Chapter 10

Endogeneity

In microeconomic analysis, exogenous variables are the factors determined outside of the economic system under consideration, and endogenous variables are those decided within the economic system.

Example 10.1. A microeconomic exercise that we encountered so many times goes as follows. If a person has a utility function $u(q_1, q_2)$ where q_1 and q_2 are the quantities of two goods. He faces a budget $p_1 q_1 + p_2 q_2 \leq C$, where p_1 and p_2 are the prices of the two goods, respectively. What is the optimal quantities q_1^* and q_2^* he will purchase? In this question the utility function $u(\cdot, \cdot)$, the prices p_1 and p_2, and the budget C are exogenous. The optimal purchase q_1^* and q_2^* are endogenous.

The terms "endogenous" and "exogenous" in microeconomics will be carried over into multiple-equation econometric models. While in a single-equation regression model

$$y_i = x_i' \beta + e_i \tag{10.1}$$

is only part of the equation system. To make it simple, in the single-equation model we say an x_{ik} is *endogenous,* or is an *endogenous variable*, if $\text{cov}(x_{ik}, \epsilon_i) \neq 0$; otherwise x_{ik} is an *exogenous variable*.

Empirical works using linear regressions are routinely challenged by questions about endogeneity. Such questions plague economic seminars and referee reports. To defend empirical strategies in quantitative economic studies, it is important to understand the sources of potential endogeneity and thoroughly discuss attempts for resolving endogeneity.

10.1 Identification

Endogeneity usually implies difficulty in identifying the parameter of interest with only (y_i, x_i). Identification is critical for the interpretation of empirical economic research. We say a parameter is *identified* if the mapping between the parameter in the model and the distribution of the observed variable is one-to-one; otherwise we say the parameter is *under-identified*. This is an abstract definition, and let us discuss it in the family linear regression context.

Example 10.2 (Identification failure due to collinearity)**.** The linear projection model implies the moment equation

$$\mathbb{E}\left[x_i x_i'\right]\beta = \mathbb{E}\left[x_i y_i\right]. \tag{10.2}$$

If $E\left[x_i x_i'\right]$ is of full rank, then $\beta = \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\mathbb{E}\left[x_i y_i\right]$ is a function of the quantities of the population moment and it is identified. On the contrary, if some x_k's are perfect collinear so that $\mathbb{E}\left[x_i x_i'\right]$ is rank deficient, there are multiple β that satisfies the k-equation system (10.2). Identification fails. □

Example 10.3 (Identification failure due to endogeneity)**.** Suppose x_i is a scalar random variable,

$$\begin{pmatrix} x_i \\ e_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{xe} \\ \sigma_{xe} & 1 \end{pmatrix}\right)$$

follows a joint normal distribution, and the dependent variable y_i is generated from (10.1). The joint normal assumption implies that the conditional mean

$$\mathbb{E}[y_i|x_i] = \beta x_i + \mathbb{E}[e_i|x_i] = (\beta + \sigma_{xe})\, x_i$$

coincides with the linear projection model, and $\beta + \sigma_{xe}$ is the linear projection coefficient. From the observable random variable (y_i, x_i), we can only learn $\beta + \sigma_{xe}$. As we cannot learn σ_{xe} from the data due to the unobservable e_i, there is no way to recover β. This is exactly the *omitted variable bias* that we have discussed earlier in this course. The gap lies between the available data (y_i, x_i) and the identification of the model. In the special case that we assume $\sigma_{xe} = 0$, the endogeneity vanishes and β is identified. □

Remark 10.1. The linear projection model is so far the most general model in this course that justifies OLS. OLS is consistent for the linear projection coefficient. By the definition of the linear projection model, $\mathbb{E}[x_i e_i] = 0$ so there is no room for endogeneity in the linear projection model. In other words, if we talk about endogeneity, we must not be working with the linear projection model, and the coefficients we pursue the structural parameter, rather than the linear projection coefficients.

In econometrics we are often interested in a model with economic interpretation. The common practice in empirical research assumes that the observed data are generated from a parsimonious model, and the next step is to estimate the unknown parameters in the model. Since it is often possible to name some factors not included in the regressors but they are correlated with the included regressors and in the mean time also affects y_i, endogeneity becomes a fundamental problem.

To resolve endogeneity, we seek extra variables or data structure that may guarantee the identification of the model. The most often used methods are (i) fixed effect model (ii) instrumental variables:

- The fixed effect model requires that multiple observations, often across time, are

collected for each individual i. Moreover, the source of endogeneity is time invariant and enters the model additively in the form

$$y_{it} = x_{it}'\beta + u_{it},$$

where $u_{it} = \alpha_i + \epsilon_{it}$ is the composite error. The panel data approach extends (y_i, x_i) to $(y_{it}, x_{it})_{i=1}^{T}$ if data are available along the time dimension.

