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

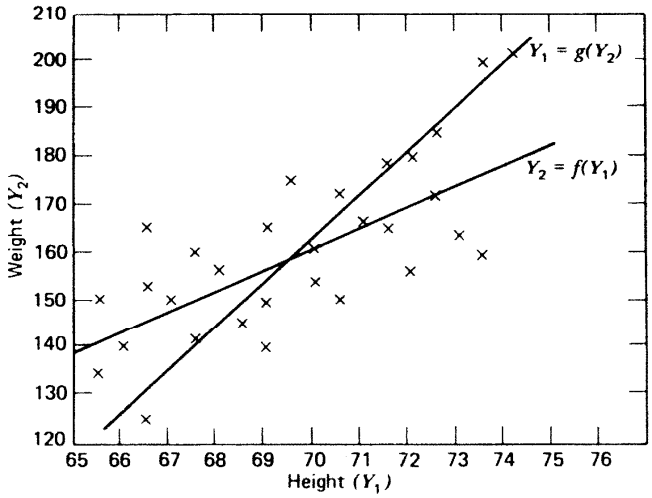

Figure 1.1. Heights and weights of 30 American males.

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.

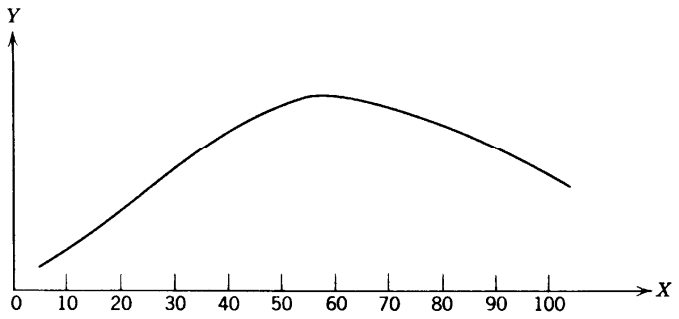

Figure 1.2. A response relationship.

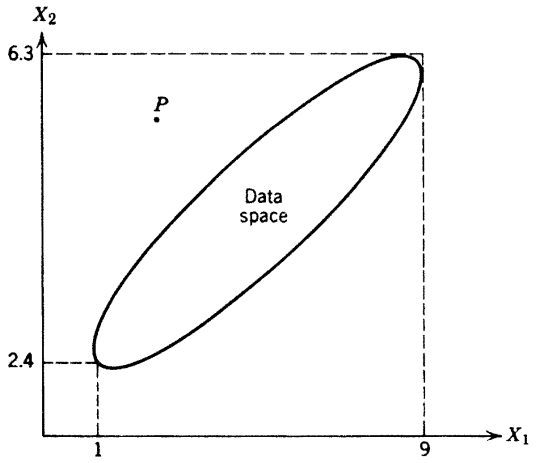

Figure 1.3. A point P outside the data space, whose coordinates nevertheless lie within the ranges of the predictor variables observed.

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

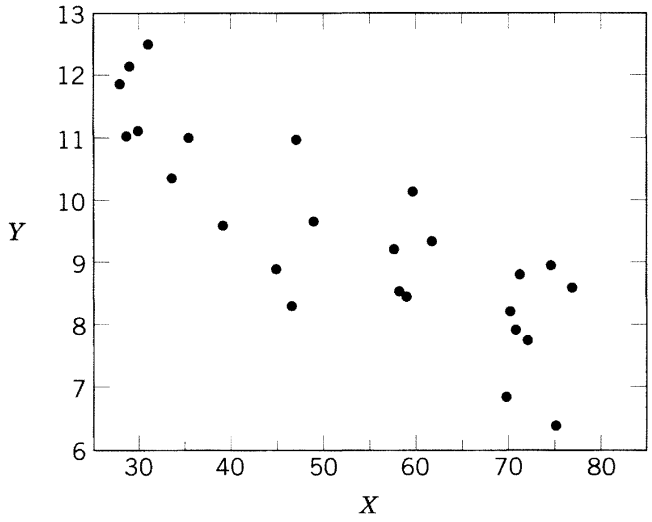

Figure 1.4. Plot of the steam data for variables 1 (Y) and 8 (X).

