Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

24 August, 2018

Assignment 1

Deadline: 29 August at 12:00pm (noon, mid-day)

Reminder: my deadlines are *very* sharp!
(If you submit at 12:01pm you will receive a mark of zero!)

No extensions given under any circumstances!

Note: I do not offer any help on solving the assignment!

Roadmap

Let's still consider the simple linear model with one endogenous regressor and one IV

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad \text{where } \mathrm{E}[u_i|X_i] \neq E[u_i]$$

Last week we learned that the TSLS estimator has the form

$$\hat{\beta}_{1,TSLS} := \frac{\sum_{i=1}^{n}(\hat{X}_i - \bar{\hat{X}})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(\hat{X}_i - \bar{\hat{X}})(\hat{X}_i - \bar{\hat{X}})} = \frac{s_{\hat{X}Y}}{s^2_{\hat{X}}}$$

We will prove that, alternatively, the TSLS estimator can be expressed like

$$\hat{\beta}_{1,TSLS} := \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} = \frac{s_{ZY}}{s_{ZX}}$$

This last result is the basis for deriving the asymptotic distribution of $\hat{\beta}_1^{TSLS}$

Starting point of that derivation are the four OLS assumptions plus one additional assumption that is required for the IV

The IV regression assumptions are

1. $\mathrm{E}[u_i|X_i] \neq \mathrm{E}[u_i]$, that is X_i is endogenous
2. (Y_i, X_i, Z_i) are i.i.d.
3. X_i, Y_i, Z_i all have nonzero, finite 4th moments
4. heteroskedasticity
5. Z_i is a valid IV

Instead of deriving the asymptotic distribution, we just state the result…

Theorem (Asymptotic Distribution of TSLS Estimator)

*The **asympototic distribution of the TSLS estimator** $\hat{\beta}_1^{TSLS}$ under IV regression assumptions 1 through 5 is*

$$\hat{\beta}_1^{TSLS} \overset{approx.}{\sim} N\left(\beta_1, \frac{1}{n}\frac{Var\left((Z_i - \mu_Z)u_i\right)}{\left(Cov(Z_i, X_i)\right)^2}\right)$$

Standard errors follow readily:

$$\text{SE}(\hat{\beta}_1^{TSLS}) = \frac{s_{uZ}}{\sqrt{n}s_{XZ}}$$

where

$$s_{uZ}^2 := \frac{1}{n}\sum_{i=1}^{n}\left((Z_i - \bar{Z})\hat{u}_i\right)^2$$

$$s_{XZ} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})$$

These are the se that Stata's ivregress command calculates

The derivation of the asymptotic distribution is not that difficult

It's similar to the derivation of the asymptotic distribution of the OLS estimator under heteroskedasticity during lecture 8 of last semester's EMET2007

A simple comparison between the asymptotic distributions of the OLS estimator from last semester and the TSLS estimator from this semester is illustrative...

Recall (from last semester) that the OLS estimator was given by

$$\hat{\beta}_{1,OLS} := \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})} = \frac{s_{XY}}{s_{XX}}$$

Its asymptotic distribution was

$$\hat{\beta}_1 \overset{approx.}{\sim} \mathrm{N}\left(\beta_1, \frac{1}{n}\frac{\mathrm{Var}\left((X_i - \mu_X)u_i\right)}{(\mathrm{Cov}(X_i, X_i))^2}\right)$$

In comparison, the TSLS estimator was given by

$$\hat{\beta}_{1,TSLS} := \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} = \frac{s_{ZY}}{s_{ZX}}$$

Its asymptotic distribution was

$$\hat{\beta}_1^{TSLS} \overset{approx.}{\sim} \mathrm{N}\left(\beta_1, \frac{1}{n}\frac{\mathrm{Var}\left((Z_i - \mu_Z)u_i\right)}{(\mathrm{Cov}(Z_i, X_i))^2}\right)$$

Digression: why are the standard errors incorrect when we calculate the TSLS estimator in two separate steps in Stata (rather than using the inbuilt ivregress command)?

