Empirical Economics

Instrumental Variables Regression (Stata Seminar 4)

Teacher: Andrew Proctor
andrew.proctor@phdstudent.hhs.se

October 6, 2017

Outline

The endogeneity problem and IV solution

- Suppose we want to estimate:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- But we know that x_i is *endogenous* (that is, $Cov(x_i, u_i) \neq 0$) and we can't reasonably find control variables to remedy this problem. What can we do?
- One possibility is to look for an 'instrument' variable z_i that only affects our outcome y_i through it's effect on x_i. So that:

  z_i **is a *relevant* instrument:** $Cov(z_i, x_i) \neq 0$ ()
  z_i **is a *valid* instrument (exogenous):** $Cov(z_i, u_i) = 0$

The Instrumental Variables equations

- Our resulting model is then:

$$x_i = \pi_0 + \pi_1 z_i + v_i \qquad \textbf{(first stage)}$$
$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \textbf{(structural equation)}$$

- Another eq. of interest is the the relationship of y_i with z_i.

$$y_i = \gamma_0 + \gamma_1 z_i + \epsilon_i \qquad \textbf{(reduced form)}$$

- How do we estimate our parameter of interest (β_1) using these equations and our assumptions about the instrument?

Deriving the IV estimator

- With a single instrument, we have:

$$Cov(y_i, z_i) = Cov(\beta_0 + \beta_1 x_i + u_i, z_i) = \beta_1 Cov(x_i, z_i)$$

$$\implies \beta_1 = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$

- Furthermore note, using the usual OLS formulas:

$$\gamma_1 = Cov(y_i, z_i)/V(z_i)$$
$$\pi_1 = Cov(x_i, z_i)/V(z_i)$$

- Hence we have:

$$\beta_1 = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \frac{Cov(y_i, z_i)/V(z_i)}{Cov(x_i, z_i)/V(z_i)} = \frac{\gamma_1}{\pi_1}$$

Generalizing IV regression

- We began by assuming a model with the same number of instruments as explanatory variables (e.g. 1). This is called the *just-identified* case.
- Sometimes, however, you may have more potential instruments than endogenous variables (*over-identified case*).
  - The *just identified* IV regression is just a special case of **two-stage least squares**, which estimate effects by using *at least as many* instruments as there are endogenous regressors.
  - The first stage with k instruments case appears as:

$$x_i = \pi_0 + \pi_1 z_i + \pi_2 z_{2,i} + ... + \pi_k z_{k,i} + v_i$$

- Deriving the 2SLS Estimates *(basic idea)*:
  - 2SLS is essentially derived by replacing x_i in the structural equation with its fitted values from the 1^{st} stage, then performing OLS (taking into account that $\hat{x}_i$ is a statistic when estimating variance).

Including other exogenous variables in 2SLS

- Typically, you will probably want to include other variables in your model besides the endogenous variable(s) you are instrumenting. That is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + u_i,$$

where x_i is an endogenous variable, and w_i is exogenous, but not an explicit instrument for x_i.

- All exogenous variables in the structural equation should also **always** appear in the first-stage.

$$x_i = \pi_0 + \pi_1 z_i + \pi_2 w_i + v_i$$

Including other exogenous variables in 2SLS ctd

- In econometrics, W_i are referred to as the *"included instruments"* (because they are included in the structural equation) and Z_i are referred to as the *"excluded instruments"* (because they do not appear in the structural equation).
- The requirement remains the same that you need as many excluded instruments as there are endogenous explanatory variables.

Performing IV regression using the -ivreg2- package

- There are multiple regression commands/packages to perform IV/2SLS in Stata, but the most comprehensive is **ivreg2**.
- Since this package was not originally written by Stata Corp (but instead was contributed by the Stata user community), you may need to install it.
  - To install a package, in the package enter **ssc install [NAME OF PACKAGE]**

Performing IV regression using the -ivreg2- package ctd

- Basic syntax of **ivreg2**:

  ivreg2 [DEPVAR] [EXOGENOUS VARS] ([ENDOG VAR] = [EXCL. INSTRUMENTS])

- Common options for **ivreg2**:
  - You can specify heteroskedastic-robust or clustered standard errors, in the same manner as **regress** or **xtregress**.
  - Include first-stage results by indicating **first**.
  - Include reduced-form results by indicating **rf**.
  - Omit results for identification tests by indicating **noid**.

