

# Endogeneity & IV Estimation

Part I

Read (Wooldridge Ch5, Ch6.1-6.2  
Greene ch10,13)

Part II

Read (Baum, Schaffer & Stillman (Boston WP 667)  
Cameron & Trivedi (2009), Microeconometrics  
Using Stata

Stata: ivreg2, xtivreg2 & ivregress  
(help messages)

(Revised 2010)

EC 671, EC 670

Lee

# Part I

## IV Estimation

- 4 sources
- requirements for IV, Examples
- 2SLS
- Testing for endogeneity
- Weak IV
- Over-identifying restrictions
- pitfalls of IV
- More issues

# IV Estimation

$$y_1 = x_1 \beta_1 + y_2 \beta_2 + e$$

$$\text{Cov}(y_2, e) \neq 0$$

Here,  $x_1$  is exogenous:  $\text{Cov}(x_1, e) = 0$

$y_2$  is endogenous:  $\text{Cov}(y_2, e) \neq 0$

when is this problem occurring?

i) measurement error in  $y_2$

(i.m.e. in  $y_1$  does not pose a problem of inconsistency.)

$y_2^*$  is a true measure:  $y_1 = x_1 \beta_1 + y_2^* \beta_2 + u$

$y_2 = y_2^* + \varepsilon$  is measured with error ( $\varepsilon$ ).

("errors in variables")

Then

$e$  includes  $-\beta_2 \varepsilon$ .

$$y_1 = x_1 \beta_1 + (y_2 - \varepsilon) \beta_2 + u = x_1 \beta_1 + y_2 \beta_2 + \underbrace{(-\varepsilon \beta_2 + u)}_e$$

Thus  $\text{Cov}(y_2, e) \neq 0$

(next page)  $\rightarrow$

$$\text{plim } \hat{\beta}_2 = \beta_2 \cdot \frac{\sigma_{y_2^*}^2}{\sigma_{y_2^*}^2 + \sigma_\varepsilon^2} < \beta_2 \quad \text{attenuation bias}$$

ii) Omitted variables (observed or unobserved)

Important variables are omitted, and thus they are included in the error term

$$e = W\gamma + u, \quad \text{Cov}(y_2, e) \neq 0 \text{ as long as}$$

$$(\text{or } e = g\gamma + u)$$

$$\text{Cov}(y_2, W) \neq 0$$

$\Rightarrow \hat{\beta}_2$  is biased  $\Rightarrow \text{bias} = \gamma \cdot c$  where  $c$  is the coef of regression of  $y_2$  on  $W$  (or  $g$ )  
(see earlier note)

Note Measurement error

(true model)  $y_1 = y_2^* \beta_2 + u$  (other variables are compressed for simplicity)

we use  $y_2 = y_2^* + \varepsilon$  measured with error

$$y_1 = y_2^* \beta_2 + u = (y_2 - \varepsilon) \beta_2 + u$$

$$\Rightarrow y_1 = y_2 \beta_2 + (u - \beta_2 \varepsilon) = y_2 \beta_2 + e$$

$$\text{with } e = u - \beta_2 \varepsilon$$

$$\hat{\beta}_2 = \frac{\sum y_{1i} y_{2i}}{\sum y_{2i}^2} = \frac{\sum y_{2i} (y_{2i} \beta_2 + u_i - \beta_2 \varepsilon_i)}{\sum y_{2i}^2}$$

Note

"A single variable measured with errors affects ALL coefficients."

$$= \beta_2 + \frac{\sum y_{2i} (u_i - \beta_2 \varepsilon_i)}{\sum (y_{2i}^* + \varepsilon_i)^2}$$

Assume  $y_{2i}^*, \varepsilon_i, u_i$  are independent  $\left\{ \begin{array}{l} \sum \varepsilon_i u_i = 0 \\ \sum y_{2i}^* u_i = 0 \\ \sum y_{2i}^* \varepsilon_i = 0 \end{array} \right.$

$$\text{numerator} = \sum (y_{2i}^* + \varepsilon_i) (u_i - \beta_2 \varepsilon_i)$$

$$= -\beta_2 \sigma_\varepsilon^2$$

$$\text{denominator} = \sum (y_{2i}^* + \varepsilon_i)^2 = \sigma_{y_2^*}^2 + \sigma_\varepsilon^2 + 2 \cdot 0$$

"attenuation bias"

(less serious problem)

then

$$\hat{\beta}_2 = \beta_2 + \frac{-\beta_2 \sigma_\varepsilon^2}{\sigma_{y_2^*}^2 + \sigma_\varepsilon^2} = \beta_2 \frac{\sigma_{y_2^*}^2}{\sigma_{y_2^*}^2 + \sigma_\varepsilon^2} < \beta_2$$

