moderately high price, owing to the inability of the producer to expand his operations more than a certain extent in any one year. In this respect this type of price study is quite distinct from other types, for in many studies of the response of prices to supplies more satisfactory results have been secured by working with the absolute figures rather than with changes from year to year.

Miscellaneous agricultural problems. Another group of studies has investigated the relation of the physical characteristics of plants or animals to their ability to produce. Studies of dairy cows, by Gowen (58), restricted to simple two-variable correlations, have indicated that most of the factors in the physical conformation of dairy cows have little or no relation to productive ability. Studies of the relation of the size and shape of corn kernels, ears, and plants to weight of the grain (59) and multiple correlations by Richey which took the actual yielding ability as the criterion (60) have led largely to the same result. These studies indicate that many of the time-honored points which have been stressed in agricultural show competitions and in breeding selection have no utilitarian significance and have led to a new stress on performance records rather than physical appearance as the ultimate test.

Correlation in psychology and education. Correlation and multiple correlation methods have been widely applied in educational and psychological investigations to the study of such problems as the

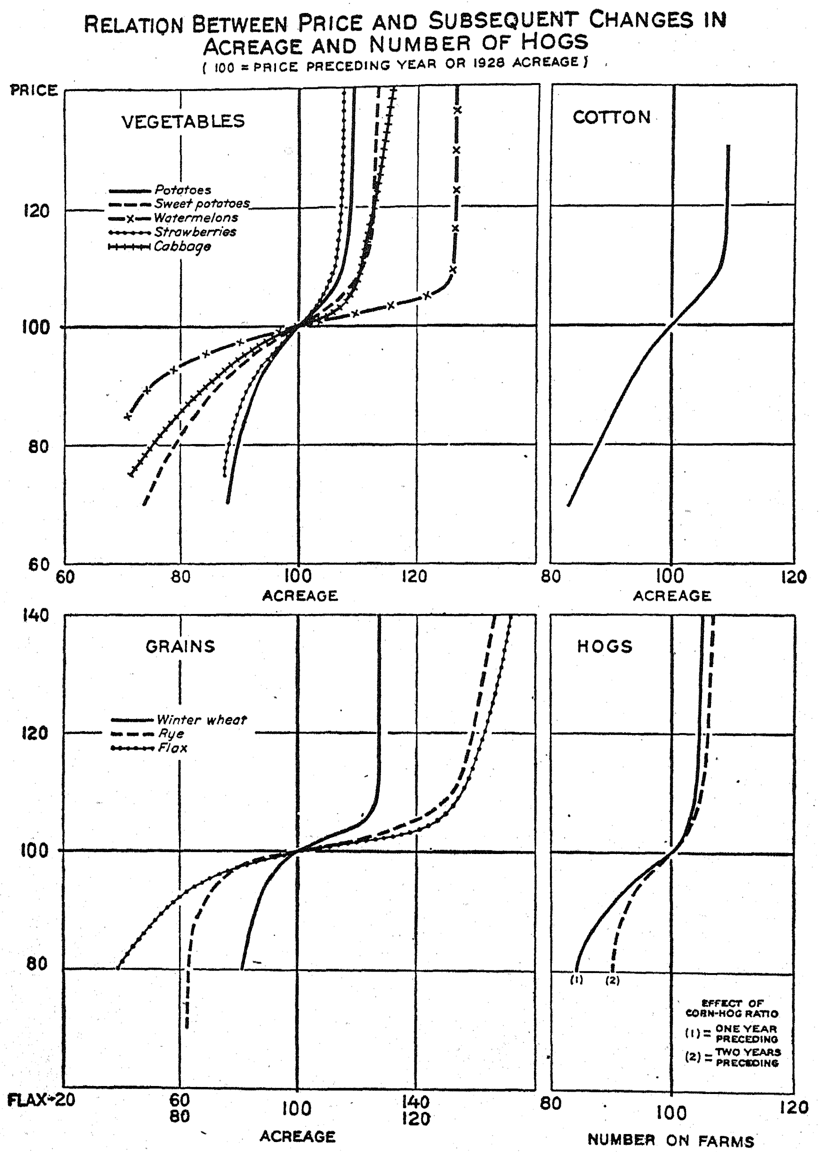


FIG. 76. Changes in acreage or production with changes in prices received, for different agricultural products. (From reference 54, by Louis H. Bean.)

relation of grades in one subject to grades in another (61), or the scores on one mental test to scores on another (62), or the relation of scores on mental tests to success in the schoolroom (63) or in later life (64). Studies have also been made of the relation of mental and physical characteristics to success in different occupations, such as the relation of the relative success of individual farmers to their training, schooling, initiative, business ability, etc. (65). This latter study, which indicated that approximately half the differences in farmers' financial success could be accounted for on the basis of individual differences in the men, has a tantalizing tie-up with the studies of farm management, which show that roughly half the differences in income can be explained by the way the farms are run. Apparently, by considering both the characteristics of the farmer and the way the farm is organized and run, it would be possible to account for *all* the differences in income. But if the men with superior mental ability are the men whose farms were organized and run in a superior manner, the ratings of the farmers and of their farming methods would be merely overlapping measures of the same thing.

In most of the cases in which correlation analysis has been applied to psychological problems, it has been used primarily to measure closeness of relationship rather than to obtain a basis for estimating one variable from another. In studies of this type even a low correlation may be important, so long as it is large enough so as not to be due to random fluctuations. Thus one study reached the conclusion that even in groups of the same economic and social status, there is a small negative correlation between number of children per family and intelligence (66). The psychologists and the biologists might have a warm argument, though, as to which was cause and which was effect! In another study, in which a given test was repeated, with twice as much time to complete it the second time, the scores made on the two trials were correlated, and correlations of 0.76 to 0.91 were found. These correlations were made the basis for concluding that the tests determine power alone, rather than speed (67). Inasmuch as a correlation of 0.76 means that nearly half the variance in the two factors is *not* associated, it might be questioned whether this interpretation is altogether satisfactory. Here the use of r (= 0.76) instead of d (= 0.58) leads to overstressing the significance of the observed correlation. Many other applications of correlation or partial correlation in psychological research (68, 69, 70) illustrate the usual tendency to depend on correlation coefficients, rather than regression equations, as the means of expressing relationship.

Interesting results have also been secured by the application of correlation methods to problems on the border line between psychology, sociology, and political science. Thus in a study of factors influencing the attitudes of mothers toward sex education, it was found that a number of measures of previous environment showed no significant correlation with the mother's attitude; but that there was a significant correlation between their opinion and the amount of sex education given their children (71). Another interesting study on the political-sociological border line determined the intercorrelations between the quantities of information, misinformation, and prejudice possessed by college students, and their grades, and their conservative or radical political positions. High prejudice, high misinformation, low grades, and conservatism were found to be associated; and likewise low prejudice, good grades, low misinformation, and radicalism. The correlations were low in all cases, however (72).

The use of correlation methods in the field of education and psychology has been hampered by the fact that in many cases the factors dealt with are not tangible facts which can be objectively measured but are intangibles which can be only roughly approximated by some process such as ranking. If anything approaching a normal distribution of the factor considered is assumed, ranking tends to make the true difference between successive individuals in the series much less in the central portions of the array than in the extreme portions. Furthermore, the ranked series is a discrete series, with the possibility always present that the sixth item in order, for example, may exceed the seventh item by 10 times the amount that the seventh exceeds the eighth, or vice versa. Both these difficulties are apparent in the accompanying set of data.

Grades Received by 24 Persons Taking an Examination in Statistics

Rank	Grade	Rank	Grade	Rank	Grade	Rank	Grade
1	99	7	94	13	85	19	81
2	98	8	91	14	85	20	80
3	98	9	90	15	85	21	77
4	98	10	88	16	83	22	75
5	97	11	87	17	83	23	69
6	94	12	85	18	82	24	68

The difficulties enumerated have made psychological workers and educators feel that the standard Pearsonian methods of correlation (those presented in Chapters 4 and 5, and 12 and 13, of this book) are not applicable to their data, and have led to the development of various alternative methods, such as the Spearman "foot-rule correlation" for ranked data (73) and other similar short cuts. It is not evident that these new measures meet the difficulties enumerated, and furthermore they give measures of correlation which differ from the Pearsonian coefficients for the same data. The use of curvilinear regressions, as discussed in Chapter 7 and subsequent chapters, partly meets the difficulties in handling such data, since the effect of a varying significance of the unit of measurement in different portions of the range may result in transforming what would otherwise be linear regressions to a non-linear shape. That does not, however, meet the difficulties of the discrete or "jumpy" quality of ranked values; nor does it seem that any other statistical treatment is likely to do so completely.

Where the dependent variable is definitely discrete, so that two or more categories can be recognized, but no continuous variation can be assumed, correlation methods are clearly inapplicable. Special statistical measures of association, parallel to the correlation coefficient, have been worked out for such problems (74).

No attempt has been made in this book to treat the special correlation methods developed in educational and psychological work. Instead, it has been restricted to the analysis of dependent variables which were continuously variable or which could logically be thrown into that form.

Correlation analysis in other fields. The types of problems which have been discussed do not begin to exhaust the uses which have been made of correlation, simple or multiple, in research work. Since they are drawn largely from the author's own range of interest, they are heavily weighted by the agricultural or even the agricultural economic field. Random examples of correlation work in other fields are the use of multiple correlation to obtain a definite formula for forecasting pig-iron production (75), to determine the extent to which freight rates are influenced by the factors of terminal charges, length of haul, expense of operation, and other factors (76), or to determine how far meat sales in different branch houses are influenced by local conditions of demand, and what a reasonable quota might be (77).

More recent applications of correlation analysis. The discussion to this point in this chapter remains substantially unchanged from that of the first edition of this book. Since that edition was published, there has been a vast expansion of research work in many of these fields. In some fields, such as commodity price analysis, an entire book would be required merely to discuss subsequent studies (78).

Here we shall simply note briefly some of the additional fields to which correlation analysis has been applied in the decade since the first edition was published, without attempting to appraise the subsequent work in the fields already mentioned.

Price-making forces for industrial commodities. The same methods used earlier with farm-product prices have more recently begun to be applied to the explanation of industrial price-making forces. Steel (79), automobiles (80), houses (81), and ships (82) illustrate some of these studies. In fields where free competition does not prevail, but the dominance of a few large concerns produces monopolistic competition, the supply and price relations may operate quite differently from the way they operate under fuller competition (83). In such cases great care is necessary to set up the statistical analyses in such terms as to represent the market situation as it really functions in the given industry.

Production functions for industries. The relation of volume of output to average cost per unit is an important consideration both in economic theory and in industrial organization. Recent overall studies for certain large concerns (84, 85, 86) have revealed the cost function for such products as steel, hosiery, and furniture. In some of these studies, multiple correlation was used to measure the influence of percentage of capacity operated on total cost or per-unit cost, while simultaneously holding constant other factors such as wage rates, price levels, or changing labor efficiency.

Size standards for children's clothes. Quite a different recent application of correlation technique was made in a study of appropriate size standards for children's clothes, conducted by the Bureau of Home Economics (87). In this study, all possible bodily dimensions were measured for thousands of children all over the country, together with their age, sex, and race. Multiple correlation was used to determine which of these measurements were most important in judging size as a whole. It was found that height and girth at hips were the most important. After these were allowed for, age was found to have no appreciable relation to the other bodily measurements. A new system of clothes sizes, based on the distribution of these two measurements, was recommended to clothing manufacturers. By using these sizes instead of the conventional age sizes, it will be possible to have ready-made clothing which can be bought merely by size, and yet have a satisfactory fit for a large proportion of all children.

Measures of components of intelligence. Certain workers in psychology and education have modified correlation procedures to investigate the problem of how many *independent* factors are involved in intelligence. Spearman introduced the theory that there was one general factor which ran through all intelligence tests, plus various specific factors in each test, and made extensive statistical studies to substantiate this claim (88). Other students advanced the theory that three or four general factors, differently weighted in each case, could explain *all* the different measures of intelligence (89). Although these investigations have led into involved calculations and highly refined mathematics, their actual significance is still in doubt.

Explanations of political behavior. During the past decade the methods of statistical analysis, especially of sampling, have been extensively applied to the field of political behavior. The earlier Literary Digest Poll, and the more refined and scientific Gallup Poll and Roper Poll, have become almost household words. Along with these, correlation analysis has been used to show the relations between votes by states and national averages, and to develop the predicting reliability of opinions or votes in particular areas (90). Correlation methods, including some of the highly involved methods of psychological studies referred to in the preceding paragraph, have also been used in detailed studies of political structure and behavior in particular cities or localities (91).

Tests of correlation results. With the passage of the years it has been possible to verify some of the earlier studies by applying them to later data, or by analyzing data for entire subsequent periods to see if they gave comparable results. Some subsequent studies of the response of milk production to prices received, however, gave quite different results from those given by earlier studies, and led to the conclusion that factors which had shown a high correlation with production while the industry was expanding in a given region failed to have the same significance after maturity was reached (92). In this case the economic growth proved to be irreversible. These studies prompted more detailed analysis of the problem and led to the development of more intensive techniques, which consider not only prices but also the whole farm-management organization of typical farms in reaching conclusions as to the long-run response of production to price (93). In a quite different case, the response of milk production to variations in feed input is being tested by elaborate feeding experiments, with the resulting data subjected to thorough statistical analyses (94). The preliminary results from these analyses show a net relation of milk output to feed input which agrees surprisingly well with the same relation as determined earlier by multiple correlation analysis (6) from cow-testing association records of actual farm experience (95).

Other applications. Other new applications of correlation methods have been made in testing the strength of materials when subjected to varying stresses, in determining the effect of various local water characteristics on the amount of inside deposit in water or steam pipes made of various materials, and in establishing sales quotas or advertising allotments for specific products in various districts in the light of the industrial and economic characteristics of each district. Since these studies were made in private research agencies for the benefit of private concerns, the results have usually not been published. In some cases the findings are regarded as valuable trade secrets. The variety of problems to which correlation, and especially multiple correlation, has been applied, does, however, indicate the significance of this technique as a means of unlocking secrets of relationship in many cases where they could be discovered in no other way.

Many more pages might be filled with the details of studies such as those discussed. But probably enough has been presented to illustrate the wide range of problems in which the use of statistical analysis sheds new light on the relationships present and their significance. It may be hoped that these illustrations have developed the necessity for careful logical analysis of each problem to which statistical analysis is to be applied, and have indicated the need both for good theoretical knowledge of the field in which the problem lies and for thorough technological knowledge of the elements involved in the particular problem. The technological knowledge is particularly important in selecting the different factors or in deciding on their statement or interpretation.

No attempt has been made here to list all the significant statistical studies in any one of the fields discussed, or to evaluate their importance. Instead, the studies mentioned have been selected solely to illustrate various specific points; in many cases a significant study has not been referred to because the point was already covered, or a relatively unimportant study has been mentioned because of its pertinence to a particular topic. This discussion should therefore not be regarded as a critical evaluation of the work in any of the fields touched upon. That has been left for experts in each field. Instead, the comments are intended solely to develop the variety, complexity, and significance of the problems to which statistical analysis may be applied and the care and thought which are even more necessary than the statistical computations, if the results are to be of lasting value.

REFERENCES

1. Haas, G. C. Sale prices as a basis for farm land appraisal. *Univ. Minn. Agr. Expt. Sta. Tech. Bul.* 9. 1922.

2. Ezekiel, Mordecai. Factors affecting farmers' earnings in Southeastern Pennsylvania. *U. S. Dept. Agr. Bul.* 1400, pp. 39–60. 1926.

3. Tolley, H. R., J. D. Black, and M. J. B. Ezekiel. Input as related to output in farm organization and cost-of-production studies. *U. S. Dept. Agr. Bul.* 1277, pp. 7–12. 1924.

4. Misner, E. G. Relation of the composition of rations on some New York dairy farms to the economics of milk production. *Cornell Univ. Agr. Expt. Sta. Memoir* 64. 1923.

5. Vernon, J. J., C. W. Holdaway, Mordecai Ezekiel, and R. S. Kifer. Factors affecting returns from the dairy enterprise in the Shenandoah Valley. *Va. Agr. Expt. Sta. Bul.* 257, pp. 32–42 1927.

6. Ezekiel, M. J. B., P. E. McNall, and F. B. Morrison. Practices responsible for variations in physical requirements and economic costs of milk production on Wisconsin dairy farms. *Wis. Agr. Expt. Sta. Research Bul.* 79. 1927.

7. Pond, George, and Mordecai Ezekiel. A study of some factors affecting the physical and economic costs of butterfat production in Pine County, Minn. *Univ. Minn. Agr. Expt. Sta. Bul.* 270. 1930.

8. Johnson, Sherman E., J. O. Tretsven, Mordecai Ezekiel, and O. V. Wells. Organization, feeding methods, and other practices affecting returns on irrigated dairy farms in Western Montana. *Univ. Montana Agr. Expt. Sta. Bul.* 264. 1932.

9. Hardenburg, E. V. A study, by the crop survey method, of factors influencing the yield of potatoes. *Cornell Univ. Agr. Expt. Sta. Memoir* 57. 1922.

10. Reference (3) above, pp. 16–18.

11. Westbrook, E. C., W. A. Minor, Jr., Kenneth Traynor, C. L. Goodrich, and W. C. Funk. An economic study of farm organization in Sumter County. *Georgia State College of Agr. Bul.* 324, pp. 82–87. December, 1927.

12. Misner, E. G. Studies of the relation of weather to the production and price of farm products. I. Corn. Cornell Univ., mimeographed publication. March, 1928.

13. Waugh, Frederick V., Chester D. Stevens, and Gustave Burmeister. Methods of forecasting New England potato yields. U. S. Dept. Agr., Bur. Agr. Econ., mimeographed report. February, 1929.

14. Moore, Henry L. *Economic Cycles; Their Law and Cause*, pp. 35–44. Macmillan. 1914.

15. Smith, Bradford B. Relation between weather conditions and yield of cotton in Louisiana. *Jour. Agr. Res.*, Vol. XXX, No. 11, pp. 1083–1086. June 1, 1925.

16. Patton, Palmer. Relationship of weather to crops in the plains region of Montana. *Mont. Expt. Sta. Bul.* 206. 1927.

17. Smith, Bradford B. The adjustment of agricultural production to demand. *Jour. Farm. Econ.*, Vol. VIII, No. 2, pp. 163–165. April, 1926.

18. Fisher, R. A. The influence of rainfall upon the yield of wheat at Rothamsted. *Phil. Trans.*, B., CCXIII, pp. 89–142. 1924.

19. Shollenberger, J. H., and Corinne F. Kyle. Correlation of kernel texture, test weight per bushel, and protein content of hard red spring wheat. *Jour. Agr. Res.*, Vol. 35, No. 12, pp. 1137–1150. Dec. 15, 1927.

20. Coleman, D. A., H. B. Dixon, and H. C. Fellows. Comparison of some physical and chemical tests for determining the quality of gluten in wheat and flour. *Jour. Agr. Res.*, Vol. 34, No. 3, pp. 241–264. Feb. 1, 1927.

21. Chatfield, Charlotte. Proximate composition of beef. U. S. Dept. Agr., *Dept. Circular* 389. 1926.

22. Pettit, Edison. Ultra-violet solar radiation. *Proc. Nat. Acad. Sciences*, 13. p. 380. 1927.

23. Reference (2) above, pp. 20–25, 54–59.

24. Taylor, C. C. A statistical analysis of farm management data. *Jour. Farm Econ.*, V, pp. 153–162. June, 1923.

25. Vernon, J. J., and M. J. B. Ezekiel. Causes of profit or loss on Virginia tobacco farms. *Va. Agr. Expt. Sta. Bul.* 241. 1925.

26. Moore, Henry L. *Economic Cycles; Their Law and Cause*, pp. 63–134. Macmillan. 1914.

—— *Forecasting the Yield and Price of Cotton.* Macmillan. 1917.

27. Working, Holbrook. Factors determining the price of potatoes in St. Paul and Minneapolis. *Univ. Minn. Agr. Expt. Sta. Tech. Bul.* 10. 1922.

28. ——Factors affecting the price of Minnesota potatoes. *Minn. Agr. Expt. Sta. Tech. Bul.* 29. 1925.

29. Waugh, Frederick V. Forecasting prices of New Jersey white potatoes and sweet potatoes. *N. J. State Dept. Agr. Circ.* 78. 1924.

30. Killough, Hugh B. What makes the price of oats. *U. S. Dept. Agr. Bul.* 1351. 1925.

31. Reference (17) above, pp. 145–153.

32. Bean, Louis H. Some interrelationships between the supply, price, and consumption of cotton. U. S. Dept. Agr., Bur. Agr. Econ., mimeographed report. April, 1928.

33. Smith, Bradford B. Factors affecting the price of cotton. *U. S. Dept. Agr. Tech. Bul.* 50. 1928.

34. Haas, G. C., and Mordecai Ezekiel. Factors affecting the price of hogs. *U. S. Dept. Agr. Bul.* 1440. 1926.

35. Ezekiel, Mordecai. Two methods of forecasting hog prices. *Jour. Amer. Stat. Assoc.*, 22, pp. 22–30. March, 1927.

36. Hanau, Arthur. Die Prognose der Schweinepreise. Vierteljahrshefte zur Konjunkturforschung, Sonderheft 7. Institut für Konjunkturforschung. Berlin February, 1928.

37. Ezekiel, Mordecai. Factors related to lamb prices. *Jour. Pol. Econ.*, Vol. XXXV, No. 2. April, 1927.

38. Hedden, W. P., and Nathan Cherniack. Measuring the melon market. Preliminary (mimeographed) report, U. S. Dept. Agr., Bur. Agr. Econ., in cooperation with the Port of N. Y. Authority, August, 1924.

39. Kantor, Harry. Factors affecting the price of peaches in the New York City market. *U. S. Dept. Agr. Tech. Bul.* 115. 1929.

40. WAUGH, FREDERICK V. *Quality as a Determinant of Vegetable Prices*, pp. 39–45. Columbia Univ. Press. 1929.

41. DIEDJENS, V. A., W. D. WHITCOMB, and R. M. KOON. Asparagus and its culture. *Mass. Agr. College Extension Leaflet* 49. April, 1929.

42. HOWE, CHARLES B. Some local market price characteristics which affect New Jersey egg producers; factors influencing the retail prices of eggs. *N. J. Agr. Expt. Sta. Bul.* 1930.

43. BENNER, CLAUDE L., and HARRY G. GABRIEL. Marketing of Delaware eggs. *Del. Agr. Expt. Sta. Bul.* 150. 1927.

44. KUHRT, W. J. A study of farmer elevator operation in the spring wheat area. Series of 1925–26. Part II. Analysis of the variation in the quality factors of the 1925 crop of spring wheat, and the relation of such variation to price received and premiums paid in 1925–26. U. S. Dept. Agr., Bur. Agr. Econ., preliminary report. October, 1927.

45. WORKING, HOLBROOK. Factors influencing price differentials between potato markets. *Jour. Farm Econ.*, pp. 377–398. October, 1925.

46. BLACK, JOHN D., and EDWARD S. GUTHRIE. Economic aspects of creamery organization. *Univ. Minn. Agr. Expt. Sta. Tech. Bul.* 26. 1924.

47. SCHOENFELD, WILLIAM A. Some economic aspects of the marketing of milk and cream in New England. *U. S. Dept. Agr. Circ.* 16, pp. 24–29. 1927.

48. WARREN, GEORGE F., and F. A. PEARSON. Interrelationships of supply and price. *Cornell Univ. Agr. Expt. Sta. Bul.* 466. 1928.

49. ROSS, H. A. The demand side of the New York milk market. *Cornell Univ. Agr. Expt. Sta. Bul.* 459. 1927.

50. BEAN, LOUIS H. A simplified method of graphic curvilinear correlation. *Jour. Amer. Stat. Assoc.* December, 1929.

51. —— Demand and supply curves on potatoes and cotton. 1929. Unpublished manuscript, on file in Bureau of Agricultural Economics Library.

52. EZEKIEL, MORDECAI. Statistical analyses and the "laws" of price. *Quart. Jour. Econ.*, Vol. XLII, pp. 199–225. February, 1928.

53. SMITH, BRADFORD B. Forecasting the acreage of cotton. *Jour. Amer. Stat. Assoc.*, Vol. 20, No. 149, pp. 31–47. 1925.

54. BEAN, LOUIS H. The farmer's response to price. *Jour. Farm. Econ.*, Vol. XI, No. 3, pp. 368–385. July, 1929.

55. ELLIOTT, FOSTER F. Adjusting hog production to market demand. *Univ. Ill. Agr. Expt. Sta. Bul.* 293. 1927.

56. GANS, A. R. Elasticity of supply of milk from Vermont plants. *Vt. Agr. Expt. Sta. Bul.* 269. 1927.

57. Reference (47) above, pp. 34–50.

58. GOWEN, JOHN W. Studies on conformation in relation to milk producing capacity in cattle. *Jour. Dairy Science*, Vol. III, No. 1, January, 1920; Vol. IV, No. 5, September, 1921.

—— Conformation and milk yield in the light of the personal equation of the dairy cattle judge. *Maine Agr. Expt. Sta. Bul.* 314. 1923

59. WOLFE, T. K. A biometrical analysis of characters of maize and of their inheritance. *Va. Agr. Expt. Sta. Tech. Bul.* 26. 1924.

60. RICHEY, FREDERICK D. A statistical study of the relation between seed-ear characters and productiveness in corn. *U. S. Dept. Agr. Bul.* 1321. 1925.

61. MENSENKAMP, L. E. Ability classification in ninth-grade algebra. *The Mathematics Teacher.* January, 1929.

62. Garrison, K. C. Correlation between intelligence test scores and success in certain rational organization problems. *Jour. Applied Psychol.* December, 1928.

63. Weeks, Angelina L. A vocabulary information test. *Archives of Psychol.* May, 1928.

64. Hull, Clark L. Prediction formulae for teams of aptitude tests. *Jour. Applied Psychol.* Vol. VII, pp. 277–284. 1923.

65. Higbie, Edgar Creighton. *An Objective Method for Determining Certain Fundamental Principles in Secondary Agricultural Education.* Published at Madison, Wis., by the author. 1924.

66. Sutherland, H. E. G. The relationship between I.Q. and size of family. *Jour. Educ. Psychol.* February, 1929.

67. Freeman, Frank S. Power and speed, their influence upon intelligence test scores. *Jour. Applied Psychol.* December, 1928.

68. Chauncey, Marlin R. The relation of the home factor to achievement and intelligence test scores. *Jour. Educ. Res.*, Vol. XX, No. 2, 88. September, 1929.

69. Winch, W. H. Accuracy in school children. Does improvement in numerical accuracy "transfer"? *Jour. Educ. Psychol.*, 1, 557–589. 1910.

70. Goodenough, F. L. The Kuhlman-Benet tests for children of pre-school age. Univ. Minn. Institute of Child Welfare, Mon. Series 2. 1928.

71. Witmer, Helen Leland. *Attitudes of Mothers Toward Sex Education.* Univ. Minn. Press. 1928.

72. Allport, Gordon W. The composition of political attitudes. *Amer. Jour. Sociology*, Vol. XXXV, 2, pp. 220–238. September, 1929.

73. Spearman, C. A footrule for measuring correlation. *British Jour. Psychol.*, Vol. II, p. 89. 1906.

74. Yule, G. Udny. *An Introduction to the Theory of Statistics*, Chapters III and IV, pp. 25–27. Sixth edition. C. Griffin and Co., Ltd., London. 1922.

75. Smith, Bradford B. Forecasting the volume and value of the cotton crop. *Jour. Amer. Stat. Assoc.*, pp. 453–458. December, 1927.

—— The use of interest rates in forecasting business activity. Proceedings of management week at Ohio State University, 1926. Published by Ohio State Bureau of Business Research.

76. Crum, W. L. The statistical allocation of joint costs. *Jour. Amer. Stat. Assoc.*, 21, pp. 9–24. March, 1926.

77. Cowan, Donald R. G. The commercial application of forecasting methods. *Jour. Farm. Econ.*, pp. 139–163. January, 1930.

78. For comprehensive bibliographies of price analysis studies, see Louise O. Bercaw. Price analysis. U. S. Dept. Agr., Bur. Agr. Econ., Bibliography 48. September, 1933. Price studies of the U. S. Dept. Agr. showing demand-supply, supply-price, and price-production relationships. U. S. Dept. Agr., Bur. Agr. Econ., Bibliography 58. October, 1938. (Both mimeographed.) Supplementary typewritten bibliographies covering later studies are also available from the Bureau of Agricultural Economics Library. See also F. L. Thomsen. *Agricultural Prices.* McGraw-Hill Book Co., Inc., New York. 1936; and Henry Schultz. *The Theory and Measurement of Demand.* Univ. Chicago Press. 1938.

79. Hearings before the Temporary National Economic Committee, Part 26. Iron and steel industry. A statistical analysis of the demand for steel, 1919–38, pp. 13,913–13,942. Washington. 1940.

80. Roos, C. F., and Victor von Szeliski. Factors governing changes in domestic automobile demand. *The Dynamics of Automobile Demand,* General Motors Corporation. New York. 1939.

81. Derksen, J. B. D. Long cycles in residential building: an explanation. *Econometrica,* Vol. VIII, pp. 97–116. October, 1940.

82. Koopmans, T. *Tanker Freight Rates and Tankship Building.* Netherlands Economic Institute. London. 1939.

83. Chamberlin, Edward. *The Theory of Monopolistic Competition.* Harvard University Press, Cambridge. 1936.

84. Hearings before the Temporary National Economic Committee, Part 26. Iron and steel industry, Exhibit 1416, an analysis of steel prices, volumes, and costs—controlling limitations on price reductions, pp. 14,032–14,082. Washington. 1940.

85. Wylie, Kathryn H., and Mordecai Ezekiel. The cost curve for steel production. *Jour. Pol. Econ.,* Vol. XLVIII, pp. 777–821. December, 1940.

86. Dean, Joel. Statistical cost curves in various industries. Report of Philadelphia meeting of Econometric Society, Dec. 27–29, 1939. *Econometrica,* Vol. VIII, p. 188. April, 1940.

87. Girshick, Meyer, and Ruth O'Brien. Children's body measurements for sizing garments and patterns. *U. S. Dept. Agr. Misc. Pub.* 365. 1940.

88. Spearman, C. The factor theory and its troubles. I. Pitfalls in the use of probable errors. *Jour. Educ. Psychol.* 1932. II. Garbling the evidence. *Jour. Educ. Psychol.* October, 1933. III. Misrepresentation of the theory. *Jour. Educ. Psychol.* November, 1933. IV. Uniqueness of *G. Jour. Educ. Psychol.* February, 1934. V. Adequacy of proof. *Jour. Educ. Psychol.* April, 1934.

———. Analysis of abilities into factors by the method of least squares. *Brit. Jour. Educ. Psychol.,* Vol. IV. June, 1934.

