ECO 310: Empirical Industrial Organization

Lecture 3: Production Functions: The simultaneity problem

Victor Aguirregabiria (University of Toronto)

September 24, 2018

Outline

1. Simultaneity problem: Definition

2. Simultaneity problem: Bias of OLS

3. Simultaneity problem: Solutions

  3.1. Restrictions in simultaneous equations model

  - **3.2.** Instrumental variables estimation
    3.3. Control function estimation

1. Endogeneity / Simultaneity Problem: Definition

Endogeneity / Simultaneity problem

- Consider the PF:

$$y_{it} = \alpha_L \ \ell_{it} + \alpha_K \ k_{it} + \omega_{it} + e_{it}$$

- We are interested in the estimation of α_L and α_K. These parameters represent "ceteris paribus" causal effects of labor and capital on output, respectively.

- When the manager decides the optimal (k_{it}, ℓ_{it}) she has some information about log-TFP ω_{it} (that we do not observe).

- This means that there is a correlation between the observable inputs (k_{it}, ℓ_{it}) are correlated with the unobservable ω_{it}.

- This correlation implies that the OLS estimators of α_L and α_K are biased and inconsistent.

Endogeneity problem: General description

- First, let's consider a Linear Regression Model (LRM) with one regressor:

$$y_i = \alpha + \beta \, x_i + \varepsilon_i$$

- We have an **endogeneity problem** if the regressor x_i is correlated with the error term ε_i. In other words,

$$\text{Endogeneity problem} \;\; \Leftrightarrow \;\; \mathbb{E}(x_i \; \varepsilon_i) \neq 0$$

- It is a problem because, in this situation, **the OLS estimator does not provide a consistent estimator of the parameter** β [of the causal effect of x on y when the rest of the variables remain constant; **ceteris paribus** effect.]

- We are going to see: (1) Bias of the OLS estimator; (2) Solution: Instrumental variables approach; and (3) Solution: Control function approach.

2. Simultaneity Problem: Bias OLS

Endogeneity problem: Bias of OLS

- The OLS estimator of the slope parameter β is defined as:

$$\widehat{\beta}_{OLS} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \ \ (x_i - \overline{x})}{\sum_{i=1}^{N} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- According to the model:

$$\begin{array}{rcl} y_i & = & \alpha + \beta \ x_i + \varepsilon_i \\ \overline{y} & = & \alpha + \beta \ \overline{x} + \overline{\varepsilon} \end{array}$$

- Such that

$$(y_i - \overline{y}) = \beta \ (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})$$

and

$$(y_i - \overline{y}) (x_i - \overline{x}) = \beta \ (x_i - \overline{x})^2 + (\varepsilon_i - \overline{\varepsilon}) (x_i - \overline{x})$$

Endogeneity problem: Bias of OLS (2)

- This implies that:

$$\sum_{i=1}^{N} (y_i - \overline{y}) (x_i - \overline{x}) = \beta \sum_{i=1}^{N} (x_i - \overline{x})^2 + \sum_{i=1}^{N} (\varepsilon_i - \overline{\varepsilon}) (x_i - \overline{x})$$

- Or:

$$S_{xy} = \beta \; S_{xx} + S_{x\varepsilon}$$

- Therefore, dividing in this expression by S_{xx}, we have that:

$$\widehat{\beta}_{OLS} \equiv \frac{S_{xy}}{S_{xx}} = \beta + \frac{S_{x\varepsilon}}{S_{xx}}$$

- $\widehat{\beta}_{OLS}$ is a measure of the correlation between x and y. In general, this measure of correlation does not give us the causal effect of x on y, as measured by the parameter β.
- Only if $S_{x\varepsilon} = 0$ we have that $\widehat{\beta}_{OLS} = \beta$ and the OLS is a consistent estimator of the causal effect β.

Endogeneity problem: How do we know?

- How do we know that $\mathbb{E}(x_i \ \varepsilon_i) = 0$ or $S_{x\varepsilon} = 0$?
- In general we don't know, but in many cases we can have serious suspicion of omitted variables that are correlated with the regressor(s).
- Only when the observable regressor comes from a randomized experiment we can be certain that $S_{x\varepsilon} = 0$.
- But data from randomized experiments is still rare in many applications in economics.

Endogeneity problem: How do we know? (2)

- In models with **simultaneous equations**, the model itself can tell us that some of the regressors are correlated with the error term: $\mathbb{E}(x_i \varepsilon_i) \neq 0$.
- For instance, this is the case in the production function model once we take into account the firm's optimal demand for inputs.

