CHAPTER 1

Fitting a Straight Line by Least Squares

1.0. INTRODUCTION: THE NEED FOR STATISTICAL ANALYSIS

In today's industrial processes, there is no shortage of "information." No matter how small or how straightforward a process may be, measuring instruments abound. They tell us such things as input temperature, concentration of reactant, percent catalyst, steam temperature, consumption rate, pressure, and so on, depending on the characteristics of the process being studied. Some of these readings are available at regular intervals, every five minutes perhaps or every half hour; others are observed continuously. Still other readings are available with a little extra time and effort. Samples of the end product may be taken at intervals and, after analysis, may provide measurements of such things as purity, percent yield, glossiness, breaking strength, color, or whatever other properties of the end product are important to the manufacturer or user.

In research laboratories, experiments are being performed daily. These are usually small, carefully planned studies and result in sets of data of modest size. The objective is often a quick yet accurate analysis, enabling the experimenter to move on to "better" experimental conditions, which will produce a product with desirable characteristics. Additional data can easily be obtained if needed, however, if the decision is initially unclear.

A Ph.D. researcher may travel into an African jungle for a one-year period of intensive data-gathering on plants or animals. She will return with the raw material for her thesis and will put much effort into analyzing the data she has, searching for the messages that they contain. It will not be easy to obtain more data once her trip is completed, so she must carefully analyze every aspect of what data she has.

Regression analysis is a technique that can be used in any of these situations. Our purpose in this book is to explain in some detail something of the technique of extracting, from data of the types just mentioned, the main features of the relationships hidden or implied in the tabulated figures. (Nevertheless, the study of regression analysis techniques will also provide certain insights into how to plan the collection of data, when the opportunity arises. See, for example, Section 3.3.)

In any system in which variable quantities change, it is of interest to examine the effects that some variables exert (or appear to exert) on others. There may in fact be a simple functional relationship between variables; in most physical processes this is the exception rather than the rule. Often there exists a functional relationship that is too complicated to grasp or to describe in simple terms. In this case we may wish to approximate to this functional relationship by some simple mathematical function, such as a polynomial, which contains the appropriate variables and which graduates

or approximates to the true function over some limited ranges of the variables involved. By examining such a graduating function we may be able to learn more about the underlying true relationship and to appreciate the separate and joint effects produced by changes in certain important variables.

Even where no sensible physical relationship exists between variables, we may wish to relate them by some sort of mathematical equation. While the equation might be physically meaningless, it may nevertheless be extremely valuable for predicting the values of some variables from knowledge of other variables, perhaps under certain stated restrictions.

In this book we shall use one particular method of obtaining a mathematical relationship. This involves the initial assumption that a certain type of relationship, linear in unknown parameters (except in Chapter 24, where nonlinear models are considered), holds. The unknown parameters are estimated under certain other assumptions with the help of available data, and a fitted equation is obtained. The value of the fitted equation can be gauged, and checks can be made on the underlying assumptions to see if any of these assumptions appears to be erroneous. The simplest example of this process involves the construction of a fitted straight line when pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are available. We shall deal with this in a simple algebraic way in Chapters 1–3. To handle problems involving large numbers of variables, however, matrix methods are essential. These are introduced in the context of fitting a straight line in Chapter 4. Matrix algebra allows us to discuss concepts in a larger linear least squares regression context in Chapters 5–16 and 19–23. Some non-least-squares topics are discussed in Chapters 17 (ridge regression), 18 (generalized linear models), 24 (nonlinear estimation), 25 (robust regression), and 26 (resampling procedures).

We assume that anyone who uses this book has had a first course in statistics and understands certain basic ideas. These include the ideas of parameters, estimates, distributions (especially normal), mean and variance of a random variable, covariance between two variables, and simple hypothesis testing involving one- and two-sided t-tests and the F-test. We believe, however, that a reader whose knowledge of these topics is rusty or incomplete will nevertheless be able to make good progress after a review of Chapter 0.

We do not intend this as a comprehensive textbook on all aspects of regression analysis. Our intention is to provide a sound basic course plus material necessary to the solution of some practical regression problems. We also add some excursions into related topics.

We now take an early opportunity to introduce the reader to the data in Appendix 1A. Here we see 25 observations taken at intervals from a steam plant at a large industrial concern. Ten variables, some of them in coded form, were recorded as follows:

1. Pounds of steam used monthly, in coded form.
2. Pounds of real fatty acid in storage per month.
3. Pounds of crude glycerin made.
4. Average wind velocity (in mph).
5. Calendar days per month.
6. Operating days per month.
7. Days below 32°F.
8. Average atmospheric temperature (°F).

9. Average wind velocity squared.
10. Number of start-ups.

We can distinguish two main types of variable at this stage. We shall usually call these *predictor variables* and *response variables*. (For alternative terms, see below.) By predictor variables we shall usually mean variables that can either be set to a desired value (e.g., input temperature or catalyst feed rate) or else take values that can be observed but not controlled (e.g., the outdoor humidity). As a result of changes that are deliberately made, or simply take place in the predictor variables, an effect is transmitted to other variables, the response variables (e.g., the final color or the purity of a chemical product). In general, we shall be interested in finding out how changes in the predictor variables affect the values of the response variables. If we can discover a simple relationship or dependence of a response variable on just one or a few predictor variables we shall, of course, be pleased. The distinction between predictor and response variables is not always completely clear-cut and depends sometimes on our objectives. What may be considered a response variable at the midstage of a process may also be regarded as a predictor variable in relation to (say) the final color of the product. In practice, however, the roles of variables are usually easily distinguished.

Other names frequently seen are the following:

$$\begin{aligned}\text{Predictor variables} &= \text{input variables} = \text{inputs}\\ &= X\text{-variables} = \text{regressors}\\ &= \text{independent variables.}\end{aligned}$$

(We shall try to avoid using the last of these names, because it is often misleading. In a particular body of data, two or more "independent" variables may vary together in some definite way due, perhaps, to the method in which an experiment is conducted. This is not usually desirable—for one thing it restricts the information on the separate roles of the variables—but it may often be unavoidable.)

$$\begin{aligned}\text{Response variables} &= \text{output variables} = \text{outputs}\\ &= Y\text{-variables}\\ &= \text{dependent variables.}\end{aligned}$$

Returning to the data in Appendix 1A, which we shall refer to as the *steam data*, we examine the 25 sets of observations on the variables, one set for each of 25 different months. Our primary interest here is in the monthly amount of steam produced and how it changes due to variations in the other variables. Thus we shall regard variable X_1 as a dependent or response variable, Y, in what follows, and the others as predictor variables, $X_2, X_3, \ldots, X_{10}$.

We shall see how the method of analysis called the *method of least squares* can be used to examine data and to draw meaningful conclusions about dependency relationships that may exist. This method of analysis is often called *regression analysis*. (For historical remarks, see Section 1.8.)

Throughout this book we shall be most often concerned with relationships of the form

$$\text{Response variable} = \text{Model function} + \text{Random error.}$$

The model function will usually be "known" and of specified form and will involve

the predictor variables as well as *parameters* to be estimated from data. The distribution of the random errors is often assumed to be a normal distribution with mean zero, and errors are usually assumed to be independent. All assumptions are usually checked after the model has been fitted and many of these checks will be described.

(*Note:* Many engineers and others call the parameters *constants* and the predictors *parameters*. Watch out for this possible difficulty in cross-discipline conversations!)

We shall present the least squares method in the context of the simplest application, fitting the "best" straight line to given data in order to relate two variables X and Y, and will discuss how it can be extended to cases where more variables are involved.

1.1. STRAIGHT LINE RELATIONSHIP BETWEEN TWO VARIABLES

In much experimental work we wish to investigate how the changes in one variable affect another variable. Sometimes two variables are linked by an exact straight line relationship. For example, if the resistance R of a simple circuit is kept constant, the current I varies directly with the voltage V applied, for, by Ohm's law, $I = V/R$. If we were not aware of Ohm's law, we might obtain this relationship empirically by making changes in V and observing I, while keeping R fixed and then observing that the plot of I against V more or less gave a straight line through the origin. We say "more or less" because, although the relationship actually is exact, our measurements may be subject to slight errors and thus the plotted points would probably not fall exactly on the line but would vary randomly about it. For purposes of predicting I for a particular V (with R fixed), however, we should use the straight line through the origin. Sometimes a straight line relationship is not exact (even apart from error) yet can be meaningful nevertheless. For example, suppose we consider the height and weight of adult males for some given population. If we plot the pair (Y_1, Y_2) = (height, weight), a diagram something like Figure 1.1 will result. (Such a presentation is conventionally called a *scatter diagram.*)

Note that for any given height there is a range of observed weights, and vice versa. This variation will be partially due to measurement errors but primarily due to variation between individuals. Thus no unique relationship between actual height and weight

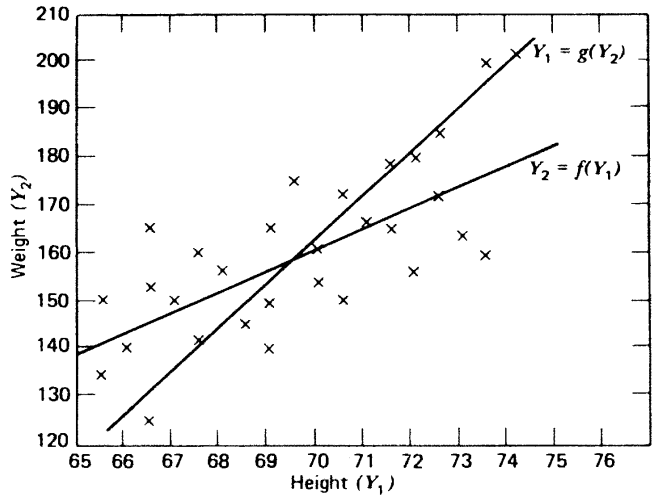

Figure 1.1. Heights and weights of 30 American males.

can be expected. But we can note that the average observed weight for a given observed height increases as height increases. This locus of average observed weight for given observed height (as height varies) is called the *regression curve* of weight on height. Let us denote it by $Y_2 = f(Y_1)$. There also exists a regression curve of height on weight, similarly defined, which we can denote by $Y_1 = g(Y_2)$. Let us assume that these two "curves" are both straight lines (which in general they may not be). In general, these two curves are *not* the same, as indicated by the two lines in the figure.

Suppose we now found we had recorded an individual's height but not his weight and we wished to estimate this weight. What could we do? From the regression line of weight on height we could find an average observed weight of individuals of the given height and use this as an estimate of the weight that we did not record.

