An Introduction to Stata for Economists

Part III
Instrumental variables in practice

Steve Bond and Stefan Hubner [1]

[1] We thank Marianne Bruins (Oxford) for sharing these slides

In this class

- IV estimation: Review
- Extended example of IV: Card (1995)
- Testing the requirements for IV
- Weak instruments

IV estimation: Review

- Linear regression model with K parameters ($\beta = (\beta_1, ..., \beta_K)$):

$$y_i = x_i'\beta + u_i$$

- Problem: $\text{cov}(x_i, u_i) \neq 0 \Rightarrow$ OLS assumption violated!
- Solution: use IV, with a vector of L instruments $\mathbf{z}_i$
- Note: the instruments z_i consist of 1) additional variables AND 2) any exogenous variables in $\mathbf{x}_i$
- The instruments z_i must:
  - Be **informative**: $\mathbb{E}[z_i x_i] \neq 0$
  - Be **valid**: $\mathbb{E}[z_i u_i] = 0$
  - Satisfy the **order condition**: $L \geq K$ (# of exogenous variables $\geq$ # of parameters)
  - Satisfy the **rank condition**: $\text{rk}\,\mathbb{E}[z_i x_i'] = K$ (each endogenous variable has at least one separate, informative instrumental variable)

Example: returns to schooling

- Consider:

$$\text{wage}_i = \beta_e \text{educ}_i + x_i' \beta_x + u_i$$

  where:

  - educ_i = years of schooling
  - x_i = exogenous variables (and a constant)

- Want to know β_e, the average effect of an additional year of schooling on wages
- But individuals with higher ability have higher levels of schooling and higher wages
  - Ability is an omitted variable $\Rightarrow$ omitted variable bias!

Example: returns to schooling

Review of Omitted Variable Bias:

- Suppose the true model is:

$$\text{wage}_i = \beta_e \text{educ}_i + \beta_a \text{ability}_i + u_i$$

  but we regress wage_i on educ_i alone.
- Results from the lecture notes imply that

$$\text{plim}\, \hat{\beta}_e = \beta_e + \beta_a \gamma_e$$

  where γ_e is the coefficient on educ_i in the regression

$$\text{ability}_i = \gamma_e \text{educ}_i + \eta_i$$

- We would expect $\beta_a > 0$ (greater ability implies a higher wage), and $\gamma_e > 0$ (ability is positively correlated with educational attainment)
- Therefore $\hat{\beta}_e$ will be biased upwards. (But remember, in general, the direction of the bias isn't clear when the other regressors x_i are also included.)

Example: returns to schooling

- Large number of IV papers in the early 90s estimating returns to schooling, we will replicate results of Card (1995):
  - Used distance from a 4-year college as instrument
  - Uncorrelated with ability
  - Correlated with likelihood of attending college

Two-stage least squares estimation

Conceptual review:

- Linear regression model:

$$y_i = x_i'\beta + u_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_i$$

- The variables x_i are divided into 2 groups: 1. endogenous variables (x_{1i}), and 2. exogenous variables (x_{2i})
  - **Remember**: z_i contains all elements of x_{2i} as well as additional instrumental variables
- Estimate using **Two-Stage Least Squares (2SLS):**
  - First stage:
    - regress x_{1i} on z_i
    - recover fitted values $\hat{x}_{1i}$
  - Second stage:
    - regress y_i on $(\hat{x}_{1i}, x_{2i})$

Two-stage least squares estimation

Implementation in STATA

- ivreg2: computes IV estimates using 2SLS
- Syntax:

  `ivreg2` depvar (endogenous variables = additional instrumental variables) exogenous variables, options

  - **options:** `robust` or `vce(r)` uses heteroskedasticity-robust standard errors
  - `first` shows the first-stage regression results and diagnostic statistics
  - `endog(`endogenous variables`)` tests for the endogeneity of the specified endogenous regressors
- Exogenous variables x_{2i} are automatically included in the first stage regression
- **Remember:** z_i consists of (original) exogenous variables + additional instrumental variables
- ivreg2 is not automatically included in the Stata library so you may need to install it (`ssc install ivreg2`)

