METHODS OF CORRELATION ANALYSIS

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
Third Printing, June, 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 edi-

tion; 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

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

CHAPTER 10

CHAPTER 11

CHAPTER 12

CHAPTER 13

CHAPTER 14

CHAPTER 15

CHAPTER 16

CHAPTER 17

CHAPTER 18

CHAPTER 19

CHAPTER 20

CHAPTER 21

CHAPTER 22

CHAPTER 23

CHAPTER 24

APPENDIX 1

APPENDIX 2

APPENDIX 3

APPENDIX 4

APPENDIX 5

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.

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:

$$M_x = \frac{\Sigma X}{n} \qquad (1)$$

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.

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

TABLE 1

YIELDS OF CORN OBTAINED BY TWENTY FARMERS*

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."

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.

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

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

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

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–

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

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

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 far 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.

[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:

$$\begin{aligned} X - M_x &= x \\ X' - M_x &= x' \\ X'' - M_x &= x'' \end{aligned} \tag{2}$$

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:

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

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:

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

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]

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

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

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:

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

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.

Thus $\quad \text{each } x = X - M_x$

$\quad \text{each } x^2 = X^2 - 2XM_x + M_x^2$

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

But $\quad \Sigma X = nM_x$

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

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

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

[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.

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.

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

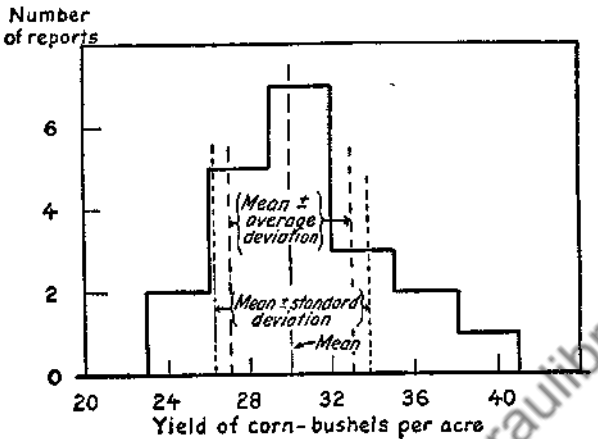

FIG. 1. Frequency distribution of corn yields, and range above and below the mean included by average and standard deviations.

the number of reports included within a range of both one average deviation and of one standard deviation on each side of the mean.

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.

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

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

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) calculated 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.

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.[a] 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

[a] 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

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 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.

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

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

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 $\pm .6745$ 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 "$\pm$."

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:

$$\sigma_M = \frac{\sigma_x}{\sqrt{n}}$$

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.

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. 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.

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.

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.

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 in-

creased 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		

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

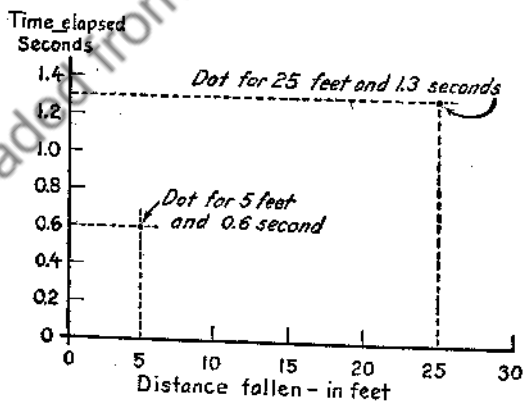

FIG. 2. Method of constructing a dot chart. Time elapsed is the dependent variable, and the distance is the independent variable.

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

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

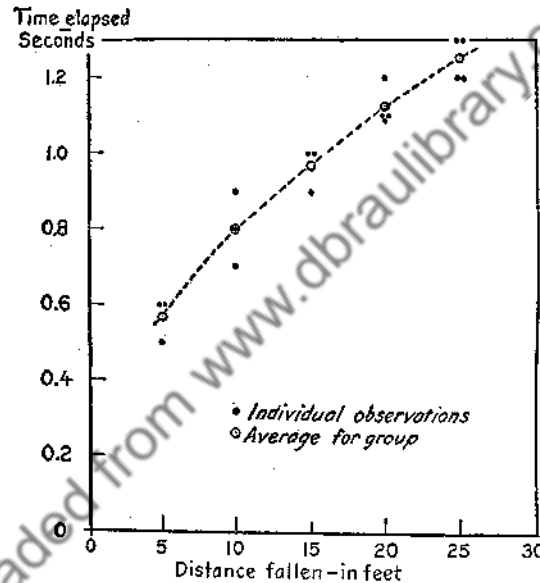

FIG. 3. Relation of distance a marble falls to time elapsed in falling, as shown by individual observations and curve of average time.

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.

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

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

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.

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

[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.

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.

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.

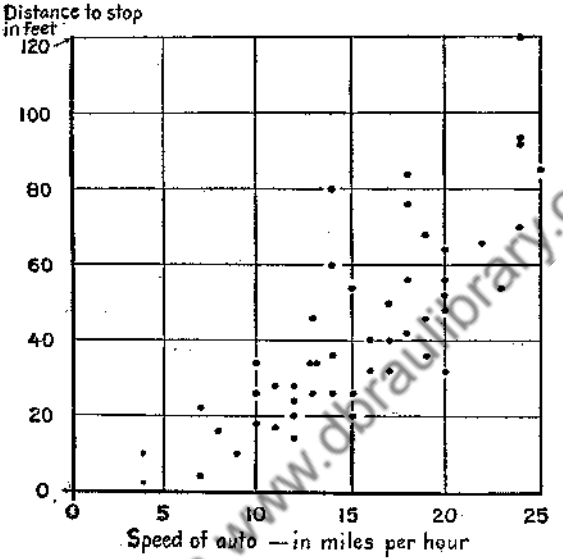

FIG. 4. Relation of speed of automobile to distance it takes to stop, as shown by individual observations.

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.

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

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.

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

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.

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.

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).

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.

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.

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-

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

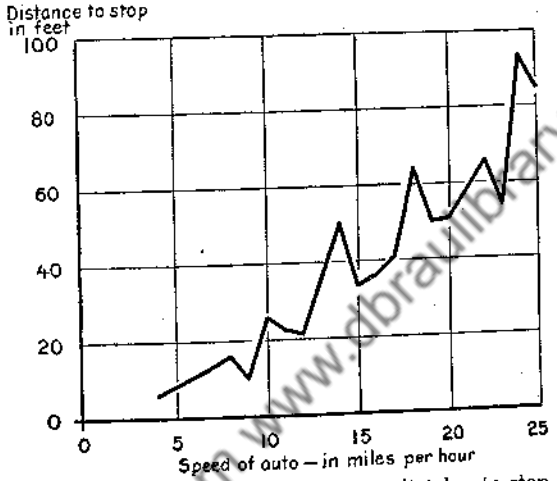

FIG. 5. Relation of speed of automobile to distance it takes to stop, as shown by averages of small groups.

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.

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.

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.*

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?

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.

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.

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.

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

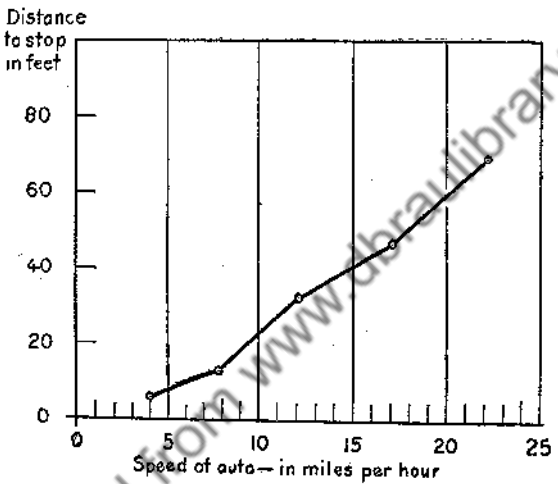

FIG. 6. Relation of speed of automobile to distance it takes to stop, as shown by averages of large groups.

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.

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.

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.

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

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	6.0	3.8	7.0	1.8
5 to 10	7.8	13.0	4.4	19.4	4.4
10 to 15	12.2	32.4	4.9	14.4	2.9
15 to 20	17.1	46.8	5.1	22.5	4.4
20 to 25	22.2	69.3			

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.

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.

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?

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-

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	.....	.....	$4.0 \pm ?$
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

* 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-

[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.

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,

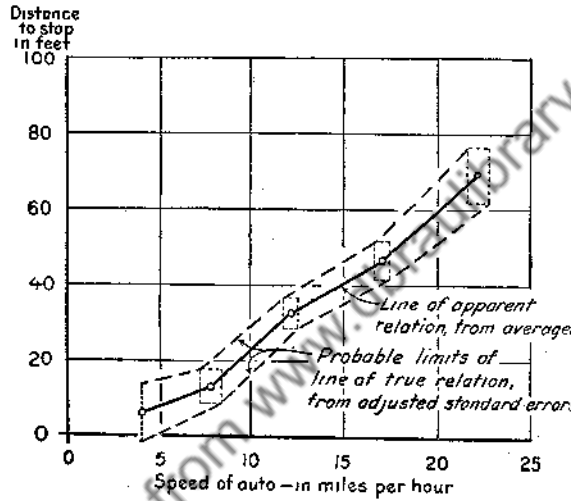

FIG. 7. Relation of speed of automobile to distance it takes to stop, as indicated by the range around group averages for which the probability is $\frac{2}{3}$ that the true average is included.

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*.

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.

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.

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.

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.

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.

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

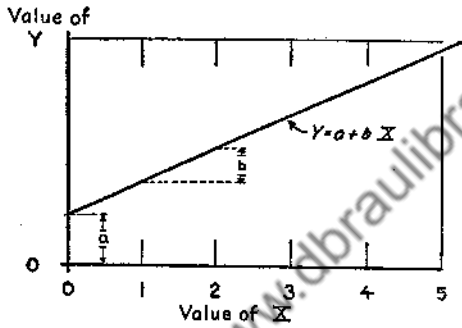

FIG. 8. Graph of the function $Y = a + bX$.

unit change in X, whether from 1 to 2, or from 0 to 1, or from 99 to 100.

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.

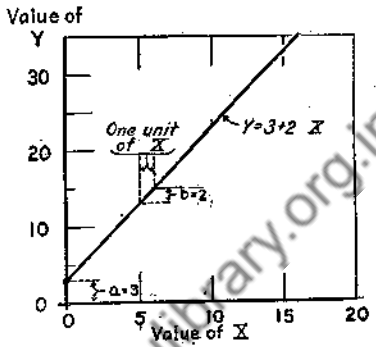

FIG. 9. Graph of the function $Y = 3 + 2X$.

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

$$2 = a + b \text{ (4)}$$

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:

$$\begin{aligned} a + 4b &= 2 \\ a + 20b &= 32 \end{aligned}$$

Solving these two equations simultaneously, we find the values

$$\begin{aligned} a &= -5\tfrac{1}{2} \\ b &= 1\tfrac{7}{8} \end{aligned}$$

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

$$a = -72$$

$$b = 10\tfrac{8}{7}$$

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

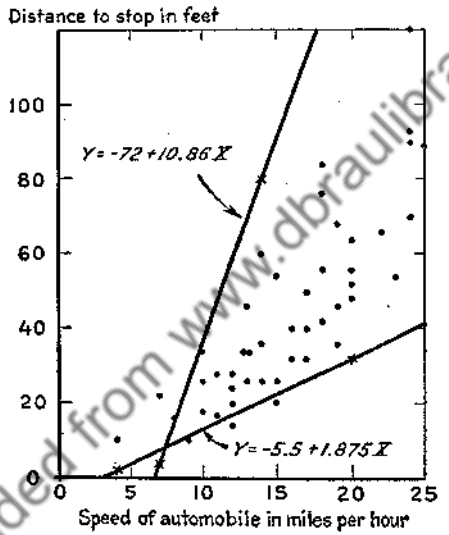

FIG. 9A. Data for automobile problem, and straight lines fitted to pairs of individual observations.