Exercise Suppose  $y_1$  is also measured with errors

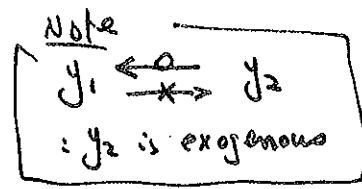
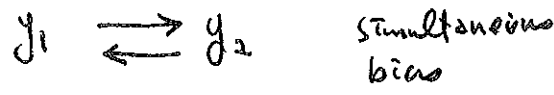
(H/W) true model  $\Rightarrow y_1^* = y_2^* \beta_2 + u$ . Assume  $v, \varepsilon, y_2^*, u$  are independent

we observe  $y_1 = y_1^* + v$ ,  $y_2 = y_2^* + \varepsilon$

find the bias of  $\hat{\beta}_2$  of the regression  $y_1 = y_2 \beta_2 + \text{error}$ .

!!!) Simultaneity

Often it is argued that  $y_2$  determines  $y_1$ , and  $y_1$  also determines  $y_2$ . Thus causality runs in both directions.



It is better to argue that

- " $y_1$  and  $y_2$  are simultaneously determined"
- "Both are related to omitted variables."

Eg1) City Crime Rate = ... +  $\beta_2$  (size of police force) + ...

( $y_1$ ) ( $y_2$ )

.. Both are related to "murder rates." they are simultaneously determined.

Eg2) Wage = .. +  $\beta_2$  Edu + ... + e

( $y_1$ ) ( $y_2$ )

.. Both are affected by family backgrounds. then we say

$Cov(Edu, e) \neq 0$

Solutions : Find IVs (instrumental variables), say,  $(z_2)$  for  $y_2$ , IV for  $x_1 = x_1$  itself.

Note  $(z_2)$  is exogenous variables in the system but it is not included in the equation to estimate. (exclusion restriction)

$\star$

$y_1 = x_1 \beta_1 + \beta_2 y_2 + \dots$

$y_2 = \beta_1 x_2 + y_1 \beta_2 + \dots$

$\rightarrow z_2$

# "Exclusion Restriction"

Examples: which eq is identified? which are IVs?

Ex1) A)  $Inf = a_0 + a_1 Open + a_2 \log Income + b_3 \log population + u$   
 B)  $Open = b_0 + b_1 Inf + e$

Ex2) A)  $Inf = a_0 + a_1 Open + a_2 \log MS + u$   
 B)  $Open = b_0 + b_1 Inf + b_2 \log Income + b_3 \log land + e$

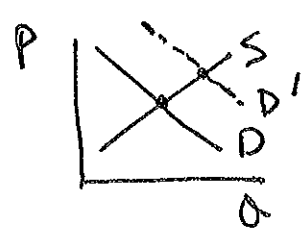
Ex3) A)  $Inf = a_0 + a_1 Open + a_2 \log pop + u$   
 B)  $Open = b_0 + b_1 Inf + b_2 \log pop + b_3 \log land + e$

Note: we don't necessarily try to estimate two equations jointly, unless we need a joint estimation (system of equations)

i) Joint estimation leads to more efficient estimators but it requires a fully correctly specified equations system. If any one equation is mis-specified, all estimators in other equations will be biased.

iii) the single equation based 2SLS needs the information of IVs in other equations, say  $z_2$ .

## Identification



$$Q^d = a + bP + cY + e_1$$

$$Q^s = a + \beta P + e_2$$

which is identified? D or S?

# Requirements for GOOD IVs

①  $Cov(z_2, e) = 0$

Think about the exclusion restriction!

$E(z_2 | e) = 0$

$\Rightarrow z'e = 0 \Rightarrow E[(y_1 - x_1\beta_1 - y_2\beta_2) z_2] = 0$

②  $z_2$  is sufficiently related to  $y_2$ .  $\Rightarrow$  partial correlation

$Rank(z'X) = k$  where  $X = (x_1, y_2)$

$z = (x_1, z_2)$

Eg) IV for Edu = "Parents' Edu"

well?!

①  $Cov(\text{Parents' edu}, e) = 0$

... Wage and Parent's edu are not jointly determined

② Also, Parent's edu is highly correlated to Edu.

$\Rightarrow$  Parent's edu is not a determinant for children's wage but it is highly correlated with children's edu.

## IV Estimation (2SLS)

$y_1 = x_1\beta_1 + y_2\beta_2 + e$

1st Run  $y_2$  on  $z = (x_1, z_2)$  and obtain  $\hat{y}_2$

Note: Do not forget to include  $x_1$  as regressors.

this is a reduced form regression. why?

