ECO375 Tutorial 8
Instrumental Variables

Matt Tudball

University of Toronto Mississauga

November 16, 2017

Review: Endogeneity

- Instrumental variables are used to deal with **endogeneity** in multiple regression models.
- Recall the assumption MLR.4

MLR.4 The error u_i has an expected value of zero conditional on all x_i $\mathbb{E}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = 0$ for $i = 1, ..., n$.

- If x_j is correlated with the error term u, perhaps because it is correlated with some omitted variable x_{k+1}, then MLR.4 will be violated and our OLS estimators will be biased and inconsistent.

Review: IV Estimator

- In the last lecture you considered the simple model

$$y_i = \beta_0 + \beta_1 x_{i1} + u_i$$

where $\text{Cov}(x_i, u_i) \neq 0$ and hence x_{i1} is endogenous.

- If we have a valid instrument z_i for x_{i1} then it must satisfy the properties

$$\begin{array}{ll} (1) & \text{Cov}(z_i, u_i) = 0 \quad \text{exogeneity condition} \\ (2) & \text{Cov}(z_i, x_{i1}) \neq 0 \quad \text{relevance condition} \end{array}$$

- The condition (1) is difficult to test and must generally by justified via economic theory.
- The condition (2) can be tested with a t-test from the regression

$$x_{i1} = \pi_0 + \pi_1 z_i + v_i$$

Review: IV Estimator

- Using the moment restrictions

$$\mathbb{E}[u_i] = \mathbb{E}[y_i - \beta_0 - \beta_1 x_{i1}] = 0$$
$$\mathbb{E}[z_i u_i] = \mathbb{E}[z_i (y_i - \beta_0 - \beta_1 x_{i1})] = 0$$

  we can show that β_1 takes the form

$$\beta_1 = \frac{\text{Cov}(z_i, y_i)}{\text{Cov}(z_i, x_{i1})}$$

  and β_0 takes the form

$$\beta_0 = \mathbb{E}(y_i) - \beta_1 \mathbb{E}(x_{i1})$$

- Replacing the covariances with their sample analogues we can obtain the **IV estimator** for β_1

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_{i1} - \bar{x_1})}$$

Review: IV Inference

- With the IV estimator for β_1 in hand, we now need to produce an estimator for the variance of $\hat{\beta}_1$ in order to conduct valid inference.
- Let's assume MLR.5 (homoscedasticity) for z_i such that

$$\text{Var}(u_i^2|z_i) = \sigma^2 = \text{Var}(u_i)$$

- Then we can show that the variance of $\hat{\beta}_1^{IV}$ takes the form

$$\text{Var}(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{n\sigma_{x_1}^2 \rho_{x_1,z}^2}$$

- Estimating σ^2 and $\sigma_{x_1}^2$ by their sample analogues and $\rho_{x_1,z}^2$ by $R_{x_1,z}^2$ (which is the R-squared from the regression of x_{i1} on z_i), we obtain the estimator for the variance of $\hat{\beta}_1^{IV}$

$$\widehat{\text{Var}}(\hat{\beta}_1^{IV}) = \frac{\hat{\sigma}^2}{SST_{x_1} R_{x_1,z}^2}$$

Weak IV

- Let's recall the sample variance of the OLS estimator:

$$\widehat{\text{Var}}(\hat{\beta}_1^{OLS}) = \frac{\hat{\sigma}^2}{SST_{x_1}}$$

- Note that in the denominator of the estimator for the variance we had the term $0 < R^2_{x_1,z} < 1$ which is the R-squared from a regression of x_{i1} on z_i.
- This suggests that the variance of the IV estimator will always be higher than the variance of the OLS estimator.
- It also indicates that the weaker the relationship between x_{i1} and z_i, the higher the variance of the IV estimator $\hat{\beta}_1^{IV}$.
- Weak IV will also exacerbate the finite sample bias of the IV estimator. As we will see below, weak IV's (even with valid instruments) may actually produce estimates that are *more* biased than OLS.