- The instrumental variable approach extends (y_i, x_i) to (y_i, x_i, z_i), where the extra random variable z_i is called the *instrument variable*. It is assumed that z_i is orthogonal to the error e_i . Therefore, along with the model it adds an extra variable z_i.

Either the panel data approach or the instrumental variable approach entails extra information beyond (y_i, x_i). Without such extra data, there is no way to resolve the identification failure. Just as the linear project model is available for any joint distribution of (y_i, x_i) with existence of suitable moments, from a pure statistical point of view a linear IV model is an artifact depends only on the choice of (y_i, x_i, z_i) without referencing to any economics. In essence, the linear IV model seeks a linear combination $y_i - \beta x_i$ that is orthogonal to the linear space spanned by z_i.

10.2 Instruments

There are two requirements for valid IVs: orthogonality and relevance. Orthogonality entails that the model is correctly specified. If relevance is violated, meaning that the IVs are not correlated with the endogenous variable, then multiple parameters can generate the observable data. Identification, as in the standard definition in econometrics, breaks down.

A structural equation is a model of economic interest. Consider the following linear

structural model

$$y_i = x_{1i}'\beta_1 + z_{1i}'\beta_2 + \epsilon_i, \tag{10.3}$$

where x_{1i} is a k_1-dimensional endogenous explanatory variables, z_{1i} is a k_2-dimensional exogenous explanatory variables with the intercept included. In addition, we have z_{2i}, a k_3-dimensional excluded exogenous variables. Let $K = k_1 + k_2$ and $L = k_2 + k_3$. Denote $x_i = \left(x_{1i}', z_{1i}'\right)'$ as a K-dimensional explanatory variable, and $z_i = \left(z_{1i}', z_{2i}'\right)$ as an L-dimensional exogenous vector.

We call the exogenous variable *instrument variables*, or simply *instruments*. Let $\beta = (\beta_1', \beta_2')'$ be a K-dimensional parameter of interest. From now on, we rewrite (10.3) as

$$y_i = x_i'\beta + \epsilon_i, \tag{10.4}$$

and we have a vector of instruments z_i.

Before estimating any structural econometric model, we must check identification. In the context of (10.4), identification requires that the true value β_0 is the only value on the parameters space that satisfies the moment condition

$$\mathbb{E}\left[z_i\left(y_i - x_i'\beta\right)\right] = 0_L. \tag{10.5}$$

The rank condition is sufficient and necessary for identification.

Assumption (Rank condition)**.** $\operatorname{rank}\left(\mathbb{E}\left[z_i x_i'\right]\right) = K$.

Note that $\mathbb{E}\left[x_i' z_i\right]$ is a $K \times L$ matrix. The rank condition implies the *order condition* $L \geq K$, which says that the number of excluded instruments must be no fewer than the number of endogenous variables.

Theorem. *The parameter in (10.5) is identified if and only if the rank condition holds.*

Proof. (The "if" direction). For any $\tilde{\beta}$ such that $\tilde{\beta} \neq \beta_0$,

$$\begin{aligned}\mathbb{E}\left[z_i\left(y_i - x_i'\tilde{\beta}\right)\right] &= \mathbb{E}\left[z_i\left(y_i - x_i'\beta_0\right)\right] + \mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right) \\ &= 0_L + \mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right).\end{aligned}$$

Because $\operatorname{rank}\left(\mathbb{E}\left[z_i x_i'\right]\right) = K$, we would have $\mathbb{E}\left[z_i x_i'\right]\left(\beta_0 - \tilde{\beta}\right) = 0_L$ if and only if $\beta_0 - \tilde{\beta} = 0_K$, which violates $\tilde{\beta} \neq \beta_0$. Therefore β_0 is the unique value that satisfies (10.5).

(The "only if" direction is left as an exercise. Hint: By contrapositiveness, if the rank condition fails, then the model is not identified. We can easily prove the claim by making an example.) □

10.3 Sources of Endogeneity

As econometricians mostly work with non-experimental data, we cannot overstate the importance of the endogeneity problem. We go over a few examples.

