Lecture 6

Solutions to Endogeneity: Instrumental Variables Estimation

Nurgul Tilenbaeva

American University - Central Asia

28.11.2022; 01.12.2022

Contents

Endogeneity Revisited

Assumption MLR.4. Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$E(u|x_1, x_2, ..., x_k) = 0.$

All factors in the unobserved error term must be uncorrelated with the explanatory variables.

If x_j is uncorrelated with u (i.e. when MLR.4. holds): **exogenous explanatory variables**
If x_j is correlated with u: **endogenous explanatory variables**

Question: Why is Assumption MLR.4 so important?

Endogeneity Revisited

Endogeneity

- Failure of Assumption MLR.4;
- Situation in which one or more of the explanatory variables are correlated with the error term.

Endogeneity Revisited

- Specification Error
- Measurement Error
- Omitted Variables
- Reversed Causality
- Sample Selection

Motivation: Omitted Variables in a Simple Regression Model

Options discussed so far when we have the **omitted variable bias**:

1. Ignore the problem and suffer the consequences of biased estimators;
   - *Example:* We concluded that there is a downward bias in the effect of job training on subsequent wages. At the same time, we have found a statistically significant positive estimate. We have still learned something: job training has a positive effect on wages, and it is likely that we have underestimated the effect.
   - Unfortunately, the opposite case, where our estimates may be too large in magnitude, often occurs, which makes it very difficult to draw any useful conclusions.
2. Try to find and use a suitable *proxy variable* for the unobserved variable.
   - It is not always possible to find a good proxy variable.

Another solution:

1. Use an Instrumental Variables Estimation Method.

Motivation:
Omitted Variables in a Simple Regression Model

Example:

$log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e$

satisfies Assumptions MLR.1 through MLR.4.

However, the data set does not contain data on ability, so we estimate β_1 from the simple regression

$log(wage) = \beta_0 + \beta_1 educ + u$

where u contains *abil*. If this equation is estimated by OLS, we have a biased β_1 if *educ* and *abil* are correlated.

However, we can still use the last equation as the basis for estimation, provided we can find an **instrumental variable** for *educ*.

Motivation: Omitted Variables in a Simple Regression Model

Instrumental Variables Estimation

$y = \beta_0 + \beta_1 x + u,$

where we think that x and u are correlated:

$Cov(x, u) \neq 0$

Suppose that we have an observable variable z that satisfies these two assumptions:

1. z is uncorrelated with u **(instrument exogeneity)**:
   $Cov(z, u) = 0$;
   - z does not have a direct effect on y;
   - z has an effect on y only through x.
2. z is correlated with x **(instrument relevance)**:
   $Cov(z, x) \neq 0$

Then, we call z an **instrumental variable (IV)** for x, or sometimes simply an **instrument** for x.

Motivation: Omitted Variables in a Simple Regression Model

Instrumental Variables Estimation

1. **Instrument exogeneity:**
   Not possible to test. We can appeal to economic behavior or introspection.
2. **Instrument relevance:**
   Can be tested:
   $x = \pi_0 + \pi_1 z + v$
   1. $\hat{\pi}_1$ has an expected sign;
   2. $\hat{\pi}_1$ is statistically significant:
      We should be able to *reject* the null hypothesis

      $H_0 : \pi_1 = 0$ against the two-sided alternative
      $H_1 : \pi_1 \neq 0$

   If this is the case, then we can be fairly confident that instrument relevance holds.

Motivation: Omitted Variables in a Simple Regression Model

Instrumental Variables Estimation

Example: $log(wage) = \beta_0 + \beta_1 educ + u$

1. **Instrument exogeneity:**
   - An instrumental variable z for *educ* must be uncorrelated with ability (and any other unobserved factors affecting wage);
   - An instrumental variable z for *educ* must not directly affect *wage*, affect only through *educ*.
2. **Instrument relevance:**

   An instrumental variable z for *educ* must be correlated with education, with an expected sign.

Do the following potential IVs satisfy these conditions?

- Last digit of an individual's Social Security Number;
- IQ;
- Mother's education;
- Number of siblings while growing up.

Motivation: Omitted Variables in a Simple Regression Model

Instrumental Variables Estimation

Example:

$score = \beta_0 + \beta_1 skipped + u$,

where

score=exam score;
skipped=total number of lectures skipped during the semester.

We are worried that *skipped* is correlated with other factors in u: more able, highly motivated students might miss fewer classes. Thus, we may get a biased estimate of the causal effect of missing classes.

Do the following potential instrumental variables (IVs) satisfy the conditions of instrument exogeneity and instrument relevance?

- Distance between living place and campus.

Motivation:
Omitted Variables in a Simple Regression Model

Example: Estimating the return to education for married women

$log(wage) = \beta_0 + \beta_1 educ + u$

1. We first obtain the OLS estimates:

   $$\widehat{log(wage)} = \underset{(0.185)}{-0.185} + \underset{(0.014)}{0.109} educ$$

   $n = 428, R^2 = 0.118$

   The estimate for β_1 implies an almost 11% return for another year of education.

