EC203: Applied Econometrics
Instrumental variables

Dr. Tom Martin

University of Warwick

Illustrative reading:

- Wooldridge: Chapters 15 and 16
- Dougherty: Chapters 8 and 9
- Gujarati: Chapter 19

Why use an instrumental variable?

Instrumental variables are typically used to address biases which arise from three main forms of endogeneity:

1. Omitted variable bias (OVB)
2. Simultaneity/reverse causality
3. Measurement error

Why use an instrumental variable?

Given the population model:

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i$$

we will have endogeneity biases if:

- if OLS is used and ONE of the explanatory variables is endogeneous, then ALL the OLS estimators will be biased: $E[b_j] \neq \beta_j$.
- This could be due to, for instance: OVB, reverse causality and/or measurement error.
- In this lecture we introduce the instrumental variable (IV) estimator as an alternative to OLS.

Schooling example: What is an instrumental variable?

Suppose the true population model is:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 A_i + u_i \ (A)$$

Where, Y_i is wage, X_i is years of schooling, A_i is ability and u_i is an error term which satisfies the usual CLRM assumptions. However, we estimate the short model:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i \ (B)$$

where $\epsilon_i = \beta_2 A_i + u_i$. Then we know estimating (B) via OLS:

$$\begin{aligned} b_1 =& cov(Y, X)/V(X) = cov(\alpha + \beta_1 X_i + \beta_2 A_i + u_i, X_i)/V(X_i) \\ =& \beta_1 + \beta_2 cov(A_i, X_i)/V(X_i) \end{aligned}$$

Which is the OVB formula again.

Schooling example: What is an instrumental variable?

How can we solve this bias? One solution is to use the IV estimator. The strategy primarily relies on the researcher being able to find another variable Z (the instrumental variable), which satisfies the following two conditions:

1. $cov(Z_i, X_i) \neq 0$ - **instrument relevance**
2. $cov(Z_i, \epsilon_i) = 0$ - **instrument exogeneity**.
   - The second condition is also known as an exclusion restriction: Z does not belong in the regression of interest. It only has an effect on Y through its effect on X.

- Intuitively, Z is used to break the the correlation between X and the error term ϵ_i, in the short regression (B).
- If we can find an IV we can use it to get an unbiased estimate of β_1.[1]

[1] Technically, IV is a consistent estimator, since it only has desirable properties in large samples. Think of consistency as the large sample equivalent to unbiasedness. We will continue to use bias to keep terminology constant throughout the course.

Schooling example: What is an instrumental variable?

In our example, we require an instrument Z, which satisfies:

1. relevance: correlated with schooling, X.
2. exogeneity: uncorrelated with ability, A_i, which is in the error term, ϵ_i, in the short regression (B).[2]

[2] More generally, we would require an instrument that is uncorrelated with everything that is unobservable in the error term, such as motivation, determination, organisation, conscientiousness, ... etc: Anything that is correlated with both wage and schooling.

Schooling example: What is an instrumental variable?

Of the two conditions:

1. $cov(Z_i, X_i) \neq 0$ - **instrument relevance**
2. $cov(Z_i, \epsilon_i) = 0$ - **instrument exogeneity**

- We can test the first: regress X on Z and see if the coefficient is statistically significant.[3]
- The second is not testable. Whether it is trusted will depend on the strength of the researcher's arguments.

[3] This is known as the test for weak instruments.

Schooling example: What is an instrumental variable?

Assuming we can find a variable that satisfies the previous assumptions, and given a dependent variable (Y), one endogenous regressor (X) and one instrument (Z), the IV estimator is:

$$\begin{aligned} b^{IV} &= \frac{cov(Y_i, Z_i)}{cov(X_i, Z_i)} \\ &= \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} \end{aligned}$$

Schooling example: What is an instrumental variable?

