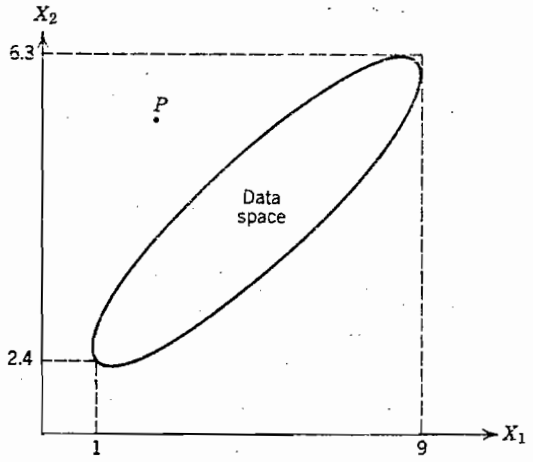

Figure 1.3 A point outside the data space.

depends on variables $X_1, X_2, \ldots, X_k$. We determine a regression equation from data which "cover" certain areas 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 multi-dimensional space. To take a simple example, consider the region defined 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 both coordinates of P lie within these ranges, P itself lies outside the region. When more dimensions are involved, misunderstandings of this sort easily arise.)

1.2. Linear Regression: Fitting a Straight Line

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,

Ref: Draper, N. R. and H. Smith, Applied Regression Analysis, Second Edition, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., 1981.

in the printout on page 616, the twenty-five 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.

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

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

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.2.1}$$

parameter that goes with $X^2 = XX$. The natural extension of this sort of notation appears, for example, in Sections 5.1 and 7.7.)

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 twenty-five observations in Table 1.1 to give us *estimates* b_0 and b_1 of β_0 and β_1; thus we can write

$$\hat{Y} = b_0 + b_1 X, \tag{1.2.2}$$

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 10.

Our estimation procedure will be that of least squares. 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 then 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*, ed. H. H. Ku, National Bureau of Standards Special Publication 300, Vol. I, 1969) and "Gauss, Carl Friedrich," *International Encyclopedia of the Social Sciences*, Vol. 6, 1968, pp. 74–81, Macmillan Co., Free Press Div., 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).

Under certain assumptions to be discussed in Chapter 2, the method of least squares has certain properties. For the moment we state it as our chosen method of estimating the parameters without justification. Suppose we have

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 given by

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

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

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

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.*

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

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

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

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

we have that

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

Substituting Eq. (1.2.10) into Eq. (1.2.2) gives the estimated regression equation

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

where b_1 is given by Eq. (1.2.9).

Note 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. Let us now perform these calculations on the data given as an example in Table 1.1. We find the following:

$$\begin{aligned}
n &= 25 \\
\sum Y_i &= 10.98 + 11.13 + \cdots + 11.08 = 235.60 \\
\overline{Y} &= 235.60/25 = 9.424 \\
\sum X_i &= 35.3 + 29.7 + \cdots + 28.6 = 1315 \\
\overline{X} &= 1315/25 = 52.60 \\
\sum X_i Y_i &= (10.98)(35.3) + (11.13)(29.7) + \cdots + (11.08)(28.6) \\
&= 11821.4320 \\
\sum X_i^2 &= (35.3)^2 + (29.7)^2 + \cdots + (28.6)^2 = 76323.42 \\
b_1 &= \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} \\
b_1 &= \frac{11821.4320 - (1315)(235.60)/25}{76323.42 - (1315)^2/25} = \frac{-571.1280}{7154.42} \\
b_1 &= -0.079829.
\end{aligned}$$

The fitted equation is thus

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

The fitted regression line is plotted in Figure 1.4. We can tabulate for each of the twenty-five 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.

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

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

or

$$y = \beta_0' + \beta_1 x + \varepsilon$$

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

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

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 the centered model as

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

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" pieces of information has been used to enable the proper adjustments to be made to the model so that the intercept term can be removed.

