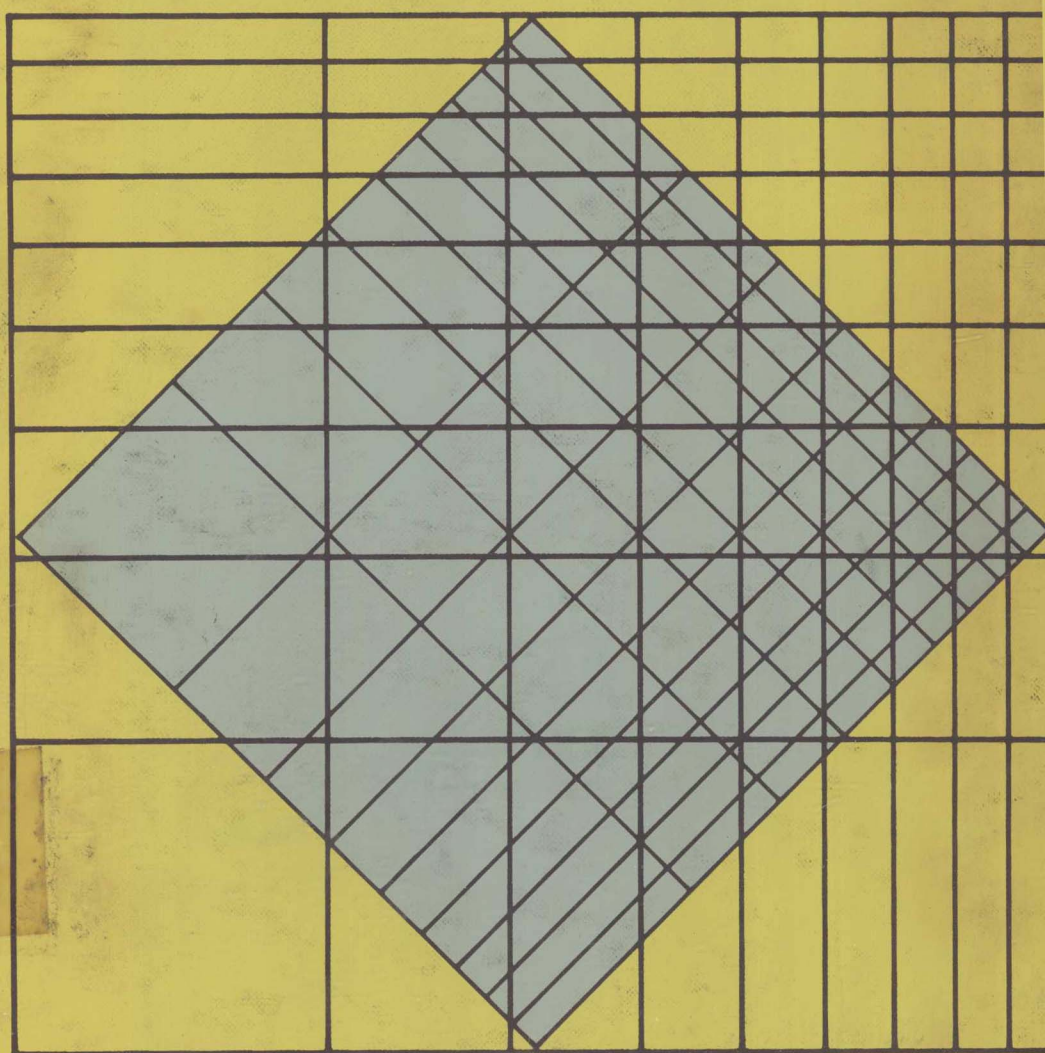
An Introduction to Linear Regression and Correlation

Allen L. Edwards



An Introduction to Linear Regression and Correlation

Allen L. Edwards
UNIVERSITY OF WASHINGTON

W. H. Freeman and Company
San Francisco

Library of Congress Cataloging in Publication Data

Edwards, Allen Louis.
An introduction to linear regression and correlation.

Includes index.
1. Regression analysis. 2. Correlation (Statistics)
I. Title.
QA278.2.E3 519.5'36 75-38811
ISBN 0-7167-0562-1
ISBN 0-7167-0561-3 pbk.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

This book was written for students of one of the behavioral sciences, psychology, but students of the other behavioral sciences may also find it of interest. What I have tried to do in the book is to provide the student with a more detailed and systematic treatment of linear regression and correlation than that ordinarily given in either a first or a second course in applied statistics for students of psychology. Correlational techniques are of importance to the student of individual differences, and regression analysis is important to the general experimentalist. In this book I attempt to show both the similarities and the differences between these two methods of data analysis.