A pair of random variables such as (height, weight) follows some sort of bivariate probability distribution. When we are concerned with the dependence of a random variable Y on a quantity X that is variable but *not* randomly variable, an equation that relates Y to X is usually called a *regression equation*. Although the name is, strictly speaking, incorrect, it is well established and conventional. In nearly all of this book we assume that the predictor variables are *not* subject to random variation, but that the response variable is. From a practical point of view, this is seldom fully true but, if it is not, a much more complicated fitting procedure is needed. (See Sections 3.4 and 9.7.) To avoid this, we use the least squares procedure only in situations where we can assume that any random variation in any of the predictor variables is so small compared with the *range* of that predictor variable observed that we can effectively ignore the random variation. This assumption is rarely stated, but it is implicit in all least squares work in which the predictors are assumed "fixed." (The word "fixed" means "not random variables" in such a context; it does not mean that the predictors cannot take a variety of values or levels.) For additional comments see Section 3.4.

We can see that whether a relationship is exactly a straight line or a line only insofar as mean values are concerned, knowledge of the relationship will be useful. (The relationship might, of course, be more complicated than a straight line but we shall consider this later.)

A straight line relationship may also be a valuable one even when we *know* that such a relationship cannot be true. Consider the response relationship shown in Figure 1.2. It is obviously not a straight line over the range $0 \leq X \leq 100$. However, if we were interested primarily in the range $0 \leq X \leq 45$, a straight line relationship evaluated from observations in this range might provide a perfectly adequate representation of the function *in this range*. The relationship thus fitted would, of course, not apply to values of X outside this restricted range and could not be used for predictive purposes outside this range.

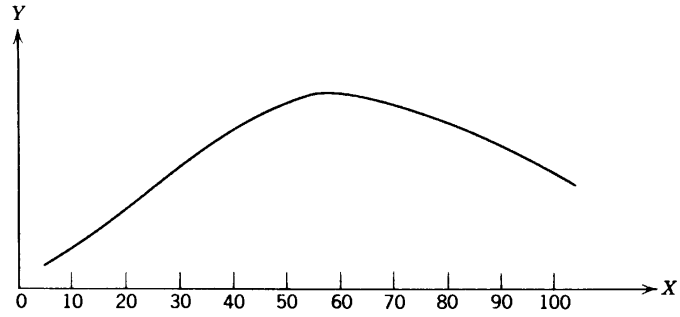

Figure 1.2. A response relationship.

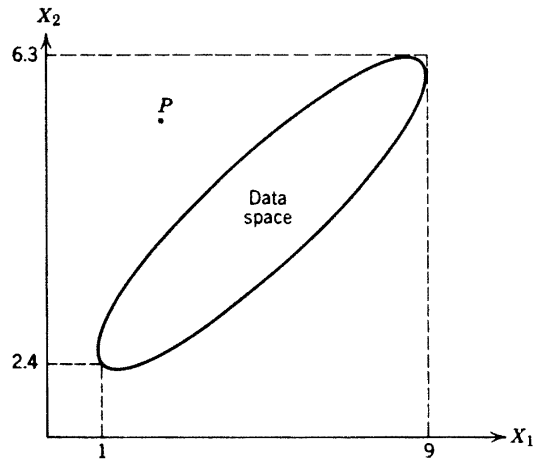

Figure 1.3. A point P outside the data space, whose coordinates nevertheless lie within the ranges of the predictor variables observed.

Similar remarks can be made when more than one predictor variable is involved. Suppose we wish to examine the way in which a response Y depends on variables X_1, X_2, ..., X_k. We determine a regression equation from data that "cover" certain regions of the "X-space." Suppose the point $\mathbf{X}_0' = (X_{10}, X_{20}, \ldots, X_{k0})$ lies *outside* the regions covered by the original data. While we can mathematically obtain a predicted value $\hat{Y}$ $(\mathbf{X}_0')$ for the response at the point $\mathbf{X}_0'$, we must realize that reliance on such a prediction is extremely dangerous and becomes more dangerous the further $\mathbf{X}_0'$ lies from the original regions, unless some additional knowledge is available that the regression equation is valid in a wider region of the X-space. Note that it is sometimes difficult to realize at first that a suggested point lies outside a region in a multidimensional space. To take a simple example, consider the region indicated by the ellipse in Figure 1.3, within which all the data points (X_1, X_2) lie; the corresponding Y values, plotted vertically up from the page, are not shown. We see that there are points in the region for which $1 \leq X_1 \leq 9$ and for which $2.4 \leq X_2 \leq 6.3$. Although the X_1 and X_2 coordinates of P lie individually within these ranges, P itself lies outside the region. A simple review of the printed data would often not detect this. When more dimensions are involved, misunderstandings of this sort easily arise.

1.2. LINEAR REGRESSION: FITTING A STRAIGHT LINE BY LEAST SQUARES

We have mentioned that in many situations a straight line relationship can be valuable in summarizing the observed dependence of one variable on another. We now show how the equation of such a straight line can be obtained by the method of least squares when data are available. Consider, in Appendix 1A, the 25 observations of variable 1 (pounds of steam used per month) and variable 8 (average atmospheric temperature in degrees Fahrenheit). The corresponding pairs of observations are given in Table 1.1 and are plotted in Figure 1.4.

Let us tentatively assume that the regression line of variable 1, which we shall denote by Y, on variable 8 (X) has the form $\beta_0 + \beta_1 X$. Then we can write the linear, first-order model

$$Y = \beta_0 + \beta_1 X + \epsilon; \tag{1.2.1}$$

TABLE 1.1. Twenty-five Observations of Variables 1 and 8

Observation Number	Variable Number	
	1 (Y)	8 (X)
1	10.98	35.3
2	11.13	29.7
3	12.51	30.8
4	8.40	58.8
5	9.27	61.4
6	8.73	71.3
7	6.36	74.4
8	8.50	76.7
9	7.82	70.7
10	9.14	57.5
11	8.24	46.4
12	12.19	28.9
13	11.88	28.1
14	9.57	39.1
15	10.94	46.8
16	9.58	48.5
17	10.09	59.3
18	8.11	70.0
19	6.83	70.0
20	8.88	74.5
21	7.68	72.1
22	8.47	58.1
23	8.86	44.6
24	10.36	33.4
25	11.08	28.6

that is, for a given X, a corresponding observation Y consists of the value $\beta_0 + \beta_1 X$ plus an amount ϵ, the increment by which any individual Y may fall off the regression line. Equation (1.2.1) is the *model* of what we believe. $\beta_0 + \beta_1 X$ is the *model function* here and β_0 and β_1 are called the *parameters* of the model. We begin by assuming that the model holds; but we shall have to inquire at a later stage if indeed it does. In many aspects of statistics it is necessary to assume a mathematical model to make progress. It might be well to emphasize that what we are usually doing is to *consider* or *tentatively entertain* our model. The model must always be critically examined somewhere along the line. It is our "opinion" of the situation at one stage of the investigation and our "opinion" must be changed if we find, at a later stage, that the facts are against it.

Meaning of Linear Model

When we say that a model is linear or nonlinear, we are referring to linearity or nonlinearity *in the parameters*. The value of the highest power of a predictor variable in the model is called the *order* of the model. For example,

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$$

is a second-order (in X) linear (in the β's) regression model. Unless a model is specifically called nonlinear it can be taken that it is linear in the parameters, and the

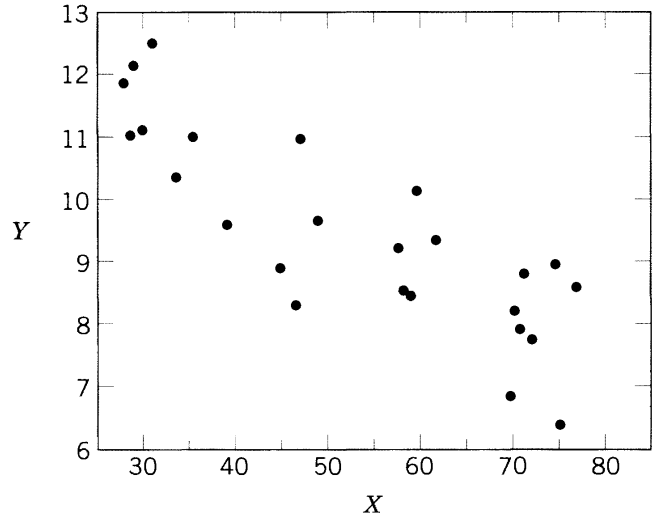

Figure 1.4. Plot of the steam data for variables 1 (Y) and 8 (X).

word linear is usually omitted and understood. The order of the model could be of any size. Notation of the form β_{11} is often used in polynomial models; β_1 is the parameter that goes with X while β_{11} is the parameter that goes with $X^2 = XX$. The natural extension of this sort of notation appears, for example, in Chapter 12, where β_{12} is the parameter associated with X_1X_2 and so on.

Least Squares Estimation

Now β_0, β_1, and ϵ are unknown in Eq. (1.2.1), and in fact ϵ would be difficult to discover since it changes for each observation Y. However, β_0 and β_1 remain fixed and, although we cannot find them exactly without examining all possible occurrences of Y and X, we can use the information provided by the 25 observations in Table 1.1 to give us *estimates* b_0 and b_1 of β_0 and β_1; thus we can write

$$\hat{Y} = b_0 + b_1X, \tag{1.2.2}$$

where $\hat{Y}$, read "Y hat," denotes the *predicted* value of Y for a given X, when b_0 and b_1 are determined. Equation (1.2.2) could then be used as a predictive equation; substitution for a value of X would provide a prediction of the true mean value of Y for that X.

The use of small roman letters b_0 and b_1 to denote estimates of the parameters given by Greek letters β_0 and β_1 is standard. However, the notation $\hat{\beta}_0$ and $\hat{\beta}_1$ for the estimates is also frequently seen. We use the latter type of notation ourselves in Chapter 24, for example.

Our estimation procedure will be that of least squares.