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.

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, $770 = \Sigma X$	$2{,}149 = \Sigma Y$	$13{,}228 = \Sigma(X^2)$	$38{,}482 = \Sigma(XY)$

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:

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

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

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]

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

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

[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

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

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.

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.

[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."

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.

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

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:

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

For 50 miles per hour it gives:

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

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.

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

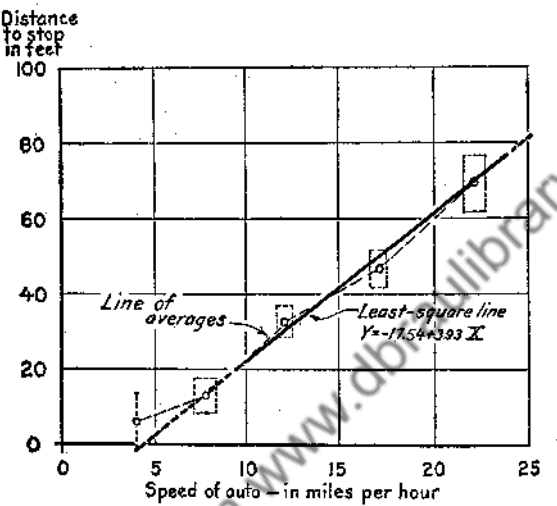

FIG. 10. Relation of speed of automobile to distance-to-stop as indicated by ranges around group averages and by least-squares straight line.

straight line may serve as a more reliable indication of the general relation than does the irregular line of the group averages.

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.

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.

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

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

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.

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.

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

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.

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.

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:

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

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

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

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

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

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

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

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

$$Y = 1 + 0.5\,(1) - 0.1\,(1^2)$$
$$= 1.4$$

When X is 2, Y will be

$$Y = 1 + 0.5\,(2) - 0.1\,(2^2)$$
$$= 1 + 1 - 0.4$$
$$= 1.6$$

Similarly, when X is 3

$$Y = 1 + 0.5\,(3) - 0.1\,(3^2)$$
$$= 1 + 1.5 - 0.9$$
$$= 1.6$$

For X equal to 4

$$Y = 1 + 0.5\,(4) - 0.1\,(4^2)$$
$$= 1.4$$

and for $X = 5$

$$Y = 1 + 0.5\,(5) - 0.1\,(5^2)$$
$$= 1$$

and for $X = 6$

$$Y = 1 + 0.5\,(6) - 0.1\,(6^2)$$
$$= 0.4$$

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-

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

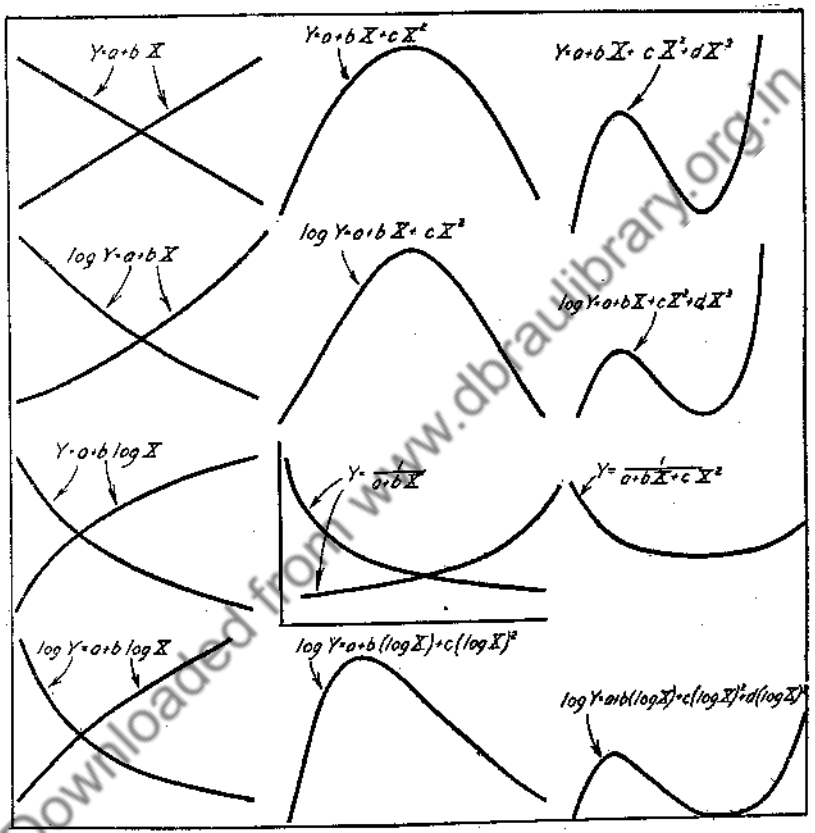

FIG. 11. Curves illustrating a number of different types of mathematical functions.

$Y = a + bX + cX^2$ 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

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.

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; the simple parabola *always* has one such change, if carried out far enough; and the cubic parabola always has two such changes. 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. No matter how small a logarithm becomes, the corresponding antilogarithm is still positive, even if only a very small decimal fraction.

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 to 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

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

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.

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]

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.

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

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

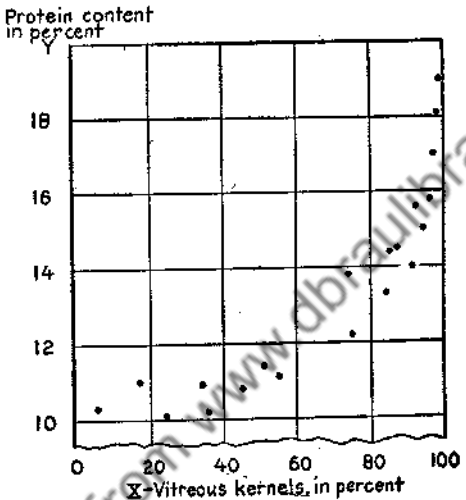

FIG. 12. Dot chart showing relation of proportion of vitreous kernels to protein content of wheat.

of vitreous kernels, along the horizontal axis and the Y values, the proportion of protein, along the vertical axis.

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

$$Y = a + bX + cX^2 \tag{12}$$

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]

$$\left.\begin{aligned}(\Sigma x^2)b + (\Sigma xu)c &= \Sigma xy\\ (\Sigma xu)b + (\Sigma u^2)c &= \Sigma uy\end{aligned}\right\} \tag{13}$$

and

$$a = M_y - b(M_x) - c(M_u) \tag{14}$$

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]

Then

$$\left.\begin{aligned}M_x &= \frac{\Sigma X}{n};\ M_u = \frac{\Sigma U}{n};\ M_y = \frac{\Sigma Y}{n}\\ \Sigma x^2 &= \Sigma X^2 - nM_x^2\\ \Sigma xu &= \Sigma XU - nM_xM_u\\ \Sigma u^2 &= \Sigma U^2 - nM_u^2\\ \Sigma xy &= \Sigma XY - nM_xM_y\\ \Sigma uy &= \Sigma UY - nM_uM_y\end{aligned}\right\} \tag{15}$$

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.

$$na + (\Sigma X)b + (\Sigma U)c = \Sigma Y$$

$$(\Sigma X)a + (\Sigma X^2)b + (\Sigma UX)c = \Sigma XY$$

$$(\Sigma U)a + (\Sigma UX)b + (\Sigma U^2)c = \Sigma YU$$

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:

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

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

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

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 $X = 6$

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

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.

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

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

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.

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,

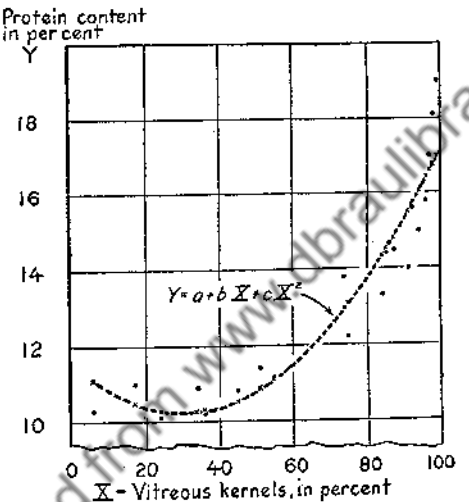

FIG. 13. Dot chart showing relation of vitreous kernels to protein content of wheat, and parabolic curve fitted to same.

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.

"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,

$$Y = a + bX + cX^2 + dX^3 \qquad (16)$$

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.

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

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

$$a = M_y - b(M_x) - c(M_u) - d(M_v) \tag{18}$$

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:

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

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:

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

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.

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,

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

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.

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

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 mathe-

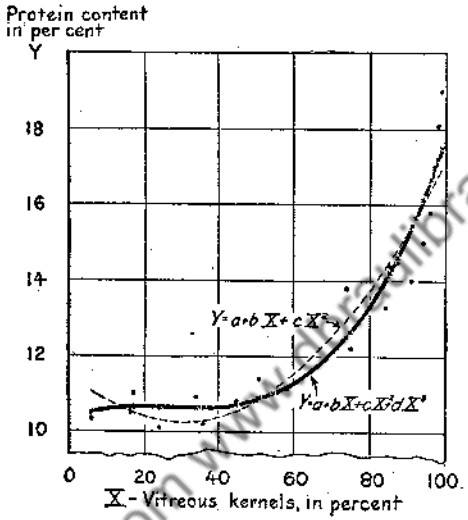

FIG. 14. Dot chart, with parabola and cubic parabola.

matical measures of this question of the sampling significance of a fitted curve.)

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.

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 Y	Per cent vitreous kernels X	Logarithms of Variables:* Protein $\overline{Y}$	Vitreous kernels $\overline{X}$
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.

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 abso-

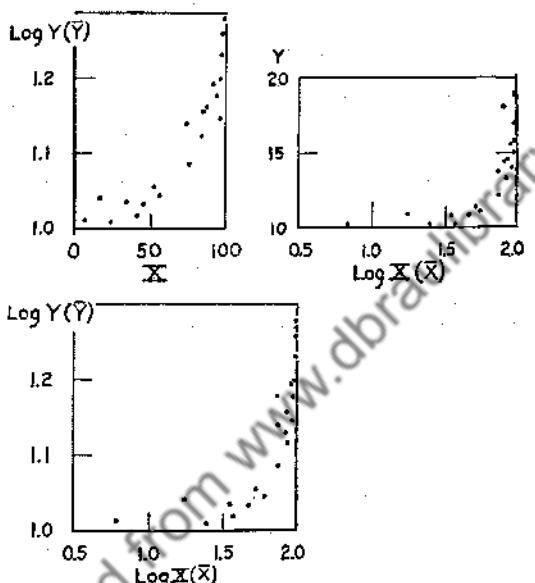

FIG. 15. Dot charts illustrating log $Y = f(X)$; $Y = f(\log X)$; log $Y = f(\log X)$.

lute 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.

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 $\overline{Y}$ and X are plotted seems to come nearest. The equation

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

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.

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

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

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 $\bar{Y}$	Extensions	
			X^2	$X\bar{Y}$
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.....	$\Sigma X = 1{,}340$	$\Sigma\bar{Y} = 22.389$	$\Sigma X^2 = 107{,}566$	$\Sigma X\bar{Y} = 1{,}545.888$

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:

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

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

Then

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

and

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

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:

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

$$\log Y = 0.9624$$

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:

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

$$\log Y = 1.1401$$

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

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

(Log $Y = 0.9469 + 0.00258\,X$)

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 $\overline{Y}'$	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

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 $\overline{Y}$, 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:

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

or in terms of natural numbers:

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

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.

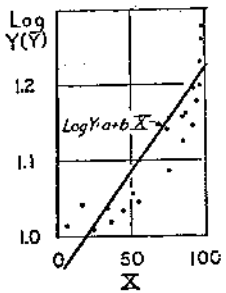

FIG. 16. Dot chart showing observations and fitted line for equation log $Y = a + bX$, in logarithms of Y.

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.