The IV estimator is unbiased if the two IV assumptions are satisfied:

$$\begin{aligned} b^{IV} =& \frac{cov(Y_i, Z_i)}{cov(X_i, Z_i)} \\ =& \frac{cov(\alpha + \beta_1 X_i + \beta_2 A_i + u_i, Z_i)}{cov(X_i, Z_i)} \\ =& \beta_1 \frac{cov(X_i, Z_i)}{cov(X_i, Z_i)} + \beta_2 \frac{cov(A_i, Z_i)}{cov(X_i, Z_i)} + \frac{cov(u_i, Z_i)}{cov(X_i, Z_i)} \end{aligned}$$

which, assuming $cov(Z_i, X_i) \neq 0$ (relevance) and $cov(A_i, Z_i) = cov(u_i, Z_i) = 0$ (exogeneity):

$$E[b^{IV}] = \beta_1$$

IV: 2 stage least squares (2SLS)

The IV estimator can be established in the following 2-step fashion. Suppose the causal relationship of interest is:

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i$$

First-stage: use OLS to regress the endogenous variable (X) on the instrument (Z)

$$X_i = \delta_0 + \delta_1 Z_i + \mu_i$$

Second-stage: use OLS to regress Y on the fitted values of X, that is, $\hat{X}$ from the first stage

$$Y_i = \alpha + \beta_1 \hat{X}_i + v_i$$

The estimator of β_1 will be equal the IV estimator, presented in the previous slide.

IV: 2 stage least squares (2SLS)

This gives us an intuitively appealing way of thinking about what an IV does:

- The 1st stage rids X (schooling) of its correlation with error term (which includes ability in our example) before carrying out the 2nd stage regression.
- Thus, IV is also known as the 2SLS estimator.
- Note, do not carry out the two stages manually because:
  1. Stata does it automatically.
  2. Although the coefficients will be identical the standard errors will be wrong.

Statistical inference with IV estimators

The standard errors of OLS estimators are smaller than those of IV estimators. In the SLR case, in large samples and assuming all CLRM and IV assumptions hold, the IV estimator b^{IV} has the following sampling distribution:

$$b^{IV} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})R^2_{X,Z}})$$

The sampling distribution can then be used to carry out population inference as usual.

Statistical inference with IV estimators

Further, we can compare IV and OLS variances:

$$V(b^{IV}) = \frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})R^2_{X,Z}}$$

$$V(b^{OLS}) = \frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})}$$

- Thus, given $0 < R^2_{X,Z} < 1$ the variance of the IV estimator is always greater than the variance of our OLS estimator.
- Further, the closer $R^2_{X,Z}$ is to zero, that is the weaker the correlation between X and Z, the greater the variance in the IVs sampling distribution.

The problem of weak instruments

Weak correlation is not only a problem for precision, it can also be problematic for bias. To see this, note:

$$\begin{aligned} b^{IV} =& \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} \\ =& \frac{Cov(Z_i, \alpha + \beta X_i + \epsilon_i)}{Cov(Z_i, X_i)} \\ =& \beta \frac{Cov(Z_i, X_i)}{Cov(Z_i, X_i)} + \frac{Cov(Z_i, \epsilon_i)}{Cov(Z_i, X_i)} \end{aligned}$$

Given $Corr(Z_i, \epsilon_i) = \frac{Cov(Z_i, \epsilon_i)}{\sigma_Z \sigma_\epsilon}$, we have:

$$\begin{aligned} b^{IV} =& \beta + \frac{Corr(Z_i, \epsilon_i)\sigma_Z \sigma_\epsilon}{Corr(Z_i, X_i)\sigma_Z \sigma_X} \\ =& \beta + \frac{Corr(Z_i, \epsilon_i)}{Corr(Z_i, X_i)} * \frac{\sigma_\epsilon}{\sigma_X} \end{aligned}$$

The problem of weak instruments

- Therefore, even if $Corr(Z, \epsilon)$ is very small, the bias of the IV estimator can be very large if Corr(Z,X) is also small (i.e. if you have a weak instrument).
- Thus, not only do weak instruments lead to large variances, but if $Corr(Z, \epsilon)$ is small then biases in IV estimates can be very large also.
- **Test for weak instruments**: an F-test of below 10 in the first stage is often seen as a indicator of weak instruments.

One endogenous variable and one IV in the MLR model

Given the general MLR model:

$$Y_{1i} = \alpha + \beta_1 Y_{2i} + \beta_2 Z_{1i} + \epsilon_i \ (A)$$

This is known as a **structural equation**: our primary equation and indicates our causal parameters of interest. Further, the new notation, Y and Z are used to highlight whether variables are endogenous (Y) or exogenous (Z). Here, Y_{2i} is modelled by the following:

$$Y_{2i} = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + \eta_i \ (B)$$

This is known as the **reduced form** for Y_{2i}, as it only contains exogenous variables.