1.3. The Precision of the Estimated Regression

We now tackle the question of what measure of precision can be attached to our estimate of 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 1.6. The residual $e_i = Y_i - \hat{Y}_i$ is the difference between two quantities: (*i*) the deviation of the observed Y_i from the overall mean $\overline{Y}$, and (*ii*) the deviation of the fitted $\hat{Y}_i$ from the overall mean $\overline{Y}$. Note that the average of the $\hat{Y}_i$, namely

$$\begin{aligned} \sum \hat{Y}_i/n &= \sum (b_0 + b_1 X_i)/n \\ &= (nb_0 + b_1 n\overline{X})/n \\ &= b_0 + b_1 \overline{X} \\ &= \overline{Y}. \end{aligned}$$

We now return to a discussion of Eq. (1.3.2). The quantity $(Y_i - \bar{Y})$ is the deviation of the ith observation from the overall mean and so the left-hand side of Eq. (1.3.2) 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*. Since $\hat{Y}_i - \bar{Y}$ 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 (the ith *residual*), we can express Eq. (1.3.2) in words as follows:

$$\begin{bmatrix}\text{Sum of squares}\\ \text{about the mean}\end{bmatrix} = \begin{bmatrix}\text{Sum of squares}\\ \text{due to regression}\end{bmatrix} + \begin{bmatrix}\text{Sum of squares}\\ \text{about regression}\end{bmatrix}$$

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, $\sum(Y_i - \hat{Y}_i)^2$, to the fact that the actual observations do not all lie on the regression line —if they all did, the sum of squares about the regression would be zero! From this procedure we can see that a 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 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.

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 (for of the numbers $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y}$, 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 $\sum(\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum(X_i - \bar{X})^2$], 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, as we can see, the sum of squares of the residuals $Y_i - \hat{Y}_i$, in fact) 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. (1.3.2), we can show the split of degrees of freedom as

$$n - 1 = 1 + (n - 2). \tag{1.3.4}$$

From Eqs. (1.3.2) and (1.3.4) 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.

Table 1.4 Analysis of Variance (ANOVA) Table Incorporating $SS(b_0)$

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

We leave it to the reader to verify the algebraic equivalence of these formulas, which follow from algebra previously given on pp. 14 and 18. Of these forms, Eq. (1.3.5) is the easiest to use on a pocket calculator because the two pieces have already been calculated to fit the straight line. However, rounding off of b_1 can cause inaccuracies, so Eq. (1.3.7) with division performed last is the formula we recommend for calculator evaluation.

Note that the total corrected SS can be written and evaluated as

$$S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2/n \qquad (1.3.9)$$

$$= \sum Y_i^2 - n\bar{Y}^2 \qquad (1.3.10)$$

The notation $SS(b_1 \mid b_0)$ is read "the sum of squares for b_1 after allowance has been made for b_0." The purpose of this notation is explained in Sections 2.2 and 2.7.

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).

2. ε_i and ε_j are uncorrelated, $i \neq j$, so that

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0.$$

Thus

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

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 (1), that is,

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

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

The situation is illustrated in Figure 1.7.

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 $\sigma^2 < \sigma^2_{Y \cdot X}$. It follows that s^2, the residual mean square which 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.

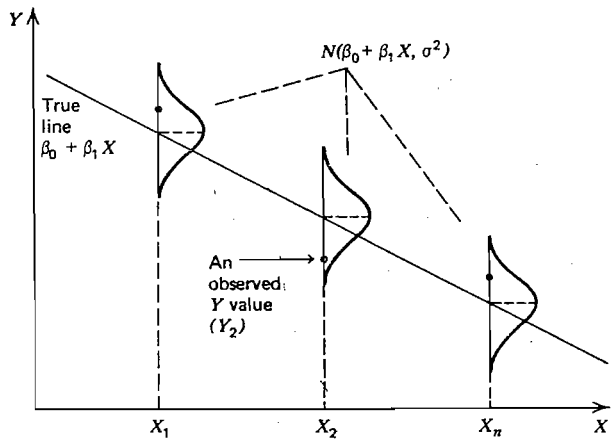

Figure 1.7 Each response observation is assumed to come from a normal distribution centered vertically at the level implied by the assumed model. The variance of each normal distribution is assumed to be the same, σ^2.