منتدى البحوث الاقتصادية
ECONOMIC RESEARCH FORUM

Identification Problem and Endogeneity: An Overview

By Chahir Zaki

Training on Applied Micro-Econometrics and Public Policy Evaluation

Economic Research Forum

Causation vs. Correlation

- Correlation: Two economic variables are correlated if they move together (example: height and weight across individuals)
- Causality: Two economic variables are causally related if the movement of one causes movement of the other (example: good nutrition as an infant increases adult height)
- There are many examples where causation and correlation can get confused.
- In statistics, this is called the identification problem: given that two series are correlated, how do you identify whether one series is causing another?

Estimating Treatment Effects

Consider treatment assignment (dummy variable) X and outcome Y

Regress Y on X

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

The estimate of β_1 is just the difference between the mean Y for X = 1 (the treatment group) and the mean Y for X = 0 (the control group)

$$\bar{Y}_{1\bullet} = \beta_0 + \beta_1 + \bar{\varepsilon}_{1\bullet}$$

$$\bar{Y}_{0\bullet} = \beta_0 + \bar{\varepsilon}_{0\bullet}$$

Thus the OLS estimate is

$$\bar{Y}_{1\bullet} - \bar{Y}_{0\bullet} = \beta_1 + \left(\bar{\varepsilon}_{1\bullet} - \bar{\varepsilon}_{0\bullet}\right)$$

Estimating Treatment Effects (With Random Assignment)

If the treatment is randomly assigned, then *X* is uncorrelated with *ε* (*X* is exogenous)

If X is uncorrelated with *ε* if and only if $\bar{\varepsilon}_{1\bullet} = \bar{\varepsilon}_{0\bullet}$

But if $\bar{\varepsilon}_{1\bullet} = \bar{\varepsilon}_{0\bullet}$, then the mean difference is

$$\bar{Y}_{1\bullet} - \bar{Y}_{0\bullet} = \beta_1 + \left(\bar{\varepsilon}_{1\bullet} - \bar{\varepsilon}_{0\bullet}\right) \qquad = \beta_1$$

This implies that standard methods (OLS) give an unbiased estimate of β_1, which is the average treatment effect

That is, the treatment-control mean difference is an unbiased estimate of β_1,

What goes wrong without randomization? (Simple Case)

If we do not have randomization, there is no guarantee that *X* is uncorrelated with *ε* (*X* may be endogenous)

Thus the OLS estimate is still

$$\bar{Y}_{1\bullet} - \bar{Y}_{0\bullet} = \beta_1 + \left(\bar{\varepsilon}_{1\bullet} - \bar{\varepsilon}_{0\bullet}\right)$$

If *X* is correlated with *ε*, then $\bar{\varepsilon}_{1\bullet} \neq \bar{\varepsilon}_{0\bullet}$

Hence $\bar{Y}_{1\bullet} - \bar{Y}_{0\bullet}$ does not estimate β_1, but some other quantity that depends on the correlation of *X* and *ε*

If *X* is correlated with *ε*, then standard methods give a biased estimate of β_1

What goes wrong without randomization?

When you regress Y on X, $Y = \beta_0 + \beta_1 X + \varepsilon$ and the OLS estimate of β_1 can be described as

$$b_{OLS} = \frac{\text{Cov}\{Y, X\}}{\text{Cov}\{X, X\}} = \frac{\text{Cov}\{\beta_0 + \beta_1 X + \varepsilon, X\}}{\text{Cov}\{X, X\}}$$

$$= \frac{\beta_1 \text{Cov}\{X, X\} + \text{Cov}\{\varepsilon, X\}}{\text{Cov}\{X, X\}} = \beta_1 + \frac{\text{Cov}\{\varepsilon, X\}}{\text{Cov}\{X, X\}}$$

But since X and ε are correlated, b_{OLS} does not estimate β_1 but some other quantity that depends on the correlation of X and ε

Assumptions of the Linear Regression Model with Strictly Exogenous Regressors

Wish to analyze the effect of all of the explanatory variables on the responses. Thus, define $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and require

- SE1. $\mathrm{E}(y_i \mid \mathbf{X}) = \mathbf{x}_i' \boldsymbol{\beta}$.
- SE2. $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are stochastic variables.
- SE3. $\mathrm{Var}(y_i \mid \mathbf{X}) = \sigma^2$.
- SE4. $\{y_i \mid \mathbf{X}\}$ are independent random variables.
- SE5. $\{y_i\}$ is normally distributed, conditional on $\{\mathbf{X}\}$.