89. Thurstone, L. L. *The Vectors of Mind, Multiple-factor Analysis for the Isolation of Primary Traits.* Univ. Chicago Press. 1935.

90. Bean, L. H. *Ballot Behavior.* American Council on Public Affairs, Washington. 1940.

91. Gosnell, Harold. *Machine Politics, Chicago Model.* Univ. Chicago Press, Chicago. 1937.

92. Cassels, J. M., and W. Malenbaum. Doubts about statistical supply analysis. *Jour. Farm Econ.,* Vol. XX, No. 2. 1938.

Mighell, R. L., and R. H. Allen. Supply schedules—"long-time" and "short-time." *Jour. Farm Econ.,* Vol. XXII, No. 3. 1940.

93. Allen, R. H., Erling Hole, and R. L. Mighell. Supply responses in milk production in Cabot-Marshfield, Vermont. *U. S. Dept. Agr. Tech. Bul.* 709. 1940.

94. Jensen, Einar. Determining input-output relationships in milk production. *U. S. Dept. Agr. Farm Management Reports,* No. 5. January, 1940.

95. Ezekiel, Mordecai. A check on a multiple correlation result. *Jour. Farm Econ.,* Vol. XXII, No. 2. 1940.

CHAPTER 24

STEPS IN RESEARCH WORK AND THE PLACE OF STATISTICAL ANALYSIS

Relation of statistical analysis to research. Statistical analysis is only a tool to be used by the investigator. The analyst must be a worker in some field, or in several; he cannot use his statistical training except in analyzing problems any more than a carpenter can use his skill without lumber and something to be made. Now that the routine of statistical analysis has been discussed, and the types of problems to which it may be applied have been surveyed, it is pertinent to ask just what are the steps in research work and just where and how does statistical analysis fit into the picture.

The research worker must have an adequate knowledge of the facts, technical and otherwise, of the field in which he is to work. This knowledge is usually insured by the situation that in most cases the worker is a biologist, an economist, a psychologist, or an agronomist, first, and then a statistician only secondarily or in addition. When his training has been primarily in mathematics or statistics, however, the statistician must acquaint himself thoroughly with the facts and theories of the field involved before he can expect to do significant and substantial work.

Stating the objective. If adequate acquaintance with the field is given, the first step in a particular research problem is setting up the objective of the project. The objective can best be stated in the form of a direct question, such as "Why does lettuce sell for more on some days than on other days?" The more exact and specific the question can be made, the more clearly is the field of the investigation defined. Thus if we make the question read "Leaf lettuce sold at retail in Boston" instead of merely "lettuce," the scope of the study is much more definitely indicated. Stating the objective as a question has the important effect of clarifying the issue, and so insuring that the worker knows what he is really trying to find out. It has the further effect of instantly challenging the attention and of instinctively calling forth mental answers which aid in the next step of the research.

Any research project which cannot be stated as a definite question has not been clearly defined. Starting out merely "to collect figures on lettuce marketing" would not constitute research. Clear formulation of the question to be answered is an essential prerequisite of good research work.

Developing an hypothesis. The second step in the development of the problem is a deductive analysis of the question raised to suggest possible answers. This deductive analysis draws on all the theoretical and practical training and experience the worker has. In addition, he may study previous work along the same lines, ask questions of those concerned in the industry, or make brief reconnaissance studies to decide on the factors which may be involved and to judge of the probable relationships. This phase of the research should lead to the setting up of a definite hypothesis as to the elements which will be involved and of the ways in which they will be related. Thus in the lettuce problem, the hypothesis might be that the supply of leaf lettuce was the most important factor determining the price and that the larger the supply, the lower the price; that the supply of Iceberg lettuce also influenced the price of leaf lettuce, large supplies of Iceberg tending to depress the price of leaf lettuce; that weather affected the demand, prices for the same supply being higher in hot weather than in cool; that prices of other vegetables, such as tomatoes and cucumbers, might also influence lettuce prices, either as competitive products tending to depress lettuce prices when their prices were low or as complementary products tending to raise lettuce prices when their prices were low. It might further be supposed that variations in the purchasing power of consumers would affect the demand and that changes in the general price level of foodstuffs would also have some effect. Finally, it might be supposed that the demand would vary regularly from day to day through the week, owing to the purchasing habits of consumers, and from time to time through the year.

The process of developing the hypothesis may be aided by breaking up the main question to be answered into a number of subquestions, each one of which may be further broken up. Thus the initial lettuce question may be broken up into questions such as "Do (the specified prices) vary because of supply? Because of demand? Supplies of what? Leaf lettuce? Iceberg lettuce? Competing products? What are competing products? What makes demand? Weather? Purchasing power? Seasonal factors?" and so on until complete details have been thought out for every phase.

In setting up the hypothesis the investigator should also attempt to think through the probable nature of the relationships. Thus, should it be assumed that the influence of supply of leaf lettuce on price will be constant, independent of other factors, or is the relation likely to change from time to time through the year, or from day to day with the weather?

In setting up his hypotheses, the investigator not only should rely on his own knowledge but also should draw upon all the skill and knowledge of others who have experience in the same field. This will involve not only a careful study of earlier investigations of the same problem but also discussions with practical men who are operating in the field to be studied. Thus the student of lettuce prices should talk with wholesale produce merchants, retail grocerymen, farmers producing lettuce, and even chefs and housewives, to get their opinions of the factors influencing lettuce prices. This will enable the student to check his hypotheses against the ideas of practical men dealing with the same problem, and often may call to his attention elements in the situation which otherwise he might completely overlook.

Measuring the factors. Once the hypothesis has been set up, and the various factors enumerated in it have been considered with much care to make sure that every important element has been included, the next step is to secure measurements of the various factors to be studied. This will involve deciding whether the data are to be taken from published records or other secondary sources, or whether they are to be secured first hand. If first-hand collection is decided upon, further detailed study is involved as to where the ultimate facts are, who has knowledge of them or records of them, and how they are to be collected—by measurement, by direct observation, by enumerators, by schedules, by mail questionnaires, etc. Extended discussions of the advantages and disadvantages of each method, and the problems involved in laying out a record form, defining the units, securing the records, and checking or editing the reports are available in standard statistical textbooks[1] and will not be repeated here.

[1] Arthur L. Bowley, *Elements of Statistics,* Chapters III and VIII, pp. 18–57, 178–195, fourth edition, P. S. King & Son., Ltd., London, C. Scribner's Sons, New York, 1920.

Horace Secrist, *Introduction to Statistical Methods,* pp. 22–52, 65–71, The Macmillan Co., New York, 1917.

Harry Jerome, *Statistical Method,* pp. 13–23, Harper and Brothers, New York, 1924.

William L. Crum and Alson C. Patton, *An Introduction to the Methods of*

Precautions also need to be observed if secondary sources are used; these precautions also are well discussed in the references just given. Only one point will be developed here, and that is the special need of *completeness* in the records, particularly if original data are to be secured. Once an enumeration or observation has been made, additional data can be secured only at much extra trouble and expense, or in many cases cannot be secured at all. That is one reason why the hypothesis must be carefully studied beforehand to make sure all relevant factors are included, and why the preliminary study and investigation are so important. Factors which are stumbled upon or which suggest themselves in the later analysis may be of value in subsequent studies of the same type, but if the essential data are lacking the suggestions are too late to be of any value in the current study.

In obtaining the basic data it is necessary to decide on the particular items to be measured to represent the hypothetical factors. Are the weather elements to be rainfall, or wind, or temperature? If temperature, average, or maximum, or minimum? If average, what kind of average? And so on through a lengthy number of details, each one of which must be carefully considered in view of the hypothetical significance of the factor, the probable relations involved, and the effect which is expected to be shown.

Studying the apparent relations. After numerical values are available for all the elements, the next step is to make a thorough study of the apparent relationships before proceeding to more elaborate analyses. Both the relation of the independent factors to each other and the relation to the dependent must be studied, for, as has been pointed out before, the relation of the dependent factor to an indpendent factor that is not related to the others can be determined by simple correlation, whereas otherwise multiple correlation might be necessary. (This does not hold, however, if joint functions are present.) It is at this point that the investigator begins to test out the various elements in the hypothetical picture and to compare the hypothesis with the observed facts. Some elements which were thought to be of importance may prove unrelated, and other variables which were thought of doubtful significance may show important relations. This preliminary examination may even prove the entire hypothesis to be wrong and necessitate a re-examination of the basic

Economic Statistics, Chapters II, III, IV, pp. 15–38, A. W. Shaw Co., Chicago and New York, 1925.

FREDERICK E. CROXTON and DUDLEY J. COWDEN, *Applied General Statistics*, Chapter II, pp. 15–48, Prentice-Hall, Inc., New York, 1939.

ideas and a reformulation of the proposed explanation more in line with the facts as observed.

Running a Correlation Analysis

The preliminary examination of the data will provide the basis for setting up the final multiple correlation analysis, if the interrelations are such that such an analysis is finally found to be needed. As included in this analysis, each variable will have a definite place in the hypothesis, and some specific kind of relation will be expected to be found when the analysis is completed. Looked at in this way, the correlation analysis is not the whole of the research project, but is merely that portion of it in which the adequacy of the theoretical hypothesis is tested and in which the exact relations, as expected in the hypothesis, are measured and determined.

Units in which variables are stated. Once the variables to be employed in the final statistical analysis are selected, the next problem is to decide in what units to state them. In studying land values, for example, the value of a given farm may be stated as total value, as value per acre of all land, or as value per acre of improved land. Which one to select depends on what other variables are included and how they are to be stated. The total value of the farm might be correlated with the value of the dwelling, the value of other buildings, the acres in cultivated land, the acres in pasture, etc. This would tend to show the contribution *per acre* of each of the acreage elements and should give a high correlation, since under normal conditions the value of the farm may be expected to approximate the value of the buildings *plus* that of the several tracts of land. In this case the simple or additive regression equation would be quite appropriate, for it would give

Farm value = value of dwelling + value of other buildings
+ (value per acre of cultivated land) (number acres of cultivated land)
+ (value per acre of pasture land) (number acres pasture land)
+ (value per acre of woodland) (number acres woodland) + etc.

But if it were desired to measure the influence of type of road, fertility of land, and distance from town on land value, they could not be so readily included in the same additive equation. For example, a 40-acre farm yielding 40 bushels of corn to the acre might be worth on the average $1,000 more than a farm of the same size yielding 30 bushels of corn per acre. Under the same conditions, it would not be reasonable to expect that a 160-acre farm yielding 40 bushels of corn to the acre would be worth only $1,000 more than a 160-acre farm yielding 30 bushels per acre. In the first case, the higher yield would add $25 per acre to the farm value, in the latter, only $6.25. Yet if yield of corn were added as a factor to the above equation, that would assume that a given increase in fertility would add the same amount to the value of the farm, no matter how large or how small the farm was.

If the value were stated as value per acre, that would partly solve the difficulty, for a given change in fertility, distance from town, or type of road would then be assumed to have the same influence upon value per acre no matter how large or how small the farm was. But that would introduce difficulty with other variables. The dwelling, for example, would not become larger in direct proportion to the size of the farm. Very large farms with good dwellings would have a very low "value-of-dwellings-per-acre," and small farms with poor dwellings would also have a low "value-of-dwellings-per-acre." Only some method of determining the effect of value of dwellings on land values separately for farms of different sizes would take care of this difficulty, as otherwise the same measurements would be used for dwelling values which might be different in their effect on land values, with consequent confusion of the results.[2]

Type of equation to be fitted. The case mentioned also illustrates the need of something other than a simple additive regression equation to express certain cases. If it is assumed that the more fertile the farm, the greater the effect of nearness to town would be, and that the nearer to town, the greater the effect of an increase in fertility would be, that could not be adequately expressed by the regression equation

$$\text{Value per acre} = f(\text{distance}) + f(\text{fertility}) + \text{etc.}$$

The multiplying effects of the two variables upon the value could be allowed for by using the equation

$$\text{Value per acre} = [f_1(\text{distance})]\,[f_2(\text{fertility})]\,[f(\text{etc.})]$$

which, for the actual process of computation, can be stated

$$\text{logarithm (value per acre)} \\ = \Phi_1\ (\text{log distance}) + \Phi_2\ (\text{log fertility}) + \Phi_3\ (\text{log etc.})$$

[2] See the appendix, pages 39–54, of *U. S. Dept. Agr. Bul.* 1400, Factors affecting farmers' earnings in southwest Pennsylvania, for an example of statistical treatment of a problem of this type.

This logarithmic equation, which puts the relations on a *relative* or *proportional* rather than an *absolute* or *arithmetic* base, is a very flexible one and one that can be used in a great many types of problems.

Finally, if the effect of fertility upon land value be found to vary with fertility, say, and the effect of building value with size of farm, not even the logarithmic equation would be applicable. Instead, an equation of the joint-function type (note Chapter 21) might be used, such as

$$\begin{aligned}\text{Log (value per acre)} = {} & f(\text{distance, fertility, roads}) \\ & + f(\text{value dwelling, size of farm, value barns}) \\ & + \text{etc.}\end{aligned}$$

One further consideration is the danger of false results or spurious correlation if the variables are improperly stated. Thus if an attempt were made to correlate the value of farms with three factors, (A) the percentage of land in corn, (B) the percentage of land in wheat, and (C) the percentage of land in all other uses, it would be impossible to solve the problem, or else it would give a spurious result. That is because the factors (A), (B), and (C) would add to exactly 100 per cent in each individual case, and after variation in (A) and (B) had been held constant by statistical means, there would not be any room left for variation in (C). Even if as the result of rounding off the variables there were slight deviations from the 100 per cent total, the results would have little significance, as the practically perfect intercorrelation between the three independent factors would make the measures of their net influence, both regression coefficients and net coefficients of correlation, exceedingly subject to error.[3] Only by dropping out one of the factors, say (C), would significant results be secured. The regressions on (A) and (B) would then also show the effect of (C); for example, the increase in value for each unit increase in (A) would mean the increase due to substituting one unit of (A) for one unit of (C); changing the sign would give the effect of substituting one unit of (C) for one of (A). The same principle would then apply as between (B) and (C); whereas the increase in the dependent variable for substituting one unit of (B) for one of (A) would be the difference between the two net regression coefficients.

[3] For an extended mathematical treatment of this problem, see Ragnar Frisch, Statistical confluence analysis by means of complete regression systems, Oslo University Økonomiske Institutt. Publikazion No. 5, (1934).

After the variables to be examined and the nature of the regression function to be used have been decided upon, at least tentatively, it is necessary to decide what type of curves are to be fitted. If mathematical regressions are to be used, this involves deciding what form of equation is to be used. (Note pages 76 to 125 of Chapter 6, and 397 to 401 of Chapter 22.) If curves are to be fitted by one of the graphic methods, limiting conditions to be applied in fitting the curves must be worked out, in the light of the hypotheses stated and of the technological and other knowledge of the relations. (See Chapter 6, pages 109 to 110, Chapter 14, page 224, and Chapter 16, pages 278 and 279.)

Steps in carrying through the computations. After the variables and the form of the equation for the statistical analysis have been decided upon, the next step is actually carrying through the computation. This involves "coding" the numerical values of the variables, that is, reducing them to simpler terms for ease of handling; calculating the extensions; setting up and solving the normal equations; and calculating the standard error of estimate, the coefficient of multiple correlation, and possibly the coefficients of separate determination or of part correlation. Then if curvilinear regressions are desired, the residuals from the linear regression equation will be computed, and the net regression curves determined by successive approximation (or by the graphic short-cut method if the conditions are favorable). After the final curves are determined, the standard error of estimate for the curvilinear regressions and the index of multiple correlation are computed. If joint functions are suspected, the residuals are grouped with respect to two or more variables, or studied with respect to compound variables of the Court type, until by successive approximations the final shape of all the functions, simple or joint, has been determined, and the new standard error of estimate and index of correlation computed. As a final step, the standard error of each of the regression coefficients, or of each portion of each regression curve, should be computed and indicated on the regression charts, to indicate the significance to be attached to the results. The standard error of the correlation coefficient or other constants likewise should be determined. All through the process, the statistical relations found should be checked back against the hypothetical expectations. If the statistical results conflict with the hypothesis, both should be re-examined to see where the conflict lies, as discussed in more detail subsequently.

Meaning of Correlation Results

It must be noted, however, that a statistical determination of the nature of any relation, no matter how complicated the methods used in making the determination or how flexible the type of function allowed for, tells nothing of the *reason* for the relation observed.

Thus the variation in potato yields with differences in early and late rainfall, as determined in Chapter 22, may be due to a large variety of different causes. The plant requires certain conditions of soil moisture, nutrients, sunshine, maximum and minimum temperature, and relative humidity to make the best growth, and the factors used reflect certain of them. Further, it may be that one set of conditions is required during the first part of the growing period while the plant is developing its leaves and top, and another set later on while it is developing the tubers; and that the rainfall factors used relate in this way to the growth periods of the plant.

There are other possibilities, however. The yield of a plant is affected by the weather conditions not only as they directly affect the development of the plant itself but also as they affect the development of insects and diseases that prey on the plant. For example, the peculiar relation of potato yields to early and late rainfall considered jointly, as shown in Figure 74, might reflect the relation of late rainfall to potato diseases. With 16 inches of early rainfall and 3 inches of late, a yield of 240 bushels would be expected; with the same early rainfall, as the late rainfall increases, the probable yield declines until with 6 inches of late rainfall it is under 180. The heavy early rainfall may stimulate good growth of the top; then if heavy late rainfall should follow it might result in conditions favorable to the development of potato blight, and so reduce an otherwise promising yield.

It is evident that a considerable range of specific technical information is necessary to interpret correctly the results of a correlation analysis, and to develop the reasons for the particular relations which have been found to exist. For best results this technical knowledge must be drawn upon in the early stages of the investigation, to aid in selecting and stating the variables to be considered in such a way that the functional relations, when found by appropriate statistical means, would adequately represent the technological elements present and so be capable of a logical technical interpretation. The correlation analysis itself can never provide the interpretation of cause and effect. It can only establish the *facts* of the relations—for the meaning of those facts. the investigator must look elsewhere.

The way in which correlation analysis establishes the facts of relationship and nothing else may be illustrated by a specific example. If the number of automobiles moving down Sixteenth Street in Washington, D. C., for each 15-minute period through a given 12 hours is correlated with the height of the water in the Potomac River during each of the same periods, a definite correlation will be obtained. On some days this correlation would be so high that its probable error would indicate that it would be very unlikely that it could have occurred by chance. However, if on the basis of this correlation one were to attempt to forecast the flow of traffic from the height of the water, he would find his forecast sadly in error if he made it for another day when the street was closed for traffic repairs, when the water was high because of a flood, or when the moon was in a different phase. This is a case in which it is perfectly obvious that there is no direct causal relation between the two phenomena. Yet there is real correlation between them because they both are influenced, though very remotely, by the same sequence of cosmic events. The rising and the setting of the sun have a very definite influence on the movements of persons and therefore on the flow of traffic, whereas the rising and the setting of the moon likewise have a definite influence on the height of the water. Washington is so close to the ocean, and has so low an elevation, that the Potomac River has a definite ebb and flood of tide. There is a certain specific though complex relation between the rising and setting of the sun and of the moon. This relation is changing constantly from day to day. This illustrates a case in which real and significant correlation between two variables reflects causation by a common factor or factors, yet gives no inference as to direct causal connections. Many similar cases are met with in practical work in which the correlation between two variables is due to both being influenced by certain common causes, although neither may in any conceivable way influence the other. This illustrates again the need for clear, logical thinking and for a technological basis for the interpretation of the statistical results, which can measure the relationships but of themselves can tell nothing of cause or effect.

Statement of results of correlation analysis. Having completed the statistical analysis of the relations—the extent and complexity of which will depend upon the nature of the problem, the number of observations available, the importance of the relations, and the facilities available with which to work—the next step is to translate the statistical results to intelligible non-technical statement. This may go only so far as simple regression charts or estimating tables of the type shown at the end of Chapter 13, or of carefully worked-out pictorial statements such as shown in Fig. 75. After the results are reduced to intelligible form—intelligible, that is, at least to the investigator—they should be carefully compared with the original hypothesis. If hypothesis and the statistical results do not agree, the hypothesis must be carefully examined to see if it may logically be restated so as to be consistent with the facts as found; and the analysis must be carefully studied to see if there are any loopholes in the way the facts are stated, or in the way the problem has been worked through, which may be responsible for the results. (The preliminary results cited at the top of page 419, in Chapter 23, are an example of mis-statement of the variables.) If the hypothesis and results are found to be consistent, or if, without doing violence to either, they can be brought into reasonable agreement, the project may be regarded as completed. If such agreement is not obtained, the results may be announced as actual observations inconsistent with what was expected and subject to further study or independent checks before being accepted as scientific conclusions.

Finally, if forecasts of future events or estimates for new observations are to be made from the results of the analysis, the methods outlined in Chapter 19 should be used to help judge how much confidence can be placed in such estimates or forecasts.

When the hypothesis and the analysis are found to be in satisfactory agreement, all that remains is to interpret the results to those who will be interested in them and have to use them. At this point many investigators fail to take into account the audience for which they are writing. If they are writing a technical paper for a scientific journal, a full discussion of the methods and techniques used, statistical and otherwise, will be quite in place, so that their fellows may pass on the adequacy of the work. If, instead, they are writing a general or a popular report for an audience which is only interested in what they have discovered and what it means, details of statistical technique may be as out of place as computations of stresses and strains would be in a magazine devoted to "The Home Beautiful." The plethora of technical terminology in some supposedly popular reports of statistical investigations has led the readers to suspect that the investigator himself did not understand what his results really meant. Unles the conclusions can be translated back into "the King's English," and stated so simply that practical men dealing with the problem investigated can understand what the results mean, the usefulness of the research may be largely wasted.

Summary. The place of statistical analysis in scientific research is no different from the place of any other technical aid the investigator may employ. It furnishes a means of measuring the elements that are involved and of examining the way in which they are related; but it does not of itself furnish an explanation of phenomena. Except insofar as the effort to reduce the variables to specific numerical statement, definitely related, forces the investigator to think more clearly and definitely about his problem, statistical analysis is not a substitute for logical analysis, clear-cut thinking, and full knowledge of a problem. The methods of analyzing complicated relations set forth in this book furnish the student keen tools for investigating complex relationships; but, like all keen tools, they may yield unsatisfactory or misleading results if employed carelessly or heedlessly. Statistical analysis is not a substitute for careful thinking and skilled workmanship in research work; instead, it is an aid which may make that thought and skill even more productive of worth-while results.

APPENDIX 1

METHODS OF COMPUTATION

Coefficients of correlation and regression. Many of the operations described in the text may be performed much more rapidly by short-cut methods. One such method, for computing σ_x has already been given. (Note 1, Chapter 1.) When the value Σx^2, required in the normal equations, is desired instead, that may be calculated from a frequency table by the same method, by use of the relation

$$\Sigma x^2 = \Sigma(d^2F) - n\left[\frac{\Sigma(dF)}{n}\right]^2 \qquad (100)$$

A similar short cut may be used in computing the product sums, Σxy, required to determine coefficients of regression or correlation. The first step is to construct a double-frequency table. Such a table is known as a correlation table, since it shows the nature of the relation between the two variables in much the same way that a correlation chart or dot chart does. The following correlation table, Table 85, is prepared from the haystack data used as an illustration in Chapter 21.

First the number of items falling in each subgroup is determined. Then the entries in each row are summed, giving the total frequencies with respect to X_1. These frequencies, denoted F_1, are shown at the right in the table, in the column headed "All values of X_2." The entries in each column are similarly added, and the totals entered at the foot. These entries give the frequency distribution with respect to X_2, and are, therefore, denoted F_2 (frequencies of X_2). A central group is then selected for each variable, and the departures of the other groups, above or below that central group, are shown in the "d_1" and "d_2" column and line, respectively. The usual extensions to compute the Σx^2 for each variable are shown in the final columns, d_1F_1 and $d_1^2F_1$, and lines d_2F_2 and $d_2^2F_2$. As their designations indicate, the entries under these heads are obtained by first multiplying the F_1 entry by d_1, giving d_1F_1, and then multiplying that again by d_1 to give $d_1^2F_1$. The similar computations for F_2 are shown at the foot of the table.

The new step in the table is the incorporation of the column Σd_2F_1 and of the line Σd_1F. The entries in the column Σd_2F show for each line the sums of the frequencies of each cell in that line multiplied by the d_2 values for each cell. Thus for the first line, the single entry has a d_2 value of -3, so the entry in Σd_2F is -3. The next line similarly has a single entry in the -2 column. The fourth line, though, has the following frequencies: 1 in the -4 column, 1 in -3, 2 in -2, 8 in -1, 4 in 0, and 1 in 1. The respective products, -4, -3, -4, -8, 0, and 1, add to -18, and this value is, therefore, entered in the Σd_2F column.

The Σd_1F line is similarly computed, showing the sum of the frequencies in each cell for each column, multiplied by the corresponding d_1 values. Thus the first column has 1 entry in the -1 line; the second, 1 in the -4, and 1 in the -1 with the sum -5. The third column has 1 in -3, 1 in -2, 2 in -1, 6 in 0, and 3 in 1.

TABLE 85

Correlation Table—Reports Classified With Respect to Both X_1 and X_2; and Extensions and Coefficients

Values of X_1	Values of X_2									All values X_2 F_1	d_1	Σd_2F	$d_1(\Sigma d_2F)$	d_1F_1	$d_1^2F_1$
	0.045–0.064	0.065–0.084	0.085–0.104	0.105–0.124	0.125–0.144	0.145–0.164	0.165–0.184	0.185–0.204	0.205–0.224						
0.020–0.099	......	1	......	......	......	......	......	......	......	1	−4	−3	12	−4	16
0.100–0.179	......	......	1	......	......	......	......	......	......	1	−3	−2	6	−3	9
0.180–0.259	......	......	1	......	1	......	......	......	......	2	−2	−2	4	−4	8
0.260–0.339	1	1	2	8	4	1	......	......	......	17	−1	−18	18	−17	17
0.340–0.419	......	......	6	9	5	3	1	......	......	24	0	−16	0	0	0
0.420–0.499	......	......	3	10	10	8	2	1	......	34	1	−1	−1	34	34
0.500–0.579	......	......	......	1	3	10	4	1	......	19	2	20	40	38	76
0.580–0.659	......	......	......	......	1	5	6	3	......	15	3	26	78	45	135
0.660–0.719	......	......	......	......	......	......	......	3	1	4	4	13	52	16	64
0.720–0.799	......	......	......	......	......	1	......	2	......	3	5	7	35	15	75
All values X_1 F_2	1	2	13	28	24	28	13	10	1	120	Σ	24	244	120	434
d_2	−4	−3	−2	−1	0	1	2	3	4	Σ					
Σd_1F	−1	−5	−4	4	13	47	28	34	4	120					
$d_2(\Sigma d_1F)$	4	15	8	−4	0	47	56	102	16	244					
d_2F_2	−4	−6	−26	−28	0	28	26	30	4	24					
$d_2^2F_2$	16	18	52	28	0	28	52	90	16	300					

Calculation of Corrections to Deviations from Mean

	Σd_1F	Σd_2F	$\Sigma d_1^2F_1$	$\Sigma d_2^2F_2$	Σd_1d_2F
Sums..............	120	24	434	300	244
Means..............	1.00	0.20			
Corrections........	......	......	−120	−4.8	−24
Corrected values....	......	......	314	295.2	220

$$b_{12} = \frac{\Sigma x_1x_2}{\Sigma x_2^2} = \frac{220}{314} = 0.701$$

$$r_{12} = \frac{\Sigma x_1x_2}{\sqrt{(\Sigma x_2^2)(\Sigma x_1^2)}} = \frac{220}{\sqrt{(314)(295.2)}} = 0.722$$

The products -3, -2, -2, 0, and 3 add to -4; and this is the value for the Σd_1F entry for that column.

After all the Σd_2F entries are made, each is multiplied by the d_1 value for the same line, giving the values entered in the $d_1(\Sigma d_2F)$ column. Similarly, the entries in the Σd_1F line are then multiplied by the corresponding d_2 values, and the products entered in the $d_2(\Sigma d_1F)$ line. Each line at the foot of the table is then summed, and the sums entered at the right of the line; and each column at the right of the table summed, and the sum entered at the foot of the column.

The arithmetic may now be checked by the following identities:

$$\begin{aligned}\Sigma F_1 &= \Sigma F_2 \text{ (=120, in the illustration)}\\ \Sigma d_2F &= \Sigma d_2F_2 \text{ (=24)}\\ \Sigma d_1F &= \Sigma d_1F_1 \text{ (=120)}\\ \Sigma d_1(\Sigma d_2F) &= \Sigma d_2(\Sigma d_1F) \text{ (=244)}\end{aligned}$$

As a matter of fact, the lines Σd_1F and $d_2(\Sigma d_1F)$ may be omitted, if this check is not to be made. Where many items are involved, however, this is a rapid and accurate check. Only the extensions $d_2^2F_2$ and $d_1^2F_1$ are then not checked, and these may be verified readily by recomputing.

The next step is to adjust the values in terms of departures from the assumed means to terms of departures from the true means: To do this in organized fashion, the five sums are entered in order, as shown at the foot of the tabulation. The first two items, Σd_1F and Σd_2F, are each divided by the number of cases (120) to give the average departure (in terms of class intervals) from the assumed mean group, that is, values for M_{d_1} and M_{d_2}.

The correction factors are then computed as follows:

$$\begin{aligned}\Sigma x_1^2 &= \Sigma d_1^2F_1 - (\Sigma d_1F)(M_{d_1}) = 434 - (120)(1.0) = 314\\ \Sigma x_2^2 &= \Sigma d_2^2F_2 - (\Sigma d_2F)(M_{d_2}) = 300 - (24)(0.20) = 295.2\\ \Sigma x_1x_2 &= \Sigma d_1d_2F - (\Sigma d_1F)(M_{d_2}) = 244 - (120)(0.20) = 220\end{aligned}$$

This process is shown in tabular form for each item.

The coefficients of regression and of correlation are then computed by the usual formulas, as shown at the foot of the table.

The coefficient of regression shown in the tabulation is of course in terms of group units. That is, the value $b_{12} = 0.7$ means that for every change of one group interval in X_2 there is on the average a change of 0.7 group intervals in X_1. Since the group intervals are 0.020 for X_2, and 0.080 for X_1, this does not apply to the actual X_1 and X_2 values. Instead, correction must be made as follows:

Regression of X_1 on X_2 in terms of original units = regression in terms of group units $\left(\frac{\text{Group interval of } X_1}{\text{Group interval of } X_2}\right)$

In this case,

$$b_{12} \text{ (for } X_1 \text{ and } X_2) = 0.7006\,\frac{0.080}{0.020} = 2.8024$$

Hence for each change of 1 unit in X_2, X_1 changes 2.8 units, on the average.

