Day 2A
Instrumental Variables, Two-stage Least Squares and Generalized Method of Moments

© A. Colin Cameron
Univ. of Calif.- Davis

Frontiers in Econometrics
Bavarian Graduate Program in Economics

.

Based on A. Colin Cameron and Pravin K. Trivedi (2009, 2010), Microeconometrics using Stata (MUS), Stata Press. and A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications (MMA), C.U.P.

March 21-25, 2011

1. Introduction

- Problem: OLS inconsistent in model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ if $\text{Cov}[\mathbf{x}_i, u_i] \neq \mathbf{0}$.
- Solution: Assume there are instruments $\mathbf{z}_i$ satisfying $\text{Cov}[\mathbf{z}_i, u_i] = \mathbf{0}$.
- If $\#$instruments $=$ $\#$regressors
  - instrumental variables (IV) estimator
- If $\#$instruments $>$ $\#$regressors then use
  - two-stage least squares (2SLS)
  - generalized method of moments (GMM).
- Complications
  - test of assumptions (exogeneity, endogeneity)
  - weak instruments

Overview

2. IV, 2SLS and GMM estimators: Definitions

- Model is $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$
  - OLS is inconsistent as $\mathrm{Cov}[\mathbf{x}_i, u_i] \neq \mathbf{0}$.
- Assume there are instruments $\mathbf{z}_i$ such that $\mathrm{Cov}[\mathbf{z}_i, u_i] = \mathbf{0}$.
  - Then $\mathrm{Cov}[\mathbf{z}_i, u_i] = 0 \Rightarrow \mathrm{E}[\mathbf{z}_i u_i] = \mathbf{0}$ given $\mathrm{E}[u_i|\mathbf{z}_i] = 0$.
- We have the population moment condition

$$\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

- Method of moments: solve the corresponding sample moment condition

$$\frac{1}{N}\sum\nolimits_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}.$$

Instrumental variables (IV) estimator

- In just-identified case (# instruments = # regressors)
  - solve k equations in k unknowns $\frac{1}{N}\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$
  - gives the instrumental variables (IV) estimator.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{IV}} &= \left(\textstyle\sum_{i=1}^N \mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(\textstyle\sum_{i=1}^N \mathbf{z}_i'y_i\right) \\ &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}\end{aligned}$$

  - estimate using Stata 10 command `ivregress 2sls`
- Often just one regressor in $\mathbf{x}_i$ is endogenous (i.e. correlated with u_i).
  - Then one variable in $\mathbf{z}_i$ is the instrument for this endogenous regressor.
  - the remaining entries in $\mathbf{z}_i$ are the exogenous variables
  - i.e. exogenous variables are instruments for themselves.

Generalized method of moments estimator

- In over-identified case (# instruments > # regressors)
  - Cannot solve $\frac{1}{N}\sum_i \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}$.
  - Instead generalized method of moments (GMM) estimator minimizes the quadratic form in $\frac{1}{N}\sum_{i=1}^N \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \left[\frac{1}{N}\sum_i \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{z}_i\right]' \times \mathbf{W}_N \times \left[\frac{1}{N}\sum_i \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{z}_i\right] \\ &= (\mathbf{Z}'\mathbf{u})'\mathbf{W}(\mathbf{Z}'\mathbf{u}) \end{aligned}$$

  - The symmetric full-rank weighting matrix $\mathbf{W}$ does not depend on $\boldsymbol{\beta}$.
- Then $\partial Q(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$ yields the GMM estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{GMM}} &= \left(\sum_i \mathbf{x}_i\mathbf{z}_i' \times \mathbf{W}_N \times \sum_{i=1}^N \mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\sum_i \mathbf{x}_i\mathbf{z}_i' \times \mathbf{W}_N \times \sum_{i=1}^N \mathbf{z}_i y_i\right) \\ &= (\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_N\mathbf{Z}'\mathbf{y}. \end{aligned}$$

