CHAPTER 4

Fitting a Line

4.0 What We Need to Know When We Finish This Chapter

This chapter develops a simple method to measure the *magnitude of the association* between two variables in a sample. The generic name for this method is *regression analysis*. The precise name, in the case of only two variables, is *bivariate regression*. It assumes that the variable X *causes* the variable Y. It identifies the *best-fitting line* as that which *minimizes the sum of squared errors* in the Y dimension. The quality of this fit is measured informally by *the proportion of the variance in Y that is explained by the variance in X*. Here are the essentials.

1. **Equation (4.1), section 4.3:** The regression line predicts y_i as a linear function of x_i:

   $$\hat{y}_i = a + bx_i.$$

2. **Equation (4.2), section 4.3:** The regression error is the difference between the actual value of y_i and the value predicted by the regression line:

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

3. **Equation (4.20), section 4.3:** The average error for the regression line is equal to zero:

   $$\bar{e} = 0.$$

4. **Equation (4.28), section 4.3:** The errors are uncorrelated with the explanatory variable:

   $$\text{CORR}\left(e, X\right) = 0.$$

5. **Equation (4.35), section 4.4:** The regression intercept is the difference between the average value of Y and the slope times the average value of X:

   $$a = \bar{y} - b\bar{x}.$$

6. **Equation (4.40), section 4.5:** The slope is a function of only the observed values of x_i and y_i in the sample:

   $$b = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)x_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)x_i} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

7. **Equation (4.57), section 4.6:** The R^2 measures the strength of the association represented by the regression line:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

8. **Equations (4.58) and (4.59), section 4.6:** The R^2 in the bivariate regression is equal to the squared correlation between X and Y and to the squared correlation between Y and its predicted values:

$$R^2 = \left(\mathrm{CORR}(X, Y)\right)^2 = \left(\mathrm{CORR}(Y, \hat{Y})\right)^2.$$