Before a_{12} can be calculated, it is necessary to have the means of X_1 and X_2. The assumed mean of X_1, at the midpoint of the group 0.340–0.419, would be at 0.380. Since the mean of d_1 is 1.00, the true mean lies exactly one group interval, or 0.080, higher than this, or at 0.460. Similarly, the assumed mean of X_2, at the midpoint of the group $0.125 - 0.144$, is at 0.135. Adding to this the mean of d_2, or 0.20, times the group interval of 0.020, gives 0.139 as the mean of X_1. The value of a_{12} may now be calculated by formula (10):

$$\begin{aligned} a &= M_1 - b_{12}M_2 \\ &= 0.139 - (2.8024)(0.0460) \\ &= -\,1.150 \end{aligned}$$

From the values shown in the table, the regression equation is, therefore, found to be

$$X_1 = -\,1.150 + 2.8024\,X_2$$

The uncorrected correlation coefficient, r_{12}, has been found to be 0.722, as shown in the table. In making this computation, however, no allowance has been made for the tendency of grouping to exaggerate the departures of the individual cases from the means, which affects Σx_1^2 and Σx_2^2, but does not affect Σx_1x_2. This may be allowed for by applying Sheppard's correction to Σx_2^2 and Σx_3^2. (See Note 1, Chapter 1.) Since the correction is (corrected σ^2) $= \sigma^2 - \dfrac{c^2}{12}$, the correction for $\Sigma x_1^2 = \Sigma x_1^2 - \dfrac{nc^2}{12}$. If we apply this correction to both Σx_1^2 and Σx_2^2, in the formula for r_{12}, the value of r_{12} comes out 0.747, a definitely higher value. In this case only 9 groups have been used for X_2 and only 10 for X_1, so the correction for grouping is important. In most practical work, at least 20 to 30 groups should be used for each variable; and when that is done, the application of the correction for the fineness of grouping becomes of much less significance.

Although the correlation computed with Sheppard's corrections (and adjusted for the number of observations by equation [25]) gives the best estimate of the true correlation in the universe from which the sample was drawn, the formulas for the standard error of correlation coefficients are all based on the uncorrected formula. If any test of the significance of the observed correlation, such as Fisher's z-transformation, is to be applied, the unadjusted value should be used.

The regression coefficient, as well as the coefficient of correlation, is changed slightly if Sheppard's correction is applied. Thus, using the correction,

$$b_{12} = \frac{\Sigma x_1x_2}{\Sigma x_2^2 - \dfrac{nc^2}{12}} = \frac{220}{314 - \dfrac{120}{12}} = 0.724$$

Just as with the correlation coefficient, the larger the number of groups, the less influence the correction has on the calculated values. With 30 or more groups, it is ordinarily neglected.

There are many other ways in which the coefficient of correlation may be calculated. Thus it may be shown (Note 1, Appendix 2) that if

$$X_3 = X_1 - X_2$$

the standard deviations are related according to the formula

$$\sigma_3^2 = \sigma_1^2 - 2r_{12}\sigma_1\sigma_2 + \sigma_2^2$$

Hence the correlation between X_1 and X_2 may be found by calculating the difference, X_3, between each pair of values for the two variables, and computing the standard deviation of each of the three series. The correlation is then given directly by the formula

$$r_{12} = \frac{\sigma_1^2 + \sigma_2^2 - \sigma_3^2}{2\sigma_1\sigma_2}$$

Coefficients of multiple correlation and net regression. When many variables are to be considered, and a large number of observations are available, the necessary extensions for multiple correlation lines or curves fitted by the least-squares method may be made most readily by the use of tabulation cards, one for each observation, with the values for each variable entered on each card. If mechanical or electrical tabulation equipment is available, the values may be designated by punched holes. The cards can then be sorted and tabulated by automatic machines.

TABLE 86

Computation of Extensions by "Digiting" and Accumulative Totaling

Value of X_1	Number of items (1)	Accumulations of X_1 (2)	$\Sigma(X_1)$ (3)	Accumulations of $(X_1)^2$ (4)	$\Sigma(X_2)$ (5)	Accumulations of X_1X_2 (6)	$\Sigma(X_3)$ (7)	Accumulations of X_1X_3 (8)	$\Sigma(X_4)$ (9)	Accumulations of X_1X_4 (10)
30–39	2	20	69	690	11	110	1	10	5	50
20–29	0	20	0	690	0	110	0	10	0	50
10–19	20	220	295	3,640	32	430	18	190	26	310
0– 9	10	(320)	66	(4,300)	14	(570)	16	(350)	23	(540)
−9	3	3	47	47	2	2	5	5	2	2
−8	2	5	28	75	3	5	7	12	3	5
−7	0	5	0	75	0	5	0	12	0	5
−6	1	6	21	96	12	17	2	14	3	8
−5	13	19	183	279	20	37	8	22	14	22
−4	5	24	70	349	6	43	3	25	7	29
−3	2	26	11	360	7	50	1	26	3	32
−2	1	27	7	367	2	52	2	28	2	34
−1	3	30	41	408	2	54	5	33	14	48
−0	2	(32)	22	(430)	3	(57)	2	(35)	6	(54)
Sums	$\Sigma X_1 = 405$		$\Sigma(X_1)^2 = 7076$		$\Sigma X_1X_2 = 915$		$\Sigma X_1X_3 = 387$		$\Sigma X_1X_4 = 595$	

The most rapid method of calculating the extensions, where card-tabulating equipment is available, is by a combination of the "digiting" method with cumulative addition. Thus if four variables are being considered, the extensions for ΣX_1^2, ΣX_1X_2, ΣX_1X_3, and ΣX_1X_4 would be secured by sorting on X_1. If each variable were tabulated in two digits (0 to 99) the cards would first be classified in ten groups from 00 to 90 on the tens column of X_1 and the total for each variable computed. The cards would then be reclassified into ten groups from 0 to 9 according to the values in the digit column, and the total for each variable computed. The totals would then be entered as shown in Table 86, starting with the highest value and running down to the smallest.

After the number of items and the sums for each group are entered (columns 1, 3, 5, 7, and 9) in the table the columns headed "accumulations" are computed as follows: The first item, times 10, is entered in the top line. Thus the first item in column 1 is 2, so 20 is entered in column 2. The second item in each column is then multiplied by 10, added to the first item in the accumulation column, and the total entered on the second line of the accumulation column. (Since the second item in column 1 is 0, the second item in column 2 is 20 + (10)(0) = 20.) The third item in each odd-numbered column is next multiplied by 10, added to the second item in each adjoining accumulation column, and the total entered on the third line of each accumulation column. (The third item in column 1 is 20, so the third item in column 2 is 20 + (10)(20) = 220.) The same operation is performed for the next line. When this process is completed for all the classifications of X_1 by tens, it is begun afresh for the classifications by digits, without multiplying by 10. The item in the "-9" class, 3, of column 1, is entered in the adjoining accumulation column. The next item, 2, is added to it, and the total, 5, entered in the accumulation column, and so on down the column, entering in column 2 the accumulated total of column 1. The same operation is performed in each of the other accumulation columns, each showing the accumulated total for the column to its left.

In each case the accumulated total for the -0 group is one-tenth that for the $0 - 9$ group, and is equal to the sum of the particular variable, checking all the computations. That is because each value appears twice: once when the observations are classified according to the first digit of X_1 (in the tens column), and once when they are sorted according to the last digit of X_1 (in the units column). If there were also a hundreds column, there would be a third sort for that, and the accumulative totals for the hundreds groups, with 00 added at each step, would be 100 times as large as for the unit groups. After the entries in the -0 line have been checked against those in the $0 - 9$ line, both are enclosed in parentheses. All the entries in each accumulation column are then added, except those enclosed in parentheses. The totals for each column are then the extensions for ΣX_1, ΣX_1^2, ΣX_1X_2, etc.[1]

By the use of this method, each variable can be carried to 3, 4, or even more digits if desired, yet the extensions be obtained with exactly the same precision as if each individual item were extended separately. The work is very greatly reduced; the extensions, even if 3 digits were used for each variable, requiring at the most only 30 lines, or for 4 digits, 40 lines.

[1] It is easily seen why this is so. Each item in the 30+ group appears 3 times in the accumulation column, times 10 each time; it, therefore, contributes 30 to the total. Likewise, each item in the -9 column appears 9 times in the accumulative totals from -9 to -1, so contributes 9 to that total. An item of $X_1 = 39$ is represented by 1 in the 30–39 class in column 1, and by 39 in the same class in column 3. It also appears as 1 in the -9 class in column 1, and as 39 in the same class in column 3. Its contribution to ΣX_1 is then (3)(10) + (1)(9), or 39, and to $\Sigma(X_1^2)$ is (39)(3)(10) + (39)(1)(9) = 1,521, or exactly equal to 39 and $(39)^2$. Similarly, an item of $X_2 = 2$ when $X_1 = 39$, appears in column 5 in the 30–39 line, and in the -9 line. It then appears 3 times in column 6, multiplied by 10 each time, and 9 times in column 6, multiplied by 1. Its contribution to $\Sigma(X_1X_2)$ is then (2)(10)(3) + (2)(9) = 78, or exactly equal to (39)(2). It is now evident why the entries in the 0 lines are not included in the total—they contribute nothing to the product with X_1.

After the extensions with respect to X_1 had been made as just shown, the cards would be reclassified with respect to X_2, and a similar tabulation and accumulative totals prepared to obtain ΣX_2, ΣX_2^2, ΣX_2X_3, and ΣX_2X_4; and so on for the other extensions required.

This method is of the greatest value where automatic card-tabulation equipment is available, and a large number of observations are to be treated. Even for hand operations however, if the number of observations are very large it can be used to advantage, with the individual items entered on cards or strips for handy classifying and adding.

Use of the check sum. Where a number of different variables are involved, every operation in making the extensions, computing the averages and corrections, and solving the normal equations through to the "back solution," can be verified by an automatic check known as the "check sum." The way in which the check sum is used will be illustrated by a small problem, carried through every step in turn, but it is equally applicable to any other method of tabulation and is especially valuable with machine tabulation, where it serves as an overall control on the accuracy of the machine processes.

The check sum as a check in extending. The values in the following table (Table 87) may be used to illustrate the use of the check sum.

The values in the columns X_2, X_3, X_4, and X_1 are the three independent factors and the dependent factor, which are to be correlated. The values in the column headed "ΣX" are the arithmetic totals of the values for the four other variables, and are designated "the check sum."

As the first step, each of these five columns is added. Since, for each line,

$$X_2 + X_3 + X_4 + X_1 = \Sigma X$$

it also holds true that

$$\Sigma X_2 + \Sigma X_3 + \Sigma X_4 + \Sigma X_1 = \Sigma(\Sigma X).$$

Adding the sums of the first four columns together gives the same value as the sum of the check sum column, which verifies all the totals.

The first set of extensions is made by multiplying the items in each line by the X_2 item in the first column of that line, giving the values shown under "Extensions with X_2." Since for each line

$$X_2 + X_3 + X_4 + X_1 = \Sigma X,$$

it also follows that

$$X_2^2 + X_2X_3 + X_2X_4 + X_2X_1 = X_2\Sigma X.$$

Then, adding each column, we find that the sums of the four other columns should total to the same value as the sum of the $X_2\Sigma X$ column. Checking up, we see that 1,994 + 25,315 + 14,158 + 14,224 = 55,691, verifying all the calculations.

The other extensions are made in similar fashion, and the sums of each column verified with the sum of the check-sum column, according to the relation

$$\Sigma X_2X_3 + \Sigma X_3^2 + \Sigma X_3X_4 + \Sigma X_3X_1 = \Sigma(X_3\Sigma X)$$

and the corresponding relation for the other extensions. It should be noted that in checking the "extensions with X_3," the value ΣX_2X_3 is taken from the previous set of extensions; in checking the "extensions with X_4," the value for ΣX_2X_4 is taken from the "extensions with X_2," and the value for ΣX_3X_4 from the "extensions with X_4"; and so on for the remaining checks.

TABLE 87

Calculation of Extensions, Using the Check Sum

Variables					Extensions with X_2				
X_2	X_3	X_4	X_1	ΣX	X_2^2	X_2X_3	X_2X_4	X_2X_1	$X_2\Sigma X$
0	136	106	103	345	0	0	0	0	0
1	140	103	108	352	1	140	103	108	352
2	86	108	102	298	4	172	216	204	596
3	115	102	111	331	9	345	306	333	993
4	115	111	95	325	16	460	444	380	1,300
12	161	91	109	373	144	1,932	1,092	1,308	4,476
13	235	109	118	475	169	3,055	1,417	1,534	6,175
14	304	118	123	559	196	4,256	1,652	1,722	7,826
15	224	123	108	470	225	3,360	1,845	1,620	7,050
16	185	108	100	409	256	2,960	1,728	1,600	6,544
17	108	100	88	313	289	1,836	1,700	1,496	5,321
18	193	88	109	408	324	3,474	1,584	1,962	7,344
19	175	109	103	406	361	3,325	2,071	1,957	7,714
134	2177	1376	1377	5064	1994	25,315	14,158	14,224	55,691

Extensions with X_3				Extensions with X_4			Extensions with X_1	
X_3^2	X_3X_4	X_3X_1	$X_3\Sigma X$	X_4^2	X_4X_1	$X_4\Sigma X$	X_1^2	$X_1\Sigma X$
18,496	14,416	14,008	46,920	11,236	10,918	36,570	10,609	35,535
19,600	14,420	15,120	49,280	10,609	11,124	36,256	11,664	38,016
7,396	9,288	8,772	25,628	11,664	11,016	32,184	10,404	30,396
13,225	11,730	12,765	38,065	10,404	11,322	33,762	12,321	36,741
13,225	12,765	10,925	37,375	12,321	10,545	36,075	9,025	30,875
25,921	14,651	17,549	60,053	8,281	9,919	33,943	11,881	40,657
55,225	25,615	27,730	111,625	11,881	12,862	51,775	13,924	56,050
92,416	35,872	37,392	169,936	13,924	14,514	65,962	15,129	68,757
50,176	27,552	24,192	105,280	15,129	13,284	57,810	11,664	50,760
34,225	19,980	18,500	75,665	11,664	10,800	44,172	10,000	40,900
11,664	10,800	9,504	33,804	10,000	8,800	31,300	7,744	27,544
37,249	16,984	21,037	78,744	7,744	9,592	35,904	11,881	44,472
30,625	19,075	18,025	71,050	11,881	11,227	44,254	10,609	41,818
409,443	233,148	235,519	903,425	146,738	145,923	539,967	146,855	542,521

While the check sum would not disclose exactly compensating errors made in different columns the possibility of such errors is so remote that, after the arithmetic has been checked by the comparisons indicated, it may be assumed that no errors have been made either in making the multiplications or adding the columns.

The check sum as a check in correcting to the means. After the values for ΣX_2^2, ΣX_2X_3, etc., have all been computed as indicated in Table 87, the process of making the corrections to get the values Σx_2^2, Σx_2x_3, etc., may be organized in regular fashion and checked by the check sum, as shown in Table 88.

TABLE 88

CALCULATION OF PRODUCT SUMS CORRECTED TO DEPARTURES FROM MEANS, WITH CHECK SUM

	X_2	X_3	X_4	X_1	ΣX	Line
Sums............	134.	2,177.	1,376.	1,377.	5,064.	1
Means............	10.30769	167.46154	105.84615	105.92308	389.53846	2
Extensions with X_2....	1,994.00	25,315.00	14,158.00	14,224.00	55,691.00	3
Corrections..........	1,381.23	22,439.84	14,183.38	14,193.69	52,198.14	4
Extensions with x_2.....	612.77	2,875.16	−25.38	30.31	3,492.86	5
Extensions with X_3....	....	409,443.00	233,148.00	235,519.00	903,425.00	6
Corrections..........	....	364,563.77	230,427.08	230,594.54	848,025.23	7
Extensions with x_3.....	....	44,879.23	2,720.92	4,924.46	55,399.77	8
Extensions with X_4....	....	....	146,738.00	145,923.00	539,967.00	9
Corrections..........	....	....	145,644.30	145,750.15	536,004.90	10
Extensions with x_4.....	....	....	1,093.70	172.85	3,962.10	11
Extensions with X_1....	....	....	....	146,855.00	542,521.00	12
Corrections..........	....	....	....	145,856.08	536,394.48	13
Extensions with x_1.....	....	....	....	998.92	6,126.52	14

The first line in Table 88 gives the sums of each of the variables, including the check sum. Dividing by the number of observations (13 in this case), gives the mean for each variable, as entered in the second line. Again, the entries for the first four columns total to equal the entry in the Σ column, checking the division.

The sums from the "extensions with X_2," of Table 87, are next entered in line 3. The sums for each variable, in line 1, are next multiplied by the mean of X_2 (10.30769), and the products entered in the corresponding column in line 4. Subtracting the entries in line 4 from those in line 3 gives the values which are entered in line 5. These values are the extensions, expressed as departures from the means.

In column X_3, for example, the entry in line 3 is ΣX_2X_3; and the entry in line 4 is ΣX_3M_2.

The entry in line 5, then, is $\Sigma X_2X_3 - \Sigma X_3M_2$

$$= \Sigma X_2X_3 - nM_3M_2$$

$$= \Sigma x_2x_3$$

Again, the values in the first four columns add to the same as the value in the check-sum column, verifying the work.

The rest of the table is entered in similar fashion. Lines 6, 9, and 12 are the extensions with X_3, X_4, and X_1, from Table 87. Lines 7, 10, and 13 are the values in the corresponding columns of line 1, multiplied by M_3, M_4, and M_1, respectively (from line 2). Lines 8, 11, and 14, obtained by subtracting the items in lines 7, 10, and 13 from those in 6, 9, and 12, show the values corrected for departures from the means.

In verifying the sum of the other entries in line 8 by the check sum, the item $\Sigma x_2 x_3$ must be included, from column X_3, line 5, before comparing with the check sum; in checking line 11, $\Sigma x_2 x_4$ and $\Sigma x_3 x_4$, from column X_4, lines 5 and 8, must be included; and in checking line 14, the values $\Sigma x_1 x_2$, $\Sigma x_1 x_3$, and $\Sigma x_1 x_4$, from column X_1, lines 5, 8, and 11, must all be included. For line 11, the other items add to 3,962.11, as against the check sum of 3,962.10; and for line 14, they add to 6,126.55, as against the check sum of 6,126.52. In both cases the discrepancies are so small as to be readily due to raising and lowering in the last digit, and, therefore, may be disregarded.

TABLE 89

SOLUTION OF NORMAL EQUATIONS BY THE DOOLITTLE METHOD, WITH CHECK SUM

Line	X_2	X_3	X_4	X_1	ΣX
I	612.77	2,875.16	−25.38	30.31	3,492.86
I′	−1.00000	−4.69207	0.04142	−0.04946	−5.70011
II	(2,875.16)	44,879.23	2,720.92	4,924.46	55,399.77
(−4.69206)(I)	(−2,875.16)	−13,490.45	119.08	−142.22	−16,388.74
Σ_2	....	31,388.78	2,840.00	4,782.24	39,011.03
II′	....	−1.00000	−0.09048	−0.15236	−1.24283
III	(−25.38)	(2,720.92)	1,093.70	172.85	3,962.10
(0.04142) (I)	(25.38)	(119.08)	−1.05	1.26	144.67
(−0.09048)(Σ_2)	....	(−2,840.00)	−256.96	−432.70	−3,529.72
Σ_3	....	....	835.69	−258.59	577.05
III′	....	....	−1.00000	0.30944	−0.69056
	BACK SOLUTION				
	$b_{12\cdot34}$	$b_{13\cdot24}$	$b_{14\cdot23}$		
	0.04946	0.15236	−0.30944		
	−0.01282	0.02800			
	−0.84626		−0.30944		
		0.18036			
	−0.80962				

Lines 5, 8, 11, and 14 now give the values required to determine the regression coefficients by simultaneous solution, according to equations (38).

The check sum as a check in solving the normal equations. The solution of the simultaneous equations by the Doolittle method has already been illustrated in Chapter 12, page 200. The check sum may be used to verify each step in the computation, as shown in Table 89.

The values from line 5 of Table 88, including the check sum, are entered as line I of Table 89. Each item is divided by the first item of the line, with its sign changed (-612.77). The quotients are entered as line I′. The sum of the first four items checks to the value in the last column, the check sum.

The values from line 8 of Table 88 are entered as line II of Table 89, beginning with column X_3 (the values enclosed in parentheses will be explained later). Line I is next multiplied by the value in column X_3 of line I′ (-4.69207), and the products entered in the corresponding columns below line II. These two lines are then summed, giving line Σ_2. These operations are now verified by adding the items of line Σ_2 in columns X_3, X_4, and X_1, and comparing the sum with the check sum in column ΣX. The three values add to 39,011.02, agreeing to 0.01 with the check sum, 39,011.03.

The values in line Σ_2 are next divided by the value in column X_3, with its sign changed ($-31{,}388.78$). The quotients are entered as line II′. Again the check sum verifies the computation.

The values from line 11, Table 88, are then entered as line III, beginning with column X_4. (Again disregard the figures in parentheses.) Line I is multiplied by the value in column X_4 of line I′ (0.04142), and the products entered in the corresponding columns below line III; and line Σ_2 is multiplied by the value in column X_4 of line II′, and the products entered in the corresponding columns in the next line. Line III and the two following lines are then summed, giving line Σ_3. The values in line Σ_3 are divided by the value in column X_4 of that line, *with its sign changed.* The quotients are entered as line III′. Again the check sum verifies the work. The values in line Σ_3 (before the check sum) add to 577.10, which agrees to 0.05 with the check sum of 577.05.

The values in lines I′, II′, and III′ of column X_1, *with the signs changed*, are then entered at the foot of columns X_2, X_3, and X_4 (designated here $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$). The value at the foot of the X_4 column, -0.30944, is the value for $b_{14.23}$. The item in column X_4, line I′ (0.04142), is then multiplied by the last of these values (-0.30944), and the product (-0.01282) entered in the X_2 column; and the item in column X_4, line II′ (-0.09048), is also multiplied by -0.30944, and the product entered in the X_3 column. The two entries at the foot of the X_3 column are then added, giving 0.18036 as the value for $b_{13.24}$. The item in column X_3, line I′ (-4.69207), is then multiplied by 0.18036, and the product (-0.84626) entered below the other two entries at the foot of the X_2 column. The sum of these three entries, -0.80962, is then the value for $b_{12.34}$.

The way the check sum works in checking the operations may be seen by filling in the missing spaces in Table 89, as indicated by the entries enclosed in parentheses. Thus in line II, the first item, 2,875.16, is the same item, Σx_2x_3, as appears in line I, column X_3. If when line I had been multiplied by -4.69207, the operation had included the X_2 column also, the product would have been $-2{,}875.16$, or exactly the same as the item, in line I, column X_3, with the sign changed. This value, entered below line II in column X_2, exactly cancels the previous value when the two lines are added, leaving line Σ_2 still the same.

Similarly, the values -25.38 and 2,720.92, from lines I and II of column X_4, may be entered in parentheses, in columns X_2 and X_3 of line III. If the previous operations had been carried out in full, below them would appear 25.38 in column X_2 (column X_4, line I, times -1), and 119.08 and $-2{,}840.00$ ([column X_4 line I] [-4.69206] and column X_4, line Σ_2, times -1). When the three lines are totaled to give line Σ_3, the items exactly cancel out, as before.

It should be noted that when all the items are entered in each line, including those in parentheses, the sum of the items in columns X_2 to X_1 exactly equals, line by line, the item in column ΣX. For that reason, if any error is found when one of the Σ lines is reached, the line in which the error occurred can be determined by adding the items line by line, and verifying the totals against the individual check sums. To do this it is not necessary to enter the missing items, as has been done in Table 89 (in parentheses); instead, the items left out can be picked out by going up the columns for the particular variable concerned. Thus all the missing terms for line III (extensions for X_4) and the next two lines appear above in the X_4 column. Once the location of the missing items in the previous work has been learned, they can be used to verify the computations line by line, and any error readily located.

The "back solution" is simply the solution, in regular form, of lines III′, II′ and I′ for b_4, b_3, and b_2. Thus line III′, if written out, is

$$-b_4 = 0.30944$$

Hence $b_4 = -0.30944$, the value at the foot of column X_4. Similarly, line II′, written out, becomes

$$-b_3 - 0.09048b_4 = -0.15236$$

Substituting the above value for b_4, and rearranging,

$$\begin{aligned} b_3 &= 0.15236 - (0.09048)(-0.30944) \\ &= 0.15236 + 0.02800 \end{aligned}$$

These last two values are the same as shown at the foot of column X_3, hence $b_3 = 0.18036$.

Similarly line I′, when written out in full,

$$-b_2 - 4.69207b_3 + 0.04142b_4 = -0.04946$$

Substituting values for b_3 and b_4, and rearranging,

$$\begin{aligned} b_2 &= 0.04946 + (0.04142)(-0.30944) + (-4.69207)(0.18036) \\ &= 0.04946 - 0.01282 - 0.84626 = -0.80962 \end{aligned}$$

exactly as shown at the foot of column X_2.

Having computed the values of the three regression coefficients, the final steps are (*a*) to check those values by substituting them in the *last* equation (line III, in full); (*b*) to compute the coefficient of multiple correlation; and (*c*) to compute the constant $a_{1.234}$ for the regression equation. These steps are all shown in Table 90.

The first operation in Table 90 is the final checking of the entire solution, including the back solution. This is done by substituting the values found for the b's in the *last* equation of the normal equations. For this problem, that equation is:

$$\Sigma(x_2x_4)b_2 + \Sigma(x_3x_4)b_3 + \Sigma(x_4^2)b_4 = \Sigma x_1x_4$$

The values of the 3 b's are entered in column 1 of the table, and the values of the corresponding coefficients of the unknowns, such as $\Sigma(x_2x_4)$ etc., are entered in column 2. The product of each b with its coefficient is then computed, and entered in column 3. These add to 172.87, checking satisfactorily with the value of $\Sigma(x_1x_4)$, 172.85, as shown at the foot of column 2.

The computation of the coefficient of multiple correlation, according to equation (46):

$$R^2_{1.234} = \frac{b_2(\Sigma x_1x_2) + b_3(\Sigma x_1x_3) + b_4(\Sigma x_1x_4)}{\Sigma(x_1^2)}$$

is shown in tabular form in columns 4 and 5.

TABLE 90

FINAL STEPS IN SOLUTION OF MULTIPLE CORRELATION PROBLEM

Variable	Regression coefficient (1)	Equation III (2)	Check (3)	Equation X_1 (4)	Computation of R^2 (5)	Means (6)	Computation of a (7)
X_2	−0.80962	−25.38	20.55	30.31	−24.54	10.308	− 8.35
X_3	0.18036	2,720.92	490.75	4,924.46	888.18	167.462	30.20
X_4	−0.30944	1,093.70	−338.43	172.85	−53.49	105.846	−32.75
Sums....	....	172.85	172.87	998.92	810.15	105.92	−10.90

The values (Σx_1x_2), etc., as shown in Table 88, lines 5, 8, and 11 of column X_1, and $\Sigma(x_1^2)$, shown in line 14, are entered in column 4 of Table 90. Each product sum is multiplied by the corresponding b, shown in column 1, and the products entered in column 5. The sum of these products is then the numerator of the fraction in equation (46). The computation is then readily completed:

$$R^2_{1.234} = \frac{810.15}{998.92} = 0.8110$$

$$R = 0.9006. \quad \text{With } n = 13, \text{ and } m = 4,$$

$$\bar{R}^2 = 1 - (1 - 0.811)\tfrac{12}{9} = 0.784, \text{ and } \bar{R} = 0.86$$

The standard error of estimate may also be readily computed

$$\Sigma(x_1^2) = n\sigma_1^2 = 998.92$$

$$\Sigma[(x_1')^2] = n\sigma^2_{x_1'} = \underline{810.15}$$

then since $n\sigma_1^2 - n\sigma^2_{x_1'} = n\sigma_z^2$

$$\Sigma z^2 = n\sigma_z^2 = 188.77$$

Since there are 13 cases and 3 independent variables,

$$\bar{S}^2_{1.234} = \frac{n\sigma_z^2}{n - m} = \frac{188.77}{9} = 20.97$$

and

$$\bar{S}_{1.234} = 4.58$$

The a for the regression equation is next computed. Using equation (39),

$$a_{1.234} = M_1 - b_2M_2 - b_3M_3 - b_4M_4$$

we may arrange the work in tabular order as shown in columns 6 and 7 of Table 90. The means, from line 2 of Table 88, are entered in column 6, then multiplied by their respective b's, and the products entered in column 7. To complete the computation, following equation (39), the sum of this column is then subtracted from the mean of X_1.

$$a_{1.234} = 105.92 - (-10.90) = 116.82$$

This completes the computation of all the linear multiple correlation constants. The results may be summarized:

$$X_1' = 116.82 - 0.810X_2 + 0.180X_3 - 0.309X_4$$
$$\bar{R}_{1.234} = 0.86$$
$$\bar{S}_{1.234} = 4.58$$

Tables 87, 88, 89, 90, and the computations following 90, have shown every arithmetic step in obtaining these results, arranged in the most convenient form for ready computation and checking. For problems involving large numbers of observations, the methods of computing the extensions, such as ΣX_2^2 and ΣX_1X_2, which are shown in Tables 85 and 86, may be used in place of the individual-item method shown in Table 87; but, thereafter, the work is the same. The check sum may be carried through in computations like those shown in Table 86 just as readily as in the longer method, thus providing a complete check on all the tabulating, multiplying, and adding.

In solving the equations in actual practice, only the items that are not enclosed in parentheses in Table 89 would be entered. Table 91 illustrates this same form of solution for a six-variable problem. It is carried out step by step, just as was Table 89. A slightly different notation is used, but the procedure through line III′ is the same. Then following line IV, line IV–1 is obtained by multiplying line I by −0.256900, the coefficient in line I′, column X_5; line IV–2 by multiplying line Σ_2 by −0.414640, the value in line II′, column X_5; and line IV–3 by multiplying line Σ_3 by −0.168107, the value in line III′, column X_5. A similar regular order is followed at the next step of the process, multiplying lines I, Σ_2, Σ_3, and Σ_4 by the coefficients in column X_6, lines I′, II′, III′, and IV′. Table 91 also illustrates the calculation of the back solution, the final check on the values of the b's by substituting in the last equation (equation V), and the computation of R.

In entering equations V and X_1, in the second and fourth columns from the right in the last section of Table 91, we note that the sequence of the values, from the top to the bottom of the columns, is reversed from the sequence at the head of the table, from left to right for equation V, and from top to bottom for equation X_1. This is because the form of the back solution at the foot of the table places $b_{16.2345}$ at the top of the regression coefficients and $b_{12.3456}$ at the foot. Hence in entering equations V and X_1, they must start off with $\Sigma(x_6^2)$ and $\Sigma(x_2x_6)$ and end up with $\Sigma(x_2x_6)$ and $\Sigma(x_1x_2)$. Reversing the order, as shown, produces this result.