Endogeneity / Simultaneity: Example

- Firms operate in the same markets for output and inputs. Same output and input prices: P and W.
- A Cobb-Douglas PF only with labor input:

$$Y_i = A_i \; L_i^{\alpha}$$

- Firm i's profit is:

$$\pi_i = P \; Y_i - W \; L_i$$

- The marginal condition of optimality for profit maximization give us the Labor Demand (LD) equation:

$$\alpha \frac{Y_i}{L_i} = \frac{W}{P}$$

- PF and LD in logarithms:

$$\begin{array}{llll} \text{(PF)} & y_i & = & \alpha \; \ell_i + \omega_i \\ \\ \text{(LD)} & \ell_i & = & y_i - w \end{array}$$

Endogeneity / Simultaneity: Example (cont)

- This is a **system of simultaneous equations** with 2 equations and 2 endogenous variables, y_i and ℓ_i:

$$\begin{array}{lrcl} \text{(PF)} & y_i & = & \alpha\ \ell_i + \omega_i \\ \\ \text{(LD)} & \ell_i & = & y_i - w \end{array}$$

- If we solve this system, we obtain y_i and ℓ_i as functions of exogenous variables only:

$$\begin{array}{rcl} y_i & = & \dfrac{\omega_i - \alpha\ w}{1-\alpha} \\ \\ \ell_i & = & \dfrac{\omega_i - w}{1-\alpha} \end{array}$$

- This expression shows that ℓ_i is correlated with the error term in the PF, ω_i.

Endogeneity / Simultaneity: Example (Bias OLS)

- If we continue with this example, we can derive the bias of the OLS estimator **as N is large**.

$$\widehat{\alpha}_{OLS} = \frac{S_{\ell y}}{S_{\ell\ell}} = \frac{\sum_{i=1}^{N} \widetilde{\ell}_i \, \widetilde{y}_i}{\sum_{i=1}^{N} \widetilde{\ell}_i \, \widetilde{\ell}_i}$$

where $\widetilde{\ell}_i \equiv \ell_i - \overline{\ell}$ and $\widetilde{y}_i \equiv y_i - \overline{y}$

- And:

$$\widetilde{y}_i \equiv y_i - \overline{y} = \frac{\omega_i - \overline{\omega}}{1-\alpha}$$

$$\widetilde{\ell}_i \equiv \ell_i - \overline{\ell} = \frac{\omega_i - \overline{\omega}}{1-\alpha}$$

- Such that $\widehat{\alpha}_{OLS} = 1$ and Bias(OLS) is $1 - \alpha$.

Simultaneity: Graphical representation

- Graphical representation of structural equations in space (ℓ, y).
- Graphical interpretation of the bias of the OLS estimator.
- With sample variation in the log-real-wage w_i the bias will be reduced, but it will be always present.

3. Simultaneity Problem: Solutions

Solutions to Endogeneity problem

- We are going to consider several (potential) solutions to the endogeneity problem.

    1. Exploiting restrictions in simultaneous equations model.
    2. Instrumental variables estimation.
    3. Control function estimation.

- First, we will see these potential solutions in a general regression model, and then we will particularize them to the estimation of PFs.

Solutions to Endogeneity: Restrictions in model

- Sometimes, the model of simultaneous equations implies restrictions that provide information of the parameter(s) of interest.

- In the case of the PF estimation these restrictions typically come from the marginal conditions of optimality in the demand for inputs.

- For illustration, consider the example with only labor input. The marginal condition is $\alpha \frac{Y_i}{L_i} = \frac{W}{P}$, and in logs:

$$y_i - \ell_i = \ln(W/P) - \ln(\alpha)$$

- Or

$$\ln(\alpha) = \overline{\ell} - \overline{y} + \ln(W/P)$$

- Mean values $\overline{\ell}$ and $\overline{y}$, together with info on $\ln(W/P)$, give us a consistent estimator of $\ln(\alpha)$ and of α.

Solutions to Endogeneity: Instrumental variables (IV)

- Consider the LRM

$$y_i = \beta_1 \; x_{1i} + ... + \beta_K \; x_{Ki} + \varepsilon_i$$

where we are concerned about the endogeneity of regressor x_{1i}, i.e., $\mathbb{E}(x_{1i} \; \varepsilon_i) \neq 0$.

- Suppose that the researcher has sample data for a variable z_i ("the instrument") that satisfies **two conditions**.

- **[Relevance]** In a regression of x_{1i} on $(z_i, x_{2i}, ..., x_{Ki})$, regressor z_i has a significant effect on x_{1i}.

- **[Independence]** z_i is NOT correlated with ε_i: $\mathbb{E}(z_i \; \varepsilon_i) = 0$.

- Under these conditions we can construct a consistent estimator of β_1, β_2, ..., β_K: the IV or Two-state Least Square (2SLS) estimator.

Two-Stage Least Square (2SLS or IV)

- The IV or 2SLS can be implemented as follows.
- **[Stage 1]** Run an OLS regression of x_{1i} on $(z_i, x_{2i}, ..., x_{Ki})$. Obtain the fitted values from this regression:

$$\widehat{x}_{1i} = \widehat{\gamma}_0 + \widehat{\gamma}_1 z_i + \widehat{\gamma}_2 x_{2i} + ... + \widehat{\gamma}_K x_{Ki}$$

- **[Stage 2]** Run an OLS regression of y_i on $(\widehat{x}_{1i}, x_{2i}, ..., x_{Ki})$. This OLS estimator is consistent for $\beta_1, \beta_2, ..., \beta_K$.
- The first stage decomposes x_{1i} in two parts: $x_{1i} = \widehat{x}_{1i} + e_{1i}$, where e_{1i} is the residual from this first-stage regression.
- Since $\widehat{x}_{1i}$ depends only on exogenous regressors, it is not correlated with ε_i.