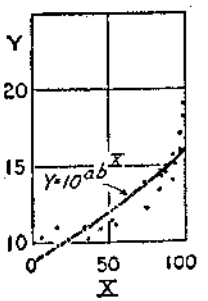

FIG. 17. Dot chart showing observations and fitted line for equation $Y = 10^{a+bX}$, in natural values of Y.

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

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

and the constants are determined from the equations

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

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

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.

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

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

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

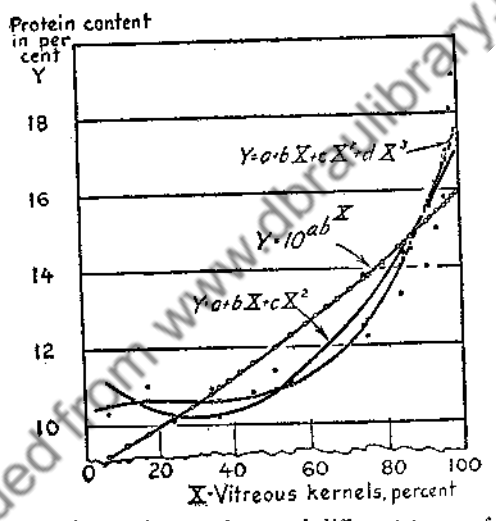

FIG. 18. Original observations, and several different types of fitted curves.

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

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.

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

[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.

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.

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.

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

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

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

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.

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, indi-

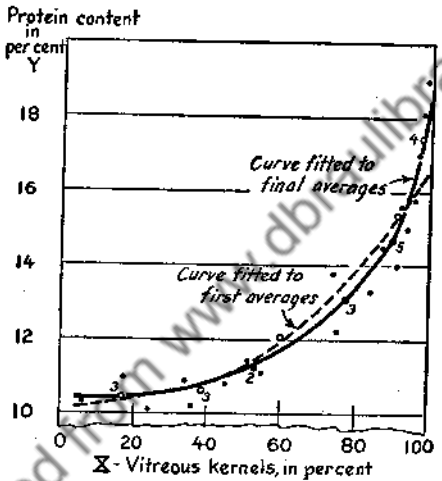

FIG. 19. Original observations and averages of protein content, and freehand curve.

cated 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.

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

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 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
	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.

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:

(*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]

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 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

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

equation stating the relation between X and Y, we still have a mathematical expression of the relation between them. For we can write

$$Y' = f(X)$$

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'

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.

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.

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 be-

ginning 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

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

$$= 16t^2$$

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

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

or

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

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

$$d = bt^2$$

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.

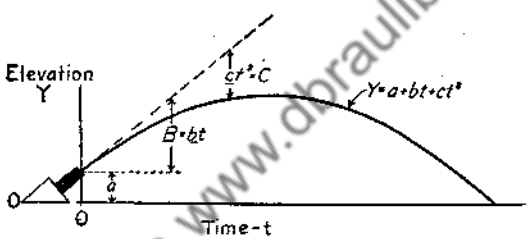

FIG. 20. The trajectory of a projectile, illustrating the equation $Y = a + bX + cX^2$.

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.

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 up-

ward 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

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

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.

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.

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.

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.

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	$\overline{C}\overline{P}$	$\overline{P}^2$
	Billion pounds	*Dollars per cwt.*	$\overline{C}$	$\overline{P}$		
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.

$$M_{\bar{c}} = \frac{\Sigma \bar{C}}{n} = \frac{6.1781}{6} = 1.02968$$

$$M_{\bar{p}} = \frac{\Sigma \bar{P}}{n} = \frac{5.8769}{6} = 0.97948$$

$$\Sigma(\bar{c}\bar{p}) = \Sigma(\overline{CP}) - nM_{\bar{c}}M_{\bar{p}}$$

$$= 6.03610 - 6(1.02968)(0.97948) = -0.01521$$

$$\Sigma(\bar{p}^2) = \Sigma(\bar{P}^2) - nM_{\bar{p}}^2 = 5.79243 - 6(0.97948)^2 = 0.03614$$

$$b_{\bar{c}\bar{p}} = \frac{\Sigma(\bar{c}\bar{p})}{\Sigma(\bar{p}^2)} = \frac{-0.01521}{0.03614} = -0.42086$$

$$a_{\bar{c}\bar{p}} = M_{\bar{c}} - bM_{\bar{p}} = 1.02968 - (-0.42086)(0.97948)$$

$$\bar{C} = a_{\bar{c}\bar{p}} + b_{\bar{c}\bar{p}}\bar{P}$$

$$= 1.4419 - 0.42086\bar{P}$$

$$\log C = 1.4419 - 0.4209 \log P$$

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]

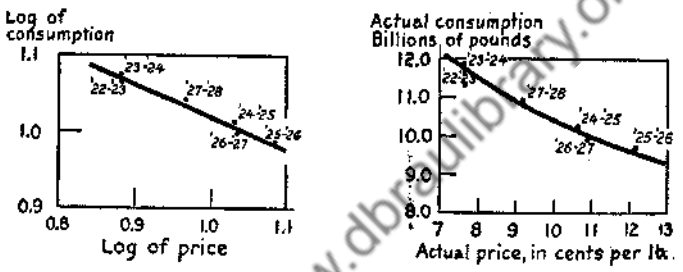

FIG. 21. The relation of consumption of hog products to hog prices, fitted by a logarithmic demand curve, both in logarithms of consumption and price and in natural numbers.

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).

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-

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

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.

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.)

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. $S_{y.x}$ indicates the standard error for estimates of Y made from a linear relation to X, by the equation $Y = a + bX$. Similarly, $S_{y.f(x)}$ 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:

$$\left.\begin{aligned} S^2_{y.x} &= \sigma^2_z = \frac{\Sigma z^2}{n} \\ S^2_{y.f(x)} &= \sigma^2_{z''} = \frac{\Sigma (z'')^2}{n} \end{aligned}\right\} \qquad (20.1)$$

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:

$$\overline{S}^2_{y\cdot x} = \frac{n\sigma^2_z}{n-2} = \frac{nS^2_{y\cdot x}}{n-2} \tag{21.1}$$

hence

$$\overline{S}^2_{y\cdot x} = \frac{\Sigma(z^2)}{n-2} = \sigma^2_z\left(\frac{n}{n-2}\right) \tag{21.2}$$

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 $\pm$ 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_s 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_s, 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.

$$\mathfrak{d}_{xy} = 1 - r_{xy}^2 \tag{24.2}$$

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

$$d_{y \cdot f(x)} = \rho_{yx}^2 \tag{24.3}$$

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 coef-

ficient 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):

$$\bar{\rho}^2_{yx} = 1 - \frac{(18-1)\left(1 - \frac{3.86^2}{7.92^2}\right)}{18-3} = 0.1389$$

$$\bar{\rho}_{yx} = 0.37$$

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:

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{\sqrt{[\Sigma(X^2) - nM_x^2]\,[\Sigma(Y^2) - nM_y^2]}} \tag{27}$$

$$= \frac{970.1 - 14(1.97)(30.64)}{\sqrt{[61.82 - 14(1.97)^2]\,[16{,}057 - 14(30.64)^2]}} = 0.847$$

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

$$b_{yx} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x^2} \quad \text{or} \quad = \frac{\Sigma(xy)}{n\sigma_x^2} \tag{27.1}$$

$$r_{xy} = \frac{\Sigma(XY) - nM_xM_y}{n\sigma_x\sigma_y} = \frac{\Sigma(xy)}{n\sigma_x\sigma_y} \tag{27.2}$$

[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).

$$\bar{r}_{xy}^2 = 1 - (1 - r_{xy}^2)\left(\frac{n-1}{n-2}\right)$$

For the present problem, that becomes

$$\bar{r}_{xy}^2 = 1 - \frac{[1 - (0.847)^2]\,(14-1)}{14-2} = 0.6939$$
$$\bar{r}_{xy} = 0.833$$

Knowing $\bar{r}_{xy}$, we may next compute the standard error of estimate by the following equation:

$$\bar{S}_{yx} = \sqrt{\frac{\Sigma(Y^2) - n(M_y)^2}{n-1}(1 - \bar{r}_{xy}^2)} \qquad (28)$$
$$= \sqrt{\frac{16{,}057 - 14(30.64^2)}{13}[1 - (0.833)^2]}$$
$$= \sqrt{68.62} = 8.28$$

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.

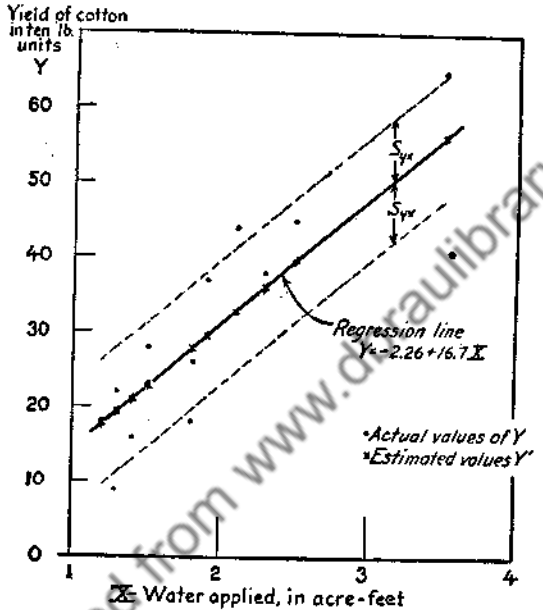

FIG. 22. Relation of yield of cotton to irrigation water applied; estimated yields from a linear regression and zone of probable yields indicated by the standard error of estimate.

The coefficient of correlation, $r_{xy} = 0.83$, 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.

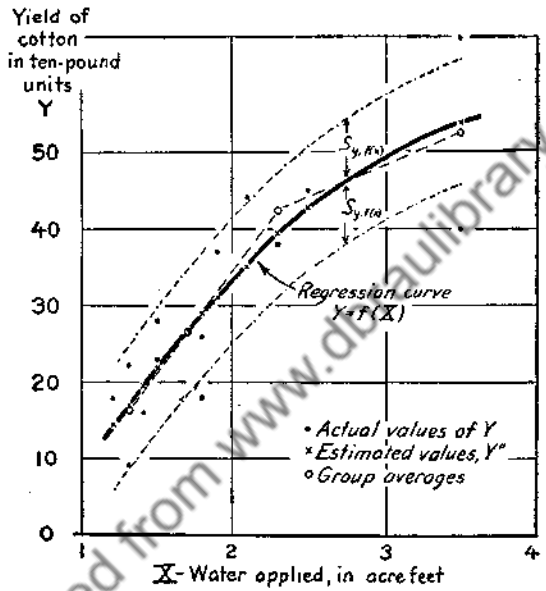

FIG. 23. Relation of yield of cotton to irrigation water applied; estimated yields from a curvilinear regression; and zone of probable yields as indicated by the standard error of estimate.

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),

$$\bar{S}^2_{y \cdot f(x)} = \sigma^2_{z''}\left(\frac{n}{n-m}\right) = (7.16^2)\left(\frac{14}{14-3}\right) = 65.23$$

$$\bar{S}_{y \cdot f(x)} = 8.07$$

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 frechand, 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 differ-

ences 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

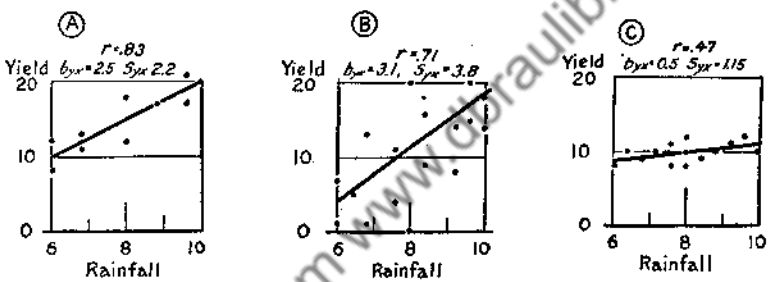

FIG. 24. Hypothetical sets of data, illustrating three types of correlation coefficients.

