Mordecai Ezekiel and
Karl A. Fox

METHODS of CORRELATION and REGRESSION ANALYSIS

Third Edition

Methods of Correlation

Mordecai Ezekiel

Head, Economics Department
Food and Agriculture Organization
of the United Nations

Karl A. Fox

Head, Department of Economics
and Sociology,
Iowa State University

THIRD EDITION

and Regression Analysis

Linear and Curvilinear

New York · John Wiley & Sons, Inc.
London · Chapman & Hall, Ltd.

Library of Congress Catalog Card Number: 59-11793

Printed in the United States of America

Preface to Third Edition

Thirty years have elapsed since the original edition of this book was written—years of political tensions and upheaval and of enormous progress in technical development. This last has been reflected in changes in some of the examples cited—from an automobile with two-wheel brakes in the 1920's to the orbit of an earth satellite in the late 1950's, and from methods for using hand calculators and card tabulators to those for electronic computers. Despite this technical progress, the basic elements of correlation analysis continue unchanged. The major emphasis, however, has shifted from correlation to regression, and the wide range of uses of the method in varied fields has led to many specialized applications or modifications. This is especially true in econometrics. Here the long controversy over mutually intercorrelated variables has finally produced an effective simultaneous-equation method for dealing with situations where single-equation solutions are inadequate; but apparently such situations are relatively infrequent.

In this third edition the senior author has been fortunate in securing the collaboration of an associate who has made distinguished contributions in these newer aspects of the field, particularly in their application to problems of actual research. The new chapter on simultaneous-equation solutions, Chapter 24, is one of his contributions. He is also responsible for the extended treatment of the analysis of variance in relation to regression problems (Chapter 23), the modernization of the chapter on standard errors in multiple regression (Chapter 17), and a complete revision of the treatment of error formulas for time

series (Chapter 20), as well as for many other contributions throughout the text.

In the revision, methods of determining regressions by algebraic equations have been given first consideration, and graphic approximation methods have been treated second, with due consideration of the limitations of each. Modern terminology, especially in the econometric field, has been recognized and used where appropriate, and the presentation of sampling and of confidence intervals has been modernized.

The order of presentation has been rearranged and grouped under seven major sections. We hope this will make the general development clearer both to students and to teachers. Modern methods of calculating correlation and regression constants are outlined in a new chapter, and methods of using electronic calculators for this purpose are briefly treated. The chapter on uses of correlation analysis has been recast and extended to cover newer fields in which the method is now widely used, as well as recent work and developments in fields long familiar. Examples from other countries, and also from the international field in which the senior author has been working this past decade, have been introduced here and elsewhere in the book. The treatment of standard errors and their meaning for statistics derived from small samples has been materially revised. Introduction of adjustments for correlation coefficients and indexes to remove bias due to limited degrees of freedom has been deferred until a late chapter in the book (Chapter 17).

Despite these innovations, the general simplicity of expression and explanation has been retained as far as possible. Mathematical derivations have been relegated to the technical appendix and many of the more obvious ones have been eliminated; the notation has been kept as simple as possible, and only a modest level of mathematical training is assumed for the reader.

As in previous editions, full attention is given to multiple *curvilinear* and *joint functional* regressions, essential for adequate treatment of many problems in both natural and social sciences. Many standard statistical texts still largely ignore the use of non-linear regressions as practical working tools. In recognition of the increased emphasis given to regression analysis in general, and the unusually full treatment of non-linear regression, the title of the book has been changed to *Methods of Correlation and Regression Analysis* with the subtitle *Linear and Curvilinear.*

This book is in part an exposition of standard statistical methods with no attempt to give proofs or to show their mathematical deriva-

tion. In substantial part, however, it is based upon procedures first developed by the senior author or by colleagues associated with him. In such cases chapter references indicate the professional papers in which the methods were first presented and proved and other important papers relating to these methods.

The authors would like to express their deep appreciation to many fellow workers in many lands who have contributed to this revision by supplying suggestions, criticisms, materials, or illustrations, and to the many students who through the years have called attention to errors in the previous printings or editions. Our special thanks are due to John H. Smith for many helpful suggestions on the entire manuscript, to Martha N. Condee for calculating the examples used in Chapters 13 and 24, and to J. P. Cavin and R. J. Foote for making their computing facilities available for this purpose. We hope readers will again call our attention to any new errors of computation or of type setting that may have slipped in in this revision and in the many new examples introduced.

In the first edition the senior author acknowledged his debt "to the spirit of research with which the Bureau of Agricultural Economics was imbued by the broad vision of Henry C. Taylor." The junior author owes a similar debt to the research environment that was maintained in the Division of Statistical and Historical Research of that Bureau under the leadership of O. C. Stine and J. P. Cavin, and to such former colleagues there as R. J. Foote, Harold F. Breimyer, and C. Kyle Randall, who contributed to its high standards in applied research during the current decade. His new colleagues at Iowa State University, including T. A. Bancroft, George Snedecor, and Emil Jebe, have helped him to appreciate the shift in emphasis from correlation to regression (and to the analysis of variance) that has taken place in a number of other sciences in addition to economics. The two authors are, of course, jointly responsible for the particular emphasis adopted in this edition and for such errors and imperfections as may exist in it.