Actually all regressors are regressed on  $z$

$\begin{cases} x_1 \text{ on } (x_1, z_2) \rightarrow \hat{x}_1 = P_z x_1 = x_1 \\ y_2 \text{ on } (x_1, z_2) \rightarrow \hat{y}_2 = P_z y_2 \end{cases}$  \*

$y_2 = x_1\gamma_1 + z_2\gamma_2 + \text{error}$  ... reduced form!

$\Rightarrow \hat{y}_2 = x_1\hat{\gamma}_1 + z_2\hat{\gamma}_2 = z\hat{\gamma}$  where  $z = (x_1, z_2)$

$= P_z y_2$  where  $P_z = z(z'z)^{-1}z'$

thus,  $\hat{X} = \begin{pmatrix} \hat{x}_1 \\ \hat{y}_2 \end{pmatrix} = P_z X$  where  $X = (x_1, y_2)$   
 $\uparrow$   
 $\hat{x}_1 = x_1$

and OLS of  $y_1$  on  $X_1$  and  $\hat{y}_2$ .

6

$$y_1 = X_1 \beta_1 + \hat{y}_2 \beta_2 + e$$

$$\text{Thus, } \hat{\beta}_{2SLS} = \begin{pmatrix} \hat{\beta}_{1,2SLS} \\ \hat{\beta}_{2,2SLS} \end{pmatrix} = (X' \hat{X})^{-1} X' y_1 \quad \text{where } \hat{X} = P_z X$$

$$= (X' P_z X)^{-1} X' P_z y_1$$

$\begin{matrix} \uparrow \\ (X_1, \hat{y}_2) \\ = (X_1, \hat{y}_2) \end{matrix}$

Note : ) this is actually 2SLS estimator, which is the case with  $L > K$ .

$$\begin{cases} L = \# \text{ of IVs} = \# \text{ of variables in } Z \\ K = \# \text{ of regressors} = \# \text{ of variables in } X. \end{cases}$$

$$\Rightarrow X = n \times K, \quad Z = n \times L$$

ii) If  $L = K$ ,  $Z/X$  is a square matrix, which is invertible.

then 2SLS becomes an IV estimator

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z' y_1$$

$$(ABC)^T = C^T B^T A^T$$

only if each is a square matrix.

in general,

$(X'Z)^T$  cannot be defined if  $L \neq K$ .

why!  $\hat{\beta}_{2SLS} = (X' P_z X)^{-1} X' P_z y_1$

$$= [X' Z (Z'Z)^{-1} Z' X]^{-1} X' Z (Z'Z)^{-1} Z' y_1$$

$$= (Z'Z)^{-1} Z' X (X'Z)^{-1} X' Z (Z'Z)^{-1} Z' y_1$$

$\underbrace{\hspace{10em}}$   
 cancel out

$$= (Z'X)^{-1} Z' y_1$$

iii) If  $L < K$ , no solution is possible. this is the identification issue.

"order condition"

← The required condition is,  $L \geq K$ . If so, the model is identified.

Identification Find "good" IVs ( $L \geq K$ ). 7

- i) which is exogenous to  $y_1$ ; and is not simultaneously determined along with  $y_1$ .
- ii) which is correlated with  $y_2$ .

Note Finding good IVs is not easy.

IVs may not be significantly correlated with  $y_2$ .

We call them weak IVs, which might be the focus of recent literatures.

Examples of IVs (Wooldridge 2, pp 7-90) "many good examples"

Ex 1) "Mother's edu" for "edu". (But, Mother's edu can be jointly determined along with edu.)

Ex 2) "First quarter birth dummy variable" for edu

.. Angrist & Krueger (1991, QJE)

; best (notorious) example for weak IVs, even with

$n = 300,000 \sim 500,000$  observations!

"At least some people are forced, by law, to attend school longer than they otherwise would. Thus edu is correlated with firstqtrt."

Ex 3) "Draft lottery number"

in the study examining the effect of service ( $y_2$ )

in the Vietnam war [Angrist (1991)] on wage ( $y_1$ ).

Ex 4) 'Natural boundary created by river' for 'concentration

in the study examining the effect of competition

among public schools; Hoxby (1996, QJE)

point

Are IVs the determinants for  $y_2$  but not for  $y_1$ ?

Ex 5) "Timing of mayoral and gubernatorial elections" for "size of police force" in a study examining the effects of police on crime rates: Levitt (1997, AER)

Ex 6) "Dummy for Catholic" for "attending Catholic school"; Evans & Schwab (1995, QJE)

Ex 7) "Regional variation in prices or taxes" in using individual-level data.

eg) # of traffic violations (of an individual) = ... +  $b_1$  Age +  $b_2$  Drinking + ...  
 ↓  
 endogenous (why?)  
 ↓

IV = state beer tax rate (which varies over different states)

(Is it a good IV?)