The book has been written at a level that can be understood by any student with a working knowledge of elementary algebra. I have *not* assumed that the reader has already been exposed to a first course in statistics. Many of the topics traditionally covered in the first course are not essential to an understanding of linear regression and correlation, and those that are essential have, I believe, been briefly but adequately covered in this book. Consequently, the book may be used as a text in either a first or a second course in statistics for psychology students. If students are exposed only to a single course in statistics, one in which a more traditional book is used as a text, then this book might be considered supplementary reading to provide a more detailed coverage of the topics of linear regression and correlation.

The book begins at a very elementary level with the equation for a straight line. There is a continuity in the development of each of the successive chapters. The second chapter treats some nonlinear functions that can be transformed into linear functions. Chapter 3 is concerned with values of a dependent variable Y that are subject to random variation. The student is shown how the method of least squares can be used to find

a line of best fit, and the residual variance and the standard error of estimate as measures of the variation of the Y values about the line of best fit are introduced.

Chapter 4 deals with the correlation coefficient as a measure of the degree to which two variables are linearly related. The relationship of the correlation coefficient to the residual variance and standard error of estimate is explained. The coefficients of determination and nondetermination are discussed. Chapter 5 begins with an explanation of how any variable can be transformed into a standardized variable. The remainder of the chapter consists of a review of correlation and regression in terms of standardized variables. Various factors that may be related to the magnitude of the correlation coefficient are discussed in Chapter 6. In Chapter 7 the phi coefficient, the point biserial coefficient, and the rank order coefficient are shown to be merely special cases of the correlation coefficient.

Chapter 8 begins with a discussion of a model for a correlational problem. There is a brief discussion of tests of significance and of the four major distributions—the normal, t, F, and χ^2 distributions—used in making such tests. The treatment is nonmathematical and intuitive, and is at a level that can be understood by the beginning student. The t test of the null hypothesis that the population correlation is zero and Fisher's z_r transformation for the correlation coefficient are then discussed. The standard normal distribution test of the difference between two independent correlation coefficients is illustrated, along with the χ^2 test of the homogeneity of several independent values of the correlation coefficient. Chapter 9 is concerned with tests of significance for the special cases of the correlation coefficient; Chapter 10 deals with tests of significance for regression coefficients.

Coefficients for orthogonal polynomials are introduced in Chapter 11. Examples of correlation and regression of mean Y values with these coefficients are discussed. In Chapter 12 an example is given of the analysis of variance for an experiment involving equally spaced values of an independent variable. Tests of significance for the linear, quadratic, and other components of the treatment sum of squares are illustrated. Chapter 13 includes a description of the analysis of variance for an experiment in which the same subjects are tested with each of the equally spaced values of an independent variable.

The book concludes with a discussion of multiple regression and correlation. For simplicity, the numerical example involves only two X variables, but the basic principles are generalized to the case of more than two X variables.

At the end of each chapter I have provided a number of simple exercises designed to test the reader's understanding of the material covered in that chapter. Answers to all of the exercises that require calculations

are given in the Answers to the Exercises section. Although the calculations are relatively simple, I strongly recommend that the reader buy and use a small electronic calculator to perform them. Excellent minicalculators with a memory can now be purchased for less than $30 and without a memory for less than $20. A minicalculator makes even difficult arithmetic a joy rather than a chore and, in addition, affords accuracy in calculations.

In some exercises I have asked for a proof. When the proof has already been given in the text, it is not repeated in the Answers to the Exercises. When the proof has not been given in the text, it is provided in the Answers to the Exercises.

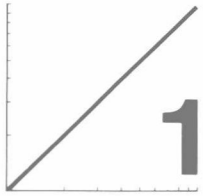
Tables III and V in the Appendix have been reprinted from R. A. Fisher, *Statistical Methods for Research Workers* (14th ed.), Copyright 1972 by Hafner Press, by permission of the publisher. Table IV is reprinted from Enrico T. Federighi, Extended tables of the percentage points of Student's t distribution, *Journal of the American Statistical Association*, 1959, *54*, 683–688, by permission of the American Statistical Association. Table VII has been reprinted from George W. Snedecor and William G. Cochran, *Statistical Methods* (6th ed.), Copyright 1967 by Iowa State University Press, Ames, Iowa, by permission of the publisher. Table VIII has been reprinted from *Essentials of Trigonometry*, by D. E. Smith, W. D. Reeve, and E. L. Morss, Copyright 1928 by W. D. Reeve, and E. L. Morss, Copyright renewed 1956 by W. D. Reeve and E. L. Morss, published by Ginn and Company, by permission of the publishers.

