Lecture 9: Relaxing the Assumptions of the Classical Model (II)

Thomas Chadefaux

Contents

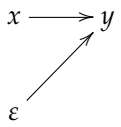
1 What is Endogeneity?

1.1 Examples

- **Example 1**. Suppose that we are interested in the effect of education on income. Our hypothesis is that additional years of education should increase people's income. I.e., we estimate

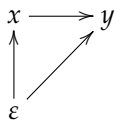
$$\text{income}_i = \beta_0 + \beta_1 \text{education}_i + \varepsilon_i$$

and probably find β_1 to be significantly greater than 0. Can we conclude that education increases revenue? Not from this model. To understand why, think of education as x and income as y, and represent our model as follows



In other words, y is a function of x and ε, where ε is a disturbance factor due (in theory) to the stochasticity of the process. Yet this model implies that the only effect of x on y is a direct effect via $\beta_1 x$. However, both education and income are probably a function of the individual's inherent ability, which in this model ends up in the error term.

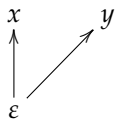
Suppose a person has a high ability (i.e., a high ε). This is expected to increase her income. But it is also expected to increase her education, since education is probably higher for those with a high level of ability. In other words, the true diagram should rather be:



So in this case we have omitted a relevant variable, ability, which causes both x and y.

- **Example 2**. You are in the forest and observe ducks flying. Suddenly you hear a loud bang, and the duck falls to the ground. Another duck, another bang, again this one falls to the ground. You observe this over and over, and most of the time a loud bang precedes the death of the duck (sometimes, but rarely, the duck escapes). In fact, you find that 90% of the time, a bang is followed by the death of the duck. You therefore write a letter to the Prime Minister, requesting that all ducks be equipped with noise-cancelling earphones. The problem here, of course, is again one of omitted variable. The sound is not causing the death, but rather the gun is. But because the gun is causing the sound as well, you again have a situation in which x (here the bang) is correlated with the error term (the gun, which you do not observe). In this case, the effect of x on y

should be 0.



2 Sources of Endogeneity

- Omitted variable
- Simultaneity and reversed causality. Consider the following system of equations:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \quad (1)$$
$$x = \gamma_0 + \gamma_1 y + \gamma_2 z + u \quad (2)$$

but note that x can be rewritten as

$$x = \gamma_0 + \gamma_1(\beta_0 + \beta_1 x + \beta_2 z + \varepsilon) + \gamma_2 z + u,$$

which makes it clear that x is a function of ε, and hence that it is correlated with the error term in (1)

- Measurement error. Suppose that the true regression model is:

$$y = \beta_0 + \beta_1 x* + \varepsilon*,$$

which can be estimated by OLS. But instead, suppose that the variable x is estimated with error :

$$x = x* + v,$$

where v is uncorrelated with $\varepsilon*$ and $x*$. What we estimate, then, is:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad = \beta_0 + \beta_1(x* + v) + \varepsilon* - \beta_1 v, \quad (3)$$

such that x is correlated with the error term.[1]

[1] Note that as the measurement error increases, b_x will tend toward 0.

3 Consequences of Endogeneity

If we estimate the model using regular OLS, then we can make NO INFERENCE about the relationship between x and y. The relationship may be positive, negative, 0, etc. There is no way to tell. OLS makes the assumption that the regressors are uncorrelated with the errors. I.e.,

$$E[X'\varepsilon] = 0$$

. This assumption is crucial. It means that the only effect of x on y is a direct effect via βx. But there is in fact also an indirect effect, from ε to y via x. The OLS estimate will instead combine these two effects, such that $b > \beta$ in the example on

education, for example. This can also be seen using calculus. Consider the situation in which $y = \beta x + u(x)$. Then:

$$\frac{\partial y}{\partial x} = \beta + \frac{\partial u}{\partial x} \neq \beta$$

More formally:

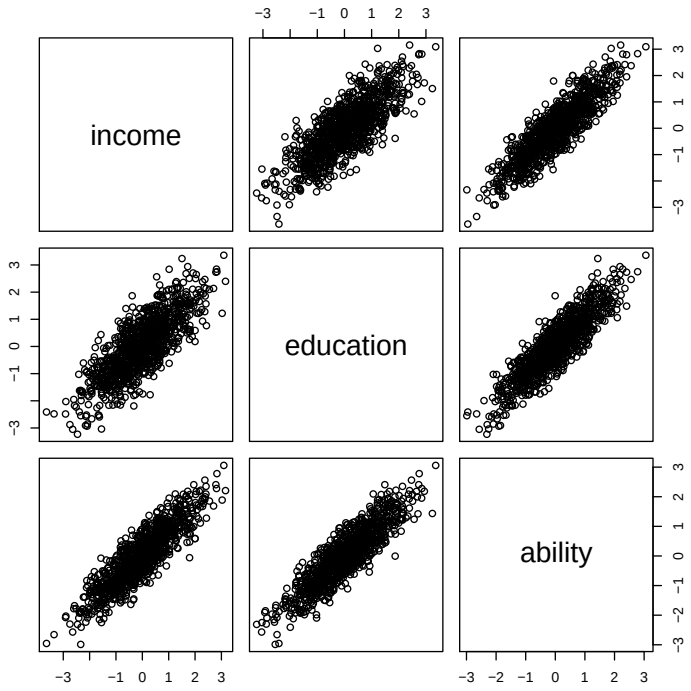
$$\begin{aligned} \beta_{OLS} &= (X'X)^{-1}X'y && (4)\\ &= (X'X)^{-1}X'(X\beta + \epsilon) && (5)\\ &= \beta + X'\epsilon && (6)\\ & && (7) \end{aligned}$$

But since it is no longer the case that $X'\epsilon = 0$, β_{OLS} is biased.

3.1 An Example with R

First, set up x, ε and y such that there is no relationship between x and y, but there is a relationship between x_1 and x_2

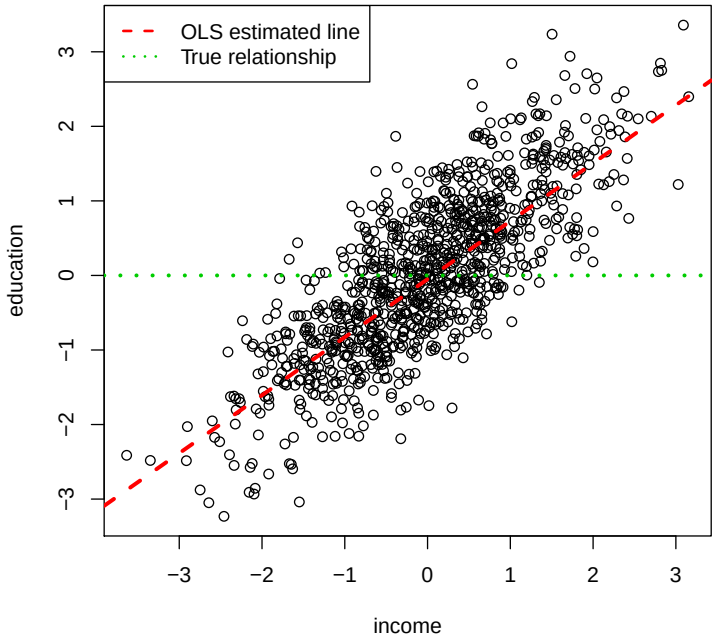
```
> n <- 1000
> ability <- rnorm(1000)
> education <- ability + rnorm(1000, sd=0.5)
> income <-  ability + rnorm(1000, sd=0.5)
> pairs(~income + education + ability)
```



Estimate the correct model and the wrong model:

```
> lm.correct <- lm(income ~ education + ability)
> summary(lm.correct)

Call:
lm(formula = income ~ education + ability)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46927 -0.33820  0.01073  0.33945  1.87989

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03920    0.01579  -2.482   0.0132 *
education    0.02374    0.03168   0.749   0.4539
ability      0.95569    0.03574  26.743   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4991 on 997 degrees of freedom
Multiple R-squared:  0.7842,        Adjusted R-squared:  0.7837
F-statistic:  1811 on 2 and 997 DF,  p-value: < 2.2e-16

> lm.incorrect <- lm(income ~ education)
> summary(lm.incorrect)

Call:
lm(formula = income ~ education)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86067 -0.46407  0.02453  0.45577  2.12947

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05052    0.02067  -2.444   0.0147 *
education    0.77803    0.01890  41.162   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6537 on 998 degrees of freedom
Multiple R-squared:  0.6293,        Adjusted R-squared:  0.6289
F-statistic:  1694 on 1 and 998 DF,  p-value: < 2.2e-16

> plot(income, education)
> abline(lm.incorrect, col=2, lwd=3, lty=2)
```

```
> abline(0,0, col=3, lwd=3, lty=3)
> legend('topleft', legend = c('OLS estimated line', 'True relationship' ), lty=c(2,3), col=c(2,3), lwd=c(2,2))
```