Under certain assumptions to be discussed in Chapter 5, the method of least squares has certain properties. For the moment we state it as our chosen method of estimating the parameters without a specific justification. Suppose we have available n sets of observations (X_1, Y_1), (X_2, Y_2), . . . , (X_n, Y_n). (In our steam data example $n = 25$.) Then by Eq. (1.2.1) we can write

$$Y_i = \beta_0 + \beta_1X_i + \epsilon_i, \tag{1.2.3}$$

for $i = 1, 2, \ldots, n$, so that the sum of squares of deviation from the true line is

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \tag{1.2.4}$$

S is also called the *sum of squares function*. We shall choose our estimates b_0 and b_1 to be the values that, when substituted for β_0 and β_1 in Eq. (1.2.4), produce the least possible value of S; see Figure 1.5. [Note that, in (1.2.4), X_i, Y_i are the fixed numbers that we have observed.] We can determine b_0 and b_1 by differentiating Eq. (1.2.4) first with respect to β_0 and then with respect to β_1 and setting the results equal to zero. Now

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i), \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i), \end{aligned} \tag{1.2.5}$$

so that the estimates b_0 and b_1 are solutions of the two equations

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) &= 0, \\ \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) &= 0, \end{aligned} \tag{1.2.6}$$

where we substitute (b_0, b_1) for (β_0, β_1), when we equate Eq. (1.2.5) to zero. From Eq. (1.2.6) we have

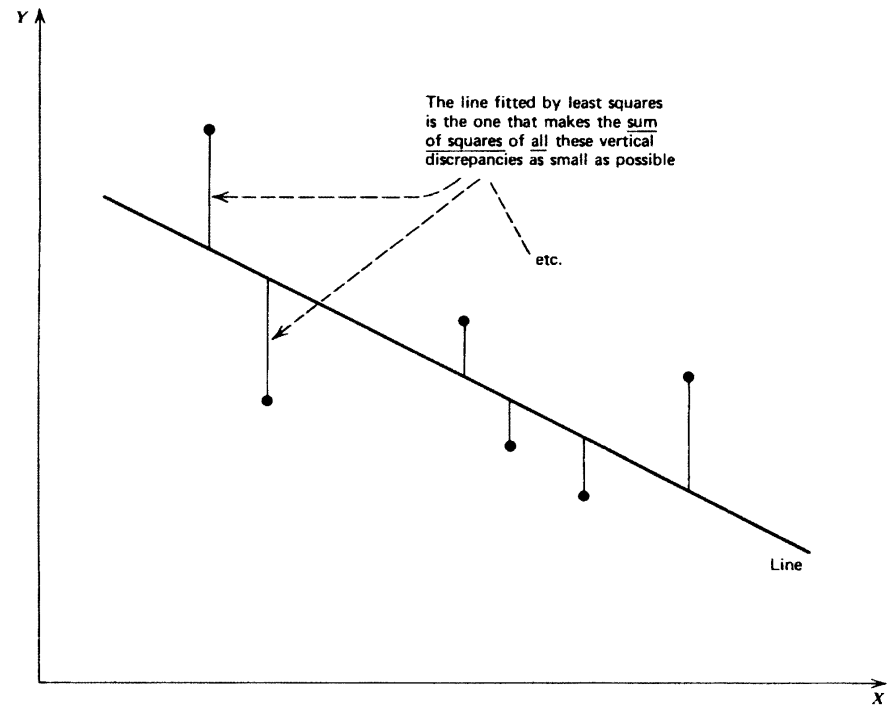

Figure 1.5. The vertical deviations whose sum of squares is minimized for the least squares procedure.

$$\begin{aligned} \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i &= 0 \\ \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 &= 0 \end{aligned} \tag{1.2.7}$$

or

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i. \end{aligned} \tag{1.2.8}$$

These equations are called the *normal equations*. (Normal here means perpendicular, or orthogonal, a geometrical property explained in Chapter 20. The normal equations can also be obtained via a geometrical argument.)

The solution of Eq. (1.2.8) for b_1, the slope of the fitted straight line, is

$$b_1 = \frac{\Sigma X_i Y_i - [(\Sigma X_i)(\Sigma Y_i)]/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}, \tag{1.2.9}$$

where all summations are from $i = 1$ to n and the two expressions for b_1 are just slightly different forms of the same quantity. For, defining

$$\begin{aligned} \overline{X} &= (X_1 + X_2 + \cdots + X_n)/n = \Sigma X_i/n, \\ \overline{Y} &= (Y_1 + Y_2 + \cdots + Y_n)/n = \Sigma Y_i/n, \end{aligned}$$

we have that

$$\begin{aligned} \Sigma (X_i - \overline{X})(Y_i - \overline{Y}) &= \Sigma X_i Y_i - \overline{X} \Sigma Y_i - \overline{Y} \Sigma X_i + n \overline{X}\, \overline{Y} \\ &= \Sigma X_i Y_i - n \overline{X}\, \overline{Y} \\ &= \Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n. \end{aligned}$$

This shows the equivalence of the numerators in (1.2.9), and a parallel calculation, in which Y is replaced by X, shows the equivalence of the denominators. The quantity ΣX_i^2 is called the *uncorrected sum of squares of the X's* and $(\Sigma X_i)^2/n$ is the *correction for the mean of the X's*. The difference is called the *corrected sum of squares of the X's*. Similarly, $\Sigma X_i Y_i$ is called the *uncorrected sum of products*, and $(\Sigma X_i)(\Sigma Y_i)/n$ is the *correction for the means*. The difference is called the *corrected sum of products of X and Y*.

Pocket-Calculator Form

The first form in Eq. (1.2.9) is normally used for pocket-calculator evaluation of b_1, because it is easier to work with and does not involve the tedious adjustment of each X_i and Y_i to $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$, respectively. To avoid rounding error, however, it is best to carry as many significant figures as possible in this computation. (Such advice is good in general; rounding is best done at the "reporting stage" of a calculation, not at intermediate stages.) Most digital computers obtain more accurate answers using the second form in Eq. (1.2.9); this is because of their round-off characteristics and the form in which most regression programs are written.

A convenient notation, now and later, is to write

$$\begin{aligned} S_{XY} &= \Sigma(X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \Sigma(X_i - \overline{X})Y_i \\ &= \Sigma X_i(Y_i - \overline{Y}) \\ &= \Sigma X_iY_i - (\Sigma X_i)(\Sigma Y_i)/n \\ &= \Sigma X_iY_i - n\overline{X}\,\overline{Y}. \end{aligned}$$

Note that all these forms are equivalent. Similarly, we can write

$$\begin{aligned} S_{XX} &= \Sigma(X_i - \overline{X})^2 \\ &= \Sigma(X_i - \overline{X})X_i \\ &= \Sigma X_i^2 - (\Sigma X_i)^2/n \\ &= \Sigma X_i^2 - n\overline{X}^2 \end{aligned}$$

and

$$\begin{aligned} S_{YY} &= \Sigma(Y_i - \overline{Y})^2 \\ &= \Sigma(Y_i - \overline{Y})Y_i \\ &= \Sigma Y_i^2 - (\Sigma Y_i)^2/n \\ &= \Sigma Y_i^2 - n\overline{Y}^2. \end{aligned}$$

The easily remembered formula for b_1 is then

$$b_1 = S_{XY}/S_{XX}. \tag{1.2.9a}$$

The solution of Eqs. (1.2.8) for b_0, the intercept at $X = 0$ of the fitted straight line, is

$$b_0 = \overline{Y} - b_1\overline{X}. \tag{1.2.10}$$

The predicted or fitted equation is $\hat{Y} = b_0 + b_1X$ as in (1.2.2), we recall. Substituting Eq. (1.2.10) into Eq. (1.2.2) gives the estimated regression equation in the alternative form

$$\hat{Y} = \overline{Y} + b_1(X - \overline{X}), \tag{1.2.11}$$

where b_1 is given by Eq. (1.2.9). From this we see immediately that if we set $X = \overline{X}$ in (1.2.11), then $\hat{Y} = \overline{Y}$. This means that the point $(\overline{X}, \overline{Y})$ lies on the fitted line. In other words, this least squares line contains the center of gravity of the data.

Calculations for the Steam Data

Let us now perform these calculations on the selected steam data given in Table 1.1. We find the following:

$$\begin{aligned} n &= 25, \\ \Sigma Y_i &= 10.98 + 11.13 + \cdots + 11.08 = 235.60, \\ \overline{Y} &= 235.60/25 = 9.424, \\ \Sigma X_i &= 35.3 + 29.7 + \cdots + 28.6 = 1315, \\ \overline{X} &= 1315/25 = 52.60, \end{aligned}$$

$$\Sigma X_i Y_i = (10.98)(35.3) + (11.13)(29.7) + \cdots + (11.08)(28.6)$$
$$= 11821.4320,$$
$$\Sigma X_i^2 = (35.3)^2 + (29.7)^2 + \cdots + (28.6)^2 = 76323.42,$$
$$b_1 = \frac{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{S_{XY}}{S_{XX}}$$
$$= \frac{11821.4320 - (1315)(235.60)/25}{76323.42 - (1315)^2/25} = \frac{-571.1280}{7154.42}$$
$$= -0.079829.$$

The fitted equation is thus

$$\hat{Y} = \overline{Y} + b_1(X - \overline{X})$$
$$= 9.4240 - 0.079829(X - 52.60)$$
$$= 13.623005 - 0.079829X.$$

The foregoing form of $\hat{Y}$ shows that $b_0 = 13.623005$. The fitted regression line is plotted in Figure 1.6. We can tabulate for each of the 25 values X_i, at which a Y_i observation is available, the fitted value $\hat{Y}_i$ and the *residual* $Y_i - \hat{Y}_i$ as in Table 1.2. The residuals are given to the same number of places as the original data. They are our "estimates of the errors ϵ_i" and we write $e_i = Y_i - \hat{Y}_i$ in a parallel notation.

Note that since $\hat{Y}_i = \overline{Y} + b_1(X_i - \overline{X})$,

$$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - b_1(X_i - \overline{X}),$$

which we can sum to give

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (Y_i - \overline{Y}) - b_1 \sum_{i=1}^{n} (X_i - \overline{X}) = 0.$$

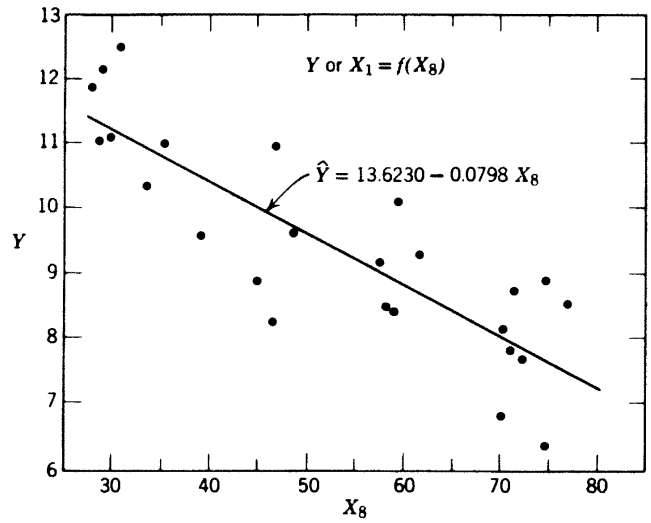

Figure 1.6. Plot of the steam data—variables 1 (Y) and 8 (X)—and the least squares line.

TABLE 1.2. Observations, Fitted Values, and Residuals

Observation Number	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	10.98	10.81	0.17
2	11.13	11.25	−0.12
3	12.51	11.17	1.34
4	8.40	8.93	−0.53
5	9.27	8.72	0.55
6	8.73	7.93	0.80
7	6.36	7.68	−1.32
8	8.50	7.50	1.00
9	7.82	7.98	−0.16
10	9.14	9.03	0.11
11	8.24	9.92	−1.68
12	12.19	11.32	0.87
13	11.88	11.38	0.50
14	9.57	10.50	−0.93
15	10.94	9.89	1.05
16	9.58	9.75	−0.17
17	10.09	8.89	1.20
18	8.11	8.03	0.08
19	6.83	8.03	−1.20
20	8.88	7.68	1.20
21	7.68	7.87	−0.19
22	8.47	8.98	−0.51
23	8.86	10.06	−1.20
24	10.36	10.96	−0.60
25	11.08	11.34	−0.26

This piece of algebra tells us that the residuals sum to zero, in theory. In practice, the sum may not be exactly zero due to rounding. The sum of residuals in any regression problem is always zero when there is a β_0 term in the model as a consequence of the first normal equation. The omission of β_0 from a model implies that the response is zero when all the predictor variables are zero. This is a very strong assumption, which is usually unjustified. In a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$, omission of β_0 implies that the line passes through $X = 0$, $Y = 0$; that is, the line has a zero *intercept* $\beta_0 = 0$ at $X = 0$.

