Econometrics I

TA Session 13

Jukina HATAKEYAMA*

July 9, 2024

Contents

*E-mail: u868710a@ecs.osaka-u.ac.jp

1 Endogeneity

1.1 Definition of Endogeneity

Now we considered the case $\mathbb{E}[u_i|x_i] = 0$, or $\mathbb{E}[u_i|x_i] = 0$. However, there are situations where a ranodm variable is **endogenous**, defined as follows.

Definition 1.1. A variable $x_{k,i}$ is said to be **endogenous** if $x_{k,i}$ is correlated with u_i.

Endogeneity arises mainly for three reasons:

- **Omitted variable:** Issue when we would like to control for one or more additional variable, but because of data unavailability, we cannot include them in the regression. Write the error $u_i = q_i + v_i$, where v_i is a centered error with variance σ^2, and q_i an additional random variable. If $x_{k,i}$ is correlated with q_i, then $x_{k,i}$ is endogeneous.
- **Measurement error:** We would like to measure the effect of a variable, say $x^*_{k,i}$, but we can observe only an imperfect measure of it, say $x_{k,i}$. When plugging $x_{k,i}$ for $x^*_{k,i}$, we put a measurement error into u_i.
- **Simultaneity:** At least one of the explanatory variables is determined simultaneoulsy along with y_i: if $x_{k,i}$ is dertermined partly as a function of y_i, that is,

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + \cdots + x_{ki}(y_i)\beta_k + u_i, \\ x_{ki}(y_i) &= f(y_i) \end{aligned}$$

  then $x_{k,i}$ and u_i are generally correlated.

1.2 Measurement Error

Suppose that the true regression model is $y = \tilde{x}\beta_0 + u$. The observed variable is $x = \tilde{x} + v$, where v is called the **measurement error**. For the element that does not include measurement errors in x, the corresponding elements in v are zeros.

As a consequence, the regression model using observed variables is

$$y = x\beta_0 + (u - v\beta_0).$$

The OLS estimate of β_0 is

$$\hat{\beta}_{OLS} = (x'x)^{-1}(x'y) = \beta_0 + (x'x)^{-1}(x'\{u - v\beta_0\}).$$

To see whether $\hat{\beta}_{OLS}$ is a consistent estimator of β, we assume the following relations:

- $\frac{1}{n}\tilde{x}'v \xrightarrow[n\to\infty]{\mathbb{P}} \mathbf{0}$. This implies

$$\begin{aligned} \frac{1}{n}x'x &= \frac{1}{n}\left[\tilde{x}'\tilde{x} + v'v + \tilde{x}'v + v'\tilde{x}\right] \\ &= \frac{1}{n}\tilde{x}'\tilde{x} + \frac{1}{n}v'v + \frac{1}{n}\tilde{x}'v + \frac{1}{n}v'\tilde{x} \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma + \Omega, \end{aligned}$$

  under the assumptions:

$$\frac{1}{n}\tilde{x}'\tilde{x} \xrightarrow[n\to\infty]{\mathbb{P}} \Sigma \quad \text{and} \quad \frac{1}{n}v'v \xrightarrow[n\to\infty]{\mathbb{P}} \Omega.$$

- $\frac{1}{n}v'u \xrightarrow[n\to\infty]{\mathbb{P}} \mathbf{0}$, and $\frac{1}{n}\tilde{x}'u \xrightarrow[n\to\infty]{\mathbb{P}} \mathbf{0}$.

