D Linear Regression簡介

D.1 Linear Regression

資料來源：

Draper, N. R., and H. Smith, ***Applied Regression Analysis***, Second Edition, John Wiley & Sons, Inc., 1981, pp.8-23.

D.2 Development of PSI Equation

資料來源：

Two Pages of S-PLUS Example Outputs

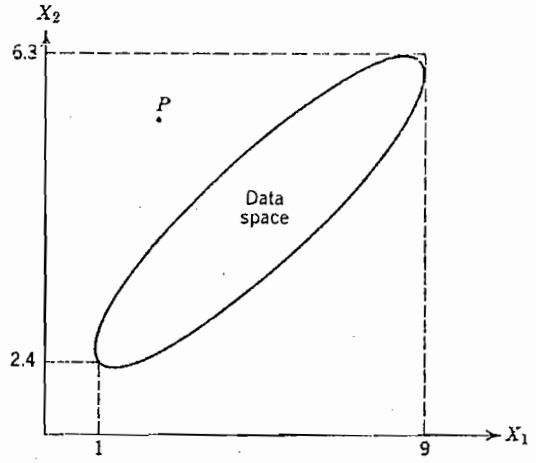

Figure 1.3 A point outside the data space.

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

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

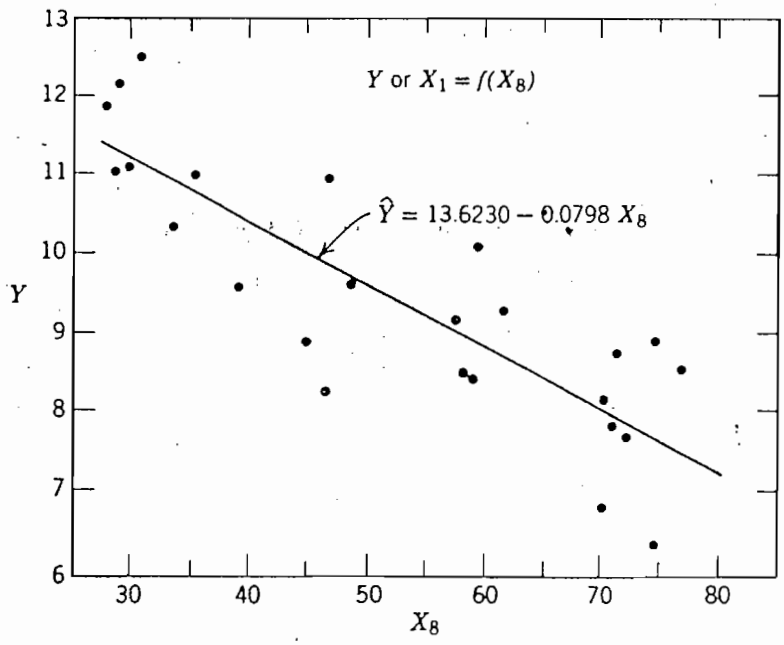

Figure 1.4 The data and the fitted straight line.

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. We begin by assuming that it 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. β_0 and β_1 are called the *parameters* of the model.

(*Note.* 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,

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

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 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 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

available n sets of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. (In our example $n = 25$.) Then by Eq. (1.2.1) we can write

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{1.2.3}$$

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

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

We shall choose our estimates b_0 and b_1 to be the values which, 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 X_i, Y_i are the fixed numbers which we have observed). We can determine b_0 and b_1 by differentiating Eq. (1.2.4) first with respect to