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]

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 . . . X_n. The

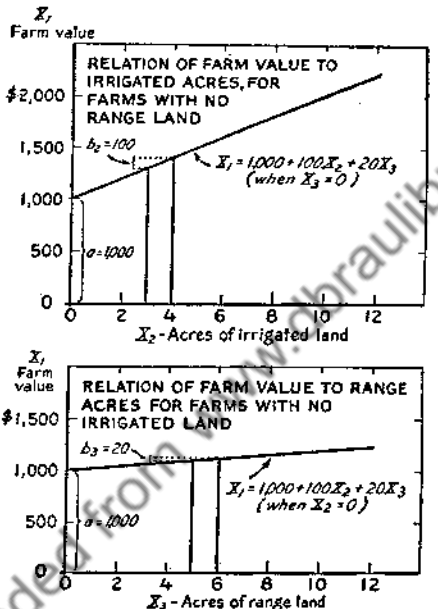

FIG. 24.1. Graph of the function $Y = 1{,}000 + 100X_2 + 20X_3$.

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 be-

tween 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

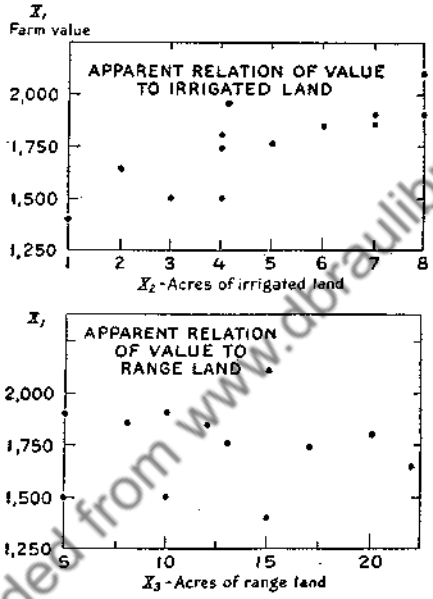

FIG. 24.2. The apparent relation of farm value to acres of irrigated land and to range land reveals little of the underlying net relationship.

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.

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.)

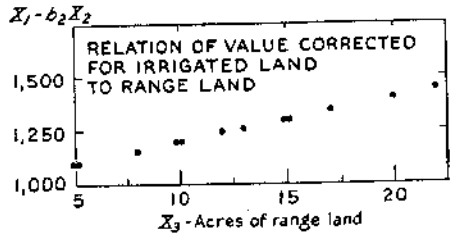

FIG. 24.3. After the net influence of irrigated land has been removed, the underlying relation of farm value to acres of range land is very clear.

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

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

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

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

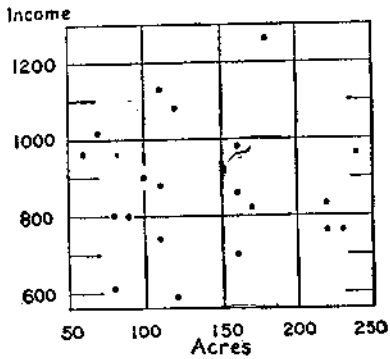

FIG. 25. Correlation chart of acres and income on individual farms.

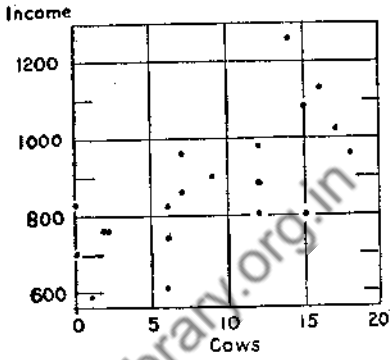

FIG. 26. Correlation chart of number of cows and income on individual farms.

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.

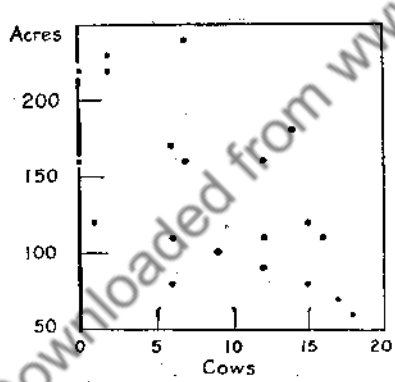

FIG. 27. Correlation chart of number of cows and number of acres on individual farms.

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.

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

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

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-

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.

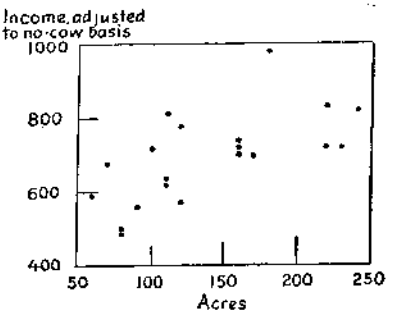

FIG. 28. Relation of income, adjusted for number of cows, to number of acres.

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:

$$\text{Income} = \$508.51 + (120)\ \$1.33 + (20)\ \$20.11$$
$$= \$1070$$

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: Estimate for acres $1.33 (acres) (A)	Computation of estimated income: Estimate for cows $20.11 (cows) (C)	Estimated income (A) + (C) +$508.51	Actual income	Actual income minus estimated income
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.

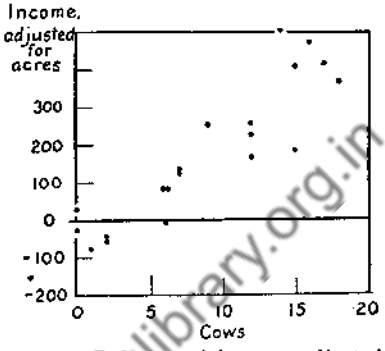

FIG. 29. Relation of income, adjusted for number of acres, to number of cows.

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).

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,
$= \$508.51 + (\text{number of acres})\ \$1.33$

(D) Income, adjusted to constant acres,
$= -\$68.77 + (\text{number of cows})\ \$27.88$

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 $= \$439.74 + (\text{number of acres})\ \$1.33 + (\text{number of cows})\ \$27.88$

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

$$\text{Estimated income} = \$439.7 + (120)\ \$1.33 + (15)\ \$27.88$$
$$= \$439.7 + \$159.6 + \$418.2 = \$1{,}018$$

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.

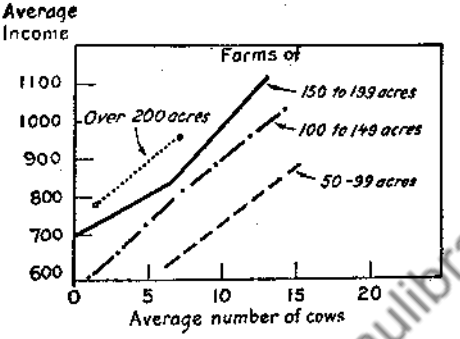

FIG. 30. Difference in average income with difference in number of cows, for farms grouped by size of farm.

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

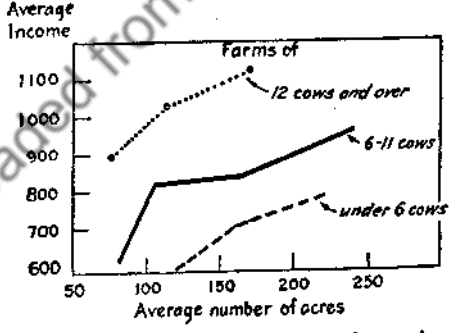

FIG. 31. Difference in average income with difference in number of acres, for farms grouped by numbers of cows.

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.

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

$$X_1 = a + b_2X_2 + b_3X_3 \tag{30}$$

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

$$\left.\begin{aligned} &\text{(I)} \quad \Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 = \Sigma x_1x_2 \\ &\text{(II)} \quad \Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 = \Sigma x_1x_3 \end{aligned}\right\} = \left\{\begin{aligned} &606.95\,b_2 - 394.15\,b_3 = 14.20 \\ &-394.15\,b_2 + 676.55\,b_3 = 1360.60 \end{aligned}\right.$$

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'):

$$(I) \qquad 606.95\, b_2 - 394.15\, b_3 = 14.20$$

$$(I') \qquad -b_2 + 0.64939\, b_3 = -0.02340$$

TABLE 45

COMPUTATION OF VALUES TO DETERMINE MULTIPLE REGRESSION EQUATION TO ESTIMATE ONE VARIABLE FROM TWO OTHERS

	1 Number of acres* X_2	2 Number of cows X_3	3 Number of dollars income* X_1	4 X_2^2	5 X_2X_3	6 X_1X_2	7 X_3^2	8 X_1X_3	9 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:

$$\begin{array}{lrl} & (\text{II}) & -394.15\,b_2 + 676.55\,b_3 = 1360.60 \\ (0.64939) & (\text{I}) & +394.15\,b_2 - 255.96\,b_3 = \quad 9.22 \\ & (\Sigma\text{II}) & 420.59\,b_3 = 1369.82 \\ & (\text{II}') & b_3 = \quad 3.25690 \end{array}$$

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.

$$\left.\begin{aligned} \Sigma(x_2^2)b_{12.34} + \Sigma(x_2x_3)b_{13.24} + \Sigma(x_2x_4)b_{14.23} &= \Sigma(x_1x_2) \\ \Sigma(x_2x_3)b_{12.34} + \Sigma(x_3^2)b_{13.24} + \Sigma(x_3x_4)b_{14.23} &= \Sigma(x_1x_3) \\ \Sigma(x_2x_4)b_{12.34} + \Sigma(x_3x_4)b_{13.24} + \Sigma(x_4^2)b_{14.23} &= \Sigma(x_1x_4) \end{aligned}\right\} \quad (38)$$

$$a_{1.234} = M_1 - b_{12.34}M_2 - b_{13.24}M_3 - b_{14.23}M_4 \quad (39)$$

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.

$$\left.\begin{aligned} \text{(I)}\quad & \Sigma(x_2^2)b_2 + \Sigma(x_2x_3)b_3 + \Sigma(x_2x_4)b_4 = \Sigma(x_1x_2) \\ \text{(II)}\quad & \Sigma(x_2x_3)b_2 + \Sigma(x_3^2)b_3 + \Sigma(x_3x_4)b_4 = \Sigma(x_1x_3) \\ \text{(III)}\quad & \Sigma(x_2x_4)b_2 + \Sigma(x_3x_4)b_3 + \Sigma(x_4^2)b_4 = \Sigma(x_1x_4) \end{aligned}\right\} = \left\{\begin{aligned} & 606.95b_2 - 394.15b_3 + 63.20b_4 = 14.20 \\ & -394.15b_2 + 676.55b_3 + 11.60b_4 = 1360.60 \\ & 63.20b_2 + 11.60b_3 + 17.20b_4 = 193.20 \end{aligned}\right.$$

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	10	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:

$$\text{(I)} \qquad 606.95b_2 - 394.15b_3 + 63.20b_4 = 14.20$$

$$\text{(I′)} \qquad -b_2 + 0.64939b_3 - 0.10413b_4 = -0.02340$$

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:

(II)	$-394.15b_2 + 676.55b_3 + 11.60b_4$	=	1360.60
(0.64939) (I)	$394.15b_2 - 255.96b_3 + 41.04b_4$	=	9.22
(Σ_2)	$420.59b_3 + 52.64b_4$	=	1369.82
(II')	$-b_3 - 0.12516b_4$	=	− 3.25690

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:

(I)	$606.95b_2 - 394.15b_3 + 63.20b_4$	=	14.20
(I')	$-b_2 + 0.64939b_3 - 0.10413b_4$	= −	0.02340
(II)	$-394.15b_2 + 676.55b_3 + 11.60b_4$	=	1360.60
(0.64939) (I)	$394.15b_2 - 255.96b_3 + 41.04b_4$	=	9.22
(Σ_2)	$420.59b_3 + 52.64b_4$	=	1369.82
(II')	$-b_3 - 0.12516b_4$	= −	3.25690
(III)	$63.20b_2 + 11.60b_3 + 17.20b_4$	=	193.20
(−0.10413) (I)	$-63.20b_2 + 41.04b_3 - 6.58b_4$	= −	1.48
(−0.12516) (Σ_2)	$-52.64b_3 - 6.59b_4$	= −	171.45
(Σ_3)	$4.03b_4$	=	20.27
(III')	$-b_4$	= −	5.02978