Ex 8) "College proximity" for "Edu"; Card (1995)

if someone grew up in the vicinity of a four-year college, her/his edu level may be higher

Ex 9) (# of new plants) = ... +  $\beta$  (adoption of env. regulation) + ...

IV = adoption of nearby county of env. regulation

(List et al. (Restat, 2002))

Ex 10) (drug use of a teen) = ... +  $\beta$  (alcohol under age) + ...  
 → IV = ??

Ex 11)  $y_t = x_{1t}\beta_1 + y_{t-1}\beta_2 + e_t$ ,  $e_t = \rho e_{t-1} + u_t$   
 ⇒  $\text{Cov}(y_t, e_t) \neq 0$  if  $\rho \neq 0$ .

point

$$y_1 = x_1 \beta_1 + \underbrace{y_2}_{IV} \beta_2 + e$$

$$IV = z_2$$

$\Rightarrow$   $IV(z_2)$  is not a determinant for  $y_1$  ("exclusion" restriction)  
but it is a determinant for  $y_2$

Two requirements

$\rightarrow$  " $z_2$  does not belong in  $y_1$ "

R1)  $z_2$  is exogenous to  $y_1$ ;  $Cov(z_2, e) = 0$

R2)  $z_2$  and  $y_2$  are partially correlated;  $Corr(\tilde{z}_2, \tilde{y}_2) \neq 0$   
(in the presence of  $x_1$ )

Check points

"Over-identification" test  $\leftarrow$  R1: We need more IVs ( $L > K$ ) to check on R1.  
If  $L = K$ , we cannot test for its validity.

"weak-IV"  $\leftarrow$  R2: We can check if the coeffs of  $z_2$  are significant.

$$y_2 = x_1 c_1 + z_2 c_2 + u$$

Ho:  $c_2 = 0$  should be rejected  
(F-test)

Another check point

Is there endogeneity?

$\Rightarrow$  Hausman test

i) Wald type: compare wim ols estimator

ii) Residual based Hausman test

$$y_2 = x_1 \beta_1 + y_2 \beta_2 + \rho \hat{V} + u$$

where  $\hat{V} = y_2 - \hat{y}_2$  from the 1st stage.

# (A) Testing for Endogeneity

(Hausman test)

1st Reduced form equation, obtain residuals

$$y_2 = x_1 \gamma_1 + z_2 \gamma_2 + v \Rightarrow \hat{v}_2$$

(Thus,  $z_2' \hat{v}_2 = 0$ . Why?)

2nd Add  $\hat{v}_2$  in the main structural form equation

$$y_1 = x_1 \beta_1 + y_2 \beta_2 + \rho \hat{v}_2 + e$$

Test  $H_0: \rho = 0$  (no endogeneity)

$H_a: \rho \neq 0$  (endogeneity)

Do t-test or F-test. (Use Robust var if heteroskedasticity is suspected)

Note i) When there are more than one endogenous regressors ( $y_2: n \times k_2, k_2 \geq 1$ ), then

$\hat{v}_2$  is  $n \times k_2$ . Thus test  $H_0: \rho_1 = \dots = \rho_{k_2} = 0$ .

... For each endogenous regressor,  $\hat{v}_2$  is obtained.

Then we use F-test (not t-test).

Move on this later!

(p.31)

$\Rightarrow$  ii)  $\hat{v}_2$  is a generated regressor (generated in the 1st stage regression). Then usual std errors are invalid, since the errors are correlated and these correlations are not taken into account. Thus, if  $H_0: \rho = 0$  is rejected, obtain std. errors by 2SLS.

Only in the (linear models)

H/w Over S.1 show this result  $\Rightarrow$

iii) the estimates of  $\beta_1$  and  $\beta_2$  are the same as those from the 2SLS. std errors are different

iv) It may be safe to use robust std. error // in testing for endogeneity.