Centering the Data

We note here, before the more general discussion in Section 16.2, that physical removal of β_0 from the model is always possible by "centering" the data, but this is quite different from setting $\beta_0 = 0$. For example, if we write Eq. (1.2.1) in the form

$$Y - \overline{Y} = (\beta_0 + \beta_1 \overline{X} - \overline{Y}) + \beta_1 (X - \overline{X}) + \epsilon$$

or

$$y = \beta_0' + \beta_1 x + \epsilon,$$

say, where $y = Y - \overline{Y}, \beta_0' = \beta_0 + \beta_1 \overline{X} - \overline{Y}$, and $x = X - \overline{X}$, then the least squares estimates of β_0' and β_1 are given as follows:

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2},$$

identical to Eq. (1.2.9), while

$$b_0' = \bar{y} - b_1\bar{x} = 0, \qquad \text{since } \bar{x} = \bar{y} = 0,$$

whatever the value of b_1. Because this always happens, we can write and fit the centered model as

$$Y - \overline{Y} = \beta_1(X - \overline{X}) + \epsilon,$$

omitting the β_0' (intercept) term entirely. We have lost one parameter but there is a corresponding loss in the data since the quantities $Y_i - \overline{Y}$, $i = 1, 2, \ldots, n$, represent only $(n - 1)$ separate pieces of information due to the fact that their sum is zero, whereas $Y_1, Y_2, \ldots, Y_n$ represent n separate pieces of information. Effectively the "lost" piece of information has been used to enable the proper adjustments to be made to the model so that the intercept term can be removed. The model fit is exactly the same as before but is written in a slightly different manner, pivoted around $(\overline{X}, \overline{Y})$.

1.3. THE ANALYSIS OF VARIANCE

We now tackle the question of how much of the variation in the data has been explained by the regression line. Consider the following identity:

$$Y_i - \hat{Y}_i = Y_i - \overline{Y} - (\hat{Y}_i - \overline{Y}). \tag{1.3.1}$$

What this means geometrically for the fitted straight line is illustrated in Figure 1.7. The residual $e_i = Y_i - \hat{Y}_i$ is the difference between two quantities: (1) the deviation of the observed Y_i from the overall mean $\overline{Y}$ and (2) the deviation of the fitted $\hat{Y}_i$ from the overall mean $\overline{Y}$. Note that the average of the $\hat{Y}_i$, namely,

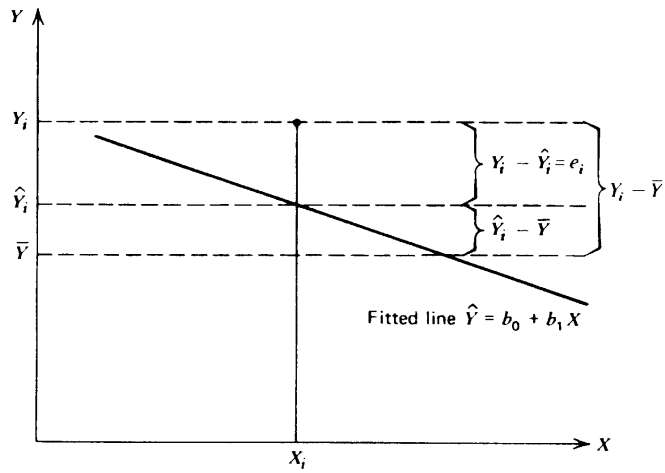

Figure 1.7. Geometrical meaning of the identity (1.3.1).

$$\begin{aligned}\Sigma \hat{Y}_i/n &= \Sigma(b_0 + b_1 X_i)/n \\ &= (nb_0 + b_1 n\overline{X})/n \\ &= b_0 + b_1\overline{X} \\ &= \overline{Y},\end{aligned} \tag{1.3.2}$$

is the same as the average of the Y_i. This fact also reconfirms that $\Sigma\, e_i = \Sigma(Y_i - \hat{Y}_i) = n\overline{Y} - n\overline{Y} = 0$, as stated previously.

We can also rewrite Eq. (1.3.1) as

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i).$$

If we square both sides of this and sum from $i = 1, 2, \ldots, n$, we obtain

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2. \tag{1.3.3}$$

The cross-product term, CPT $= 2\,\Sigma(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i)$, can be shown to vanish by applying Eq. (1.2.11) with subscript i, so that

$$\begin{aligned}\hat{Y}_i - \overline{Y} &= b_1(X_i - \overline{X}) \\ Y_i - \hat{Y}_i &= Y_i - \overline{Y} - b_1(X_i - \overline{X}).\end{aligned} \tag{1.3.4}$$

It follows that the cross-product term is

$$\begin{aligned}\text{CPT} &= 2\,\Sigma\, b_1(X_i - \overline{X})\{(Y_i - \overline{Y}) - b_1(X_i - \overline{X})\} \\ &= 2b_1\{S_{XY} - b_1 S_{XX}\} \\ &= 0\end{aligned} \tag{1.3.5}$$

by Eq. (1.2.9a). It is also clear that

$$\begin{aligned}\Sigma(\hat{Y}_i - \overline{Y})^2 &= b_1^2 S_{XX} \\ &= b_1 S_{XY} \\ &= S_{XY}^2/S_{XX}.\end{aligned} \tag{1.3.6}$$

This provides a simple pocket calculator way of computing this quantity when calculating b_1.

Sums of Squares

We now return to a discussion of Eq. (1.3.3). The quantity $(Y_i - \overline{Y})$ is the deviation of the ith observation from the overall mean and so the left-hand side of Eq. (1.3.3) is the sum of squares of deviations of the observations from the mean; this is shortened to *SS about the mean* and is also the *corrected sum of squares of the Y's*, namely, S_{YY}. Since $\hat{Y}_i - \overline{Y}$ is the deviation of the predicted value of the ith observation from the mean, and $Y_i - \hat{Y}_i$ is the deviation of the ith observation from its predicted or fitted value (i.e., the ith *residual*), we can express Eq. (1.3.3) in words as follows:

$$\begin{pmatrix}\text{Sum of squares} \\ \text{about the mean}\end{pmatrix} = \begin{pmatrix}\text{Sum of squares} \\ \text{due to regression}\end{pmatrix} + \begin{pmatrix}\text{Sum of squares} \\ \text{about regression}\end{pmatrix}. \tag{1.3.7}$$

This shows that, of the variation in the Y's about their mean, some of the variation can be ascribed to the regression line and some, $\Sigma(Y_i - \hat{Y}_i)^2$, to the fact that the actual observations do not all lie on the regression line: if they all did, the sum of squares

TABLE 1.3. Analysis of Variance (ANOVA) Table: The Basic Split of S_{YY}

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)
Due to regression	1	$\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	MS_{Reg}
About regression (residual)	$n - 2$	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$s^2 = \frac{SS}{(n-2)}$ [a]
Total, corrected for mean $\overline{Y}$	$n - 1$	$\sum_{i=1}^{n} (Y_i - \overline{Y})^2$	

[a] Some older regression programs have documentation that labels the quantity $\Sigma(Y_i - \overline{Y})^2/(n-1) = S_{YY}/(n-1)$ as s^2. For us, this would be true *only* if the model fitted were $Y = \beta + \epsilon$. In this case, the regression sum of squares due to b_0 would be (as it is in general—e.g., Table 1.4) $n\overline{Y}^2 = (\Sigma Y_i)^2/n$ and S_{YY} would be the appropriate residual sum of squares for the corresponding fitted model $\hat{Y} = \overline{Y}$.

about the regression (soon to be called the *residual sum of squares*) would be zero! From this procedure we can see that a sensible first way of assessing how useful the regression line will be as a predictor is to see how much of the SS about the mean has fallen into the SS due to regression and how much into the SS about the regression. We shall be pleased if the SS due to regression is much greater than the SS about regression, or what amounts to the same thing, if the ratio R^2 = (SS due to regression)/(SS about mean) is not too far from unity.

Degrees of Freedom (df)

Any sum of squares has associated with it a number called its degrees of freedom. This number indicates how many independent pieces of information involving the n independent numbers $Y_1, Y_2, \ldots, Y_n$ are needed to compile the sum of squares. For example, the SS about the mean needs $(n - 1)$ independent pieces [of the numbers $Y_1 - \overline{Y}, Y_2 - \overline{Y}, \ldots, Y_n - \overline{Y}$, only $(n - 1)$ are independent since all n numbers sum to zero by definition of the mean]. We can compute the SS due to regression from a single function of $Y_1, Y_2, \ldots, Y_n$, namely, b_1 [since $\Sigma(\hat{Y}_i - \overline{Y})^2 = b_1^2 \Sigma(X_i - \overline{X})^2$ as in Eq. (1.3.6)], and so this sum of squares has one degree of freedom. By subtraction, the SS about regression, which we shall in future call the residual sum of squares (it is, in fact, the sum of squares of the residuals $Y_i - \hat{Y}_i$) has $(n - 2)$ degrees of freedom (df). This reflects the fact that the present residuals are from a fitted straight line model, which required estimation of *two* parameters. In general, the residual sum of squares is based on (number of observations − number of parameters estimated) degrees of freedom. Thus corresponding to Eq. (1.3.3), we can show the split of degrees of freedom as

$$n - 1 = 1 + (n - 2). \tag{1.3.8}$$

Analysis of Variance Table

From Eqs. (1.3.3) and (1.3.8) we can construct an *analysis of variance* table in the form of Table 1.3. The "Mean Square" column is obtained by dividing each sum of squares entry by its corresponding degrees of freedom.