One endogenous variable and one IV in the MLR model

Here Z_{2i} can be used as an instrument for Y_{2i} where we require:

- **Instrument relevance**: π_2 the partial effect of Z_2 on Y_2 is non-zero. This can be tested with a simple t-test.
- **Instrument exogeneity**: $Cov(Z_2, \epsilon) = 0$. This needs to be argued for.

Further note, the instrument cannot be part of the structural equation (original equation). This is known as the **exclusion restriction**. That is, Z_2 should not have a direct impact on Y_1, it should only have an impact on Y_1 **through** its impact on Y_2.

One endogenous variable and one IV in the MLR model

The above model can be estimated in the 2SLS fashion:

In the first stage: regress Y_2 on Z_1, Z_2 and compute the fitted values $\hat{Y}_2$.

In the second stage: regress Y_1 on the fitted values $\hat{Y}_2$ and Z_1.

One endogenous variable and one IV in the MLR model

Suppose our population model of interest is:

$$ln(wage_i) = \alpha + \beta_1 S_i + \beta_2 male_i + \beta_3 south_i + \epsilon_i \ (A)$$

We are concerned schooling (S_i) is endogenous. Further, suppose we think college proximity is a good instrument.[4] Then we need to check for **relevance** by running the reduced form regression:

$$S_i = \pi_0 + \pi_1 male_i + \pi_2 south_i + \pi_3 prox_i + \eta_i \ (B)$$

Then test $H_0 : \pi_3 \neq 0$ using a standard t-test, or F-test. To have a non-weak instrument we require an F-statistic greater than 10.

[4] Card (1995) uses a dummy for whether an individual grew up near a four-year college as an instrument for education.

One endogenous variable and one IV in the MLR model

Then in the second stage, we take the predicted values from regression (B), $\hat{S}_i$, and use these in place of the schooling variable in regression (A), to obtain our IV estimate.

As always you should ask yourself: does the instrument satisfy the **exogeneity condition**?

In Stata the 2SLS estimation can be run in one command:
ivregress 2sls lwages male south $(sch = prox)$

More than one IV for one endogenous variable

It is no more complicated to have more than one IV for an endogenous variable. Suppose we have the general MLR model:

$$Y_{1i} = \alpha + \beta_1 Y_{2i} + \beta_2 Z_{1i} + \epsilon_i \ (A)$$

Where Z_{1i} is exogenous but Y_{2i} is endogenous, now the reduced form for Y_{2i} is:

$$Y_{2i} = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + \pi_3 Z_{3i} + \eta_i \ (B)$$

Such that the two IVs are Z_{2i} and Z_{3i}. Where now we want to test if $H_0 : \pi_2 = 0, \pi_3 = 0$. To have non-weak instruments we require an F-statistic greater than 10.

More than one endogenous variable

Above we have considered one endogenous variable. However, we can extend the idea to more than one endogenous variable. For instance, consider the model:

$$Y_{1i} = \alpha + \beta_1 Y_{2i} + \beta_2 Y_{3i} + \beta_3 Z_{1i} + \beta_4 Z_{2i} + \beta_5 Z_{3i} + \epsilon_i$$

- The two endogenous variables are Y_2 and Y_3.
- We now need two IVs, say Z_4 and Z_5, where we require either Z_4 or Z_5 to have a significant effect in each reduced form for Y_2 and Y_3.
- Note of caution: suppose Z_4 and Z_5 only have a significant effect on Y_2 but no effect on Y_3, then we do not really have two exogenous variables partially correlated with Y_2 and Y_3. In this case we say our equation is under-identified and we will get inconsistent IV estimates.

Under, just or over

Summarising, if you have E endogenous variables and I instruments, then:

- if $E > I$ you have an under-identified model.
- if $E = I$ you have an just-identified model.
- if $E < I$ you have an over-identified model.[5]

[5] If you have an over-identified model then there are also tests for instrument exogeneity. However, this implies you can find more than one instrument for each endogenous variable, this it highly unlikely. It is more important to be able to argue for the exogeneity of your instrument (as well as its relevance, which you can test for).

