Notes 2

Maryclare Griffin

2/9/2023

These notes are based on Chapters 1 and 6 of KNNL.

The linear regression model for a dependent variable or response Y and independent variables, predictors, or covariates $X_1, \dots X_{p-1}$ is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where:

- $\beta_0, \beta_1, \dots, \beta_{p-1}$ are parameters
- $X_{i1}, \dots, X_{i,p-1}$ are known constants
- ϵ_i is a random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = \sigma^2$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\epsilon_i, \epsilon_j\} = 0$ for all i, j; $i \neq j$)
- $i = 1, \dots, n$

  Note: When we just have one independent variable or predictor ($p = 2$) and we will call this a **simple** linear regression model. When we have more than one predictor, we will call this a **multiple** linear regression model.

We call this a **linear** regression model because it is linear in the parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$.

This model has several important features:

- The response Y_i in the i-th trial is the sum of two components: (1) the constant term $\beta_0 + \sum_{k=1}^{p} \beta_k X_{ik}$ and (2) the random term ϵ_i. Hence, Y_i is a random variable.

$$Y_i = \underbrace{\beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}}_{(1)} + \overbrace{\epsilon_i}^{(2)}$$

- Since $E\{\epsilon_i\} = 0$, it follows from properties of the expected value that:

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

- The response Y_i exceeds or falls short of the regression function $E\{Y_i\}$ by the error term amount ϵ_i.
- The error terms ϵ_i are assumed to have constant variance σ^2. It then therefore follows that the responses Y_i have the same constant variance:

$$\sigma^2\{Y_i\} = \sigma^2.$$

  Thus, the regression model assumes that the probability distributions of Y have the same variance σ^2, regardless of the level of the predictor variable X.
- The error terms are assumed to be uncorrelated. Since the error terms ϵ_i and ϵ_j are uncorrelated, so are the responses Y_i and Y_j.

To summarize, the regression model implies that the responses Y_i come from probability distributions whose means are $E\{Y_i\} = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik}$ and whose variances are σ^2, the same for all levels of X. Further, any two responses Y_i and Y_j are uncorrelated.