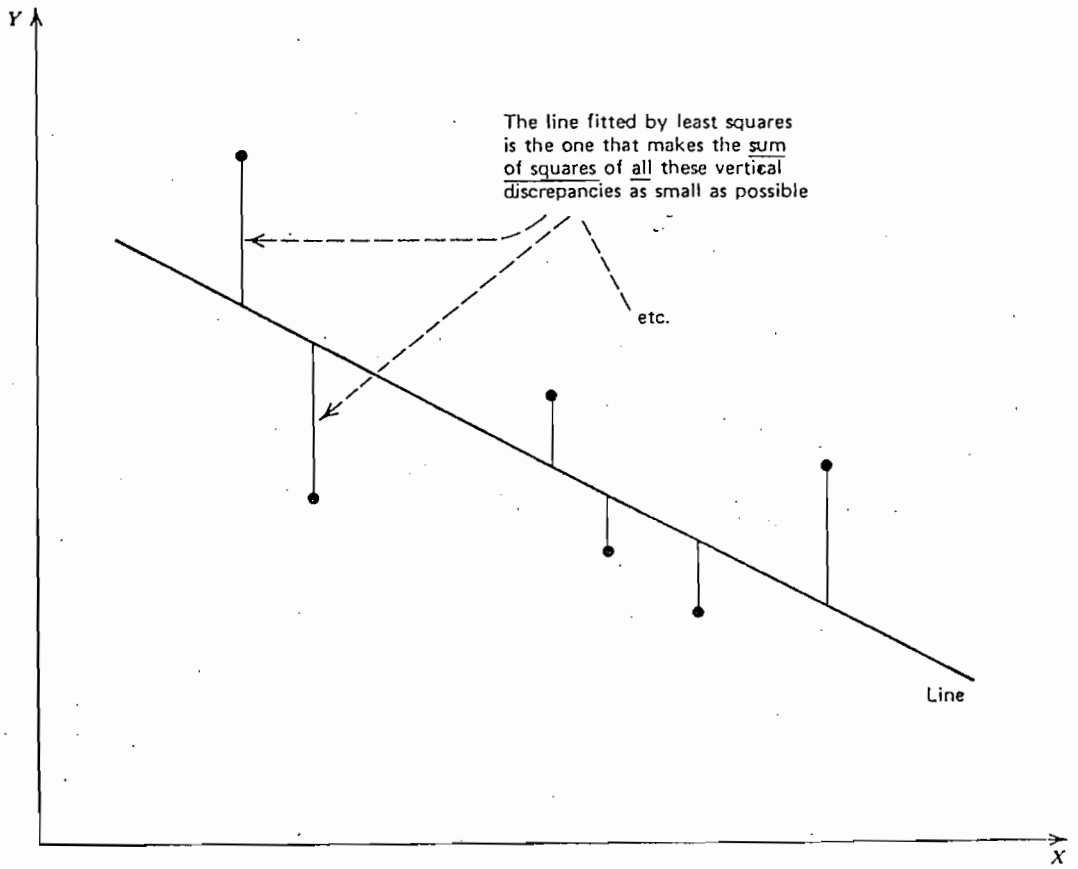

Figure 1.5 The vertical deviations whose sum of squares is minimized for the least squares procedure.

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

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

we have that

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

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 $\sum X_i^2$ is called the *uncorrected sum of squares of the X's* and $(\sum X_i)^2/n$ is the *correction for the mean of the X's*. The difference is called the *corrected sum of squares of the X's*. Similarly, $\sum X_i Y_i$ is called the *uncorrected sum of products*, and $(\sum X_i)(\sum Y_i)/n$ is the *correction for the means*. The difference is called the *corrected sum of products of X and Y*.

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 $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ 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 roundoff characteristics.

A convenient notation, now and later, is to write

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

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

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

and

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

The easily remembered formula for b_1 is then

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

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

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

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

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

where b_1 is given by Eq. (1.2.9).

Note that if we set $X = \bar{X}$ in (1.2.11), then $\hat{Y} = \bar{Y}$. This means that the point $(\bar{X}, \bar{Y})$ 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:

$$n = 25$$

$$\sum Y_i = 10.98 + 11.13 + \cdots + 11.08 = 235.60$$

$$\bar{Y} = 235.60/25 = 9.424$$

$$\sum X_i = 35.3 + 29.7 + \cdots + 28.6 = 1315$$

$$\bar{X} = 1315/25 = 52.60$$

$$\sum X_i Y_i = (10.98)(35.3) + (11.13)(29.7) + \cdots + (11.08)(28.6)$$

$$= 11821.4320$$

$$\sum X_i^2 = (35.3)^2 + (29.7)^2 + \cdots + (28.6)^2 = 76323.42$$

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n}$$

$$b_1 = \frac{11821.4320 - (1315)(235.60)/25}{76323.42 - (1315)^2/25} = \frac{-571.1280}{7154.42}$$

$$b_1 = -0.079829.$$

The fitted equation is thus

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

$$\hat{Y} = 9.4240 - 0.079829(X - 52.60)$$

$$\hat{Y} = 13.623005 - 0.079829X.$$

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 $\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X})$,