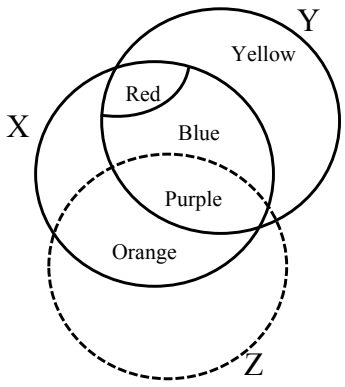


4 Dealing with Endogeneity

The inconsistency of OLS is due to the endogeneity of x. Because changes in x are associated not only with changes in y, but also with changes in ε, we cannot interpret b alone. There are two widely ways to deal with endogeneity: instrumental variables, and fixed effects.

4.1 Instrumental Variables

- The clearest explanation of IV comes from Kennedy (2008), pp. 147–8. "Suppose that Y is determined by X and an error term ε (ignore the dashed circle Z for the moment), but that X and ε are not independent. The lack of independent between X and ε means that the yellow area (representing the influence of the error term) must now overlap with the X circle. This is represented by the red area. The action from the error is represented by the red plus yellow areas. Variation in Y in the red area is due to the influence of *both* the error term and the explanatory variable X. If Y were regressed on X, the information in the red-plus-blue-plus-purple area would be used to estimate β_x This estimate is biased because the red area does not reflect variation in Y arising solely from variation in X. Some way must be found to get rid of the red area.



The circle Z represents an IV for X. It is drawn to reflect the two properties it must possess:

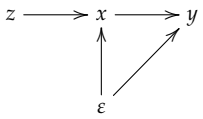
- It must be independent of the error term, so it is draw such that it does not intersect the yellow or red areas.
- It must be as highly correlated as possible with X

Suppose X is regressed on Z. The predicted X from this regression, $\hat{X}$, is represented b the purple-plus-orange area. Now regress Y on $\hat{X}$ to produce an estimate of β_x. This in fact defines the IV estimator. The overlap of the Y circle with the purple-plus-orange area is the purple area, so information in the purple area is used to form this estimate; since the purple area corresponds to variation in Y arising entirely from variation in X, the resulting estimate of β_X is unbiased.

Notice that, in constructing this estimate, although the bias arising from the red area is eliminated, the information set used to estimate β_X has shrunk from the red-plus-blue-plus-purple area to just the purple area. This implies that the variance of the IV estimator will be considerable higher than the variance of the OLS estimator, a reason why many researchers prefer to stick with OLS despite its asymptotic bias. It should now be apparent why the IV should be as highly correlated with X as possible: this makes the purple area as large as possible, reducing the variance of the IV estimator.
[2]

- **Examples of Instruments**
  - Returning to the example of education: Angrist and Krueger (1991, Quarterly Journal of Economics) suggest birth month as instrument for education. In U.S. school system, students are categorized into school year system. As a consequence of this education system, there are first and last school categorization months. It is reported that students who were born in earlier months have higher school grades and SAT scores compared to students who were born in later. As a consequent, students who are born earlier are more likely to go to colleges. So, birth month is correlated with the length of education (Angrist and Krueger (1991))
  - (from Wooldrige 2010, p. 87) A variable such as the last digit of one?s social security number makes a poor IV candidate for the opposite reason. Because the last digit is randomly determined, it is independent of other factors that affect earnings. But it is also independent of education.
- **Derivation of the IV estimator** Return to the earnings-schooling example. Suppose a one unit change in the instrument z is associated with 0.2 more years of schooling and with a \$500 increase in annual earnings. This increase in earnings is a consequence of the indirect effect that increase in z led to increase in schooling which in turn increases income. Then it follows that 0.2 years additional schooling are associated with a \$500 increase in earnings, so that a one year

[2] Another way to picture IV is as follows:



increase in schooling is associated with a \$500/0.2 = \$2500 increase in earnings. The causal estimate of

is therefore 2500. In mathematical notation we have estimated the changes dx/dz and dy/dz and calculated the causal estimator as

$$\beta_{IV} = \frac{dy/dz}{dx/dz} = \frac{500}{0.2}.$$

All we have to do, then, is to estimate dx/dz and dy/dz, which we can do by OLS. We estimate dx/dz using $(Z'Z)^{-1}Z'X$ and dy/dz using $(Z'Z)^{-1}Z'y$. Then

$$\beta_{IV} = \frac{(Z'Z)^{-1}Z'y}{(Z'Z)^{-1}Z'X} = (Z'X)^{-1}Z'y$$

Another way to write the same thing is

$$\beta_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y,$$

where $\hat{X}$ is the fitted values obtained from regression X on Z (i.e., the orange-plus-purple area on the ballentine plot)

- **Testing Instruments' Relevance.** The most basic thing to do is to test the correlation between X and Z. For more complicated setups, with several instruments, run the first stage regression, in which you regress X on your instrument(s) Z. At this point, you should run an F-test that at least one of the coefficients on the instruments is not 0. As a rule of thumb, you should have $F > 10$, or corresponding p-value ≤ 0.0016. You should report this F-test when reporting IV estimates.