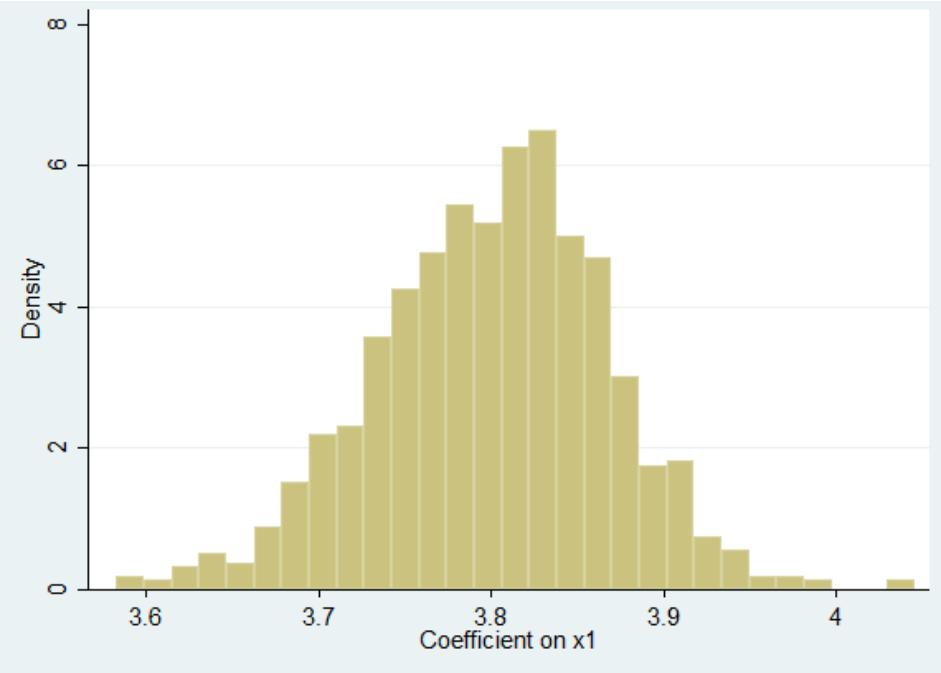
Monte Carlo Results: OVB Correction

- Let's generate some Monte Carlo results to visually compare the IV and OLS estimators in the case where there is a relevant omitted variable.
- First let's visualise the distribution of the OLS estimates in this case.
- In our simulation we are going to generate 1000 replications of samples $n = 50$ produced from the following data-generating process:

$$
\begin{aligned}
x_{i2} &= \mathcal{N}(1, 2) \\
x_{i1} &= 2x_{i2} + \mathcal{N}(2, 2) \\
y_i &= 3x_{i1} + 2x_{i2} + \mathcal{N}(0, 1)
\end{aligned}
$$

- In each replication we are going to calculate the OLS estimator from the regression of y_i on x_{i1} (i.e. omitting x_{i2} which is correlated with both x_{i1} and y_i).
- You can therefore think of $u_i = 2x_{i2} + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$.

Monte Carlo Results: OVB Correction



- Notice that the estimates are centred around $\hat{\beta}_1^{OLS} \approx 3.8$ while the true value of β_1 is 3.

Monte Carlo Results: OVB Correction

- Let's now try running a simulation with the same data-generating process as before but introducing a valid IV z_i for x_{i1}.