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

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

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

Thus the residuals sum to zero. 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 $Y = \beta_0 + \beta_1 X + \varepsilon$ omission of β_0 implies that the line passes through $X = 0$, $Y = 0$—that is, that the line has a zero *intercept* $\beta_0 = 0$ at $X = 0$. We note here, before the more general discussion in Section 5.4, that physical removal of β_0 from the model is always possible by "centering" the data, but that this is quite different from setting $\beta_0 = 0$. For example, if we write Eq. (1.2.1) in the form

$$Y - \bar{Y} = (\beta_0 + \beta_1 \bar{X} - \bar{Y}) + \beta_1 (X - \bar{X}) + \varepsilon$$

or

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

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

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

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

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

omitting the β_0' (intercept) term entirely. We have lost one parameter but there is a corresponding loss in the data since the quantities $Y_i - \bar{Y}$, $i = 1, 2, \ldots, n$ 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:

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

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 $\bar{Y}$, and (*ii*) the deviation of the fitted $\hat{Y}_i$ from the overall mean $\bar{Y}$. Note that the average of the $\hat{Y}_i$, namely

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

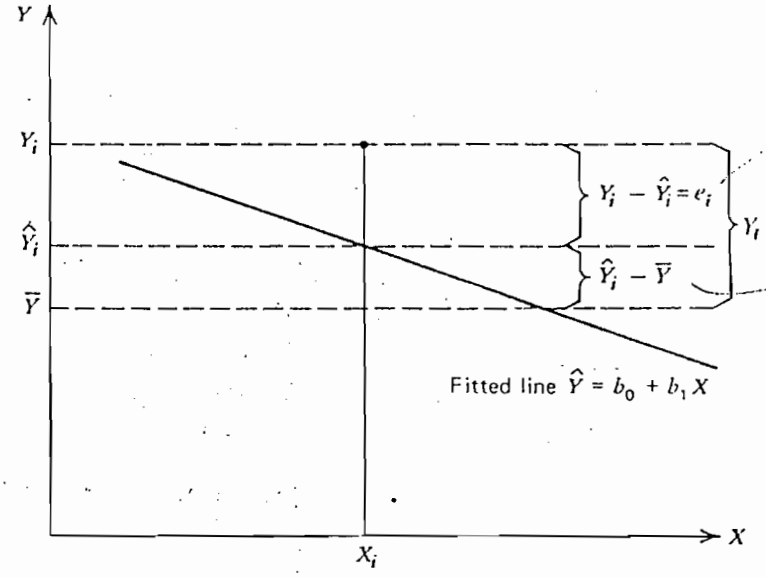

Figure 1.6 Geometrical meaning of the identity (1.3.1).

In other words, the average of the $\hat{Y}_i$'s is the same as the average of the Y_i's. This fact also reconfirms that $\sum e_i = \sum (Y_i - \hat{Y}_i) = n\bar{Y} - n\bar{Y} = 0$, as previously stated.

