Lecture Summary: Jan. 11, 2023

• Simple linear regression model: $Y = \beta_0 + \beta_1 x + \epsilon$, where y is the response (also called dependent variable), x is a predictor (also called independent variable), β_0 and β_1 are unknown constants called intercept and slope, respectively, and ϵ is a random error.

The data are collected in pairs: $(x_1, Y_1), \ldots, (x_n, Y_n)$, where n is the sample size. All pairs of data satisfy the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

• Assumptions:

(i) $\epsilon_1, \ldots, \epsilon_n$ are independent;

(ii) $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$, where σ^2 is an unknown constant;

(iii) (normality assumption): normality is often (but not always) assumed when making inference, that is, ϵ_i is normally distributed.

Under Assumptions (i)–(iii), Y_i is normally distributed with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. A plot of the distributions of Y_i, for different x_i, is presented on page 11 of the text book (Figure 1.6).

• Least squares (LS) estimates: The intercept and slope of the simple linear regression are estimated by minimizing

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

(to find the optimal β_0, β_1).

- Geometric interpretation of least squares: Projection.

- Formulae for LS estimates:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x},
\end{aligned}$$

where $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ (sample mean of the x_i's) and $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ (sample mean of the Y_i's).

- Computing the LS estimates

- Regression line: $y = \hat{\beta}_0 + \hat{\beta}_1 x$. It is the line that is closest to the scatter points in an overall sense.

- Residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$. The residual $\hat{\epsilon}_i$ may be viewed as an estimate of the rergession error ϵ_i.