Example 10.4 (Dynamic Panel Model)**.** We know that the first-difference (FD) estimator is consistent for (static) panel data model. Nevertheless, the FD estimator encounters difficulty in a dynamic panel model

$$y_{it} = \beta_1 + \beta_2 y_{i,t-1} + \beta_3 x_{it} + \alpha_i + \epsilon_{it}, \tag{10.6}$$

even if we assume

$$\mathbb{E}\left[\epsilon_{is} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}\right] = 0, \ \ \forall s \geq t \tag{10.7}$$

When taking difference of the above equation (10.6) for periods t and $t-1$, we have

$$\left(y_{it} - y_{i,t-1}\right) = \beta_2\left(y_{it-1} - y_{i,t-2}\right) + \beta_3\left(x_{it} - x_{i,t-1}\right) + \left(\epsilon_{it} - \epsilon_{i,t-1}\right). \tag{10.8}$$

Under (10.7), $\mathbb{E}\left[\left(x_{it} - x_{i,t-1}\right)\left(\epsilon_{it} - \epsilon_{i,t-1}\right)\right] = 0$, but

$$\mathbb{E}\left[\left(y_{i,t-1} - y_{i,t-2}\right)\left(\epsilon_{it} - \epsilon_{i,t-1}\right)\right] = -\mathbb{E}\left[y_{i,t-1}\epsilon_{i,t-1}\right] = -\mathbb{E}\left[\epsilon_{i,t-1}^2\right] \neq 0.$$

Therefore the coefficients β_2 and β_3 cannot be identified from the linear regression model (10.8). □

Remark 10.2. Instruments for the above example is easy to find. Notice that the linear relationship (10.6) implies

$$\begin{aligned} & \mathbb{E}\left[y_{i,t-1} - y_{i,t-2} | \alpha_i, x_{i1}, \ldots, x_{iT}, \epsilon_{i,t-3}, \epsilon_{i,t-4}, \ldots, \epsilon_{i1}, y_{i0}\right] \\ = \; & \mathbb{E}\left[y_{i,t-1} - y_{i,t-2} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{i,t-3}, y_{i,t-4}, \ldots, y_{i0}\right] = 0 \end{aligned}$$

according to the assumption (10.7). The above relationship gives orthogonal condition in the form

$$\mathbb{E}\left[\left(y_{i,t-1} - y_{i,t-2}\right) f\left(\epsilon_{i,t-3}, \epsilon_{i,t-4}, \ldots, \epsilon_{i1}\right)\right] = 0.$$

In other words, any function of $\epsilon_{i,t-3}, \epsilon_{i,t-4}, \ldots, \epsilon_{i1}$ is orthogonal to the endogenous variable $(y_{i,t-1} - y_{i,t-2})$. Here the excluded IVs are naturally generated from the model itself.

Another classical source of endogeneity is the measurement error.

Example 10.5 (Classical Measurement Error)**.** Endogeneity also emerges when an explanatory variables is not directly observable but is replaced by a measurement with error. Suppose the true linear model is

$$y_i = \beta_1 + \beta_2 x_i^* + u_i, \tag{10.9}$$

with $\mathbb{E}\left[u_i | x_i^*\right] = 0$. We cannot observe x_i^* but we observe x_i, a measurement of x_i^*, and they are linked by

$$x_i = x_i^* + v_i$$

with $\mathbb{E}\left[v_i|x_i^*, u_i\right] = 0$. Such a formulation of the measurement error is called the *classical measurement error*. Substitute out the unobservable x_i^* in (10.9),

$$y_i = \beta_1 + \beta_2 (x_i - v_i) + u_i = \beta_1 + \beta_2 x_i + e_i \tag{10.10}$$

where $e_i = u_i - \beta_2 v_i$. The correlation

$$\mathbb{E}[x_i e_i] = \mathbb{E}[(x_i^* + v_i)(u_i - \beta_2 v_i)] = -\beta_2 \mathbb{E}\left[v_i^2\right] \neq 0.$$

OLS (10.10) would not deliver a consistent estimator. □

Next, we give two examples of equation systems, one from microeconomics and the other from macroeconomics.

Example 10.6 (Demand-Supply System)**.** Let p_i and q_i be a good's log-price and log-quantity on the i-th market, and they are iid across markets. We are interested in the demand curve

$$p_i = \alpha_d - \beta_d q_i + e_{di} \tag{10.11}$$

for some $\beta_d \geq 0$ and the supply curve

$$p_i = \alpha_s + \beta_s q_i + e_{si} \tag{10.12}$$

for some $\beta_s \geq 0$. We use a simple linear specification so that the coefficient β_d can be interpreted as demand elasticity and β_s as supply elasticity. Undergraduate microeconomics teaches the deterministic form but we add an error term to cope with the data. Can we learn the elasticities by regression p_i on q_i?