We can also rewrite Eq. (1.3.1) as

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

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

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2. \tag{1.3.2}$$

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

$$\hat{Y}_i - \bar{Y} = b_1(X_i - \bar{X})$$

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - b_1(X_i - \bar{X}).$$

It follows that the cross-product term is

$$\begin{aligned} \text{CPT} &= 2\sum b_1(X_i - \bar{X})\{(Y_i - \bar{Y}) - b_1(X_i - \bar{X})\} \\ &= 2b_1\{S_{XY} - b_1 S_{XX}\} \\ &= 0 \end{aligned}$$

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

$$\begin{aligned} \sum (\hat{Y}_i - \bar{Y})^2 &= \sum b_1{}^2(X_i - \bar{X})^2 \\ &= b_1{}^2 S_{XX} \\ &= b_1 S_{XY}. \end{aligned} \tag{1.3.3}$$

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:

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

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 $R^2 = (\text{SS due to regression})/(\text{SS about mean})$ 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 1.3 Analysis of Variance (ANOVA) Table; the Basic Split

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

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

A more general form of the analysis of variance table, which we do not need here but which is useful for comparison purposes later (see Section 2.2), is obtained by incorporating the correction factor for the mean of the Y's into the table where, for reasons explained in Section 2.2, 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 labelled "Total, corrected" and the rule above it. The "Total" line is then the sum of the remaining three entries.)

When the calculations for Tables 1.3 and 1.4 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." 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 $i = 1, 2, \ldots, n$).

$$SS(b_1|b_0) = \sum(\hat{Y}_i - \bar{Y})^2 = b_1\{\sum(X_i - \bar{X})(Y_i - \bar{Y})\} = b_1 S_{XY} \tag{1.3.5}$$

$$= \frac{\{\sum(X_i - \bar{X})(Y_i - \bar{Y})\}^2}{\sum(X_i - \bar{X})^2} = \frac{S_{XY}^2}{S_{XX}} \tag{1.3.6}$$

$$= \frac{\{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n\}^2}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{S_{XY}^2}{S_{XX}} \tag{1.3.7}$$

$$= \frac{\{\sum(X_i - \bar{X})Y_i\}^2}{\sum(X_i - \bar{X})^2}. \tag{1.3.8}$$

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

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

$$= \sum Y_i^2 - n\bar{Y}^2 \tag{1.3.10}$$

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).

We shall now carry out the calculations of this section for our example and then discuss a number of ways the regression equation can be examined. The SS due to regression is, using (1.3.7),

$$= \frac{\{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n\}^2}{\{\sum X_i^2 - (\sum X_i)^2/n\}}$$

$$= (-571.1280)^2/7154.42$$

$$= 45.5924.$$

The Total (corrected) SS is $\sum Y_i^2 - (\sum Y_i)^2/n$

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

$$= 63.8158$$

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

Table 1.5 The Analysis of Variance Table for the Example

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

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 1.8 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.

1.4. Examining the Regression Equation

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 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$,

1. ε_i is a random variable with mean zero and variance σ^2 (unknown), that is, $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$.

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.