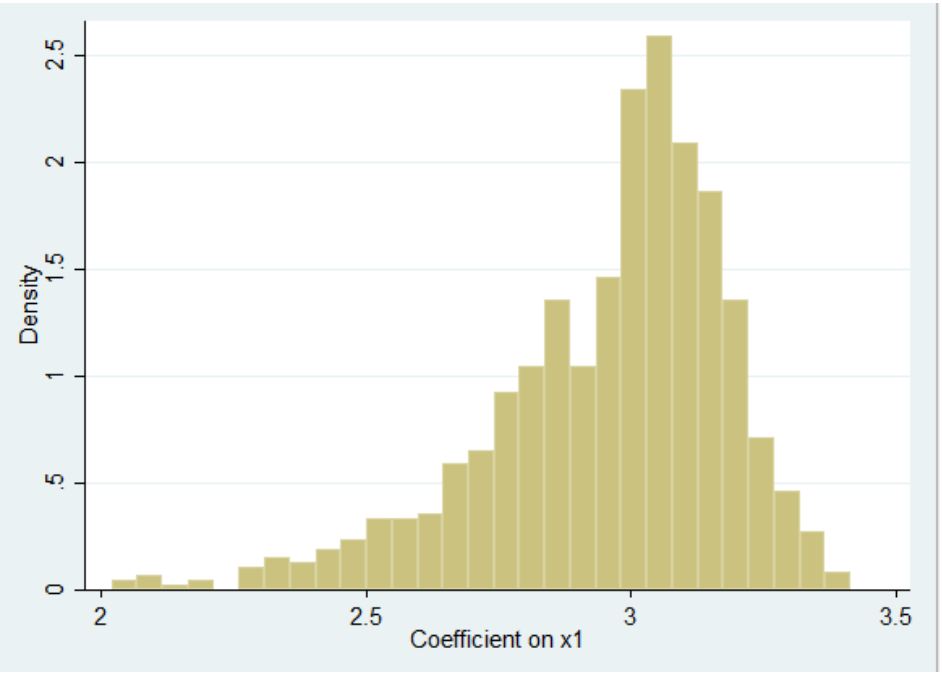
$$
\begin{aligned}
x_{i2} &\sim \mathcal{N}(1, 2) \\
z_i &\sim \mathcal{N}(3, 1) \\
x_{i1} &= 3z_i + 2x_{i2} + \mathcal{N}(2, 2) \\
y_i &= 3x_{i1} + 2x_{i2} + \mathcal{N}(0, 1)
\end{aligned}
$$

- Notice that z_i is uncorrelated with x_{i2}. Since $u_i = 2x_{i2} + \epsilon_i$ we can show that

$$\text{Cov}(z_i, u_i) = \text{Cov}(z_i, 2x_{i2} + \epsilon_i) = 2\text{Cov}(z_i, x_{i2}) + \text{Cov}(z_i, \epsilon_i) = 0$$

  and therefore z_i satisfies condition (1) of IV validity in slide 3.
- Since z_i enters into the equation which determines x_{i1} we also know that it is relevant. It therefore also satisfies condition (2) of IV validity.
- Therefore z_i is a valid instrument and it should return a consistent estimate of $\beta_1 = 3$.

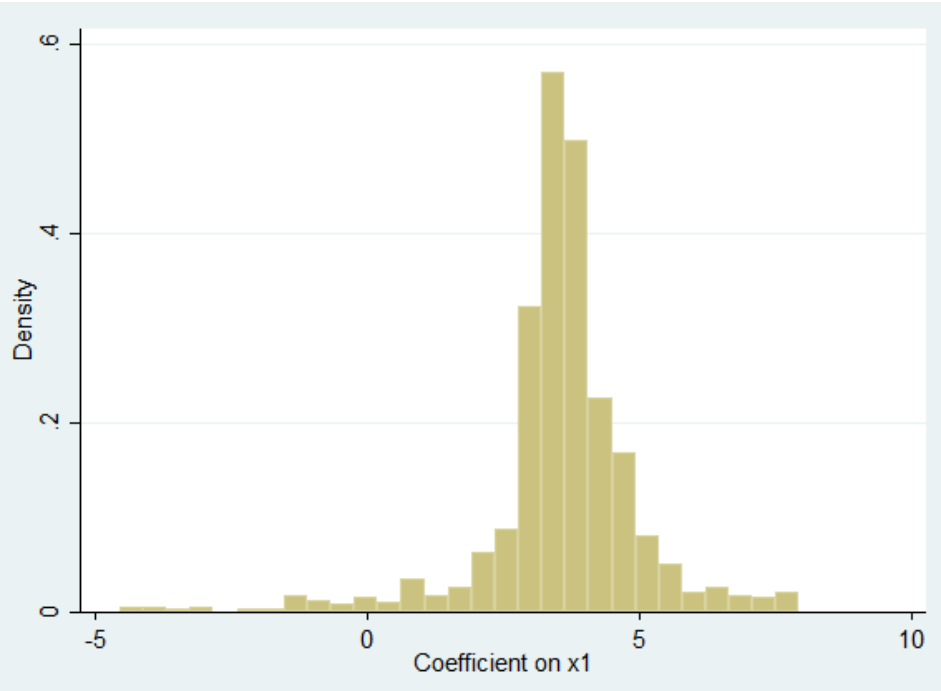
Monte Carlo Results: OVB Correction



- Notice that, although the estimates are clustered much more around the true value of $\beta_1 = 3$, the distribution is skewed very much to the left.
- It's also true that the average $\hat{\beta}_1^{IV} \approx 2.9$ indicating that the IV estimator is biased.