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 $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. (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

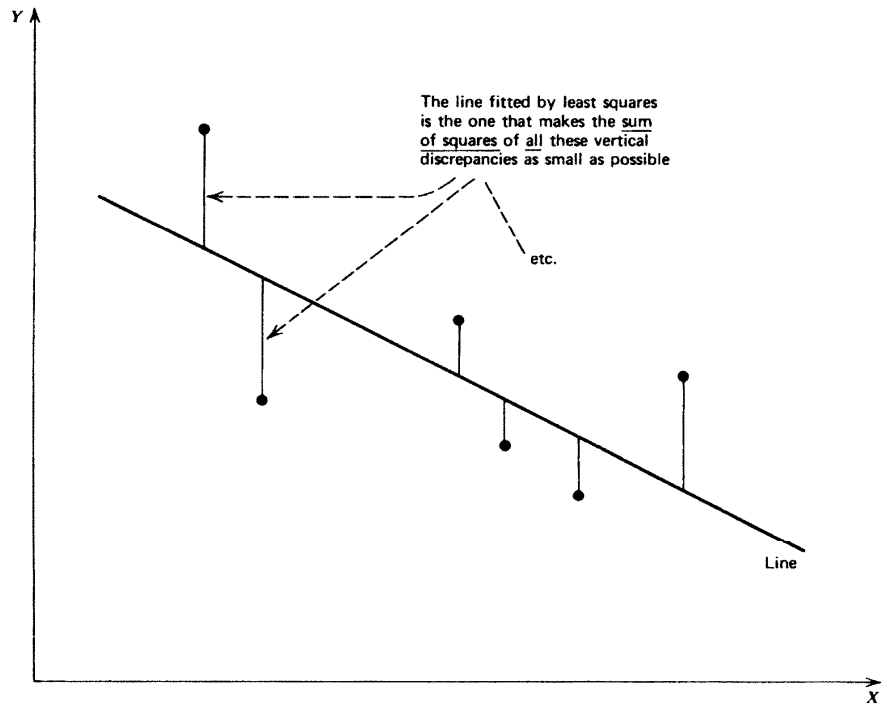

Figure 1.5. The vertical deviations whose sum of squares is minimized for the least squares procedure.

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

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

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

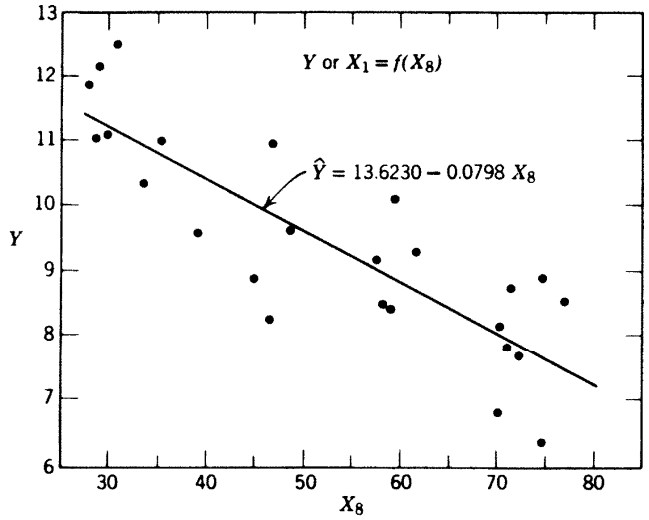

Figure 1.6. Plot of the steam data—variables 1 (Y) and 8 (X)—and the least squares line.

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,

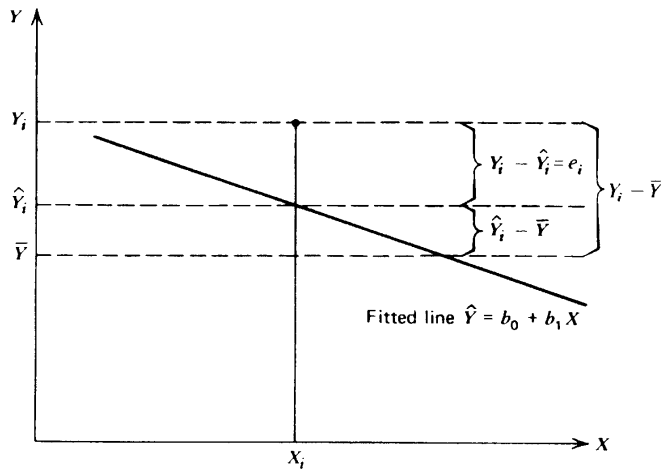

Figure 1.7. Geometrical meaning of the identity (1.3.1).