When automatic calculating machines are available to perform the calculations shown in Tables 89 and 91, many of the operations shown can be performed in the machine without separate recording in the tables. Details of this short cut are given on page 478.

TABLE 91

Doolittle Solution of Normal Equations for Six Variables

Line Designation	Column Designation						
	X_2	X_3	X_4	X_5	X_6	X_1	ΣX
	Equations to be Solved						
Eq. I	100.00	23.32	19.86	25.69	10.64	40.17	219.68
Eq. II	(23.32)	100.00	17.47	45.20	21.39	60.03	267.41
Eq. III	(19.86)	(17.47)	100.00	26.28	0.33	23.79	187.73
Eq. IV	(25.69)	(45.20)	(26.28)	100.00	29.89	68.07	295.13
Eq. V	(10.64)	(21.39)	(0.33)	(29.89)	100.00	35.53	197.78
	Front Solution						
I	100.0000	23.3200	19.8600	25.6900	10.6400	40.1700	219.6800
I′	−1.000000	−0.233200	−0.198600	−0.256900	−0.106400	−0.401700	−2.196800
II	....	100.0000	17.4700	45.2000	21.3900	60.0300	267.4100
II-1	....	−5.4382	−4.6314	−5.9909	−2.4812	−9.3676	−51.2294
Σ_2	....	94.5618	12.8386	39.2091	18.9088	50.6624	216.1806
II′	....	−1.000000	−0.135769	−0.414640	−0.199962	−0.535760	−2.286130
III	....	....	100.0000	26.2800	0.3300	23.7900	187.7300
III-1	....	....	−3.9442	−5.1020	−2.1131	−7.9778	−43.6284
III-2	....	....	−1.7431	−5.3234	−2.5672	−6.8784	−29.3506
Σ_3	....	....	94.3127	15.8546	−4.3503	8.9338	114.7510
III′	....	....	−1.000000	−0.168107	+0.046126	−0.094725	−1.216706
IV	....	....	....	100.0000	29.8900	68.0700	295.1300
IV-1	....	....	....	−6.5993	−2.7334	−10.3197	−56.4358
IV-2	....	....	....	−16.2577	−7.8403	−21.0067	−89.6371
IV-3	....	....	....	−2.6653	+0.7313	−1.5018	−19.2904
Σ_4	....	....	....	74.4772	20.0476	35.2418	129.7667
IV′	....	....	....	−1.000000	−0.269178	−0.473189	−1.742367
V	....	....	....	....	100.0000	35.5300	197.7800
V-1	....	....	....	....	−1.1321	−4.2741	−23.3740
V-2	....	....	....	....	−3.7310	−10.1306	−43.2279
V-3	....	....	....	....	−0.2007	+0.4121	+5.2930
V-4	....	....	....	....	−5.3964	−9.4863	−34.9303
Σ_5	....	....	....	....	89.4898	12.0511	101.5408
V′	....	....	....	....	−1.00000	−0.134664	−1.134664

	Back Solution					Eq. V	Check	Eq. X_1	Computation of R^2
	$b_{12\cdot3456}$	$b_{13\cdot2456}$	$b_{14\cdot2356}$	$b_{15\cdot2346}$	$b_{16\cdot2345}$				
	+0.401700	+0.535759	+0.094725	+0.473189	+0.134664				
	−0.014328	−0.026928	+0.006212	−0.036249	+0.134664	100.00	13.4664	35.53	4.7846
	−0.112250	−0.181173	−0.073453	+0.436940		29.89	13.0601	68.07	29.7425
	−0.005458	−0.003731	+0.027484			0.33	0.0091	23.79	0.6538
	−0.075540	+0.323927				21.39	6.9288	60.03	19.4453
	+0.194124					10.64	2.0655	40.17	7.7980
					Eq. V = 35.53		35.5299		62.4242

$$R^2 = \frac{62.4242}{\sigma_1^2} = \frac{62.4242}{100.00} \qquad R = \sqrt{0.624242} = 0.790090$$

Standard errors of partial regression coefficients and standard error of an individual estimate. The computation of standard errors of net or partial regression coefficients by equation (74), as discussed in Chapter 18, and of standard errors of an individual estimate, by equations (77) or (81), as described in Chapter 19, may be much simplified by the following procedure:

For three independent variables, set up the normal equations:

$$\begin{aligned}
\Sigma(x_2^2)c_{22} &+ \Sigma(x_2x_3)c_{23} + \Sigma(x_2x_4)c_{24} = 1\\
\Sigma(x_2x_3)c_{22} &+ \Sigma(x_3^2)c_{23} + \Sigma(x_3x_4)c_{24} = 0\\
\Sigma(x_2x_4)c_{22} &+ \Sigma(x_3x_4)c_{23} + \Sigma(x_4^2)c_{24} = 0
\end{aligned}$$

Solve simultaneously to obtain the values for c_{22}, c_{23}, and c_{24}. Then set up exactly the same set of equations, with c_{32}, c_{33}, and c_{34} as the unknowns, and with 0, 1, and 0 to the right of the equal signs, in the first, second, and third equations, respectively, and solve. Then set up again, with c_{42}, c_{43}, and c_{44} as the unknowns, and with 0, 0, and 1 to the r.ght of the equal signs, and solve again. The standard errors of the regression coefficients may then be found by the following equations (for proof, see Note 13, Appendix 2):

$$\left.\begin{aligned}
\sigma_{b_{12.34}} &= \bar{S}_{1.234}\sqrt{c_{22}}\\
\sigma_{b_{13.24}} &= \bar{S}_{1.234}\sqrt{c_{33}}\\
\sigma_{b_{14.23}} &= \bar{S}_{1.234}\sqrt{c_{44}}
\end{aligned}\right\} \qquad (101)$$

It will be noted that, except for the values to the right of the equal sign, the coefficients of the equations are exactly the same as those required to obtain the values of $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$. For that reason the values for c_{22}, c_{33}, and c_{44} may be most readily calculated by introducing as many new columns in the form of the Doolittle solution (Table 91) as there are independent factors, between the columns for X_1 and Σ. These columns will be

Line	Error b_2	Error b_3	Error b_4	Error b_5
(Eq. I)	1	0	0	0
(Eq. II)	0	1	0	0
(Eq. III)	0	0	1	0
(Eq. IV)	0	0	0	1
etc.				

These values can be included in the check sum, and the operations carried through for them just as for the other columns until the "back solution" to find the b's is reached. Then a separate "back solution" can be run for each set of "c" values, starting with the values in each "Error" column just as the back solution to find the b's started with the values in the X_1 column.[2]

[2] For an explanation of why this process and equation (01) gives the standard error of the b's see Note 14, Appendix 2. For other uses of the "c" constants, see R. A. Fisher, *Statistical Methods for Research Workers*, seventh edition, pages 160–168. Oliver and Boyd, Edinburgh and London, 1938.

TABLE 92

SOLUTION OF NORMAL EQUATIONS BY THE DOOLITTLE METHOD, TO CALCULATE REGRESSION COEFFICIENTS AND THEIR STANDARD ERRORS

Line Designation	X_2	X_3	X_4	X_1	c_2	c_3	c_4	$\Sigma(X + c)$
	EQUATIONS TO BE SOLVED							
Eq. I........	612.77	2,875.16	−25.38	30.31	1	0	0	3,493.86
II........	2,875.16	44,879.23	2,720.92	4,924.46	0	1	0	55,400.77
III........	−25.38	2,720.92	1,093.70	172.85	0	0	1	3,963.09
	FRONT SOLUTION							
I.............	612.77	2,875.16	−25.38	30.31	1.00000	0	0	3,493.86
I′.............	−1.0000000	−4.6920704	0.0414185	−0.0494639	−0.0016319	0	0	−5.7017477
II.............		44,879.23	2,720.92	4,924.46	0	1.00000	0	55,400.77
I(−4.6920704)		−13,490.45	119.08	−142.22	−4.69207	0	0	−16,393.44
Σ_2............		31,388.78	2,840.00	4,782.24	−4.69207	1.00000	0	39,007.33
II′............		−1.0000000	−0.0904782	−0.1523551	0.0001495	−0.0000319	0	−1.2427157
III............			1,093.70	172.85	0	0	1.0000	3,963.09
I(0.0414185)			−1.05	1.26	0.0414185	0	0	144.71
Σ_2(−0.0904782)			−256.96	−432.69	0.4245300	−0.0904782	0	−3,529.31
Σ_3............			835.69	−258.58	0.4659485	−0.0904782	1.0000000	578.49
III′...........			−1.0000000	0.3094210	−0.0005576	0.0001083	−0.0011966	−0.6922304

BACK SOLUTION ON c_2

	c_{22}	c_{23}	c_{24}	Eq. II—c_2	Check
	0.0016319	−0.0001495	0.0005576		
	0.0000231	−0.0000505	0.0005576	2,720.92	1.52
	0.0009384	−0.0002000		44,879.23	−8.98
	0.0025934			2,875.16	7.46
				0.00	0.00

BACK SOLUTION ON c_3

	c_{32}	c_{33}	c_{34}	Eq. II—c_3	Check
		0.0000319	−0.0001083		
		0.0000098	−0.0001083	2,720.92	−0.29
		00.0000417		44,879.23	1.87
	−0.0002000			2,875.16	−0.58
				1.00	1.00

BACK SOLUTION ON c_4

	c_{24}	c_{34}	c_{44}	Eq. II—c_4	Check
			0.0011966		
			0.0011966	2,720.92	3.26
		−0.0001083		44,879.23	−4.86
	0.0005576			2,875.16	1.60
				0.00	0.00

Table 92 shows all the computations necessary to compute all the b's and c's from the product sums calculated in Table 88, except for the back solution on X_1, as shown in the lower section of Table 89. Table 92 thus replaces all of Table 89, except this last section. In practice, this back solution would be included in Table 92 ahead of the three back solutions on c_2, c_3, and c_4.

In computing Table 92, the work in the c columns is carried out to two more decimal places than in the other columns. This is necessary because of the small size of the values involved. It should also be noticed that in the back solution on c_3, only c_{34} and c_{33} are calculated directly. Since c_{32} is identical with c_{23}, the value previously calculated for the latter is inserted instead. Similarly, the back solution on c_{44} involves no additional calculating at all, since c_{44} is copied down (with the sign changed) from line III′, c_{24} is written down for c_{42}, and c_{34} for c_{43}. Only the computation by substitution in the check equations is involved. Even that computation can be omitted for the c_4 values, since each of them has been checked earlier—c_{42} and c_{43} by substitution and c_{44} by the check sum in lines Σ_3 and III′.

As a result of these computations, the following values are secured:

$$c_{22} = 0.00259;\ c_{33} = 0.000042;\ c_{44} = 0.00120$$

Since $\bar{S}_{1.234} = 4.58$, the standard error of the b's may be readily calculated by equation (101)

$$\sigma_{b_{12.34}} = 4.58\sqrt{0.00259} = 0.233$$

$$\sigma_{b_{13.24}} = 4.58\sqrt{0.000042} = 0.030$$

$$\sigma_{b_{14.23}} = 4.58\sqrt{0.00120} = 0.159$$

The net regression coefficients may then be stated

$$b_{12.34} = -\ 0.810 \pm 0.233$$

$$b_{13.24} = \quad 0.180 \pm 0.030$$

$$b_{14.23} = -\ 0.309 \pm 0.159$$

Just as in the illustrations discussed in Chapter 18, some of the net regression coefficients are much more reliable than are others. If we assume that the conditions of random sampling are fulfilled, there is some possibility that the regression for $b_{14.23}$ in the universe from which the sample was drawn is really positive instead of negative; but there is only a very slight chance that $b_{12.34}$ is really positive, and it is almost a certainty that $b_{13.24}$ is really positive, and above 0.1.

The computation of the standard errors of the net regression coefficients, by the method just presented, is not a difficult one. It should be made an integral part of every multiple correlation solution, so that not only will the regression coefficients be obtained, but also the amount of confidence that can be placed in each value will be determined. Only if that is done can the regressions be interpreted with confidence.

The computations shown in Table 92 also give all the values needed to estimate the standard error of an individual estimate. Substituting these values in equation (77), and using the value for $\bar{S}_{1.234}$ previously calculated on page 467 (in practice, the calculations on that page would all be made after Table 92 was calculated, including the back solution on X_1), we have:

$$\text{(X)} \quad \sigma^2_{x_1'-x_1} = 4.58\left[1 + \frac{1}{13} + .00259x_2^2 + .000042x_3^2 + .00120x_4^2 + 2(-.00020)x_2x_3 + 2(.00056)x_2x_4 + 2(-.00011)x_3x_4\right]$$

The use of this equation may be shown as follows: Suppose we draw a new observation from the same universe as that from which the original sample (shown in Table 87) was drawn, and the new observation has values of 18 for X_2, 300 for X_3, and 90 for X_4. After we estimate the probable X_1 value from the regression equation, how much confidence can we place in that estimate?

The estimated value works out as follows:

The regression equation (from Table 90) is

$$\begin{aligned} X_1' &= -\ 10.90 - 0.80962X_2 + 0.18036X_3 - 0.30944X_4 \\ &= -\ 10.90 - 0.80962(18) + 0.18036(300) - 0.30944(90) \\ &= 0.79 \end{aligned}$$

Before the values of X_2, X_3, and X_4 for this new observation can be substituted in equation (X), they must be put in the form of x_2, x_3, or x_4. Using the means shown in Table 86, we calculate

$$\begin{aligned} x_2 &= X_2 - M_2 = 18 - 10.31 = 7.69 \\ x_3 &= X_3 - M_3 = 300 - 167.46 = 132.64 \\ x_4 &= X_4 - M_4 = 90 - 105.84 = -\ 15.84 \end{aligned}$$

Substituting these values in equation (X), we have

$$\sigma^2_{x_1'-x_1} = 4.58\left[1 + \frac{1}{13} + .00259\ (7.69)^2 + .000042(132.64)^2 + .00120(-15.84)^2 + 2\ (-.00020)(7.69)(132.64) + 2(.00056)(7.69)(-15.84) + 2(-.00011)(132.64)(-15.84)\right] = 10.0205$$

$$\sigma_{x_1'-x_1} = 3.17$$

We can now say that our estimate of X_1, for the new observation with $X_2 = 18$, $X_3 = 300$, and $X_4 = 90$, is $X_1' = 0.79 \pm 3.17$. Alternatively, applying the method explained on pages 343 and 344 of Chapter 19, we can say we feel confident that the *true value* of X_1 for the new observation will lie between -6.37 and 7.95, knowing that such a statement will be wrong in only one out of twenty such statements, on the average.

It is evident from this illustration that the standard error of this particular estimate, 3.17, is larger than the standard error of estimate for the sample, 2.14. That is because the values of the independent variables for this new observation lay near the extremes of their several ranges in the sample. It is also evident that the value of $\sigma_{x'_1-x_1}$ will vary with each new observation, depending on the combination of values for the independent variables in each observation.

Coefficients of partial correlation. Computation of the coefficients of partial correlation by equation (50) involves the calculation of the multiple correlation of the dependent va iable with successive sets of the independent variables, with a different independent va iable left out in each set. Thus, for the four-variable problem whose solution is shown in Table 89 (and 92) the three coefficients of partial correlation involve not only the value $\bar{R}_{1.234}$, but also $\bar{R}_{1.23}$, $\bar{R}_{1.24}$, and $\bar{R}_{1.34}$. These may be calculated readily by the same process shown in Table 89. It is not necessary to repeat the process three times, however, as the several columns may be rearranged with little additional calculation to omit each independent variable in turn. The first two stages in this process are illustrated in Table 93. The values for lines I and I′ and Σ_2 and II′ are copied from Table 89, as shown in the first four lines of Table 93. The columns X_4 and ΣX are dropped, however, as they are not needed at this step.

Lines I′ and II′ give all the information needed for the "back solution" with X_4 omitted. This is accordingly given in the second section of Table 93, using the same form as in the back solution of Table 89. The verification of the b's by substitution in equation II, and the calculation of $R_{1.23}$ by use of equation I are also shown, organized the same as in Table 91.

The next step is to enter the values necessary for the "front solution" with X_3 omitted. This is shown in the third block of Table 43. Lines I and I′ are entered again, with the column for X_3 omitted. Lines III and (0.04142) (I) are copied from Table 89 for columns X_4 and X_1. All that is necessary to complete the front solution is to add the new totals, Σ_3, and to divide col. X_1 by col. X_4, to get the new line ΣIII′, and then to proceed with the back solution, as before. The check on equation III and the calculation of $R_{1.24}$ by substitution in equation I are also shown under the back solution on X_3.

TABLE 93

Doolittle Solution of Normal Equations, to Find Coefficients of Partial Correlation, for Three Independent Variables

Line Designation	Column Designation		
	X_2	X_3	X_1
I	612.77	2,875.16	30.31
I′	−1.00000	−4.69207	−0.04946
Σ_2		31,388.78	4,782.24
II′		−1.00000	−0.15236

Back Solution, X_4 Omitted			Eq. II	Check	Eq. I	Computation of $R^2_{1.23}$
	$b_{12.3}$	$b_{13.2}$				
	+0.04964	+0.15236				
	−0.71488	+0.15236	44,879.12 2,875.16	6,837.78 −1,912.67	4,924.46 30.31	750.29 −20.16
	−0.66524		4,924.46	4,925.11	998.92	730.13

$$R^2_{1.23} = \frac{730.13}{998.92} = 0.730919$$

Front Solution, X_3 Omitted			
	X_2	X_4	X_1
I	612.77	−25.38	30.31
I′	−1.00000	0.04142	−0.04946
III		1,093.70	172.85
(0.04142)(I)		−1.05	1.26
Σ_3		1,092.65	174.11
III′		−1.00000	−.15935

Back Solution, X_3 Omitted			Eq. III	Check	Eq. I	Computation of $R^2_{1.24}$
	$b_{12.4}$	$b_{14.2}$				
	0.04946 0.00660	0.15935				
	0.05606	0.15935	1,093.70 −25.38	174.28 −1.42	172.85 30.31	27.54 1.70
			172.85	172.86	998.92	29.24

$$R^2_{1.24} = \frac{29.24}{998.92} = 0.029272$$

The final step in computing the needed coefficients of multiple correlation involves calculating $R_{1.34}$. Since this involves rearranging Table 89 to omit X_2, which appears in the first column, it is necessary to carry through an entire new front solution, with X_2 omitted. This process is shown in Table 94. (The column ΣX is a new Σ, obtained by adding the values in columns X_3, X_4, and X_1 for lines II and III, and then using it as a check thereafter.)

In problems where there are four independent variables, this new back solution should be arranged in this column order X_4, X_5, X_3, X_2, X_1. After the entries were calculated through the front solution, two back solutions could then be run, one leaving out the X_2 column, and one the X_3 column. Where there are six independent variables, a third step could be used by repeating the last two steps of the front solution for the third independent variable to be dropped out; or a complete new solution could be run with X_3 and X_4 occupying the last columns before X_1. In many-variable problems, various other time-saving combinations can be worked out by the ingenious computer.

TABLE 94

Doolittle Solution of Normal Equations, to Find Coefficients of Partial Correlation, for Three Independent Variables (*continued*)

Line Designation	Column Designation			
	X_3	X_4	X_1	ΣX
II	44,879.23	2,720.92	4,924.46	52,524.61
II′	−1.00000	−0.06063	−0.10973	−1.17046
III		1,093.70	172.85	3,987.47
(−0.06063)(II)		−164.97	−298.57	−3,184.45
Σ_2		928.73	−125.72	803.01
		−1.00000	0.13537	−0.86463

Back Solution, X_2 Omitted

	$b_{13.4}$	$b_{14.3}$	Eq. III	Check	Eq. I	Computation of $R^2_{1.34}$
	0.10973	−0.13537				
	0.00821	−0.13537	2,720.92	−368.33	172.85	−23.39
	0.11794		44,879.23	5,293.06	4,924.46	580.79
			4,924.46	4,924.73	998.92	557.40

$$R^2_{1.34} = \frac{557.40}{998.92} = 0.558002$$

Tables 90, 93, and 94 provide all the values necessary for the calculation of the partial correlation coefficients, using equations (47) and (50). These calculations may be tabled as follows:

Variables	(1) Uncorrected R^2	(2) $1 - R^2$	(3) $\dfrac{n-1}{n-m}$	(4) $1 - \bar{R}^2$ (3) × (2)
1.234	0.8110	0.1890	$\frac{12}{9} = 1.3333$	0.2520
1.23	0.7309	0.2691	$\frac{12}{10} = 1.2000$	0.3229
1.24	0.0293	0.9707	$\frac{12}{10} = 1.2000$	* 1.0000
1.34	0.5580	0.4420	$\frac{12}{10} = 1.2000$	0.5304

* Taken as 1.0 (largest possible value), when estimate of probable value exceeds 1.0.

$$\bar{r}^2_{12.34} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.34}} = 1 - \frac{0.2520}{0.5304} = 1 - 0.4751 = 0.5249$$

$$\bar{r}^2_{13.24} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.24}} = 1 - \frac{0.2520}{1.0000} = 1 - 0.2520 = 0.7480$$

$$\bar{r}^2_{14.23} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.23}} = 1 - \frac{0.2520}{0.3229} = 1 - 0.7804 = 0.2196$$

$$\bar{r}_{12.34} = -0.72$$

$$\bar{r}_{13.24} = 0.87$$

$$\bar{r}_{14.23} = -0.47$$

The signs of the partial correlation coefficients are taken from the signs of the corresponding net regression coefficients, as shown in Table 89 or 90.

Alternative methods of solving normal equations. The methods for solving normal equations and obtaining the various constants necessary in correlation analysis, which have been presented in Tables 89 to 94, inclusive, employ the so-called Doolittle method of solving equations, first developed by Dr. M. H. Doolittle, a computer in the Geodetic Survey.[3] His method involved slight modifications of the methods originally suggested by Gauss, the discoverer of the least-squares technique. (The solutions shown in Tables 93 and 94 involve short cuts added by the author of this book.) The use of the 0–1–0, etc., method of calculating error formulas (the reciprocal matrix), was also first developed by Gauss, and was revived by R. A. Fisher. Its further application to calculating the standard errors of an individual estimate was developed by Dr. Meyer Girshick of the U. S. Department of Agriculture, at the author's request.

[3] M. H. Doolittle, Adjustment of the primary triangulation between Kent Island and Atlanta base lines (Paper No. 3, Method employed in the solution of normal equations and the adjustment of a triangulation), Report of the Superintendent, Coast and Geodetic Survey, 1878, pp. 115–120.

Since the normal equations are in the form of a symmetrical determinant, the methods of determinantal algebra can be applied in their solution. Those familiar with determinantal and matrix algebra may find forms of solutions based on these principles, developed by Frederick V. Waugh,[4] more convenient than the methods illustrated here. Careful comparisons of the Waugh solutions with the Doolittle solutions by the author of this book, however, have revealed that the Doolittle method involves somewhat fewer calculations if all that is desired are the constants whose calculations have been illustrated to this point—the coefficients of net regression and multiple correlation, the coefficients of partial correlation, and the standard errors for the regression coefficients and for individual estimates. If one also wishes to determine all the other possible coefficients of multiple and partial correlation ($R_{2.134}$ as well as $R_{1.234}$, and $r_{23.14}$ as well as $r_{12.34}$, etc.) and all the other net regression coefficients and equations ($b_{23.14}$ as well as $b_{12.34}$, etc.), the Waugh method is faster. Since these additional coefficients are rarely used, and since the Doolittle method can be understood more readily by students whose mathematical training has not extended beyond relatively simple algebra, the presentation here has been restricted to the Doolittle method.

A somewhat simpler short cut in the solution of the normal equations has been suggested by P. S. Dwyer.[5] He points out that much of the "front solution" involves subtracting a series of products from, or adding them to, a given figure. In Table 91, for example, the item that appears in line Σ_4 of column X_5 is simply the value:

$$100.0000 + (25.6900)(-0.256900) + (39.2091)(-0.414640) + (15.8546)(-0.168107) = 74.4772$$

With modern computing machines this value can be computed directly without clearing the total dial, using the reverse lever whenever the product is to be subtracted instead of added. This method saves reading off and entering in the table the values that appear in line IV–1, IV–2, and IV–3. It is slightly more accurate as it obviates the possible errors in rounding off each of the products as they are entered in the table. The calculations in the machine, for example, may be carried to ten decimal places, and only the final sum is rounded off. The method does involve one handicap, however, in that the multiplier (as for example -0.256900 for line IV–1) has to be set up on the keys separately for each column in turn, whereas in the usual method it can be set up and left in the machine as a constant multiplier as all the products clear across line IV–1 are computed. Whether this additional operation and possibility of error offset the other savings each computer can determine for himself.

Using this Dwyer short cut all the way through, the front solution of Table 89 would show only the lines I and I′, Σ_2 and II′, and Σ_3 and III′. Similarly Table 91 would show in the front solution only I and II′, Σ_2 and II′, Σ_3 and III′, Σ_4 and IV′, and Σ_5 and V′. Various other possible modifications of the Doolittle solution, all based on the same basic principle, are shown in Dwyer's article referred to above.

[4] Frederick V. Waugh, *Journal of the American Statistical Association*, December, 1935, and December, 1936.

[5] P. S. Dwyer, The solution of simultaneous equations, *Psychometrika*, Vol. 6, No. 2, April, 1941.

Computing residuals for multiple curvilinear correlations. Where there are a large number of individual observations, the average residual around the net regression line may be computed from group averages, instead of calculated for each individual observation as described in Chapter 14. This may save much time in calculating the average residuals to obtain the first approximation regression curves.

After the net linear regression coefficients are computed, the observations are thrown into groups with respect to the first independent factor, say X_2, and averages of each factor are computed for the records falling in each group. If there are four groups, for example, there will be four sets of averages.

Value of X_2	Average X_2	Average X_3	Average X_1
0– 9	M_{2-1}	M_{3-1}	M_{1-1}
10–19	M_{2-2}	M_{3-2}	M_{1-2}
20–29	M_{2-3}	M_{3-3}	M_{1-3}
30 and over	M_{2-4}	M_{3-4}	M_{1-4}

The average estimated value, $M_{x_1'}$, may then be calculated for each group by substituting the means for that group in the regression equation. Thus for the first group,

$$M_{x'} = a + b_2(M_{2-1}) + b_3(M_{3-1})$$

and

$$M_z = M_{1-1} - M_{x'}$$

In a similar manner the average residual may be calculated from the group averages for each of the other groups, and then plotted as a departure from the net regression line, as illustrated in Figure 35 of Chapter 14. After the computation is completed for X_2, the records may be reclassified with respect to X_3, new means calculated for each variable for each group, and the process continued just as for X_2. The same steps are carried out for each other independent variable in turn. This method may be used to determine the net residuals around curvilinear regressions fitted by mathematical curves just as well as for linear regressions.

Once the first set of freehand approximation curves has been drawn, the remainder of the work has to be carried forward just as described in Chapter 14, as the average of values along a curve do not precisely represent that curve in the same way that the average of values along a straight line will represent that line.

Auxiliary graphic processes with the short-cut graphic method. The short-cut method of determining a net curvilinear regression, described in Chapter 16, may be materially aided by using graphic methods in transferring departures from one figure to another, and in calculating the averages of the values as plotted.

After the original observations are plotted and the first approximation to the regression line or curve is drawn (as in Figure 51 of Chapter 16), the departures from that line must be plotted against the next variable. A procedure for making those transfers graphically is shown in Figures 77 to 81. The first step is to place an arrow in the middle of a strip of blank paper. Using the arrow to indicate the position of the regression line or curve, we mark off on the paper the vertical departures of each observation from that line, with each observation indicated by its number. Figure 77 shows this process just as the first observation (29) is marked on

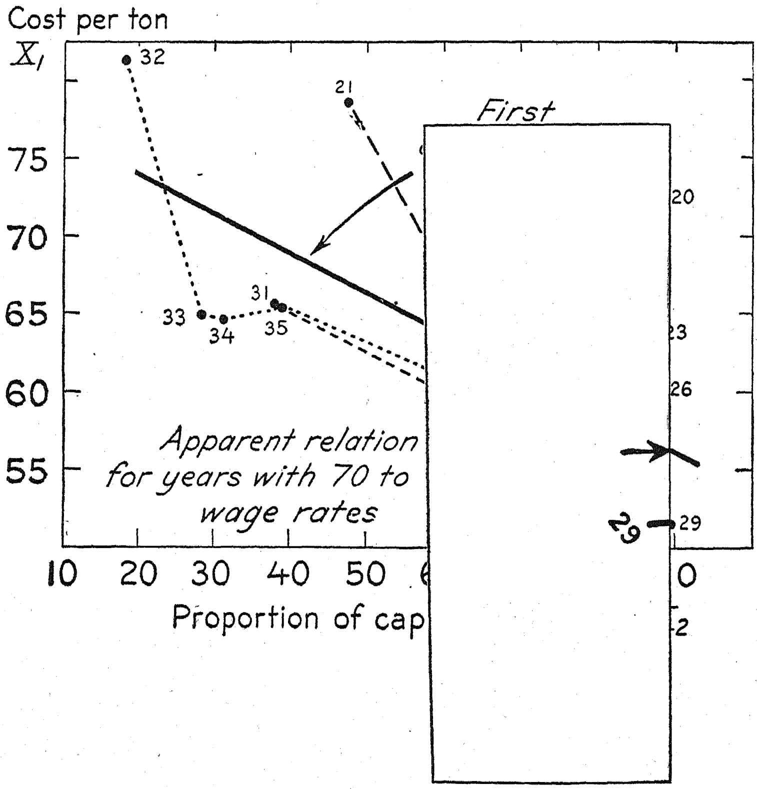


FIG. 77. This shows the start of the process of scaling off graphically the departures from the first approximation to the net regression line or curve.

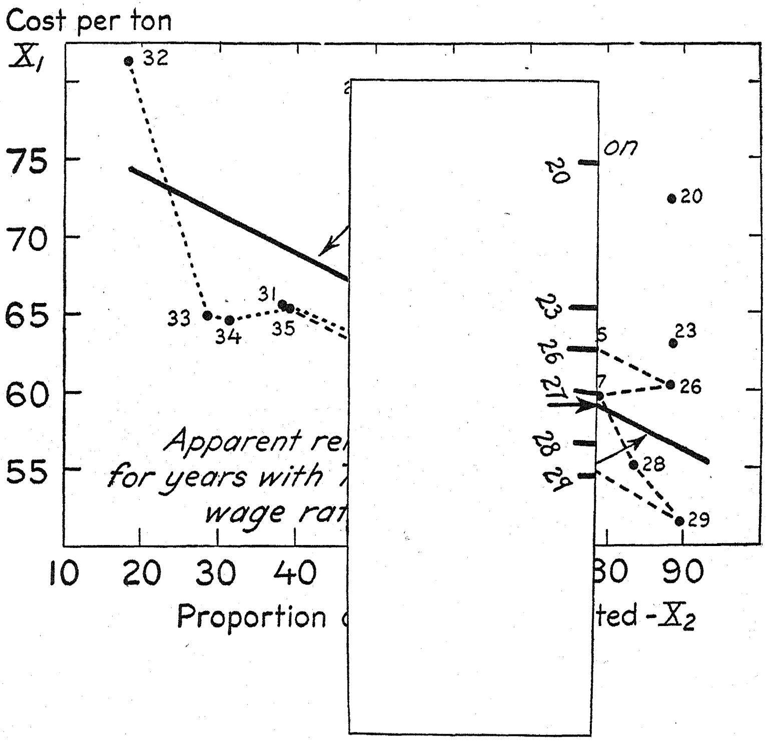


FIG. 78. This shows the process of scaling off the departures partially completed.

