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
Consider the PF:

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

We are interested in the estimation of \( \alpha_L \) and \( \alpha_K \). These parameters represent "ceteris paribus" causal effects of labor and capital on output, respectively.

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

This means that there is a correlation between the observable inputs \((k_{it}, \ell_{it})\) are correlated with the unobservable \(\omega_{it}\).

This correlation implies that the OLS estimators of \(\alpha_L\) and \(\alpha_K\) 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 \( \varepsilon_i \). In other words,
  \[ \text{Endogeneity problem } \iff \mathbb{E}(x_i \varepsilon_i) \neq 0 \]

- It is a problem because, in this situation, the **OLS estimator does not provide a consistent estimator of the parameter** \( \beta \) [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 $\beta$ is defined as:

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

- According to the model:

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

$$\bar{y} = \alpha + \beta \bar{x} + \bar{\epsilon}$$

- Such that

$$(y_i - \bar{y}) = \beta (x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon})$$

and

$$(y_i - \bar{y}) (x_i - \bar{x}) = \beta (x_i - \bar{x})^2 + (\epsilon_i - \bar{\epsilon}) (x_i - \bar{x})$$
Endogeneity problem: Bias of OLS (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:
  \[
  S_{xy} = \beta S_{xx} + S_{x\varepsilon}
  \]

- Therefore, dividing in this expression by \( S_{xx} \), we have that:
  \[
  \hat{\beta}_{OLS} \equiv \frac{S_{xy}}{S_{xx}} = \beta + \frac{S_{x\varepsilon}}{S_{xx}}
  \]

- \( \hat{\beta}_{OLS} \) 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 \( \beta \).

- Only if \( S_{x\varepsilon} = 0 \) we have that \( \hat{\beta}_{OLS} = \beta \) and the OLS is a consistent estimator of the causal effect \( \beta \).
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.
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:
  \[
  (\text{PF}) \quad y_i = \alpha \ l_i + \omega_i \\
  (\text{LD}) \quad l_i = y_i - w
  \]
Endogeneity / Simultaneity: Example (cont)

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

  \[
  \begin{align*}
  \text{(PF)} & \quad y_i = \alpha \ell_i + \omega_i \\
  \text{(LD)} & \quad \ell_i = y_i - w
  \end{align*}
  \]

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

  \[
  \begin{align*}
  y_i &= \frac{\omega_i - \alpha w}{1 - \alpha} \\
  \ell_i &= \frac{\omega_i - w}{1 - \alpha}
  \end{align*}
  \]

- This expression shows that $\ell_i$ is correlated with the error term in the PF, $\omega_i$. 
If we continue with this example, we can derive the bias of the OLS estimator as $N$ is large.

$$\hat{\alpha}_{OLS} = \frac{S_{ly}}{S_{ll}} = \frac{\sum_{i=1}^{N} \tilde{l}_i \tilde{y}_i}{\sum_{i=1}^{N} \tilde{l}_i \tilde{l}_i}$$

where $\tilde{l}_i \equiv l_i - \bar{l}$ and $\tilde{y}_i \equiv y_i - \bar{y}$

And:

$$\tilde{y}_i \equiv y_i - \bar{y} = \frac{\omega_i - \bar{\omega}}{1 - \alpha}$$

$$\tilde{l}_i \equiv l_i - \bar{l} = \frac{\omega_i - \bar{\omega}}{1 - \alpha}$$

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

Simultaneity: Graphical representation

- Graphical representation of structural equations in space \((\ell, y)\).
- 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
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 $\frac{Y_i}{L_i} = \frac{W}{P}$, and in logs:

  $$y_i - \ell_i = \ln\left(\frac{W}{P}\right) - \ln(\alpha)$$

- Or

  $$\ln(\alpha) = \bar{\ell} - \bar{y} + \ln\left(\frac{W}{P}\right)$$

- Mean values $\bar{\ell}$ and $\bar{y}$, together with info on $\ln\left(\frac{W}{P}\right)$, give us a consistent estimator of $\ln(\alpha)$ and of $\alpha$. 
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 \( \varepsilon_i \): \( \mathbb{E}(z_i \varepsilon_i) = 0 \).

- Under these conditions we can construct a consistent estimator of \( \beta_1, \beta_2, ..., \beta_K \): 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 $(z_i, x_{2i}, \ldots, x_{Ki})$. Obtain the fitted values from this regression:
  \[
  \hat{x}_{1i} = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + \hat{\gamma}_2 x_{2i} + \ldots + \hat{\gamma}_K x_{Ki}
  \]
- **[Stage 2]** Run an OLS regression of $y_i$ on $(\hat{x}_{1i}, x_{2i}, \ldots, x_{Ki})$. This OLS estimator is consistent for $\beta_1, \beta_2, \ldots, \beta_K$.
- 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 $\varepsilon_i$. 
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 \(N\):
  \[
  \frac{S_{zy}}{S_{zx}} = \beta
  \]
This means that the estimator \( \hat{\beta}_{IV} = \frac{S_{zy}}{S_{zx}} = \frac{\sum_{i=1}^{N} (y_i - \bar{y}) (z_i - \bar{z})}{\sum_{i=1}^{N} (x_i - \bar{x}) (z_i - \bar{z})} \) is a consistent estimator of \( \beta \).

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

By definition:

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

where \( \hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i \), with \( \hat{\gamma}_1 = \frac{S_{zx}}{S_{zz}} \).
Therefore, \( \hat{x}_i - \bar{x} = \hat{\gamma}_1 (z_i - \bar{z}) = \frac{S_{zx}}{S_{zz}} (z_i - \bar{z}) \).

Such that

\[
\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{N} (y_i - \bar{y}) \frac{S_{zx}}{S_{zz}} (z_i - \bar{z})}{\sum_{i=1}^{N} \frac{S_{zx}}{S_{zz}} (z_i - \bar{z})} = \frac{S_{yz} \frac{S_{zx}}{S_{zz}}}{S_{zz} \frac{S_{zx}}{S_{zz}} S_{zz}} = \frac{S_{yz}}{S_{zx}}
\]

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.

\[
\text{(PF)} \quad y_i = \alpha \ell_i + \omega_i
\]

\[
\text{(LD)} \quad \ell_i = \ln(\alpha) + y_i - w_i
\]

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 \( \mathbb{E}(w_i \mid \omega_i) = 0 \) it is a valid instrument.
Control Function (CF) Method

- Consider the LRM

\[ y_i = \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \]

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 (CF) estimator

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

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

- 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.