v) the usual Hausman test uses:

$$(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})' [\text{Var}(\hat{\beta}_{2SLS}) - \text{Var}(\hat{\beta}_{OLS})]^{-1} (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$$

or using the coeff. of endogenous regressors

$$(\hat{\beta}_{2,2SLS} - \hat{\beta}_{2,OLS})' [\text{Var}(\hat{\beta}_{2,2SLS}) - \text{Var}(\hat{\beta}_{2,OLS})]^{-1} (\hat{\beta}_{2,2SLS} - \hat{\beta}_{2,OLS})$$

where

$$\text{Var}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$$

$$= \text{Var}(\hat{\beta}_{2SLS}) + \text{Var}(\hat{\beta}_{OLS}) - 2\text{Cov}(\hat{\beta}_{2SLS}, \hat{\beta}_{OLS})$$

$$= \text{Var}(\hat{\beta}_{2OLS}) - \text{Var}(\hat{\beta}_{OLS}) \text{ since}$$

$$\text{Cov}(\hat{\beta}_{2SLS}, \hat{\beta}_{OLS}) = \text{Var}(\hat{\beta}_{OLS})$$

under  $H_0$  of no endogeneity

### Hausman test

$H_0$ :  $\hat{\beta}_{OLS}$  is unbiased (no endogeneity)

$H_a$ :  $\hat{\beta}_{OLS}$  is biased ( $\hat{\beta}_{2SLS}$  is preferred)

$\Rightarrow H_0$ : more efficient estimator is unbiased.

this occurs when  $\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}$  is small.

i) if  $H_0$  is not rejected

$$\hat{\beta}_{OLS} \approx \hat{\beta}_{2SLS}$$

Thus  $\hat{\beta}_{2SLS}$  can be also used

ii) if  $H_0$  is rejected

$\hat{\beta}_{2SLS}$  should be used.

But the inverse matrix may not exist, and it's often negative definite, except when  $X_1$  is not included. The regression-based test is asymptotically equivalent to this formal test.

vi) the motivation for the regression based test is:

$$y_1 = x_1 \beta_1 + y_2 \beta_2 + e = x_1 \beta_1 + y_2 \beta_2 + e(V) + u$$

$$y_2 = z \gamma + V$$

$$\text{write } e = \rho V + u$$

$\hat{V}$  is used.

vii) LM test is also available.

1st OLS residuals of the main eq.  $y_1 = x_1 \beta_1 + y_2 \beta_2 + u$   
 $\Rightarrow \hat{u}$

2nd Regress  $\hat{u}$  on  $x_1, y_2$  and  $\hat{V}$  (1st stage residual)  
 $LM = n R^2$

(B) Weak IVs "hot" issue, these days)

HOT!

12

$\text{Cov}(y_2, z_2)$  is low.

- Consider a simple model,

$$y_1 = \beta_1 y_2 + e \quad \Rightarrow \hat{\beta}_1 = (z_2' y_2)^{-1} z_2' y_1$$

$\downarrow$   
endogenous. IV =  $z_2$

$$= \beta_1 + (z_2' y_2)^{-1} z_2' e$$

$$\text{plim } \hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(z_2, e)}{\text{Cov}(z_2, y_2)} = \beta_1 + \left(\frac{\sigma_e}{\sigma_{y_2}}\right) \frac{\text{Cov}(z_2, e)}{\text{Cov}(z_2, y_2)}$$

i)  $\text{plim } \hat{\beta}_1 \neq \beta_1$  if  $\text{Cov}(z_2, e) \neq 0$  (1st requirement for IV)  
 $\text{Cov}(z_2, e) = 0$

ii) If  $z_2$  is weakly correlated with  $y_2$ ,

$\text{Cov}(z_2, y_2)$  gets smaller.

The bias term gets larger.

- It's not easy to check  $\text{Cov}(z_2, e) = 0$ . But we can check the significance of (partial) correlations,  $\text{Cov}(z_2, y_2)$  from the 1st stage reduced form equation.

$$y_2 = x_1 y_1 + z_2 (y_2) + e_{11}$$

Test  $H_0: \gamma_2 = 0$        $H_a: H_0$  is not true

use F-test. [use Robust variance and Wald tests if needed: if  $y_2$  is discrete, p92]

$\Rightarrow$  If  $y_2$  is insignificant,  $z_2$  is not a good IV.

Note Staiger & Stock (1997, Econometrica) suggest that  $F^* > 10$ .

**DO NOT** put IVs in the main structural equation

Note One cannot check the significance of  $\beta_2(z_2)$  in the main structural form equation.

(Ex 5.5, p. 110, Wooldridge)

## (c) Over-identifying restrictions

13

Using more IVs leads to more efficient estimator.

But the question is whether they are valid IVs.

Testing on validity of IVs, when there are more IVs ( $L > k$ ) is the over-identifying restriction test, to see if  $\text{Cov}(Z, e) = 0$ .

1st. Obtain 2SLS residuals using all IVs,  $Z$ ; say  $\hat{e}$

2nd. Regress  $\hat{e}$  on  $Z$ , and obtain  $R^2$ .

$$\hat{e} = Z\alpha + \epsilon$$

$$LM = nR^2 \sim k^2_{\# \text{ of over-IVs}} \quad ; \text{Sargan's test}$$

where # of over-IVs =  $L - k$

$H_0$ : IVs are valid

$H_a$ : IVs are invalid.