A more general form of the analysis of variance table, which we do not need here but which is useful for comparison purposes later, is obtained by incorporating the

TABLE 1.4. Analysis of Variance (ANOVA) Table Incorporating SS(b_0)

Source	df	SS	MS = SS/df
Due to $b_1\|b_0$	1	$\text{SS}(b_1\|b_0) = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	MS_{Reg}
Residual	$n - 2$	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	s^2
Total, corrected	$n - 1$	$\sum_{i=1}^{n} (Y_i - \overline{Y})^2$	
Correction factor (due to b_0)	1	$\text{SS}(b_0) = \left(\sum_{i=1}^{n} Y_i\right)^2 / n = n\overline{Y}^2$	
Total	n	$\sum_{i=1}^{n} Y_i^2$	

correction factor for the mean of the Y's into the table, where, for reasons explained below, it is called SS(b_0). The table takes the form of Table 1.4. (Note the abbreviated headings.) An alternative way of presenting Table 1.4 is to drop the line labeled "Total, corrected" and the rule above it. The "Total" line is then the sum of the remaining three entries. If we rewrite the entries in the order of Table 1.5, we can see a logical development. The first SS entry "Due to b_0" is the amount of variation $n\overline{Y}^2$ explained by a horizontal straight line $\hat{Y} = \overline{Y}$. In fact, if we fit the model $Y = \beta_0 + \epsilon$ via least squares, the fitted model is $\hat{Y} = \overline{Y}$. If we subsequently fit the "with slope also" model $Y = \beta_0 + \beta_1 X + \epsilon$, the "Due to $b_1|b_0$" SS entry S_{XY}^2/S_{XX} is the *extra* variation picked up by the slope term *over and above* that picked up by the intercept alone. This is a special case of the "extra sum of squares" principle to be explained in Chapter 6. Note, however, that most computer programs produce an analysis of variance table in a form that omits SS(b_0), that is, a table like Table 1.4 down to the "Total, corrected" line. Often the word "corrected" is omitted in a printout, but if the "Total" df is only $n - 1$, the source should be labeled as "Total, corrected."

When the calculations for Tables 1.3–1.5 are actually carried out on a pocket calculator, the residual SS is rarely calculated directly as shown, but is usually obtained by subtracting "SS($b_1|b_0$)" from the "total, corrected, SS." As already mentioned in (1.3.6), the sum of squares due to regression SS($b_1|b_0$) can be calculated a number of ways as follows (all summations are over $i = 1, 2, \ldots, n$).

TABLE 1.5. Analysis of Variance Table in Extra SS Order

Source	df	SS	MS = SS/df
b_0	1	$n\overline{Y}^2$	—
$b_1\|b_0$	1	S_{XY}^2/S_{XX}	MS_{Reg}
Residual	$n - 2$	By subtraction	s^2
Total	n	$\sum_{i=1}^{n} Y_i^2$	

TABLE 1.6. Analysis of Variance Table for the Example

Source	df	SS	MS = SS/df	Calculated F-Value
Regression	1	45.5924	45.5924	$57.54 = \frac{45.5924}{0.7923}$
Residual	23	18.2234	$s^2 = 0.7923$	
Total, corrected	24	63.8158		

$$\mathrm{SS}(b_1|b_0) = \Sigma(\hat{Y}_i - \overline{Y})^2 = b_1\{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\} = b_1 S_{XY} \tag{1.3.9}$$

$$= \frac{\{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\}^2}{\Sigma(X_i - \overline{X})^2} = \frac{S_{XY}^2}{S_{XX}} \tag{1.3.10}$$

$$= \frac{\{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n\}^2}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{S_{XY}^2}{S_{XX}} \tag{1.3.11}$$

$$= \frac{\{\Sigma(X_i - \overline{X})Y_i\}^2}{\Sigma(X_i - \overline{X})^2}. \tag{1.3.12}$$

We leave it to the reader to verify the algebraic equivalence of these formulas, which follows from the algebra previously given. Of these forms, Eq. (1.3.9) is perhaps the easiest to use on a pocket calculator because the two pieces have already been calculated to fit the straight line. However, rounding off b_1 can cause inaccuracies, so Eq. (1.3.11) with division performed last is the formula we recommend for calculator evaluation.

Note that the total corrected SS, $\Sigma(Y_i - \overline{Y})^2$, can be written and evaluated either as

$$S_{YY} = \Sigma Y_i^2 - (\Sigma Y_i)^2/n \tag{1.3.13}$$

or as

$$S_{YY} = \Sigma Y_i^2 - n\overline{Y}^2. \tag{1.3.14}$$

The notation $\mathrm{SS}(b_1|b_0)$ is read "the sum of squares for b_1 after allowance has been made for b_0" or, more simply, as "the sum of squares for b_1 given b_0." This notation is further expanded in Section 6.1.

The mean square about regression, s^2, will provide an estimate *based on* $(n - 2)$ *degrees of freedom* of the *variance about the regression*, a quantity we shall call $\sigma^2_{Y \cdot X}$. If the regression equation were estimated from an indefinitely large number of observations, the variance about the regression would represent a measure of the error with which any observed value of Y could be predicted from a given value of X using the determined equation (see note 1 of Section 1.4).

Steam Data Calculations

We now carry out and display in Table 1.6 the calculations of this section for our steam data example and then discuss a number of ways in which the regression equation can be examined. The SS due to regression given b_0 is, using (1.3.11),

$$\mathrm{SS}(b_1|b_0) = \frac{\{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n\}^2}{\{\Sigma X_i^2 - (\Sigma X_i)^2/n\}} = \frac{S_{XY}^2}{S_{XX}}$$

$$= \frac{(-571.1280)^2}{7154.42}$$

$$= 45.5924.$$

The Total (corrected) SS is

$$\Sigma Y_i^2 - (\Sigma Y_i)^2/n$$

$$= 2284.1102 - (235.60)^2/25$$

$$= 63.8158.$$

Our estimate of $\sigma_{Y \cdot X}^2$ is $s^2 = 0.7923$ based on 23 degrees of freedom. The F-value will be explained shortly.

Skeleton Analysis of Variance Table

A *skeleton* analysis of variance table consists of the "source" and "df" columns only. In many situations, for example, as in Section 3.3 when comparing several possible arrangements of experimental runs not yet performed, it is useful to compare the corresponding skeleton analysis of variance tables to see which might be most desirable.

R^2 Statistic

A useful statistic to check is the R^2 value of a regression fit. We define this as

$$R^2 = \frac{(\text{SS due to regression given } b_0)}{(\text{Total SS, corrected for the mean } \overline{Y})}$$

$$= \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2}, \tag{1.3.15}$$

where both summations are over $i = 1, 2, \ldots, n$. (This is a general definition.) Then R^2 measures the "proportion of total variation about the mean $\overline{Y}$ explained by the regression." It is often expressed as a percentage by multiplying by 100. In fact, R is the correlation [see Eq. (1.6.5)] between Y and $\hat{Y}$ and is usually called the multiple correlation coefficient. R^2 is then "the square of the multiple correlation coefficient." For a straight line fit

$$R^2 = \mathrm{SS}(b_1|b_0)/S_{YY}$$

$$= S_{XY}^2/(S_{XX}S_{YY}).$$

Example *(Continued)*. From Table 1.6,

$$R^2 = \frac{45.5924}{63.8158} = 0.7144.$$

Thus the fitted regression equation $\hat{Y} = 13.623 - 0.0798X$ explains 71.44% of the total variation in the data about the average $\overline{Y}$. This is quite a large proportion.

R^2 can take values as high as 1 (or 100%) when all the X values are different. When

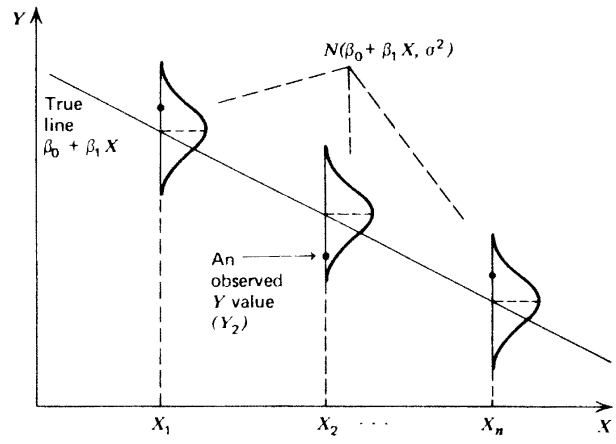

Figure 1.8. Each response observation is assumed to come from a normal distribution centered vertically at the level implied by the assumed model. The variance of each normal distribution is assumed to be the same, σ^2.

repeat runs exist in the data, however, as described in Section 2.1, the value of R^2 cannot attain 1 no matter how well the model fits. This is because no model, however good, can explain the variation in the data due to pure error. An algebraic proof of this fact is given in the solution to Exercise M in "Exercises for Chapters 1–3."

A number of criticisms have been leveled at R^2 in the literature and we discuss them in Section 11.2. In spite of these criticisms, we continue to like R^2 as a "useful first thing to look at" in a regression printout.

1.4. CONFIDENCE INTERVALS AND TESTS FOR β_0 AND β_1

Up to this point we have made no assumptions at all that involve probability distributions. A number of specified algebraic calculations have been made and that is all. We now make the basic assumptions that, in the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $i = 1, 2, \ldots, n$:

1. ϵ_i is a random variable with mean zero and variance σ^2 (unknown); that is, $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.
2. ϵ_i and ϵ_j are uncorrelated, $i \neq j$, so that $\text{cov}(\epsilon_i, \epsilon_j) = 0$. Thus

$$E(Y_i) = \beta_0 + \beta_1 X_i, \qquad V(Y_i) = \sigma^2, \tag{1.4.1}$$

and Y_i and Y_j, $i \neq j$, are uncorrelated.

A further assumption, which is not immediately necessary and will be recalled when used, is that:

3. ϵ_i is a normally distributed random variable, with mean zero and variance σ^2 by assumption 1; that is,

$$\epsilon_i \sim N(0, \sigma^2).$$

Under this additional assumption, ϵ_i, ϵ_j are not only uncorrelated but necessarily independent.

The situation is illustrated in Figure 1.8.

Notes

1. σ^2 may or may not be equal to $\sigma^2_{Y\cdot X}$, the variance about the regression mentioned earlier. If the postulated model is the true model, then $\sigma^2 = \sigma^2_{Y\cdot X}$. If the postulated model is not the true model, then $\sigma^2 < \sigma^2_{Y\cdot X}$. If follows that s^2, the residual mean square that estimates $\sigma^2_{Y\cdot X}$ in any case, is an estimate of σ^2 if the model is correct, but not otherwise. If $\sigma^2_{Y\cdot X} > \sigma^2$ we shall say that the postulated model is incorrect or *suffers from lack of fit*. Ways of deciding this will be discussed later.

2. There is a tendency for errors that occur in many real situations to be normally distributed due to the Central Limit Theorem. If an error term such as ϵ is a sum of errors from several sources, and no source dominates, then no matter what the probability distribution of the separate errors may be, their sum ϵ will have a distribution that will tend more and more to the normal distribution as the number of components increases, by the Central Limit Theorem. An experimental error, in practice, may be a composite of a meter error, an error due to a small leak in the system, an error in measuring the amount of catalyst used, and so on. Thus the assumption of normality is not unreasonable in most cases. In any case we shall later check the assumption by examining residuals (see Chapters 2 and 8).

We now use these assumptions in examining the regression coefficients.

