Applied Statistics and Econometrics

Giuseppe Ragusa

Lecture 15: Instrumental Variables

Outline

Three important threats to internal validity are:

- omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression
- simultaneous causality bias (X causes Y, Y causes X)
- errors-in-variables bias (X is measured with error)

Instrumental Variables Regression

Three important threats to internal validity are:

- omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression
- simultaneous causality bias (X causes Y, Y causes X)
- errors-in-variables bias (X is measured with error)

Condition violated

- Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ — using an **instrumental variable**, Z

IV Regression with one regressor and one instrument

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

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.

IV Regression with one regressor and one instrument

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

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an instrumental variable, Z_i, which is uncorrelated with u_i.

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

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an instrumental variable, Z_i, which is uncorrelated with u_i.
- The instrumental variable detects movements in X_i that are uncorrelated with u_i, and uses these to estimate β_1.

Endogeneity and exogeneity

- An endogenous variable is one that is correlated with u
- An exogenous variable is one that is uncorrelated with u

Historical note

"Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality. However, this definition is narrow and IV regression can be used to address OV bias and errors-in-variable bias, not just to simultaneous causality bias.

Two conditions for a valid instrument

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

For an instrumental variable (an "instrument") Z to be valid, it must satisfy two conditions:

1. **Instrument relevance:** $corr(Z_i, X_i) \neq 0$
2. **Instrument exogeneity:** $corr(Z_i, u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later).

How can you use Z_i to estimate β_1?

First Stage

- First isolates the part of X that is uncorrelated with u:

$$X_i = \pi_0 + \pi_1 Z_i + \upsilon_i$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we can estimate them by regressing X_i on Z_i using OLS, and taking the predicted value

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

Second Stage

- Replace X_i by $\hat{X}_i$ in the regression of interest and estimate β_1 by OLS

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- Because Z_i is uncorrelated with u_i (if n is large), the first least squares assumption holds (if n is large)
- This the resulting estimator is called the Two Stage Least Squares (TSLS) estimator.

TSLS with one X and one Z: Summary

Giiven a valid instrument Z_i

1. (First Stage): Regress X_i on Z_i to obtain $\hat{X}_i$
2. (Second Stage): Regress Y_i on $\hat{X}_i$; the coefficient $\hat{\beta}_1$ is the TSLS estimator, that we denote as $\hat{\beta}_1^{TSLS}$

Consistency Property

- $\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1

It can be shown that the TSLS estimator of β_1 is

$$\hat{\beta}_1^{TSLS} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})} = \frac{\widehat{cov}(Z_i, Y_i)}{\widehat{cov}(X_i, Z_i)}$$

Notice that from

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

follows that

$$\begin{aligned} cov(Y_i, Z_i) &= cov(\beta_0 + \beta_1 X_i + u_i, Z_i) \\ &= cov(\beta_0, Z_i) + \beta_1 cov(X_i, Z_i) + cov(u_i, Z_i) \\ &= 0 + \beta_1 cov(X_i, Z_i) + 0 \end{aligned}$$

Thus,

$$\beta_1 = \frac{cov(Y_i, Z_i)}{cov(X_i, Z_i)}$$

General IV regression model

So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).

We need to extend this to:

- multiple endogenous regressors ($X_1, \ldots, X_k$)
- multiple included exogenous variables ($W_1, \ldots, W_r$)
- multiple instrumental variables ($Z_1, \ldots, Z_m$)

The general IV regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Y_i is the dependent variable
- $X_{1i}, \ldots, X_{ki}$ are the **endogenous variables**
- $W_{1i}, \ldots, W_{ri}$ are the **included exogenous variables**
- $\beta_0, \ldots, \beta_{k+r}$ are the **unknown regression coefficients**

In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (m) and the number of endogenous regressors (k)

Identification and overidentification

The coefficients $\beta_1, \ldots, \beta_k$ are said to be:

- **exactly identified** if $m = k$
- **overidentified** if $m > k$
- **underidentified** if $m < k$