Usual Properties Hold

- Under SE1-SE4, we retain most of the desirable properties of our ordinary least square estimators of $\boldsymbol{\beta}$. These include:
  - the unbiasedness and
  - the Gauss-Markov property of ordinary least square estimators of $\boldsymbol{\beta}$.
- If, in addition, SE5 holds, then the usual *t* and *F* statistics have their customary distributions, regardless as to whether or not **X** is stochastic.
- Define the disturbance term to be $\varepsilon_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$ and
  - write SE1 as $\mathrm{E}(\varepsilon_i \mid \mathbf{X}) = 0$
  - is known as *strict exogeneity* in the econometrics literature.

Orthogonality assumption

- If other OLS assumptions fail, we can still estimate $\beta 1$ consistently, and we can make inference with minor adjustments
- Orthogonality is the most important assumption, where X is not correlated to the error term.
- Endogeneity does not allow to estimate $\beta 1$ consistently
- If there is strong endogeneity bias, our estimated models can be poor descriptions of reality
- Stronger notions of exogeneity require:
  - ε to be independent from x
  - or conditional mean independence $E[\varepsilon | x] = 0$
- Exogeneity guarantees that the factors that are not accounted for (the error term) do not interfere with the estimation of the parameters of interest

Weak exogeneity

- A set of variables are said to be *weakly exogenous* if, when we condition on them, there is no loss of information about the parameters of interest.
- Weak endogeneity is sufficient for efficient estimation.
- Suppose that we have random variables ($\mathbf{x}_1$, y_1), …, ($\mathbf{x}_T$, y_T) with joint probability density (or mass) function for f(y_1, …, y_T, $\mathbf{x}_1$, …, $\mathbf{x}_T$).
- By repeated conditioning, we write this as:

$$\mathrm{f}\left(y_1,\ldots,y_T,\mathbf{x}_1,\ldots,\mathbf{x}_T\right)=\prod_{t=1}^{T}\mathrm{f}\left(y_t,\mathbf{x}_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_{t-1}\right)$$

$$=\prod_{t=1}^{T}\left\{\mathrm{f}\left(y_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_t\right)\mathrm{f}\left(\mathbf{x}_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_{t-1}\right)\right\}$$

Weak exogeneity

- Suppose that this joint distribution is characterized by vectors of parameters **θ** and **ψ** such that

$$\mathrm{f}\left(y_1,\ldots,y_T,\mathbf{x}_1,\ldots,\mathbf{x}_T\right)$$

$$=\left(\prod_{t=1}^{T}\mathrm{f}\left(y_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_t,\boldsymbol{\theta}\right)\right)\left(\prod_{t=1}^{T}\mathrm{f}\left(\mathbf{x}_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\boldsymbol{\psi}\right)\right)$$

- We can ignore the second term for inference about **θ**, treating the **x** variables as essentially fixed.
- If this relationship holds, then we say that the explanatory variables are *weakly exogenous*.

Strong Exogeneity

- Suppose, in addition, that

$$\mathrm{f}\left(\mathbf{x}_t \mid y_1,\ldots,y_{t-1},\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\boldsymbol{\psi}\right)=\mathrm{f}\left(\mathbf{x}_t \mid \mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\boldsymbol{\psi}\right)$$

- that is, conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_{t\text{-}1}$, that the distribution of $\mathbf{x}_t$ does not depend on past values of y, $y_1, \ldots, y_{t\text{-}1}$. Then, we say that $\{y_1, \ldots, y_{t\text{-}1}\}$ does not *Granger-cause* $\mathbf{x}_t$.
- This condition, together with weak exogeneity, suffices for *strong exogeneity*.
- This is helpful for prediction purposes.

Causal effects

- Researchers are interested in causal effects, often more so than measures of association among variables.
- Statistics has contributed to making causal statements primarily through randomization.
  - Data that arise from this random assignment mechanism are known as *experimental*.
  - In contrast, most data from the social sciences are *observational*, where it is not possible to use random mechanisms to randomly allocate observations according to variables of interest.

Structural Models

- A *structural model* is a stochastic model representing a causal relationship, as opposed to a relationship that simply captures statistical associations.
- A sampling based model is derived from our knowledge of the mechanisms used to gather the data.
  - The sampling based model directly generates statistics that can be used to estimate quantities of interest
  - It is also known as an *estimable* model.
- Structural vs. reduced form.

References

- Slides of Ragui Assaad and Caroline Krafft.

- Rubin, D. B. (1974). Estimating causal effects in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688-701.

- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.

Thank you for your attention