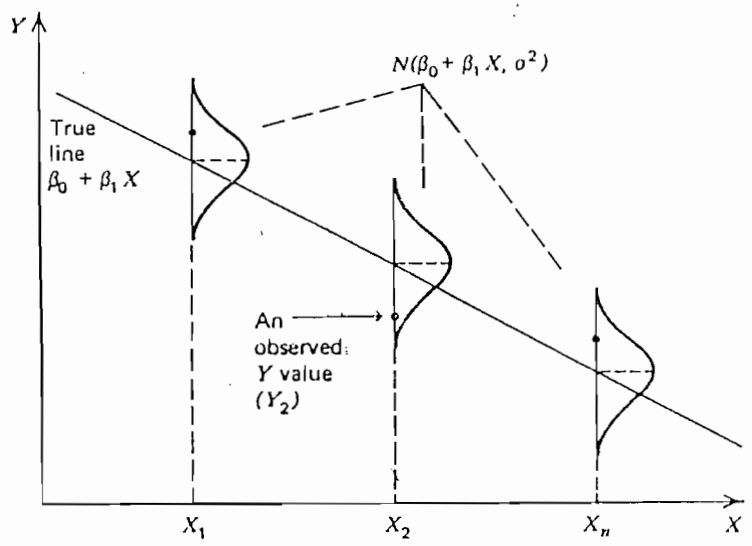

Figure 1.7 Each response observation is assumed to come from a normal distribution centered vertically at the level implied by the assumed model. The variance of each normal distribution is assumed to be the same, σ^2.

```
flex_psi> lm2 <- lm(PSI ~ log10(1 + SV) + RD^2 + CP^0.5)
flex_psi> print(summary(lm2))

Call: lm(formula = PSI ~ log10(1 + SV) + RD^2 + CP^0.5)
Residuals:
      Min        1Q    Median        3Q     Max
 -0.5458 -0.3319 -0.04131 0.2205 1.107

Coefficients:
                Value Std. Error   t value Pr(>|t|)
  (Intercept)  5.0331     0.1263   39.8634   0.0000
log10(1 + SV) -1.9179     0.1395  -13.7457   0.0000
      I(RD^2) -1.3889     0.3326   -4.1756   0.0001
   I(CP^0.5)  -0.0087     0.0070   -1.2417   0.2185

Residual standard error: 0.3858 on 70 degrees of freedom
Multiple R-Squared: 0.8442

Correlation of Coefficients:
             (Intercept) log10(1 + SV) I(RD^2)
log10(1 + SV) -0.8506
      I(RD^2) -0.2268       0.0134
   I(CP^0.5)   0.3000      -0.6496      0.0270
flex_psi> print(anova(lm2))
Analysis of Variance Table

Response: PSI

Terms added sequentially (first to last)
             Df Sum of Sq Mean Sq F Value      Pr(F)
log10(1 + SV)  1  53.6505 53.6505 360.538 0.000000
      I(RD^2)  1   2.5549  2.5549  17.169 0.000094
   I(CP^0.5)   1   0.2294  0.2294   1.542 0.218502
   Residuals  70  10.4165  0.1488
flex_psi> lm3 <- lm(PSI ~ log10(1 + SV) + log10(1 + RDV) + RD^2 + CP^0.5)
flex_psi> print(summary(lm3))

Call: lm(formula = PSI ~ log10(1 + SV) + log10(1 + RDV) + RD^2 + CP^0.5)
Residuals:
      Min        1Q    Median        3Q     Max
 -0.5645 -0.2148 -0.05779 0.204 0.7287

Coefficients:
                Value Std. Error   t value Pr(>|t|)
  (Intercept)  4.8316     0.1188   40.6643   0.0000
log10(1 + SV) -1.2857     0.1822   -7.0576   0.0000
log10(1 + RDV) -1.1083    0.2365   -4.6856   0.0000
      I(RD^2) -1.1902     0.2949   -4.0365   0.0001
   I(CP^0.5)  -0.0114     0.0062   -1.8517   0.0683

Residual standard error: 0.3384 on 69 degrees of freedom
Multiple R-Squared: 0.8818

Correlation of Coefficients:
             (Intercept) log10(1 + SV) log10(1 + RDV) I(RD^2)
log10(1 + SV) -0.8008
log10(1 + RDV)  0.3619      -0.7406
      I(RD^2) -0.2612       0.1154       -0.1438
   I(CP^0.5)   0.3126      -0.5045        0.0945      0.0131
flex_psi> print(anova(lm3))
Analysis of Variance Table

Response: PSI

Terms added sequentially (first to last)
             Df Sum of Sq Mean Sq F Value      Pr(F)
log10(1 + SV)  1  53.6505 53.6505 468.466 0.0000000
log10(1 + RDV)  1   3.0620  3.0620  26.737 0.0000022
      I(RD^2)  1   1.8440  1.8440  16.101 0.0001503
   I(CP^0.5)   1   0.3927  0.3927   3.429 0.0683462
   Residuals  69   7.9021  0.1145

flex_psi> print(anova(lm2, lm3))
Analysis of Variance Table

Response: PSI

                                                  Terms Resid. Df     RSS
1                          log10(1 + SV) + RD^2 + CP^0.5        70 10.4165
2 log10(1 + SV) + log10(1 + RDV) + RD^2 + CP^0.5        69  7.9021

                  Test Df Sum of Sq F Value          Pr(F)
1