Basics of IV/2SLS

```
.              ivreg2 LWKLYWGE YR2* AGEQ c.AGEQ#c.AGEQ (EDUC = QTR1* QTR22* Q
Warning - collinearities detected
Vars dropped:         YR29 QTR129 QTR328 QTR329

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =   247199
                                                      F( 12,247186) =     8.68
                                                      Prob > F      =   0.0000
Total (centered) SS     =  104853.0198                Centered R2   =   0.1053
Total (uncentered) SS   =  6674371.774                Uncentered R2 =   0.9859
Residual SS             =  93812.46542                Root MSE      =     .616

------------------------------------------------------------------------------
    LWKLYWGE |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        EDUC |   .1299233   .0333514     3.90   0.000     .0645558    .1952908
        YR20 |  -.1115319   .0708754    -1.57   0.116    -.2504451    .0273812
        YR21 |  -.1065764    .064942    -1.64   0.101    -.2338604    .0207076
        YR22 |  -.1026378   .0556624    -1.84   0.065    -.2117341    .0064584
        YR23 |  -.0928895   .0495341    -1.88   0.061    -.1899747    .0041956
        YR24 |   -.079894   .0425343    -1.88   0.060    -.1632597    .0034717
        YR25 |  -.0569173   .0337398    -1.69   0.092     -.123046    .0092115
        YR26 |  -.0423755   .0271461    -1.56   0.119    -.0955809    .0108299
        YR27 |  -.0187343   .0174877    -1.07   0.284    -.0530095    .0155409
        YR28 |   .0003651   .0102736     0.04   0.972    -.0197708     .020501
        YR29 |          0  (omitted)
        AGEQ |    .142279   .0703048     2.02   0.043     .0044843    .2800738
             |
c.AGEQ#c.AGEQ|  -.0013786   .0007866    -1.75   0.080    -.0029202     .000163
             |
       _cons |    .121397   1.650096     0.07   0.941    -3.112731    3.355525
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):          25.305
                                                 Chi-sq(26) P-val =    0.5018
------------------------------------------------------------------------------
Collinearities detected among instruments: 1 instrument(s) dropped
Instrumented:         EDUC
Included instruments: YR20 YR21 YR22 YR23 YR24 YR25 YR26 YR27 YR28 AGEQ
                      c.AGEQ#c.AGEQ
Excluded instruments: QTR1 QTR120 QTR121 QTR122 QTR123 QTR124 QTR125 QTR126
                      QTR127 QTR128 QTR220 QTR221 QTR222 QTR223 QTR224 QTR225
```

Evaluating instrument relevance

- Recall that we in order for a variable, z_i to be a relevant instrument for endogenous variable, x_i, we require:

$$Cov(z_i, x_i) \neq 0$$

- Or, in terms of $\beta_{IV} = \frac{\gamma_1}{\pi_1}$, we require: $\pi_1 \neq 0$
- Generally speaking, we can test this condition with a t-test (or F-test in the case of multiple instruments) of the excluded instruments in the first-stage.
- **ivreg2** by default reports more advanced tests for the relevance condition, which it calls "underidentification tests"
  - Because if instruments are irrelevant, then the number of relevant instruments is less than the number of endogenous explanatory variables and so the model is *underidentified*.

Weak Instruments

- Two-stage least squares, however, in fact has a problem not only if $\pi = 0$ but even if $\pi \approx 0$. Why?
    - Although IV regression is consistent, it does suffer from finite sample bias. This bias is inversely related to the correlation of x_i with z_i.
    - As the number of instruments increase for a given endogenous variable (each with $\pi \approx 0$), the bias of 2SLS becomes as large as OLS.
- Returning again to the F-test for the excluded instruments, a good "rule of thumb" is that (for a single endogenous explanatory variable) the F-Stat should be at least 10.
    - $F^{crit} = 10$ in this case corresponds to a relative bias of ∼10% compared to the bias of OLS.
    - **ivreg2** presents weak identification test statistics and the critical values corresponding to different levels of relative bias.