4.2 Two-staged least squares (2SLS)

Suppose that there are more instrument variables than regressors. Then why not use them all? In fact, IV estimation is a special case of 2SLS. Suppose our model is

$$y = X\beta + Z_1\delta + \varepsilon$$

, where Z denotes exogenous variables, whereas X are endogenous variables. Suppose moreover that we have two possible instruments for X: Z_2 and Z_3. The procedure works in 2 stages:

- Stage 1: Regress each endogenous variable on all the exogenous variables, and obtain estimated values of these endogenous variables. In our example,

$$\hat{X} = Z_1\hat{\delta}_1 + Z_2\hat{\delta}_2 + Z_3\hat{\delta}_3 + \varepsilon$$

- Stage 2: Use these estimated values and the exogenous variables as regressors in an OLS regression.

Note that this is equivalent to IV estimation, except that with IV we'd only use Z_2.

How to do it in R?

```
> library(foreign)
> hsng2 <- read.dta("http://www.stata-press.com/data/r11/hsng2.dta")
> #regress median monthly rents (rent) of census divisions on the share
> #of urban population (pcturban) and the median housing value (hsngval)
> #Housing values are likely endogeneous and therefore instrumented by median
> #family income (faminc):
>
> #--- OLS estimate:
> lm.ols <- lm(rent ~ hsngval + pcturban, data=hsng2)
> summary(lm.ols)

Call:
lm(formula = rent ~ hsngval + pcturban, data = hsng2)

Residuals:
    Min      1Q  Median      3Q     Max
-41.856 -11.313  -2.221   7.420  93.985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.259e+02  1.419e+01   8.876 1.30e-11 ***
hsngval     1.520e-03  2.276e-04   6.681 2.49e-08 ***
pcturban    5.248e-01  2.491e-01   2.107   0.0405 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 20.76 on 47 degrees of freedom
Multiple R-squared:  0.6692,        Adjusted R-squared:  0.6551
F-statistic: 47.54 on 2 and 47 DF,  p-value: 5.128e-12

> #--- IV estimation by 2SLS (manual version)
> #- first stage:
> iv1 <- lm(hsngval ~ pcturban + faminc, data=hsng2)
> x.hat <- fitted(iv1)
> #- second stage:
> iv2 <- lm(rent ~ pcturban + x.hat, data=hsng2)
> summary(iv2)

Call:
lm(formula = rent ~ pcturban + x.hat, data = hsng2)

Residuals:
    Min      1Q  Median      3Q     Max
-28.188 -12.596  -3.184   5.600  49.614
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.814331  12.999622   8.755 1.95e-11 ***
pcturban     -0.506412   0.304971  -1.661    0.103
x.hat         0.003194   0.000393   8.127 1.65e-10 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 18.69 on 47 degrees of freedom
Multiple R-squared:  0.7318,        Adjusted R-squared:  0.7204
F-statistic: 64.13 on 2 and 47 DF,  p-value: 3.693e-14
```

```
> #--- IV estimation, canned version, including heteroskedasticity corrected SEs
> library(AER)
> iv.canned <- ivreg(rent~hsngval+pcturban|pcturban+faminc,
+       data = hsng2)
> library(sandwich)
> library(lmtest)
> coeftest(iv.canned, vcov=sandwich)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  1.1381e+02  2.1622e+01  5.2639 3.428e-06 ***
hsngval      3.1938e-03  7.3798e-04  4.3278 7.818e-05 ***
pcturban    -5.0641e-01  5.4283e-01 -0.9329    0.3556
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

If you want the canned version, use

```
1 ivreg(y ~ x1 + x2 + w1 + w2 | z1 + z2 + z3 + w1 + w2)
```

Listing 1:

where x_1 and x_2 are endogenous regressors, w_1 and w_2 exogeneous regressors, and z_1 to z_3 are excluded instruments.

4.3 Fixed Effects

Covered in the lecture on panel data