April 4 – The linear model

1. Review:
   (a) Given: bivariate data: \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).
   (b) Problem: fit a line \(y = a + bx\) to the data.
   (c) Solution: Choose \(a\) and \(b\) to minimize the sum of the squared residuals. That is
      \[
      SS_{\text{Resid}} = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \sum_{i=1}^{n} e_i^2
      \]
      (d) A measure of fit: \(R^2\):
      \[
      R^2 = \frac{SS_{\text{Regress}}}{SS_{\text{Tot}}}
      \]
      \[
      SS_{\text{TOT}} = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad SS_{\text{Regress}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
      \]

2. A statistical model.
   \[
   y = \alpha + \beta x + \epsilon
   \]
   where
   (a) \(\epsilon\) is a random variable with mean 0 and variance \(\sigma^2\)
   (b) \(\alpha, \beta, \sigma^2\) are (unknown) parameters
   (c) Additionally, when we want to write confidence intervals, \(\epsilon\) is normal

3. The data \((x_1, y_1), \ldots, (x_n, y_n)\) are assumed to arise by choosing a random sample \(\epsilon_1, \ldots, \epsilon_n\). The random variables \(\epsilon_1, \ldots, \epsilon_n\) are assumed to be independent and all have the same distribution as \(\epsilon\). (Note that this really makes \(y_i\) a random variable but \(x_i\) is not.) We can write
   \[
   \mu_{Y|x} = \alpha + \beta x \quad \sigma^2_{Y|x} = \sigma^2
   \]

4. This model is just a generalization of one that we have seen before: we could view a random sample \(y_1, \ldots, y_n\) as arising from the model
   \[
   y = \mu + \epsilon
   \]

Homework

Read Devore and Farnum, Section 11.1, pages 488–492. No problems!