As a consequence, the OLS estimator satisfies

$$\begin{aligned}\hat{\beta}_{OLS} &= \beta_0 + (x'x)^{-1}(x'\{u - v\beta_0\}) \\ &= \beta_0 + (x'x)^{-1}(\tilde{x} + v)'(u - v\beta_0) \\ &= \beta_0 + (x'x)^{-1}\tilde{x}'(u - v\beta_0) + (x'x)^{-1}v'(u - v\beta_0) \\ &= \beta_0 + (x'x)^{-1}\tilde{x}'u - (x'x)^{-1}\tilde{x}'v\beta_0 + (x'x)^{-1}v'u - (x'x)^{-1}v'v\beta_0.\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\hat{\beta}_{OLS} &= \operatorname*{plim}_{n\to\infty}\left[\beta_0 + (\frac{1}{n}x'x)^{-1}\frac{1}{n}\tilde{x}'u - (\frac{1}{n}x'x)^{-1}\frac{1}{n}\tilde{x}'v\beta_0 + (\frac{1}{n}x'x)^{-1}\frac{1}{n}v'u - (\frac{1}{n}x'x)^{-1}\frac{1}{n}v'v\beta_0\right] \\ &= \beta_0 + \operatorname*{plim}_{n\to\infty}\underbrace{(\frac{1}{n}x'x)^{-1}\frac{1}{n}\tilde{x}'u}_{\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{0}} - \operatorname*{plim}_{n\to\infty}\underbrace{(\frac{1}{n}x'x)^{-1}\frac{1}{n}\tilde{x}'v\beta_0}_{\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{0}} + \operatorname*{plim}_{n\to\infty}\underbrace{(\frac{1}{n}x'x)^{-1}\frac{1}{n}v'u}_{\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{0}} - \operatorname*{plim}_{n\to\infty}(\frac{1}{n}x'x)^{-1}\frac{1}{n}v'v\beta_0 \\ &= \beta_0 - \operatorname*{plim}_{n\to\infty}\left\{\frac{1}{n}(\tilde{x} + v)'(\tilde{x} + v)\right\}^{-1}\frac{1}{n}v'v\beta_0 \\ &= \beta_0 - \operatorname*{plim}_{n\to\infty}\Big(\underbrace{\frac{1}{n}\tilde{x}'\tilde{x}}_{\xrightarrow[n\to\infty]{\mathbb{P}}\Sigma} + \underbrace{\frac{1}{n}v'\tilde{x}}_{\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{0}} + \underbrace{\frac{1}{n}\tilde{x}'v}_{\xrightarrow[n\to\infty]{\mathbb{P}}\mathbf{0}} + \underbrace{\frac{1}{n}v'v}_{\xrightarrow[n\to\infty]{\mathbb{P}}\Omega}\Big)^{-1}\underbrace{\frac{1}{n}v'v}_{\xrightarrow[n\to\infty]{\mathbb{P}}\Omega}\beta_0 \\ &= \beta_0 - \big(\Sigma + \Omega\big)^{-1}\Omega\beta_0,\end{aligned}$$

by the **continuous mapping theorem**. Hence, we have:

$$\hat{\beta}_{OLS} \xrightarrow[n\to\infty]{\mathbb{P}} \beta_0 - \big(\Sigma + \Omega\big)^{-1}\Omega\beta_0. \tag{1}$$

Remark 1.1 (Measurement Error for Response/Dependent Variable)**.** Let us consider the case that the true regression model becomes

$$\tilde{y} = x\beta_0 + u,$$

and the measurement error exists only in the dependent variable:

$$y = x\beta_0 + u, \tag{2}$$

where $y = \tilde{y} + \underline{w}$ is the observed variable and $\underline{w}$ is the measurement error. Then the regression model is reduced to:

$$\tilde{y} = x\beta_0 + u^* \tag{3}$$

where $u^* = u - \underline{w}$. In this case, the error term is independent from $\tilde{x}$. Therefore, we can apply the usual OLS method to derive the estimator since we can consider that $\underline{w}$ has the same properties as u(, which allows us to use the OLS method to the regression model). Note that the variacnce of u^* is larger than that of u.

