Chapter 6

Causal Inference and the Sources of Endogeneity

6.1 Introduction

Causal inference in regression analysis is concerned with estimating the effect of an explanatory variable X on an outcome Y in a way that reflects a true causal effect rather than a spurious correlation. The ideal scenario for causal inference is a randomized experiment, where the explanatory variable is assigned independently of any other factors. In such an experiment, any observed relationship between X and Y can be interpreted causally because X is exogenous by design. In observational studies, however, we must rely on assumptions to ensure that our estimated relationship is not driven by confounding factors.

In the context of ordinary least squares (OLS) regression, a key requirement for a causal interpretation is the exogeneity of the regressors. This means that each explanatory variable must be uncorrelated with the error term in the regression equation. Equivalently, the error term should have zero conditional mean given the regressors. We state this formally:

Assumption 6.1 (Exogeneity (Zero Conditional Mean))**.** For the true regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i$, the error term satisfies $\mathbb{E}[u_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}] = 0$.

Assumption 6.1 is one of the central requirements for OLS to identify causal effects. If this assumption holds, OLS estimators are unbiased and consistent for the true parameters.[1] However, if Assumption 6.1 is violated, then one or more regressors are *endogenous*, meaning they are correlated with the error term. In that case, the OLS estimates will generally be biased and inconsistent, leading us away from the true causal effect.

There are several common sources of endogeneity in regression models. In this chapter, we examine three key sources:

1. **Omitted variable bias** – Important explanatory factors are left out of the model.

2. **Simultaneity (reverse causality)** – One of the regressors is jointly determined with the dependent variable.

3. **Measurement error** – A regressor is measured with error, contaminating the regression with noise.

All of these scenarios result in a violation of exogeneity and thus bias the OLS coefficient estimates. In the following sections, we discuss each source of endogeneity in detail, providing intuitive examples and formal derivations where appropriate. Our goal is to understand *why* these problems lead to biased estimates and how to recognize them in practice. (Methods for addressing these issues—such as instrumental variables or experimental designs—will be touched upon briefly and covered more fully in subsequent chapters.)

6.2 Omitted Variable Bias

One of the most prevalent sources of endogeneity is **omitted variable bias (OVB)**. This bias occurs when a relevant variable that influences the dependent variable is omitted from the regression, and that omitted variable

[1] We assume throughout that all other standard regression assumptions (linearity of the model, random sampling, no perfect multicollinearity, etc.) hold as well. The focus here is on violations of exogeneity, which is often the most critical assumption for causal interpretation.

is correlated with one or more included regressors. In such cases, the coefficient on the included regressor will capture not only its own direct effect on the outcome, but also the indirect effect that operates through the omitted factor. Intuitively, the regression is attributing to X some of the effect that actually belongs to the omitted variable, since X is serving as a proxy for that omitted factor in the model.

OVB can arise if two conditions are met:

1. The omitted variable is a *determinant of the dependent variable* (i.e. it truly belongs in the regression model for Y).

2. The omitted variable is *correlated with at least one included regressor*.

If an omitted factor does not affect Y, then leaving it out does not bias the estimation of other coefficients (it would simply inflate the error variance). Likewise, if the omitted factor is unrelated to all included X variables, its omission will not violate exogeneity (because the omitted factor, though in the error term, would be uncorrelated with X by construction). Only when both conditions hold do we get omitted variable bias.

To formalize the idea, consider a true model with two regressors X and Z:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u, \tag{6.1}$$

where u is an error term with $\mathbb{E}[u \mid X, Z] = 0$. Here Z represents an important variable determining Y alongside X. Suppose we erroneously omit Z and estimate a regression of Y on X alone:

$$Y = \alpha_0 + \alpha_1 X + e, \tag{6.2}$$

where e is the new error term in this misspecified regression. The omitted variable Z is now part of the error: specifically, comparing (6.2) with the true model (6.1), we have

$$e = \beta_2 Z + u.$$

If Z is correlated with X, then X will be correlated with e (since Z resides in e), violating Assumption 6.1. As a result, the OLS estimator $\hat{\alpha}_1$ in the omitted model (6.2) will be biased.

We can derive an expression for this bias. The OLS estimator from (6.2) satisfies

$$\hat{\alpha}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)},$$

the ratio of the sample covariance between X and Y to the variance of X. Consider the covariance term using the true model (6.1):

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(X,\ \beta_0 + \beta_1 X + \beta_2 Z + u).$$

Since β_0 is a constant and $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$, this expands to

$$\operatorname{Cov}(X, Y) = \beta_1 \operatorname{Var}(X) + \beta_2 \operatorname{Cov}(X, Z) + \operatorname{Cov}(X, u).$$

By the exogeneity of the true model, $\operatorname{Cov}(X, u) = 0$. Therefore,

$$\operatorname{Cov}(X, Y) = \beta_1 \operatorname{Var}(X) + \beta_2 \operatorname{Cov}(X, Z).$$

Plugging this into the formula for $\hat{\alpha}_1$ and simplifying gives:

$$\hat{\alpha}_1 = \beta_1 + \beta_2 \frac{\operatorname{Cov}(X, Z)}{\operatorname{Var}(X)}.$$

We summarize this result formally:

Proposition 6.2 (Omitted Variable Bias)**.** *Under the true model* (6.1)*, if the regressor Z is omitted and we estimate* (6.2)*, then*

$$\textit{plim}\ \hat{\alpha}_1 = \beta_1 + \beta_2 \frac{\operatorname{Cov}(X, Z)}{\operatorname{Var}(X)},$$

i.e. the probability limit of the estimator includes an asymptotic bias term $\beta_2 \frac{\operatorname{Cov}(X,Z)}{\operatorname{Var}(X)}$.[2]

Proof. Using the true model (6.1), the population covariance between X and Y is

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \operatorname{Cov}\big(X,\ \beta_0 + \beta_1 X + \beta_2 Z + u\big) \\
&= \beta_1 \operatorname{Cov}(X, X) + \beta_2 \operatorname{Cov}(X, Z) + \operatorname{Cov}(X, u) \\
&= \beta_1 \operatorname{Var}(X) + \beta_2 \operatorname{Cov}(X, Z),
\end{aligned}$$

[2]We use the term "probability limit" (plim) to denote the value to which an estimator converges as the sample size grows large (consistency). If the estimator is unbiased and consistent, its plim equals the true parameter. Under endogeneity, the plim will generally be different from the true parameter.

since $\text{Cov}(X, u) = 0$ by exogeneity of the true model. Dividing both sides by $\text{Var}(X)$ yields

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta_1 + \beta_2 \frac{\text{Cov}(X, Z)}{\text{Var}(X)}.$$

The left-hand side is the probability limit of the OLS estimator $\hat{\alpha}_1$ from the misspecified regression (as $n \to \infty$), and the right-hand side is β_1 plus the bias term. □

The second term in the above expression, $\beta_2 \frac{\text{Cov}(X, Z)}{\text{Var}(X)}$, represents the bias induced by omitting Z. From this formula we can determine the direction of the bias in different scenarios:

- If X and the omitted variable Z are positively correlated ($\text{Cov}(X, Z) > 0$) and Z has a positive effect on Y ($\beta_2 > 0$), then the bias term is positive. In this case, $\hat{\alpha}_1$ will overestimate β_1 (an upward bias).

- If X and Z are positively correlated but Z has a negative effect on Y ($\beta_2 < 0$), the bias term is negative. In this scenario, $\hat{\alpha}_1$ will underestimate β_1.

- In general, the sign of the bias is $\text{sign}(\beta_2) \times \text{sign}(\text{Cov}(X, Z))$. If X and Z are negatively correlated, the reasoning reverses.

- If either $\beta_2 = 0$ (the omitted variable has no true effect on Y) or $\text{Cov}(X, Z) = 0$ (the omitted variable is uncorrelated with X), then the bias term is zero. This is consistent with our earlier statement that both conditions must hold for OVB to occur.

It is worth noting that omitted variable bias can sometimes be severe enough to even flip the sign of an estimated coefficient relative to its true value. For example, if $\beta_1 > 0$ (a positive true effect of X on Y) but $\beta_2 < 0$ and $\text{Cov}(X, Z)$ is sufficiently large in magnitude, the term $\beta_2 \frac{\text{Cov}(X,Z)}{\text{Var}(X)}$ could be more negative than β_1 is positive, resulting in $\hat{\alpha}_1$ that is negative. In that case, the regression would misleadingly suggest a negative relationship between X and Y even though the true causal effect of X is positive. This underscores how dangerous omitted confounders can be for causal inference.

Example 6.3 (Education, Ability, and Wages)**.** A classic example of omitted variable bias arises in estimating the returns to education on earnings. Suppose we regress individuals' wage (or log wage) on their years of education. The coefficient on education in such a regression is meant to capture the causal effect of schooling on earnings (e.g. the percentage increase in wages per additional year of education). However, not everyone pursues education to the same extent, and one important factor is *ability*. More able or talented individuals may both achieve higher education and earn higher wages. If we omit innate ability from the wage regression (perhaps because it is hard to measure), the schooling variable X will partly proxy for ability. Here Z (the omitted variable) is ability, which likely satisfies both conditions for OVB: (1) ability clearly affects wages (more able workers are more productive and thus command higher pay), and (2) ability is positively correlated with education (more able individuals tend to obtain more schooling, due to factors like better academic performance or more opportunities).

In this scenario, the omission of ability biases the estimated coefficient on education upward. Education appears to have a larger effect on wages than it truly does, because the regression is attributing the wage gains due to ability to the additional schooling. This is sometimes referred to as *ability bias* in the returns-to-education literature. If we could measure and include ability in the regression, we expect the coefficient on education would decrease (perhaps substantially). In practice, researchers attempt to address this OVB problem by finding proxies for ability or using methods like twin studies and instrumental variables (for example, using variations in schooling caused by differences in school access or laws) to isolate the true effect of education on earnings.

The education example illustrates the threat that an omitted confounding factor poses: it undermines the causal interpretation of the regression coefficient. Whenever we run a regression, we should carefully think about what factors might have been omitted and whether they could induce correlation between our regressors and the error term. If so, OLS estimates will not recover the true causal parameters.

6.3 Measurement Error

Another major source of endogeneity is **measurement error** in the regressors. Measurement error occurs when the variable we use in the regression is measured or recorded with noise relative to the true quantity we want to capture. For instance, a survey might ask respondents for their income or debt levels, but respondents could misreport these either intentionally or due to forgetfulness. In economic data, key variables like wealth, earnings, or leverage can be measured with error. When the mismeasured variable is used as an X in regression, the estimation can suffer from what is known as *errors-in-variables* bias.

Not all measurement error causes bias. If the error is in the dependent variable Y (for example, if Y is self-reported and measured with noise), the OLS coefficients remain unbiased—such measurement error in Y only adds extra randomness to the model and typically inflates the residual variance (making estimates less precise) but does not bias the coefficients.[3] The more problematic case is measurement error in an independent variable X.

For simplicity, consider a single regressor model $Y = \beta_0 + \beta_1 X^* + u$, where X^* is the true, unobserved regressor and u is the error term (assume $\mathbb{E}[u \mid X^*] = 0$ so that X^* itself satisfies exogeneity). We do not observe X^* directly; instead, we observe $X = X^* + w$, where w is a measurement error. We assume *classical measurement error*: w is independent of X^* and u, with mean zero (so the observed X is an unbiased measure of the true X^*). Now consider the regression of Y on X (the observed, error-ridden regressor). Substituting $X^* = X - w$ into the true model:

$$Y = \beta_0 + \beta_1(X - w) + u = \beta_0 + \beta_1 X + (u - \beta_1 w).$$

The regression we actually estimate is

$$Y = \beta_0 + \tilde{\beta}_1 X + \tilde{u},$$

where $\tilde{u} = u - \beta_1 w$ is the composite error term. Importantly, this new error term $\tilde{u}$ includes the term $-\beta_1 w$, and w is part of X. Even though w is

[3] Classical measurement error in Y increases the variance of the error term but does not violate the zero conditional mean assumption, since the mismeasurement is part of the error term and is independent of X. However, non-classical measurement error in Y (e.g. if the error is systematically related to X) could introduce bias.

independent of X^* and u, it is *not* independent of X (because X contains w). In fact, $\text{Cov}(X, \tilde{u}) = \text{Cov}(X, -\beta_1 w + u) = -\beta_1 \text{Cov}(X, w) + \text{Cov}(X, u)$. Now, $\text{Cov}(X, u) = 0$ by assumption (since X^* and u are independent and $X = X^* + w$). But $\text{Cov}(X, w) = \text{Cov}(X^* + w,\ w) = \text{Cov}(X^*, w) + \text{Var}(w) = 0 + \text{Var}(w)$, since X^* and w are independent. Thus $\text{Cov}(X, \tilde{u}) = -\beta_1 \text{Var}(w)$, which is generally nonzero (unless $\beta_1 = 0$). This means X is endogenous in the observed regression, and OLS will be biased.

We can derive the attenuation effect explicitly. The population OLS estimator with X measured with error is:

$$\tilde{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

Using $Y = \beta_0 + \beta_1 X^* + u$ and $X = X^* + w$, we find:

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X^* + w,\ \beta_1 X^* + u) \\ &= \beta_1 \text{Cov}(X^*, X^*) + \beta_1 \text{Cov}(w, X^*) + \text{Cov}(X^*, u) + \text{Cov}(w, u). \end{aligned}$$

Given our assumptions: $\text{Cov}(w, X^*) = 0$, $\text{Cov}(X^*, u) = 0$, and $\text{Cov}(w, u) = 0$. Therefore, $\text{Cov}(X, Y) = \beta_1 \text{Var}(X^*)$. Meanwhile, the variance of X is

$$\text{Var}(X) = \text{Var}(X^* + w) = \text{Var}(X^*) + \text{Var}(w),$$

since X^* and w are independent. Thus:

$$\tilde{\beta}_1 = \frac{\beta_1 \text{Var}(X^*)}{\text{Var}(X^*) + \text{Var}(w)} = \beta_1 \times \frac{\text{Var}(X^*)}{\text{Var}(X^*) + \text{Var}(w)}.$$

Because $\text{Var}(w) > 0$, the fraction $\frac{\text{Var}(X^*)}{\text{Var}(X^*)+\text{Var}(w)}$ is less than 1. Denoting $\lambda = \frac{\text{Var}(X^*)}{\text{Var}(X^*)+\text{Var}(w)}$ (sometimes called the reliability ratio of the measured variable), we have $\tilde{\beta}_1 = \lambda \beta_1$. In expectation, the estimated coefficient is attenuated (shrunk towards zero) by the factor λ. In the limit of large samples,

$$plim\ \tilde{\beta}_1 = \lambda \beta_1,$$

so unless $\text{Var}(w) = 0$ (no measurement error) or $\beta_1 = 0$ (the true effect is zero), $plim \tilde{\beta}_1 \neq \beta_1$. Typically λ is significantly below 1 when measurement error is substantial, meaning the regression understates the true effect. This phenomenon is known as **attenuation bias**.

In summary, classical measurement error in a regressor biases the estimated coefficient toward zero. A positive true effect will be underestimated (closer to zero), and a negative true effect will also be underestimated in magnitude (closer to zero from below). The severity of the bias depends on how noisy the measurement is: if $\text{Var}(w)$ is large relative to $\text{Var}(X^*)$, then λ is small and the bias is severe. If the measurement is fairly precise (small $\text{Var}(w)$), λ is closer to 1 and the bias is mild. It is important to note that this derivation assumes the measurement error is random (classical). In reality, some measurement errors are systematic or correlated with other variables ("non-classical" measurement error), which can lead to more complex forms of bias that do not necessarily attenuate toward zero.[4]

Example 6.4 (Measurement Error in Leverage)**.** Researchers in finance often study the effect of a firm's leverage (debt-to-equity ratio or a similar metric) on outcomes such as its stock returns, risk, or investment behavior. Leverage is typically measured from balance sheet data (book leverage) or market data (market leverage), but both measures can be prone to error or may not perfectly capture the concept of true leverage at every moment. Suppose a researcher regresses a measure of firm performance (say, stock return) on the firm's leverage ratio. If the leverage data contains measurement error (for instance, due to accounting differences, reporting lags, or approximation in using book values), the estimated coefficient on leverage will likely suffer from attenuation bias.

In practice, this means the regression might find only a weak or insignificant relationship between leverage and performance, even if the true effect of leverage is substantial. The noise in the leverage variable dilutes the signal. Some studies indeed attribute the difficulty in detecting strong leverage effects to measurement error. For example, if true leverage changes are not fully captured in the reported data, or if firms' off-balance-sheet debts are not counted, the observed leverage is a noisy proxy for true leverage. The resulting coefficient estimate is biased toward zero. Researchers must be cautious interpreting such results: a lack of a significant coefficient on leverage could be due to measurement error rather than a truly negligible effect. In response, they might seek instruments for leverage or use techniques like

[4]For example, if less educated respondents systematically misreport their income more than highly educated respondents, the measurement error in income would be correlated with education, violating the classical assumption and potentially biasing the estimated effect of education on income in unpredictable ways.

averaging over time (to reduce noise) as remedies for the errors-in-variables problem.

This example highlights that measurement error can mask real economic relationships. Whenever we suspect that a key regressor is measured with error, we should be wary of coefficient estimates that are surprisingly small or insignificant. Correcting for measurement error often requires additional information, such as validation datasets or instrumental variables, which provide an external source of variation in the mismeasured variable.

6.4 Simultaneity (Reverse Causality)

The third major source of endogeneity we consider is **simultaneity bias**, also known as reverse causality or joint determination of variables. Simultaneity arises in situations where one of the regressors is not truly independent of the outcome, but rather is determined simultaneously with the outcome through an equilibrium relationship or a feedback mechanism. In other words, X and Y mutually influence each other, making it ambiguous which one is the cause and which is the effect in a simple regression framework.

When simultaneity is present, the regressor X is endogenous because changes in X may be caused by changes in Y (or by common factors that affect both). As a result, the OLS estimation of Y on X will capture a mix of effects, and it generally cannot isolate the pure causal effect of X on Y. The estimated coefficient may be biased and inconsistent because Assumption 6.1 fails: the variation in X is not external or as-good-as-random; instead, X responds to shocks in Y.

A canonical example of simultaneity comes from supply and demand in economics:

Example 6.5 (Simultaneous Determination of Price and Quantity)**.** Consider a market for a commodity where we have a demand equation and a supply equation:

$$Q^d = \alpha_0 - \alpha_1 P + u \qquad \text{(Demand)},$$

$$Q^s = \gamma_0 + \gamma_1 P + v \qquad \text{(Supply)}.$$

Here Q^d is quantity demanded, Q^s is quantity supplied, P is the price, and u, v are demand and supply shocks respectively. In equilibrium, the observed price P and quantity Q satisfy $Q = Q^d = Q^s$. If one were to naively regress Q on P using observational data from this market, the coefficient would not recover the true demand parameter $-\alpha_1$ (nor the supply parameter γ_1). The reason is that price and quantity are jointly determined by both equations. When demand u shifts, it changes Q and P; when supply v shifts, it also changes Q and P. The OLS regression of Q on P essentially mixes up these demand and supply movements, producing an estimate that lies somewhere between the true demand and supply slopes. Moreover, the regressor P is correlated with the composite error (which would include both u and v). Thus, P is endogenous. The estimation suffers from simultaneity bias, and we cannot interpret the estimated coefficient as the demand elasticity (or the supply slope) on its own.

This example illustrates that when both X and Y are determined by a system of equations, a single-equation regression cannot disentangle causality. In the supply-demand case, more advanced methods (such as two-stage least squares with instrumental variables for price) are required to consistently estimate the demand or supply parameters separately.

Simultaneity bias is essentially a form of omitted variable bias, where the omitted factors are the forces that simultaneously determine the regressor and the outcome. In the above example, the demand shock u is an omitted factor in the supply equation and is correlated with price P, and the supply shock v is omitted in the demand equation and is also correlated with P. Each equation alone omits the influence of the other side of the market.

Another way to think about simultaneity is as a reverse causation problem: we typically write $Y = f(X)$ but in reality X might also be a function of Y. For instance, consider the relationship between a firm's advertising expenditure and its sales. One might posit a model $\text{Sales} = \beta_0 + \beta_1 \text{Advertising} + u$, assuming that more advertising causes higher sales. However, it could also be true that firms adjust their advertising budgets in response to sales trends (when sales are expected to increase, they spend more on ads, or perhaps when sales are poor, they spend more to boost them). If advertising is partly determined by anticipated sales, then causality runs both ways: advertising affects sales and sales affects advertising. A simple OLS regression of sales on advertising would be endogenous, as the regressor (advertising) is influenced

by the error term (which in this context includes shocks to sales or demand). In such a case, the estimated β_1 will be biased, and we cannot interpret it straightforwardly as the effect of advertising on sales.

The core issue in simultaneity is that an explanatory variable is *not purely external* to the system but is itself an outcome of some other simultaneous process. As a result, we lose the clear causal ordering that regression requires for interpretation. OLS regression alone cannot resolve "who is causing whom." In fact, with simultaneity, it is fundamentally impossible to discern the direction of causality using only the variation in X and Y that we observe: all the variation in X is potentially contaminated by feedback from Y. This is why additional techniques are needed to study causal relationships in simultaneous settings. One approach is to find an *instrumental variable* that affects X but not Y except through X, and use it to tease out exogenous variation in X. (Instrumental variables are the subject of the next chapter.) Another approach is to build and estimate a system of equations (as in structural modeling) to account for the joint determination explicitly.

6.4.1 Testing for Endogeneity: The Durbin–Wu–Hausman Test

Given an econometric model, it is often useful to test whether a regressor is in fact endogenous (violating Assumption Exogeneity) or whether OLS might be reliable. One widely used approach is the Durbin–Wu–Hausman (DWH) test for endogeneity, which in regression form is sometimes called the Davidson–MacKinnon test.[5] This test provides a way to check if an OLS coefficient differs significantly from what it would be under an alternative method that accounts for endogeneity (such as an instrumental variables estimator). A significant difference suggests that the regressor is endogenous.

[5]The Durbin–Wu–Hausman test is named after James Durbin (1954), De-Min Wu (1973), and Jerry Hausman (1978), who developed tests for endogeneity/consistency. Davidson and MacKinnon (1993) describe a regression-based implementation of this test. The intuition is to compare an estimator that is always consistent (under both H_0 and H_a, such as IV) to one that is only consistent under H_0 (such as OLS). A significant difference between the two estimates signals that H_0 (exogeneity of X) is likely false. The procedure described above (sometimes called a Hausman regression) is a convenient way to perform this test in practice by checking if the OLS residual contains any remaining predictive power after accounting for the instrument.

The basic implementation of the DWH test for a single suspect regressor X is as follows:

1. **First-stage regression:** Find an additional variable (or set of variables) that can serve as an instrument Z for the suspect regressor X. The instrument Z should be correlated with X (relevant) but uncorrelated with the original error term ε in the Y equation (valid). Regress X on all exogenous variables in the model, including Z, to obtain the fitted values $\hat{X}$ and the residual $r = X - \hat{X}$. This first stage effectively isolates the part of X that is exogenous (as explained by Z and other controls) from the part that might be endogenous.

2. **Augmented regression:** Next, take the original regression model for Y and add the residual r from the first stage as an additional regressor. For example, if our original model was $Y = \beta_0 + \beta_1 X + \text{controls} + \varepsilon$, we now estimate an augmented model

$$Y = \beta_0 + \beta_1 X + \theta\, r + (\text{other controls}) + \text{error}.$$

   This augmented regression allows us to directly test whether any leftover variation in X (the part not explained by the instrument and other exogenous variables) has an effect on Y.

3. **Test:** Now we test whether the coefficient $\hat{\theta}$ on the residual r is significantly different from zero. The null hypothesis is $H_0 : \theta = 0$ (which would imply X is exogenous). The alternative is that $\theta \neq 0$ (which implies X is endogenous).

   - If the test fails to reject H_0 (i.e. $\hat{\theta}$ is not significantly different from zero), we do not find evidence that X is endogenous. The OLS estimate for β_1 may be considered reliable (at least with respect to endogeneity concerns).
   - If the test rejects H_0 (i.e. $\hat{\theta}$ is significantly non-zero), this is evidence that X is endogenous. Intuitively, the residual r contains the variation in X that is unrelated to the instruments and other controls—essentially the variation that could be driven by Y itself or omitted factors. A significant θ means that variation is indeed affecting Y, implying X was picking up that effect (and thus correlated with the error term). In such a case, the OLS estimate

of β_1 is biased, and one should rely on instrumental variables or other techniques to obtain a consistent estimate.

Example 6.6 (Testing Endogeneity of Education)**.** Continuing with the theme of education and earnings, suppose we want to test whether years of education X is endogenous in a wage regression (perhaps due to omitted ability or other factors). We might have a candidate instrument Z for education—common examples include variables like distance to the nearest college, changes in schooling laws that affect certain cohorts, or other background characteristics that influence education but are plausibly unrelated to individual wage potential aside from their effect on schooling.

We can perform a Durbin–Wu–Hausman test as follows:

1. First, we would regress X (education) on the instrument Z and any other exogenous controls (such as experience, demographic variables, etc.). This yields a fitted value $\hat{X}$ capturing the part of education predicted by these exogenous factors, and a residual $r = X - \hat{X}$ which represents the part of education not explained by the instrument and controls.

2. Second, we run the augmented wage regression including r: wage $= \beta_0 + \beta_1 X + \theta r +$ (other controls) $+u$.

3. Finally, we test whether $\hat{\theta} = 0$. If, for instance, we find $\hat{\theta}$ is positive and statistically significant, it suggests that the portion of education not explained by the instrument has a positive association with wages—likely capturing the effect of ability or other omitted factors. This indicates education is endogenous (the OLS estimate of β_1 was picking up ability bias), and we would reject the null of exogeneity. On the other hand, if $\hat{\theta}$ is near zero (and statistically insignificant), we would not reject exogeneity, providing some reassurance that OLS was not severely biased by omitted factors (at least those correlated with the instrument).

In practice, this testing procedure helps determine whether an instrumental variable approach is necessary. If the test suggests endogeneity, one would proceed with IV estimation to get a consistent estimate of β_1. If the test does not find evidence of endogeneity, researchers might be more confident in the OLS results (though one must always consider the strength of the instrument and power of the test when interpreting such outcomes).

The Davidson–MacKinnon (Durbin–Wu–Hausman) test provides a formal way to detect endogeneity, but it relies on having at least one valid instrument for the suspect regressor. Without an instrument or some alternative source of identification, simultaneity bias cannot be definitively diagnosed or resolved. Nonetheless, the test is a useful diagnostic when instruments are available, and it serves as a reminder that whenever we suspect a regressor might be jointly determined with the outcome, we should not blindly trust the OLS estimate. Instead, we must seek additional evidence or more robust strategies to pin down the causal effect.

6.5 Conclusion

In this chapter, we explored three fundamental sources of endogeneity—omitted variable bias, measurement error, and simultaneity—and examined how each leads to a violation of the exogeneity assumption required for causal inference in OLS regression. Through examples and derivations, we saw that:

- Omitting a relevant confounding variable can bias our coefficient estimates, sometimes severely, by attributing the effect of the missing factor to the included regressors.
- Measurement error in regressors, especially when random (classical), tends to attenuate estimated effects, making it harder to detect true relationships.
- Simultaneity or reverse causality means that the direction of influence between X and Y is blurred, so OLS estimates capture a mixture of effects and cannot be given a straightforward causal interpretation.

All these issues result in OLS estimates that are biased and inconsistent for the true causal parameters. In practice, recognizing potential endogeneity is a crucial part of empirical research. Researchers must combine economic reasoning (to suspect when an explanatory variable might be endogenous) with statistical tests (like the Davidson–MacKinnon test) and, importantly, techniques to address endogeneity.

Having identified the problems, the next step is to solve them. There are several strategies to obtain credible causal estimates in the presence of endogeneity. One powerful approach is the use of **instrumental variables (IV)**, which provide an external source of variation in the endogenous regressor that can isolate the true causal effect. Another strategy is to use panel data and fixed-effects models to difference out omitted variables that are fixed within entities. Additionally, randomized experiments or natural experiments can offer exogenous variation that circumvents endogeneity concerns. These topics will be explored in subsequent chapters.

In sum, causal inference requires careful attention to the sources of endogeneity. By understanding omitted variable bias, measurement error, and simultaneity, and by employing appropriate methods to mitigate these issues, we can move closer to uncovering true causal relationships in economic data, rather than being misled by mere correlations.