Optimal GMM and 2SLS

- The variance of $\widehat{\beta}_{\text{GMM}}$ is smallest when the optimal weighting matrix $\mathbf{W}_N$ is consistent for $(\text{Var}[\mathbf{Z}'\mathbf{u}])^{-1}$
  - Though in the just-identified $(r = K)$ GMM = IV for any $\mathbf{W}_N$.
- For homoskedastic errors $\text{Var}[\mathbf{Z}'\mathbf{u}] = \sigma^2 \sum_{i=1}^N \mathbf{z}_i'\mathbf{z}_i$
  - Two-stage least squares (2SLS) estimator sets $\mathbf{W}_N = (\sum_{i=1}^N \mathbf{z}_i'\mathbf{z}_i)^{-1}$
  - Yields $\widehat{\beta}_{2\text{SLS}} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} \times \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$
  - Estimate using Stata 10 command `ivregress 2sls`
  - but use robust VCE to guard against errors not homoskedastic.
- For heteroskedastic errors $\text{Var}[\mathbf{Z}'\mathbf{u}] = \sigma^2 \sum_{i=1}^N \mathbf{z}_i'\mathbf{z}_i$
  - "Optimal" GMM estimator if errors are heteroskedastic errors sets

$$\mathbf{W}_N = (\textstyle\sum_{i=1}^N \widehat{u}_i^2 \mathbf{z}_i'\mathbf{z}_i)^{-1}, \widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}_{2\text{SLS}}$$

  - estimate using Stata 10 command `ivregress gmm`.

More on 2SLS

- 2SLS gets it's name because it can be computed in two-stages.
- Suppose y_1 depends in part on scalar y_2 which is endogenous

$$\begin{array}{ll} \text{Structural equation for } y_1 & y_{1i} = \beta_1 y_{2i} + \mathbf{x}_{1i}'\boldsymbol{\beta}_2 + u_i \\ \text{First-stage equation for } y_2 & y_{2i} = \mathbf{x}_{1i}'\boldsymbol{\pi}_1 + \mathbf{x}_{2i}'\boldsymbol{\pi}_2 + v_i \end{array}$$

  - where $\mathbf{x}_2$ is one or more instruments for y_2
  - in earlier notation $\mathbf{x}_i = (y_{2i}\ \mathbf{x}_{1i}')'$ and $\mathbf{z}_i = (\mathbf{x}_{1i}'\ \mathbf{x}_{2i}')'$.
- OLS of y_1 on y_2 and $\mathbf{x}_1$ is inconsistent.
- 2SLS can be computed as follows
  - 1. First-stage: $\widehat{y}_2$ as prediction from OLS of y_2 on $\mathbf{x}_1$ and $\mathbf{x}_2$.
  - 2. Structural: Do OLS of y_2 on $\widehat{y}_2$ and $\mathbf{x}_2$.
- But this method does not generalize to nonlinear models.

3. Data Example: Drug expenditures

- Example from MUS chapter 6.
- Drug expenditures for U.S. elderly (`ldrugexp`) regressed on
  - endogenous private health insurance dummy (`hi_empunion`) and
  - exogenous regressors defined by global `x2list`.

```
. * Read data, define global x2list (exogenous regressors), and summarize d
. use mus06data.dta

. global x2list totchr age female blhisp linc

. keep if linc != .
(302 observations deleted)

. describe ldrugexp hi_empunion $x2list

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------
ldrugexp        float  %9.0g                  log(drugexp)
hi_empunion     byte   %8.0g                  Insured thro emp/union
totchr          byte   %8.0g                  Total chronic cond
age             byte   %8.0g                  Age
female          byte   %8.0g                  Female
blhisp          float  %9.0g                  Black or Hispanic
linc            float  %9.0g                  log(income)
```

- Summary statistics

```
. summarize ldrugexp hi_empunion $x2list

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    ldrugexp |     10089    6.481361    1.362052          0   10.18017
  hi_empunion|     10089    .3821984    .4859488          0          1
      totchr |     10089    1.860938    1.292858          0          9
         age |     10089    75.05174    6.682109         65         91
      female |     10089    .5770641    .4940499          0          1
-------------+--------------------------------------------------------
      blhisp |     10089    .1635445      .36988          0          1
        linc |     10089    2.743275    .9131433  -6.907755   5.744476
```

- Sample is 65+.
  38% have employer or union-sponsored health insurance.

