8.5.2 The First Method for Finding β_0 and β_1

Here, we assume that x_i's are observed values of a random variable X. Therefore, we can summarize our model as

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is a $N(0, \sigma^2)$ random variable independent of X. First, we take expectation from both sides to obtain

$$\begin{aligned} EY &= \beta_0 + \beta_1 EX + E[\epsilon] \\ &= \beta_0 + \beta_1 EX \end{aligned}$$

Thus,

$$\beta_0 = EY - \beta_1 EX.$$

Next, we look at $\mathrm{Cov}(X, Y)$,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, \beta_0 + \beta_1 X + \epsilon) \\ &= \beta_0 \mathrm{Cov}(X, 1) + \beta_1 \mathrm{Cov}(X, X) + \mathrm{Cov}(X, \epsilon) \\ &= 0 + \beta_1 \mathrm{Cov}(X, X) + 0 \quad (\text{since } X \text{ and } \epsilon \text{ are independent}) \\ &= \beta_1 \mathrm{Var}(X). \end{aligned}$$

Therefore, we obtain

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \quad \beta_0 = EY - \beta_1 EX.$$

Now, we can find β_0 and β_1 if we know EX, EY, $\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$. Here, we have the observed pairs (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), so we may estimate these quantities. More specifically, we define

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n}, \\ \overline{y} &= \frac{y_1 + y_2 + \ldots + y_n}{n}, \\ s_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2, \\ s_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \end{aligned}$$

We can then estimate β_0 and β_1 as

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}},$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

The above formulas give us the regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

For each x_i, the **fitted value** $\hat{y}_i$ is obtained by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Here, $\hat{y}_i$ is the predicted value of y_i using the regression formula. The errors in this prediction are given by

$$e_i = y_i - \hat{y}_i,$$

which are called the **residuals**.

Simple Linear Regression

Given the observations (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), we can write the regression line as

$$\hat{y} = \beta_0 + \beta_1 x.$$

We can estimate β_0 and β_1 as

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}},$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

where

$$s_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2,$$
$$s_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

For each x_i, the **fitted value** $\hat{y}_i$ is obtained by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The quantities

$$e_i = y_i - \hat{y}_i$$

are called the **residuals**.

Example 8.31
Consider the following observed values of (x_i, y_i):

$$(1, 3) \quad (2, 4) \quad (3, 8) \quad (4, 9)$$

1. Find the estimated regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

based on the observed data.

2. For each x_i, compute the fitted value of y_i using

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

3. Compute the residuals, $e_i = y_i - \hat{y}_i$ and note that

$$\sum_{i=1}^{4} e_i = 0.$$

Solution

1. We have

$$\begin{aligned}
\overline{x} &= \frac{1+2+3+4}{4} = 2.5, \\
\overline{y} &= \frac{3+4+8+9}{4} = 6, \\
s_{xx} &= (1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2 = 5, \\
s_{xy} &= (1-2.5)(3-6) + (2-2.5)(4-6) + (3-2.5)(8-6) + (4-2.5)(9-6) = 11.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} = \frac{11}{5} = 2.2 \\
\hat{\beta}_0 &= 6 - (2.2)(2.5) = 0.5
\end{aligned}$$

2. The fitted values are given by

$$\hat{y}_i = 0.5 + 2.2 x_i,$$

so we obtain

$$\hat{y}_1 = 2.7, \quad \hat{y}_2 = 4.9, \quad \hat{y}_3 = 7.1, \quad \hat{y}_4 = 9.3$$

3. We have

$$\begin{aligned}
e_1 &= y_1 - \hat{y}_1 = 3 - 2.7 = 0.3, \\
e_2 &= y_2 - \hat{y}_2 = 4 - 4.9 = -0.9, \\
e_3 &= y_3 - \hat{y}_3 = 8 - 7.1 = 0.9, \\
e_4 &= y_4 - \hat{y}_4 = 9 - 9.3 = -0.3
\end{aligned}$$

So, $e_1 + e_2 + e_3 + e_4 = 0$.