For their careful reading of the manuscript, for completing the exercises at the end of the chapters, and for providing me with their reactions to and evaluations of the material contained in this book, I owe a special debt of gratitude to Clark Ashworth, Mary Cerreto, Randall M. Chestnut, Virginia deWolf, Donald Eismann, Kenneth Johnson, Lynda King, Nana Lowell, Patricia Pedigo, Gary Quarfoth, Francine Rose, Judith Siegel, Frances Thompson, Vicki Wilson, and Thomas Zieske.

Seattle, Washington
August 1975

Allen L. Edwards

Contents



1

Linear Relationships

1.1 Introduction

Many experiments are concerned with the relationship between an independent variable X and a dependent variable Y. The values of the independent variable may represent measures of time, number of trials, varying levels of illumination, varying amounts of practice, varying dosages of a drug, different intensities of shock, different levels of reinforcement, or any other quantitative variable of experimental interest. Ordinarily, the values of the X variable *in an experiment* are selected by the experimenter and are limited in number. They are usually measured precisely and can be assumed to be without error. In general, we shall refer to the values of the X variable in an experiment as fixed in that any conclusions based on the outcome of the experiment will be limited to the particular X values investigated.

For each of the X values, one or more observations of a relevant dependent Y variable are obtained. The objective of the experiment is to determine whether the Y values (or the average Y values, if more than one observation is obtained for each value of X) are related to the X values. In this chapter we shall be concerned with the case where the Y values are linearly related to the X values. By "linearly related" we mean that if the Y values are plotted against the X values, the resulting trend of the plotted points can be represented by a straight line. If the Y values are linearly related to the X values, then we also want to determine the equation for the straight line. We may regard this equation as a rule that relates the Y values to the X values.

1.2 The Equation of a Straight Line

Consider the values of X and Y shown in Table 1.1. What is the rule that relates the values of Y to the values of X? Examination of the pairs of

values will show that for each value of X, the corresponding value of Y is equal to $-.4X$. We may express this rule in the following way:

$$Y = bX \tag{1.1}$$

where $b = -.4$ is a constant that multiplies each value of X. If each value of Y in Table 1.1 were exactly equal to the corresponding value of X, then the value of b would have to be equal to 1.00. If each value of Y were numerically equal to X, but opposite in sign, then the value of b would have to be equal to -1.00.

Now examine the values of X and Y in Table 1.2. The rule or equation relating the Y values to the X values in this case has the general form

$$Y = a + bX \tag{1.2}$$

where b is again a constant that multiplies each value of X and a is a constant that is added to each of the products. For the values of X and Y given in Table 1.2, the value of b is equal to .3 and the value of a is equal to 2. Thus when $X = 10$, we have $Y = 2 + (.3)(10) = 5$. When $X = 8$, we have $Y = 2 + (.3)(8) = 4.4$.

Both (1.1) and (1.2) are equations for a straight line. For example, we could take any arbitrary constants for a and b. Then for any given set of X values we could substitute in (1.2) and obtain a set of Y values. If these values of Y are plotted against the corresponding X values, the set of plotted points will fall on a straight line.

1.3 Graph of $Y = a + bX$

Table 1.3 gives another set of X and Y values. Let us plot the Y values against the corresponding X values. The resulting graph will provide some additional insight into the nature of the constant b that multiplies each value of X as well as the nature of the constant a that is added to the

TABLE 1.1
$Y = -.4X$

X	Y
10	−4.0
9	−3.6
8	−3.2
7	−2.8
6	−2.4
5	−2.0
4	−1.6
3	−1.2
2	− .8
1	− .4

TABLE 1.2
$Y = 2 + .3X$

X	Y
10	5.0
9	4.7
8	4.4
7	4.1
6	3.8
5	3.5
4	3.2
3	2.9
2	2.6
1	2.3

TABLE 1.3
$Y = 3 + .5X$

X	Y
10	8.0
9	7.5
8	7.0
7	6.5
6	6.0
5	5.5
4	5.0
3	4.5
2	4.0
1	3.5

product. In making the graph we set up two axes at right angles to each other. It is customary to let the horizontal axis represent the independent or X variable and the vertical axis represent the dependent or Y variable. We need not begin our scale on the X and Y axes at zero. We may begin with any convenient values that permit us to plot the smallest values of X and Y. In Figure 1.1, for example, we begin the X scale with 0 and the Y scale with 2.0. Nor is it necessary that the X and Y scales be expressed in the same units, as they are in Figure 1.1.

You will recall that a pair of (X,Y) values represents the coordinates of a point. To find the point on the graph corresponding to (10, 8.0), we go out the X axis to 10 and imagine a line perpendicular to the X axis erected at this point. We now go up the Y axis to 8.0 and imagine another line perpendicular to the Y axis erected at this point. The intersection of the two perpendiculars will be the point (10, 8.0) on the graph. It is obviously not necessary to draw the perpendiculars in order to plot a set of points.