OLS estimates

- OLS is inconsistent if `hi_empunion` endogenous

```
. * OLS
. regress ldrugexp hi_empunion $x2list, vce(robust)

Linear regression                                      Number of obs =   10089
                                                       F(  6, 10082) =  376.85
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1770
                                                       Root MSE      =   1.236

------------------------------------------------------------------------------
             |               Robust
    ldrugexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 hi_empunion |   .0738788   .0259848     2.84   0.004     .0229435    .1248141
      totchr |   .4403807   .0093633    47.03   0.000     .4220268    .4587346
         age |  -.0035295    .001937    -1.82   0.068    -.0073264    .0002675
      female |   .0578055   .0253651     2.28   0.023     .0080848    .1075262
      blhisp |  -.1513068   .0341264    -4.43   0.000    -.2182013   -.0844122
        linc |   .0104815   .0137126     0.76   0.445    -.0163979     .037361
       _cons |   5.861131   .1571037    37.31   0.000     5.553176    6.169085
------------------------------------------------------------------------------
```

- Drug expenditure increases by 7.4% if have private insurance.

Instruments

- A valid instrument for private health insurance (`hi_empunion`) must
  - not be directly a cause of `ldrugexp` (so uncorrelated with u_i)
  - i.e. must not belong in the model for `ldrugexp`
  - and to be relevant should be correlated with `hi_empunion`
- Possible instrument 1
  - `ssiratio` = social security income ÷ income from all other sources
  - need to assume that the direct role of income is adequately captured by the regressor `linc`
- Possible instrument 2
  - `multlc` = 1 if firm has multiple locations
  - need to assume that firm size does not effect `ldrugexp`

- Two possible instruments `ssiratio` and `multlc`

```
. * Two available instruments for hi_empunion
. describe ssiratio multlc

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
ssiratio        float  %9.0g                  SSI/Income ratio
multlc          byte   %8.0g                  Multiple locations

. summarize ssiratio multlc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    ssiratio |     10089    .5365438    .3678175          0    9.25062
      multlc |     10089    .0620478    .2412543          0          1

. correlate hi_empunion ssiratio multlc
(obs=10089)

             | hi_emp~n ssiratio   multlc
-------------+---------------------------
 hi_empunion |   1.0000
    ssiratio |  -0.2124   1.0000
      multlc |   0.1198  -0.1904   1.0000
```

- Correlation between z and x is low
  - e.g. $\text{Cor}[z, x] = -0.21$ for `ssiratio`

IV estimates

- IV estimates using the single instrument `ssiratio` for `hi_empunion`

```
. * IV estimator with ssiratio as single instrument for hi_empunion
. ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)

Instrumental variables (2SLS) regression               Number of obs =    10089
                                                       Wald chi2(6)  = 2000.86
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.0640
                                                       Root MSE      =  1.3177

------------------------------------------------------------------------------
             |               Robust
    ldrugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 hi_empunion |  -.8975913   .2211268    -4.06   0.000    -1.330992   -.4641908
      totchr |   .4502655   .0101969    44.16   0.000       .43028     .470251
         age |  -.0132176   .0029977    -4.41   0.000    -.0190931   -.0073421
      female |   -.020406   .0326114    -0.63   0.531    -.0843232    .0435113
      blhisp |  -.2174244   .0394944    -5.51   0.000     -.294832   -.1400167
        linc |   .0870018   .0226356     3.84   0.000     .0426368    .1313668
       _cons |    6.78717   .2688453    25.25   0.000     6.260243    7.314097
------------------------------------------------------------------------------
Instrumented:  hi_empunion
Instruments:   totchr age female blhisp linc ssiratio
```

- Coefficient even changes sign, from 0.074 (OLS) to -0.898 (IV). Standard error increases from 0.026 (OLS) to 0.221 (IV).