Example: returns to schooling

Suppose, in the population we know:

$$Y_i = \alpha + \beta_1 S_i + \beta_2 A_i + u_i$$

Where Y_i is log wage, S_i is number of years in school, and A_i is ability. However, we don't have information on ability so we have the short model:

$$Y_i = \alpha + \beta_1 S_i + \epsilon_i$$

Where $\epsilon_i = \beta_2 A_i + u_i$. Estimating via OLS we get:

$$\hat{Y}_i = \underset{(0.185)}{-0.185} + \underset{(0.014)}{0.109} S_i$$

Example: returns to schooling

We are concerned with the CIA ($cov(X, \epsilon) = 0$) holding. Which here:

- we have $cov(S, \epsilon) = cov(S, \beta_2 A_i + u_) \neq 0$.
- So we are concerned with $cov(S, A_i) \neq 0$, since we are assuming we have only omitted ability from the regression.[6]
- Thus, we require the IV:
  1. relevant: correlated with education
  2. exogenous: uncorrelated with ability
- Any ideas for an IV?

[6] However, more generally an IV should be uncorrelated with any unobservable that jointly determines wage and education.

Example: returns to schooling

OLS regression:

$$\hat{Y}_i = \underset{(0.185)}{-0.185} + \underset{(0.014)}{0.109} S_i$$

- Example I: father's education as an IV for individual education.
- Instrument relevance: $\hat{S}_i = \underset{(0.28)}{10.24} + \underset{(0.029)}{0.269} fe_i$
- IV estimate: $\hat{Y}_i = \underset{(0.446)}{0.441} + \underset{(0.035)}{0.059} S_i$
- Instrument exogeneity: is father's education and individual ability uncorrelated?
- ANSWER: probably not a good IV.

Example: returns to schooling

OLS regression:

$$\hat{Y}_i = \underset{(0.185)}{-0.185} + \underset{(0.014)}{0.109} S_i$$

- Example II: number of siblings as an IV for education.
- Instrument relevance: $\hat{S}_i = \underset{(0.11)}{14.41} - \underset{(0.030)}{0.228} sibs_i$
- IV estimate: $\hat{Y}_i = \underset{(0.36)}{5.31} + \underset{(0.026)}{0.122} S_i$
- Instrument exogeneity: is number of siblings and individual ability uncorrelated?
- ANSWER: probably not a good IV.

Example: returns to schooling

As the above examples of IVs illustrate, finding a convincing IV is not easy. Angrist and Krueger (1991) is an influential IV paper in the returns to schooling literature. The basic idea is the following:

- In the US, people can drop out from school as soon as they turn 16.
- Given people start schooling at different ages, the length of time people have spent at school when they reach 16 varies.
- Thus, the paper proposes quarter of birth as an IV for years of schooling.

The first stage:

$$S_i = \pi_0 + \pi_1 qtr_i + \pi_2 X_2 + ... + \pi_k X_k + u_i$$

The second stage:

$$ln(wage_i) = \alpha + \beta_1 \hat{S}_i + \pi_2 X_2 + ... + \pi_k X_k + v_i$$

Policy and natural experiments

Studying policy and having knowledge about the institutional settings can provide useful for finding valid IVs:

1. In the previous example, although month of birth is not an explicit experiment, it can be viewed as a natural experiment: naturally occurring random variation.[7]
2. More generally, you could note that choice of schooling level is based on the costs and benefits of alternative choices. Thus, good instruments could come from changes in loans or subsidies that vary (increase/decrease schooling) independently of ability. That is, policy can offer potential IVs.
3. Or institutional constraints, such as changes in compulsory schooling laws, as the Angrist and Krueger example (1991) discussed.

[7] That is, if you believe month of birth is random.

Summary

In this lecture we have considered IV as a solution to problems of endogeneity, caused by for example: omitted relevant variables, measurement error, or simultaneity:

- Indeed, the IV estimator is often used by econometricians.
- In summary, an instrument is a variable Z, which satisfies:
  1. Instrument relevance: correlated with the endogenous variable
  2. Instrument exogeneity: uncorrelated with the error term. (It doesn't belong in the regression of interest.)
- However, it is no panacea:
  1. It requires a lot of thought to find an instrument that satisfies the exogeneity condition.
  2. If there is even a small degree of correlation between the instruments and error term and the instruments are weak then there can be very large inconsistencies (bias) in the IV estimates.