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.

$$a = M_1 - b_2M_2 - b_3M_3 - b_4M_4$$
$$= 87.2 - (1.20584)(13.95) - (2.62737)(8.85) - (5.02978)(2.20)$$
$$= 36.06$$

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:

$$\left.\begin{aligned}
\Sigma(x_2^2)b_{12.345} &+ \Sigma(x_2x_3)b_{13.245} + \Sigma(x_2x_4)b_{14.235} \\
&+ \Sigma(x_2x_5)b_{15.234} = \Sigma(x_1x_2) \\
\Sigma(x_2x_3)b_{12.345} &+ \Sigma(x_3^2)b_{13.245} + \Sigma(x_3x_4)b_{14.235} \\
&+ \Sigma(x_3x_5)b_{15.234} = \Sigma(x_1x_3) \\
\Sigma(x_2x_4)b_{12.345} &+ \Sigma(x_3x_4)b_{13.245} + \Sigma(x_4^2)b_{14.235} \\
&+ \Sigma(x_4x_5)b_{15.234} = \Sigma(x_1x_4) \\
\Sigma(x_2x_5)b_{12.345} &+ \Sigma(x_3x_5)b_{13.245} + \Sigma(x_4x_5)b_{14.235} \\
&+ \Sigma(x_5^2)b_{15.234} = \Sigma(x_1x_5)
\end{aligned}\right\} \tag{40}$$

$$a_{1.2345} = M_1 - b_{12.345}M_2 - b_{13.245}M_3 - b_{14.235}M_4 - b_{15.234}M_5 \tag{41}$$

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_2	x_3	x_4	x_5	x_6	x_7	x_8	x_1
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:

$$S^2_{1.234} = \frac{n\sigma^2_{z_{1.234}}}{n - m} \tag{42}$$

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 $\overline{S}^2_{1.234}$ 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	$\overline{S}_{1.2} = 165.15$
X_2, X_3	70.48	20	3	$\overline{S}_{1.23} = 76.45$
X_2, X_3, X_4	66.77	20	4	$\overline{S}_{1.234} = 74.65$

*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, $\overline{S}_{1.2} = \sigma_1$.) 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:

$$\overline{S}^2_{1.234\ldots n} = \frac{\left\{\begin{array}{c}\Sigma(x_1^2) - [b_{12.34\ldots n}(\Sigma x_1 x_2) + b_{13.24\ldots n}(\Sigma x_1 x_3) \\ + \ldots + b_{1n.23\ldots(n-1)}(\Sigma x_1 x_n)]\end{array}\right\}}{n - m} \tag{43}$$

Substituting the values for the regression equation computed with two independent variables, pages 193 and 194, the equation becomes

$$\overline{S}^2_{1.23} = \frac{\Sigma(x_1^2) - [b_{12.3}(\Sigma x_1 x_2) + b_{13.2}(\Sigma x_1 x_3)]}{n - 3}$$

In terms of coded values for X_1,

$$\frac{\overline{S}^2_{1.23}}{10^2} = \frac{5{,}455.20 - (2.1384)(14.20) - (3.2569)(1{,}360.60)}{20 - 3}$$

$$\frac{\overline{S}_{1.23}}{10} = \sqrt{\frac{993.50}{17}} = 7.645; \quad \overline{S}_{1.23} = 76.45$$

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:

$$X_{1(234)} = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{44}$$

$$R_{1.234} = \frac{\sigma_{1(234)}}{\sigma_1} \tag{45}$$

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]

$$R^2_{1.234\ldots n} = \frac{\left\{\begin{matrix} b_{12.34\ldots n}(\Sigma x_1 x_2) + b_{13.24\ldots n}(\Sigma x_1 x_3) \\ + \ldots + b_{1n.23\ldots(n-1)}(\Sigma x_1 x_n) \end{matrix}\right\}}{\Sigma(x_1^2)} \qquad (46)$$

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:

$$\bar{R}^2_{1.234\ldots n} = 1 - (1 - R^2_{1.234\ldots n})\left(\frac{n-1}{n-m}\right) \qquad (47)$$

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:

$$\bar{R}^2_{1.234\ldots n} = 1 - \left(\frac{\bar{S}^2_{1.234\ldots n}}{\sigma_1^2}\right)\left(\frac{n-1}{n}\right) \qquad (48)$$

$$\bar{S}^2_{1.234\ldots n} = \sigma_1^2(1 - \bar{R}^2_{1.234\ldots n})\left(\frac{n}{n-1}\right) \qquad (49)$$

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 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

$$r^2_{12\overline{34}} = \frac{b^2_{12.34}\sigma^2_2}{b^2_{12.34}\sigma^2_2 + \sigma^2_1(1 - R^2_{1.234})} \qquad (51)$$

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 $R_{1.24}$

$$(\Sigma x_2^2)b_{12.4} + (\Sigma x_3 x_4)b_{14.2} = (\Sigma x_1 x_2)$$
$$(\Sigma x_2 x_4)b_{12.4} + (\Sigma x_4^2)b_{14.2} = (\Sigma x_1 x_4)$$

Similarly for $R_{1.34}$

$$(\Sigma x_3^2)b_{13.4} + (\Sigma x_3 x_4)b_{14.3} = (\Sigma x_1 x_3)$$
$$(\Sigma x_3 x_4)b_{13.4} + (\Sigma x_4^2)b_{14.3} = (\Sigma x_1 x_4)$$

RELATIVE IMPORTANCE OF INDIVIDUAL FACTORS AFFECTING INCOME, AS INDICATED BY COEFFICIENTS OF PARTIAL CORRELATION

Factors already considered	Factor added	Coefficient of partial correlation ($r_{12.34}$, etc.)	Reduction in unexplained variance ($r^2_{12.34}$, 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 inter-

correlations 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 ($r_{12.34}$)	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]

$$\left.\begin{array}{l}\text{Multiple correlation}\\ \text{squared of}\\ (X_1 - b_{12.34}X_2)\\ \text{with } X_3 \text{ and } X_4\end{array}\right\} = 1 - \frac{\sigma_1^2(1 - R_{1.234}^2)}{\sigma_1^2 - 2b_{12.34}(\Sigma x_1 x_2/n) + b_{12.34}^2\sigma_2^2} \qquad (53)$$

[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

$$X_1 = a' + f_2(X_2) + f_3(X_3) + f_4(X_4) + \text{etc.} \qquad (54)$$

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

$$X_1 = a + b_2X_2 + b_{2'}(X_2^2) + b_3X_3 + b_{3'}(X_3^2) + b_4X_4 + b_{4'}(X_4^2) \qquad (55)$$

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

$$X_1 = a + b_2(X_2) + b_{2'}(X_2^2) + b_{2''}(X_2^3) + b_3(X_3) + b_{3'}(X_3^2) + b_{3''}(X_3^3) + b_4(X_4) + b_{4'}(X_4^2) + b_{4''}(X_4^3) \qquad (56)$$

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 inde-

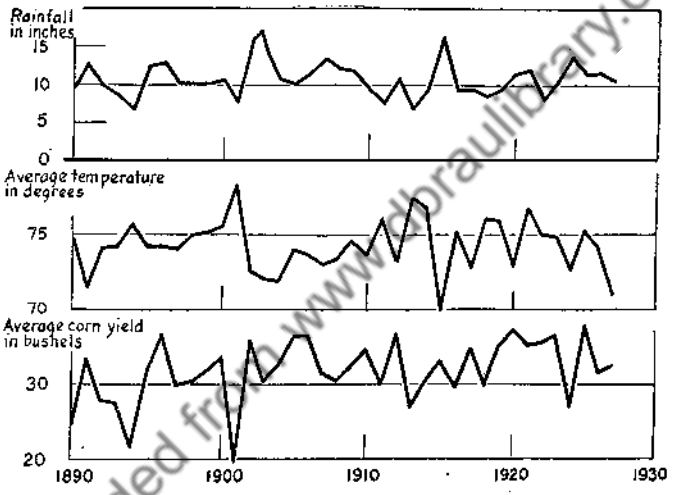

FIG. 32. Rainfall, temperature, and corn yields in the Corn Belt, 1890 to 1927.

pendent 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.

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 aver-

age 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.

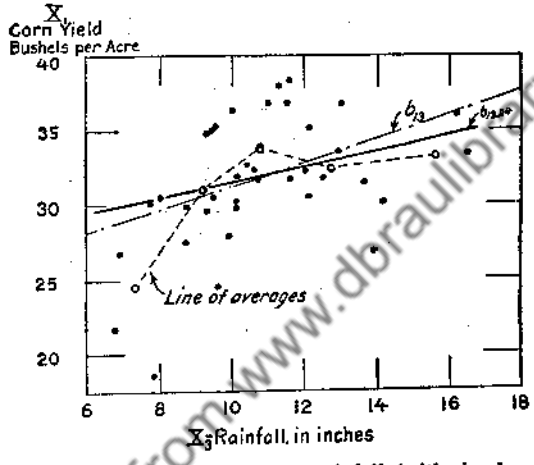

FIG. 33. Apparent relation of corn yields to rainfall (with simple and net regression lines).

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:

(A) $$X_1 = a_{1.234} + b_{12.34}M_2 + b_{13.24}X_3 + b_{14.23}M_4$$

To these values have been added the residuals, z. These residuals equal $X_1 - X_1'$, and therefore for each observation are equal to

(B) $$X_1 - a_{1.234} - b_{12.34}X_2 - b_{13.24}X_3 - b_{14.23}X_4$$

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

$$\pi = a_{1.234} + b_{12.34}M_2 + b_{13.24}X_3 + b_{14.23}M_4 + X_1 - a_{1.234} - b_{12.34}X_2 - b_{13.24}X_3 - b_{14.23}X_4$$

$$\pi = X_1 - b_{12.34}(X_2 - M_2) - b_{14.23}(X_4 - M_4)$$

$$\pi = X_1 - b_{12.34}x_2 - b_{14.23}x_4$$

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.

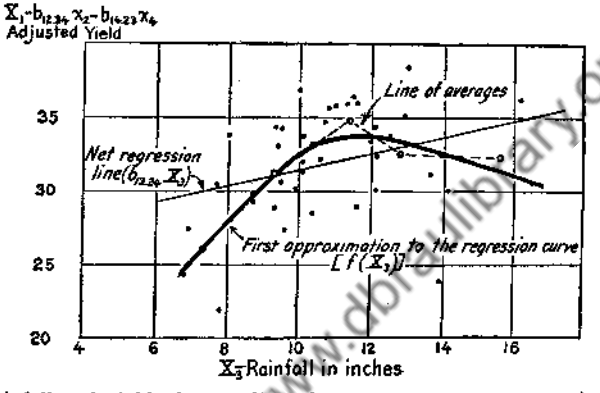

FIG. 34. Rainfall and yield of corn adjusted to average temperature and year, and first approximation curve fitted to the averages. [The notation $f(X_3)$ on the figure corresponds to $f_3(X_3)$.]

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.

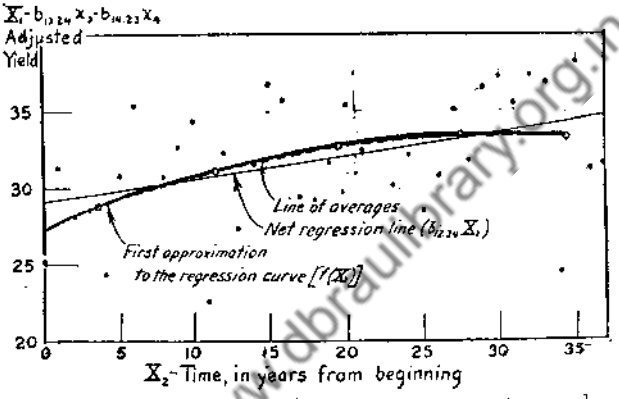

FIG. 35. Time and yield of corn adjusted to average temperature and rainfall, and first approximation curve fitted to the averages. [The notation $f(X_2)$ on the figure corresponds to $f_2(X_2)$.]