2 Measurement Error: Example

Here we exhibit an example in the case of one explanatery variable. Consider the following regression model:

$$y_i = \alpha + \beta x_i + u_i, \quad x_i = \tilde{x}_i + v_i, \quad i = 1, \ldots, n.$$

Let $\mathbb{E}[\tilde{x}_i] = \mu$ and $\mathbb{V}[\tilde{x}_i] = \sigma^2$ for all $i \in \{1, \ldots, n\}$ and

$$\tilde{x} := \begin{pmatrix} 1 & \tilde{x}_1 \\ 1 & \tilde{x}_2 \\ \vdots & \vdots \\ 1 & \tilde{x}_n \end{pmatrix}; \quad v := \begin{pmatrix} 0 & v_1 \\ 0 & v_2 \\ \vdots & \vdots \\ 0 & v_n \end{pmatrix}.$$

Then, from the fact

$$\tilde{x}'\tilde{x} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \tilde{x}_1 & \tilde{x}_2 & \ldots & \tilde{x}_n \end{pmatrix} \begin{pmatrix} 1 & \tilde{x}_1 \\ 1 & \tilde{x}_2 \\ \vdots & \vdots \\ 1 & \tilde{x}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} \tilde{x}_i \\ \sum_{i=1}^{n} \tilde{x}_i & \sum_{i=1}^{n} \tilde{x}_i^2 \end{pmatrix},$$

we obtain

$$\frac{1}{n}\tilde{x}'\tilde{x} = \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \\ \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i & \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^2 \end{pmatrix} \xrightarrow[n\to\infty]{\mathbb{P}} \begin{pmatrix} 1 & \mu \\ \mu & \mu^2 + \sigma_v^2 \end{pmatrix} =: \Sigma.$$

since from the **Weak Law of Large Numbers**,

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[\tilde{x}_i] =: \mu, \quad \text{and}$$
$$\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[\tilde{x}_i^2] = \mathbb{V}[\tilde{x}_i] + (\mathbb{E}[\tilde{x}_i])^2 =: \mu^2 + \sigma^2$$

hold. Moreover, with the assumptions $\mathbb{V}[v_i] = \sigma_v^2$ and $\mathbb{E}[v_i] = 0$,

$$\frac{1}{n}v'v = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^{n} v_i^2 \end{pmatrix} \xrightarrow[n\to\infty]{\mathbb{P}} \begin{pmatrix} 0 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} =: \Omega.$$

Then, according to Eq. (1), we have

$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &\xrightarrow[n\to\infty]{\mathbb{P}} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \left(\begin{pmatrix} 1 & \mu \\ \mu & \mu^2\sigma^2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ &= \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \frac{1}{\sigma^2 + \sigma_v^2} \begin{pmatrix} -\mu\sigma_v^2\beta \\ \sigma_v^2\beta \end{pmatrix} \end{aligned}$$

Consequently, $\hat{\beta}$ is **not consistent**:

$$\hat{\beta}_{OLS} \xrightarrow[n\to\infty]{\mathbb{P}} \beta - \frac{\sigma_v^2}{\sigma^2 + \sigma_v^2}\beta \neq \beta.$$

3 Instrumental Variable

In this section we treat instrumental variables estimation, which is probably second only to ordinary least squares in terms of methods used in empirical economic research.

3.1 Exogenous and Endogeneous

Consider the **structual** regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \tag{4}$$

with $\mathbb{E}[u_i] = 0$ and $\mathrm{Cov}[x_{ij}, u_i] = 0$ for $j \leq k-1$, but x_{ik} might be correlated with u_i. The explanatory variables $x_{i1}, \ldots, x_{i(k-1)}$ are **exogenous** but x_{ik} is potentially **endgeneous** in this equation. The method of instrumental variables (IV) provides a general solution to the problem of an endogeneous variable. To use the IV approach with x_{iK} endogeneous, we need an observable variable z_i, which should satisfy two conditions

- $\mathrm{Cov}[z_i, u_i] = 0$ and $x_{i1}, \ldots, x_{i(k-1)}, z_i$ are **exogenous**;
- $x_{ik} = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{i(k-1)} + \theta z_i + v_i$, where the error term v_i satisfies $\mathbb{E}[v_i] = 0$ and v_i is uncorrelated with $x_{i1}, \ldots, x_{i(k-1)}, z_i$.

The key assumption is that $\theta \neq 0$. This condition means that z_i is partially correlated with x_{ik} once the other variables $x_{i1}, \ldots, x_{i(k-1)}$ have been controlled.