TSLS with a single endogenous regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- m instruments: $Z_{1i}, \ldots, Z_{mi}$
- (First Stage): Regress X_{1i} on **all** the exogenous regressors: regress X_{1i} on $W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}$ by OLS; end compute the predicted values, $\hat{X}_{1i}$, $i = 1, \ldots, n$
- (Second Stage): Regress Y_i on $\hat{X}_{1i}, W_{1i}, \ldots, W_{ri}$. The coefficients from this second stage regression are the TSLS estimators

Although we can obtain the TSLS estimator by using the two stage procedure described earlier, the standard errors of the second stage are not correct

- The second stage OLS does not account for the fact that one of the regressor, namely $\hat{X}_{1i}$ is a generated variable being the predicted values of a previous regression
- R has a command that carries out the two stages automatically and provides the correct SE

Cigarette Data

- state: Factor indicating state.
- year: Factor indicating year.
- cpi: Consumer price index.
- population: State population.
- packs: Number of packs per capita.
- income: State personal income (total, nominal).
- taxs: Average tax due to the broad-based state sales tax applied to all consumption good and to cigarette specific tax (in cents per pack)
- tax: Cigarette specific, that is, tax applied to cigarette only (in cents per pack)
- price: Average price during fiscal year, including sales tax.

Example: Demand for cigarettes

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

- $GenTax_i$ = general sales tax in state i
- $CigTax_i$ = cigarette-specific tax in state i

In this case we have:

- One endogenous variable $P_i^{cigarettes}$, $(k = 1)$
- One included exogenous variable $Income_i$
- Two instruments $(m = 2)$, thus the demand elasticity is overidentified

Example: Demand for cigarettes

```r
data("CigarettesSW", package = "ase")  ## from ASE package
library(AER)  ## Package has iv regression functions
## Create real variables
CigarettesSW$rprice <- with(CigarettesSW, price/cpi)
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
CigarettesSW$tdiff <- with(CigarettesSW, (taxs - tax)/cpi)
c1985 <- subset(CigarettesSW, year == "1985")
c1995 <- subset(CigarettesSW, year == "1995")
```

```r
## First stage
fm_s1 <- lm(log(rprice) ~ tdiff, data = c1995)
## Predict
tdiff_hat <- fitted(fm_s1)
fm_s2 <- lm(log(packs) ~ tdiff_hat, data = c1995)
fm_s2
```

```
##
## Call:
## lm(formula = log(packs) ~ tdiff_hat, data = c1995)
##
## Coefficients:
## (Intercept)    tdiff_hat
##        9.72        -1.08
```

Example: Demand for cigarettes

```r
## convenience function: need this to obtain heteroskedastic standard
## error
hc1 <- function(x) vcovHC(x, type = "HC1")

fm_ivreg <- ivreg(log(packs) ~ log(rprice) | tdiff, data = c1995)
coeftest(fm_ivreg, vcov = hc1)  ## need to pass hc1
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.720      1.528    6.36  8.3e-08 ***
## log(rprice)   -1.084      0.319   -3.40   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Demand for cigarettes, Ctd.