Standard Deviation of the Slope b_1; Confidence Interval for β_1

We know that

$$\begin{aligned} b_1 &= \Sigma(X_i - \overline{X})(Y_i - \overline{Y})/\Sigma(X_i - \overline{X})^2 \\ &= \Sigma(X_i - \overline{X})Y_i/\Sigma(X_i - \overline{X})^2 \end{aligned}$$

[since the other term removed[1] from the numerator is $\Sigma(X_i - \overline{X})\overline{Y} = \overline{Y}\,\Sigma(X_i - \overline{X}) = 0$]

$$b_1 = \{(X_1 - \overline{X})Y_1 + \cdots + (X_n - \overline{X})Y_n\}/\Sigma(X_i - \overline{X})^2. \tag{1.4.2}$$

Now the variance of a function

$$a = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

is

$$V(a) = a_1^2V(Y_1) + a_2^2V(Y_2) + \cdots + a_n^2V(Y_n), \tag{1.4.3}$$

if the Y_i are pairwise uncorrelated and the a_i are constants; furthermore, if $V(Y_i) = \sigma^2$,

$$\begin{aligned} V(a) &= (a_1^2 + a_2^2 + \cdots + a_n^2)\sigma^2 \\ &= (\Sigma\, a_i^2)\sigma^2. \end{aligned} \tag{1.4.4}$$

In the expression for b_1, $a_i = (X_i - \overline{X})/\Sigma(X_i - \overline{X})^2$, since the X_i can be regarded as constants. Hence after reduction

$$V(b_1) = \frac{\sigma^2}{\Sigma(X_i - \overline{X})^2} = \frac{\sigma^2}{S_{XX}}. \tag{1.4.5}$$

[1]This term could be left off altogether, but it is conventional to write the numerator of b_1 in a symmetrical form. See the definition of S_{XY} above Eq. (1.2.9a).

The standard deviation of b_1 is the square root of the variance, that is,

$$\text{sd}(b_1) = \frac{\sigma}{\{\Sigma(X_i - \bar{X})^2\}^{1/2}} = \frac{\sigma}{S_{XX}^{1/2}} \tag{1.4.6}$$

or, if σ is unknown and we use the estimate s in its place, assuming the model is correct, the *estimated* standard deviation of b_1 is given by

$$\text{est. sd}(b_1) = \frac{s}{\{\Sigma(X_i - \bar{X})^2\}^{1/2}} = \frac{s}{S_{XX}^{1/2}}. \tag{1.4.7}$$

An alternative terminology for the estimated standard deviation is the *standard error*, se. We shall use mostly this alternative terminology.

Confidence Interval for β_1

If we assume that the variations of the observations about the line are normal—that is, the errors ϵ_i are all from the same normal distribution, $N(0, \sigma^2)$—it can be shown that we can assign $100(1 - \alpha)\%$ confidence limits for β_1 by calculating

$$b_1 \pm \frac{t(n-2, 1-\frac{1}{2}\alpha)s}{\{\Sigma(X_i - \bar{X})^2\}^{1/2}} \tag{1.4.8}$$

where $t(n - 2, 1 - \frac{1}{2}\alpha)$ is the $100(1 - \frac{1}{2}\alpha)$ percentage point of a t-distribution, with $(n - 2)$ degrees of freedom (the number of degrees of freedom on which the estimate s^2 is based).

Test for H_0: $\beta_1 = \beta_{10}$ Versus H_1: $\beta_1 \neq \beta_{10}$

On the other hand, if a test is appropriate, we can test the null hypothesis that β_1 is equal to β_{10}, where β_{10} is a specified value that could be zero, against the alternative that β_1 is different from β_{10} (usually stated "$H_0: \beta_1 = \beta_{10}$ versus $H_1: \beta_1 \neq \beta_{10}$") by calculating

$$\begin{aligned} t &= \frac{(b_1 - \beta_{10})}{\text{se}(b_1)} \\ &= \frac{(b_1 - \beta_{10})\{\Sigma(X_i - \bar{X})^2\}^{1/2}}{s} \end{aligned} \tag{1.4.9}$$

and comparing $|t|$ with $t(n - 2, 1 - \frac{1}{2}\alpha)$ from a t-table with $(n - 2)$ degrees of freedom—the number on which s^2 is based. The test will be a two-sided test conducted at the $100\alpha\%$ level in this form. Calculations for our example follow.

Example *(Continued)*. Confidence interval for β_1.

$$\begin{aligned} V(b_1) &= \sigma^2/\Sigma(X_i - \bar{X})^2 \\ &= \sigma^2/7154.42 \\ \text{est. } V(b_1) &= s^2/7154.42 \\ &= 0.7923/7154.42 \\ &= 0.00011074 \\ \text{se}(b_1) &= \sqrt{\text{est. } V(b_1)} = 0.0105. \end{aligned}$$

Suppose $\alpha = 0.05$, so that $t(23, 0.975) = 2.069$. Then 95% confidence limits for β_1 are

$$b_1 \pm t(23, 0.975) \cdot s/\{\Sigma(X_i - \overline{X})^2\}^{1/2}, \text{ or}$$

$$-0.0798 \pm (2.069)(0.0105),$$

providing the interval

$$-0.1015 \leq \beta_1 \leq -0.0581.$$

In words, the true value β_1 lies in the interval (-0.1015 to -0.0581), and this statement is made with 95% confidence.

Example *(Continued)*. Test of $H_0: \beta_1 = 0$.

We shall also test the null hypothesis that the true value β_1 is zero, or that there is no straight line sloping relationship between atmospheric temperature (X) and the amount of steam used (Y). As noted above, we write (using $\beta_{10} = 0$)

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

and evaluate

$$\begin{aligned} t &= (b_1 - 0)/\text{se}(b_1) \\ &= -0.0798/0.0105 \\ &= -7.60. \end{aligned}$$

Since $|t| = 7.60$ exceeds the appropriate critical value of $t(23, 0.975) = 2.069$, H_0: $\beta_1 = 0$ is rejected. [Actually, 7.60 also exceeds $t(23, 0.9995)$; we chose a two-sided 95% level test here, however, so that the confidence interval and the t-test would both make use of the same probability level. In this case we can effectively make the test by examining the confidence interval to see if it includes zero, as described below.] The data we have seen cause us to reject the idea that a linear relationship between Y and X might not exist.

Reject or Do Not Reject

If it had happened that the observed $|t|$ value had been smaller than the critical value, we would have said that we *could not reject* the hypothesis. Note carefully that we do not use the word "accept," since we normally cannot accept a hypothesis. The most we can say is that on the basis of certain observed data we cannot reject it. It may well happen, however, that in another set of data we can find evidence that is contrary to our hypothesis and so reject it.

For example, if we see a man who is poorly dressed we may hypothesize, H_0: "This man is poor." If the man walks to save bus fare or avoids lunch to save lunch money, we have no reason to reject this hypothesis. Further observations of this kind may make us feel H_0 is true, but we still cannot accept it unless we know all the true facts about the man. However, a single observation against H_0, such as finding that the man owns a bank account containing $500,000, will be sufficient to reject the hypothesis.

Confidence Interval Represents a Set of Tests

Once we have the confidence interval for β_1 we do not actually have to compute the $|t|$ value for a particular two-sided t-test at the same α-level. It is simplest to examine

the confidence interval for β_1 and see if it contains the value β_{10}. If it does, then the hypothesis $\beta_1 = \beta_{10}$ cannot be rejected; if it does not, the hypothesis is rejected. This can be seen from Eq. (1.4.9), for $H_0: \beta_1 = \beta_{10}$ is rejected at the α-level if $|t| > t(n - 2, 1 - \frac{1}{2}\alpha)$, which implies that

$$|b_1 - \beta_{10}| > t(n - 2, 1 - \tfrac{1}{2}\alpha) \cdot s/\{\Sigma(X_i - \overline{X})^2\}^{1/2};$$

that is, β_{10} lies outside the limits of Eq. (1.4.8).

Standard Deviation of the Intercept; Confidence Interval for β_0

A confidence interval for β_0 and a test of whether or not β_0 is equal to some specified value can be constructed in a way similar to that just described for β_1. We can show that

$$\text{sd}(b_0) = \left\{\frac{\Sigma X_i^2}{n\,\Sigma(X_i - \overline{X})^2}\right\}^{1/2}\sigma. \tag{1.4.10}$$

Replacement of σ by s provides the estimated $\text{sd}(b_0)$, that is, $\text{se}(b_0)$. Thus $100(1 - \alpha)\%$ confidence limits for β_0 are given by

$$b_0 \pm t(n - 2, 1 - \tfrac{1}{2}\alpha)\left\{\frac{\Sigma X_i^2}{n\,\Sigma(X_i - \overline{X})^2}\right\}^{1/2} s. \tag{1.4.11}$$

A t-test for the null hypothesis $H_0: \beta_0 = \beta_{00}$ against the alternative $H_1: \beta_0 \neq \beta_{00}$, where β_{00} is a specified value, will be rejected at the α-level if β_{00} falls outside the confidence interval, or will not be rejected if β_{00} falls inside, or may be conducted separately by finding the quantity

$$t = \frac{(b_0 - \beta_{00})}{\left\{\dfrac{\Sigma X_i^2}{n\,\Sigma(X_i - \overline{X})^2}\right\}^{1/2} s} \tag{1.4.12}$$

and comparing it with percentage points $t(n - 2, 1 - \frac{1}{2}\alpha)$ since $n - 2$ is the number of degrees of freedom on which s^2, the estimate of σ^2, is based. Testing the value of the intercept is seldom of practical interest.

Note: It is also possible to get a *joint confidence region* for β_0 and β_1 simultaneously. See Exercise M in "Exercises for Chapters 5 and 6."

1.5. *F*-TEST FOR SIGNIFICANCE OF REGRESSION

Since the Y_i are random variables, any function of them is also a random variable; two particular functions are MS_{Reg}, the mean square due to regression, and s^2, the mean square due to residual variation, which arise in the analysis of variance tables, Tables 1.3–1.6. These functions then have their own distribution, mean, variance, and moments. It can be shown that the mean values are as follows:

$$\begin{aligned} E(\text{MS}_{\text{Reg}}) &= \sigma^2 + \beta_1^2\Sigma(X_i - \overline{X})^2 \\ E(s^2) &= \sigma^2, \end{aligned} \tag{1.5.1}$$

where, if Z is a random variable, $E(Z)$ denotes its mean value or expected value. Suppose that the errors ϵ_i are independent $N(0, \sigma^2)$ variables. Then it can be shown that if $\beta_1 = 0$, the variable MS_{Reg} multiplied by its degrees of freedom (here, one) and

divided by σ^2 follows a χ^2 distribution with the same (one) number of degrees of freedom. In addition, $(n-2)s^2/\sigma^2$ follows a χ^2 distribution with $(n-2)$ degrees of freedom. Also, since these two variables are independent, a statistical theorem tells us that the ratio

$$F = \frac{\text{MS}_{\text{Reg}}}{s^2} \tag{1.5.2}$$

follows an F-distribution with (here) 1 and $(n-2)$ degrees of freedom provided (recall) that $\beta_1 = 0$. This fact can thus be used as a test of $\text{H}_0: \beta_1 = 0$ versus $\text{H}_1: \beta_1 \neq 0$. We compare the ratio $F = \text{MS}_{\text{Reg}}/s^2$ with the $100(1-\alpha)\%$ point of the tabulated $F(1, n-2)$ distribution in order to determine whether β_1 can be considered nonzero on the basis of the data we have seen.