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

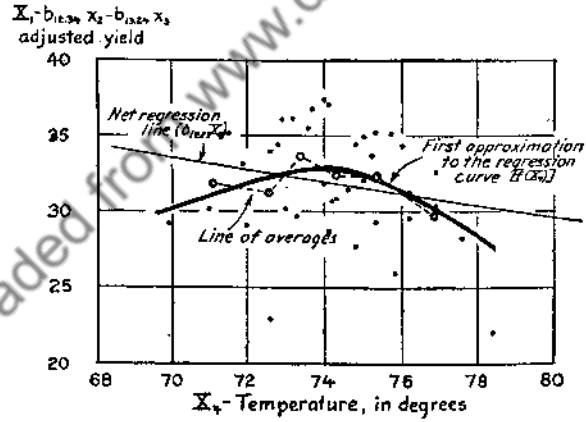

FIG. 36. Temperature and yield of corn adjusted to average rainfall and year, and first approximation curve fitted to the averages. [The notation $f(X_4)$ on the figure corresponds to $f_4(X_4)$.]

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.

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) cor-

responding 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

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				

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]

[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

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 X_2	2 X_3	3 X_4	4 $f_2'(X_2)$	5 $f_3'(X_3)$	6 $f_4'(X_4)$	7 $f_2'(X_2)+f_3'(X_3)+f_4'(X_4)$	8 $\Sigma(f)+a'=X_1''$	9 X_1	10 $X_1-X_1''=x''$
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	10.5	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			

to equation (57), and the constant which has just been computed. When equations (57) and (58) are compared, it is evident that, except

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'' di-

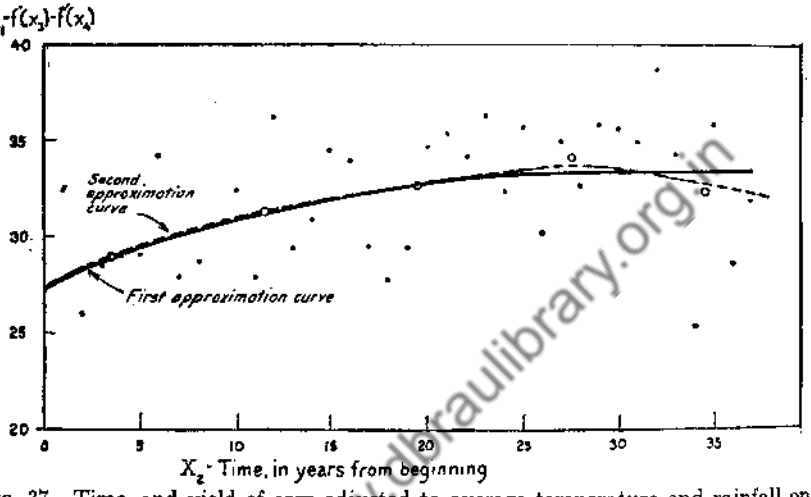

FIG. 37. Time, and yield of corn adjusted to average temperature and rainfall on basis of first approximation curves; and second approximation to $f_2(X_2)$.

rectly above or below from the corresponding point on the regression curve.

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,

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

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.

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

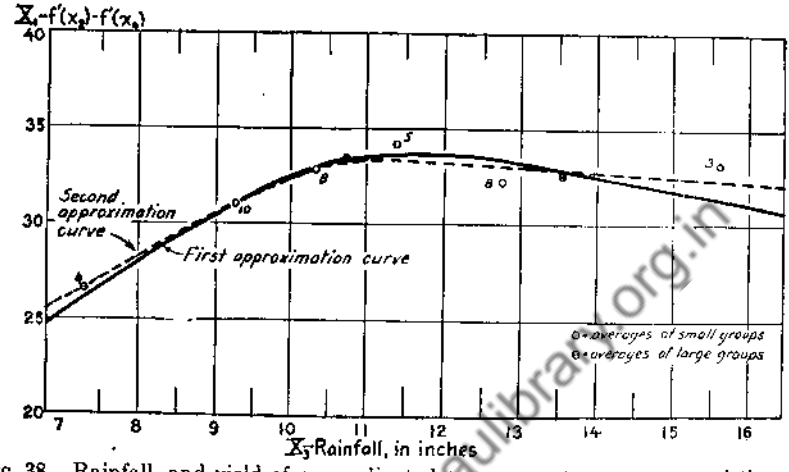

FIG. 38. Rainfall, and yield of corn adjusted to average temperature and time on the basis of first approximation curves; and second approximation to $f_3(X_3)$.

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

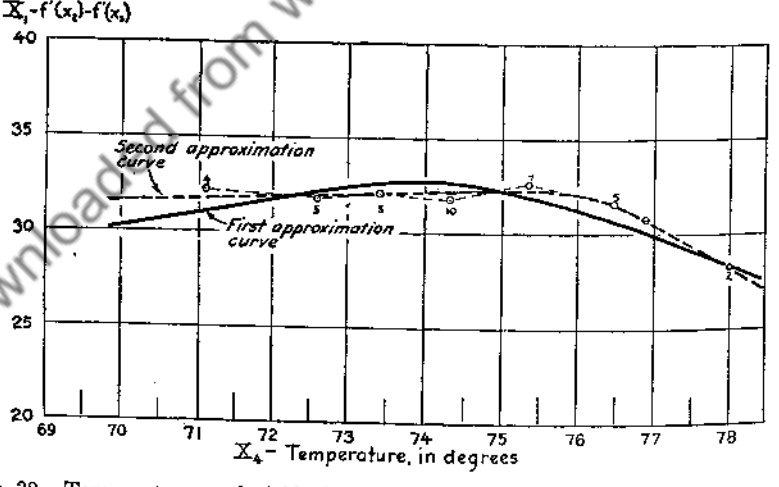

FIG. 39. Temperature, and yield of corn adjusted to average rainfall and time on the basis of first approximation curves; and second approximation to $f_4(X_4)$.

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

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		

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-

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

$$a''_{1.234} = 31.916 - 35.300$$
$$= -3.384$$

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	12.3	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

$$X_1' = F_2(X_2) = f_2(X_2) - M_{f(2)} + M_1 \tag{60}$$

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

$$x_1' = F_2(x_2) = f_2(X_2) - M_{f(2)} \tag{61}$$

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

$$X_1' = F_2(x_2) + F_3(X_3) + F_4(x_4) + \cdots + F_n(x_n) \tag{62}$$

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 Table 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

$$F_4(x_4) = f_4''(X_4) - M_{f(X_4)}$$
$$= f_4''(X_4) - 11.81$$

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

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.

influence of various quantities of rain and degrees of temperature. These estimates would then be defined by the equation

$$X_1' = F_2(x_2) + F_3(X_3) + F_4(x_4)$$
$$= 0.4 + F_3(X_3) + F_4(x_4)$$

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

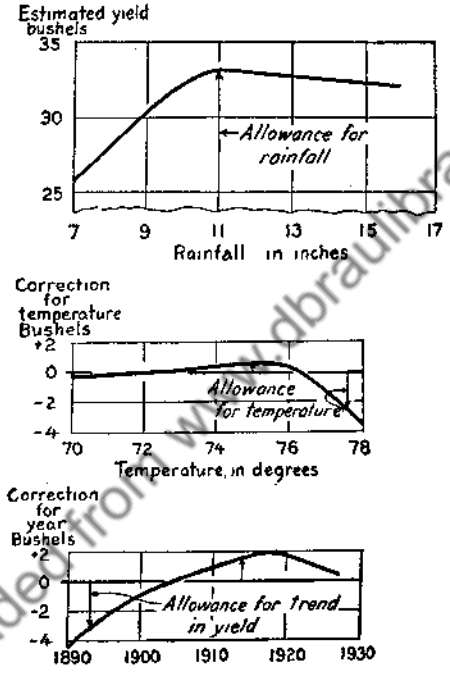

FIG. 40. Relation of yield of corn to rainfall, temperature, and time.

on the surface and to provide a basis for estimating the probable effect on the dependent factor of new combinations of the independent factors.

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.

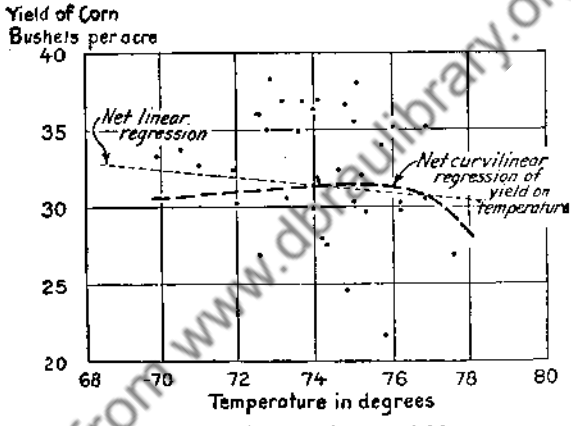

FIG. 41. Comparison of apparent relation of corn yields to temperature with net relation after eliminating influence of rainfall and of trend in yield.

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:

$$\bar{S}^2_{1.f(2,3,\text{etc.})} = \frac{\sigma^2_{z\,1.f(2,3,\text{etc.})}}{1 - m/n} \tag{64}$$

or

$$\bar{S}^2_{1.f(2,3,4,\text{etc.})} = \frac{n\sigma_z^2}{n - m} = \frac{\Sigma(z^2)}{n - m} \tag{65}$$

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

$$\mathrm{P}^2 = 1 - \frac{\sigma_{z''}^2}{\sigma_1^2} \tag{66.1}$$

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.

$$\overline{P}^2_{1.234} = 1 - \left[\left(\frac{\overline{S}^2_{1.f(2,3,4)}}{\sigma_1^2}\right)\left(\frac{n-1}{n}\right)\right] \tag{66.2}$$

$$\left.\begin{aligned}\overline{P}^2_{1.234} &= 1 - \left[\left(\frac{\sigma^2_{z_{1.f(2,3,4)}}}{\sigma_1^2}\right)\left(\frac{n-1}{n-m}\right)\right] \\ &= 1 - \left[\left(\frac{\Sigma(z^2_{1.f(2,3,4)})}{\Sigma(x_1^2)}\right)\left(\frac{n-1}{n-m}\right)\right]\end{aligned}\right\} \tag{66.3}$$

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, $\overline{P}$
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

$$\bar{P}^2_{1.234} = 1 - \frac{\bar{S}^2_{1.f(234)}}{\sigma_1^2}\left(\frac{n-1}{n}\right)$$

$$= 1 - \frac{(3.46)^2}{(4.30)^2}\left(\frac{37}{38}\right)$$

$$= 0.369$$

$$\bar{P}_{1.234} = 0.61$$

The index of multiple correlation is therefore 0.61, as compared with the coefficient of multiple correlation of 0.49. The total determination, which 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, $P_{1.234}$. 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:

$$\bar{P}^2_{1.234} = 1 - (1 - R^2_{1.2'3'4'}) \frac{n-1}{n-m} \tag{67}$$

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.)