2SLS Estimates

- Overidentified as two instruments `ssiratio` and `multlc`

```
. * 2SLS estimator with ssiratio and multlc as instruments for hi_empunion
. ivregress 2sls ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)

Instrumental variables (2SLS) regression               Number of obs =    10089
                                                       Wald chi2(6)  = 1955.36
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.0414
                                                       Root MSE      =  1.3335

------------------------------------------------------------------------------
             |               Robust
    ldrugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 hi_empunion |  -.9899269   .2045907    -4.84   0.000    -1.390917   -.5889365
      totchr |   .4512051   .0103088    43.77   0.000     .4310001      .47141
         age |  -.0141384      .0029    -4.88   0.000    -.0198223   -.0084546
      female |  -.0278398   .0321743    -0.87   0.387    -.0909002    .0352207
      blhisp |  -.2237087   .0395848    -5.65   0.000    -.3012934   -.1461239
        linc |   .0942748   .0218841     4.31   0.000     .0513828    .1371668
       _cons |   6.875188   .2578855    26.66   0.000     6.369741    7.380634
------------------------------------------------------------------------------
Instrumented:  hi_empunion
Instruments:   totchr age female blhisp linc ssiratio multlc
```

- Coefficient changes from -0.898 (IV) to -0.990 (2SLS). Standard error decreases from 0.221 (IV) to 0.205 (2SLS).

Optimal GMM

- Two instruments `ssiratio` and `multlc`
  - optimal GMM if errors are heteroskedastic and start with $\mathrm{E}[\mathbf{z}u] = \mathbf{0}$.

```
. * GMM estimator with ssiratio and multlc as instruments for hi_empunion
. ivregress gmm ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)

Instrumental variables (GMM) regression               Number of obs =    10089
                                                      Wald chi2(6)  =  1952.65
                                                      Prob > chi2   =   0.0000
                                                      R-squared     =   0.0406
GMM weight matrix: Robust                             Root MSE      =   1.3341

------------------------------------------------------------------------------
             |               Robust
    ldrugexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 hi_empunion |  -.9932795   .2046731    -4.85   0.000    -1.394431   -.5921275
     totchr  |   .4509508   .0103104    43.74   0.000     .4307428    .4711588
        age  |  -.0141509   .0029014    -4.88   0.000    -.0198375   -.0084644
     female  |  -.0281716   .0321881    -0.88   0.381    -.0912592     .034916
     blhisp  |  -.2231048   .0395972    -5.63   0.000    -.3007139   -.1454957
       linc  |   .0944632   .0218959     4.31   0.000     .0515481    .1373783
      _cons  |   6.877821   .2579974    26.66   0.000     6.372155    7.383486
------------------------------------------------------------------------------
Instrumented:  hi_empunion
Instruments:   totchr age female blhisp linc ssiratio multlc
```

- Estimate and standard error for `hi_empunion` are very similar to 2SLS
  - Little efficiency gain compared to 2SLS.

Estimator comparison

- Compare OLS, IV, 2SLS (over-identified), GMM (over-identified)

```
. * Compare estimators
. quietly regress ldrugexp hi_empunion $x2list, vce(robust)

. estimates store OLS

. quietly ivregress 2sls ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)

. estimates store IV

. quietly ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)

. estimates store TWOSLS

. quietly ivregress gmm ldrugexp (hi_empunion = ssiratio multlc) $x2list, vce(robust)

. estimates store GMM

. estimates table OLS IV TWOSLS GMM, b(%9.4f) se(%9.3f) stats(N r2 F)
```

Variable	OLS	IV	TWOSLS	GMM
hi_empunion	0.0739	-0.9899	-0.8976	-0.9933
	0.026	0.205	0.221	0.205
totchr	0.4404	0.4512	0.4503	0.4510
	0.009	0.010	0.010	0.010
age	-0.0035	-0.0141	-0.0132	-0.0142
	0.002	0.003	0.003	0.003
female	0.0578	-0.0278	-0.0204	-0.0282
	0.025	0.032	0.033	0.032
blhisp	-0.1513	-0.2237	-0.2174	-0.2231
	0.034	0.040	0.039	0.040
linc	0.0105	0.0943	0.0870	0.0945
	0.014	0.022	0.023	0.022
_cons	5.8611	6.8752	6.7872	6.8778
	0.157	0.258	0.269	0.258
N	10089	10089	10089	10089
r2	0.1770	0.0414	0.0640	0.0406
F	376.8458			

legend: b/se

4. Instrumental variables methods in practice

- Do we need to use instruments?
  - Hausman test of endogeneity.
- Is the instrument valid (uncorrelated with the error)?
  - If model is over-identified can do over-identifying restrictions test.
- What if the instrument is weakly correlated with regressor instrumented
  - Lose efficiency
  - If really weak can have finite-sample bias and wrong test size.
- How many instruments?
  - Need # instruments $\geq$ # endogenous regressors.
  - In theory more is better but too many can have finite-sample bias.