Example *(Continued)*. From Table 1.6, we see that the required ratio is $F = 45.5924/0.7923 = 57.54$. If we look up percentage points of the $F(1, 23)$ distribution, we see that the 95% point $F(1, 23, 0.95) = 4.28$. Since the calculated F exceeds the critical F-value in the table—that is, $F = 57.54 > 4.28$—we reject the hypothesis $H_0: \beta_1 = 0$, running a risk of less than 5% of being wrong.

p-Values for *F*-Statistics

Many computer printouts give the tail area beyond the observed F-value, typically to three or four decimal places. For our example, this is $\text{Prob}\{F(1, 23) > 57.54\} = 0.0000$. (This simply means that the tail area, when rounded, is smaller than 0.0001 or 0.01%.) This can then be judged in relation to the risk level adopted by the person making the test. Thus a 5% person (one prepared to run the risk of rejecting H_0: $\beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ wrongly once in every 20 tests, on average) and a 10% person (. . . once in every 10 . . .) can look at the same printout and make decisions appropriate to their own percentage level. (In this example, both would come to the same conclusion.)

$F = t^2$

The reader will have noticed that we have had two tests for the test of $\text{H}_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$, a t-test and an F-test. In fact, the two tests are equivalent and mathematically related here, due to the theoretical fact that $F(1, v) = \{t(v)\}^2$; that is, the square of a t-variable with v df is an F-variable with 1 and v df. (*Note:* This only happens when the first df of F is 1.) For the test statistic we have

$$\begin{aligned} F = \frac{\text{MS}_{\text{Reg}}}{s^2} &= \frac{b_1\{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})\}}{s^2} \\ &= \frac{b_1^2\Sigma(X_i - \overline{X})^2}{s^2} \qquad \text{(by the definition of } b_1\text{)} \\ &= \left[\frac{b_1\{\Sigma(X_i - \overline{X})^2\}^{1/2}}{s}\right]^2 \\ &= t^2 \end{aligned} \tag{1.5.3}$$

from Eq. (1.4.9) with $\beta_{10} = 0$. Since the variable $F(1, n-2)$ is the square of the $t(n-2)$ variable, and this carries over to the percentage points (upper α tail of the

F and two-tailed t, total of α), we find exactly the same test results. When there are more regression coefficients the overall F-test for regression, which is the extension of the one given here, does not correspond to the t-test of a coefficient. This is why we have to know about both t- and F-tests, in general. However, tests for *individual coefficients* can be made either in t or $t^2 = F$ form by a similar argument. The t form is often seen in computer programs.

In the example we had an observed F-value of 57.54 and an observed t-value of -7.60. Note that $(-7.6)^2 = 57.76$. Were it not for round-off error this would be equal to the observed value of F. In all such comparisons, the effects of round-off must be taken into consideration.

p-Values for *t*-Statistics

As in the case of the F-statistic discussed above, many computer programs print out a tail area for the observed t-statistic. This is typically the two-tailed probability value, that is, the area outside the t-value observed and outside minus the t-value observed. Each user can then decide on the message he or she reads from this, depending on the user's chosen α-level. In regression contexts where $F(1, \nu) = t^2(\nu)$, the one-sided F-level corresponds to the two-sided t-level.

1.6. THE CORRELATION BETWEEN *X* AND *Y*

When we fit a postulated straight line model $Y = \beta_0 + \beta_1 X + \epsilon$, we are tentatively entertaining the idea that Y can be expressed, apart from error, as a first-order function of X. In such a relationship, X is usually assumed to be "fixed," that is, does not have a probability distribution, while Y is usually assumed to be a random variable that follows a probability distribution with mean $\beta_0 + \beta_1 X$ and variance $V(\epsilon)$. (Even if this is not true exactly for X, in many practical circumstances we can behave as though it is, as discussed in Section 1.1.)

More generally for the moment, let us consider two random variables, U and W, say, which follow some continuous joint bivariate probability distribution $f(U, W)$. Then we define the correlation coefficient between U and W as

$$\rho_{UW} = \frac{\operatorname{cov}(U, W)}{\{V(U)V(W)\}^{1/2}}, \tag{1.6.1}$$

where

$$\operatorname{cov}(U, W) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{U - E(U)\}\{W - E(W)\} f(U, W)\, dU\, dW, \tag{1.6.2}$$

and

$$V(U) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{U - E(U)\}^2 f(U, W)\, dU\, dW, \tag{1.6.3}$$

where

$$E(U) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U f(U, W)\, dU\, dW. \tag{1.6.4}$$

$V(W)$ and $E(W)$ are similarly defined in terms of W. (If the distributions are discrete, summations replace integrals in the usual way.)

It can be shown that $-1 \leq \rho_{UW} \leq 1$. The quantity ρ_{UW} is a measure of the linear association between the random variables U and W. For example, if $\rho_{UW} = 1$, U and W are perfectly positively correlated and the possible values of U and W all lie on a straight line with positive slope in the (U, W) plane. If $\rho_{UW} = 0$, the variables are said to be uncorrelated, that is, linearly unassociated with each other. This does *not* mean that U and W are statistically independent, as most elementary textbooks emphasize. If $\rho_{UW} = -1$, U and W are perfectly negatively correlated and the possible values of U and W again all lie on a straight line, this time with negative slope, in the (U, W) plane.

If a sample of size n, $(U_1, W_1), (U_1, W_2), \ldots, (U_n, W_n)$, is available from the joint distribution, the quantity

$$r_{UW} = \frac{\sum_{i=1}^{n}(U_i - \overline{U})(W_i - \overline{W})}{\{\sum_{i=1}^{n}(U_i - \overline{U})^2\}^{1/2}\{\sum_{i=1}^{n}(W_i - \overline{W})^2\}^{1/2}}, \tag{1.6.5}$$

where $n\overline{U} = \Sigma U_i$ and $n\overline{W} = \Sigma W_i$, called the *sample correlation coefficient between U and W*, is an estimate of ρ_{UW} and provides an empirical measure of the *linear association* between U and W. [If factors $1/(n-1)$ are placed before all summations, then r_{UW} has the form of ρ_{UW} with the covariance and variances replaced by sample values.] Like ρ_{UW}, $-1 \leq r_{UW} \leq 1$.

When the U_i and W_i, $i = 1, 2, \ldots, n$, are all constants, rather than sample values from some distribution, r_{UW} can still be used as a measure of linear association. Because the set of values (U_i, W_i), $i = 1, 2, \ldots, n$, can be thought of as a complete finite distribution, r_{UW} is, effectively, a population rather than a sample value; that is, $r_{UW} = \rho_{UW}$ in this case. [If factors of $1/n$ are inserted before all summations in Eq. (1.6.5), we obtain Eq. (1.6.1) for the discrete case.]

When we are concerned with the situation where $X_1, X_2, \ldots, X_n$ represent the values of a finite X-distribution and corresponding observations $Y_1, Y_2, \ldots, Y_n$ are observed values of random variables whose mean values depend on the corresponding X's (as in this chapter), the correlation coefficient ρ_{XY} can still be defined by Eq. (1.6.1), provided that all integrations with respect to X in expressions like Eqs. (1.6.2)–(1.6.4) are properly replaced by summations over the discrete values $X_1, X_2, \ldots, X_n$. Equation (1.6.5), with X and Y replacing U and W, of course, can still be used to estimate ρ_{XY} by r_{XY} if a sample of observations $Y_1, Y_2, \ldots, Y_n$ at the n X-values $X_1, X_2, \ldots, X_n$, respectively, is available.

In this book we shall make use of expressions of the form of r_{UW} in Eq. (1.6.5); their actual names and roles will depend on whether the quantities that stand in place of U and W are to be considered sample or population values. We shall call all such quantities r_{UW} the *correlation (coefficient) between U and W* and use them as appropriate measures of linear association between various quantities of interest. The distinction made above as to whether they are actually population or sample values is not necessary for our purpose and will be ignored throughout.

When the correlation r_{XY} is nonzero, this means that there exists a linear association between the specific values of the X_i and the Y_i in the data set, $i = 1, 2, \ldots, n$. In our regression situation, we assume that the X_i are values not subject to random error (to a satisfactory approximation at least; such assumptions are rarely strictly true, as discussed in Section 1.1) and the Y_i are random about mean values specified by the model. Later, when we get involved with more than one predictor variable, we shall use the correlation coefficient [e.g., Eq. (1.6.5) with X_1 and X_2 replacing U and W, which we can then call r_{12}] to measure the linear association between the specific

values (X_{1i}, X_{2i}) that occur in a data set. In neither of these cases are we sampling a bivariate distribution.

One final and extremely important point is the following. The value of a correlation r_{XY} shows only the extent to which X and Y are linearly associated. It does not by itself imply that any sort of causal relationship exists between X and Y. Such a false assumption has led to erroneous conclusions on many occasions. (For some examples of such conclusions, including "Lice make a man healthy," see Chapter 8 of *How to Lie with Statistics* by Darrell Huff, W.W. Norton, New York, 1954. Outside North America this book is available as a Pelican paperback.)

Correlation and Regression

Suppose that data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are available. We can obtain $r_{XY} = r_{YX}$ by applying Eq. (1.6.5) and, if we postulate a model $Y = \beta_0 + \beta_1 X + \epsilon$, we can also obtain an estimated regression coefficient b_1 given by Eq. (1.2.9). Our emphasis in the foregoing on the fact that r_{XY} is a measure of linear association between X and Y invites the question of how r_{XY} and b_1 are connected. We first note that the formula in Eq. (1.6.5) is unaffected by changes in the origins or the scales of U and W. Comparing Eq. (1.6.5), with X and Y replacing U and W, with Eq. (1.2.9) we see that

$$b_1 = \left\{\frac{\Sigma(Y_i - \overline{Y})^2}{\Sigma(X_i - \overline{X})^2}\right\}^{1/2} r_{XY}, \tag{1.6.6}$$

where summations are over $i = 1, 2, \ldots, n$. In other words, b_1 is a scaled version of r_{XY}, scaled by the ratio of the "spread" of the Y_i divided by the spread of the X_i. If we write

$$(n-1)s_Y^2 = \Sigma(Y_i - \overline{Y})^2,$$
$$(n-1)s_X^2 = \Sigma(X_i - \overline{X})^2,$$

then

$$b_1 = \frac{s_Y}{s_X} r_{XY}. \tag{1.6.7}$$

Thus b_1 and r_{XY} are closely related but provide different interpretations. The unit-free and scale-free correlation r_{XY} measures linear association between X and Y, while b_1 measures the size of the change in Y which can be predicted when a unit change is made in X. Scale changes in the data will affect b_1 but *not* r_{XY}. In more general regression problems, the regression coefficients are also related to simple correlations of the type of Eq. (1.6.5) but in a more complicated manner. These relationships are rarely of practical interest.

r_{XY} and R Connections

Two additional relationships should be noted:

$$r_{XY} = (\text{sign of } b_1)R = (\text{sign of } b_1)(R^2)^{1/2} \tag{1.6.8}$$

for a straight line fit only, where R is the positive multiple correlation coefficient whose square is

$$R^2 = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} \tag{1.6.9}$$

defined in Section 1.3. Also,

$$r_{Y\hat{Y}} = R; \tag{1.6.10}$$

that is, R is equal to the correlation between the given observations Y_i and the predicted values $\hat{Y}_i$. Equation (1.6.10) is true for any linear regression with any number of predictors [whereas Eq. (1.6.8) holds only for the straight line case]. The reader can confirm these relationships by some straightforward algebra (see the solution to Exercise P in "Exercises for Chapters 1–3.").