2 +log10(1 + RDV)  1  2.51434 21.9547 0.0000135804
```

```
source("try-mtrx.s")
> #source(flexdat.s")
attach(flex.frame.new, 2)
> Y <- PSI
> X <- cbind(1, log10(1 + SV), RD^2, CP^0.5)
> dimnames(X)[[2]] <- c("1", "log10(1+SV)", "RD^2", "CP^0.5")
> XT <- t(X)
> XTX <- XT %*% X
> XTXI <- solve(XTX)
> H <- X %*% XTXI %*% XT
> beta <- XTXI %*% XT %*% Y
> print(X[1:4, 1:4])
       1 log10(1+SV)     RD^2       CP^0.5
IL-F3 1     0.5797836 0.0100    0.000000
IL-F4 1     1.3324385 0.0484 18.520259
IL-F5 1     1.0086002 0.0064    2.828427
IL-F6 1     0.6532125 0.0064    0.000000
> print(Y[1:4])
 IL-F3 IL-F4 IL-F5 IL-F6
   4.3   2.4   3.3   4.4
> print(XT[1:4, 1:4])
                  IL-F3       IL-F4       IL-F5       IL-F6
             1 1.0000000    1.000000 1.000000 1.0000000
log10(1+SV) 0.5797836    1.332438 1.008600 0.6532125
         RD^2 0.0100000    0.048400 0.006400 0.0064000
       CP^0.5 0.0000000 18.520259 2.828427 0.0000000
> print(XTX[1:4, 1:4])
                          1 log10(1+SV)         RD^2        CP^0.5
             1    74.00000     75.243560    5.623900    565.93854
log10(1+SV)   75.24356     89.753868    5.546208    746.98887
         RD^2     5.62390      5.546208    1.775583     39.05244
       CP^0.5 565.93854    746.988867 39.052441 9582.00000
> print(XTXI[1:4, 1:4])
                              1   log10(1+SV)              RD^2            CP^0.5
             1    0.107125429 -0.100699961 -0.0639961226    0.0017840269
log10(1+SV) -0.100699961    0.130828344    0.0041787264 -0.0042684738
         RD^2 -0.063996123    0.004178726    0.7435218138    0.0004237183
       CP^0.5    0.001784027 -0.004268474    0.0004237183    0.0003300256
> print(beta)
                         [,1]
             1    5.03306638
log10(1+SV) -1.91792059
         RD^2 -1.38893699
       CP^0.5 -0.00870147
> lm2 <- lm(PSI ~ log10(1 + SV) + RD^2 + CP^0.5)
> print(summary(lm2))

Call: lm(formula = PSI ~ log10(1 + SV) + RD^2 + CP^0.5)
Residuals:
      Min        1Q    Median        3Q     Max
 -0.5458 -0.3319 -0.04131 0.2205 1.107

Coefficients:
                    Value Std. Error   t value Pr(>|t|)
   (Intercept)     5.0331     0.1263     39.8634     0.0000
log10(1 + SV)     -1.9179     0.1395    -13.7457     0.0000
        I(RD^2)   -1.3889     0.3326     -4.1756     0.0001
     I(CP^0.5)    -0.0087     0.0070     -1.2417     0.2185

Residual standard error: 0.3858 on 70 degrees of freedom
Multiple R-Squared: 0.8442
F-statistic: 126.4 on 3 and 70 degrees of freedom, the p-value is 0

Correlation of Coefficients:
             (Intercept) log10(1 + SV) I(RD^2)
log10(1 + SV) -0.8506
        I(RD^2) -0.2268         0.0134
     I(CP^0.5)   0.3000        -0.6496          0.0270
> print(anova(lm2))
Analysis of Variance Table

Response: PSI

Terms added sequentially (first to last)
                Df Sum of Sq   Mean Sq   F Value        Pr(F)
log10(1 + SV)    1   53.65053 53.65053 360.5376 0.0000000
I(RD^2)          1    2.55490  2.55490  17.1692 0.0000943
I(CP^0.5)        1    0.22942  0.22942   1.5418 0.2185019
Residuals       70   10.41649  0.14881
> detach(2)
sink()
```