Hausman test

- In general a Hausman test considers two different estimators $\widehat{\boldsymbol{\theta}}$ and$\widehat{\boldsymbol{\theta}}$ that have the same plim under H_0 and different $\mathrm{plim}'s$ under H_a.
  - $H_0 : \mathrm{plim}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) = \mathbf{0}$ versus $H_a : \mathrm{plim}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \neq \mathbf{0}$.
- We reject H_0 if the difference is large, using

$$\mathsf{H} = (\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}})'(\widehat{\mathsf{V}}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}])^{-1}(\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2(q).$$

- Tricky bit is estimating $\mathsf{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}] = \mathsf{V}[\widehat{\boldsymbol{\theta}}] + \mathsf{V}[\widetilde{\boldsymbol{\theta}}] - 2 \times \mathsf{Cov}[\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}]$
  - usual Hausman test implementation assumes one of $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ is fully efficient under the null. Say $\widetilde{\boldsymbol{\theta}}$: then $\mathsf{V}[\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}] = \mathsf{V}[\widehat{\boldsymbol{\theta}}] - \mathsf{V}[\widetilde{\boldsymbol{\theta}}]$
  - such efficiency is not usually the case in practice
    - ★ e.g. if errors are heteroskedastic then OLS is inefficient
  - instead need to use a robust form of the Hausman test.

- Hausman test of endogeneity: 2SLS ($\widehat{\boldsymbol{\theta}}$) versus OLS ($\widehat{\boldsymbol{\theta}}$)
  - $H_0 : \text{plim}(\widehat{\boldsymbol{\theta}}_{2\text{SLS}} - \widetilde{\boldsymbol{\theta}}_{\text{OLS}}) = \mathbf{0}$ if exogeneity
    vs. $H_a : \text{plim}(\widehat{\boldsymbol{\theta}}_{2\text{SLS}} - \widetilde{\boldsymbol{\theta}}_{\text{OLS}}) \neq \mathbf{0}$ if endogeneity
- Use heteroskedasticity-robust version of Hausman test
  - this is command `estat endogenous` and not `hausman`

```
. * Robust version of Hausman test using augmented regression
. quietly ivregress 2sls ldrugexp (hi_empunion = ssiratio) $x2list, vce(robust)

. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Robust score chi2(1)            =   24.935  (p = 0.0000)
  Robust regression F(1,10081)    =  26.4333  (p = 0.0000)
```

- Reject H_0 as $p = 0.000$.
  Conclude that `hi_empunion` is endogenous. Need to do IV.

Test of instrument validity

- Cannot test validity in a just identified model
  - Intuition: Test based on $\mathrm{Cov}[\mathbf{z}_i, \widehat{u}_i] \simeq 0$ requires $\widehat{u}_i$ based on a consistent estimator of $\boldsymbol{\beta}$ which requires at least just-identified model.
- Test of overidentifying restrictions (for over-identified model)
  - Test $H_0 : \mathrm{E}[\mathbf{z}_i' u_i] = \mathbf{0}$ by testing if $N^{-1} \sum_i \mathbf{z}_i' \widehat{u}_i \simeq \mathbf{0}$.
  - Limited test as assumes instruments in just-identified model are valid.
- In Stata command `estat overid` after command `ivregress gmm`.
  - Here one over-identifying restriction (2 instruments for 1 endogenous)

```
. * Test of overidentifying restrictions following ivregress gmm
. quietly ivregress gmm ldrugexp (hi_empunion = ssiratio multlc) $x2list, wmatrix(robust)

. estat overid

  Test of overidentifying restriction:

  Hansen's J chi2(1) = 1.04754 (p = 0.3061)
```

- Do not reject H_0 as $p = 0.31 < 0.05$.
  Conclude that, assuming the just-identifying restriction is valid, then the over-identifying restriction is also valid.

