UG Empirical Industrial Organization

Lecture 3: Production Functions: The Simultaneity Problem

Anubhav Jha[1]
Ashoka University

March 11, 2024

[1] This course is based on Victor Aguirregabiria's Empirical IO book. The slides (in my first year of teaching) are extremely close to his slides. These will change in the future iterations of this course.

Outline

- Simultaneity problem: Definition
- Simultaneity problem: Bias of OLS
- Simultaneity problem: Solutions
    - Control Function estimation
    - Instrumental variables estimation

Simultaneity Problem

- Consider the Cobb-Douglas PF in logarithms:

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

- Note here $\omega_{it} = \log(A_{it})$. Note that $\mathbb{E}\left[\omega_i\right] \neq 0$ in this regression.[2]
- 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.

[2] The reason I am not using an intercept here is because in Lecture-4 we allow for Fixed-Effects which is more general than a constant term. In examples given in these notes, we allow for intercepts but that will change in Lecture-4.

Simultaneity Problem: General Description

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

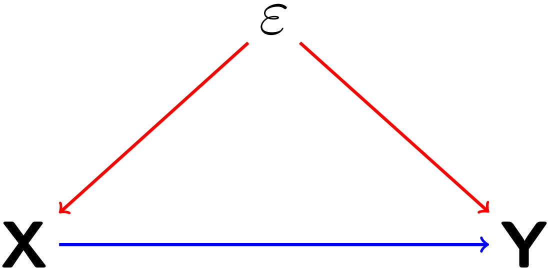
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

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

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

- 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



Simultaneity Problem: Bias of OLS

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

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

- According to the model:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$
$$\bar{y} = \alpha + \beta \bar{x} + \varepsilon$$

- Such that:

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

- and

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

Simultaneity Problem: Bias of OLS (2/2)

- This implies that:

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

- Or:

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

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

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

- $\hat{\beta}_{\text{OLS}}$ 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 $S_{\varepsilon x} = 0$ we have that $\hat{\beta}_{\text{OLS}} = \beta$ 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:

$$\mathbb{E}(x_i \varepsilon_i) \neq 0.$$

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

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:

$$\pi_i = PY_i - WL_i$$

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

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

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 $\alpha_0 = \mathbb{E}\left[\ln(A_i)\right]$; $\omega_i = \ln(A_i) - \alpha_0 \Rightarrow \mathbb{E}\left[\omega_i\right] = 0$; and $\gamma_0 = \ln(\alpha_L P/W)$.

Simultaneity Problem: Example (3/3)

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

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

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

- 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:

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

Simultaneity Problem: Example – Biased OLS

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

$$\hat{\alpha}_L^{\mathrm{OLS}} = \frac{S_{y\ell}}{S_{\ell\ell}} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\ell_i - \bar{\ell})}{\sum_{i=1}^{N}(\ell_i - \bar{\ell})^2}$$

- The model implies that:

$$y_i - \bar{y} = \frac{\omega_i}{1 - \alpha_L} \quad \text{and} \quad \ell_i - \bar{\ell} = \frac{\omega_i}{1 - \alpha_L}$$

- Such that

$$\hat{\alpha}_L^{\mathrm{OLS}} = \frac{S_{y\ell}}{S_{\ell\ell}} = 1, \quad \text{and} \quad \mathrm{Bias(OLS)} = 1 - \alpha_L.$$

Simultaneity Problem: Graphical Representation

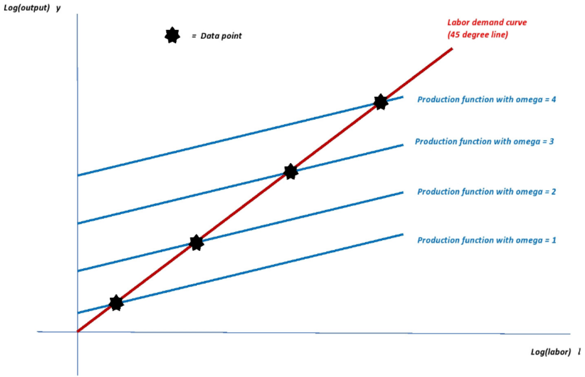

Log(output) y
= Data point
Labor demand curve
(45 degree line)
Production function with omega = 4
Production function with omega = 3
Production function with omega = 2
Production function with omega = 1
Log(labor) l