```
fm_ivreg2 <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
    tdiff, data = c1995)
coeftest(fm_ivreg2, vcov = hc1)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.431      1.259    7.49  1.9e-09 ***
## log(rprice)    -1.143      0.372   -3.07   0.0036 **
## log(rincome)    0.215      0.312    0.69   0.4949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Demand for cigarettes, Ctd.

```
fm_ivreg3 <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
    tdiff + I(tax/cpi), data = c1995)
coeftest(fm_ivreg3, vcov = hc1)
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.895      0.959   10.32  1.9e-13 ***
## log(rprice)       -1.277      0.250   -5.12  6.2e-06 ***
## log(rincome)       0.280      0.254    1.10     0.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Demand for cigarettes

TSLS Estimate, general tax as instrument ($m = 1$)

$$log(packs) = \underset{(1.25)}{9.43} - \underset{(0.37)}{1.14} \times log(price) + \underset{(0.31)}{0.21} \times log(income)$$

TSLS Estimate, general tax and cig-only tax ($m = 2$)

$$log(packs) = \underset{(.96)}{9.89} - \underset{(0.24)}{1.27} \times log(price) + \underset{(0.25)}{0.28} \times log(income)$$

1. Smaller SEs for $m = 2$: Using two instruments gives more information — more “as-if random variation”
2. Low income elasticity (not a luxury good); income elasticity not statistically significantly different from 0
3. Surprisingly high price elasticity

General IV model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+m} W_{mi} + u_i$$

1. **Instrument exogeneity**: $corr(Z_{1i}, u_i) = 0, \ldots, corr(Z_{mi}, u_i)$
2. **Instrument relevance**: General case, multiple X's
    - Suppose the second stage regression could be run using the

predicted values from the population first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression

General IV model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i|W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}) = 0$
2. $(Y_i, X_{1i}, , X_{ki}, W_{1i}, , W_{ri}, Z_{1i}, , Z_{mi})$ are i.i.d.
3. The X's, W's, Z's, and Y have nonzero, finite 4th moments
4. The instruments $(Z_{1i}, \ldots, Z_{mi})$ are valid.

Note

Under 1-4, TSLS and its t-statistic are normally distributed

Instrument Validity

Recall the two requirements for valid instruments:

1. Relevance (special case of one X) At least one instrument must enter the population counterpart of the first stage regression.
2. Exogeneity All the instruments must be uncorrelated with the error term: $corr(Z_{1i}, u_i) = 0, \ldots, corr(Z_{mi}, ui) = 0$

What happens...

- if one of these requirements isn't satisfied? How can you check? What do you do?
- if you have multiple instruments, which should you use?

Exogeneity of taxes as intrument

- Wealthier states may have higher consumption of cigarettes (income effect)
  - controlling for income should solve this potential endogeneity issue
- states that grow tobacco have higher consumption, but these states could extert greater political influence and thus have lower cigarette specific tax rates

$$\ln(Q_{it}^{cigarettes}) = \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + \alpha_i + u_{it}$$

where α_i is the unobservable variable summarizing the political clout of the tobacco industry in state i (state fixed effect)

To get rid of the the fixed effect we can estimate the model in time differences:

$$
\begin{aligned}
\Delta \ln(Q_{it}^{cigarettes}) &= \gamma_0 + \beta_1 \Delta \ln(P_{it}^{cigarettes}) + \beta_2 \Delta \ln(Income_{it}) + \Delta u_{it} \\
\Delta \ln(P_{it}^{cigarettes}) &= \pi_0 + \pi_1 \Delta CigTax_{it} + \pi_2 GenTax_{it} + \pi_3 \Delta \log(Income_{it}) + \Delta \eta_{it}
\end{aligned}
$$

where Δ denotes time difference, e.g. $\Delta \ln(Q_{it}^{cigarettes}) = ln(Q_{i,1995}^{cigarettes}) - ln(Q_{i,1985}^{cigarettes})$.

Panel data IV

…or we can simply include state and year dummies

$$
\begin{aligned}
ln(Q_{it}^{cigarettes}) &= \beta_0 + \beta_1 \ln(P_{it}^{cigarettes}) + \beta_2 \ln(Income_{it}) + \beta_3 DAL_{it} + \ldots + \beta_{50} DWI_{it} + u_{it} \\
\ln(P_{it}^{cigarettes}) &= \pi_0 + \pi_1 CigTax_{it} + \pi_2 \Delta GenTax_{it} + \pi_3 \log(Income_{it}) + \pi_4 DAL_{it} + \ldots + \pi_{51} DWI_{it} + \Delta
\end{aligned}
$$