Note - If  $H_0$  is rejected, it does not say which IVs are invalid. One may test subsets of orthogonality condition, say  $\text{Cov}(Z_i, e) = 0$ , using a subset of IVs.

- Over-identifying restriction tests are being done in the GMM framework (J-test of Hansen) 1982

- When  $k = L$ , this over-identifying restriction test is not defined.

Note  
If  $L = k$ ,  
 $df = 0$ .  
No such  
test is possible.

## Remarks

i) Computing  $\hat{\sigma}^2$  (RSS)

use  $\hat{e}_i = y_{1i} - x_{1i}\hat{\beta}_1 - y_{2i}\hat{\beta}_2$  correct

$\Rightarrow \hat{\sigma}^2 = \frac{1}{n-k} \sum \hat{e}_i^2$  ; the initial model  
 $y_1 = x_1\beta_1 + y_2\beta_2 + e$

$\hat{e}_i = y_{1i} - x_{1i}\hat{\beta}_1 - \hat{y}_{2i}\hat{\beta}_2$  incorrect

(as in your own 2 stage)  
 $y_1 = x_1\beta_1 + \hat{y}_2\beta_2 + e$

ii) Can we replace  $y_2$  with  $z_2$ ?

$y_1 = x_1\beta_1 + (y_2)\beta_2 + e$

↳ endogenous. IV =  $z_2$

$\Rightarrow$  One may be tempted to use

$y_1 = x_1\beta_1 + (z_2)\beta_2 + e$

↑  $y_2$  is replaced with  $z_2$ .

This is a reduced form. OLS is fine, but ...

iii) How about more than 1 endogenous regressors?

Hausman test  $\Rightarrow$  use the residuals of each of the reduced form (1st stage) regression.

iv) Panel data: Finding IVs is EASY.

$y_{it} = x_{it}\beta_1 + (y_{2it})\beta_2 + e_{it} \Rightarrow$  use lagged variables

;  $(y_{1,t-1}, y_{1,t-2} \dots), (y_{2,t-1}, y_{2,t-2} \dots)$  are IVs

$\Rightarrow$  More in ec671

(more next page)



## 2SLS Exercise

(1) OLS: biased

```
. use "2sls_mroz_428.dta"
. regress lwage educ exper expersq
```

Source	SS	df	MS	Number of obs =	428
Model	35.0223023	3	11.6741008	F( 3, 424) =	26.29
Residual	188.305149	424	.444115917	Prob > F =	0.0000
				R-squared =	0.1568
				Adj R-squared =	0.1509
Total	223.327451	427	.523015108	Root MSE =	.66642

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1074896	.0141465	7.60	0.000	.0796837	.1352956
exper	.0415665	.0131752	3.15	0.002	.0156697	.0674633
expersq	-.0008112	.0003932	-2.06	0.040	-.0015841	-.0000382
_cons	-.5220407	.1986321	-2.63	0.009	-.9124668	-.1316145

(2) Hausman test

```
. regress educ exper expersq motheduc fatheduc huseduc
```

Source	SS	df	MS	Number of obs =	428
Model	955.830608	5	191.166122	F( 5, 422) =	63.30
Residual	1274.36565	422	3.01982382	Prob > F =	0.0000
				R-squared =	0.4286
				Adj R-squared =	0.4218
Total	2230.19626	427	5.22294206	Root MSE =	1.7378

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0374977	.0343102	1.09	0.275	-.0299424	.1049379
expersq	-.0006002	.0010261	-0.58	0.559	-.0026171	.0014167
motheduc	.1141532	.0307835	3.71	0.000	.0536452	.1746613
fatheduc	.1060801	.0295153	3.59	0.000	.0480648	.1640955
huseduc	.3752548	.0296347	12.66	0.000	.3170049	.4335048
_cons	5.538311	.4597824	12.05	0.000	4.634562	6.44206

```
. predict edu_res, res
```

```
. regress lwage educ exper expersq edu_res
```

Source	SS	df	MS	Number of obs =	428
				F( 4, 423) =	20.48

Model		36.230504	4	9.05762599	Prob > F	=	0.0000	17
Residual		187.096947	423	.442309568	R-squared	=	0.1622	
-----								
Total		223.327451	427	.523015108	Adj R-squared	=	0.1543	
-----								
Root MSE = .66506								

lwage		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ		.0803918	.0216362	3.72	0.000	.0378639 .1229197
exper		.0430973	.013181	3.27	0.001	.017189 .0690057
expersq		-.0008628	.0003937	-2.19	0.029	-.0016366 -.000089
edu_res		.047189	.0285519	1.65	0.099	-.0089322 .1033102
_cons		-.1868574	.2835905	-0.66	0.510	-.7442794 .3705647

