Lecture 7: Endogeneity

Zhentao Shi

November 15, 2018

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

$$y_i = x_i'\beta + e_i, \tag{1}$$

we say an x_{ik} is *endogenous*, or is an *endogenous variable*, if $\mathrm{cov}\left(x_{ik}, \epsilon_i\right) \neq 0$; otherwise x_{ik} is an *exogenous variable*. These terms here are irrelevant to those in microeconomic analysis.

Empirical economic works using only linear regressions are routinely challenged with questions about endogeneity. We hear such questions so often in economic seminars and on journal referee reports. In this sense, understanding the source of potential endogeneity and thoroughly discussion of effort to resolve endogeneity are important in defending empirical strategies in quantitative economic studies.

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

Example 1 (Identification failure due to collinearity)**.** We know in the linear projection model we have

$$\mathbb{E}\left[x_i x_i'\right]\beta = \mathbb{E}\left[x_i y_i\right]. \tag{2}$$

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 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 (2). Identification fails. □

Example 2 (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 is generated from (1). The joint normal assumption implies that the conditional mean

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

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 generative model: we simply cannot identify such a generative model from the available data. In the special case that we assume $\sigma_{xe} = 0$, the endogeneity vanishes and β is identified.

So far in this course the most general model that justifies the OLS is the linear projection model. Under very general condition, a population linear projection model exists, and the OLS is consistent for the linear projection coefficient. By the definition of the linear projection model, $\mathbb{E}\left[x_i e_i\right] = 0$ 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 are after are not the linear projection coefficients. □

In econometrics we are often interested in the generative model which contains richer economic interpretation, instead of the liner projection model which is a merely statistical artifact. The common practice in empirical research is that we assume 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 contain information to guarantee the identification of the generative 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 generative 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 the data along the time dimension are available.

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 generative model it adds an extra variable z_i. The IV method will be the topic of the next chapter.

Before closing this section, we stress that either the panel data approach or the instrumental variable approach requires extra information beyond (y_i, x_i). Without such extra data, there is no way to resolve the identification failure. If we take (y_i, x_i) as our budget at hand, with such a mean budget we cannot afford to identify a general generative model (with the error possibly correlated with the regressors).

1.1 Examples 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 3 (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

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

even if we assume

$$\mathbb{E}\left[\epsilon_{it} | \alpha_i, x_{i1}, \ldots, x_{iT}, y_{it-1}, y_{it-2}, \ldots, y_{i0}\right] = 0. \tag{3}$$

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

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

Under (3), $\mathbb{E}\left[\left(x_{it} - x_{it-1}\right)\left(\epsilon_{it} - \epsilon_{it-1}\right)\right] = 0$, but

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

Example 4 (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

$$C_i = \beta_1 + \beta_2 Y_i + u_i \tag{4}$$

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 $\mathbb{E}\left[u_i | I_i\right] = 0$ as investment is determined in advance. OLS (4) will be inconsistent because in the reduced-form $Y_i = \frac{1}{1-\beta_2}\left(\beta_1 + u_i + I_i\right)$ implies $\mathbb{E}\left[Y_i u_i\right] = \mathbb{E}\left[u_i^2\right] / \left(1 - \beta_2\right) \neq 0$. □

Example 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

$$y_i = \beta_1 + \beta_2 x_i^* + u_i, \tag{5}$$

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*. When we substitute out the unobservable x_i^* in (5), we have

$$y_i = \beta_1 + \beta_2\left(x_i - v_i\right) + u_i = \beta_1 + \beta_2 x_i + e_i \tag{6}$$

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

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

OLS (6) would not deliver a consistent estimator. □

Example 6 (Demand-Supply System)**.** Let p_i and q_i be a good's log-price and log-quantity on the i-th market. In microeconomics, we are interested in the demand curve

$$p_i = \alpha_d - \beta_d q_i + e_{di} \tag{7}$$

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

$$p_i = \alpha_s + \beta_s q_i + e_{si} \tag{8}$$

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. Moreover, undergraduate microeconomics teaches the deterministic form but we use add an error term to cope with the data. Can we learn the elasticities by regression p_i on q_i? Notice that the two equations can be written in a matrix form

$$\begin{pmatrix} 1 & \beta_d \\ 1 & -\beta_s \end{pmatrix} \begin{pmatrix} p_i \\ q_i \end{pmatrix} = \begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix}. \tag{9}$$

(9) 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

$$\begin{pmatrix} p_i \\ q_i \end{pmatrix} = \frac{1}{\beta_s + \beta_d} \begin{pmatrix} \beta_s & \beta_d \\ 1 & -1 \end{pmatrix} \left[\begin{pmatrix} \alpha_d \\ \alpha_s \end{pmatrix} + \begin{pmatrix} e_{di} \\ e_{si} \end{pmatrix} \right]. \tag{10}$$

According to the use of microeconomics, here we can call (p_i, q_i) endogenous variables and (e_{di}, e_{si}) exogenous variables. This equation (10) 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 in general correlated with both e_{si} and e_{di}. As q_i is endogenous (in the econometric sense) in either (7) or (8), 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), the linear projection coefficient of p_i on q_i is

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

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

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 instrument that only shifts the supply curve; and vice versa. □