Consistency of IV / 2SLS

- To illustrate how this approach give us a consistent estimator, consider the model with a single regressor: $y_i = \alpha + \beta \; x_i + \varepsilon_i$.

- Remember that:

$$(y_i - \overline{y}) = \beta \; (x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})$$

- Such that multiplying by $(z_i - \overline{z})$:

$$(y_i - \overline{y}) \, (z_i - \overline{z}) = \beta \, (x_i - \overline{x}) \, (z_i - \overline{z}) + (\varepsilon_i - \overline{\varepsilon}) \, (z_i - \overline{z})$$

- And summing over observations i:

$$S_{zy} = \beta \; S_{zx} + S_{z\varepsilon}$$

- Since $S_{z\varepsilon} = 0$, we have that, for large N:

$$\frac{S_{zy}}{S_{zx}} = \beta$$

Consistency of IV / 2SLS (cont)

- This means that the estimator $\widehat{\beta}_{IV} = \frac{S_{zy}}{S_{zx}} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \ (z_i - \overline{z})}{\sum_{i=1}^{N} (x_i - \overline{x}) (z_i - \overline{z})}$ is a consistent estimator of β.

- It remains to show that this $\widehat{\beta}_{IV} = \frac{S_{zy}}{S_{zx}}$ is identical to the 2SLS described above.

- By definition:

$$\widehat{\beta}_{2SLS} = \frac{S_{\widehat{x}y}}{S_{\widehat{x}\widehat{x}}} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \ \left(\widehat{x}_i - \overline{\widehat{x}}\right)}{\sum_{i=1}^{N} \left(\widehat{x}_i - \overline{\widehat{x}}\right) \left(\widehat{x}_i - \overline{\widehat{x}}\right)}$$

where $\widehat{x}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 z_i$, with $\widehat{\gamma}_1 = \frac{S_{zx}}{S_{zz}}$.

Consistency of IV / 2SLS (cont)

- Therefore, $\widehat{x}_i - \overline{\widehat{x}} = \widehat{\gamma}_1 (z_i - \overline{z}) = \dfrac{S_{zx}}{S_{zz}} (z_i - \overline{z})$.

- Such that

$$\widehat{\beta}_{2SLS} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \ \dfrac{S_{zx}}{S_{zz}} (z_i - \overline{z})}{\sum_{i=1}^{N} \dfrac{S_{zx}}{S_{zz}} (z_i - \overline{z}) \ \dfrac{S_{zx}}{S_{zz}} (z_i - \overline{z})} = \frac{S_{yz} \ \dfrac{S_{zx}}{S_{zz}}}{S_{zz} \dfrac{S_{zx}}{S_{zz}} \dfrac{S_{zx}}{S_{zz}}} = \frac{S_{yz}}{S_{zx}}$$

- The 2SLS is equivalent to the IV estimator as defined above.

How to obtain instruments?

- A simultaneous equation model may suggest valid instruments.
- For instance, consider the PF with only labor input, but now firms operate in different output/labor markets with different prices.

$$\begin{array}{lrcl} \text{(PF)} & y_i & = & \alpha \, \ell_i + \omega_i \\ \text{(LD)} & \ell_i & = & \ln(\alpha) + y_i - w_i \end{array}$$

with $w_i = \ln(W_i / P_i)$.

- Suppose that the researcher observes w_i.
- It is clear that w_i satisfies the **relevance condition**: it does not enter in the PF as a regressor; it has an effect on labor.
- Under the condition $\mathbb{E}(w_i \ \omega_i) = 0$ it is a valid instrument.

Control Function (CF) Method

- Consider the LRM

$$y_i = \beta_1 \ x_{1i} + ... + \beta_K \ x_{Ki} + \varepsilon_i$$

where we are concerned about the endogeneity of regressor x_{1i}, i.e., $\mathbb{E}(x_{1i} \ \varepsilon_i) \neq 0$.

- Suppose that the researcher has sample data for a variable c_i ("the control") that satisfies **two conditions**.

- **[Control]** $\varepsilon_i = \gamma \ c_i + u_i$ such that u_i is independent of x_{1i} and c_i.

- **[No multicollinearity]** We cannot write c_i as a linear combination of the exogenous regressors x_{2i}, ..., x_{Ki}.

- Under these conditions we can construct a consistent estimator of β_1, β_2, ..., β_K: the Control Function (CF) estimator.

Control Function (CF) estimator

- To obtain the CF estimator we simply include the CF variable c_i in the regression and apply OLS:

$$y_i = \beta_1 \ x_{1i} + ... + \beta_K \ x_{Ki} + \gamma c_i + u_i$$

- Under the "Control" condition, the new error term u_i is not correlated with the regressors.
- And under the "No multicollinearity" condition all the regressors (including c_i) are not linearly independent.
- Therefore, this OLS estimator is consistent.
- The CF approach uses observables to control for the part of the error that is correlated with the regressor.