Weak instruments

- Weak instrument means that instrument(s) are weakly correlated with endogenous regressor(s), after controlling for exogenous regressors.
- Then
  - 1. standard errors $\uparrow$ greatly as 2SLS much less efficient than OLS.
  - 2. even slight correlation between error and the instrument can lead to 2SLS more inconsistent than OLS.
  - 3. even if instrument(s) are valid so 2SLS is inconsistent, in typical sample sizes usual asymptotic theory can provide a poor approximation e.g. bias.
- Consequences
  - 1. key coefficient estimate(s) can become statistically insignificant.
  - 2. even more important to ensure that the instrument is valid.
  - 3. focus of the weak instrument literature.
- In Stata for 3. use
  - command `estat firststage` after command `ivregress`
  - add-on commands `condivreg` and `ivreg2`

5. IV estimator properties: consistency

- Stacking all observations

$$\widehat{\beta}_{\mathrm{IV}} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{y}.$$

- Substitute $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ for $\mathbf{y}$ yields

$$\begin{aligned}\widehat{\beta}_{\mathrm{IV}} &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'[\mathbf{X}\beta + \mathbf{u}] \\ &= \beta + \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{u} \\ &= \beta + \left(\tfrac{1}{N}\mathbf{Z}'\mathbf{X}\right)^{-1}\tfrac{1}{N}\mathbf{Z}'\mathbf{u}\end{aligned}$$

- So $\widehat{\beta}_{\mathrm{IV}} \xrightarrow{p} \beta$ and $\widehat{\beta}_{\mathrm{IV}}$ is consistent for β if
  - $\operatorname{plim} \frac{1}{N}\mathbf{Z}'\mathbf{u} = \mathbf{0}$ (instruments are valid) and
  - $\operatorname{plim} \frac{1}{N}\mathbf{Z}'\mathbf{X} \neq \mathbf{0}$ (instruments are relevant).

IV estimator: asymptotic distribution

- Informal derivation:

$$\begin{aligned}\widehat{\beta}_{\text{GMM}} - \beta &= (\mathbf{Z}'\mathbf{X})^{-1} \times \mathbf{Z}'\mathbf{u} \\ &\overset{a}{\sim} (\mathbf{Z}'\mathbf{X})^{-1} \times \mathcal{N}[\mathbf{0},\ \mathsf{V}[\mathbf{Z}'\mathbf{u}]] \\ &\overset{a}{\sim} (\mathbf{Z}'\mathbf{X})^{-1} \times \mathcal{N}[\mathbf{0},\ \mathbf{Z}'\mathsf{V}[\mathbf{u}|\mathbf{Z}]\mathbf{Z}] \\ &\overset{a}{\sim} (\mathbf{Z}'\mathbf{X})^{-1} \times \mathcal{N}[\mathbf{0},\ \mathbf{Z}'\Omega\mathbf{Z}]\end{aligned}$$

- Thus

$$\widehat{\beta}_{\text{IV}} \overset{a}{\sim} \mathcal{N}[\beta,\ (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\Omega\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}]; \quad \Omega = \ \mathsf{V}[\mathbf{u}|\mathbf{Z}].$$

- With independent heteroskedastic errors (Stata option `vce(robust)`)

$$\widehat{\mathsf{V}}[\widehat{\beta}_{\text{IV}}] = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\widehat{\Omega}\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}; \quad \widehat{\Omega} = \ \text{Diag}[\widehat{u}_i^2].$$

- Note: $\text{Cor}[\mathbf{Z}, \mathbf{X}] \Rightarrow \mathbf{Z}'\mathbf{X}$ small $\Rightarrow (\mathbf{Z}'\mathbf{X})^{-1}$ large $\Rightarrow \widehat{\beta}_{\text{IV}}$ is imprecise.