We wish to thank all our helpers in Ames, Rome, and at Wiley & Sons in New York, for their part in making this new edition possible.

MORDECAI EZEKIEL
KARL A. FOX

Rome, Italy
Ames, Iowa
August, 1959

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

SECTION VI

Miscellaneous Special Regression Methods

SECTION VII

Uses and Philosophy of Correlation and Regression Analysis

SECTION I

Introductory Concepts

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. But even in these fields, the modern work on atomic subparticles has involved the use of statistical concepts. In fact, the statistical concept of true value owes its existence to the reproducibility of measurements in fields like these.

In many natural sciences, 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. In such sciences, statistical methods can be used to plan experiments in such a way as to make the conclusions most reliable with a minimum of effort, and they can be used to measure the interrelations in sciences like astronomy, where the phenomena can be observed but not controlled.[1]

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 importance 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

[1] W. G. Cochran and Gertrude M. Cox, *Experimental Designs*, 2nd ed., John Wiley and Sons, New York, 1957.

R. A. Fisher, *The Design of Experiments*, 5th ed., Oliver and Boyd, Edinburgh and London, 1949.

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 special need for analytical methods in the social sciences has been clearly stated by an eminent Englishman, as follows:[2]

> 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 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 concern the organization of quantitative information as a basis for drawing inferences. Some of the basic work involves averaging and classifying data. Thus, if a person were studying the yield of corn in one year in some area, say a county, he might talk with 20 farmers picked at random and obtain figures, such as those in Table 1.1, showing the yield of corn each farmer had obtained.[3]

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

[3] "Picked at random" means so selected that, for each observation, there is just as great a chance of any one farm in the universe (as here the county) being selected as of any other farm. One way of making a random selection would be to put slips with the names of all the farmers in the county into a bowl, mix the slips thoroughly, and then have a blindfolded person draw out slips one at a time, repeating the mixing before each new drawing. Data would then be obtained from the n farmers represented by the slips so drawn. A sample so selected is known as a "random sample" or an "equal probability sample." (See V. G. Panse and P. V. Sukhatme, *Statistical Methods for Agricultural Workers*, pp. 36–40, Indian Council of Agricultural Research, New Delhi, 1954.) 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.

Table 1.1

YIELDS OF CORN OBTAINED BY 20 FARMERS*

Farmer	Yield	Farmer	Yield	Farmer	Yield	Farmer	Yield
	Bushels per acre		*Bushels per acre*		*Bushels per acre*		*Bushels per acre*
1	39	6	43	11	39	16	43
2	35	7	30	12	45	17	41
3	48	8	38	13	36	18	47
4	40	9	40	14	33	19	38
5	37	10	39	15	41	20	42

* In making 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, 43 bushels represents any report of 42.5 bushels or more, and up to but not including 43.5 bushels. If the original reports were secured to the nearest tenth bushel, this might be indicated by writing 42.5–43.4 instead of 43; or if secured to the nearest hundredth bushel, by writing 42.50–43.49.

In performing arithmetic calculations on rounded off data, the results may always have a certain range of inaccuracy due to the effects of rounding. See Appendix 3, Note 7.

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 800 bushels, or an average of 40 bushels.[4] This provides a single figure into which one characteristic of the whole group is condensed.

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

$$M_x = \frac{\Sigma X}{n} \tag{1.1}$$

This formula can be applied to anything we are studying, no matter whether X means bushels of corn, inches in height, degrees of temperature, grade in a school examination, distance of a star, height of a flood, or any other measurable quantity; or whether there are 2 cases or 2 million. This is thus 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 (1.1), 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 that might be of interest. The average would still be 40 if every one of the 20 farmers had had instead a yield of 40 bushels per acre; yet the mean of 20 reports each of 40 bushels would certainly be more reliable than the mean of 20 reports ranging from 33 to 48 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 1.2.

Table 1.2

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

Bushels per acre			
33	38	40	43
35	38	40	43
36	39	41	45
36	39	41	47
37	39	42	48

It is now easier to tell from the series something about the group of reports. One can now see that only 1 farmer had a yield of less than 35 bushels per acre, and only 2 had more than 45, so that 17 out of the 20 had 35 to 45, inclusive. The series shows, too, that 10 of the farmers had less than 40 bushels of corn per acre and 10 had 40 or more, so that the figures 39 and 40 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 37 bushels or less, 5 had yields of 43 bushels or more, whereas 10 men—half of those reporting—had yields of 38 to 42 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.[5]

[5] In statistical terminology, the figure that divides the number of reports into halves—39.5 in this case—is termed the *median*; and the figures that divide the numbers into quarters—37.5 and 42.5—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*.