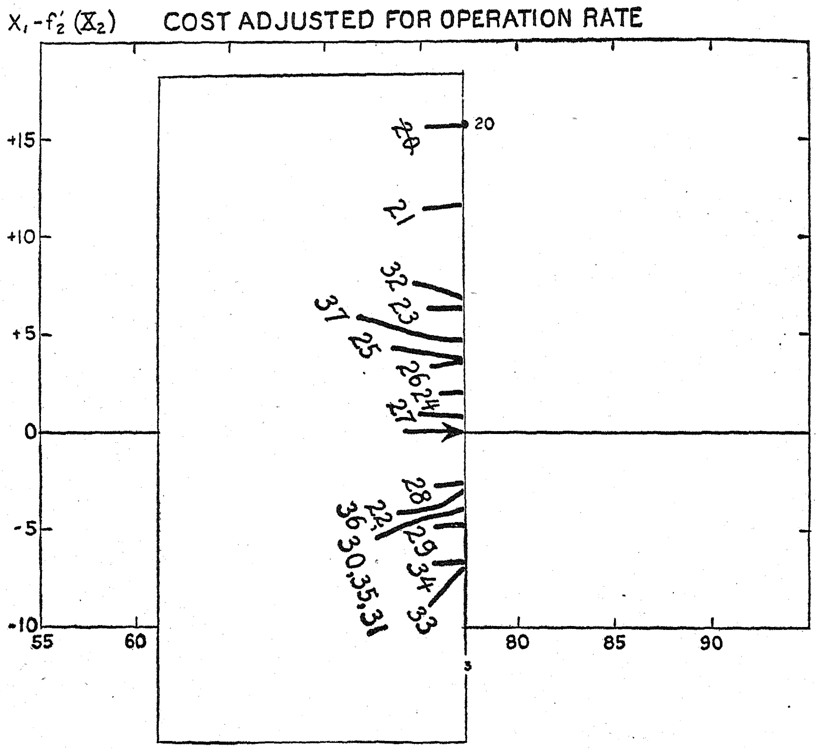


FIG. 79. Here the slip with the departures from the first approximation is moved to the next chart, ready to start transferring the departures to get the first approximation for the next variable. The first observation, for 1920, has been entered and checked off.

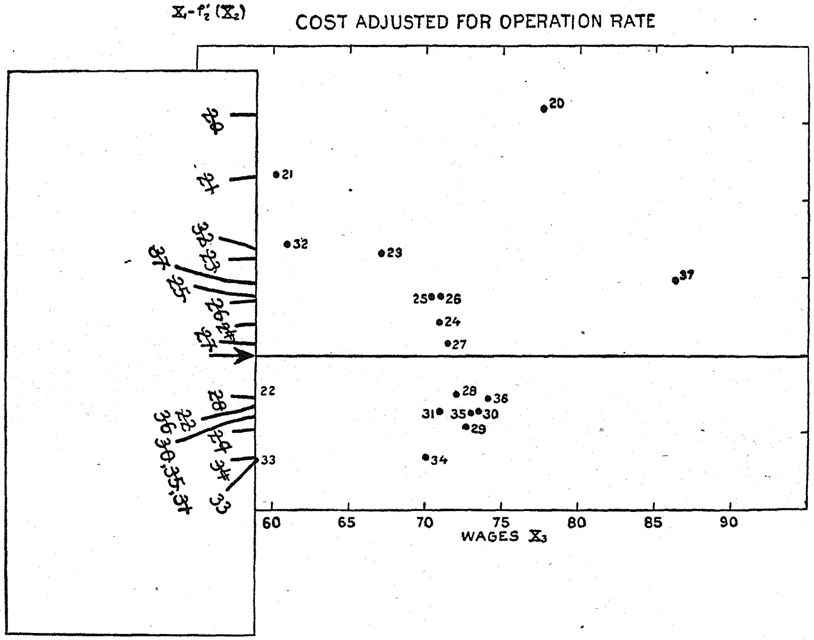


FIG. 80. The process shown at the start in Fig. 79 is here shown completed.

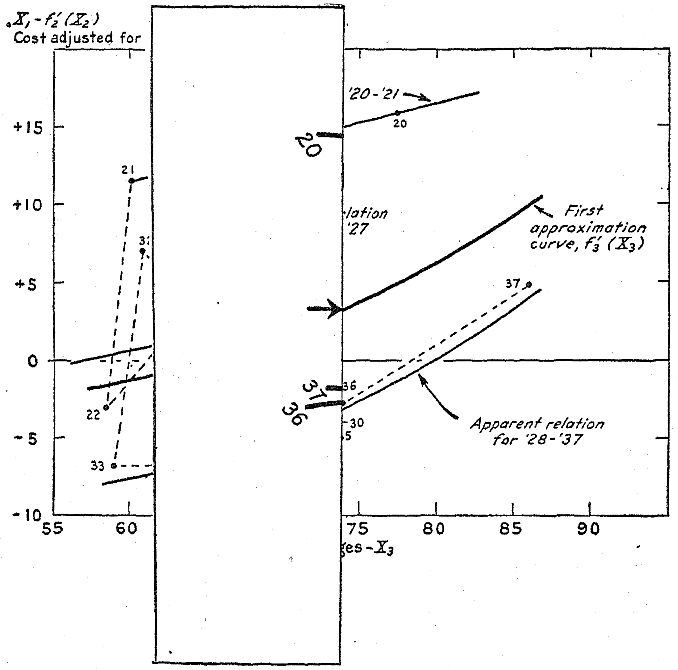


FIG. 81. After the first (or subsequent) approximation curves are drawn in, the departures from the curve can be scaled off as shown.

the strip. Figure 78 shows it after several such values have been marked on the strip. The process is continued (with one or more strips of paper) until the vertical departures have been marked off for each observation.

The next step in the process is to transfer these departures to the next figure, Figure 52 of Chapter 16. After the chart form has been prepared, the arrow on the slip is centered on the zero line, and the departures marked on the figure, with the slip moved to the corresponding X_3 value as ordinate. Figure 79 shows this step just after the value for the 1920 observation was entered on the chart. After the value is marked on the new figure, it is crossed off on the slip, to prevent confusion. Figure 80 shows the process completed, just as the last value on the slip—that for the 1933 observation—is entered. (It will be noticed that the values are transferred in sequence from top to bottom of the slip, to prevent confusion.)

After the new curve is inserted on the chart, the next step is to transfer residuals from the new curve to the next figure. The departures can be scaled off from a curve as readily as from a line. Figure 81 shows the start of the next stage of the process, after the departures for the observations for 1920 and 1937 have been scaled off from the first approximation curve on Figure 52, and just as the value for 1936 is entered. The process is completed and carried on to the next chart (Figure 53) just as illustrated above. The same process is used in transferring the departures for each stage in the approximation process, always scaling off the residuals *from* the last approximation curve, and plotting them as departures from the last curve on the next chart, prior to drawing in the new curve.

After the departures are entered, averages of departures are sometimes needed. In such cases, graphic means can be used to average each group of observations. To do this, an approximate average is inserted by eye. Then all the positive departures of the observations in that group from the approximate average are accumulated on one slip, scaled off each in turn as an addition to the other departures, and all the negative departures from the approximate average are accumulated on another slip. The difference between the two accumulations is divided by the number of cases, giving a plus or minus correction to the approximate average. At later stages when average deviations from a previous line or curve are desired, graphic accumulations can be used similarly, with the previous line used as the first approximation to the new average.

APPENDIX 2

TECHNICAL NOTES

Note 1 (Chapter 2). The formula for estimating the standard error of an average is derived as follows:

Assume that we have N random samples of n observations each, all expressed in deviations, x, from the true mean of the variable, X, we are sampling. Designate the successive observations in each sample x', x'', x''', etc., as shown in the following tabular statement:

Observation	Sample 1	Sample 2	Sample 3	Sample N
1	x'	x'	x'	x'
2	x''	x''	x''	x''
3	x'''	x'''	x'''	x'''
4	x''''	x''''	x''''	x''''
....	....	....	....	....
....	....	....	....	....
....	....	....	....	....
n	x_n	x_n	x_n	x_n
	Σx_1	Σx_2	Σx_3	Σx_N

The first observations in the several samp'es (line 1) will have a standard deviation ($\sigma_{x'}$) which will tend to approach the true standard deviation (σ_x) of the universe of x's from which the observations are drawn. As N, the number of samples, becomes larger and larger, $\sigma_{x'}$ will tend to agree more and more closely with σ_x.

The second observations in the several samples (line 2), will have a standard deviation ($\sigma_{x''}$) which will also tend to agree more and more closely with the standard deviation of the universe (σ_x) as N becomes larger and larger.

Suppose now that the first and second observations in each sample are added. Let this sum be designated x_{1+2}. That is,

$$x' + x'' = x_{1+2}$$

The standard deviation of all the x_{1+2}'s from the N samples is, by definition,

$$\sqrt{\frac{\Sigma(x_{1+2})^2}{N}}$$

But each

$$\begin{aligned}(x_{1+2})^2 &= (x' + x'')^2\\ &= (x')^2 + 2x'x'' + (x'')^2\end{aligned}$$

Hence,

$$\Sigma(x_{1+2})^2 = \Sigma(x')^2 + 2\Sigma x'x'' + \Sigma(x'')^2$$

If the successive observations in each sample, x', x'', x''', are obtained by true random sampling, they will not be correlated with each other; that is, a large value for x' will be just as likely to be followed by a small value for x'' as by another large value.

The correlation of x' with x'' will tend to approach 0 as the number of samples, N, is increased. But if $r_{x'x''} = 0$, $\Sigma x'x''$ will equal 0. Hence, if the successive observations x' and x'' are uncorrelated, the last equation becomes

$$\Sigma(x_{1+2})^2 = \Sigma(x')^2 + \Sigma(x'')^2$$

Dividing by N

$$\sigma_{1+2}^2 = \sigma_{x'}^2 + \sigma_{x''}^2$$

But $\sigma_{x'}$ and $\sigma_{x''}$ both tend to equal σ_x

Hence

$$\sigma_{1+2}^2 = 2\sigma_x^2, \text{ when } N \text{ is very large.}$$

Similarly, if the first three observations are added, the standard deviation of their sum, x_{1+2+3}, will be

$$\sigma_{1+2+3}^2 = \sigma_{x'}^2 + \sigma_{x''}^2 + \sigma_{x'''}^2$$
$$= 3\sigma_x^2, \text{ when } N \text{ is very large.}$$

So if all n observations in each sample are added, the standard deviation of the sum, $x_{1+2+\cdots+n}$, will be

$$\sigma_{1+2+\cdots+n}^2 = \sigma_{x'}^2 + \sigma_{x''}^2 + \ldots + \sigma_{x_n}^2$$
$$= n\sigma_x^2, \text{ when } N \text{ is large.}$$

If each observation in each sample and also the sum of all the observations in each sample are divided through by n (the total number of observations in each sample), each x' will become x'/n, and each Σx will become $\Sigma x/n$. The standard deviation for the series of values x'/n for the first observation in each sample may then be computed

$$\text{each } \left(\frac{x'}{n}\right)^2 = \frac{(x')^2}{n^2}$$

and

$$\Sigma\left(\frac{x'}{n}\right)^2 = \frac{\Sigma(x')^2}{n^2} = \frac{N\sigma_{x'}^2}{n^2}$$

Dividing through by N,

$$\sigma_{x'/n}^2 = \frac{\sigma_{x'}^2}{n^2}$$

and since $\sigma_{x'}$ tends to equal σ_x

$$\sigma_{x'/n}^2 = \frac{\sigma_x^2}{n^2}, \text{ when } N \text{ is very large.}$$

The standard deviation for the second series of observations x''/n, similarly will be

$$\sigma_{x''/n}^2 = \frac{\sigma_{x''}^2}{n^2} = \frac{\sigma_x^2}{n^2}$$

The standard deviation of the sums of $x'/n + x''/n$ likewise will be

$$\sigma^2_{x'/n+x''/n} = \frac{\sigma^2_x}{n^2} + \frac{\sigma^2_x}{n^2} = \frac{2\sigma^2_x}{n^2}$$

So the standard deviation of the sums of all the n values x'/n to x_n/n will be

$$\sigma^2_{x'/n+x''/n+x'''/n+\cdots+x_n/n} = \frac{\sigma^2_x}{n^2} + \frac{\sigma^2_x}{n^2} + \ldots + \frac{\sigma^2_x}{n^2}$$

$$= \frac{n\sigma^2_x}{n^2}$$

$$= \frac{\sigma^2_x}{n}$$

But the sums $\Sigma\left(\frac{x}{n}\right)$ of all the values x'/n to x_n/n, that is the values $\frac{\Sigma x}{n}$, are the arithmetic averages for each sample, M_x. Hence, under conditions of random sampling, the standard deviation of the arithmetic means of samples of n observations is given by the last equation, which may be written

$$\sigma^2_M = \frac{\sigma^2_x}{n}, \text{ and hence, } \sigma_M = \frac{\sigma_x}{\sqrt{n}}$$

It should be noted that σ_x is the standard deviation of all the items in the universe, not the standard deviation of the items in a random sample. It has been shown by "Student" (Biometrika, Vol. 6, p. 1, 1908) that the standard deviation, σ_s, of the items in a small random sample, calculated not from the true mean but from the mean of the values in that sample, tends to be less than the standard deviation, σ_x, of all the items in the universe from which that sample was drawn. To obtain an estimate of the true standard deviation which is free from this bias, the standard deviation calculated from a sample of n observations must be adjusted by the equation[1]

$$\bar{\sigma}^2_x = \frac{\Sigma x^2}{n-1} = \frac{n\sigma^2_s}{n-1}$$

This is the origin of equation (6.1), given in the text.

Hence

$$\sigma_M = \frac{\bar{\sigma}_x}{\sqrt{n}}$$

It should be noted that an essential assumption made in deriving this formula is that the successive observations x', x'', etc., are not correlated with each other. In many types of economic problems, such as time series, for example, this assump-

[1] For a more extended discussion of various adjustments to obtain an unbiased estimate of σ_x, see W. Edwards Deming and Raymond T. Birge, On the statistical theory of errors, *Reviews of Modern Physics*, Vol. 6, 119–161, July, 1934.

tion may be incorrect.[2] The effect of that fact upon the usefulness of error formulas for time series has already been discussed in Chapter 19, pages 349 to 356.

Note 2 (Chapter 5). In fitting a straight line, the requirement is to determine, from the series of paired observations of X and Y, values for a and b in the equation

$$Y' = a + bX$$

which will make the sum of the squares of the residuals, $Y - Y'$, as small as possible.

The values whose sum is to be minimized are

$$\begin{aligned}(Y - Y')^2 &= (Y - a - bX)^2 \\ &= Y^2 + a^2 + b^2X^2 - 2aY - 2bYX + 2abX\end{aligned}$$

The sum of these values is therefore

$$\Sigma(Y - Y')^2 = \Sigma Y^2 + na^2 + b^2\Sigma(X^2) - 2a\Sigma(Y) - 2b\Sigma(YX) + 2ab\Sigma(X)$$

To determine the values of a and b which will make $\Sigma(Y - Y')^2$ a minimum, the partial derivatives with respect to a and b must be obtained and set equal to zero

$$\frac{\partial[\Sigma(Y - Y')^2]}{\partial a} = 2na - 2\Sigma(Y) + 2b\Sigma(X)$$

$$\frac{\partial[\Sigma(Y - Y')^2]}{\partial b} = 2b\Sigma(X^2) - 2\Sigma(YX) + 2a\Sigma(X)$$

Setting each equal to zero

$$2na - 2\Sigma(Y) + 2b\Sigma(X) = 0$$

(I) $$na + \Sigma(X)b = \Sigma(Y)$$

and

$$2b\Sigma(X^2) - 2\Sigma(YX) + 2a\Sigma(X) = 0$$

(II) $$\Sigma(X)a + \Sigma(X^2)b = \Sigma(XY)$$

Equations (I) and (II) are then the required normal equations, to be solved simultaneously, as given in footnote 2 of Chapter 5.

Solving the normal equations for a and b, the steps are as follows:

(I) $$na + \Sigma(X)b = \Sigma(Y)$$

$$a + \frac{\Sigma(X)}{n}b = \frac{\Sigma(Y)}{n}$$

$$a = \frac{\Sigma(Y)}{n} - \frac{\Sigma(X)}{n}b$$

(I′) $$a = M_y - M_xb$$

and

(II) $$\Sigma(X)a + \Sigma(X^2)b = \Sigma(XY)$$

[2] This development follows that suggested in G. U. Yule, *An Introduction to the Theory of Statistics*, Chapter XVII, ¶10, pp. 344–345, and Chapter XI, ¶2, pp. 210–211, in the sixth edition, C. Griffin & Co., Ltd., London, 1922.

substituting the value of a given above

$$\Sigma(X)M_y - \Sigma(X)M_x b + \Sigma(X^2)b = \Sigma(XY)$$

$$nM_xM_y - nM_x^2 b + \Sigma(X^2)b = \Sigma(XY)$$

$$\Sigma(X^2)b - nM_x^2 b = \Sigma(XY) - nM_xM_y$$

hence

(II′) $$b = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - nM_x^2}$$

Equations (I′) and (II′) are then the equations given in the text as equations (10) and (9), to compute b and a.

Note 3 (Chapter 6). In computing Σx^2 for a variable X, or Σxy for two variables X and Y, the values are not at all affected if some constant value is subtracted from each item of the series before the computations are made.

Let c represent the constant subtracted, and D_1 represent $X - c$.

Then

$$D_1^2 = (X - c)^2 = X^2 - 2Xc + c^2$$

$$\Sigma D_1^2 = \Sigma X^2 - 2c\Sigma X + nc^2$$

and

$$M_{D_1} = M_x - c$$

Now from equation (5)

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{n} - M_x^2}$$

$$\Sigma x^2 = \Sigma X^2 - nM_x^2$$

Let

$$d = D_1 - M_{D_1}$$

Then

$$\begin{aligned}\Sigma d^2 &= \Sigma D_1^2 - nM_{D_1}^2\\ &= \Sigma X^2 - 2c\Sigma X + nc^2 - n(M_x - c)^2\\ &= \Sigma X^2 - 2cnM_x + nc^2 - nM_x^2 + 2ncM_x - nc^2\\ &= \Sigma X^2 - nM_x^2\end{aligned}$$

Hence

$$\Sigma d^2 = \Sigma x^2$$

By an exactly similar process, if $D_1 = X_1 - c_1$ and $D_2 = X_2 - c_2$, it can be proved that

$$\Sigma(d_1d_2) = \Sigma(x_1x_2)$$

Similarly, if any variable, X, is multiplied or divided by a constant, c, the effect will be to multiply or divide Σx^2 by c^2, and Σxy by c. Hence, where X is divided by c, σ_x and b_{xy} will be divided by c, but b_{yx} will be multiplied by c.

For

$$\Sigma\frac{X}{c} = (\Sigma X)\frac{1}{c}$$

and

$$M_{x/c} = (M_x)\left(\frac{1}{c}\right)$$

And

$$\Sigma\left(\frac{X}{c}\right)^2 = \Sigma(X^2)\,\frac{1}{c^2}$$

Hence

$$\Sigma\left(\frac{X}{c}\right) - nM^2_{x/c} = \Sigma(X^2)\,\frac{1}{c^2} - n(M^2_x)\left(\frac{1}{c^2}\right)$$

$$= \frac{\Sigma X^2 - nM^2_x}{c^2}$$

$$= \frac{\Sigma x^2}{c^2}$$

So

$$\sigma_{x/c} = \frac{\sigma_x}{c}$$

Likewise,

$$\Sigma\left(\frac{X}{c}\right)Y = \Sigma(XY)\,\frac{1}{c}$$

Hence

$$\Sigma\left(\frac{X}{c}\right)Y - nM_{x/c}M_y = \Sigma(XY)\,\frac{1}{c} - nM_x\left(\frac{1}{c}\right)M_y$$

$$= \frac{\Sigma XY - nM_xM_y}{c}$$

$$= \frac{\Sigma xy}{c}$$

$$b_{y(x/c)} = \frac{\Sigma\left(\dfrac{x}{c}\right)y}{\Sigma\left(\dfrac{x}{c}\right)^2} = \frac{\dfrac{\Sigma xy}{c}}{\Sigma\left(\dfrac{x^2}{c^2}\right)} = \frac{\Sigma xy}{\dfrac{\Sigma x^2}{c}} = b_{yx}c$$

Similarly,

$$b_{(x/c)y} = \frac{\Sigma\left(\dfrac{x}{c}\right)y}{\Sigma y^2} = \frac{\dfrac{\Sigma xy}{c}}{\Sigma y^2} = \frac{b_{xy}}{c}$$

Note 3a (**Chapter 7**). To prove that r_{yx}, as computed by equation (23.1), equals r_{yx}, as computed by equation (27),

In equation (23.1)

$$r_{yx} = \frac{\sigma_{y'}}{\sigma_y}$$

$\sigma_{y'}$ is the σ of the series of values of Y' estimated from the equation

$$Y' = a + b_{yx}X$$

$$\text{Each } y' = Y' - M_{y'}$$

$$= (a + b_{yx}X) - \left(\frac{\Sigma a}{n} + \frac{\Sigma(b_{yx}X)}{n}\right)$$

(I) $$= b_{yx}(X - M_x) = b_{yx}x$$

and

$$b_{yx} = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - n(M_x^2)} \qquad \text{(equation [9])}$$

or in terms of departures from the mean

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

Substituting this value in equation (I) above, we find that

$$\text{each } y' = \left(\frac{\Sigma xy}{\Sigma x^2}\right)x$$

$$\text{each } (y')^2 = \left[\frac{(\Sigma xy)^2}{(\Sigma x^2)^2}\right]x^2$$

$$\Sigma(y')^2 = \left[\frac{(\Sigma xy)^2}{(\Sigma x^2)^2}\right]\Sigma x^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

and

(II) $$\sigma^2_{y'} = \frac{(\Sigma xy)^2}{n\Sigma x^2}$$

By equation (23.1)

(III) $$r^2_{yx} = \frac{(\sigma_{y'})^2}{\sigma^2_y}$$

Substituting in equation (III) the values of $\sigma^2_{y'}$ given in equation (II), we have

$$r^2_{yx} = \frac{(\Sigma xy)^2}{n\Sigma x^2}/\sigma^2_y$$

$$= \frac{(\Sigma xy)^2}{n\Sigma x^2\sigma^2_y}$$

hence

$$r^2_{yx} = \frac{(\Sigma xy)^2}{n^2\sigma^2_x\sigma^2_y}$$

and

$$r_{yx} = \frac{\Sigma xy}{n\sigma_x\sigma_y}$$

But equation (27) gives

$$r_{yx} = \frac{\Sigma(XY) - nM_xM_y}{\sqrt{[\Sigma(X^2) - nM_x^2]\,[\Sigma(Y^2) - nM_y^2]}}$$

which, stated in terms of departures from the means, becomes

$$r_{yx} = \frac{\Sigma xy}{\sqrt{n\sigma_x^2 n\sigma_y^2}}$$

$$= \frac{\Sigma xy}{n\sigma_x\sigma_y}$$

Hence equations (23.1) and (27) are identical.

Note 4 (Chapter 7). To prove for a very simple case that r_{yx}^2 measures the proportion of variance in Y explained by X. Let a, b, c, etc., be series of variables with $\sigma_a = \sigma_b = \sigma_c$, and with all intercorrelations such as r_{ab}, r_{ac}, etc., $= 0$.

Let

$$Y = a + b + c$$

$$X = a + b$$

Then

$$r_{yx}^2 = \frac{p_{yx}^2}{\sigma_x^2\sigma_y^2}$$

$$\left(\text{Here the symbol } p_{yx} \text{ is used to represent } \frac{\Sigma yx}{n}\cdot\right)$$

$$\text{each } (y)(x) = (a + b + c)(a + b)$$

$$= a^2 + 2ab + ac + b^2 + bc$$

Since

$$\Sigma(ab), \Sigma(ac), \Sigma(bc) = 0$$

$$\Sigma(y)(x) = \Sigma a^2 + \Sigma b^2$$

Similarly,

$$(y^2) = (a + b + c)^2$$

$$= a^2 + 2ab + 2ac + b^2 + 2bc + c^2$$

$$\Sigma(y^2) = \Sigma a^2 + \Sigma b^2 + \Sigma c^2$$

By similar proof,

$$\Sigma(x)^2 = \Sigma a^2 + \Sigma b^2$$

Hence

$$r_{yx}^2 = \frac{(\sigma_a^2 + \sigma_b^2)^2}{(\sigma_a^2 + \sigma_b^2 + \sigma_c^2)(\sigma_a^2 + \sigma_b^2)}$$

$$= \frac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_c^2}$$

$$= \tfrac{2}{3} \text{ (since } \sigma_a = \sigma_b = \sigma_c)$$

Similar results will be obtained for other simple combinations of elements.

Note 5 (Chapter 7). Instead of measuring the presence of correlation by comparing the standard deviation of the estimated values with the standard deviation of the actual, the amount of *absence* of correlation may be measured by comparing the standard deviation of the residuals—the standard error of estimate—with the original standard deviation. Thus in the horse-feed problem, where the coefficient of correlation—r_{yx}—was found (before adjusting) to be equal to 3.47/7.92 or +0.44, for the straight-line relation, the standard error of the residuals was equal to 7.13. If we express this in proportion to the original standard deviation, it gives us the ratio σ_z/σ_y, or, in this case, 7.13/7.92, which equals 0.90. This term has been given the name *coefficient of alienation*,[3] since it measures the *lack of correlation* in exactly the same way that the coefficient of correlation measures the presence of correlation.

In this particular case, the (unadjusted) coefficient of correlation is 0.44, and the coefficient of alienation is 0.90. The total of the two is considerably greater than 1.00. This, therefore, warns us that we cannot regard the coefficient of correlation as giving the *percentage* of correlation, or the coefficient of alienation as giving the *percentage* of absence of correlation. Except when one of the two values is equal to 0, the sum of the two will always be greater than unity.

If we look back to the values from which these coefficients were computed—the standard deviation of the original dependent series, the standard deviation of the estimated values, and the standard deviation of the residuals, it is easy to see why this is so. The standard deviation of the original values, $\sigma_y = 7.92$; the standard deviation of the estimated values, $\sigma_{y'}$, $= 3.47$; and the standard deviation of the residuals, σ_z, $= 7.13$. When we add the last two together, we find they equal more than the original standard deviation. Therefore, when we express each of them as a percentage of this original standard deviation, the sum of the two values is more than unity. But if we square these standard deviations, we find that $\sigma_y^2 = 62.73$, $\sigma_{y'}^2 = 12.04$, and $\sigma_z^2 = 50.84$. Here $\sigma_{y'}^2$ plus the σ_z^2, 12.04 plus 50.84, is equal to 62.88, practically identical with σ_y^2, 62.73.[4] This will always hold true, as each individual $Y' + z = Y$. It has already been proved (Note 1), that when Y' and z are not correlated, and $Y = Y' + z$, that then

$$\sigma_{y'}^2 + \sigma_z^2 = \sigma_y^2$$

which we have just observed to hold true in this case.

If we measure the amount of correlation by dividing $\sigma_{y'}^2$ by σ_y^2, and the lack of correlation by dividing σ_z^2 by σ_y^2, we shall have two measures whose sum will always equal unity, so that when we know what one is, we can tell the other immediately. These values, $\sigma_{y'}^2/\sigma_y^2$, and σ_z^2/σ_y^2, are known respectively as the *coefficient of determination* and the *coefficient of non-determination*. They equal the square of the coefficient of correlation (r^2), and the square of the coefficient of alienation (k^2). In this case these values, $(0.44)^2$ and $(0.90)^2$, are 0.19 and 0.81, showing that (if the adjustment for the number of cases is ignored) 19 per cent of the variance in feed is associated with days worked, and 81 per cent is not so associated.

The adjustment of the coefficient of non-determination for the effect of small samples, to obtain an unbiased estimate of the most probable value of k in the universe from which the sample is drawn, is

$$\bar{k}^2 = \frac{(n-1)\left(\dfrac{\sigma_z^2}{\sigma_y^2}\right)}{n-m} \tag{102}$$

[3] Truman L. Kelley, *Statistical Method*, pp. 173–175, The Macmillan Co., New York, 1924.

[4] The slight difference is due to rounding off decimals in entering the Y' values.

Applying this adjustment, $\bar{k}^2$ for the horse-feed problem becomes 0.86, indicating that in the universe, it is likely that 86 per cent of the variance is independent of the days worked.

Since the several measures of correlation are all derived from the standard deviations of Y, Y', and z, certain mathematical relations always hold among the unadjusted coefficients, as is shown following:

$$\sigma^2_{y'} + \sigma^2_z = \sigma^2_y$$

$$\frac{\sigma^2_{y'}}{\sigma^2_y} + \frac{\sigma^2_z}{\sigma^2_y} = 1$$

$$r^2_{xy} + k^2_{xy} = 1 \text{ } (k = \text{the coefficient of alienation})$$

or

$$d_{xy} + k^2_{xy} = 1$$

$$r = \sqrt{1 - k^2}$$

also,

$$\frac{\sigma^2_z}{\sigma^2_y} = 1 - \frac{\sigma^2_{y'}}{\sigma^2_y} = 1 - r^2_{xy}$$

hence

$$\sigma^2_z = (1 - r^2_{xy})\sigma^2_y$$

or

$$\sigma_z = \sigma_y\sqrt{1 - r^2_{xy}}$$

This last equation is useful in calculating S_{yx}, the standard error in estimating Y from known values of X, when only the standard deviation of Y and the correlation of X with Y are known. As is shown in Chapter 8, the coefficient of correlation can be computed directly without first computing all the estimated values of Y (the Y' values) or without computing the individual residuals. When the correlation coefficient is thus computed, this last equation provides a short cut to tell what the errors in estimating values of Y from the known values of X according to the straight-line relation are likely to be.