Asymptotic Distribution of GMM

- Informal derivation:

$$
\begin{aligned}
\widehat{\beta}_{\text{GMM}} &= (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'(\mathbf{X}\beta+\mathbf{u}) \\
\widehat{\beta}_{\text{GMM}}-\beta &= (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{u} \\
&\overset{a}{\sim} (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\times\mathcal{N}[\mathbf{0},\ \mathrm{V}[\mathbf{Z}'\mathbf{u}]] \\
&\overset{a}{\sim} (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\times\mathcal{N}[\mathbf{0},\ \mathbf{Z}'\mathrm{V}[\mathbf{u}|\mathbf{Z}]\mathbf{Z}] \\
&\overset{a}{\sim} (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\times\mathcal{N}[\mathbf{0},\ \mathbf{Z}'\Omega\mathbf{Z}]
\end{aligned}
$$

- Thus

$$
\begin{aligned}
\widehat{\beta}_{\text{GMM}} &\overset{a}{\sim} \mathcal{N}[\beta,\ (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\Omega\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}] \\
\Omega &= \mathrm{V}[\mathbf{u}|\mathbf{Z}].
\end{aligned}
$$

- Optimal **W** is a consistent estimate of Ω^{-1}. Then

$$
\widehat{\beta}_{\text{OptGMM}} \overset{a}{\sim} \mathcal{N}[\beta,\ (\mathbf{X}'\mathbf{Z}\Omega^{-1}\mathbf{Z}'\mathbf{X})^{-1}]
$$

6. Nonlinear GMM estimator: Definition

- Nests LS, MLE, IV, GMM, ... The way to view estimation.
- Population unconditional moment condition

$$\mathrm{E}[\mathbf{h}(\mathbf{w},\boldsymbol{\theta}_0)] = \mathbf{0}; \quad \mathbf{w} = (\mathbf{y},\mathbf{x},\mathbf{z}) \text{ is all observables.}$$

- $\widehat{\boldsymbol{\theta}}$ solves the corresponding sample moment condition

$$\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\widehat{\boldsymbol{\theta}}) = \mathbf{0}.$$

  - just-identified case $(r = q)$ can solve for $\boldsymbol{\beta}$
  - over-identified case $(r > q)$ cannot as r equations in k unknowns.
- The generalized method of moments (GMM) estimator (for $r > q$) minimizes the quadratic form in $N^{-1}\sum_i \mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})$

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right]'\mathbf{W}_N\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i,\boldsymbol{\theta})\right] \\ &= \mathbf{g}(\boldsymbol{\theta})'\mathbf{W}_N\mathbf{g}(\boldsymbol{\theta}) \end{aligned}$$

  - where $\mathbf{g}(\boldsymbol{\theta}) = \sum_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\theta})$ and $\frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_N$ is a symmetric full-rank weighting matrix that does not depend on $\boldsymbol{\theta}$.

Nonlinear GMM estimator: Properties

- $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ is asymptotically normally distributed with

$$\mathrm{V}[\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}] = N(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}.$$

  where

$$\begin{aligned}\mathbf{G} &= \lim \ \mathrm{E}\left[\frac{\partial \mathbf{g}_N(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\right] = \lim \ \mathrm{E}\left[\frac{1}{N}\sum\nolimits_{i=1}^N \frac{\partial \mathbf{h}_i(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\right] \\ \mathbf{S} &= \mathrm{Var}[\sqrt{N}\mathbf{g}_N(\boldsymbol{\theta})] = \mathrm{Var}\left[\frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^N \mathbf{h}(\mathbf{w}_i, \boldsymbol{\theta})\right].\end{aligned}$$

- Optimal GMM: $\mathbf{W}_N = \widehat{\mathbf{S}}^{-1}$ where $\widehat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$.
- Similar issues as for weighted LS in the linear model.
  - Model choice: specify moment conditions for estimation.
  - Estimator choice: specify a weighting function.
  - Statistical inference: use robust standard errors.
  - Most efficient estimator: a particular choice of weighting function.
  - In Stata 11 use the new command `gmm`.