(In one unusual application, it was desired to compare the Y_i values also with prespecified W_i values, as well as with the $\hat{Y}_i$. This was done by computing r_{YW} in addition to $r_{Y\hat{Y}}$.)

Testing a Single Correlation

Suppose we have evaluated a single correlation coefficient r_{XY} (which we shall call r without subscripts in what follows), which is an estimate of a true (but unknown) parameter ρ. We can obtain a confidence interval for ρ, or test the null hypothesis $H_0: \rho = \rho_0$, where ρ_0 is the specified value (perhaps zero) versus any of the alternative hypotheses $H_1: \rho \neq \rho_0$ or $\rho > \rho_0$, or $\rho < \rho_0$, using the approximation known as Fisher's[2] z-transformation. This is

$$z' = \tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \equiv \tanh^{-1}(r) \sim N\left(\tanh^{-1}\rho, \frac{1}{n-3}\right) \tag{1.6.11}$$

approximately. Thus an approximate $100(1 - \alpha)\%$ confidence interval for ρ is given by solving

$$\tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm z\left(1 - \frac{\alpha}{2}\right)\left\{\frac{1}{n-3}\right\}^{1/2} = \tfrac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \tag{1.6.12}$$

where $z(1 - \alpha/2)$ is the upper $\alpha/2$ percentage point of the $N(0, 1)$ distribution, for the two values of ρ that satisfy the $\pm$ alternatives on the left. A test statistic for testing H_0 is

$$z = (n-3)^{1/2}\left\{\tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \tfrac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right\}, \tag{1.6.13}$$

which is compared to preselected percentage points of the $N(0, 1)$ distribution. The three alternative hypotheses require a two-tailed test for $H_1: \rho \neq \rho_0$, an upper-tailed test for $H_1: \rho > \rho_0$, and a lower-tailed test for $\rho < \rho_0$ respectively. Because a t-distribution with infinite degrees of freedom is a unit normal distribution, a selection of percentage points is given in the bottom row of the t-table. For other percentage points, use the table of the normal distribution itself.

[2]For more about R. A. Fisher, see *R. A. Fisher: The Life of a Scientist*, by Joan Fisher Box. Wiley, New York, 1978.

Example. Suppose $n = 103$, $r = 0.5$. Choose $\alpha = 0.05$. Then Eq. (1.6.12) reduces to

$$\tfrac{1}{2}\ln 3 \pm 0.196 = \tfrac{1}{2}\ln\{(1 + \rho)/(1 - \rho)\}$$

and the 95% confidence interval for ρ is 0.339–0.632. Any value ρ_0 of ρ outside this interval would thus be rejected in a 5% level two-sided test of $H_0: \rho = \rho_0$ versus $H_1: \rho \neq \rho_0$, due to the parallel arithmetic involved.

Suppose we wish to test $H_0: \rho = 0.6$ versus $H_1: \rho < 0.6$ at the 1% level. The percentage point needed is -2.326 from the 0.02 column (since we need only a lower-tail test) of the t-table with infinite degrees of freedom. From Eq. (1.6.13) we find the test statistic

$$z = 10\{\tfrac{1}{2}\ln 3 - \tfrac{1}{2}\ln (1.6/0.4)\} = -1.438,$$

which falls above the percentage point -2.326; we do not reject H_0 at the one-sided 1% level.

A good reference for this material is *Biometrika Tables for Statisticians*, Vol. I, by E. S. Pearson and H. O. Hartley, Cambridge University Press, 1958. See pp. 28–32 and 139.

1.7. SUMMARY OF THE STRAIGHT LINE FIT COMPUTATIONS

Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
Model: $Y = \beta_0 + \beta_1 X + \epsilon$.

Pocket-Calculator Computations

Evaluate

$$S_{XX} = \Sigma X_i^2 - (\Sigma X_i)^2/n,$$
$$S_{XY} = \Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n,$$
$$S_{YY} = \Sigma Y_i^2 - (\Sigma Y_i)^2/n,$$
$$b_1 = S_{XY}/S_{XX},$$
$$b_0 = \overline{Y} - b_1\overline{X},$$
$$\text{SS}(b_1|b_0) = S_{XY}^2/S_{XX},$$
$$\text{Residual SS} = S_{YY} - S_{XY}^2/S_{XX} = (n - 2)s^2.$$

Get Table 1.5 or 1.3. Then:

$$R^2 = S_{XY}^2/(S_{XX}S_{YY}),$$
$$F = \{S_{XY}^2/S_{XX}\}/s^2 \quad \text{with } (1, n - 2) \text{ df},$$
$$t = F^{1/2}.$$

(Either F or t can be used to test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$.)

1.8. HISTORICAL REMARKS

It appears that Sir Francis Galton (1822–1911), a well-known British anthropologist and meteorologist, was responsible for the introduction of the word "regression." Originally he used the term "reversion" in an unpublished address "Typical laws of heredity in man" to the Royal Institution on February 9, 1877. The later term "regression" appears in his Presidential address made before Section H of the British Association at Aberdeen, 1885, printed in *Nature*, September 1885, pp. 507–510, and also in a paper "Regression towards mediocrity in hereditary stature," *Journal of the Anthropological Institute,* **15,** 1885, 246–263. In the latter, Galton reports on his initial discovery (p. 246) that the offspring of seeds "did *not* tend to resemble their parent seeds in size, but to be always more mediocre [i.e., more average] than they—to be smaller than the parents, if the parents were large; to be larger than the parents if the parents were very small.... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it." Galton then describes how the same features were observed in the records of "heights of 930 adult children and of their respective parentages, 205 in number." Essentially, he shows that, if Y = child's height and X = "parents height" (actually a weighted average of the mother's and father's heights; see the original paper for the details), a regression equation something like $\hat{Y} = \overline{Y} + \frac{2}{3}(X - \overline{X})$ is appropriate, although he does not phrase it in this manner. (The notation is explained in Section 1.1.) Galton's paper makes fascinating reading, as does the account in *The History of Statistics*, by S. M. Stigler, 1986, pp. 294–299, Belknap Press of Harvard University. Galton's analysis would be called a "correlation analysis" today, a term for which he is also responsible. The term "regression" soon came to be applied to relationships in situations other than the one from which it originally arose, including situations where the predictor variables were *not* random, and its use has persisted to this day. In most model-fitting situations today, there is no element of "regression" in the original sense. Nevertheless, the word is so established that we continue to use it.

There has been a dispute about who first discovered the method of least squares. It appears that it was discovered independently by Carl Friedrich Gauss (1777–1855) and Adrien Marie Legendre (1752–1833), that Gauss started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date), and that the first account was published by Legendre in 1805. When Gauss wrote in 1809 that he had used the method earlier than the date of Legendre's publication, controversy concerning the priority began. The facts are carefully sifted and discussed by R. L. Plackett in "Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares," *Biometrika,* **59,** 1972, 239–251, a paper we enthusiastically recommend. Also recommended are accounts by C. Eisenhart, "The meaning of 'least' in least squares," *Journal of the Washington Academy of Sciences,* **54,** 1964, 24–33 (reprinted in *Precision Measurement and Calibration*, edited by H. H. Ku, National Bureau of Standards Special Publication 300, Vol. 1, 1969), and "Gauss, Carl Friedrich," in the *International Encyclopedia of the Social Sciences*, Vol. 6, 1968, pp. 74–81, Macmillan Co., Free Press Division, New York, and the *Encyclopedia of Statistical Sciences*, Vol. 3, 1983, pp. 305–309, Wiley, New York; and a related account by S. M. Stigler, "Gergonne's 1815 paper on the design and analysis of polynomial regression experiments," *Historia Mathematica,* **1,** 1974, 431–447 (see p. 433). See also Stigler's book *The History of Statistics,* 1986, Belknap Press of Harvard University.

APPENDIX 1A. STEAM PLANT DATA

The response is X1, the predictors X2–X10.

1	10.98	5.20	0.61	7.4	31	20	22	35.3	54.8	4
2	11.13	5.12	0.64	8.0	29	20	25	29.7	64.0	5
3	12.51	6.19	0.78	7.4	31	23	17	30.8	54.8	4
4	8.40	3.89	0.49	7.5	30	20	22	58.8	56.3	4
5	9.27	6.28	0.84	5.5	31	21	0	61.4	30.3	5
6	8.73	5.76	0.74	8.9	30	22	0	71.3	79.2	4
7	6.36	3.45	0.42	4.1	31	11	0	74.4	16.8	2
8	8.50	6.57	0.87	4.1	31	23	0	76.7	16.8	5
9	7.82	5.69	0.75	4.1	30	21	0	70.7	16.8	4
10	9.14	6.14	0.76	4.5	31	20	0	57.5	20.3	5
11	8.24	4.84	0.65	10.3	30	20	11	46.4	106.1	4
12	12.19	4.88	0.62	6.9	31	21	12	28.9	47.6	4
13	11.88	6.03	0.79	6.6	31	21	25	28.1	43.6	5
14	9.57	4.55	0.60	7.3	28	19	18	39.1	53.3	5
15	10.94	5.71	0.70	8.1	31	23	5	46.8	65.6	4
16	9.58	5.67	0.74	8.4	30	20	7	48.5	70.6	4
17	10.09	6.72	0.85	6.1	31	22	0	59.3	37.2	6
18	8.11	4.95	0.67	4.9	30	22	0	70.0	24.0	4
19	6.83	4.62	0.45	4.6	31	11	0	70.0	21.2	3
20	8.88	6.60	0.95	3.7	31	23	0	74.5	13.7	4
21	7.68	5.01	0.64	4.7	30	20	0	72.1	22.1	4
22	8.47	5.68	0.75	5.3	31	21	1	58.1	28.1	6
23	8.86	5.28	0.70	6.2	30	20	14	44.6	38.4	4
24	10.36	5.36	0.67	6.8	31	20	22	33.4	46.2	4
25	11.08	5.87	0.70	7.5	31	22	28	28.6	56.3	5

EXERCISES

Exercises for Chapter 1 are located in the section "Exercises for Chapters 1–3", at the end of Chapter 3.