Motivation:
Omitted Variables in a Simple Regression Model

Example: Estimating the return to education for married women

$log(wage) = \beta_0 + \beta_1 educ + u$

2. Next, we use father's education (*fatheduc*) as an instrumental variable for *educ*:
   - First, we have to maintain that *fatheduc* is uncorrelated with *u* **(instrument exogeneity)**. Next, we check that *educ* and *fatheduc* are correlated **(instrument relevance)**:

     $$\hat{educ} = \underset{(0.28)}{10.24} + \underset{(0.029)}{0.269} fatheduc$$
     $$n = 428, R^2 = 0.173$$

     The *t* statistic on *fatheduc* is 9.28, which indicates that *educ* and *fatheduc* have a statistically significant positive correlation.

Motivation:
Omitted Variables in a Simple Regression Model

Example: Estimating the return to education for married women

$log(wage) = \beta_0 + \beta_1 educ + u$

2. Next, we use father's education (*fatheduc*) as an instrumental variable for *educ*:
   - Using *fatheduc* as an IV for *educ* gives

     $$log(\hat{wage}) = \underset{(0.446)}{0.441} + \underset{(0.035)}{0.059} educ$$
     $$n = 428, R^2 = 0.093$$

     The IV estimate of the return to education is 5.9%, which is about one-half of the OLS estimate. This suggests that the OLS estimate is too high and is consistent with omitted ability bias.

Motivation: Omitted Variables in a Simple Regression Model

Example: Estimating the effect of smoking on birth weight

$log(bwght) = \beta_0 + \beta_1 packs + u$

We might worry that *packs* is correlated with other health factors or the availability of good prenatal care, so that *packs* and u might be correlated. A possible IV for *packs* is the average price of cigarettes in the state of residence, *cigprice*. We will assume that *cigprice* and u are uncorrelated.

Motivation:
Omitted Variables in a Simple Regression Model

Example: Estimating the effect of smoking on birth weight

$log(bwght) = \beta_0 + \beta_1 packs + u$

- We first check **instrument relevance**:
  $\hat{packs} = \underset{(0.103)}{0.067} + \underset{(0.0008)}{0.0003} cigprice$
  $n = 1,388, R^2 = 0.0000$

  This indicates no relationship between smoking during pregnancy and cigarette prices. Because *packs* and *cigprice* are not correlated, we should not use *cigprice* as an IV for *packs*. But what happens if we do?

- The IV results would be
  $log(\hat{bwght}) = \underset{(0.91)}{4.45} + \underset{(8.70)}{2.99} packs$
  $n = 1,388$

Motivation:
Omitted Variables in a Simple Regression Model

Computing R-Squared after IV Estimation

Most regression packages compute an R-squared after IV estimation, using the standard formula:

$R^2 = 1 - \frac{SSR}{SST}$, where

SSR is the sum of squared IV residuals
SST is the total sum of squares of y.

Unlike in the case of OLS, the R-squared from IV estimation can be negative because SSR for IV can actually be larger than SST.

Although it does not really hurt to report the R-squared for IV estimation,it is not very useful, either.

In addition, these R-squareds cannot be used in the usual way to compute F tests of joint restrictions.

IV Estimation of the Multiple Regression Model

$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$

This is called a **structural equation**.

z_1 is *exogenous*;

y_2 is *endogenous*.

If we estimate this model by OLS, all of the estimators will be biased.

Thus, we look for an instrumental variable z_2 for y_2.

1. z_2 is uncorrelated with u_1 **(instrument exogeneity)**:
   $Cov(z_2, u_1) = 0$;
   - z_2 does not have a direct effect on y_1;
   - z_2 has an effect on y_1 only through y_2.
2. z_2 is correlated with y_2 **(instrument relevance)**:
   $Cov(z_2, y_2) \neq 0$

   $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$

   This is called a **reduced form equation**.

   The key identification condition is that

   $\pi_2 \neq 0$

IV Estimation of the Multiple Regression Model

Example: Using College Proximity as an IV for Education

$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 black + \beta_4 south + ... + u$

exper, *black*, *south*, etc. are *exogenous*;
educ is *endogenous*.

Card (1995) used a dummy variable for whether someone grew up near a four-year college (*nearc*4) as an instrumental variable for *educ*.

We estimate the **reduced form equation**:

$$\hat{educ} = \underset{(0.24)}{16.64} + \underset{(0.088)}{0.320} nearc4 - \underset{(0.034)}{0.413} exper + ...$$

$n = 3,010, R^2 = 0.477.$

We are interested in the coefficient and t statistic on *nearc*4.

IV Estimation of the Multiple Regression Model

Example: Using College Proximity as an IV for Education

TABLE 15.1 Dependent Variable: log(*wage*)

Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations	3,010	3,010
R-squared	.300	.238
Other controls: *smsa66*, *reg662*, ..., *reg669*		