```r
mod1 <- ivreg(log(packs) ~ log(rprice) + log(rincome) + state + year |
    log(rincome) + tdiff + state + year, data = CigarettesSW)

mod2 <- ivreg(log(packs) ~ log(rprice) + log(rincome) + state + year |
    log(rincome) + I(tax/cpi) + state + year, data = CigarettesSW)

mod3 <- ivreg(log(packs) ~ log(rprice) + log(rincome) + state + year |
    log(rincome) + tdiff + I(tax/cpi) + state + year, data = CigarettesSW)
```

Table 1: Two Stage Least square estimate: Demand for Cigarettes using panel data

	Dependent variable:		
	log(packs)		
	(1)	(2)	(3)
log(rprice)	−0.940***	−1.300***	−1.200***
	(0.210)	(0.230)	(0.200)
log(rincome)	0.530	0.430	0.460
	(0.340)	(0.300)	(0.310)
Constant	7.800***	9.800***	9.100***
	(1.300)	(1.400)	(1.200)
Instrumental variables	Sales Tax	Cigarette-specific tax	Both

Weak instruments

An instrument is said to be weak if its correlation with the variable is intended to instrument is very small...

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

The IV estimator is $\hat{\beta}^{TSLS} = \frac{\widehat{cov}(Y,Z)}{\widehat{cov}(X,Z)}$

- If $cov(X, Z)$ is small, then $\widehat{cov}(X, Z)$ will be small: with a weak instrument the denominator is nearly zero
- If so, the sampling distribution of (and its t-statistic) is not well approximated by its large-n normal approximation

$$\hat{\beta}^{TSLS} = \frac{\widehat{cov}(Y, Z)}{\widehat{cov}(X, Z)}$$

- If $cov(X, Z)$ is small, then small changes in $\widehat{cov}(X, Z)$ (from one sample to the next) can induce big changes in $\hat{\beta}^{TSLS}$ Suppose in one sample you calculate
- Suppose you calculate $\widehat{cov}(X, Z) = .00001$
- Thus the large-n normal approximation is a poor approximation to the sampling distribution of
- As a consequence, if instruments are weak, the usual methods of inference are unreliable – potentially very unreliable.

Measuring the strength of the instruments

Consider the first stage regression

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots, + \pi_m Z_{mi} + \pi_{m+1,i} W_{1i} + \ldots + \pi_{m+r,i} W_{ri} + u_i$$

Testing for totally irrelevant instruments

$$H_0 : \pi_1 = \pi_2 = \ldots = \pi_m = 0$$

Use a *Wald*-test using the first stage *Wald*-statistic.

Weak instruments

Weak instruments imply a small first stage Wald-statistic.

Weak instruments

Rule-of-thumb:

If the first stage Wald-statistic is less than $10 \times m$, then the set of instruments is weak.

- If so, the TSLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
- Note that simply rejecting the null hypothesis that the coefficients on the Z's are zero isn't enough – you actually need substantial predictive content for the normal approximation to be a good one.
- If there is only one instrument, the first stage *Wald*-statistic is the square of the t-statistic
- There are more sophisticated things to do than just compare Wald to $10 \times m$ but they are beyond this course.

Example: Demand for cigarettes

```r
fs1 <- lm(log(rprice) ~ log(rincome) + tdiff + state + year, data = CigarettesSW)
fs2 <- lm(log(rprice) ~ log(rincome) + I(tax/cpi) + state + year, data = CigarettesSW)
fs3 <- lm(log(rprice) ~ log(rincome) + tdiff + I(tax/cpi) + state + year,
    data = CigarettesSW)