. test edu\_res

( 1) edu\_res = 0

*Handman test*

F( 1, 423) = 2.73  
 Prob > F = 0.0991

(3) 2SLS

. ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq

Instrumental variables (2SLS) regression

Source		SS	df	MS	Number of obs	=	428
Model		33.3927427	3	11.1309142	F( 3, 424)	=	11.52
Residual		189.934709	424	.447959218	Prob > F	=	0.0000
-----							
Total		223.327451	427	.523015108	R-squared	=	0.1495
-----							
Adj R-squared = 0.1435							
Root MSE = .6693							

lwage		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
educ		.0803918	.021774	3.69	0.000	.0375934 .1231901
exper		.0430973	.0132649	3.25	0.001	.0170242 .0691704
expersq		-.0008628	.0003962	-2.18	0.030	-.0016415 -.0000841
_cons		-.1868574	.2853959	-0.65	0.513	-.7478243 .3741096

Instrumented: educ  
 Instruments: exper expersq motheduc fatheduc huseduc

*(★★)*  
 } Compare these with (★)

*Coeffs = same  
 Ad. err = diff*

(4) My own 2 stages

1st stage

. regress educ exper expersq motheduc fatheduc huseduc

Source		SS	df	MS	Number of obs	=	428
--------	--	----	----	----	---------------	---	-----

Model		955.830608	5	191.166122
Residual		1274.36565	422	3.01982382
Total		2230.19626	427	5.22294206

F( 5, 422) = 63.30  
 Prob > F = 0.0000  
 R-squared = 0.4286  
 Adj R-squared = 0.4218  
 Root MSE = 1.7378

18

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0374977	.0343102	1.09	0.275	-.0299424	.1049379
expersq	-.0006002	.0010261	-0.58	0.559	-.0026171	.0014167
motheduc	.1141532	.0307835	3.71	0.000	.0536452	.1746613
fatheduc	.1060801	.0295153	3.59	0.000	.0480648	.1640955
huseduc	.3752548	.0296347	12.66	0.000	.3170049	.4335048
_cons	5.538311	.4597824	12.05	0.000	4.634562	6.44206

. predict edu\_pre  
 (option xb assumed; fitted values)

2nd stage

. regress lwage edu\_pre exper expersq edu\_res

Source	SS	df	MS
Model	36.2305033	4	9.05762583
Residual	187.096948	423	.44230957
Total	223.327451	427	.523015108

Number of obs = 428  
 F( 4, 423) = 20.48  
 Prob > F = 0.0000  
 R-squared = 0.1622  
 Adj R-squared = 0.1543  
 Root MSE = .66506

or  
 edu →

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edu_pre	0.0803918	.0216362	3.72	0.000	.0378639	.1229197
exper	.0430973	.013181	3.27	0.001	.017189	.0690057
expersq	-.0008628	.0003937	-2.19	0.029	-.0016366	-.000089
edu_res	.1275808	.0186301	6.85	0.000	.0909616	.1642
_cons	-.1868573	.2835905	-0.66	0.510	-.7442793	.3705647

Incorrect

(~~STAR~~)

Compare these with (~~STAR~~) & (~~STAR~~)

Note that the coefficients are the same as those from the 2SLS, but Std. Err. values are different. It is advised to use the command, 2SLS, which corrects for Std. Err.

- (1)  $y_1 = x_1\beta_1 + y_2\beta_2 + p\hat{v} + e_{1NV}$  .. 2SLS, Hausman
- (2)  $y_1 = x_1\beta_1 + \hat{y}_2\beta_2 + p\hat{v} + e_{1NV}$  .. same as (1)? why?
- (3)  $y_1 = x_1\beta_1 + \hat{y}_2\beta_2 + e_{1NV}$  .. 2SLS (my own)
- (4) 2SLS using  $\hat{\sigma}^2$  with  $\hat{e}_x = y_{1x} - x_{1x}\hat{\beta}_1 - \hat{y}_2\hat{\beta}_2$  from 2SLS estimator  
 [not  $\hat{y}_2$ ]

```

. *** Hausman test (using Wald type test)
.
. regress lwage educ exper expersq
. est store ols

. ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
. hausman ols .