Similarly, with curvilinear relations,

$$\sigma^2_{y''} + \sigma^2_{z''} = \sigma^2_y$$

and

$$\frac{\sigma^2_{y''}}{\sigma^2_y} + \frac{\sigma^2_{z''}}{\sigma^2_y} = 1$$

Hence

$$\rho^2 + \frac{\sigma^2_{z''}}{\sigma^2_y} = 1$$

and

$$\rho^2 = 1 - \frac{\sigma^2_{z''}}{\sigma^2_y}$$

These relations hold precisely true only when ρ is calculated for a mathematically determined regression curve. For freehand curves, they are only approximately correct.

Note 6 (Chapter 12). The normal equations, to determine the "best" regression values for two or more independent variables, are derived by exactly the same process as given in full in Note 2. Thus to determine the constants in the equation

$$X_1 = a + b_2X_2 + b_3X_3 + b_4X_4$$

The value to be made a minimum is

$$\Sigma(X_1 - a - b_2X_2 - b_3X_3 - b_4X_4)^2$$

Differentiating this with respect to a, b_2, b_3, and b_4, setting the partial derivatives equal to zero, and transposing give the normal equations, stated in terms of sums of X_1, X_2, etc. The equation may also be stated in terms of deviations from the means

$$x_1 = b_2x_2 + b_3x_3 + b_4x_4$$

This, by the same derivation, gives the normal equations as given in the text (equation [38]). The constant, a, is then found separately by equation (39), which represents simply the separate solution of the first normal equation given by the first derivation,

$$\Sigma X_1 = na + \Sigma(X_2)b_2 + \Sigma(X_3)b_3 + \Sigma(X_4)b_4$$

Hence

$$a = \frac{\Sigma X_1}{n} - \frac{\Sigma X_2}{n}b_2 - \frac{\Sigma(X_3)}{n}b_3 - \frac{\Sigma(X_4)}{n}b_4$$

This readily reduces to equation (39).

Note 7 (Chapter 13). Coefficients of partial correlation are usually defined by the formula

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}}$$

For coefficients with more variables eliminated, such as $r_{12.345}$, for example, this becomes

$$r_{12.345} = \frac{r_{12.34} - r_{15.34}r_{25.34}}{\sqrt{(1 - r_{15.34}^2)(1 - r_{25.34}^2)}}$$

To determine the coefficients with several factors held constant by this method involves a lengthy process of elimination, variable by variable; and for that reason the method presented in the text is preferred as shorter, simpler, and more readily subject to checking.

Note 8 (Chapter 13). It can be proved that the coefficient of multiple determination (R^2) measures the percentage of variance ascribable to the several independent factors for certain simple cases. Thus assume four variables, A, B, C, D, with all intercorrelations equal to 0, and all σ equal. Let $Y = A + B + C$. Then correlate Y with A, B, and D. The regression equation will work out

$$Y = a + A + B[+0(D)]$$

Computing $R_{Y.ABD}$ by equation (46),

$$R_{Y.ABD}^2 = \frac{b_{ya.bd}\Sigma ya + b_{yb.ad}\Sigma yb + b_{yd.ab}\Sigma yd}{\Sigma y^2}$$

Each

$$(y)(a) = (a + b + c)(a) = a^2 + ab + ac$$

$$\Sigma(ya) = \Sigma a^2 + \Sigma ab + \Sigma ac = \Sigma a^2 \text{ (Since } r_{ab} = r_{ac} = 0)$$

Similarly

$$\Sigma(yb) = \Sigma b^2; \ \Sigma(yd) = \Sigma d^2$$

And each

$$(y^2) = (a + b + c)^2 = a^2 + 2ab + b^2 + 2ac + 2bc + c^2$$

and

$$\Sigma(y^2) = \Sigma a^2 + \Sigma b^2 + \Sigma c^2$$

Hence

$$R^2_{Y.ABD} = \frac{(1)(\Sigma a^2) + 1(\Sigma b^2) + 0(\Sigma d^2)}{\Sigma a^2 + \Sigma b^2 + \Sigma c^2}$$

And since all σ's are identical,

$$R^2_{Y.ABD} = \tfrac{2}{3}$$

In this case, then, when Y, composed of three equally variable non-correlated elements, is correlated with two of those elements, and with one other equal element which is not represented in Y and which is not correlated with elements present in Y, the multiple determination of Y by the two elements (A and B) is found to be $\frac{2}{3}$.

Similar results will be secured for other experimental cases which may be set up.

Note 9 (Chapter 13). The coefficient of part correlation was first worked out by B. B. Smith, in collaboration with the present author, and was first published in Correlation theory and method applied to agricultural research, pp. 57–60, Bureau of Agricultural Economics, U. S. Department of Agriculture, August, 1926, (a mimeographed publication).

The formula for part correlation coefficients is derived as follows:

Let ${}_{12}r_{34}$ equal correlation of x_2 with $x_1 - b_{13.24}x_3 - b_{14.23}x_4$

Let

$$z = x_1 - b_2x_2 - b_3x_3 - b_4x_4$$

$$z + b_2x_2 = x_1 - b_3x_3 - b_4x_4$$

Then

$$({}_{12}r_{34})^2 = \frac{[\Sigma(z + b_2x_2)(x_2)]^2}{[\Sigma(z + b_2x_2)^2][\Sigma x_2^2]}$$

each

$$(z + b_2x_2)(x_2) = zx_2 + b_2x_2^2$$

$$\Sigma(z + b_2x_2)(x_2) = b_2\Sigma x_2^2 \qquad (r_{zx_2} = 0, \text{ hence } \Sigma zx_2 = 0)$$

each

$$(z + b_2x_2)^2 = z^2 + 2b_2x_2z + b_2^2x_2^2$$

$$\Sigma(z + b_2x_2)^2 = \Sigma z^2 + b_2^2\Sigma x_2^2$$

So

$$({}_{12}r_{34})^2 = \frac{b_2^2(\Sigma x_2^2)^2}{(\Sigma z^2 + b_2^2\Sigma x_2^2)(\Sigma x_2^2)} = \frac{b_2^2\Sigma x_2^2}{\Sigma z^2 + b_2^2\Sigma x_2^2}$$

$$= \frac{1}{\dfrac{\Sigma z^2}{b_2^2\Sigma x_2^2} + 1} = \frac{1}{\dfrac{n\sigma_1^2(1 - \bar{R}^2_{1.234})}{b_2^2n\sigma_2^2} + 1}$$

$$= \frac{1}{1 + \dfrac{\sigma_1^2(1 - \bar{R}^2_{1.234})}{b_2^2\sigma_2^2}}$$

Note 10 (Chapter 13). The coefficient of partial correlation, as used in biometric work, is employed to measure the correlation in an hypothetical universe, from which all variation due to changes in the eliminated factors has been excluded. This correlation may then be compared to the simple correlation of the two factors, to see whether the closeness of relation is improved or not by excluding the variation *in both variables* associated with other factors. With the coefficient of part correlation, on the contrary, all the original variation in the independent factor is left in it, and only the dependent factor is adjusted.

Note 11 (Chapter 13). Separate determination. Neither the coefficients of partial determination nor the coefficients of part determination equal, when totaled, the multiple determination of X_1 by X_2, X_3, and X_4. That is b cause both these types of measures are computed on bases which change from variable to variable; hence their sums, when several are added together, have no mathematical significance. There is, however, a third type of coefficient, which parcels out among each of the several independent variables that part of the variation in the dependent variable which each one of them seems able to account for, when estimates of the dependent variable are made from all of the independent variables. To distinguish this type from the coefficients previously presented, they may be termed *coefficients of separate determination.*

Using $d_{12.34}$ to represent the separate determination of X_1 by X_2, when X_3 and X_4 are also considered, we may compute it by the formula

$$\bar{d}_{12.34} = \left[\frac{b_{12.34}(\Sigma x_1x_2)}{\Sigma(x_1^2)}\right]\left[\frac{\bar{R}^2_{1.234}}{R^2_{1.234}}\right] \tag{103}$$

Each of these values has been used previously in computing the coefficient of multiple correlation, so the determination coefficient may be readily calculated.

When the several coefficients of separate determination are added together, their sum is equal to the coefficient of multiple determination, $\bar{R}^2$. Comparing the last equation with equation (46) for R^2, we readily see that

$$\bar{R}^2_{1.234} = \left[\frac{b_{12.34}(\Sigma x_1x_2)}{\Sigma(x_1^2)} + \frac{b_{13.24}(\Sigma x_1x_3)}{\Sigma(x_1^2)} + \frac{b_{14.23}(\Sigma x_1x_4)}{\Sigma(x_1^2)}\right]\left[\frac{\bar{R}^2_{1.234}}{R^2_{1.234}}\right]$$

The three terms on the right are the three coefficients of separate determination, $d_{12.34}$, $d_{13.24}$, and $d_{14.23;}$ These coefficients are the simplest to compute of any of the three types which have been discussed, the computation of their values being readily made a part of the process of working out R and S.

Working out the last equation for acres and income, in the 4-variable problem, we find that it becomes

$$\bar{d}_{12.34} = \left[\frac{(1.20584)(0.71)}{(272.76)}\right]\left[\frac{0.806}{0.8366}\right] = \frac{0.690}{228.19} = 0.00302$$

The corresponding values are 0.630 for the separate determination of income by cows and 0.171 for the separate determination of income by number of men.

The coefficients of "separate" determination are the easiest of all to compute, and have the further advantage of adding to a definite sum ($\bar{R}^2$), and hence being directly comparable one with another. The disadvantage in their use, however, is that under certain conditions the value of one or more coefficients will prove to be negative. Off-hand it seems difficult to explain how the "determination" of any variable can be less than nothing. (This result will be obtained whenever the gross or apparent correlation and the coefficient of net or partial regression are of opposite sign.) The explanation is simple, however. Although the total variation in the estimates of the dependent variable is obtained by adding the contributions from the several independent variables, it does not follow that all variables will be influencing the estimate in the same direction at the same time—all tending to give low values when the actual value is low, or all tending to give high values when the actual value is high. It sometimes happens that one variable may tend to work counter to the other variables, usually preventing the final estimate from going so low as it otherwise would when the general effect is downward, and tending to keep it from going so high as it otherwise would when the others are forcing it up. It is under such conditions that negative coefficients of separate determination are obtained; they do not mean that the variable has no significance, but that its influence is usually exerted counter to the influence of other variables.

When there is very high intercorrelation between the several independent variables, the coefficients of separate determination may vary quite erratically, and hence become of little significance. Under such conditions other measures of the individual importance of the several factors will need to be employed.

Although nothing is known of the sampling error involved in determining coefficients of separate determination, since they are computed from standard deviations, product sums, and net regression coefficients, their standard error must be some function of the standard error of these other coefficients. It is under the conditions noted in the preceding paragraph that net regression coefficients have the least reliability, so it may be that a problem which fails to yield reasonable separate determination coefficients may also fail to yield reliable values for the other measures of determination.

There seems to be evidence that coefficients of separate determination are less stable, and more subject to random error, than any other measure of the importance of individual factors. On account of their ease of computation, they have been much used in the past; but it is doubtful how much confidence can be placed in them. For that reason it seems best to use the other measures, and discard this measure until its reliability has been more definitely determined.

The relation between beta coefficients and coefficients of separate determination may be shown algebraically.

The normal equations for determining the regression coefficients

$$(\mathrm{I})\quad \left\{\begin{aligned} \Sigma x_2^2 b_2 + \Sigma x_2x_3b_3 + \Sigma x_2x_4b_4 &= \Sigma x_1x_2 \\ \Sigma x_2x_3b_2 + \Sigma x_3^2 b_3 + \Sigma x_3x_4b_4 &= \Sigma x_1x_3 \\ \Sigma x_2x_4b_2 + \Sigma x_3x_4b_3 + \Sigma x_4^2 b_4 &= \Sigma x_1x_4 \cdot \end{aligned}\right.$$

may also be written after dividing through by n, the number of cases, and dividing each line and column by the corresponding standard deviation. Solution of these equations, which are shown below, then gives the values for the partial (net) beta coefficients.

$$(\mathrm{II})\quad \left\{\begin{aligned} \beta_2 + r_{23}\beta_3 + r_{24}\beta_4 &= r_{12} \\ r_{23}\beta_2 + \beta_3 + r_{34}\beta_4 &= r_{13} \\ r_{24}\beta_2 + r_{34}\beta_3 + \beta_4 &= r_{14} \end{aligned}\right.$$

For the first set of equations (I) the separate determination of X_1 by X_2, $d_{12.34}$, is given by the equation

$$d_{12.34} = \frac{b_2 \Sigma x_1 x_2}{\Sigma x_1^2} \left(\text{i.e.,} = \frac{b_{12.34} \Sigma x_1 x_2}{\Sigma x_1^2} \right)$$

In terms of the values given in the second set of equations (II), the coefficient of separate determination would be computed

$$d_{12.34} = \beta_2 r_{12} (\text{i.e.,} = \beta_{12.34} r_{12})$$

Substituting the value for r_{12} given by the first equation of the second set, we see that this becomes

$$d_{12.34} = \beta_2(\beta_2 + r_{23}\beta_3 + r_{24}\beta_4)$$

$$= \beta_2^2 + r_{23}\beta_2\beta_3 + r_{24}\beta_2\beta_4$$

Similarly,

$$d_{13.24} = \beta_3^2 + r_{23}\beta_2\beta_3 + r_{34}\beta_3\beta_4$$

and

$$d_{14.23} = \beta_4^2 + r_{24}\beta_2\beta_4 + r_{34}\beta_3\beta_4$$

It is evident from this that each coefficient of separate determination consists of one portion, β^2, which is, as Dr. Sewall Wright named it, the "direct determination" by that independent variable; plus (or minus) a pro-rated share of the joint determination of that variable with each other independent variable. Since $r_{23}\beta_2\beta_3$ contributes equally to both $d_{12.34}$ and $d_{13.24}$, this "joint determination" is simply divided equally between both independent variables. As only the direct determination (β^2) can be said to reflect the *separate influence* of the particular independent variable, the further attempt to allocate or split up the joint influence is unsatisfactory. For that reason, the several betas (or their squares) seem the best measures of the separable importance of each variable, the combined influence of variables acting jointly being left out of the distribution.

This explanation follows that developed by H. R. Tolley, and presented in full in the bulletin by F. F. Elliott, Adjusting hog production to market demand, *University of Illinois Agricultural Experiment Station Bulletin* 293, 1927. See also Sewall Wright, Correlation and Causation, *Journal Agricultural Research*, Vol. XX, No. 7, pp. 557–575.

Note 12 (Chapter 13). Given the multiple regression equation

$$X_1 = a + b_2X_2 + b_3X_3 + b_4X_4$$

let $R_{(X_1 - b_2X_2).3,4}$ represent the correlation between $[X_1 - b_2X_2]$, and X_3 and X_4.

To find the formula for this correlation:

Using departures from their means for all variables

$$z = x_1 - b_2x_2 - b_3x_3 - b_4x_4$$

hence

$$b_3x_3 + b_4x_4 = x_1 - b_2x_2 - z$$

The required correlation is therefore that between

$$[x_1 - b_2x_2] \text{ and } [x_1 - b_2x_2 - z]$$

From equation (27) it is equal to

$$\frac{\Sigma[(x_1 - b_2x_2)(x_1 - b_2x_2 - z)]}{\sqrt{\Sigma(x_1 - b_2x_2)^2\Sigma(x_1 - b_2x_2 - z)^2}}$$

Each

$$(x_1 - b_2x_2)(x_1 - b_2x_2 - z) = (x_1 - b_2x_2)^2 - zx_1 + b_1zx_2$$

and

$$\Sigma[(x_1 - b_2x_2)(x_1 - b_2x_2 - z)] = \Sigma(x_1 - b_2x_2)^2 - \Sigma zx_1 + b_1\Sigma zx_2$$

Since z is uncorrelated with x_2, $\Sigma zx_2 = 0$.
The value of Σzx_1 may be evaluated as follows:

$$x_1 = b_2x_2 + b_3x_3 + b_4x_4 + z$$

$$zx_1 = b_2zx_2 + b_3z_3 + b_4zx_4 + z^2$$

$$\Sigma zx_1 = b_2\Sigma zx_2 + b_3\Sigma zx_3 + b_4\Sigma zx_4 + \Sigma z^2$$

But

$$r_{z_2} = r_{z_3} = r_{z_4} = 0$$

hence

$$\Sigma zx_1 = \Sigma z^2$$

Therefore,

$$\Sigma[(x_1 - b_2x_2)(x_1 - b_2x_2 - z)] = \Sigma(x_1 - b_2x_2)^2 - \Sigma z^2$$

Similarly,

$$\Sigma(x_1 - b_2x_2 - z)^2$$

may be shown to equal

$$\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2$$

Hence

$$R_{(X_1 - b_2X_2).3,\,4} = \frac{\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2}{\sqrt{\Sigma(x_1 - b_2x_2)^2}\sqrt{\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2}}$$

And

$$R^2_{(X_1 - b_2X_2).3,\,4} = \frac{[\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2]^2}{[\Sigma(x_1 - b_2x_2)^2][\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2]}$$

$$= \frac{\Sigma(x_1 - b_2x_2)^2 - \Sigma z^2}{\Sigma(x_1 - b_2x_2)^2}$$

$$= 1 - \frac{\Sigma z^2}{\Sigma(x_1 - b_2x_2)^2}$$

$$= 1 - \frac{\sigma_z^2}{\sigma_1^2 - 2b_2p_{12} + b_2^2x_2^2}$$

$$= 1 - \frac{\sigma_1^2(1 - R^2_{1.234})}{\sigma_1^2 - 2b_2\dfrac{\Sigma x_1x_2}{n} + b_2^2\sigma_2^2}$$

Note 13 (for Appendix on Methods of Computation.) To prove that the equation (101)

$$\sigma_{b_{13.24}} = \bar{S}_{1.234}\sqrt{c_{33}}$$

gives the same value as that given by equation (74)

$$\sigma_{b_{13.24}} = \sqrt{\frac{\bar{S}^2_{1.234}}{\sigma_3^2(1 - R^2_{3.24})n}} \tag{74}$$

For a problem in two independent variables, c_{33} is obtained by the simultaneous solution of the equations

$$\begin{aligned} \Sigma(x_2^2)c_{32} + \Sigma(x_2x_3)c_{33} &= 0 \\ \Sigma(x_2x_3)c_{32} + \Sigma(x_3^2)c_{33} &= 1 \end{aligned}$$

Solving by the Doolittle method, we have

$$\begin{array}{r} \Sigma(x_2^2)c_{32} + \Sigma(x_2x_3)c_{33} = 0 \\ -\, c_{32} - \dfrac{\Sigma(x_2x_3)}{\Sigma x_2^2}\, c_{33} = 0 \\ \hline \Sigma(x_2x_3)c_{32} + \Sigma(x_3^2)c_{33} = 1 \\ -\, \Sigma(x_2x_3)c_{32} - \dfrac{[\Sigma(x_2x_3)]^2}{\Sigma(x_2^2)}\, c_{33} = 0 \\ \hline c_{33}\left[\Sigma(x_3^2) - \dfrac{[\Sigma(x_2x_3)]^2}{\Sigma(x_2^2)}\right] = 1 \end{array}$$

and

$$\begin{aligned} \frac{1}{c_{33}} &= \Sigma(x_3^2) - \frac{[\Sigma(x_2x_3)]^2}{\Sigma(x_2)^2} \\ &= \Sigma x_3^2\left[1 - \frac{[\Sigma(x_2x_3)]^2}{\Sigma(x_2)^2\Sigma(x_3)^2}\right] \\ &= n\sigma_3^2(1 - r_{23}^2) \end{aligned}$$

Hence

$$c_{33} = \frac{1}{n\sigma_3^2(1 - r_{23}^2)}$$

Substituting this value for c_{33} in equation (101), we obtain

$$\begin{aligned} \sigma_{b_{13\cdot 2}} &= \bar{S}_{1.23}\sqrt{c_{33}} \\ &= \bar{S}_{1.23}\sqrt{\frac{1}{n\sigma_3^2(1 - r_{23}^2)}} \\ &= \sqrt{\frac{\bar{S}^2_{1.23}}{n\sigma_3^2(1 - r_{23}^2)}} \end{aligned}$$

This is seen to be identical with the value given by equation (74), when written for the corresponding coefficient. The equations to determine c_{22} for a problem of three independent variables are

$$\begin{aligned}
\Sigma(x_2^2)c_{22} + \Sigma(x_2x_3)c_{23} + \Sigma(x_2x_4)c_{24} &= 1 \\
\Sigma(x_2x_3)c_{22} + \Sigma(x_3^2)c_{23} + \Sigma(x_3x_4)c_{24} &= 0 \\
\Sigma(x_2x_4)c_{22} + \Sigma(x_3x_4)c_{23} + \Sigma(x_4^2)c_{24} &= 0
\end{aligned}$$

If these equations are solved simultaneously, it will be found that the value for c_{22} will be

$$c_{22} = \frac{1}{n\sigma_2^2(1 - R_{2.34}^2)}$$

and substituting this in equation (101) will again show that result to be identical with equation (74). This same proof may be carried through for any number of independent variables.

APPENDIX 3

CHARTS FOR INTERPRETING OR ADJUSTING CORRELATION CONSTANTS

Reliability of small samples. The accompanying figures will serve to facilitate and simplify many of the computations which are discussed in the text.

Figure A is an extension of Table A, in Chapter 2. For random samples of varying sizes, it gives the average proportion of samples of each given size in which the observed mean will miss the true mean in the universe by more than the stated multiple of the standard error of the mean, as computed from each sample. The figure is drawn for 2, 3, 4, 5, 6, 7, 8, 10, 12, 16, 20, 30, and ∞ observations, and may be used for any desired multiple of the standard error from 1.0 to 6.0. By interpolation, values for intermediate sizes of samples may be read. The figure is read by entering with the desired multiple of the standard error (shown at the bottom) and noting the ordinate where the line for the given number of observations intersects that abscissa. The ordinate then gives the average proportion of samples in which such a departure will occur solely by chance. The figure may also be entered with a desired probability and the given number of observations, to determine what multiple of the standard error must be taken to give that degree of reliability. Thus if with 10 observations a reliability of 0.95 was desired, the figure indicates 2.26 times the standard error. That is, with 10 observations, in 5 samples out of 100, on the average, the true mean will not come within the range covered by observed mean ±2.26 S.E., if the sample was drawn under the conditions assumed in random sampling.

Just as with Table A, Figure A may be used to judge the reliability of certain other coefficients, by subtracting 1 from the number of observations for each additional degree of freedom removed in determining the constant. For coefficients of simple regression, 1 must be subtracted; for partial or multiple regression coefficients, subtract the number of independent variables from n; for curvilinear regressions, subtract $(m - 1)$.

Figure A is based upon the results given by "Student" in his article, New tables for testing the significance of observations, *Metron* V, No. 3, pp. 105–120, 1925. It is comparable to Fisher's t table, with n here equal to Fisher's n', or his $n + 1$.

Reliability of observed correlations. Figures B, C, D, and E have been discussed in Chapter 18, pages 319 to 324. These figures provide a ready means of judging the probable minimum value for the correlation in the universe, with any observed value and any given size of sample. The chart is entered with the observed correlation as abscissa; the ordinate for the intersection of that abscissa with the curve for the given size of sample gives the probable correlation. Thus if a coefficient of simple correlation, $r_{xy} = 0.65$, is obtained from a sample of 22 cases, the researcher will know from Figure B that, if he makes the statement that the true correlation in the universe is at least 0.38, he will be wrong in only 5 per cent of such statements,

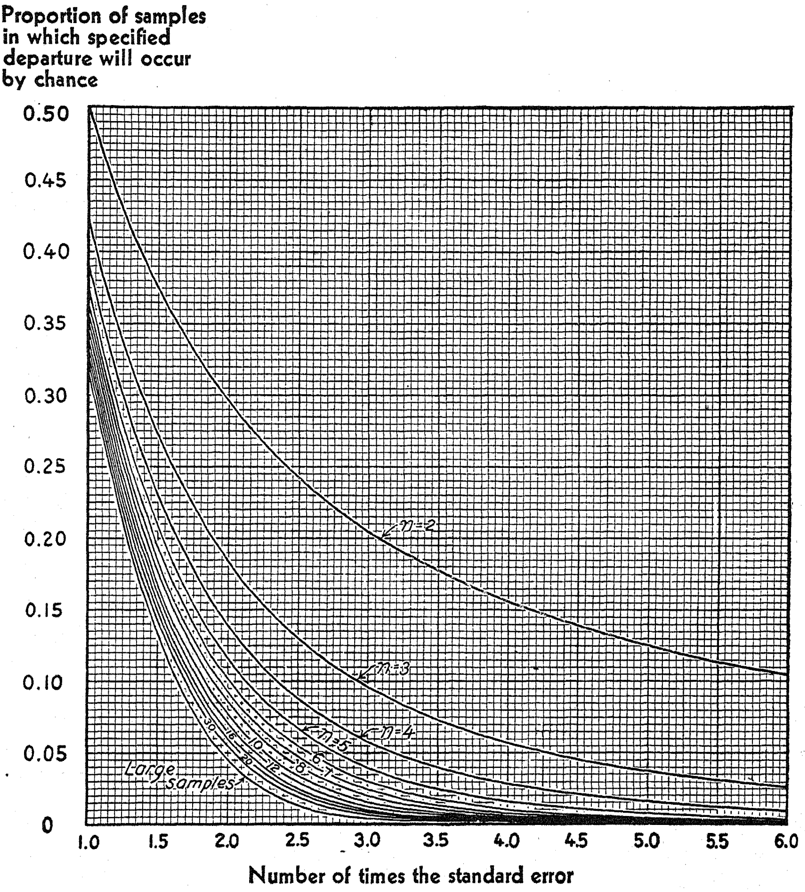


Fig. A. The proportion of random samples in which the observed mean will miss the true mean by more than the stated multiple of the standard error computed from the sample, for samples with 2, 3, 4, 5, 6, 7, 8, 10, 12, 16, 20, 30, or ∞ observations. (To apply to coefficients of regression, see footnote to Table A, page 23.)

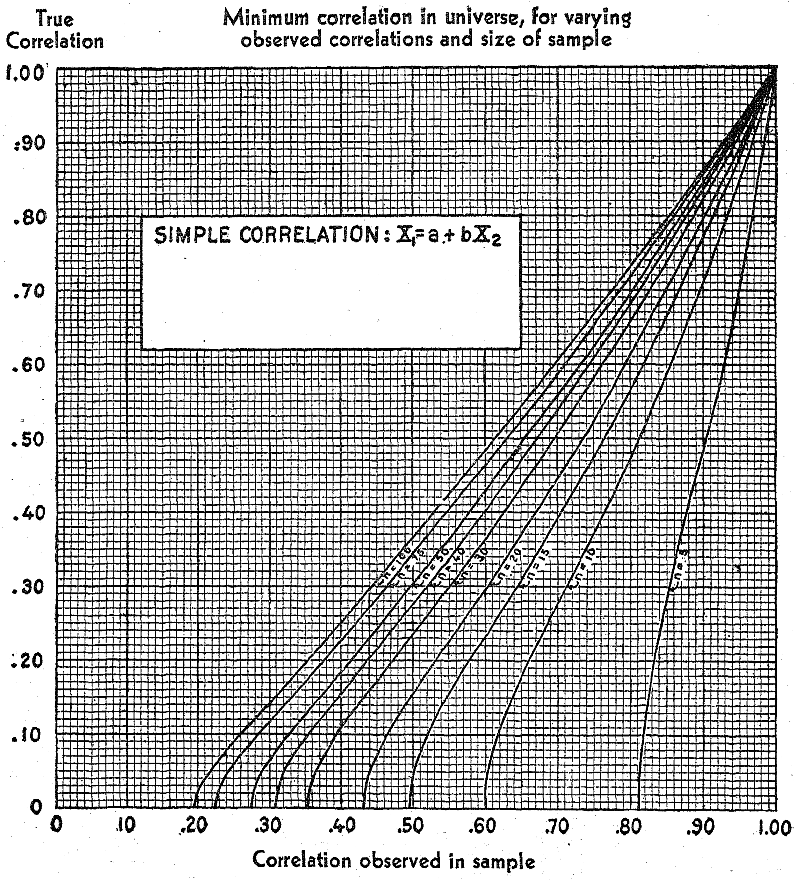


FIG. B. Under conditions of random sampling, one sample out of twenty, on the average, will show a correlation coefficient with a ± value as high as that "observed in sample," when drawn from a universe with the stated true correlation.

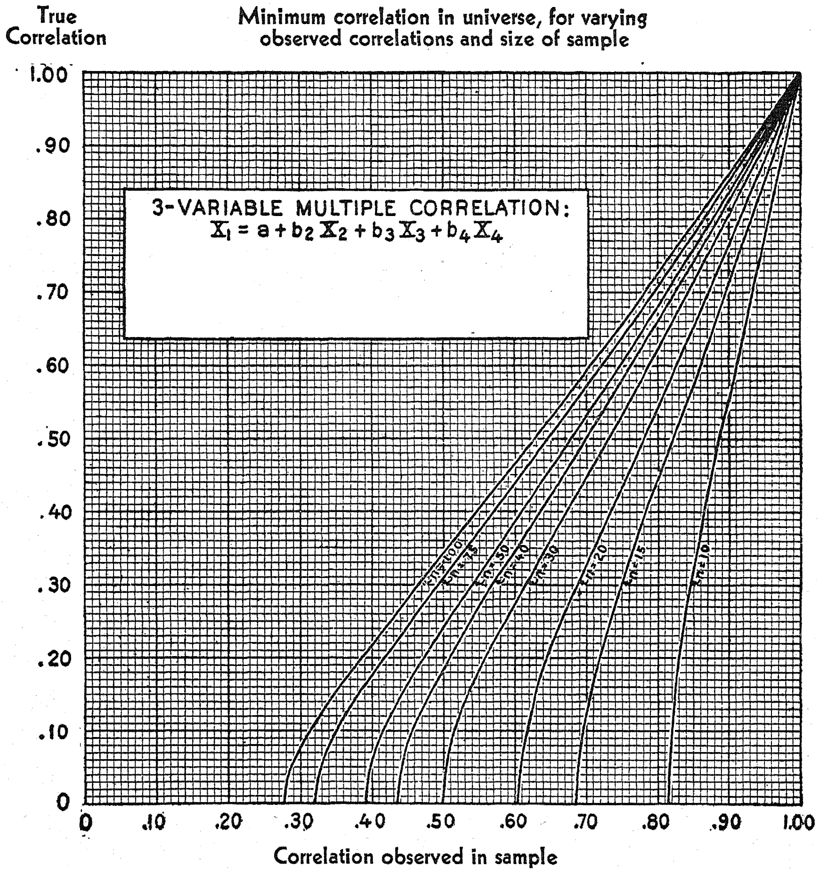


FIG. C. Under conditions of random sampling, one sample out of twenty, on the average, will show a multiple correlation as high as that "observed in sample," when drawn from a universe with the stated true multiple correlation, in the case of multiple correlation with three independent variables.