Example: Card (1995)

Exercise 1

- Open the Card dataset by selecting File, then Open
- The dataset can be found here: http://hubner.info/#teaching
- Run the OLS regression:

```
regress lwage educ exper expersq black south
    smsa reg661 reg662 reg663 reg664 reg665
    reg666 reg667 reg668 smsa66, vce(r)
```

- Run the 2SLS regression:

```
ivreg2 lwage (educ=nearc4) exper expersq black
    south smsa reg661 reg662 reg663 reg664
    reg665 reg666 reg667 reg668 smsa66, robust
```

- Note: The coefficient on educ is actually larger in 2SLS

Example: Card (1995)

Exercise 1: Solutions

- Download and open the Card dataset (`card.dta`) from http://www.hubner.info/#teaching
  - Run the OLS regression: (Column (2), Table 2 in Card (1995) paper)

```
regress lwage educ exper expersq black south
   smsa reg661 reg662 reg663 reg664 reg665
   reg666 reg667 reg668 smsa66, vce(r)
```

  - Using 2SLS: (first IV estimate in Table 4)

```
ivreg2 lwage (educ=nearc4) exper expersq black
   south smsa reg661 reg662 reg663 reg664
   reg665 reg666 reg667 reg668 smsa66, robust
```

  - Note: The coefficient on educ is actually larger in 2SLS

Example: Card (1995)

Exercise 2:

- We can get the same coefficient on education by doing the 2-stage process explicitly.
- Instead of using the `ivreg2` command, obtain the same coefficients using OLS (hint: regress `educ` on exogenous variables, obtain predicted values of `educ`, and use these values in the second-stage regression).
- Compare the standard errors from the second-stage OLS regression with those from `ivreg2`. Why might they be different?

Example: Card (1995)

Exercise 2: Solutions

- 2SLS is equivalent to the following:
  - Run the first-stage OLS regression:

```
regress educ exper expersq black south smsa
   reg661 reg662 reg663 reg664 reg665 reg666
   reg667 reg668 smsa66 nearc4, vce(r)
```

  - Predict education

```
predict educhat
```

  - Run the second-stage OLS regression:

```
regress lwage educhat exper expersq black
   south smsa reg661 reg662 reg663 reg664
   reg665 reg666 reg667 reg668 smsa66, vce(r)
```

- Note: the coefficient on educ is the same as in 2SLS (from ivreg2)
  - but s.e. are different (above does not take into account the fact that `educhat` is an estimate)
- **Main takeaway:** For correct SE's, use `ivreg2`.

Verifying the required conditions

- Does z_i satisfy the requirements of an instrument?
- We can test the following:
  - Over-identifying restrictions (if # instruments $\geq$ # of endogenous variables): $H_0 : \mathbb{E}[z_i u_i] = 0$
  - Endogeneity/simultaneity bias: $H_0 : \mathbb{E}[x_{1i} u_i] = 0$
  - Rank test: $\operatorname{rk} \mathbb{E}[z_i x_i'] = K$
  - Finite-sample problems:
    - Weak instruments
    - Too many instruments (overfitting)

Verifying the required conditions

- Tests can be conducted using the options of `ivreg2`:

  `ivreg2` depvar (endogenous variables = additional instrumental variables) exogenous variables, options

  - Overidentification test (automatic)
  - Rank test (automatic)
  - Endogeneity/simultaneity (the option `endog`)
  - Weak instruments (the option `first`)

1. Instrument validity

Conceptual review:

- Hansen's test for overidentifying restrictions:

$$H_0 : \mathbb{E}[z_i u_i] = 0$$
$$H_A : \mathbb{E}[z_i u_i] \neq 0$$

- Test statistic:

$$\left(\sum_{i=1}^{n} z_i \hat{u}_i \right)' \sum_{i=1}^{n} \hat{u}_i^2 z_i z_i' \left(\sum_{i=1}^{n} z_i \hat{u}_i \right) \xrightarrow{d} \chi^2[L - K]$$

  - Limit distribution is χ^2 with degrees of freedom equal to the number of overidentifying restrictions
  - This is reported in Stata output as the `Hansen J statistic.`

1. Instrument validity

Exercise 3:

- Run the 2SLS regression from Exercise 2 again, this time using both `nearc4` and `nearc2` as instruments.
- Based on the Hansen J statistic, can you reject the null hypothesis that the instruments are valid?

1. Instrument validity

Exercise 3: Solutions

- Run the 2SLS regression:

```
ivreg2 lwage (educ=nearc4 nearc2) exper expersq black
   south smsa reg661 reg662 reg663 reg664
   reg665 reg666 reg667 reg668 smsa66, robust
```

- Output:

```
------------------------------------------------------------------------------
Hansen J statistic (overidentification test of all instruments):         1.269
                                                 Chi-sq(1) P-val =    0.2600
------------------------------------------------------------------------------
```

  - We cannot reject the null hypothesis that the instruments are valid
  - Here $L = 16$, $K = 15$, so test distribution has 1 d.f.

2. (Non-)Endogeneity of the regressors

- It's possible that the regressors we think are endogenous (x_{1i}) may not actually be endogenous. We can test for that!
- Durbin-Wu-Hausman test for (non-)endogeneity of x_{1i}

$$H_0 : \mathbb{E}[x_{1i} u_i] = 0$$
$$H_A : \mathbb{E}[x_{1i} u_i] \neq 0$$

- This test involves the hypothesis test of $H_0 : \rho = 0$ in the regression:

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + \hat{\epsilon}_i'\rho + u_i,$$

where $\hat{\epsilon}_i$ is the vector of residuals obtained from regressing each endogenous variable (x_{1i}) on all instruments (z_i).

- **Remember:** The null is that the variable(s) are exogenous.

2. (Non-)Endogeneity of the regressors

Exercise 4:

- Run the 2SLS regression from Exercise 3 again, this time including the option `endog` to test for endogeneity of the variable `educ` (Remember the syntax for the option is `endog(`name of endogenous variable`)`).
- Can you reject the null hypothesis that `educ` is exogenous?

2. (Non-)Endogeneity of the regressors

Exercise 4: Solutions

- 2SLS regression with the `endog()` option:

```
ivreg2 lwage (educ=nearc4 nearc2) exper expersq
   black south smsa reg661 reg662 reg663 reg664 reg665
   reg666 reg667 reg668 smsa66, robust endog(educ)
```

- Output:

```
------------------------------------------------------------------------------
Endogeneity test of endogenous regressors:                             2.831
                                                 Chi-sq(1) P-val =    0.0925
Regressors tested: educ
------------------------------------------------------------------------------
```

  - We cannot reject the null hypothesis that educ is exogenous
  - **Remember:** The null is that the variable(s) are exogenous.

3. Rank condition

Conceptual review:

- First stage: with K_1 endogenous regressors,

$$\underset{(K_1 \times 1)}{x_{1i}} = \underset{(K_1 \times L)(L \times 1)}{\Pi} \underset{}{z_i} + \underset{(L \times 1)}{\epsilon_i}$$

- The Rank condition can be equivalently stated as: $\text{rk}\,\Pi = K_1$ (the number of endogenous variables)
- **Kleibergen-Paap rank test:** The null hypothesis is that the model is under-identified

$$H_0 : \text{rk}\,\Pi = K_1 - 1$$
$$H_A : \text{rk}\,\Pi = K_1$$

- This test is implemented in Stata using the option `first`.

3. Rank condition

Exercise 5:

- Rerun the 2SLS regression from Exercise 3, using the option `first` to test for under-identification.
- Is the Rank condition satisfied?

