ECO 310: Empirical Industrial Organization

Lecture 3: PRODUCTION FUNCTIONS: THE SIMULTANEITY PROBLEM

Victor Aguirregabiria (University of Toronto)

September 22, 2022

OUTLINE

1. Simultaneity problem: Definition
2. Simultaneity problem: Bias of OLS
3. Simultaneity problem: Solutions
   - 3.1. Fixed Effects estimation
   - 3.2. Instrumental variables estimation

1. Simultaneity Problem: Definition

SIMULTANEITY PROBLEM

- Consider the Cobb-Douglas PF in logarithms:

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

- We want to estimate parameters α_L and α_K. These parameters represent the causal effects of labor and capital on output.

- When the manager decides the optimal (k_{it}, ℓ_{it}) she has some information about log-TFP ω_{it}.

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

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

SIMULTANEITY PROBLEM: GENERAL DESCRIPTION

- Consider a Linear Regression Model (LRM) with one regressor:

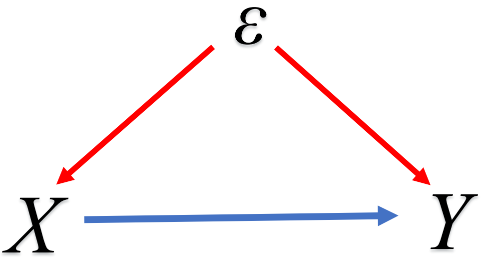
$$y_i = \alpha + \beta \; x_i + \varepsilon_i$$

- We have an **simultaneity problem** (or **endogeneity problem**) if the regressor x_i is correlated with the error term ε_i.

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

- It is a problem because it implies that the **OLS estimator of β is not consistent**: it does not give us the causal effect of x on y.

SIMULTANEITY PROBLEM: GENERAL DESCRIPTION



2. Simultaneity Problem: Biased OLS

SIMULTANEITY PROBLEM: BIAS OF OLS

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

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

- 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

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

SIMULTANEITY PROBLEM: BIAS OF OLS (2/2)

- This implies that:

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

- 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.

SIMULTANEITY PROBLEM: HOW DO WE KNOW?

- How do we know whether $\mathbb{E}(x_i\ \varepsilon_i) = 0$ or $\mathbb{E}(x_i\ \varepsilon_i) \neq 0$?
- 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 $\mathbb{E}(x_i\ \varepsilon_i) = 0$.
- But data from randomized experiments are still rare in many applications in economics.

SIMULTANEITY PROBLEM: HOW DO WE KNOW? (2/2)

- In models with **simultaneous equations**, the model itself can tell us that some regressors are correlated with the error term: $\mathbb{E}(x_i \, \varepsilon_i) = 0$.

- For instance, this is the case in the production function model once we take into account the firm's optimal demand for inputs.

SIMULTANEITY PROBLEM: EXAMPLE (1/3)

- A Cobb-Douglas PF only with labor input:

$$Y_i = A_i \ L_i^{\alpha_L}$$

- The amounts of output (Y_i) and labor (L_i) are endogenous variables which are determined by the conditions of profit maximization.

- Firms operate in the same markets for output and inputs. Same output and input prices: P and W.

- A firm's profit is:

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

- A firm's Labor Demand is the amount L_i that maximizes profit:

$$\frac{d\pi_i}{dL_i} = 0 \ \rightarrow \ MP_{L_i} = \frac{W}{P} \rightarrow \alpha_L \ \frac{Y_i}{L_i} = \frac{W}{P}$$

SIMULTANEITY PROBLEM: EXAMPLE (2/3)

- The complete model consists of the Production Function (PF) and the Labor Demand equation (LD):

$$\text{(PF)} \quad Y_i \;=\; A_i \; L_i^{\alpha_L}$$

$$\text{(LD)} \quad L_i \;=\; \alpha_L \, \frac{Y_i}{W/P}$$

- This is a system of two equations with two endogenous variables.

- We can take logarithms in these equations to have a model that is linear in parameters (a linear regression model):

$$\text{(log-PF)} \quad y_i \;=\; \alpha_0 + \alpha_L \; \ell_i + \omega_i$$

$$\text{(log-LD)} \quad \ell_i \;=\; \gamma_0 + y_i$$

where α_0 is the mean value of $\ln(A_i)$; $\omega_i = \ln(A_i) - \alpha_0$; and $\gamma_0 = \ln(\alpha_L P/W)$.

SIMULTANEITY PROBLEM: EXAMPLE (3/3)

- Solving for the endogenous variables in the **system of equations**,

$$\begin{array}{lrcl} \text{(log-PF)} & y_i & = & \alpha_0 + \alpha_L \; \ell_i + \omega_i \\ \text{(log-LD)} & \ell_i & = & \gamma_0 + y_i \end{array}$$

- we obtain the solution:

$$y_i = \frac{\omega_i + \alpha_0 + \alpha_L \; \gamma_0}{1 - \alpha_L}$$

$$\ell_i = \frac{\omega_i + \alpha_0 + \gamma_0}{1 - \alpha_L}$$

- This solution shows that ℓ_i is correlated with ω_i:

$$Cov(\ell_i, \omega_i) = \frac{Var(\omega_i)}{(1 - \alpha_L)^2} > 0$$

SIMULTANEITY PROBLEM: EXAMPLE – BIASED OLS

- Following up with this example, we can show that the OLS estimator of α_L is biased.

$$\widehat{\alpha}_L^{OLS} = \frac{S_{\ell y}}{S_{\ell\ell}} = \frac{\sum_{i=1}^{N} (y_i - \overline{y}) \left(\ell_i - \overline{\ell}\right)}{\sum_{i=1}^{N} \left(\ell_i - \overline{\ell}\right)^2}$$

- The model implies that:

$$y_i - \overline{y} = \frac{\omega_i}{1 - \alpha_L}$$

$$\ell_i - \overline{\ell} = \frac{\omega_i}{1 - \alpha_L}$$

- Such that $\widehat{\alpha}_L^{OLS} = \frac{S_{\ell y}}{S_{\ell\ell}} = 1$, and $Bias(OLS) = 1 - \alpha_L$.

SIMULTANEITY PROBLEM: 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 THE SIMULTANEITY PROBLEM

- We are going to consider two possible solutions to the endogeneity problem.

    1. Control function / Fixed effects estimation
    2. Instrumental variables estimation.

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

CONTROL FUNCTION METHOD

- Consider the LRM

$$y_i = \beta_0 + \beta_1 \ x_{1i} + ... + \beta_K \ x_{Ki} + \varepsilon_i$$

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 ESTIMATOR

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

$$y_i = \beta_0 + \beta_1 \ x_{1i} + ... + \beta_K \ x_{Ki} + \gamma \ c_i + u_i$$

- 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.

SOLUTIONS TO SIMULTANEITY: INSTRUMENTAL VARIABLES

- 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 SQUARES (2SLS or IV ESTIMATOR)

- 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 β_1, β_2, ..., β_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 (2/3)

- 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 (3/3)

- 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 DO WE FIND 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.