7. Endogeneity in nonlinear models

- Example is $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) + u_i$ and $\mathrm{Cov}[\mathbf{x}_i, u_i] \neq 0$.
- Several very different methods (and associated models) exist.
- 1. Nonlinear IV (often called nonlinear 2SLS) is nonlinear GMM based on $\mathrm{E}[\mathbf{z}_i u_i] = \mathbf{0}$ and $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$.
- 2. Control function: add estimated first-stage error $\widehat{v}_i$ as regressor.
  - differs from 1. in nonlinear models
- 3. Fully structural approach adds an equation for endogenous regressors and estimates the model
  - Differs from 1. and 2. in most nonlinear models and is computationally difficult.
- 4. The following is inconsistent in nonlinear models: get $\widehat{\mathbf{x}}_i$ from first stage regressions and estimate $y_i = \exp(\widehat{\mathbf{x}}_i'\boldsymbol{\beta}) + error$.
  - The two-stage LS interpretation of 2SLS does not carry over to nonlinear models.

8. Stata commands

IV (just-identified)	`ivregress 2sls`
2SLS (over-identified)	`ivregress 2sls`
GMM (over-identified)	`ivregress gmm`
Overidentifying restrictions test	`xtoverid`
Hausman test (if i.i.d. error)	`hausman`
Hausman test (if heteroskedastic error)	`estat endogenous`
Weak instruments	`estat firststage`
(plus user written commands)	`condivreg; ivreg2`
Static panel IV	`xtivreg; xthaustaylor`
Dynamic panel IV	`xtabond; xtdpdsys; xtdpd`
Nonlinear GMM (new in Stata 11)	`gmm`

9. Appendix: Instrumental variables Intuition

- Simplify to scalar regression of y on single regressor x (no intercept).
- Linear regression model
  - $y = \beta x + u$ where u is an error term.
- In general
  - $\mathrm{E}[y|x] = \beta x + \mathrm{E}[u|x]$.
- Standard regression:
  - assume $\mathrm{E}[u|x] = 0$ i.e. regressors uncorrelated with error
  - implies the following path analysis diagram

$$\begin{array}{ccc} x & \longrightarrow & y \\ & \nearrow & \\ u & & \end{array}$$

where there is no association between x and u.

- But there may be an association between regressors and errors.
- Example: regression of earnings (y) on years of schooling (x).
- The error u embodies all factors other than schooling that determine earnings, such as ability.
- Suppose a person has high u, due to high (unobserved) ability.
  - This increases earnings, since $y = \beta x + u$.
  - But it may also increase x, since schooling is likely to be higher for those with high ability.
- So high u
  - (1) directly increases y and
  - (2) indirectly increases y via higher x.

- The path analysis diagram becomes

$$
\begin{array}{ccc}
x & \longrightarrow & y \\
\uparrow & \nearrow & \\
u & &
\end{array}
$$

where now there is an association between x and u.

- Then $y = \beta x + u(x)$ implies

$$\frac{dy}{dx} = \beta + \frac{du}{dx}.$$

- OLS is inconsistent for β as it measures dy/dx, not just β.

- Assume there exists an **instrument** z that has the properties
  - changes in z do not directly lead to changes in y
  - changes in z are associated with changes in x
- The path analysis diagram becomes

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$$

- Note: z does not directly cause y, though z and y are correlated via indirect path of z being correlated with x which in turn determines y.
- Formally, z is an instrument for regressor x if
  - (1) z is uncorrelated with the error u; and
  - (2) z is correlated with the regressor x.

- Example: a one unit change in the instrument z is associated with
  - 0.2 more years of schooling (x) and
  - \$500 increase in annual earnings (y) (due to $z \uparrow \Rightarrow x \uparrow \Rightarrow y \uparrow$.)
- Then 0.2 years extra schooling is associated with \$500 extra earnings.
  - So a one year increase in schooling is associated with a $\$500/0.2 = \$2,500$ increase in earnings.
- The causal estimate of β is therefore 2500.

- Mathematically we estimated changes dx/dz and dy/dz and calculated the causal estimator as

$$\beta_{\text{IV}} = \frac{dy/dz}{dx/dz}.$$

  - dy/dz estimated by OLS of y on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$
  - dx/dz estimated by OLS of x on z with slope estimate $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}$.

- The IV estimator is

$$\begin{aligned}
\widehat{\beta}_{\text{IV}} &= \frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}}{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}} \\
&= (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} \\
&= \left(\textstyle\sum_{i=1}^{N} z_i x_i\right)^{-1} \textstyle\sum_{i=1}^{N} z_i y_i.
\end{aligned}$$