W1 <- wald_test(fs1, testcoef = "tdiff")
W2 <- wald_test(fs2, testcoef = "I(tax/cpi)")
W3 <- wald_test(fs3, testcoef = c("I(tax/cpi)", "tdiff"))
```

Table 2: Two Stage Least square estimate: Demand for Cigarettes using panel data

	Dependent variable:		
	log(packs)		
	(1)	(2)	(3)
log(rprice)	−0.940***	−1.300***	−1.200***
	(0.210)	(0.230)	(0.200)
log(rincome)	0.530	0.430	0.460
	(0.340)	(0.300)	(0.310)
Constant	7.800***	9.800***	9.100***
	(1.300)	(1.400)	(1.200)
Instrumental variables	Sales Tax	Cigarette-specific tax	Both
First stage Wald	33.67	107.2	177.2

- The first stage *Wald*-statistic is large in all three cases
- We reject the null hypothesis that the coefficient of the instruments are equal to 0
- Since *Wald*-statistic is way larger than $10 \times m$, we also have confidence that we can use the normal critical values in conducting inference

Weak instruments

- Get better instruments (!)
- If you have many instruments, some are probably weaker than others and it's a good idea to drop the weaker ones (dropping an irrelevant instrument will increase the first-stage W)
- If you only have a few instruments, and all are weak, then you need to do some IV analysis other than TSLS
    - Separate the problem of estimation of β_1 and construction of confidence intervals
    - This seems odd, but if TSLS isn't normally distributed, it makes sense (right?)

Checking instrument exogeneity

- **Instrument exogeneity**: All the instruments are uncorrelated with the error term: $corr(Z_{1i}, u_i) = 0, \ldots, corr(Z_{mi}, u_i) = 0$
- If the instruments are correlated with the error term, the first stage of TSLS doesn't successfully isolate a component of X that is uncorrelated with the error term, so the prediction of X_i is correlated with u_i and TSLS is inconsistent.
- If there are more instruments than endogenous regressors, it is possible to test – partially – for instrument exogeneity

Consider the simplest case:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- Suppose there are m valid instruments: $Z_{1i}, Z_{2i}, \ldots, Z_{mi}$
- Then you could compute m separate TSLS estimates.
- Intuitively, if these m TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The J-test of overidentifying restrictions makes this comparison in a statistically precise way. This can only be done if #Z's > #X's (overidentified).

Checking instrument exogeneity

- First estimate the equation of interest using TSLS and all m instruments; compute the predicted values of Y_i, using the actual X's (not the X's used to estimate the second stage)
- Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
- Regress $\hat{u}_i$ on $Z_{1i}, \ldots, Z_{mi}, W_{1i}, \ldots, W_{ri}$
- Compute the Wald-statistic testing the hypothesis that the coefficients on $Z_{1i}, \ldots, Z_{mi}$ are all zero
- The J-statistic is $J = W$, the Wald stat.

- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m - k$ degrees of freedom
- If $m = k$, $J = 0$ (does this make sense?)
- If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.
- How do we calculate the J-statistics with R? `wald_test()`.....

J-TEST

```
ivm <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff +
    I(tax/cpi), data = c1995)
## Calcluate residual
uhat <- resid(ivm)
## Regress residuals on instruments
Jreg <- lm(uhat ~ log(rincome) + tdiff + I(tax/cpi), data = c1995)
wald_test(Jreg, testcoef = c("tdiff", "I(tax/cpi)"))
```

```
## Wald test
##
## Null hypothesis:
## tdiff = 0
## I(tax/cpi) = 0
##
##   q    W pvalue
##   2 0.32   0.85
```

- Careful here: the *p*-value is calculated assuming that $W \sim \chi^2_m$. Instead, in this case $W \sim \chi^2_{m-k}$

J-test

- Careful here: the p-value is calculated assuming that $W \sim \chi^2_m$. Instead, in this case $W \sim \chi^2_{m-k}$

```
wald_test(Jreg, testcoef = c("tdiff", "I(tax/cpi)"))
```

```
## Wald test
##
## Null hypothesis:
## tdiff = 0
## I(tax/cpi) = 0
##
##   q    W pvalue
##   2 0.32   0.85
```

- $m = 2$ number instruments - $k = 1$ number of endogenous variables - Critical value of χ^2_1: - $\alpha = 0.10 \implies cv = 2.70$ - $\alpha = 0.05 \implies cv = 3.84$