CHAPTER 11

What If There Is More Than One *X*?

11.0 What We Need to Know When We Finish This Chapter

If the population relationship includes two explanatory variables, but our sample regression contains only one, our estimate of the effect of the included variable is almost surely biased. The best remedy is to include the omitted variable in the sample regression. Minimizing the sum of squared errors from a regression with two explanatory variables yields two slopes, each of which represents the relationship between the parts of the dependent variable and the associated explanatory variable that are not related to the other explanatory variable. These slopes are unbiased estimators of the population coefficients. Here are the essentials.

1. **Equation (11.1), section 11.2:** If there are two explanatory variables that affect y_i, the population relationship is

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

2. **Equations (11.6), (11.10), and (11.11), section 11.2:** If the population relationship is equation (11.1), but we still run the regression of equation (4.4), the slope of that regression is a biased estimator of the true effect of x_{1i} on y_i, β_1:

$$\mathrm{E}(b) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)x_{21i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)x_{1i}} = \beta_1 + \beta_2 b_{x_2x_1} \neq \beta_1.$$

   This bias is usually referred to as *specification error*, *omitted-variable bias*, or *left-out-variable error* (LOVE). The regression of equation (4.4) mistakenly attributes some of β_2, the effect of the omitted variable x_{2i}, to the included variable, x_{1i}. The extent of this mistaken attribution is determined by the extent to which x_{2i} looks like x_{1i}.
3. **Section 11.2:** If the omitted variable is not available but a suitable instrument, z_i, is, LOVE can be fixed. The instrumental variables (IV) strategy is often referred to as addressing *unobserved heterogeneity*. If an instrument is not available, it may be possible to *sign* the bias and determine whether b is an over- or underestimate of β_1.
4. **Equations (11.12), (11.19), (11.24), and (11.27), section 11.3:** If we minimize the sum of squared errors for the multivariate sample relationship,

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i,$$

   we get errors that have an average value of zero and are unrelated, at least linearly, to either of the two explanatory variables.
5. **Equation (11.65), section 11.4:** Regression estimates the effect of x_{1i} on y_i as equivalent to

$$b_1 = \frac{\sum_{i=1}^{n}\left(e_{(x_1x_2)i} - \bar{e}_{(x_1x_2)}\right)\left(e_{(yx_2)i} - \bar{e}_{(yx_2)}\right)}{\sum_{i=1}^{n}\left(e_{(x_1x_2)i} - \bar{e}_{(x_1x_2)}\right)^2},$$

the effect of the part of x_{1i} that is not related to x_{2i} on the part of y_i that is not related to x_{2i}. This is why we can interpret a regression slope as measuring the effect of an explanatory variable ceteris paribus, holding constant all other explanatory variables. Analogously, regression estimates the effect of x_{2i} on y_i as equivalent to the effect of the part of x_{2i} that is not related to x_{1i} on the part of y_i that is not related to x_{1i}.

6. **Equation (11.85), section 11.5, and exercises 11.16 and 11.17:** The slope and intercept estimators from the regression of equation (11.12) are unbiased estimators of the coefficients and constant in the population relationship of equation (11.1): $E(b_1) = \beta_1$, $E(b_2) = \beta_2$, and $E(a) = \alpha$.