CHAPTER 5

FROM SAMPLE TO POPULATION

5.0 What We Need to Know When We Finish This Chapter

This chapter discusses the reasons why we can *generalize* from what we observe in one sample to what we might expect in others. It defines the *population*, distinguishes between *parameters* and *estimators*, and discusses why we work with samples when populations are what we are interested in: *The sample is an example.* It demonstrates that, with the appropriate assumptions about the structure of the population, a and b, as calculated in chapter 4, are best linear unbiased (BLU) estimators of the corresponding parameters in the population relationship. Under these population assumptions, regression is frequently known as *ordinary least squares (OLS) regression*. Here are the essentials.

1. **Section 5.1:** We work with samples, even though populations are what we are interested in, because samples are available to us and populations are not. Our intent is to generalize from the sample that we see to the population that we don't.
2. **Equation (5.1), section 5.2:** The population relationship between x_i and y_i is

   $$y_i = \alpha + \beta x_i + \varepsilon_i.$$

3. **Equation (5.5), section 5.3:** Our first assumption about the disturbances is that their expected values are zero:

   $$\mathrm{E}(\varepsilon_i) = 0.$$

4. **Equation (5.16), section 5.4:** The deterministic part of y_i is

   $$\mathrm{E}(y_i) = \alpha + \beta x_i.$$

5. **Equation (5.6), section 5.3, and equation (5.20), section 5.4:** Our second assumption about the disturbances is that their variances are the same. Their variances are equal to those of the dependent variable:

   $$\mathrm{V}(\varepsilon_i) = \mathrm{V}(y_i) = \sigma^2.$$

6. **Equation (5.11), section 5.3, and equation (5.22), section 5.4:** Our third assumption about the disturbances is that they are uncorrelated in the population. This implies that the values of the dependent variable are uncorrelated in the population:

   $$\mathrm{COV}(\varepsilon_i, \varepsilon_j) = \mathrm{COV}(y_i, y_j) = 0.$$

7. **Equation (5.37), section 5.6:** Our slope, b, is an unbiased estimator of the population coefficient, β:

   $$\mathrm{E}(b) = \beta.$$

8. **Equation (5.42), section 5.6:** Our intercept, a, is an unbiased estimator of the population constant, α:

$$E(a) = \alpha.$$

9. **Section 5.7:** A *simulation* is an analysis based on artificial data that we construct from parameter values that we choose.
10. **Equation (5.50), section 5.8:** The variance of b is

$$V(b) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

11. **Equation (5.51), section 5.8:** The variance of a is

$$V(a) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

12. **Equation (5.60), section 5.9:** According to the Gauss-Markov theorem, no linear unbiased estimator of β is more precise than b. In other words, the variance of any other linear unbiased estimator of β, d, is at least as large as that of b:

$$V(d) \geq V(b).$$

13. **Section 5.10:** Consistent estimators get better as the amount of available information increases. With enough information, consistent estimators are perfect.