The main idea of the first stage in TSLS is to extract the exogenous part of X_i

Literally, that exogenous part is $\tilde{X}_i := \pi_0 + \pi_1 Z_i$

But we do not know the π_0 and π_1

Instead we have to estimate them by $\hat{\pi}_0$ and $\hat{\pi}_1$

Then we calculate $\hat{X}_i := \hat{\pi}_0 + \hat{\pi}_1 Z_i$

Because $\hat{\pi}_0 \neq \pi_0$ and $\hat{\pi}_1 \neq \pi_1$ it follows that $\hat{X}_i \neq \tilde{X}_i$

This shows that $\hat{X}_i$ is a noisy measurement of $\tilde{X}_i$

The estimate $\hat{X}_i$ itself has a sampling distribution which needs to be accounted for when using it in the second stage regression

When doing the two stages of TSLS estimation by hand, we do not make this adjustment

Stata's ivregress, in contrast, does

Roadmap

Here is a more general version of the model we studied last week:

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

- Y_i is the dependent variable
- $X_{1i}, \ldots, X_{ki}$ are endogenous regressors (correlated with u_i)
- $W_{1i}, \ldots, W_{ri}$ are exogenous regressors (uncorrelated with u_i)
- $\beta_1, \ldots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i}, \ldots, Z_{mi}$ are instrumental variables (the excluded exogenous variables)

We therefore have

- k endogenous explanatory variables
- r exogenous explanatory variables
- m instrumental variables

Reality check:
In practice you will mostly have $k \leq 2$ and $m \leq 5$

Depending on the relationship between k and m, we can or cannot estimate the model from the previous slide

New terminology:

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

- **exactly identified** if m = k

  There are just enough instruments

  can estimate all coefficients

- **overidentified** if m $>$ k

  There are more than enough instruments

  can estimate all coefficients

- **underidentified** if m $<$ k

  There are too few IV;

  cannot estimate any coefficients; need more IV

  for example: $k = 1$ and $m = 0$

The five IV regression assumptions in the general model:

1. $\mathrm{E}[u_i|X_{ji}] \neq \mathrm{E}[u_i]$ for $j = 1, \ldots, k$
   $\mathrm{E}[u_i|W_{1i}, \ldots, W_{ri}] = \mathrm{E}[u_i]$
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ are i.i.d.
3. The W_i, X_i, Y_i, Z_i all have nonzero, finite 4th moments
4. heteroskedasticity
5. $(Z_{1i}, \ldots, Z_{mi})$ are valid IV

How does two-stage estimation work in the general model?

Let's study a more tractable version of the general model:

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

- 1 endogenous explanatory variable
- r exogenous explanatory variables
- m instruments

(The model with k endogenous variables is just too tedious for the purpose of presenting it in a lecture!)

Is β_1 here underidentified, exactly identfied or overidentified?

Instrument validity in that model means:

- Instrument relevance:
  at least one of the $Z_{1i}, \ldots, Z_{mi}$ has a non-zero coefficient in the equation

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

- Instrument exogeneity: $\rho_{Z_{1i} u_i} = \cdots = \rho_{Z_{mi} u_i} = 0$

Estimation of the more general model again follows two stages

1. Run OLS of X_i on a constant and all exogenous regressors

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

(don't forget to include the W's here!)

- Obtain OLS estimates $\hat{\pi}_0, \ldots, \hat{\pi}_{m+r}$
- Compute the predicted values of X_i:
  $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + \cdots + \hat{\pi}_{m+r} W_{ri}$
- The predicted $\hat{X}_i$ can be viewed as the exogenous part of X_i

2. Run OLS of Y_i on a constant, $\hat{X}_i$ as well as $W_{1i}, \ldots, W_{ri}$
   - Obtain OLS estimates and denote them by $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{1+r}^{TSLS}$
   - These are the TSLS estimates
   - This procedure is based on the following rewriting of the regression model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i \\ &= \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + (u_i + \beta_1 (X_i - \hat{X}_i)) \\ &=: \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + w_i \end{aligned}$$

   (Will prove soon why this is a good idea)

Roadmap

We still focus on a model with one endogenous regressor:

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

First stage regression:

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

- The instruments are relevant if at least one of $\pi_1, \ldots, \pi_m$ are nonzero
- The instruments are said to be weak if all the $\pi_1, \ldots, \pi_m$ are either zero or nearly zero
- Weak instruments explain very little of the variation in X, beyond that explained by the W's

If instruments are weak, the sampling distribution of TSLS and its t-statistic are not normal, even with n large

Consider the simplest case:

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

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

Recall that the IV estimator is $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX}$

- If π_1 is zero or small, then s_{ZX} will be small:
  With weak instruments, the denominator is nearly zero
- If so, the asymptotic distribution of $\hat{\beta}_1^{TSLS}$ is not normal
- standard errors, t-statistic, p-values and confidence intervals are all wrong

How can you know if your IV are weak? It can be tested!

Simply run the first stage regression

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

and test statistical significance of $\pi_1, \ldots, \pi_m$

Simple F-test for $\pi_1, \ldots, \pi_m$

Small F is indicative of weak instruments, large F suggests that instruments are not weak

Rule of thumb: If the first stage F-statistic exceeds 10, then the set of instruments is not weak

If so, the TSLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading

Simply rejecting the joint null hypothesis that the $\pi_1, \ldots, \pi_m$ are zero is not enough: need the $\pi_1, \ldots, \pi_m$ to be sufficiently far away from zero to provide predictive content

Where exactly does the rule of thumb ("F-stat larger than 10") come from? A: Heavy econometric/mathematical machinery

What should you do if you have weak IV?

- Get better instruments (often easier said than done!)
- 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 F)
- If you only have a few instruments, and all are weak, then you need to do some IV analysis other than TSLS
  - Limited Information Maximum Likelihood (LIML) estimation
  - Moreira's (2003) conditional likelihood test
  - Anderson-Rubin (1949) test

Roadmap

Instrument exogeneity:
All the instruments are uncorrelated with the error term
$\rho_{Z_1 u} = \cdots = \rho_{Z_m u} = 0$

If the instruments are correlated with the error term, the first stage of TSLS cannot isolate a component of X that is uncorrelated with the error term, so $\hat{X}$ is correlated with u 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:

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

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}
- Then you could compute two separate TSLS estimates
- Intuitively, if these 2 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 $m > k$ (overidentified)

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

Conducting the "J-test":

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$ according to
$$\hat{Y}_i := \hat{\beta}_0^{TSLS} + \hat{\beta}_1^{TSLS} X_{1i} + \ldots + \hat{\beta}_k^{TSLS} X_{ki} + \hat{\beta}_{k+1}^{TSLS} W_{1i} + \ldots + \hat{\beta}_{k+r}^{TSLS} W_{ri}$$
2. Compute the residuals: $\hat{u}_i = Y_i - \hat{Y}_i$, where
3. Regress $\hat{u}_i$ on $Z_{1i}, \ldots, Z_{mi}, W_{1i}, \ldots, W_{ri}$
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i}, \ldots, Z_{mi}$ are all zero
5. Define the J-statistic is $J := m \cdot F$

Testing the J-statistic

- Null hypothesis H_0: all the instruments are exogenous
- Asymptotic distribution of the J-statistic under null: chi-squared distribution with $m - k$ degrees of freedom
- The J-test is a one-sided test (rejection region on right)
- The 95% cut-off value depends on the degrees of freedom

m	k	df	95% cut-off
2	1	1	3.84
3	1	2	5.99
4	1	3	7.81
5	1	4	9.49

- If some instruments are exogenous and others are endogenous, the J-statistic will be large, leading to a rejection of H_0
- If $m = k$, then $J = 0$

Roadmap

Example: demand for cigarettes

We want to estimate price and income elasticity of cigarette demand

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + \beta_2 \ln(Income_i) + u_i$$

Data: 48 states in the US, year 1995

- Q_i: Annual per capita cigarette sales (in packs) in state i
- P_i: real average retail cigarette price per pack in state i
- $Income_i$: real per capita income in state i
- Which one is/are the endogenous variable/s?
- Instruments:
  - Z_{1i}: general sales tax in state i
  - Z_{2i}: cigarette specific tax in state i
- Is β_1 over-, under-, or exactly identified?