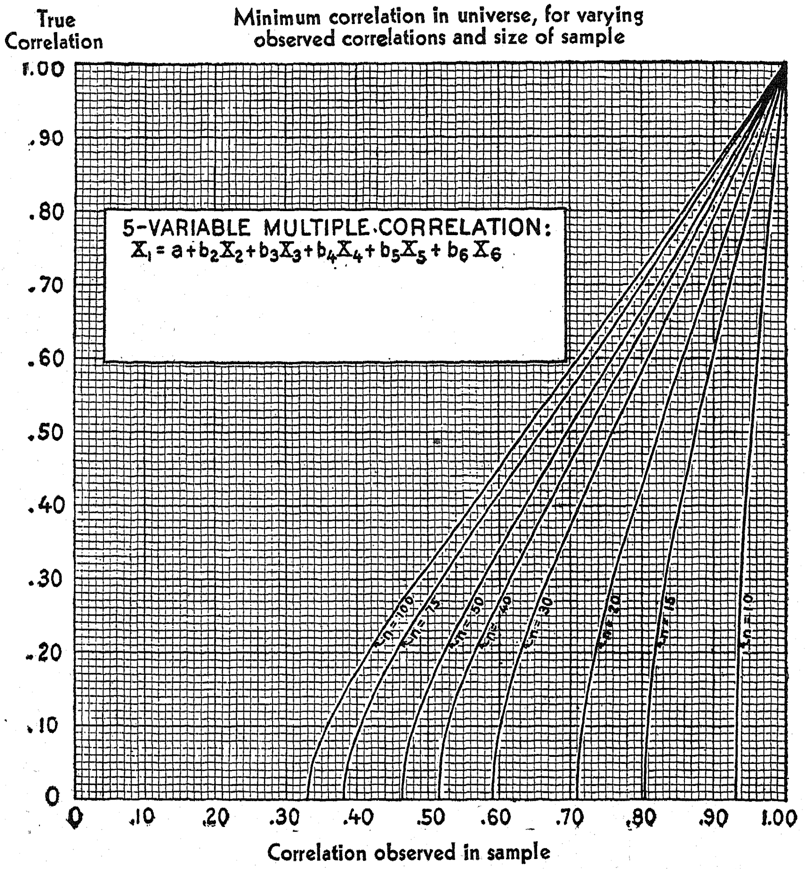


FIG. D. Under conditions of random sampling, one sample out of twenty, on the average, will show a multiple correlation as high as that "observed in sample," when drawn from a universe with the stated true multiple correlation, in the case of multiple correlation with five independent variables.

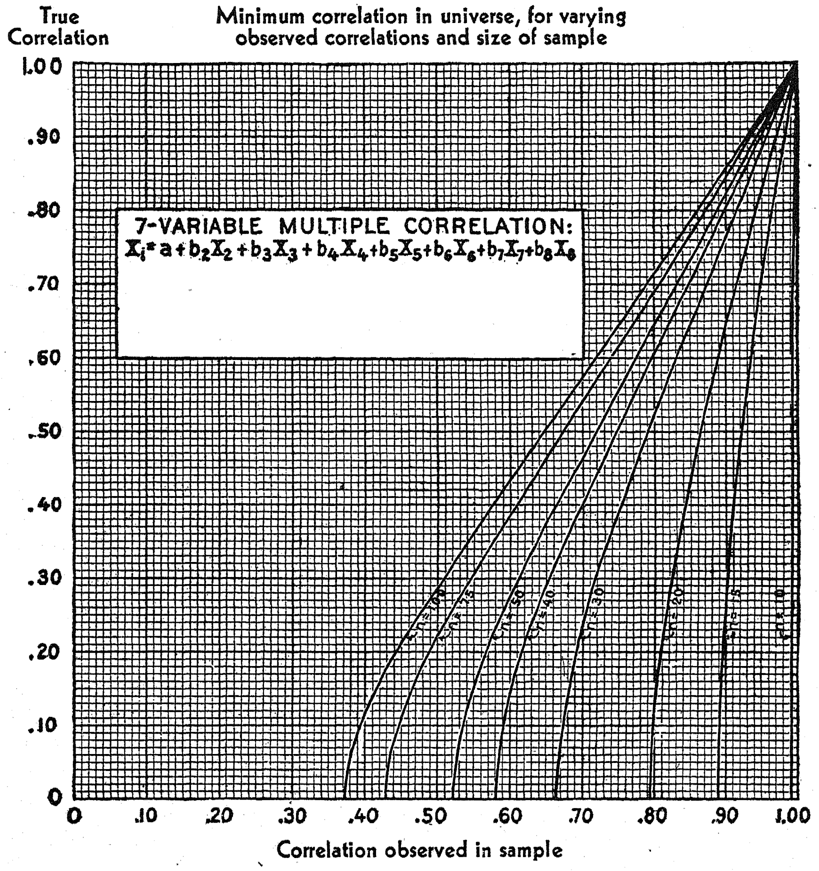
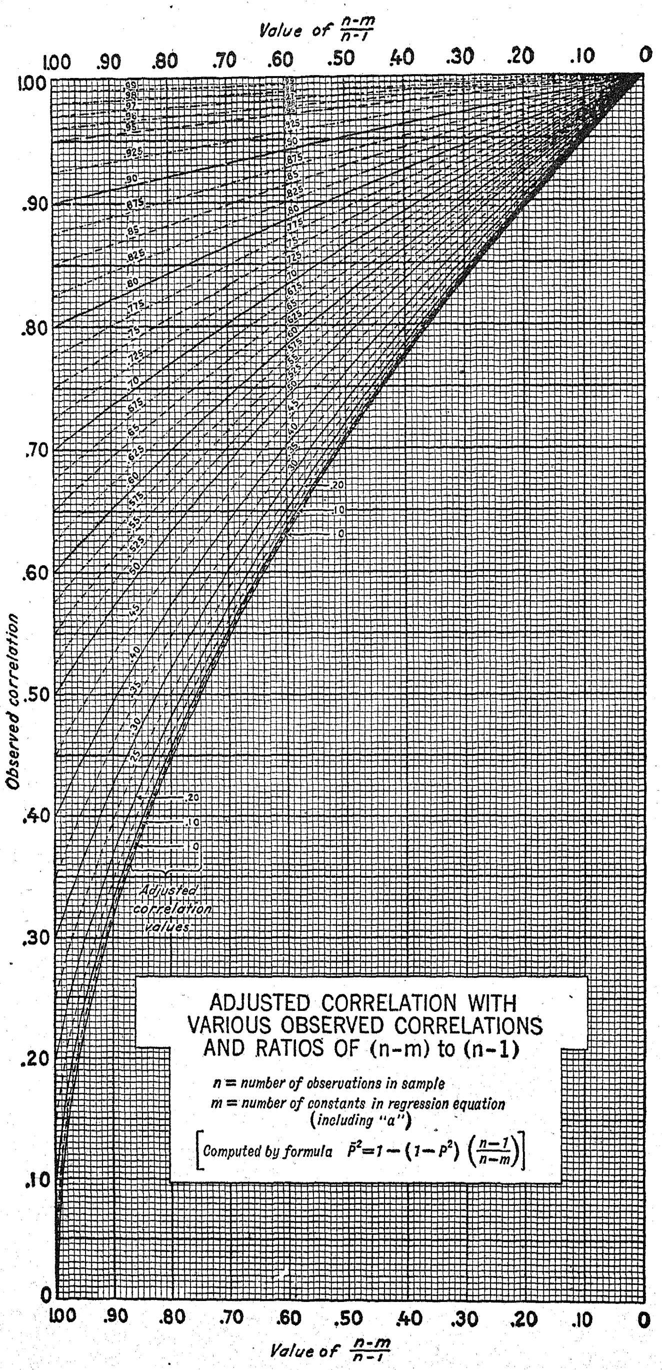


FIG. E. Under conditions of random sampling, one sample out of twenty, on the average, will show a multiple correlation as high as that "observed in sample," when drawn from a universe with the stated true multiple correlation, in the case of multiple correlation with seven independent variables.



ADJUSTED CORRELATION WITH VARIOUS OBSERVED CORRELATIONS AND RATIOS OF (n-m) to (n-1)

n = number of observations in sample

m = number of constants in regression equation (including "a")

[*Computed by formula* $\bar{P}^2 = 1 - (1 - P^2)\left(\frac{n-1}{n-m}\right)$]

Fig. F. This chart provides a graphic means of calculating the adjusted coefficient or index of correlation, as shown in formulas (25), (26), (47), and (66.3).

on the average. Figure C applies to $R_{1.234}$, Figure D to $R_{1.23456}$, and Figure E to $R_{1.2345678}$. Values for 2, 4, and 6 independent variables may be obtained by interpolation. The figures are based upon the researches of R. A. Fisher, summarized in his publication, The general sampling distribution of the multiple correlation coefficient, *Proceedings of the Royal Society,* A, Vol. 121, pp. 655–673, 1928. The tables in that article assume a large sample, and therefore give only approximately correct values when applied to small samples. For that reason the values shown by Figures B to E do not agree precisely with the exact values given in the corresponding tables in Fisher's article or in Wishart's tables (cited on page 522) for the value of observed correlations when the samples are drawn from a universe with zero correlation. The differences are so slight, however, that Figures B to E are quite adequate for practical purposes.

Adjustment of correlation for size of sample. Figure F may be used to facilitate the calculation of adjusted coefficients and indexes of correlation, $\bar{r}$, $\bar{R}$, $\bar{\rho}$, or $\overline{\mathrm{P}}$, from the unadjusted values r, R, ρ, or P. All that is necessary is to calculate the ratio $\frac{n - m}{n - 1}$, and enter the chart with that as abscissa and the observed correlation as ordinate. Then note the value of $\bar{r}$ given by the curve which lies nearest the intersection of the two coordinates, or interpolate between the two nearest curves. Thus with $\mathrm{P} = 0.70$, $n = 30$, and $m = 8$, $\frac{n - m}{n - 1} = 0.759$, and $\overline{\mathrm{P}} = 0.57$. Likewise foı $R_{1.234} = 0.88$, and $n = 20$; $\frac{n - m}{n - 1} = \frac{16}{19} = 0.842$, and $\bar{R}_{1.234} = 0.85$.

APPENDIX 4

LIST OF IMPORTANT EQUATIONS

For convenience in referring to the most important of the equations which are introduced from time to time in the text, all numbered equations are repeated here in numerical order.

$$M_x = \frac{\Sigma X}{n} \tag{1}$$

$$X - M_x = x \tag{2}$$

$$\delta = \frac{\Sigma x \text{ (without regard to sign)}}{n} \tag{3}$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n}} \tag{4}$$

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{n} - M_x^2} \tag{5}$$

$$\sigma_u = \sqrt{\frac{\Sigma(d^2F)}{n} - \left[\frac{\Sigma(dF)}{n}\right.^{2} - \frac{c^2}{12}} \tag{6}$$

$$\bar{\sigma} = \sigma\sqrt{\frac{n}{n-1}} \tag{6.1}$$

$$\bar{\sigma} = \sqrt{\frac{\Sigma x^2}{n-1}} \tag{6.2}$$

$$\bar{\sigma}_x = \sqrt{\frac{\Sigma X^2 - nM_x^2}{n-1}} \tag{6.3}$$

$$\sigma_M = \frac{\bar{\sigma}_x}{\sqrt{n}} \tag{7.1}$$

$$\sigma_{\sigma_M} = \frac{1}{\sqrt{2(n-1)}} \tag{7.2}$$

$$Y = a + bX \tag{8}$$

$$b = \frac{\Sigma(XY) - nM_xM_y}{\Sigma(X^2) - n(M_x)^2} \tag{9}$$

$$a = M_y - bM_x \tag{10}$$

$$\Sigma(XY) - nM_xM_y = \Sigma(xy) \tag{11}$$

$$Y = a + bX + cX^2 \tag{12}$$

(With X used for X, x for $X - M_x$, U for X^2, u for $U - M_u$, equation [12] becomes $Y = a + bX + cU$. These symbols are used in equations [13] to [15], inclusive.)

$$\left.\begin{aligned}(\Sigma x^2)b + (\Sigma xu)c &= \Sigma xy\\ (\Sigma xu)b + (\Sigma u^2)c &= \Sigma uy\end{aligned}\right\} \tag{13}$$

$$a = M_y - b(M_x) - c(M_u) \tag{14}$$

$$\left.\begin{aligned}M_x &= \frac{\Sigma X}{n};\ M_u = \frac{\Sigma U}{n};\ M_y = \frac{\Sigma Y}{n}\\ \Sigma x^2 &= \Sigma X^2 - nM_x^2\\ \Sigma xu &= \Sigma XU - nM_xM_u\\ \Sigma u^2 &= \Sigma U^2 - nM_u^2\\ \Sigma xy &= \Sigma XY - nM_xM_y\\ \Sigma uy &= \Sigma UY - nM_uM_y\end{aligned}\right\} \tag{15}$$

$$Y = a + bX + cX^2 + dX^3 \tag{16}$$

(With U for X^2, V for X^3, equation [16] becomes $Y = a + bX + cU + dV$. These symbols are used in equations [17] to [19], inclusive.)

$$\left.\begin{aligned}(\Sigma x^2)b + (\Sigma xu)c + (\Sigma xv)d &= \Sigma xy\\ (\Sigma xu)b + (\Sigma u^2)c + (\Sigma uv)d &= \Sigma uy\\ (\Sigma xv)b + (\Sigma uv)c + (\Sigma v^2)d &= \Sigma vy\end{aligned}\right\} \tag{17}$$

$$a = M_y - b(M_x) - c(M_u) - d(M_v) \tag{18}$$

$$\left.\begin{aligned} M_v &= \frac{\Sigma V}{n} \\ \Sigma uv &= \Sigma UV - nM_uM_v \\ \Sigma xv &= \Sigma XV - nM_xM_v \\ \Sigma v^2 &= \Sigma V^2 \;\; - nM_v^2 \\ \Sigma vy &= \Sigma VY - nM_vM_y \end{aligned}\right] \tag{19}$$

$$\left.\begin{aligned} S^2_{y.x} &= \sigma^2_z = \frac{\Sigma z^2}{n} \\ S^2_{y.f(x)} &= \sigma^2_{z''} = \frac{\Sigma(z'')^2}{n} \end{aligned}\right] \tag{20.1}$$

$$\bar{S}^2_{y.x} = \frac{n\sigma^2_z}{n-2} = \frac{nS^2_{y.x}}{n-2} \tag{21.1}$$

$$\bar{S}^2_{y.x} = \frac{\Sigma(z^2)}{n-2} = \sigma^2_z\left(\frac{n}{n-2}\right) \tag{21.2}$$

$$\bar{S}^2_{y.f(x)} = \frac{n\sigma^2_{z''}}{n-m} = \frac{nS^2_{y.f(x)}}{n-m} \tag{22.1}$$

$$\bar{S}^2_{y.f(x)} = \frac{\Sigma\,(z''^2)}{n-m} = \sigma^2_{z''}\left(\frac{n}{n-m}\right) \tag{22.2}$$

$$r_{yx} = \frac{\sigma_{y'}}{\sigma_y} \tag{23.1}$$

$$\rho_{yx} = \frac{\sigma_{y''}}{\sigma_y} \tag{23.2}$$

$$d_{xy} = r^2_{xy} \tag{24.1}$$

$$\bcancel{d}_{xy} = 1 - r^2_{xy} \tag{24.2}$$

$$d_{y.f(x)} = \rho^2_{yx} \tag{24.3}$$

$$r_{yx} = \sqrt{b_{yx}b_{xy}} \tag{24.4}$$

$$\bar{r}^2_{yx} = 1 - (1 - r^2_{yx})\left(\frac{n-1}{n-2}\right) \tag{25}$$

$$\bar{\rho}^2_{yx} = 1 - (1 - \rho^2_{yx})\left(\frac{n-1}{n-m}\right) \tag{26}$$

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{\sqrt{[\Sigma(X^2) - nM_x^2]\,[\Sigma(Y^2) - nM_y^2]}} \tag{27}$$

$$b_{yx} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x^2} = \frac{\Sigma(xy)}{n\sigma_x^2} \tag{27.1}$$

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x\sigma_y} = \frac{\Sigma(xy)}{n\sigma_x\sigma_y} \tag{27.2}$$

$$\overline{S}_{y.x} = \sqrt{\frac{\Sigma(Y^2) - n(M_y)^2}{n-1}\,(1 - \bar{r}_{xy}^2)} \tag{28}$$

$$\bar{\rho}_{yx}^2 = 1 - \left(\frac{\sigma_{z''}^2}{\sigma_y^2}\right)\left(\frac{n-1}{n-m}\right) \tag{29}$$

$$X_1 = a + b_2X_2 + b_3X_3 + \ldots b_nX_n \tag{29.1}$$

$$X_1 = a + b_2X_2 + b_3X_3 \tag{30}$$

$$\left.\begin{aligned} \Sigma(x_2^2)b_2 \quad &+ \Sigma(x_2x_3)b_3 = \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_2 &+ \Sigma(x_3^2)b_3 \quad = \Sigma(x_1x_3) \end{aligned}\right\} \tag{31}$$

$$a = M_1 - b_2M_2 - b_3M_3 \tag{32}$$

$$X_1' = a + b_2X_2 + b_3X_3 \tag{33}$$

$$z = X_1 - X_1 \tag{34}$$

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \tag{35}$$

$$X_1 = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{36}$$

$$X_1 = a_{1.2345} + b_{12.345}X_2 + b_{13.245}X_3 + b_{14.235}X_4 + b_{15.234}X_5 \tag{37}$$

$$\left.\begin{aligned} \Sigma(x_2^2)b_{12.34} \quad &+ \Sigma(x_2x_3)b_{13.24} + \Sigma(x_2x_4)b_{14.23} = \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_{12.34} &+ \Sigma(x_3^2)b_{13.24} \quad + \Sigma(x_3x_4)b_{14.23} = \Sigma(x_1x_3) \\ \Sigma(x_2x_4)b_{12.34} &+ \Sigma(x_3x_4)b_{13.24} + \Sigma(x_4^2)b_{14.23} \quad = \Sigma(x_1x_4) \end{aligned}\right\} \tag{38}$$

$$a_{1.234} = M_1 - b_{12.34}M_2 - b_{13.24}M_3 - b_{14.23}M_4 \tag{39}$$

$$\left.\begin{aligned} \Sigma(x_2^2)b_{12.345} \quad &+ \Sigma(x_2x_3)b_{13.245} + \Sigma(x_2x_4)b_{14.235} \\ &\qquad + \Sigma(x_2x_5)b_{15.234} = \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_{12.345} &+ \Sigma(x_3^2)b_{13.245} \quad + \Sigma(x_3x_4)b_{14.235} \\ &\qquad + \Sigma(x_3x_5)b_{15.234} = \Sigma(x_1x_3) \end{aligned}\right\} \tag{40}$$

Etc.

$$a_{1.2345} = M_1 - b_{12.345}M_2 - b_{13.245}M_3 - b_{14.235}M_4 - b_{15.234}M_5 \tag{41}$$

$$\bar{S}^2_{1.234} = \frac{n\sigma^2_{z_{1.234}}}{n - m} \tag{42}$$

$$\bar{S}^2_{1.234\,\ldots\,n} = \frac{\left\{\begin{matrix}\Sigma(x_1^2) - [b_{12.34\,\ldots\,n}(\Sigma x_1x_2) + b_{13.24\,\ldots\,n}(\Sigma x_1x_3) \\ + \ldots + b_{1n.23\,\ldots\,(n-1)}(\Sigma x_1x_n)]\end{matrix}\right\}}{n - m} \tag{43}$$

$$X_{1(234)} = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{44}$$

$$R_{1.234} = \frac{\sigma_{1(234)}}{\sigma_1} \tag{45}$$

$$R^2_{1.234\,\ldots\,n} = \frac{\left\{\begin{matrix}b_{12.34\,\ldots\,n}(\Sigma x_1x_2) + b_{13.24\,\ldots\,n}(\Sigma x_1x_3) + \ldots \\ + b_{1n\cdot23\,\ldots\,(n-1)}(\Sigma x_1x_n)\end{matrix}\right\}}{\Sigma(x_1^2)} \tag{46}$$

$$\bar{R}^2_{1.234\,\ldots\,n} = 1 - (1 - R^2_{1.234\,\ldots\,n})\left(\frac{n - 1}{n - m}\right) \tag{47}$$

$$\bar{R}^2_{1.234\,\ldots\,n} = 1 - \left(\frac{\bar{S}^2_{1.234\,\ldots\,n}}{\sigma_1^2}\right)\left(\frac{n - 1}{n}\right) \tag{48}$$

$$\bar{S}^2_{1.234\,\ldots\,n} = \sigma_1^2(1 - \bar{R}^2_{1.234\,\ldots\,n})\left(\frac{n}{n - 1}\right) \tag{49}$$

$$\bar{r}^2_{14.23} = 1 - \frac{1 - \bar{R}^2_{1.234}}{1 - \bar{R}^2_{1.23}} \tag{50}$$

$${}_{12}\bar{r}^2_{34} = \frac{b^2_{12.34}\sigma_2^2}{b^2_{12.34}\sigma_2^2 + \sigma_1^2(1 - \bar{R}^2_{1.234})} \tag{51}$$

$$\beta_{12.34} = b_{12.34}\,\frac{\sigma_2}{\sigma_1} \tag{52}$$

$$\left.\begin{matrix}\text{Multiple correlation squared} \\ \text{of } (X_1 - b_{12.34}X_2) \text{ with } X_3 \\ \text{and } X_4\end{matrix}\right\} = 1 - \frac{\sigma_1^2(1 - R^2_{1.234})}{\sigma_1^2 - 2b_{12.34}\left(\dfrac{\Sigma x_1x_2}{n}\right) + b^2_{12.34}\sigma_2^2} \tag{53}$$

$$X_1 = a' + f_2(X_2) + f_3(X_3) + f_4(X_4) + \ldots \tag{54}$$

$$\left.\begin{aligned}X_1 = a + b_2X_2 + b_{2'}(X_2^2) + b_3X_3 + b_{3'}(X_3^2) \\ + b_4X_4 + b_{4'}(X_4^2)\end{aligned}\right\} \tag{55}$$

$$\left.\begin{aligned} X_1 = a + b_2(X_2) + b_{2'}(X_2^2) + b_{2''}(X_2^3) + b_3(X_3) + b_{3'}(X_3^2) \\ + b_{3''}(X_3^3) + b_4(X_4) + b_{4'}(X_4^2) + b_{4''}(X_4^3) \end{aligned}\right\} \tag{56}$$

$$X_1'' = a'_{1.234} + f'_2(X_2) + f'_3(X_3) + f'_4(X_4) \tag{57}$$

$$a'_{1.234} = M_1 - \frac{\Sigma[f'_2(X_2) + f'_3(X_3) + f'_4(X_4)]}{n} \tag{58}$$

$$z'' = X_1 - X_1'' \tag{59}$$

$$X_1' = F_2(X_2) = f_2(X_2) - M_{f(2)} + M_1 \tag{60}$$

$$x_1' = F_2(x_2) = f_2(X_2) - M_{f(2)} \tag{61}$$

$$X_1' = F_2(x_2) + F_3(X_3) + F_4(x_4) \ldots + F_n(x_n) \tag{62}$$

$$S_{1.f(2,3,4)} = \sigma_{z_{1.f(2,3,4)}} \tag{63}$$

$$\bar{S}^2_{1.f(2,3,4,\text{ etc.})} = \frac{\sigma^2_{z_{1.f(2,3,4,\text{ etc.})}}}{1 - m/n} \tag{64}$$

$$\bar{S}^2_{1.f(2,3,4,\text{ etc.})} = \frac{n\sigma_z^2}{n - m} = \frac{\Sigma(z^2)}{n - m} \tag{65}$$

$$\mathrm{P}^2 = 1 - \frac{\sigma^2_{z''}}{\sigma_1^2} \tag{66.1}$$

$$\bar{\mathrm{P}}^2_{1.234} = 1 - \left[\left(\frac{\bar{S}^2_{1.f(2,3,4)}}{\sigma_1^2}\right)\left(\frac{n-1}{n}\right)\right] \tag{66.2}$$

$$\bar{\mathrm{P}}^2_{1.234} = 1 - \left[\left(\frac{\sigma^2_{z_{1.f(2,3,4)}}}{\sigma_1^2}\right)\left(\frac{n-1}{n-m}\right)\right] = 1 - \left[\left(\frac{\Sigma(z^2_{1.f(2,3,4)})}{\Sigma(x_1^2)}\right)\left(\frac{n-1}{n-m}\right)\right] \tag{66.3}$$

$$\bar{\mathrm{P}}^2_{1.234} = 1 - (1 - R^2_{1.2'3'4'})\left(\frac{n-1}{n-m}\right) \tag{67}$$

$$\eta_{yx} = \sqrt{\frac{\Sigma[n_0(M_0)^2] - n(M_y)^2}{n\sigma_y^2}} \tag{68}$$

$$\sigma_{b_{yx}} = \frac{\bar{S}_{y.x}}{\sigma_x\sqrt{n}} \tag{69}$$

$$\sigma_{M_{y'}} = \frac{\bar{S}_{y.x}}{\sqrt{n}} \tag{70}$$

$$\sigma_{y'} = \sqrt{\sigma^2_{M_{y'}} + (\sigma_{b_{yx}}x)^2} \tag{70.1}$$

$$\sigma_{r_{yx}} = \frac{1 - r^2}{\sqrt{n - 2}}\text{, for large values of } n \qquad (71)$$

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \qquad (71.1)$$

$$\sigma_{\rho_{xy}} = \frac{1 - \rho^2}{\sqrt{n - m}}\text{, for large values of } n \qquad (72)$$

$$\sigma_{R_{1\cdot 234 \ldots n}} = \frac{1 - R^2_{1.234 \ldots n}}{\sqrt{n - m}}\text{, for large values of } n \qquad (73)$$

$$\sigma_{b_{12\cdot 34 \ldots n}} = \sqrt{\frac{\bar{S}^2_{1.234 \ldots n}}{n\sigma_2^2(1 - R^2_{2.34 \ldots n})}} \qquad (74)$$

$$t = \frac{R_{1.234}\sqrt{n - m}}{\sqrt{1 - R^2_{1.234}}} \qquad (74.05)$$

$$\sigma_{f(X) - f(X_M)} = \sqrt{\frac{\bar{S}^2_{y.f(x)}ux}{\sigma_x^2 n_u}} \qquad (74.1)$$

$$\sigma_{f_{12\cdot 34}(X_2) - f_{12\cdot 34}(X_{M_2})} = \sqrt{\frac{\bar{S}^2_{1.f(2,3,4)}ux_2}{\sigma_2^2 n_u(1 - \bar{P}^2_{2.34})}} \qquad (74.2)$$

$$\sigma_{f(X) - f(X_M)} = \sqrt{k\frac{x}{n_u}}; \quad \text{where} \quad k = \frac{\bar{S}^2_{y.f(x)}u}{\sigma_x^2} \qquad (74.11)$$

$$\sigma_{f_{12\cdot 34}(X_2) - f_{12\cdot 34}(X_{M_2})} = \sqrt{k'\frac{x_2}{n_u}}, \text{ where } k' = \frac{\bar{S}^2_{1.f(2,3,4)}u}{\sigma_2^2(1 - \bar{P}^2_{2.34})} \qquad (74.21)$$

$$\sigma^2_{Y' - Y} = \sigma^2_{M_{y'}} + (\sigma_{b_{yx}}x)^2 + \bar{S}^2_{y.x} \qquad (75)$$

$$Y = Y' \pm t\,\sigma_{y' - y} \qquad (76)$$

$$\sigma^2_{x'_{1\cdot 234} - x_1} = \bar{S}^2_{1.234}\Big[1 + \frac{1}{n} + c_{22}x_2^2 + c_{33}x_3^2 + c_{44}x_4^2 + 2c_{23}x_2x_3 + 2c_{24}x_2x_4 + 2c_{34}x_3x_4\Big] \qquad (77)$$

$$\left.\begin{aligned} (\Sigma x_2^2)c_{22} + (\Sigma x_2x_3)c_{23} + (\Sigma x_2x_4)c_{24} &= 1 \\ (\Sigma x_2x_3)c_{22} + (\Sigma x_3^2)c_{23} + (\Sigma x_3x_4)c_{24} &= 0 \\ (\Sigma x_2x_4)c_{22} + (\Sigma x_3x_4)c_{23} + (\Sigma x_4^2)c_{24} &= 0 \end{aligned}\right\} \qquad (78)$$

$$\left.\begin{aligned}(\Sigma x_2^2)c_{32} + (\Sigma x_2x_3)c_{33} + (\Sigma x_2x_4)c_{34} &= 0\\ (\Sigma x_2x_3)c_{32} + (\Sigma x_3^2)c_{33} + (\Sigma x_3x_4)c_{34} &= 1\\ (\Sigma x_2x_4)c_{32} + (\Sigma x_3x_4)c_{33} + (\Sigma x_4^2)c_{34} &= 0\end{aligned}\right\} \tag{79}$$

$$\left.\begin{aligned}(\Sigma x_2^2)c_{42} + (\Sigma x_2x_3)c_{43} + (\Sigma x_2x_4)c_{44} &= 0\\ (\Sigma x_2x_3)c_{42} + (\Sigma x_3^2)c_{43} + (\Sigma x_3x_4)c_{44} &= 0\\ (\Sigma x_2x_4)c_{42} + (\Sigma x_3x_4)c_{43} + (\Sigma x_4^2)c_{44} &= 1\end{aligned}\right\} \tag{80}$$

$$\left.\begin{array}{l}\sigma^2_{x'_{1.23\ldots n}-x_1} = \bar{S}^2_{1.23\ldots n}\left[1 + \dfrac{1}{n} + (c_2x_2 + c_3x_3 + \ldots + c_nx_n)^2\right]\\ \text{on condition that } (c_2c_2) = c_{22},\quad c_2c_n = c_{2n}, \text{ etc.}\end{array}\right\} \tag{81}$$

$$\sigma^2_{y'_{y.f(x)}-y} = \bar{S}^2_{y.f(x)}\left(1 + \frac{1}{n}\right) + [\text{standard error of } f(X) - f(X_M)]^2 \tag{82}$$

$$\sigma^2_{x'_{1.f(234)}-x} = \bar{S}^2_{1.f(234)}\left(1 + \frac{1}{n}\right) + \sigma^2_{f_2(X_2)} + \sigma^2_{f_3(X_3)} + \sigma^2_{f_4(X_4)} \tag{83}$$

$$X_1 = f(X_2, X_3) \tag{84}$$

$$X_1 = f_{2,3}(X_2, X_3) + f_4(X_4) \tag{85}$$

$$X_1 = f_{2,3}(X_2, X_3) + f_{4,5}(X_4, X_5) + f_6(X_6) \tag{86}$$

$$X_1 = f(X_2, X_3, X_4, \ldots X_n) \tag{87}$$

$$z'''' = a_{z.234} + b_{z2.34}X_2 + b_{z3.24}X_3 + b_{z4.23}X_4 \tag{88}$$

$$\left.\begin{aligned}X_1 = a_{1.2'3'4'} + b_{12'.3'4'}\,[f_2'''(X_2)] &+ b_{13'.2'4'}\,[f_3'''(X_3)]\\ &+ b_{14'.2'3'}\,[f_4'''(X_4)]\end{aligned}\right\} \tag{89}$$

$$X_1'''' = \theta(X_1''') \tag{90}$$

$$X_1'''' = \theta\,[a + f_2'''(X_2) + f_3'''(X_3) + f_4'''(X_4)] \tag{91}$$

$$X_1 = a + eX_3 + g(X_2X_3) \tag{92}$$

$$X_1 = a + eX_3 + g(X_2X_3) + h(X_2) \tag{93}$$

$$X_1 = a + f_2(X_2) + f_{2,3}(X_2X_3) + f_3(X_3) \tag{94}$$

$$X_1 = f_2(X_2) + f_3(X_3) + f_{2+3}(X_2 + X_3) \tag{95}$$

$$X_1 = f_2(X_2) + f_3(X_3) + f_{2+3}\left(\frac{X_2}{a} + \frac{X_3}{b}\right) + f_{2-3}\left(\frac{X_2}{a} - \frac{X_3}{c}\right) \tag{96}$$

$$\left.\begin{aligned}X_1 &= f_2(X_2) + f_3(X_3) + f_4(X_4)\\ &+f_{2+3+4}\left(\frac{X_2}{\sigma_2}+\frac{X_3}{\sigma_3}+\frac{X_4}{\sigma_4}\right)+f_{2+3-4}\left(\frac{X_2}{\sigma_2}+\frac{X_3}{\sigma_3}-\frac{X_4}{\sigma_4}\right)\\ &+f_{2-3+4}\left(\frac{X_2}{\sigma_2}-\frac{X_3}{\sigma_3}+\frac{X_4}{\sigma_4}\right)+f_{2-3-4}\left(\frac{X_2}{\sigma_2}-\frac{X_3}{\sigma_3}-\frac{X_4}{\sigma_4}\right)\end{aligned}\right\} \tag{97}$$

$$R_{1.2,\ 2^2,\ 2^3,\ 3,\ 3^2,\ 3^3,\ \ldots\ n,\ n^2,\ n^3} = \mathrm{P}_{1.23\ \ldots\ n} \tag{98}$$

$$\bar{S}^2_{1.f(23\ \ldots\ n)} = \sigma_1^2(1 - \bar{\mathrm{P}}^2_{1.23\ \ldots\ n})\left(\frac{n}{n-1}\right) \tag{99}$$

$$\Sigma x^2 = \Sigma(d^2F) - n\left[\frac{\Sigma(dF)}{n}\right]^2 \tag{100}$$

$$\left.\begin{aligned}\sigma_{b_{12.34}} &= \bar{S}_{1.234}\sqrt{c_{22}}\\ \sigma_{b_{13.24}} &= \bar{S}_{1.234}\sqrt{c_{33}}\\ \sigma_{b_{14.23}} &= \bar{S}_{1.234}\sqrt{c_{44}}\end{aligned}\right\} \tag{101}$$

$$\bar{k}^2 = \frac{(n-1)(\sigma_z^2/\sigma_y^2)}{n-m} \tag{102}$$

$$\bar{d}_{12.34} = \left[\frac{b_{12.34}(\Sigma x_1 x_2)}{\Sigma(x_1^2)}\right]\left[\frac{\bar{R}^2_{1.234}}{R^2_{1.234}}\right] \tag{103}$$