Evaluating instrument validity

- In addition to instrument relevance, we also require that the instrument is *valid* (that is, exogenous).

$$Cov(z_i, u_i) = 0$$

- In general, it is not possible to test instrumental validity.
- But if you have more than one instrument for a single endogenous variable, you can perform a type of falsification test of your instrumental validity assumptions:
  - The **test of overidentifying restrictions** assumes (under the null) that all instruments are valid.
    - If only some of the proposed instruments are valid, the test will tend to reject the null.
    - If none of the instruments are valid, however, the test is not helpful.
- **ivreg2** reports the results of the test of overidentifying restrictions by default.

Instrument Validity

LWKLYWGE	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
EDUC	.0517289	.0157576	3.28	0.001	.0208445	.0826133
c.AGEQ#c.AGEQ	-.0014031	.000026	-53.96	0.000	-.0014541	-.0013521

```
Underidentification test (Anderson canon. corr. LM statistic):     176.566
                                          Chi-sq(30) P-val =    0.0000

Weak identification test (Cragg-Donald Wald F statistic):            5.886
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    21.42
                                         10% maximal IV relative bias    11.32
                                         20% maximal IV relative bias     6.09
                                         30% maximal IV relative bias     4.29
                                         10% maximal IV size             86.17
                                         15% maximal IV size             44.78
                                         20% maximal IV size             30.72
                                         25% maximal IV size             23.65
Source: Stock-Yogo (2005).  Reproduced by permission.

Sargan statistic (overidentification test of all instruments):      93.212
                                          Chi-sq(29) P-val =    0.0000

Instrumented:         EDUC
```

Intro to the -estout- package

- A key component of effective data analysis is professional presentation of results.
- Likely the best Stata package for preparing professional-looking descriptive statistic and regression output tables is the **-estout- package**.
- We will briefly go over the basics of using this package to produce regression output tables.

Basics of the -estout- package

- The first step to creating table with -estout- is to store your the results of your table.
  - To do this, simply put **eststo:** immediately before your regression command (on the same line).
- After saving any regressions that you want to appear in a table, use the **esttab** command to save generate the table (and save it to disk).
  - The syntax is **esttab using "[YOURFILENAME]",** ***options***
- To clear the saved results, use the command **eststo clear**.

Specifying Table Output with **esttab**

- There are several common options:
  - Most of the time, you will want to add a title for your table by specifying the option **title(“[YOUR TITLE HERE]”)**
  - Titles for each regression model saved in a table can be specified by **mtitles("[TITLE1]" "[TITLE2]" ...)**
  - You should also specify the format that you want your table saved as. Formats include **scml csv rtf html** & **tex**.
  - A great option is **onecell**, which specifies that estimates and standard errors should appear in the same cells of your table.
  - **label** displays variable labels instead of names (generally preferable).
  - If you have long variable names/labels, use **wrap** and **noabbrev** to word-wrap and not abbreviate these labels.
  - If your table is big, you can make it more compact by specifying the option **compress**.
  - Finally, use the option **replace** to overwrite the table if it has been previously saved.

Local Macros and Scalars

Sometimes you may want to store information not as a variable in a dataset, but just as a short-hand for values or text that you use frequently. A good way to do this is with **scalars** and **locals**.

- To save a specific number (not as a variable), you can define it as a sclar. To do this, the syntax is **scalar [SCALARNAME] = [VALUE/EXPRESSION]**. Scalars can then be multiplied or used in expressions just like a normal variable.
- Other times, you may want to refer to a long expression (of any sort, a list of variables, values, part of a regression code etc) in a shorthand to avoid having to rewrite values repeatedly.
  - To do this, define a local via the syntax:
    local [NAMEFORLOCAL] [EXPRESSION...]
  - To then refer to the local, put it in quotes just like the index values in a loop, like this: `[NAMEFORLOCAL]'.