x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Table 7.1: Data set for simple linear regression

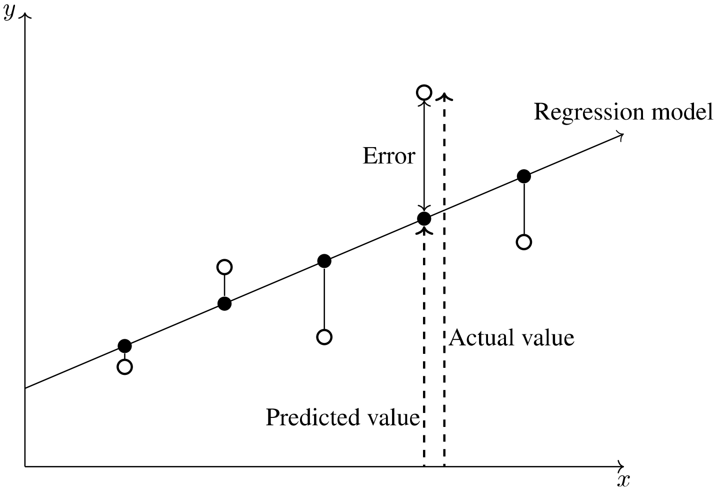

Figure 7.1: Errors in observed values

7.3 Simple linear regression

Let x be the independent predictor variable and y the dependent variable. Assume that we have a set of observed values of x and y:

A simple linear regression model defines the relationship between x and y using a line defined by an equation in the following form:

$$y = \alpha + \beta x$$

In order to determine the optimal estimates of α and β, an estimation method known as *Ordinary Least Squares* (OLS) is used.

The OLS method

In the OLS method, the values of y-intercept and slope are chosen such that they minimize the sum of the squared errors; that is, the sum of the squares of the vertical distance between the predicted y-value and the actual y-value (see Figure 7.1). Let $\hat{y_i}$ be the predicted value of y_i. Then the sum of squares of errors is given by

$$\begin{aligned} E &= \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \\ &= \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 \end{aligned}$$

So we are required to find the values of α and β such that E is minimum. Using methods of calculus, we can show that the values of a and b, which are respectively the values of α and β for which E is minimum, can be obtained by solving the following equations.

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Formulas to find a and b

Recall that the means of x and y are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$
$$\bar{y} = \frac{1}{n}\sum y_i$$

and also that the variance of x is given by

$$\text{Var}\,(x) = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2.$$

The *covariance of x and y*, denoted by Cov(x, y) is defined as

$$\text{Cov}(x,y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

It can be shown that the values of a and b can be computed using the following formulas:

$$b = \frac{\text{Cov}\,(x,y)}{\text{Var}\,(x)}$$
$$a = \bar{y} - b\bar{x}$$

Remarks

It is interesting to note why the least squares method discussed above is christened as "ordinary" least squares method. Several different variants of the least squares method have been developed over the years. For example, in the *weighted least squares* method, the coefficients a and b are estimated such that the weighted sum of squares of errors

$$E = \sum_{i=1}^{n} w_i [y_i - (a + bx_i)]^2,$$

for some positive constants $w_1, \ldots, w_n$, is minimum. There are also methods known by the names *generalised least squares* method, *partial least squares* method, *total least squares* method, etc. The reader may refer to *Wikipedia*, a free online encyclopedia, to obtain further information about these methods.

The OLS method has a long history. The method is usually credited to Carl Friedrich Gauss (1795), but it was first published by Adrien-Marie Legendre (1805).

Example

Obtain a linear regression for the data in Table 7.2 assuming that y is the independent variable.

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

Table 7.2: Example data for simple linear regression

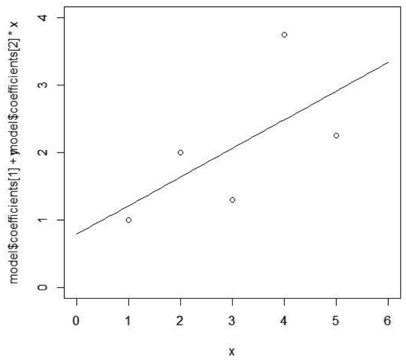

Figure 7.2: Regression model for Table 7.2

Solution

In the usual notations of simple linear regression, we have

$$\begin{aligned}
n &= 5\\
\bar{x} &= \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)\\
&= 3.0\\
\bar{y} &= \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)\\
&= 2.06\\
\text{Cov}\,(x, y) &= \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]\\
&= 1.0625\\
\text{Var}\,(x) &= \frac{1}{4}[(1.0 - 3.0)^2 + \cdots + (5.0 - 3.0)^2]\\
&= 2.5\\
b &= \frac{1.0625}{2.5}\\
&= 0.425\\
a &= 2.06 - 0.425 \times 3.0\\
&= 0.785
\end{aligned}$$

Therefore, the linear regression model for the data is

$$y = 0.785 + 0.425x. \tag{7.1}$$

Remark

Figure 7.2 in page 76 shows the data in Table 7.2 and the line given by Eq. (7.1). The figure was created using R.