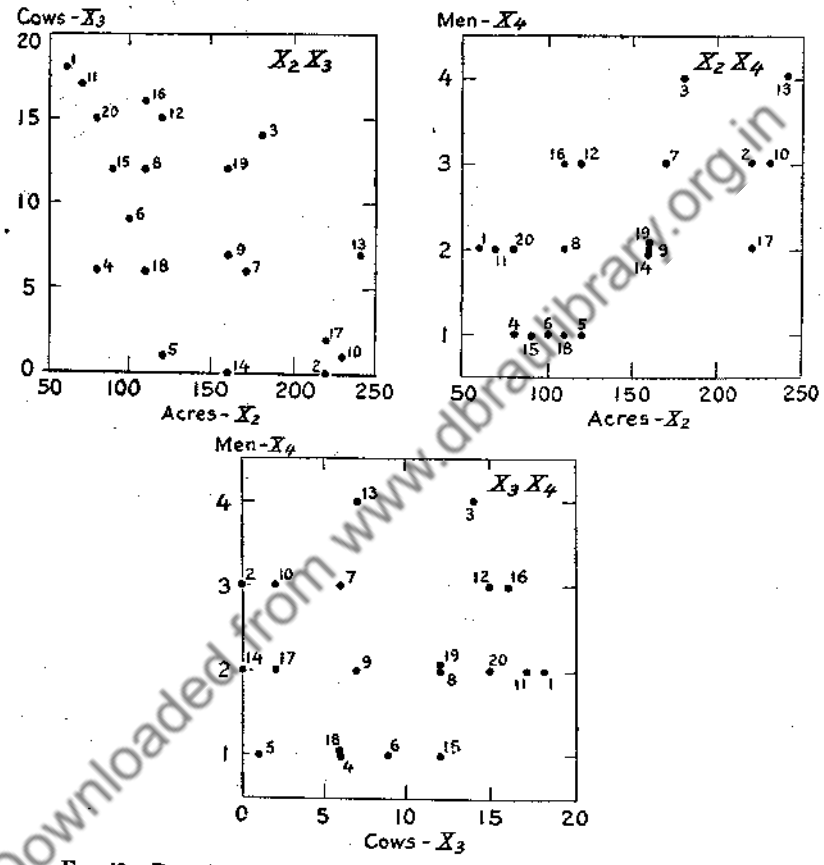

FIG. 42. Dot charts showing the intercorrelations of the independent variables, X_2, X_3, and X_4.

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

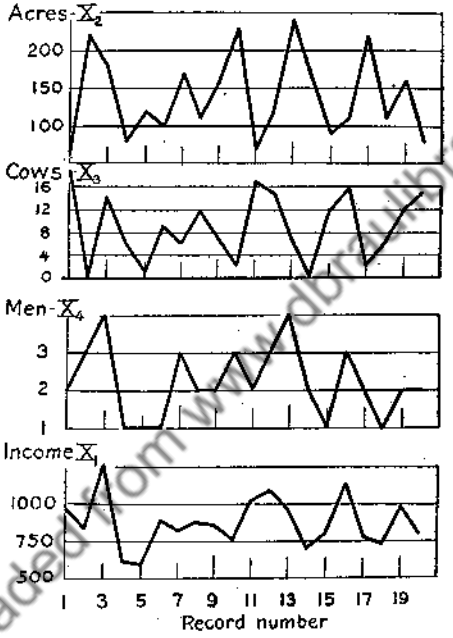

FIG. 43. Acres, cows, men, and income, on 20 farms.

in cows, as shown by these four observations. (The values may be taken from Table 69, page 277.)

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

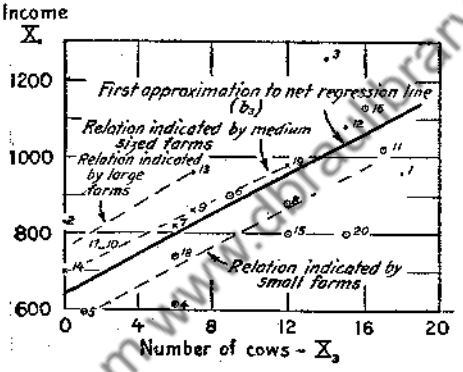

FIG. 44. Income plotted against cows, on specified farms, and first approximation to net linear regression on cows.

dots for the remaining farms, numbers 1, 3, and 12, are then plotted in, with the numbers to indicate their identity.

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 cor-

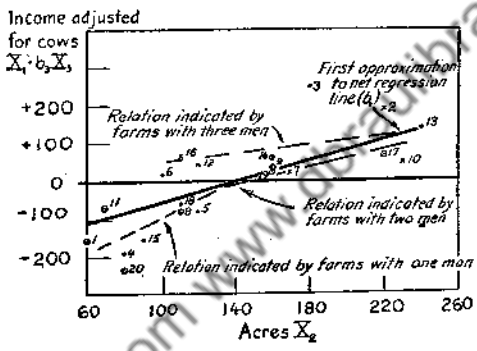

FIG. 45. Income adjusted for cows (by first approximate regression), plotted against acres on specified farms, and first approximation to net linear regression on acres.

responding 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.

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

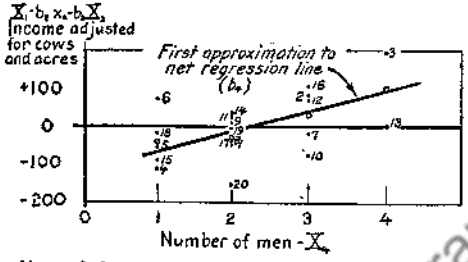

FIG. 46. Income adjusted for cows and acres (by first approximate regressions), plotted against number of men on specified farms, and first approximation to net linear regression on men.

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

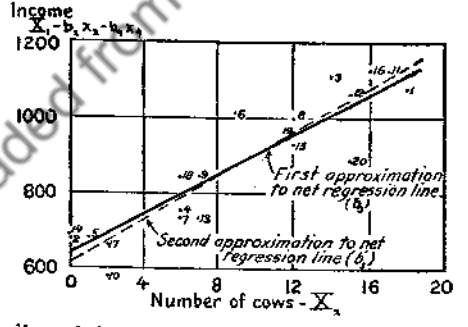

FIG. 47. Income adjusted for acres and men (by first approximate regressions), plotted against cows, and first and second approximations to net regressions on cows.

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

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

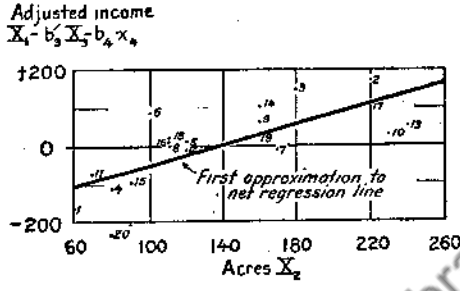

FIG. 48. Income adjusted for cows (by second approximate regression) and men (by first approximation), plotted against acres.

process, carried out by graphic plotting just as before, is shown in Figure 48.

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

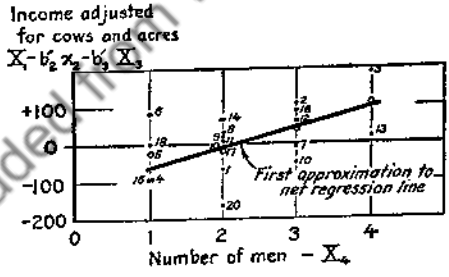

FIG. 49. Income adjusted for cows and acres (by second approximate regressions), plotted against men.

the straight line, but for the present we are considering linear relations only.

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:

$$\overline{S}^2_{1.234} = \frac{n\sigma_z^2}{n-m} = \frac{20(4{,}632)}{16} = 5{,}790$$

$$\overline{S}_{1.234} = 76.09$$

$$\overline{R}^2_{1.234} = 1 - \frac{\overline{S}^2_{1.234}}{\sigma_1^2}\left(\frac{n-1}{n}\right) = 1 - \frac{5{,}790}{27{,}276}\left(\frac{19}{20}\right) = 0.798$$

$$\overline{R}_{1.234} = 0.893.$$

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.)

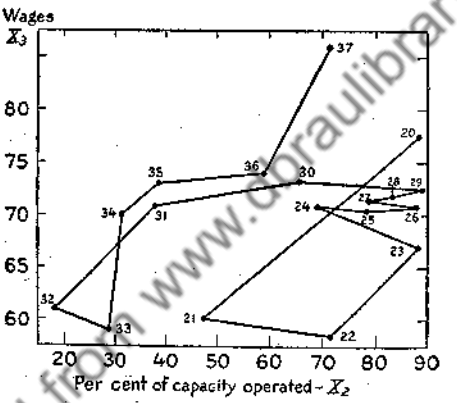

FIG. 50. Wages and per cent of capacity operated, with successive observations connected to indicate shift in the X_2X_3 relationship with time.

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

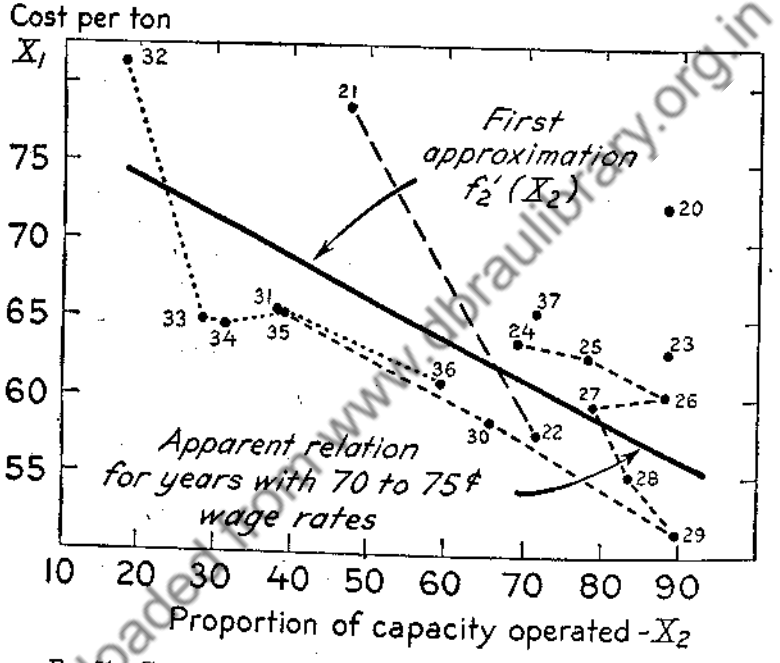

FIG. 51. Cost per ton and per cent of capacity operated, and first approximation to $f_2(X_2)$.

1921 may be ignored for the moment, as possibly reflecting the high price levels at the end of the first World War inflation.

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

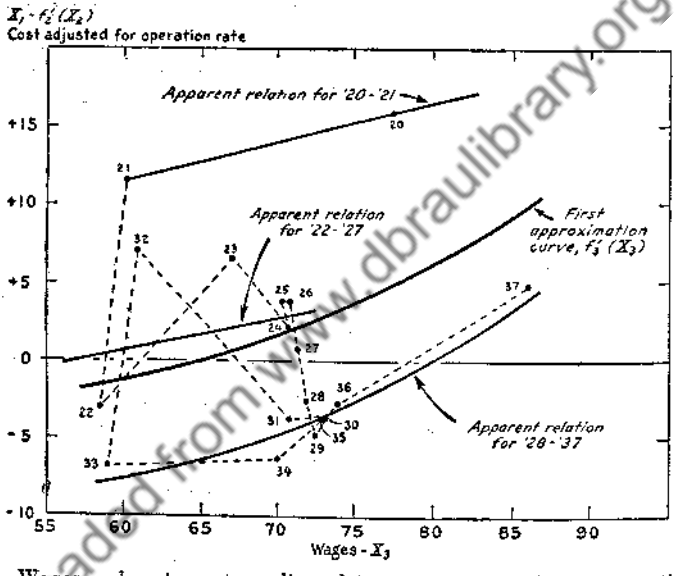

FIG. 52. Wages and cost per ton adjusted to average operation rate on the basis of the first approximation, and first approximation to $f_3(X_3)$.

