ECO 372: Introduction to Econometrics

Spring 2025

Lecture 7: Endogeneity and Emergence of Instrumental Variable Regressions-I

Sakib Bin Amin, Ph.D.

Associate Professor in Economics
Director, Accreditation Project Team (APT)

Outline

Our objectives for this lecture will be to learn

- Revisiting the concept of endogeneity
- Addressing endogeneity
- Issue with OLS when endogeneity prevails
- Instrumental Variable (IV) approach
- Conditions for valid instruments
- General Framework of IV Models

Recall the Basics of Endogeneity

- Zero mean condition of errors and OVB conditions.
- This condition asserts that the conditional distribution of u_i given X_i has a mean of 0
- In simple, it means factors contained in u_i are not related with X_i. But what if $E(u_i|X_i) \neq 0$?
- The the left out variable is related to both X_i and Y_i, when both OVB conditions are met.
- Presence of these both concepts leads to the emergence of endogeneity.
- Apart from the mentioned measurement error, selection bias, simultaneity sometimes lead to endogeneity.
- Endogeneity is the one of the most feared problems in econometric analysis, when data is not experimental in nature.
- The OLS estimates becomes highly biased and inconsistent when there is presence of endogeneity.

Addressing Endogeneity

- So, how to approach this fatal problem?
- One of the best ways to address endogeneity issue (if found) is to use Instrumental Variable (IV) approach.
- The instrumental variables estimator provides a way to nonetheless obtain consistent and unbiased (a trivial amount may persist) parameter estimates.
- This method, widely used in econometrics and rarely used elsewhere, is conceptually difficult and easily misused.

Issue with OLS When Endogeneity Prevails

- Consider the regression model with dependent variable Y_i and single regressor X_i:

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

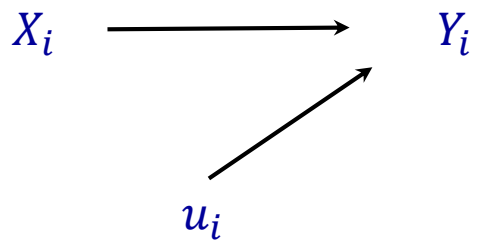
- The goal of this regression analysis is to estimate the conditional mean function and obtain consistent estimate of β_1:

$$E(Y_i|X_i) = \beta_1$$

- Standard regression results make the assumption that the regressors are uncorrelated with the errors in the model [i.e., $E(u_i|X_i) = 0$]
- Then the only effect of X_i on Y_i is a direct effect via the term β_1
- Remember $E(u_i|Y_i)$ should not be zero
- Let's use a path diagram to continue.

Issue with OLS When Endogeneity Prevails

- We have the following path analysis diagram



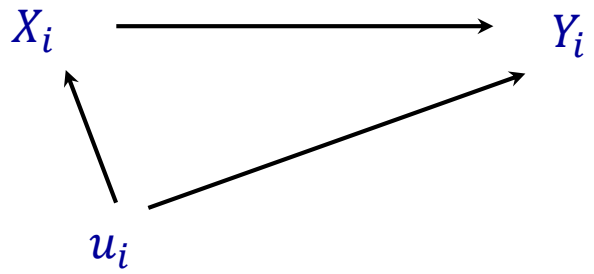
- Where there is no association between X_i and u_i.
- So, X_i and u_i are independent causes of Y_i
- However, in some situations, there may be an association between regressors and errors.
- For example, consider regression of earnings on years of schooling:

$$Wage_i = \alpha_0 + \beta_1 EDU_i + u_i$$

- The error term u_i embodies all factors other than schooling that determine earnings
  - such as say skills.