First stage, testing for weak IV

```
regress lavgprice lpcincome cigtax salestax if year==1995, robust

Linear regression                               Number of obs =      48
                                                F(  3,    44) =  263.12
                                                Prob > F      =  0.0000
                                                R-squared     =  0.9403
                                                Root MSE      =  .03226

------------------------------------------------------------------------------
             |               Robust
  lavgprice  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lpcincome  |   .1083446   .0396525     2.73   0.009     .0284302     .188259
     cigtax  |   .0093517   .0008698    10.75   0.000     .0075987    .0111047
   salestax  |   .0108898   .0021366     5.10   0.000     .0065838    .0151958
      _cons  |   4.103035   .0883802    46.42   0.000     3.924916    4.281153
------------------------------------------------------------------------------

testparm cigtax salestax                    F-TEST OF IV

( 1)  cigtax = 0
( 2)  salestax = 0

F(  2,    44) =  209.68                     EXCEEDS RULE OF THUMB
     Prob > F =    0.0000
```

Using Stata's "ivregress"

```
ivregress 2sls lpcpack lpcincome (lavgprice = cigtax salestax) if year==1995, robust

Instrumental variables (2SLS) regression              Number of obs =      48
                                                      Wald chi2(2)  =   34.51
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.4294
                                                      Root MSE      =  .18189

------------------------------------------------------------------------------
             |               Robust
    lpcpack  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lavgprice  |  -1.277424   .2416838    -5.29   0.000    -1.751115   -.8037324
  lpcincome  |   .2804045   .2458274     1.14   0.254    -.2014083    .7622174
      _cons  |   9.894955   .9287578    10.65   0.000     8.074623    11.71529
------------------------------------------------------------------------------
Instrumented:  lavgprice
Instruments:   lpcincome cigtax salestax
```

Large price elasticity, insignificant income elasticity

Test of Overidentifying Restrictions, lazy way

Immediately after running "ivregress" execute:

```
estat overid

Test of overidentifying restrictions:

Score chi2(1)           =  .334736  (p = 0.5629)
```

J-statistic equals 0.33

From table (a few slides earlier), critical cut-off is 3.84 (because $k = 1$ and $m = 2$)

Cannot reject the null hypothesis that IV are exogenous

Test of Overidentifying Restrictions, clumsy way

```
predict uhat, residuals
regress uhat lpcincome cigtax salestax if year==1995, robust

Linear regression                               Number of obs =      48
                                                F(  3,    44) =    0.11
                                                Prob > F      =  0.9530
                                                R-squared     =  0.0069
                                                Root MSE      =  .18932
------------------------------------------------------------------------------
             |               Robust
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lpcincome |    .031566   .2333859     0.14   0.893    -.4387923    .5019243
      cigtax |   -.001507    .003303    -0.46   0.650    -.0081638    .0051498
    salestax |    .006332   .0121997     0.52   0.606    -.0182548    .0309188
       _cons |  -.0654967   .5900834    -0.11   0.912    -1.254732    1.123738
------------------------------------------------------------------------------

testparm cigtax salestax                    F-TEST OF IV

( 1)  cigtax = 0
( 2)  salestax = 0

      F(  2,    44) =    0.16                COMPUTE J-STAT BY MULTIPLYING
           Prob > F =    0.8536              BY m=2 WHICH GIVES EXACTLY
                                             SAME RESULT AS PREVIOUS SLIDE
```

Problem Solving Exercises

1. A sociologist studies if a mother's prenatal smoking causes her child to become more violent during adolesence. Using household survey data that collects information on mothers and their children, the sociologist runs the regression:

   $$\text{Fights}_i = \beta_0 + \beta_1 \text{Smoking}_i + \beta_2 \text{Age}_i + u_i, \text{where}$$

   - Fights_i is the number of fights teenager i was involved in during the month before the survey interview
   - Smoking_i is a dummy variable indicating if the mother of teenager i smoked during pregnancy
   - Age_i is the age of teenager i

   Does $\hat{\beta}_1$ estimate the causal effect of a mother's smoking during pregnancy on her child's violence later in life?