Monte Carlo Results: Weak IV

- In the last Monte Carlo simulation x_{i1} and z_i were related according to the coefficient $\pi_1 = 3$.
- Let's see what happens to the mean and variance of our estimates when we reduce that to $\pi_1 = 0.3$.



- Here the average $\hat{\beta}_1^{IV} \approx 4.1$ which is actually *worse* than the estimate produced by OLS.

IV With Additional Exogenous Variables

- Let's extend this model somewhat. Suppose there is an additional explanatory variable x_{i2} which satisfies MLR.4 (i.e. $Cov(x_{i2}, u_i) = 0$) such that our regression takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + u_i$$

- How does this change our estimation procedure?
- The relevance condition can be restated as testing that $\pi_1 \neq 0$ in a multiple regression

$$x_{i1} = \pi_0 + \pi_1 z_i + \pi_2 x_{i2} + v_i$$

- The moment restrictions needed for identification are now

$$\mathbb{E}[u_i] = 0 \quad \text{as before}$$
$$\mathbb{E}[z_{i1} u_i] = 0 \quad \text{as before}$$
$$\mathbb{E}[x_{i2} u_i] = 0 \quad \text{which is the standard MLR.4 condition}$$

IV With Additional Exogenous Variables

- We can replace these moment restrictions with their sample analogues in order to obtain the FOCs that we need to solve

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2 x_{i2} \right) \\
0 &= \sum_{i=1}^{n} z_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2 x_{i2} \right) \\
0 &= \sum_{i=1}^{n} x_{i2} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1^{IV} x_{i1} - \hat{\beta}_2 x_{i2} \right)
\end{aligned}
$$

- We can solve this system of equations to obtain estimates for $\hat{\beta}_0$, $\hat{\beta}_1^{IV}$ and $\hat{\beta}_2$.

Two-Stage Least Squares (2SLS): Motivation

- Notice that the IV estimator held for the case in which we had one endogeneous variable x_{i1} and one instrument z_i.
- What happens if we have more instruments z_{i1} and z_{i2} than we have endogenous variables x_{i1} (let's ignore having an exogenous x_{i2} for now?
- We cannot use the standard IV estimator since it is unclear what instrument(s) should be contained in the covariances $\text{Cov}(z_i, x_{i1})$ and $\text{Cov}(z_i, y_i)$.
- We need to find some way of aggregating the information contained in z_{i1} and z_{i2}.

Two-Stage Least Squares (2SLS): Predicted Instrument

- Note that since z_{i1} and z_{i2} are both individually uncorrelated with u_i, any linear combination of z_{i1} and z_{i2} is also uncorrelated with u_i.
- The linear combination that is most correlated with x_{i1} (i.e. the one that maximises the relevance of the IVs) is obtained from the **first stage regression**

$$\begin{aligned} x_{i1} &= \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \epsilon_i \\ &= x_{i1}^* + \epsilon_i \end{aligned}$$

- The "best IV" for x_{i1} is therefore the linear combination

$$x_{i1}^* = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2}$$

- We can estimate x_{i1}^* by OLS since all of the explanatory variables z_{i1} and z_{i2} are exogenous:

$$\hat{x}_{i1} = \hat{\pi}_0 + \hat{\pi}_1 z_{i1} + \hat{\pi}_2 z_{i2}$$

- We can then use $\hat{x}_{i1}$ as an IV for x_{i1} and obtain $\hat{\beta}_1^{IV}$.

Two-Stage Least Squares (2SLS): Procedure

- As suggested in the last slide, the estimator we obtain from this procedure is called the **Two-Stage Least Squares (2SLS)** estimator.
- The reason for this name is that the estimator can be obtained by running two OLS regressions.
  1. Run OLS regression on the first stage regression

     $$x_{i1} = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \epsilon_i$$

     and calculate the predicted values $\hat{x}_{i1}$.
  2. Run OLS in the **second stage regression** which replaces x_{i1} with $\hat{x}_{i1}$

     $$y_i = \beta_0 + \beta_1 \hat{x}_{i1} + u_i$$

     obtaining $\hat{\beta}_1^{2SLS}$.
- It can be shown that $\hat{\beta}_1^{2SLS} = \hat{\beta}_1^{IV}$.
- Between step 1 and step 2 we can also compute an F-test for $H_0 : \pi_1 = \pi_2 = 0$ to test the relevance condition.