Issue with OLS When Endogeneity Prevails

- Suppose a person has a high level of u_i, as a result of high skills, which is not considered in the model.
- It increases earnings. This is because $Wage_i = \alpha_0 + \beta_1 EDU_i + u_i$
- But it may also lead to higher levels of education, since schooling is likely to be higher for those with high skills.
- Now a more appropriate path diagram is then the following



- Now there is an association between X_i and u_i

Issue with OLS When Endogeneity Prevails

- What are the consequences of this correlation between X_i and u_i?
- Now higher levels of X_i have two effects on Y_i
- There is both a direct effect via $\beta_1 X_i$ and an indirect effect via u_i affecting X_i, which in turn affects Y_i.
- The goal of regression is to estimate only the first effect, yielding an estimate of β_1.
- The OLS estimate will instead combine these two effects.
- Giving $\hat{\beta}_1 > \beta_1$ in this example where both effects are positive.
  - You can think of other examples where $\hat{\beta}_1 < \beta_1$.
- Using calculus, we can come up with a simple expression of the above discussion:

$$\frac{dY_i}{dX_i} = \beta_1 + \frac{du_i}{dX_i}$$

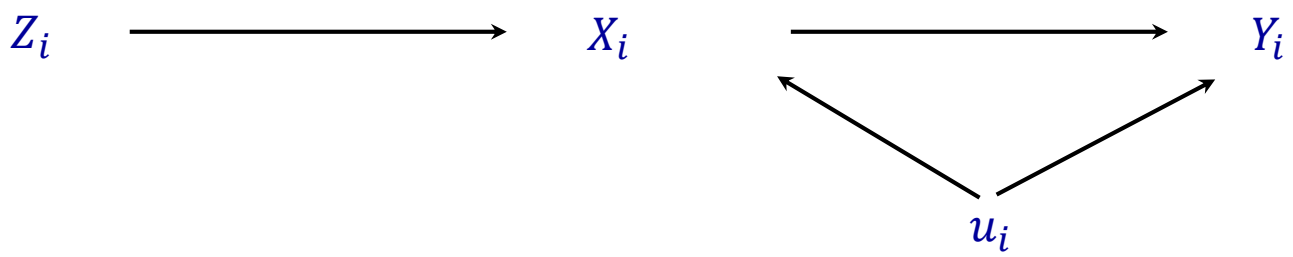
- The OLS estimator is therefore biased and inconsistent.

Instrumental Variable (IV) Approach

- The issues of OLS is due to endogeneity of X_i meaning that changes in X_i are associated not only with changes in Y_i but also changes in the error u_i.
- What is needed here is a method to generate only exogenous variation in X_i.
- So, econometricians have come up with an approach known IV approach.
- In this approach, there exists an instrument Z_i.
- This Z_i has the property that changes in Z_i is associated with changes in X_i but do not directly lead to change in Y_i (i.e., aside from the indirect route via X_i).

Instrumental Variable (IV) Approach

- Now, Let's use the trusty path diagram again



- This introduces a variable Z_i that is causally associated with X_i but not u_i
- It is still the case that Z_i and Y_i will be correlated, but the only source of such correlation is the indirect path of Z_i being correlated with X_i, which in turn determines Y_i.
- The more direct path of Z_i being a regressor in the model for Y_i is ruled out.
- More formally, a variable Z_i is called an instrument or instrumental variable or IV for the regressor X_i.
- Nevertheless, there are conditions for valid Z_i

Conditions for Valid Instruments

- A valid instrument needs to satisfy two conditions:

    1. Instrument relevance: $corr(Z_i|X_i) \neq 0$ [the magnitude must be is very high]

    2. Instrument exogeneity: $corr(Z_i|u_i) = 0$ [if not zero then it must be very close to zero]

- The first assumption requires that there is some association between the instrument and the variable being instrumented.

- The second assumption requires that instrument is exogenous. Then that part of the variation of X_i captured by the instrumental variable is exogenous.

General Framework of IV Models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \psi_1 R_{1i} + \cdots + \psi_j R_{ji} + u_i$$

i=runs over all observations, i={1, 2, 3,…,N}

Y_i=dependent variable

X_i= independent variable

R_i= control variables

β_0= intercept of the regression line

$\beta_1, \ldots, \beta_k$= slopes/predictors [any or all can be endogenous]

$\psi_1, \ldots, \psi_j$= slopes of exogenous controls

u_i= error term

General Framework of IV Models

- If the number of instruments equals (e.g., skills or any other variables) the number of endogenous regressors (e.g., education): exactly identified model.
- If the number of instruments greater than the number of endogenous regressors: overidentified model.
- If the number of instruments less than the number of endogenous regressors: underidentified model.
- Having (at least) one instrument for any single endogenous regressor is essential. Otherwise, computation is not possible.