Figure: Graphical Representation of Simultaneity Problem

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

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

  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 $x_{2i}, \ldots, x_{Ki}$.
- Under these conditions we can construct a consistent estimator of $\beta_1, \beta_2, \ldots, \beta_K$: 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:

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

- 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 dependent.
- 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

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

  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 $(z_i, x_{2i}, \ldots, x_{Ki})$, 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 $\beta_1, \beta_2, \ldots, \beta_K$: the IV or Two-stage 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 $(z_i, x_{2i}, \dots, x_{Ki})$. Obtain the fitted values from this regression:

$$\hat{x}_{1i} = \gamma_0 + \gamma_1 z_i + \gamma_2 x_{2i} + \cdots + \gamma_K x_{Ki}$$

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

Consistency of IV / 2SLS

- To illustrate how this approach gives us a consistent estimator, consider the model with a single regressor:

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

- Remember that:

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

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

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

- 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

  $$\hat{\beta}_{\text{IV}} = \frac{S_{zy}}{S_{zx}} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}$$

  is a consistent estimator of β.
- It remains to show that this $\hat{\beta}_{\text{IV}} = \frac{S_{zy}}{S_{zx}}$ is identical to the 2SLS described above.
- By definition:

  $$\hat{\beta}_{\text{2SLS}} = \frac{S_{\hat{x}y}}{S_{\hat{x}\hat{x}}} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{x}_i - \bar{\hat{x}})}{\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2}$$

- where $\hat{x}_i = \gamma_0 + \gamma_1 z_i$, with $\gamma_1 = \frac{S_{xz}}{S_{zz}}$.

Consistency of IV / 2SLS (3/3)

- Therefore,

$$\hat{x}_i - \bar{\hat{x}} = \gamma_1(z_i - \bar{z}) = \frac{S_{xz}}{S_{zz}}(z_i - \bar{z}).$$

- Such that

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})\frac{S_{xz}}{S_{zz}}(z_i - \bar{z})}{\sum_{i=1}^{N}\left(\frac{S_{xz}}{S_{zz}}(z_i - \bar{z})\right)^2} = \frac{S_{yz} \cdot S_{xz}/S_{zz}}{S_{xz}^2/S_{zz}} = \frac{S_{yz}}{S_{zx}}.$$

- 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.

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

$$\text{(LD)} \quad \ell_i = \ln(\alpha_L) + y_i - w_i$$

  with $w_i = \ln(W_i/P_i)$ and $\mathbb{E}[\omega_i] = 0$ here because the intercept term is included.
- 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.

Implementation in R

$$Y_i = \beta_0 + \beta_1 * X_i + \beta_2 * D_i + e_i$$

$$X_i = \gamma_0 + \gamma_1 * Z_i + \gamma_2 * D_i + u_i$$

$$Cov(u_i, e_i) \neq 0 \Rightarrow Cov(X_i, e_i) \neq 0$$

```
# Run OLS (naive regression, biased if x is
   endogenous)
ols_model <- feols(y ~ x+D, data = df)
# Run IV regression: instrument x with z
iv_model <- feols(y ~ 1 + D| x ~ z, data = df)
```

Implementation in R with Fixed Effects

$$Y_{it} = \beta_i + \beta_1 * X_{it} + \beta_2 * D_{it} + e_{it}$$

$$X_{it} = \gamma_i + \gamma_1 * Z_{it} + \gamma_2 * D_{it} + u_{it}$$

$$Cov(u_{it}, e_{it}) \neq 0 \Rightarrow Cov(X_{it}, e_{it}) \neq 0$$

where β_i and $gamma_i$ are Firm level fixed effects.

```
# OLS with Firm level FE (biased if x is
   endogenous)
ols_model <- feols(y ~ x+D | firm_id, data = df)
# Instrument x with z with Firm level fixed
   effects
iv_model <- feols(y ~ 1 + D| firm_id | x ~ z,
   data = df)
```