APPENDIX 5

GLOSSARY

The Greek letters used as symbols in this text, and the most important other symbols, are as follows:

δ	(small *delta*)	= coefficient of average deviation.
σ	(small *sigma*)	= coefficient of standard deviation.
Σ	(capital *sigma*)	= sum of the items specified.
n	(Latin)	= number of observations in a sample.
b	(Latin)	= coefficient of regression.
$f\,(\,)$	(Latin)	= function of the variable in the parenthesis.
r	(Latin)	= coefficient of correlation.
ρ	(small *rho*)	= index of (curvilinear) correlation.
S	(Latin)	= standard error of estimate.
m	(Latin)	= number of constants in the regression equation.
z	(Latin)	= residual, or difference between observed and estimated values of a dependent variable.
R	(Latin)	= coefficient of multiple correlation.
β	(small *beta*)	= " beta " coefficient of regression, in terms of unit standard deviations.
P	(capital *rho*)	= index of multiple (curvilinear) correlation.
η	(small *eta*)	= correlation ratio.
θ	(small *theta*)	= function of (used here for the Bruce adjustment function).
Δ	(capital *delta*)	= arbitrary symbol.
π	(small *pi*)	= arbitrary symbol.
Φ	(capital *phi*)	= function of.
X, Y	(Latin)	= variables, as observed.
x, y	(Latin)	= variables, in terms of departures from their means.
d	(Latin)	= coefficient of determination.
k	(Latin)	= coefficient of alienation.

REFERENCES

Excellent bibliographies covering the basic development of the theory of statistics are given in G. U. Yule and M. G. Kendall's *Introduction to the Theory of Statistics*, and in the special study, *Studies in the History of Statistical Method*, by Helen M. Walker (Williams & Wilkens Co., Baltimore, 1929). In addition, a brief list of references covering especially articles on the theory of sampling is given in R. A. Fisher's *Statistical Methods for Research Workers*. No attempt will be made here to repeat these bibliographies; instead, the student of statistical theory is referred to the sources mentioned.

Articles used as the basis for specific points have already been cited at various places in this book, particularly in Chapters 22 and 23. In addition to those, the methods discussed which go beyond most statistical textbooks are based upon the following technical articles:

Brandt, A. E. Use of machine factoring in multiple correlation, *Jour. Amer. Stat. Assoc.*, XXIII, p. 291. September, 1928.

Ezekiel, Mordecai. A method of handling curvilinear correlation for any number of variables, *Quart. Pub., Amer. Stat. Assoc.*, Vol. XIX, pp. 431–453. December, 1924.

——, The assumptions implied in the multiple regression equation, *Jour. Amer. Stat. Assoc.*, Vol. XX, pp. 405–408. September, 1925.

——, The determination of curvilinear regression "surfaces" in the presence of other variables, *Jour. Amer. Stat. Assoc.*, Vol. XXI, pp. 310–320. September, 1926.

——, The application of the theory of error to multiple and curvilinear correlations, *Proceedings Amer. Stat. Assoc.*, pp. 99–104. March, 1929.

——, A first approximation to the sampling reliability of multiple correlation curves obtained by successive graphic approximations, *Annals of Mathematical Statistics*, Vol. I. September, 1930.

Mendenhall, Robert M., and Richard Warren. The Mendenhall-Warren-Hollerith correlation method. *Columbia Univ. Stat. Bur. Doc.* 1. 1929.

Mills, Frederick C., The measurement of correlation and the problem of estimation, *Quart. Pub., Amer. Stat. Assoc.*, pp. 273–300. September, 1924.

Schultz, Henry. The standard error of a forecast from a curve, *Jour. Amer. Stat. Assoc.*, pp. 139–185. June, 1930.

Smith, Bradford B. Forecasting the acreage of cotton, *Jour. Amer. Stat. Assoc.*, pp. 31–47, especially footnotes on pp. 41 and 42. March, 1925.

——, The use of punched card tabulating equipment in multiple correlation problems. Bur. of Agri. Econ., mimeographed report. 1923.

——, Correlation theory and method applied to agricultural research .... Dept. of Agr., Bur. Agr. Econ., mimeographed report. August, 1926.

Tolley, H. R., and Mordecai Ezekiel. A method of handling multiple correlation problems. *Quart. Pub., Amer. Stat. Assoc.*, pp. 994–1003. Dec., 1923.

—— and ——, The Doolittle method for solving multiple correlation equations versus the Kelley-Salisbury "iteration" method, *Jour. Amer. Stat. Assoc.*, pp. 497–500. December, 1927.

Wallace, H. A., and George W. Snedecor. Correlation and machine calculation, *Iowa State College Bul.* 35. 1925.

Wishart, John. Table of significant values of the multiple correlation coefficient. *Quart. Jour. Royal Meteorological Society*, pp. 258–259. July, 1928.

INDEX

Abscissa, defined, 10
Adjustment for number of observations, *See* Number of observations, adjustment for
Alienation coefficient, 494
Allen, R. H., refs. on supply analysis, 441
Allport, Gordon W., ref., 440
Apple prices, as illustration of joint functions, 393
Arithmetic average, standard error of, 19-22
See also Average, arithmetic
Asparagus prices, illustration, 425
Assumptions, in sampling, 15
on free-hand curves, 109, 152, 224, 278
Auto-stopping, as illustrative problem, 42
Average, arithmetic, defined. 2
standard error of, 19
Average deviation, defined, 3
Averages, in determining functional relation, 47

Bartels Technique, footnote. 355
Bean, Louis H., acknowledgment to, vii
ref. on cotton prices, 438
on graphic correlation, 439
on graphic *vs.* mathematical methods, 110
on potato and cotton prices. 439
on production response. 439
on short-cut method, 268, 300
on voting, 441
Been, Richard O., cited, 349
ref. on standard error, 358
Benner, Claude L., ref. on egg prices, 439
Bercaw, Louise O., ref. on price analyses, 440
Beta coefficient, defined, 159
in multiple correlation, 217
Bias, in sampling, 28, 370
Birge, Raymond T., ref., 24, 488
Black, John D., influence of, x
ref. on creamery costs, 439
on graphic *vs.* mathematical methods, 110
on input-output, 437
on short-cut method, 296, 300
Bowley, Arthur L., ref., 444
Brandt, A. E., ref. on computation method, 522
Bruce, Donald, ref., 414
Bruce Adjustment, 403
Bureau of Standards, 34
Burmeister, Gustave, ref. on potato yields, 437

Cassels, J. M., ref. on supply analysis, 441
Chamberlain, Edward, 441
Chatfield, Charlotte, ref. on beef composition, 438
Chauncey, Marlin R., ref., 440
Check sum, 461
Cherniack, Nathan, cited, 424
ref. on watermelon prices, 438
Children's clothes, size standards for, 434
Class interval, defined, 5
Coding, in fitting logarithms, 95
in fitting parabolas, 84
mathematical effect of, 490
Coefficient of alienation, defined, 139
Coefficient of correlation, *See* Correlation coefficient
Coefficient of determination, defined, 139
Coefficient of multiple correlation, 210

Coefficient of non-determination, defined, 140
Coefficient of partial correlation, *See* Partial correlation coefficient
Coefficient of regression, *See* Regression coefficient
Coleman, D. A., ref., 438
Computation methods, 455
  for standard deviation, 6, 12
  steps in, 449
Contours, for joint regressions, 404
Coordinates, defined, 10
Corn yield, example of forecasting, 255
  illustrating sampling, 17
  multiple curvilinear regression, 225
Correlation coefficient, adjustment for size of sample, 510
  compared to other measures, 160
  computation methods, 455
  defined, 137
  practical equation for, 149
  standard error of, 318
Correlation coefficients, reliability of, 504
Correlation equations, proof of formulas, 491
Correlation index, defined, 138
  practical equation for, 156
  standard error of, 320
Correlation measures, for short-cut linear method, 276
  mathematical meaning of, 493
  meaning of, 159
Correlation methods, place in research, 442
Correlation ratio, defined, 306
Correlation results, meaning of, 450
Correlation table, 456
Correlations, 136
Cotton yield, illustrating standard error of regression curve, 335
  study of fertilizing, 417
  used in illustrative problem, 147
Court, Andrew T., illustration from, 374
  ref. on joint functions, 414
Court Method, for joint functions, 409
Cowan, Donald R. G., ref., 440.
Cowden, Dudley J., refs., 81, 93, 189, 445
Cross classification, for many variables, 186
  for three variables, 181
Croxton, Frederick E., refs., 81, 93, 189, 445
Crum, W. L., ref., 444
    on joint costs, 440
Cubic parabola, method of fitting, 89
Curves, equations for, 76, 79
  logarithmic, *See* Logarithmic curves
Curvilinear correlation, interpretation of results, 157
  practical methods for, 152
  standard error of forecast, 345
Curvilinear functions, illustrations of, 75
Curvilinear multiple correlation, measure of, 264
  standard error of estimate for, 259
Curvilinear relation, fitted by freehand curve, 105, 152, 222, 277
    by mathematical curves, 83, 89, 93, 121

Dairy cows, studies of feeding, 417
Davis, Floyd E., ref., 413, 414
Dean, Joel, ref. on cost curves, 441
Degrees of freedom, reduction in, 142
Deming, W. Edwards, acknowledgment to, vii
  ref., 24
    on theory of errors, 488
Dependent and independent variables, functions for, 58
  straight lines for, 74
Dependent variable, defined, 50
  mathematical representation of, 59
  result of selecting values of, 361
Derksen, J. B. D., ref., 441
Determination coefficient, 494
Determinantal solution, for normal equations, 478
Deviation, average, *See* Average deviation
  standard, *See* Standard deviation
Diedjens, V. A., ref., 439

Differential regressions, 413
Digiting, 459
DIXON, H. B., ref., 438
DOOLITTLE, M. H., 477
Doolittle Method, 200, 464
Dot chart, method of constructing, 36
DWYER, P. S., ref. on computation methods, 478

Education, use of correlation in, 429, 434
Egg price problem, illustrating non-quantitative variable, 302
ELLIOTT, FOSTER F., ref. on correlation method, 500
  on hog production, 439
Equation, for straight line, 59
  linear, meaning of constants in, 60
  parabolic, meaning of constants in, 118
  selecting the type, 447
Equations, for different types of curves, 76, 79
  limitations to, 102
  list of, 512
Error, probable, *See* Probable error
Error of estimate, *See* Standard error of estimate
Errors of observation, in both variables, 366
  in dependent variable, 365
  in independent variable, 366
Estimate, *See* Forecast
Estimating dependent variable, limitations on, 125
Extrapolation of regression equations, 347
EZEKIEL, MORDECAI, data from, 278
  refs. on correlation methods, 110, 522
    on dairy farming, 437
    on error of regression coefficients, 327
    on farmers' earnings, 437
    on hog prices, 438
    on input-output, 437
    on lamb prices, 438
    on milk production, 441
    on price analysis, 439
EZEKIEL, MORDECAI, refs. on short-cut method, 300
    on steel costs, 441
    on tobacco farming, 438

Factors, methods of measuring, 444
Falling body, illustrating a mathematical function, 114
Farm income, illustrating multiple correlation, 169
  related to farm organization, 421
Farm values, example of multiple relations, 164
FELLOWS, H. C., ref., 438
FISHER, R. A., cited, 10
    on standard error of *r*, 319
  ref., 93
    on analysis of variance, 189
    on error of multiple correlation, 323
    on sampling, 522
    on standard errors of *b*, 470
    on use of differential regression, 419
    on wheat yield, 438
  statistical contributions of, x
FOOTE, RICHARD J., ref. on short-cut method, 300
Forecast, change in universe, 31
  from multiple correlation, standard error of, 344
  from simple correlation, standard error of, 342
  practical procedures for, 356
Forecasting corn yields, illustrating method, 255
Freehand curve, as method of expressing relation, 105, 152, 222, 277
  assumptions for, 109, 152, 224, 278
FREEMAN, FRANK S., ref. on intelligence tests, 440
Frequency distribution, chart of, 11
Frequency table, defined, 3
FRISCH, RAGNAR, ref., 448
Function, defined, 39
  linear, assumptions of, 73
Functional relation, graphic illustration of, 38
  mathematical expression of, 39

Functional relation, statistical measurement of, 42
  use of averages to determine, 47
Funk, W. C., ref., 437

Gabriel, Harry G., ref. on egg prices, 439
Gallup poll, 435
Gans, A. R., ref. on milk elasticity, 439
Garrison, K. C., ref. on intelligence tests, 440
Girschick, Meyer A., acknowledgment to, vii
  cited, 301, 477
  ref. on clothing sizes, 441
Glossary, 521
Goodenough, F. L., ref. on intelligence tests, 440
Goodrich, C. L., ref., 437
Gosnell, Harold, ref. on politics, 441
Gowen, John W., refs. on dairying, 439
Graphic chart, for adjusting observed correlations, 510
  for small samples, 505
Graphic charts, for reliability of observed correlation, 506-9
Graphic correlation, standard error of, 327
Graphic curve, fitting freehand, 105
  limited by logical conditions, 109
Graphic curves, *See* Regression curves and Multiple regression curves
Graphic interpolation, 87
Graphic method, 222, 268
  auxiliary graphic processes, 479
Graphic regression curves, adjustments by least-squares, 401
Gravity, illustration of mathematical function, 114
Greek letters, meaning of, 21
Guthrie, Edward S., ref. on creamery costs, 439.

Haas, G. C., cited, 415
  ref. on hog prices, 438
    on land prices, 437
Hafstad, L. R., ref., 355
Hainsworth, R. G., acknowledgments to, vii, xi
Hanau, Arthur, cited, 423
  ref. on hog prices, 438
Hardenburg, E. V., ref. on potato yields, 437
Harper, F. A., ref., 186
Hay-stack problem, illustrating joint correlation, 376
Hedden, W. P., cited, 424
  ref. on watermelon prices, 438
Higbie, Edgar Creighton, ref., 440
Hog prices, illustrating a mathematically fitted equation, 121
Holdaway, C. W., ref. on dairy farming, 437
Hole, Erling, ref. on dairying, 441
Horse feed, illustrative problem, 129
Hosterman, W. H., data from, 378
Hotelling, Harold, ref., 316
Howe, Charles B., data from, 302
  ref. on egg prices, 439
Hull, Clark L., ref. on prediction formulas, 440
Huxley, Julian, quotation from, 1
Hyperbolas, characteristics of, 80
Hypothesis, checking against relations, 452
  method of developing, 443

Independent variable, defined, 50
  mathematical representation of, 59
  result of selecting values, 360
Index of correlation, *See* Correlation index
Index of determination, defined, 140
Index of multiple correlation, defined, 264
  for joint correlation, 387
  from short-cut method, 293
Individual forecast, standard error of, 341, 469
Industrial commodities, cost curves for, 432
  prices of, 434
Inference, statistical, *See* Statistical inference
Input-output relations, studies of, 416
Intelligence, components of, 434
Interpolation, graphic, 87

Interpretation of correlation results, 71, 151, 157, 254, 450
Interquartile range, 4
Irrigation water, illustrative problem, 147
IVES, J. RUSSELL, ref. on short-cut method, 300

JENSEN, EINAR, ref. on milk production, 441
JEROME, HARRY, ref., 444
JOHNSON, SHERMAN E., ref. on dairy farming, 437
Joint correlation, for n variables, 391
Joint function, for three variables, 376
 mathematical functions, 407
Joint regression, 372
Joint relations, identifying by short-cut method, 296

KANTOR, HARRY, ref. on peach prices, 438
KELLEY, TRUMAN L., cited, 159
 refs., 282, 494
KENDALL, M. G., refs., 189, 522
KIFER, R. S., ref. on dairy farming, 437
KILLOUGH, HUGH B., ref., 103
 on oat prices, 438
KOON, R. M., ref., 439
KOOPMANS, T., ref., 441
KUHRT, W. J., ref. on wheat prices, 439
KYLE, CORRINE F., ref. on wheat protein, 438

Land values, 415
Least squares, derivation of normal equations, 489
 in fitting straight line, 64
 note on, 67
Lettuce price, as illustration of hypothesis, 443
Line of best fit, defined, 67
Linear correlation, interpretation of results, 150
 practical methods for, 147, 455
Linear equation, interpretation of, 71, 151
 meaning of constants in, 60
Linear net regressions, by short-cut method, 269
Linear regression equation, 145
 for three variables, 191, 198
Literary Digest poll, 435
Logarithmic curve, method of fitting, 93
Logarithmic curves, characteristics of, 79
Logarithmic equation, fitted to hog prices, 121
 for minimizing absolute departures, 99
Logical conditions, on free-hand curve, 109
 on multiple regression curves, 224, 278
 on regression curves, example of, 152
Logical significance, of mathematical functions, 113

McNALL, P. E., ref., 327
 on dairy farming, 437
MALENBAUM, WILFRED, ref. graphic *vs.* mathematical methods, 110
 on short-cut method, 296, 300
 on supply analysis, 441
Market prices, studies of, 422
Mathematical curves, limitations to, 102
Mathematical equations, for joint functions, 407
 in an economic problem, 121
 when to fit, 120
Mathematical functions, for net regression curves, 396
 logical significance, 113
Mathematical net regression curves, method of fitting, 221, 396
 standard error of, 339
Mean, arithmetic, defined, 2
 standard error of, 19-22
Median, defined, 4
MENDENHALL, ROBERT M., ref. on computation methods, 522
MENSENKAMP, L. E., ref., 439
MIGHELL, R. L., refs. on supply analysis, 441
MILLS, FREDERICK C., ref., 92
 on correlation method, 522
MINOR, W. A., JR., ref., 440
MISES, RICHARD VON, ref., 350

MISNER, E. G., data from, 225
ref. on dairy rations, 437
on weather and prices, 437
MOORE, HENRY L., cited, 422
ref. on cotton forecasting, 438
on cycles, 437, 438
MORRISON, F. B., ref., 327
on dairying, 437
MUIR, JAMES C., data from, 147
Multiple correlation, affected by errors of observation, 367
by successive elimination method, 169
computation methods, 459
theoretical illustration of, 164
Multiple correlation coefficient, 210
reliability of, 504
standard error of, 321
Multiple curvilinear correlation, index of, 264
standard error of, 327
Multiple curvilinear regressions, by mathematical equations, 221, 396
by short-cut method, 277
by successive approximations, 222
Multiple determination, coefficient of, 211
Multiple regression curves, graphic, by successive approximations, 222
limitations on use, 254
mathematical, 221, 396
stating the conclusions, 247
Multiple regression equation, defined, 165, 196
interpretation of, 205
Multiple relations, examples of, 163

Net regression coefficient, defined, 165, 196-7
standard error of, 321
Net regression coefficients, computation method, 459
meaning of, 202, 205
standard error for, 321, 469
Net regression curves, by mathematical functions, 221, 396
standard error of, 327, 339
Non-determination coefficient, 494
Non-quantitative variable, 302
Normal distribution, description of, 9
Normal equations, alternative methods of solving, 477
derivation of, 496
method of solving, 464
Number of observations, adjustments for, 133, 141, 209, 211

Objective, method of stating, 442
O'BRIEN, RUTH, ref. on clothing sizes, 441
Observation equations, defined, 61
Observed range, extrapolation beyond, 347
Ordinate, defined, 10
Orthogonal polynomials, in fitting trends, 93

PALLESEN, J. E., refs., 413, 414
Parabola, characteristics of, 78, 80
cubic, *See* Cubic parabola
equation for, 76
method of fitting, 83
Parabolas for time series, method of fitting, 92
Parabolic equation, meaning of constants, 118
Part correlation, coefficient of, 213 (footnote)
derivation of, 497
Partial correlation, coefficient of, 213
method of computing, 474
Partial correlation indexes, 267
Partial regression coefficient, 197
PATTON, ALSON C., ref., 444
PATTON, PALMER, ref. on crop yields, 437
PEARSON, F. A., ref. on prices, 439
Perpendicular regressions, 367
PETTIT, EDISON, ref., 438
Physical characteristics, studies of, 420
Political behavior, studies of, 435
POND, GEORGE, ref. on dairy farming, 437
Potato yield problem, illustrating joint correlation, 407
Practical correlation methods, 146, 442, 455
Practical procedures, for correlation computations, 146, 455
for reliability of forecasts, 356

Price analysis, examples, 422
  simplified illustration, 124
Prices of farm products, 422
  of industrial commodities, 434
  influence on production, 428
Probable error, 26
Product moment, defined, 149
Production, response to price, 428, 430
Projectile, trajectory of, 118
Protein content, wheat illustration, 82
Psychology, use of correlation in, 429, 434
PURVES, CLARENCE M., acknowledgment to, vii

Quadratic equation, illustration of, 114
Quartiles, lower and upper, defined, 4

RAEBURN, JOHN R., ref., 393
Ratio, correlation, *See* Correlation ratio
Regression coefficient, compared to other measures, 161
  computation methods, 455
  defined, 146
  gross, 165
  net, 165, 196
  practical equation for, 148, 149
  standard error of, 312
Regression curve, defined, 146
Regression curves, limitations on, *See* Logical limitations on regression curves
Regression equation, defined, 146
  extrapolation of, 347
  for *n* variables, 203
  See also Linear and multiple regression equation
Regression line, defined, 146
  standard error of, 315
Reliability, *See* Standard error
RICHEY, FREDERICK D., cited, 429
  ref. on corn breeding, 439
ROOS, C. F., ref. on automobile demand, 441
Roper poll, 435
ROSS, H. A., ref. on milk marketing, 439

Sample, defined, 14
  reliability of, 18
  selection of, 359
  size required, 29
  spot and stratified, 16
Samples, small, *See* Small samples
Sampling, assumptions of, 15
  bias in, 28
  of time series, 349
Sampling theories, assumptions in, 486
SARLE, CHARLES F., acknowledgment to, vii
SASULY, MAX, ref., 92
SCHOENFELD, WILLIAM A., ref. on milk marketing, 439
SCHULTZ, HENRY, ref., 346
    on prices, 440
    on standard errors, 340, 358, 522
SECRIST, HORACE, ref., 444
Separate determination, 498
SHEPHERD, GEOFFREY, ref. on short-cut method, 300
Sheppard's Correction, 12, 458
SHOLLENBERGER, J. H., ref. on wheat kernels, 438
Short-cut method of multiple correlation, 268
  auxiliary graphic processes for, 479
  refs. on, 300
Short-cut method for multiple curvilinear regressions, 277
Small samples, correction for, 23
  reliability of, 22, 504
SMITH, BRADFORD B., cited, 423, 497
  refs. on correlation methods, 522
    on cotton acreage, 439
    on cotton and interest rates, 440
    on cotton prices, 438
    on cotton yield, 437
    on production, 438
SMITH, G. E. P., data from, 147
SNEDECOR, GEORGE W., ref., 189
    on computation methods, 522
SPEARMAN, C., cited, 435
  refs. on factor theory, 441
    on "footrule correlation," 440
SPILLMAN, WILLIAM J., cited, 152
Spot sample, 16

Standard deviation, defined, 8
method of computing, 8
short-cut method of computing, 12
Standard error, 25
of arithmetic average, 19-22
of correlation coefficient, 318
of correlation index, 320
of group averages, 49, 54
of individual estimate, method of computing, 469
of mathematical regression curves, 339
of mean, derivation of, 486
of multiple correlation coefficient, 321
of net regression coefficient, 321
of partial regression coefficients, computation method, 469
of regression coefficient, 312
of regression line, 315
of standard error, 24-25
for time series, 349
Standard error of estimate, 128
compared to other measures, 160
for curvilinear multiple correlation, 259
defined, 131
for joint correlation, 387
for linear multiple correlation, 208
practical equation for, 149
for short-cut method, 293
for simple curvilinear correlation, 131
for simple linear correlation, 129
units for, 135
Standard error of forecast, 341
from curvilinear correlation, 345
from multiple correlation, 344
from simple correlation, 342
Statistical inference, assumptions of, 15, 486
Steel cost problem, as illustration for short-cut method, 278
as illustration of standard error of net regression curves, 338
Stevens, Chester D., ref., 437
Stine, O. C., acknowledgment to, vii
Straight line, equation for, 59
fitted by least squares, 64
graph of, 60, 61
interpretation of equation for, 71
Straight line, methods of fitting, 59
Stratified sample, defined, 16
"Student," ref. on correlation, 504
statistical contributions of, x
Sturges, Alexander, ref., 350
Successive approximation method, for net regression curves, 222
Supplementary methods, for net regression curves, 401
Sutherland, H. E. G., ref. on I. Q., 440

t-test, for correlation coefficient, 318
for multiple correlation coefficient, 323
Table, frequency, *See* Frequency table
Taylor, Henry C., influence of, x
Taylor, C. C., ref. on farm management, 438
Thomsen, F. L., ref. on prices, 440
Thurstone, L. L., ref., 441
Time series, error formula for, 349
method of fitting parabolas to, 92
Time-series analysis, examples of changing, 426
Tolley, Howard R., acknowledgments to, vii, x
cited, 500
refs. on correlation methods, 522
on input-output, 437
Traynor, Kenneth, ref., 437
Trends, fitting to time series, 92
Tretsven, J. O., ref. on dairy farming, 437

Universe, changing, 426
defined, 14
Universes, past and present, 30

Variable, dependent, *See* Dependent variable
independent, *See* Independent variable
Variables, relations between, 34
units of statement, 446
Variance, analysis of, 189
defined, 10
related to correlation coefficient, 493
Vernon, J. J., ref. on dairy farming, 437
on tobacco farming, 438

VON SZELISKI, VICTOR, ref. on automobile demand, 441

WAITE, WARREN C., refs. on short-cut method, 300
WALD, ABRAHAM, ref., 367
WALKER, HELEN M., ref., 522
WALLACE, HENRY A., ref. on computation methods, 522
WARREN, GEORGE F., ref. on prices, 439
WARREN, RICHARD, ref. on computation methods, 522
WAUGH, FREDERICK V., acknowledgments, vii, xi
  cited, 347
  data from, 393, 407
  ref. on computation methods, 478
    on potato prices, 438
    on potato yields, 437
    on standard error, 358
    on vegetable prices, 439
Waugh Method, for joint regressions, 404
Weather conditions, influence on crop yields, 418
WEEKS, ANGELINA L., ref., 440
WELLMAN, H. R., ref. on short-cut method, 300
WELLS, O. V., ref. on dairy farming, 437
WESTBROOK, E. C., ref., 437
Wheat, vitreous kernels, as illustrative problem, 82
Wheat protein, illustrating free-hand curve fitting, 106
  ref. on, 438
Wheat yields, study of, 419
WHITCOMB, W. D., ref., 439
WINCH, W. H., ref., 440
WISHART, JOHN, ref. on standard errors, 522
WITMER, HELEN LELAND, ref. on sex education, 440
WOLFE, T. K., ref. on corn breeding, 439
WORKING, ELMER J., acknowledgment to, vii
  ref. on short-cut method, 300
WORKING, HOLBROOK, ref., 316
    on potato differentials, 439
    on potato prices, 438
WRIGHT, SEWALL, ref. on correlation method, 500
WYLIE, KATHRYN H., data from, 278
  ref., 327
    on steel costs, 441

Yield, corn, *See* Corn yield
YULE, G. UDNY, cited, 16
  ref., 189, 440, 522
    on error theory, 489