1.4 The Slope and Intercept of a Straight Line

It is clear that the points plotted in Figure 1.1 fall along a straight line. The equation of this line, as given by (1.2), is

$$Y = a + bX$$

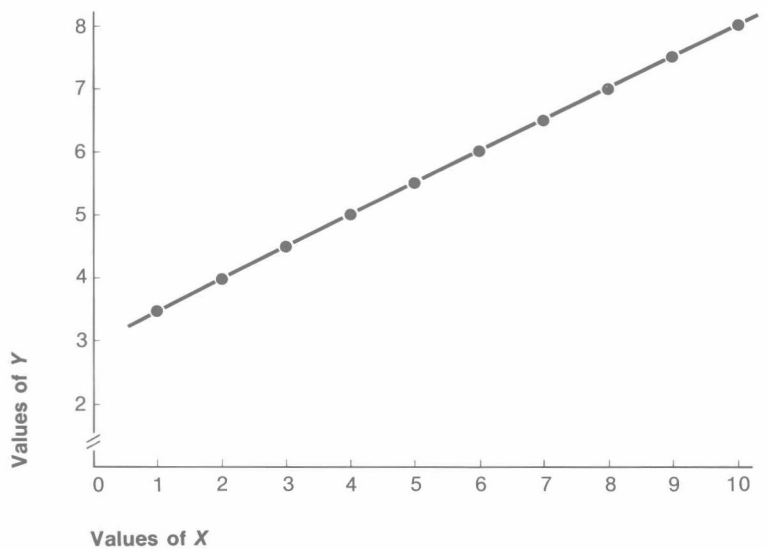


Figure 1.1 Plot of the (X,Y) values given in Table 1.3.

What is the nature of the multiplying constant b? Note, for example, that as we move from 7 to 8 on the X scale, the corresponding increase on the Y scale is from 6.5 to 7.0. An increase of one unit in X, in other words, results in .5 of a unit increase in Y. The constant b is simply the rate at which Y changes with unit change in X.

The value of b can be determined directly from Figure 1.1. For example, if we take any two points on the line with coordinates (X_1,Y_1) and (X_2,Y_2), then

$$b = \frac{Y_2 - Y_1}{X_2 - X_1} \tag{1.3}$$

Substituting in (1.3) the coordinates (2, 4.0) and (3, 4.5), we have

$$b = \frac{4.5 - 4.0}{3 - 2} = .5$$

In geometry (1.3) is known as a particular form of the equation of a straight line, and the value of b is called the slope of the straight line.

The nature of the additive constant a in (1.2) can readily be determined by setting X equal to zero. The value of a must then be the value of Y when X is equal to zero. If the straight line in Figure 1.1 were to be extended downward, we would see that the line would intersect the Y axis at the point $(0,a)$. The number a is called the Y-intercept of the line. In our example, it is easy to see that the value of a is equal to 3. If a straight line passed through the point (0,0), then a would be equal to zero, and the equation of the straight line would be $Y = bX$.

1.5 Positive and Negative Relationships

We may conclude that if the relationship between two variables is linear, then the values of a and b can be determined by plotting the values and finding the Y-intercept and the slope of the line, respectively. A single equation may then be written that will express the nature of the relationship. When the value of b is positive, the relationship is also described as positive; that is, an increase in X is accompanied by an increase in Y and a decrease in X is accompanied by a decrease in Y. When the value of b is negative, the relationship is also described as negative. A negative relationship means that an increase in X is accompanied by a decrease in Y, and a decrease in X is accompanied by an increase in Y. When two variables are positively related, the line representing the relationship will extend from the lower left of the graph to the upper right, and the slope of the line will be positive. When the relationship is negative, the line will extend from the upper left of the graph to the lower right, and the slope of the line will be negative.

Exercises

1.1. Find the values of a and b in the equation $Y = a + bX$ for the following paired (X,Y) values:

X	Y
1	2.2
2	2.6
3	3.0
4	3.4
5	3.8

1.2. Find the values of a and b in the equation $Y = a + bX$ for the following paired (X,Y) values:

X	Y
1	−5.4
2	−5.8
3	−6.2
4	−6.6
5	−7.0

1.3. Find the values of a and b in the equation $Y = a + bX$ for the following paired (X,Y) values:

X	Y
2	10.6
4	11.2
6	11.8
8	12.4
10	13.0

1.4. Find the values of a and b in the equation $Y = a + bX$ for the following paired (X,Y) values:

X	Y
1.0	4.8
1.5	4.7
2.0	4.6
2.5	4.5
3.0	4.4

1.5. Find the values of a and b in the equation $Y = a + bX$ for the following paired (X,Y) values: