# 8 Endogeneity

## 8.1 The Linear Model and Exogeneity

So far we have written the conditional mean of an outcome $Y_i$ as a linear function of observed covariates $\boldsymbol{X}_i$:

$$\begin{aligned} Y_i &= \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, \\ E[u_i \mid \boldsymbol{X}_i] &= 0 \end{aligned} \tag{A1}$$

If (A1) holds, then $E[Y_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i'\boldsymbol{\beta}$, which makes $\boldsymbol{X}_i'\boldsymbol{\beta}$ the best predictor of $Y_i$ given $\boldsymbol{X}_i$. Each coefficient $\beta_j$ is a **conditional marginal effect**:

> **Interpretation:** *"Among individuals who share the same values of all included control variables, those whose $X_{ij}$ is higher by one unit have, on average, a $Y_i$ that is higher by $\beta_j$."*

So far the course has provided three empirical tactics to narrow the gap between correlation and causation:

- Add observed confounders. Whenever economic theory identifies a variable that influences both $X_{ij}$ and $Y_i$, we try to measure it and augment $\boldsymbol{X}_i$.
- Exploit panel structure. With panel data data we include individual and time fixed effects to control for unobserved factors that are constant across individuals or time periods.
- Use flexible functional forms. Polynomials, interactions, or other transformations can absorb nonlinearities that would otherwise leak into $u_i$.

Even after taking these steps, important issues remain. For example, there may be reverse causality, which occurs when $Y_i$ feeds back into $X_i$. Additionally, there may be control variables with a dual role that act as both confounders and mediators/colliders simultaneously.

Nothing in (A1) – nor in the additional assumptions (A2)–(A4) about i.i.d. sampling, finite moments, and full rank – guarantees that $\beta_j$ is **causal**. It represents only a conditional **correlative** relationship unless $X_{ij}$ is uncorrelated with all unobserved determinants of $Y_i$.

---

## 8.2 Conditional vs Causal Effects: Price Elasticities

Economists often want *causal* price effects, not merely conditional associations. Consider the following structural system in a competitive market written in logs so that slopes are elasticities:

$$\begin{aligned} \text{Demand:} \quad & \log(Q_i) = \beta_1 + \beta_2 \log(P_i) + u_i, \\ \text{Supply (pricing rule):} \quad & \log(P_i) = \gamma_1 + \gamma_2 \log(C_i) + \gamma_3 u_i + \eta_i. \end{aligned}$$

We have $\beta_2 < 0$ by theory.

- Index $i$ denotes a market (e.g., city or store) observed at a single point in time; the data are cross-sectional and i.i.d.
- $Q_i$ is the total quantity demanded in market $i$.
- $P_i$ is price.
- $C_i$ is the exogenous wholesale cost of the product.
- $u_i$ captures consumers' taste shocks unobserved by the econometrician (though retailers may infer them and respond when setting prices); $\eta_i$ captures supply-side shocks.

Because higher demand (large $u_i$) in a particular store leads retailers to charge higher prices ($\gamma_3 > 0$), we have $Cov(\log(P_i), u_i) > 0$. Hence, (A1) is violated in the demand equation.

Suppose a researcher estimates

$$\log(Q_i) = \alpha_1 + \alpha_2 \log(P_i) + \varepsilon_i$$

or

$$\log(Q_i) = \theta_1 + \theta_2 \log(P_i) + \theta_3 \log(C_i) + v_i$$

Both regressions (one simple and one with wholesale-cost controls) deliver conditional marginal effects $\alpha_2$ or $\theta_2$. They answer

> *"Among markets with the same wholesale cost (and any other included controls), how does observed quantity co-move with observed price?"*

But the policy-relevant question is different:

> *"By how much would quantity fall if we exogenously raised price – say, via a* 1% *tax – holding everything else constant?"*

That causal elasticity is $\beta_2$. Because $P_i$ responds to $u_i$, OLS estimates suffer simultaneity bias and $\alpha_2$ or $\theta_2$ generally differ from $\beta_2$.

Endogeneity arises because we want the parameter to be causal, not because the regression is mechanically misspecified. Even if the conditional mean is correctly linear, interpreting $\beta_2$ causally implies $Cov(\log(P_i), u_i) \neq 0$.

## 8.3 Measurement Error

Another important source of endogeneity arises from measurement error. Suppose we consider the structural model:

$$Y_i^0 = \beta_1 + \beta_2 X_i^0 + u_i^0, \quad i = 1, \dots, n, \quad u_i^0 \sim \text{i.i.d.}(0, \sigma^2),$$

but we do not observe the latent variables $Y_i^0$ and $X_i^0$ directly. Instead, we observe:

$$Y_i = Y_i^0 + \eta_i, \quad X_i = X_i^0 + \zeta_i,$$

where $\eta_i \sim \text{i.i.d.}(0, \sigma_\eta^2)$ and $\zeta_i \sim \text{i.i.d.}(0, \sigma_\zeta^2)$ denote classical measurement errors that are assumed independent of each other and of $X_i^0, Y_i^0$, and $u_i^0$.

Plugging the observed variables into the structural equation yields:

$$Y_i - \eta_i = \beta_1 + \beta_2 (X_i - \zeta_i) + u_i^0,$$

which can be rearranged as:

$$Y_i = \beta_1 + \beta_2 X_i + \underbrace{(u_i^0 + \eta_i - \beta_2 \zeta_i)}_{\text{composite error term}}.$$

The composite error term is problematic:

$$E[u_i^0 + \eta_i - \beta_2 \zeta_i \mid X_i] \neq 0,$$

because $X_i$ contains $\zeta_i$, which also appears in the error term. This violates the exogeneity condition, resulting in a biased and inconsistent OLS estimator. Specifically, the bias tends to attenuate the coefficient estimate $\widehat{\beta}_2$ toward zero (known as attenuation bias). For positive true coefficients, this leads to underestimation; for negative coefficients, overestimation.

By contrast, if only the dependent variable $Y_i$ is measured with error, OLS remains unbiased, although the variance of the error term increases.

## 8.4 Endogeneity as a Violation of (A1)

Formally, a regressor $X_{ij}$ is **endogenous** if it correlates with the structural error term:

$$Cov(X_{ij}, u_i) \neq 0 \quad \Rightarrow \quad E[u_i \mid X_i] \neq 0$$

When this happens, OLS estimates remain descriptive but lose their causal interpretation. Whether you care depends on your goal:

| Purpose | Is (A1) needed? | Parameter meaning |
|---|---|---|
| **Prediction / description** | *No.* Bias relative to causal truth is irrelevant if forecasting is the aim. | Conditional marginal effect |
| **Causal policy evaluation** | **Yes!** You need $E[u\|X] = 0$ in the causal sense, or an alternative identification strategy. | Structural (causal) effect |

---

## 8.5 Sources of Endogeneity

Besides the functional-form misspecification that we have already discussed in previous sections, there are four other common sources of endogeneity in practice:

| Mechanism | Typical manifestation |
|---|---|
| **Omitted-variable bias** | Unobserved ability affects both schooling ($X$) and wages ($Y$) |
| **Simultaneity / reverse causality** | Price and quantity determined jointly in markets |
| **Measurement error in $X$** | Measurement error inflates the variance of the regressor, so OLS slopes are biased toward zero (attenuation bias) |
| **Dual role controls** | A variable (e.g., health) acts as both confounder and mediator/collider |

All four cases yield $E[\boldsymbol{u}|\boldsymbol{X}] \neq 0$ and threaten causal inference.

We have

$$E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{u}|\boldsymbol{X}] \neq \boldsymbol{\beta}.$$