Ch5. Simple Linear Regression

1 The model

The *simple linear regression* model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where y is the dependent or response variable and x is the independent or predictor variable. The designation *simple* indicates that there is only one predictor x, and *linear* means that the model 1 is linear in parameters β_0 and β_1. For the model, we assume that y_i and ϵ_i are random variables and that the values of x_i are known constants. In addition, we have the following three assumptions for the model:

1. $E(\epsilon_i) = 0$ for all $i = 1, 2, \cdots, n$, or, equivalently, $E(y_i) = \beta_0 + \beta_1 x_i$.

2. $var(\epsilon_i) = \sigma^2$ for all $i = 1, 2, \cdots, n$, or, equivalently, $var(y_i) = \sigma^2$.

3. $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, or, equivalently, $cov(y_i, y_j) = 0$.

2 Estimation of β_0, β_1 and σ^2

Given n observations $(x_1, y_1), \cdots, (x_n, y_n)$, the least squares approach seeks estimators β_0 and β_1 that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Note that $\hat{y}_i$ is an estimator of $E(y_i)$ instead of y_i. Differentiate $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$ and set the results equal to 0:

$$\frac{\partial \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{2}$$

$$\frac{\partial \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{3}$$

(4)

Solve the system (2), we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Note that the three model assumptions were not used in deriving the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. However, if the assumptions hold, the estimators are unbiased and have minimum variance among all linear unbiased estimators (this will be developed further in the latter chapters). Under the assumptions, we have

$$E(\hat{\beta}_0) = \beta_0,$$
$$E(\hat{\beta}_1) = \beta_1,$$
$$var(\hat{\beta}_0) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}],$$
$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The method of least squares does not yield an estimator of σ, the variance can be estimated by

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}.$$

This estimator is unbiased,

$$E(s^2) = \sigma^2.$$

The deviation $\hat{\epsilon} = y_i - \hat{y}_i$ is often called the *residual* of y_i, and SSE is called the residual sum of squares or error sum of squares.

3 Hypothesis Test and Confidence Interval for β_1

Linear relationship testing:

$$H_0: \quad \beta_1 = 0.$$

In order to obtain a test for $H_0 : \beta_1 = 0$, we make a further assumption (normality assumption) that y_i is $N(\beta_0 + \beta_1 x_i, \sigma^2)$. Then we have

1. $\hat{\beta}_1$ is $N(\beta_1, \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2)$.

2. $(n-2)s^2/\sigma^2$ is $\chi^2(n-2)$.

3. $\hat{\beta}_1$ and s^2 are independent.

(These properties will be further developed in the next chapter.)

The three properties follows that

$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_i (x_i - \bar{x})^2}}$$

is distributed as $t(n-2, \delta)$, the noncentral t with noncentrality parameter δ,

$$\delta = \frac{E(\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} = \frac{\beta_1}{\sigma/\sqrt{\sum_i (x_i - \bar{x})^2}}.$$

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

4 Coefficient of Determination

The *coefficient of determination* r^2 is defined as

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the regression sum of squares and $SST = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the total sum of squares. The total sum of squares can be partitioned into SST=SSR+SSE, i.e.,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Thus r^2 gives the proportion of variation in y that is explained by the model, or, equivalently, accounted for by regression on x.

Note that r^2 is the same as the square of the *sample correlation coefficient* r between y and x,

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $s_{xy} = [\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})]/(n-1)$.