The two equations can be written in a matrix form

$$\begin{pmatrix} 1 & \beta_d \\ 1 & -\beta_s \end{pmatrix} \begin{pmatrix} p_i \\ q_i \end{pmatrix} = \begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix}. \tag{10.13}$$

Microeconomic terminology calls (p_i, q_i) endogenous variables and (e_{di}, e_{si}) exogenous variables. (10.13) is a *structural equation* because it is motivated from economic theory so that the coefficients bear economic meaning. If we rule out the trivial case $\beta_d = \beta_s = 0$, we can solve

$$\begin{pmatrix} p_i \\ q_i \end{pmatrix} = \frac{1}{\beta_s + \beta_d} \begin{pmatrix} \beta_s & \beta_d \\ 1 & -1 \end{pmatrix} \left[\begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix} \right]. \tag{10.14}$$

This equation (10.14) is called the *reduced form*—the endogenous variables are expressed as explicit functions of the parameters and the exogenous variables. In particular,

$$q_i = (\alpha_d + e_{di} - \alpha_s - e_{si}) \,/\, (\beta_s + \beta_d)$$

so that the log-price is correlated with both e_{si} and e_{di}. As q_i is endogenous (in the econometric sense) in either (10.11) or (10.12), neither the demand elasticity nor the supply elasticity is identified with (p_i, q_i). Indeed, as

$$q_i = (\beta_s \alpha_d + \beta_d \alpha_s + \beta_s e_{di} + \beta_d e_{si}) \,/\, (\beta_s + \beta_d)$$

from (10.14), the linear projection coefficient of p_i on q_i is

$$\frac{\operatorname{cov}(p_i, q_i)}{\operatorname{var}(q_i)} = \frac{\beta_s \sigma_d^2 - \beta_d \sigma_s^2 + (\beta_d - \beta_s)\, \sigma_{sd}}{\beta_d^2 \sigma_d^2 + \beta_d \sigma_s^2 + 2\beta_d \beta_s \sigma_{sd}},$$

where $\sigma_d^2 = \operatorname{var}(e_{di})$, $\sigma_s^2 = \operatorname{var}(e_{si})$ and $\sigma_{sd} = \operatorname{cov}(e_{di}, e_{si})$.

This is a classical example of the demand-supply system. The structural parameter cannot be directly identified because the observed (p_i, q_i) is the outcome of an equilibrium—

the crossing of the demand curve and the supply curve. To identify the demand curve, we will need an instrument that shifts the supply curve only; and vice versa. □

Example 10.7 (Keynesian-Type Macro Equations)**.** This is a model borrowed from Hayashi (2000, p.193) but originated from Haavelmo (1943). An econometrician is interested in learning β_2, the *marginal propensity of consumption*, in the Keynesian-type equation

$$C_i = \beta_1 + \beta_2 Y_i + u_i \tag{10.15}$$

where C_i is household consumption, Y_i is the GNP, and u_i is the unobservable error. However, Y_i and C_i are connected by an accounting equality (with no error)

$$Y_i = C_i + I_i,$$

where I_i is investment. We assume $\mathbb{E}[u_i|I_i] = 0$ as investment is determined in advance. In this example, $(Y_i C_i)$ are endogenous and (I_i, u_i) are exogenous. OLS (10.15) will be inconsistent because in the reduced-form $Y_i = \frac{1}{1-\beta_2}(\beta_1 + u_i + I_i)$ implies $\mathbb{E}[Y_i u_i] = \mathbb{E}\left[u_i^2\right] / (1 - \beta_2) \neq 0$. □

10.4 Summary

Even though we often deal with a single equation model with potential endogenous variables, the underlying structural system may involve multiple equations. The simultaneous equation model is a classical econometric modeling approach, and it is still actively applied in structural economic studies. When our economic model is "structural", we keep in mind a causal mechanism. Instead of identifying the causal effect by control group and treatment group as in Chapter 2, here we look at causality from the economic structural perspective.

Historical notes: Instruments originally appeared in Philip Wright (1928) for identifying the coefficient of an endogenous variables. It is believed to be a collaborative idea with Philip's son Sewall Wright. The demand and supply analysis is attributed to Working (1927), and the measurement error study is dated back to Fricsh (1934).

Further reading: Causality is the holy grail of econometrics. Pearl and Mackenzie (2018) is a popular book with philosophical depth. It is a delight to read. Chen et al. (2011) is a survey for modern nonlinear measurement error models.

Zhentao Shi. Nov 18, 2020.

Bibliography

Chen, X., H. Hong, and D. Nekipelov (2011). Nonlinear models of measurement errors. *Journal of Economic Literature 49*(4), 901–37.

Fricsh, R. (1934). Statistical confluence study. *Oslo: University Institute of Economics*.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 1–12.

Pearl, J. and D. Mackenzie (2018). *The book of why: the new science of cause and effect*. Basic Books.

Working, E. J. (1927). What do statistical "demand curves" show? *The Quarterly Journal of Economics 41*(2), 212–235.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. Macmillan Company, New York.