Two-Stage Least Squares (2SLS): Intuition

- Note that when we are doing 2SLS, we are replacing the endogenous variable x_{i1} with its predicted value from an OLS regression with z_{i1} and z_{i2} as explanatory variables.
- We can think of these predicted values as containing the variation in x_{i1} that is uncorrelated with u_i.
- Therefore, in the second stage regression, we are able to estimate β_1 by looking only at the correlation between y_i and x_{i1} that is uncorrelated with u_i.
- Since we are only looking at a portion of the total variation in x_{i1}, this is why our IV estimators have a higher variance than the corresponding OLS estimators.

Two-Stage Least Squares (2SLS): `ivregress`

- To implement 2SLS in Stata we need the command `ivregress`.
- Suppose we want to run the regression

$$lwage_i = \beta_0 + \beta_1 educ_i + u_i$$

  and we have two instruments $fatheduc_i$ and $motheduc_i$.
- Then to estimate $\hat{\beta}_1^{2SLS}$ in Stata we would type the command

```
ivregress 2sls lwage (educ = fatheduc motheduc)
```

- Note that this command does not estimate 2SLS in the way we outlined in slide 16. The procedure that Stata uses accounts for the variance in the first stage predicted values when calculating the standard errors for $\hat{\beta}_1^{2SLS}$. The procedure in slide 16 takes those predicted values as given in the second stage and so it will underestimate the true variance.

In-Class Exercise 1

In this exercise we are going to use 2SLS to estimate the effect of education on wages. Download the dataset CARD.dta from my website (matthewtudball.com). Consider the following regression:

$$lwage_i = \beta_0 + \beta_1 educ_i + u_i$$

where $lwage_i$ is the natural logarithm of wage and $educ_i$ is years of education.

1. Estimate the model by OLS and record the estimate $\hat{\beta}_1^{OLS}$.
2. Consider the three potential instruments $nearc4_i$ (distance to nearest 4-year college), $fatheduc_i$ (father's years of education) and $motheduc_i$ (mother's years of education). Recall that each of these three instruments must be uncorrelated with u_i, which contains all of the variables other than $educ_i$ which are related to wages $lwage_i$. Break into small groups of 2 or 3 and discuss which (if any) of these instruments will satisfy that assumption.

In-Class Exercise 1

3. Estimate the first stage of this regression by regressing $educ_i$ on $nearc4_i$, $fatheduc_i$ and $motheduc_i$. Run an F-test to test whether the instruments are jointly significant. What do you conclude about whether these instruments satisfy the relevance condition?
4. Estimate the model using the command `ivregress`.
5. Estimate the model again using the procedure on slide 16. You should end up running two OLS regressions. How do your estimates compare to those in the previous question? How about the standard errors?

In-Class Exercise 2

This is a variation on Computer Exercise 2 from Wooldridge Chapter 15. Download the dataset FERTIL2.dta from my website. This dataset includes, for women in Botswana during 1988, information on numbers of children, years of education, age and religious and economic status.

1. Estimate the model

$$children_1 = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + u_i$$

   by OLS. Holding age fixed, what is the estimated effect of another year of education on fertility? Do you think this estimate has a causal interpretation? You may talk in small groups about this.

2. The variable $frsthalf_i$ is a dummy equal to 1 if the woman was born during the first six months of the year. Do you think $frsthalf_i$ is a reasonable IV for $educ_i$? Test the relevance condition (Hint: you need to run a regression).

In-Class Exercise 2

3. Estimate the model from part 1 by using $frsthalf_i$ as an IV for $educ_i$. Compare the estimated effect of education with the OLS estimate from part 1. Interpret the coefficient on $educ_i$. Do you think this estimate has a causal interpretation?
4. Add the binary variables $electric_i$, tv_i and $bicycle_i$ to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compared the estimated coefficient on $educ_i$. Interpret the coefficient on tv_i and explain why television ownership has a negative effect on fertility. You may talk in small groups.