CHAPTER 10

What If the Disturbances and the Explanatory Variables Are Related?

10.0 What We Need to Know When We Finish This Chapter

This chapter deals with the possibility that the disturbances and the explanatory variables are related. This problem is called *endogeneity* or *simultaneity*. If we have it, our ordinary least squares (OLS) estimators are biased and inconsistent. Worse, they are unsalvageable. As in the previous two chapters, the solution can be thought of as a two-step procedure consisting of two applications of OLS. For this reason, it is often called *two-stage least squares*. It is also called, more generally, *instrumental variables*. This technique provides estimators that are not unbiased, only consistent. Therefore, the success of the solution in this chapter is much more sensitive to the size of the sample than it was in the previous two chapters. Here are the essentials.

1. **Section 10.2:** The three most common sources of endogeneity are *reverse causality*, *measurement error* in x_i, and *dynamic choice*.
2. **Equation (10.12), section 10.2:** Endogeneity means that the explanatory variable is a random variable and it's correlated with the disturbances:

   $$\text{COV}\left(x_i, \varepsilon_i\right) \neq 0.$$

3. **Equation (10.13), section 10.3:** The consequence of endogeneity is that b is neither unbiased nor consistent for β:

   $$\text{E}\left(b\right) = \beta + \text{E}\left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\varepsilon_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)x_i}\right) \neq \beta.$$

4. **Equation (10.18), section 10.3:** With measurement error, b is approximately equal to

   $$b \to \beta \frac{\text{V}\left(x_i^*\right)}{\text{V}\left(x_i^*\right) + \text{V}\left(v_i\right)} < \beta.$$

   OLS tends to understate the true magnitude of β.
5. **Equations (10.19) and (10.20), section 10.4:** An *instrument*, or an *instrumental variable*, z_i, has, roughly speaking, the following properties, at least as the sample size approaches infinity:

   $$\text{COV}\left(x_i, z_i\right) \neq 0$$

   and

   $$\text{COV}\left(\varepsilon_i, z_i\right) = 0.$$

6. **Equation (10.21), section 10.4:** The first step of our two-stage procedure is the OLS *instrumenting equation*

$$x_i = c + dz_i + f_i.$$

It provides an estimator of x_i, $\hat{x}_i$, that is purged of the correlation between x_i and ε_i.

7. **Equation (10.23), section 10.4:** We obtain the *two-stage least squares* (2SLS) estimators of α and β through OLS estimation of our second-step equation,

$$y_i = a_{2\text{SLS}} + b_{2\text{SLS}}\hat{x}_i + e_i.$$

8. **Equation (10.24), section 10.4, and equations (10.30) and (10.32), section 10.5:** The instrumental variables (IV) estimator of β is the same as the 2SLS estimator,

$$b_{\text{IV}} = b_{2\text{SLS}} = \frac{\sum_{i=1}^{n}\left(\hat{x}_i - \overline{\hat{x}}\right)y_i}{\sum_{i=1}^{n}\left(\hat{x}_i - \overline{\hat{x}}\right)\hat{x}_i} = \frac{\sum_{i=1}^{n}\left(z_i - \overline{z}\right)y_i}{\sum_{i=1}^{n}\left(z_i - \overline{z}\right)x_i}.$$

9. **Section 10.6:** The IV estimator b_{IV} is a consistent estimator of β.

10. **Equation (10.41), section 10.6, and equation (10.45), section 10.7:** The estimated variance of the IV slope estimator is

$$\text{V}\left(b_{\text{IV}}\right) = \frac{s_{\text{IV}}^2 \sum_{i=1}^{n}\left(z_i - \overline{z}\right)^2}{\left(\sum_{i=1}^{n}\left(z_i - \overline{z}\right)x_i\right)^2} = \frac{\text{V}(b)}{\left[\text{CORR}\left(z_i, x_i\right)\right]^2}.$$

11. **Section 10.7:** It's often difficult to find an appropriate instrument. The more likely it is that z_i satisfies one of the assumptions in equations (10.19) and (10.20), the less likely it is that it satisfies the other. The best instruments are moderately correlated with x_i, in order to satisfy equation (10.19) without invalidating equation (10.20). The *Staiger-Stock rule of thumb* stipulates that the *F*-statistic for the first-stage regression should exceed 10 in order for b_{IV} to be useful.

12. **Equation (10.47), section 10.8:** The *Hausman test* for endogeneity consists of the auxiliary regression

    $$y_i = a + b_1 x_i + b_2 \hat{x}_i + e_i.$$

    Endogeneity is present if the coefficient on $\hat{x}_i$, b_2 is statistically significant.
13. **Section 10.9:** It's always necessary to have a behavioral argument as to why an instrument might be appropriate. It's usually possible to offer a counterargument as to why it might not be. Great instruments are hard to come by. Even acceptable instruments often require considerable ingenuity to construct and justify.