3. Rank condition

Exercise 5: Solutions

- Run the 2SLS regression:

```
ivreg2 lwage (educ=nearc4 nearc2) exper expersq black
    south smsa reg661 reg662 reg663 reg664 reg665
    reg666 reg667 reg668 smsa66, robust first
```

- Output:

```
------------------------------------------------------------------------------
    Underidentification test
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Kleibergen-Paap rk LM statistic          Chi-sq(2)=16.37    P-val=0.0003
------------------------------------------------------------------------------
```

- We reject the null hypothesis of reduced rank (K_1 denotes number of endogeneous regressors) i.e. the rank condition is satisfied.
- **Remember:** The null hypothesis is that the rank condition is NOT satisfied.

4. Weak instruments

- Weak instruments problem = when the additional instruments (z_i) have only a small amount of explanatory power for the endogenous variables $\Rightarrow$ Finite sample bias!
- How can we detecting weak instruments?
  - Rule of thumb: F-test for significance of excluded instruments in first stage > 10
  - Additional conditions necessary with more than one endogenous variable:
    - Problem if only one instrument has explanatory power for all endogenous variables
    - Check using Shea partial correlation
  - These statistics are both saved in `e(first)` when the `ivreg2` command is run.

4. Weak instruments

Exercise 6

- Retrieve the F-statistic and Shea partial correlation from the regression in Exercise 5 (hint: use `matrix list e(first)`).
- Does there appear to be a weak instruments problem?

4. Weak instruments

Exercise 6: Solutions

- Run the 2SLS regression:

```
ivreg2 lwage (educ = nearc4 nearc2) exper expersq
    black south smsa reg661 reg662 reg663 reg664
    reg665 reg666 reg667 reg668 smsa66, robust first
```

  - Simple F-stat and Shea partial correlation saved in matrix
  - type in the command `matrix list e(first)`

- Output:

```
------------------------------------------------------------------------------
             educ
sheapr2   .0052467
    pr2   .0052467
      F  8.3189747
     df          2
   df_r       2993
 pvalue  .00024953       ...
------------------------------------------------------------------------------
```

  - F-stat (`F`) < 10, weak instruments could be a problem, the partial R-squared (`pr2`) and Shea partial correlation (`sheapr2`) are also low.

4. Weak instruments: Stock & Yogo (2005) tests (An aside)

- In the output for `ivreg2`, Stata reports "`Stock-Yogo weak ID test critical values`". What are they and what do those values mean?
- Basically, it's another way to test for weak instruments.
- **Recall:** if instruments are weak, then the IV estimator will be biased; the bias can even be bigger than that of the OLS estimator.
- But how big does the difference between 2SLS and OLS estimates have to be for there to be a weak instruments problem?
  - Stock and Yogo (2005) provide critical values for the F-stat by comparing the bias of the 2SLS and OLS estimators
  - These critical values depend on what relative bias the researcher thinks is acceptable, the number of endogenous variables, and the number of overidentifying restrictions.
  - A lower acceptable bias means that the first-stage F-statistic has to be higher
  - If our F-statistic is smaller than the critical value, then there is a weak instruments problem.

Review

Stata skills covered in this session:

1. How to use the ivreg2 command
2. How to interpret the output from ivreg2
3. Options in ivreg2: robust, endog(), first
4. Testing for instrument validity, non-endogeneity of regressors, the rank condition, and weak instruments

Some references

- Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 'Instrumental variables and GMM: Estimation and testing.' *Stata Journal* 3.1 (2003): 1-31.
- Bound, John, David A. Jaeger, and Regina M. Baker. 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak.' *Journal of the American Statistical Association* 90.430 (1995): 443-450.
- Cameron, A. Colin, and Pravin K. Trivedi. *Microeconometrics Using Stata*. Vol. 5. College Station, TX: Stata Press, 2009.
- Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2010.