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

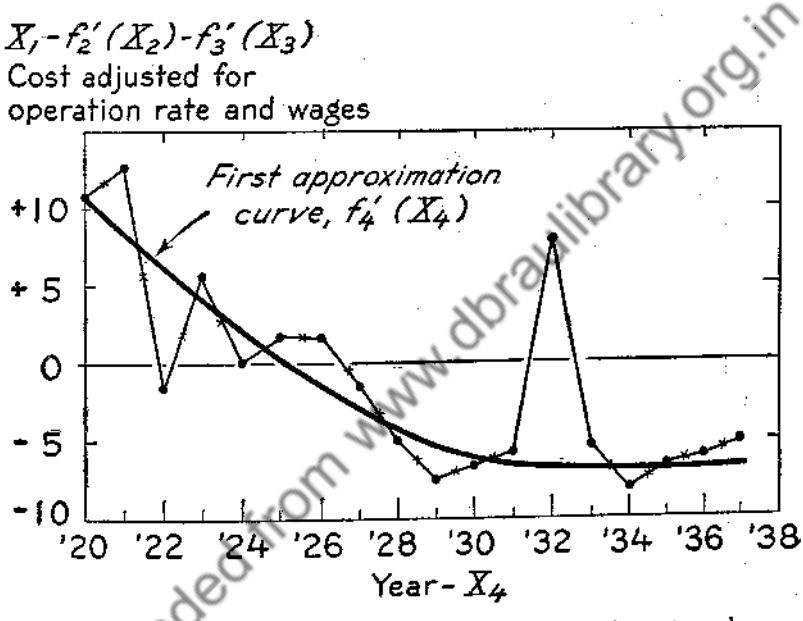

FIG. 53. Time and cost per ton adjusted to average operation rate and wages, on the basis of the first approximation curves, and first approximation to $f_4(X_4)$.

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.

[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 correc-

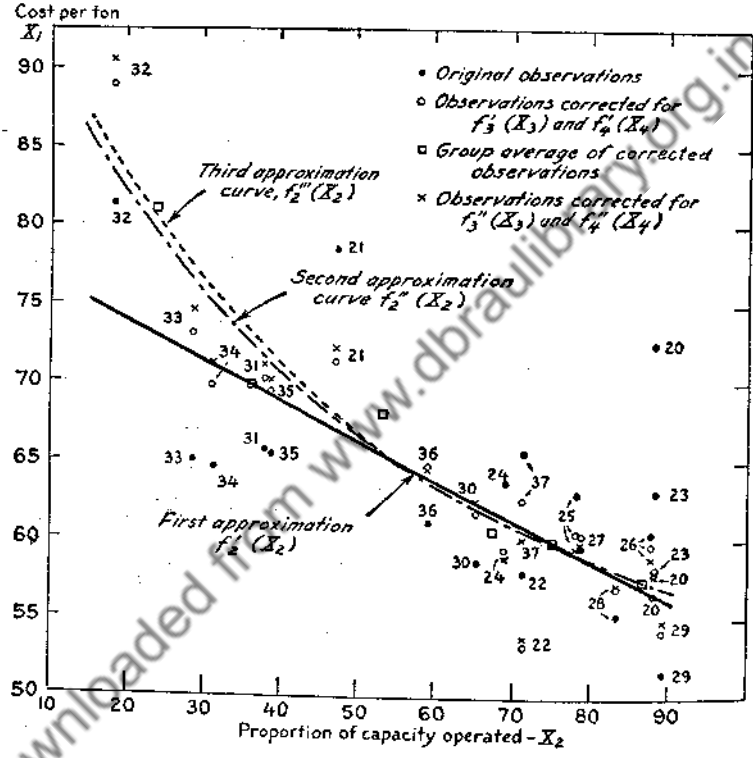

FIG. 54. Per cent of capacity operated, and cost per ton unadjusted and adjusted to average values of other variables, and second and third approximations to $f_2(X_2)$.

tion 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.

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

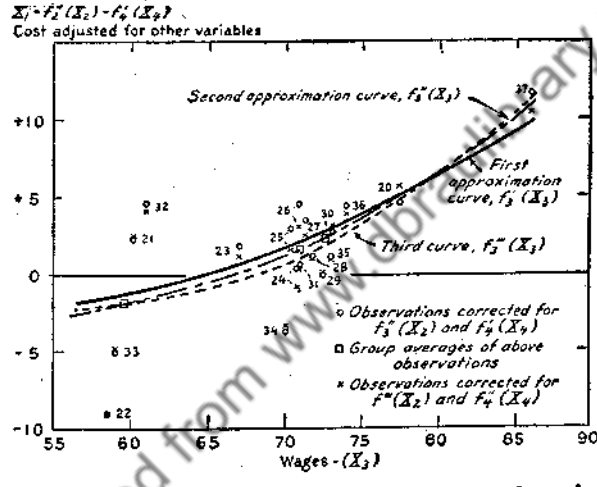

FIG. 55. Wages, and cost per ton adjusted to average values of all other variables, and second and third approximations to $f_3(X_3)$.

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

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

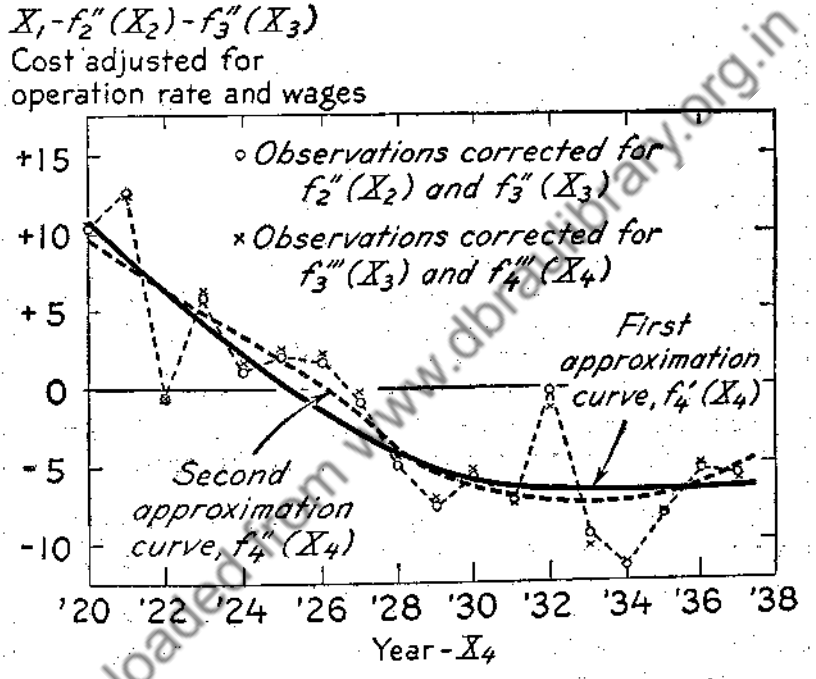

FIG. 56. Time, and cost per ton adjusted to average operation rate and wages on basis of second approximation curves; and second approximation to $f_4(X_4)$.

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 be-

tween 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 mathe-

matical curves to the regression functions graphically determined. Accordingly, $n = 18$ and $m = 8$. With these values, we can now compute $\overline{S}$ and $\overline{P}$ by equations (65) and (66.2).

$$\overline{S}^2_{1.f(2,3,4)} = \frac{n\sigma^2_{z'''}}{n-m} = \frac{18(2.88^2)}{18-8} = 14.9299$$

$$\overline{S}_{1.f(2,3,4)} = 3.86$$

$$\overline{P}^2_{1.234} = 1 - \frac{\overline{S}^2_{1.f(2,3,4)}}{\sigma^2_1}\left(\frac{n-1}{n}\right) = 1 - \frac{14.9299}{(7.19)^2}\left(\frac{17}{18}\right)$$

$$= .7272$$

$$\overline{P}_{1.234} = 0.85$$

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 $\bar{S}_{1.f(2,3,4)}$. 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.

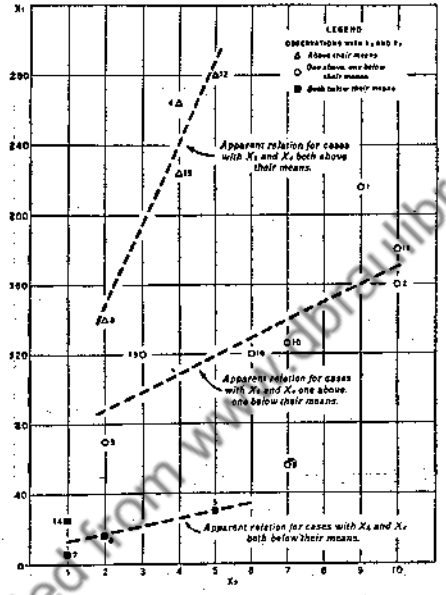

FIG. 57. Relation of X_1 to X_2, with observations classified on X_3 and X_4. When natural numbers are used, the net regression of X_1 on X_2 appears to shift with the accompanying values of X_3 and X_4.

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.

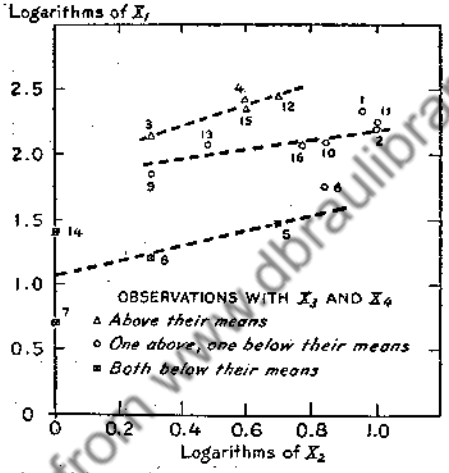

FIG. 58. When the logarithms of the data shown in Figure 57 are used, the net regression of X_1 on X_2 is found to be about the same, regardless of the accompanying values of X_3 and X_4.

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 indi-

vidual 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	$(M_0)^2 n_0$
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

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

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.

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 $\bar{P}_{1.234} = 0.59$; $\bar{P}_{1.2345} = 0.62$. 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

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

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

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 from samples drawn from the same universe, such as those listed in Table 74, with each sample regression followed by $\pm$ 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 ($1\ \sigma_b$); about 8 out of each 100 would miss by as much as 0.24 ($2\ \sigma_b$); and 15 samples out of each 1,000 would miss by as much as $3\ \sigma_b$, 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

$$\sigma_{M_{y'}} = \frac{\overline{S}_{y.x}}{\sqrt{n}} \tag{70}$$

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 $b_{yx} = 16.70$, $a_{yx} = -2.261$, $M_x = 1.97$, $\overline{S}_{yx} = 8.28$, $\sigma_x = 0.73$, $M_y = M_{y'} = 30.64$, $\sigma_y = 14.43$.

$$M_{y'} = -2.261 + (16.70)(1.97) = 30.64$$

$$\sigma_{M_{y'}} = \frac{8.28}{\sqrt{14}} = 2.21$$

$$\sigma_{b_{yx}} = \frac{8.28}{0.73\sqrt{14}} = 3.03$$

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]

$$\sigma_{y'} = \sqrt{\sigma^2_{M_{y'}} + (\sigma_{b_{yx}} x)^2} \qquad (70.1)$$

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.

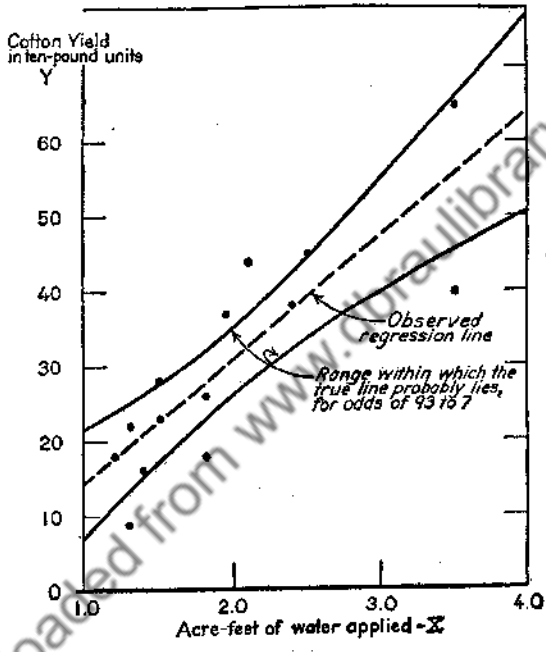

FIG. 59. Linear regression of cotton yield on irrigation water applied, and range within which the true relation probably lies.

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.