When z_i satisfies both assumptions, it is said to be an **instrumental variable** candidate for x_{ik} (or instrument). Since $x_{i1}, \ldots, x_{i(k-1)}$ are uncorrelated with u_i, they serve as their own instrument: the full list of IV variables is the same as the list of exogenous variables.

3.2 Reduced–Form

Plugging

$$x_{ik} = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_{k-1} x_{i(k-1)} + \theta z_i + v_i$$

into the structural equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i,$$

we obtain the **reduced–form** projection:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{k-1} x_{i(k-1)} + \lambda_k z_i + w_i, \tag{5}$$

where $w_i := u_i + \beta_k v_i$ is the reduced–form error and

$$\begin{aligned} \alpha_j &:= \beta_j + \beta_k \delta_j, \quad j = 1, \ldots, k-1; \\ \lambda &:= \beta_k \theta. \end{aligned}$$

4 Identification Problem

Estimating the structural parameters in (4) is generally more useful than the reduced form parameters. By idenficication, we can write the β_j of Eq. (4) in terms of population moments in observable variables. To do so, write Eq. (4) as

$$y_i = \mathbf{x_i}\beta + u_i,$$

with $\mathbf{x_i} = (1, x_{i2}, \ldots, x_{ik})$. Write the $1 \times k$ exogenous variables as $\mathbf{z_i} = (1, x_{i2}, \ldots, x_{i(k-1)}, z_i)$. The assumptions $\mathbb{E}[u_i] = 0$ and $\text{Cov}[x_{ij}, u_i] = 0$ for $j \in \{1, \ldots, k-1\}$ imply the k orthogonality (or exogeneity) conditions:

$$\mathbb{E}[\mathbf{z_i}'u_i] = \mathbf{0}.$$

Thus, multiplying $y_i = \mathbf{x_i}\beta + u_i$ by $\mathbf{z_i}'$ and taking expectations yields

$$\mathbb{E}[\mathbf{z_i}'\mathbf{x_i}]\beta = \mathbb{E}[\mathbf{z_i}'y_i],$$

where $\mathbb{E}[\mathbf{z_i}'\mathbf{x_i}]$ is a $k \times k$ matrix and $\mathbb{E}[\mathbf{z_i}'y_i]$ is a $k \times 1$ vector. If $\mathbb{E}[\mathbf{z_i}'\mathbf{x_i}]$ is non–singular (rank condition), then we obtain

$$\beta_{IV} = \left(\mathbb{E}[\mathbf{z_i}'\mathbf{x_i}]\right)^{-1} \mathbb{E}[\mathbf{z_i}'y_i]. \tag{6}$$

5 Instrumental Variable Estimator

To derive the previous relationship, we used $\text{Cov}[z_i, u_i] = 0$ and $\theta \neq 0$. Indeed, $\text{Cov}[z_i, u_i] = 0 \iff \theta \neq 0$. Thus, according to Eq. (6), the instrument variable estimator is

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{n} \mathbf{z_i}'\mathbf{x_i}\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{z_i}'y_i\right) \tag{7}$$

The important point is that the condition $\theta \neq 0$ can be tasted (e.g. student test) but $\text{Cov}[z_i, u_i] = 0$ must be maintained: indeed the covariance involves the unobservable u_i and therefore we cannnot test anything about $\text{Cov}[z_i, u_i]$.

6 Partial Identification

In general, it is difficult for researchers to identify an instrumental variable (IV). Moreover, even if an IV is found, it might still be correlated with the error term. To address this issue, we apply partial identification. In partial identification, we estimate the range (band) of parameters instead of a specific value (point estimation). If you are interested in partial identification, please read the following articles:

- Edward E. Leamer (1981) "Is it a Demand Curve, Or Is It A Supply Curve? Partial Identification through Inequality Constraints," The Review of Economics and Statistics, `https://www.jstor.org/stable/1924348`
- Nevo, Aviv and Rosen, Adam (2012) "Identification With Imperfect Instruments," The Review of Economics and Statistics, `https://EconPapers.repec.org/RePEc:tpr:restat:v:94:y:2012:i:3:p:659-671`