```



---- Coefficients ----				
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	ols	.	Difference	S.E.
educ	.1074896	.0803918	.0270979	.
exper	.0415665	.0430973	-.0015308	.
expersq	-.0008112	-.0008628	.0000516	.

b = consistent under Ho and Ha; obtained from regress  
 B = inconsistent under Ha, efficient under Ho; obtained from ivreg

Test: Ho: difference in coefficients not systematic

$$\begin{aligned}
 \text{chi2}(3) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\
 &= -2.68 \quad \text{chi2} < 0 \implies \text{model fitted on these} \\
 &\quad \text{data fails to meet the asymptotic} \\
 &\quad \text{assumptions of the Hausman test;} \\
 &\quad \text{see suest for a generalized test}
 \end{aligned}$$

← evvvv!  
why?

```

. ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
. est store keep

. regress lwage educ exper expersq
. hausman keep .

```

---- Coefficients ----				
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	keep	.	Difference	S.E.
educ	.0803918	.1074896	-.0270979	.0165524
exper	.0430973	.0415665	.0015308	.0015398
expersq	-.0008628	-.0008112	-.0000516	.0000482

b = consistent under Ho and Ha; obtained from ivreg  
 B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

$$\begin{aligned}
 \text{chi2}(3) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\
 &= 2.68 \\
 \text{Prob} > \text{chi2} &= 0.4436
 \end{aligned}$$

```
. *** Hausman test (using IVENDOG)
.
. ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
. ivendog
```

Tests of endogeneity of: educ

H0: Regressor is exogenous  
 Wu-Hausman F test: 2.73157 F(1,423) P-value = 0.09912  
 Durbin-Wu-Hausman chi-sq test: 2.74613 Chi-sq(1) P-value = 0.09749

```
. ivreg lwage (educ exper = motheduc fatheduc huseduc)
. ivendog
```

Tests of endogeneity of: educ exper

H0: Regressors are exogenous  
 Wu-Hausman F test: 1.53128 F(2,423) P-value = 0.21746  
 Durbin-Wu-Hausman chi-sq test: 3.07648 Chi-sq(2) P-value = 0.21476

```
. *** F-test for weak IVs
```

```
. ivreg lwage (educ = motheduc fatheduc huseduc)
. regress educ exper expersq motheduc fatheduc huseduc
```

Source	SS	df	MS	Number of obs =	428
Model	955.830608	5	191.166122	F( 5, 422) =	63.30
Residual	1274.36565	422	3.01982382	Prob > F =	0.0000
				R-squared =	0.4286
				Adj R-squared =	0.4218
Total	2230.19626	427	5.22294206	Root MSE =	1.7378

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0374977	.0343102	1.09	0.275	-.0299424	.1049379
expersq	-.0006002	.0010261	-0.58	0.559	-.0026171	.0014167
motheduc	.1141532	.0307835	3.71	0.000	.0536452	.1746613
fatheduc	.1060801	.0295153	3.59	0.000	.0480648	.1640955
huseduc	.3752548	.0296347	12.66	0.000	.3170049	.4335048
_cons	5.538311	.4597824	12.05	0.000	4.634562	6.44206

```
. test motheduc fatheduc huseduc
```

- ( 1) motheduc = 0
- ( 2) fatheduc = 0
- ( 3) huseduc = 0

F( 3, 422) = 104.29  
 Prob > F = 0.0000

```
. ivreg lwage (educ exper = motheduc fatheduc huseduc)
. regress educ  expersq  motheduc fatheduc huseduc
. test motheduc fatheduc huseduc
```

- ( 1) motheduc = 0
- ( 2) fatheduc = 0
- ( 3) huseduc = 0

F( 3, 423) = 104.75  
 Prob > F = 0.0000

```
. regress exper  expersq  motheduc fatheduc huseduc
. test motheduc fatheduc huseduc
```

- ( 1) motheduc = 0
- ( 2) fatheduc = 0
- ( 3) huseduc = 0

F( 3, 423) = 0.15  
 Prob > F = 0.9328

```
. *** Overidentification restriction test
```

```
. ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
. overid
```

Tests of overidentifying restrictions:

Sargan N*R-sq test	1.115	Chi-sq(2)	P-value = 0.5726
Basmann test	1.102	Chi-sq(2)	P-value = 0.5763

```
. ivreg lwage (educ exper = motheduc fatheduc huseduc) expersq
. overid
```

Tests of overidentifying restrictions:

Sargan N*R-sq test	0.040	Chi-sq(1)	P-value = 0.8424
Basmann test	0.039	Chi-sq(1)	P-value = 0.8433

(Summary)

Check points (three results to show)

22

(1) Is there evidence of endogeneity?

⇒ Hausman tests!

i) residual based Hausman test

"ivendog" module (installation required)

or ; Durbin-Wu-Hausman (DWH) test

Do it by yourself; see example.

$H_0$ : Regressors are exogenous (OLS is valid)

(if rejected, IV estimation is valid.)  
(OLS is invalid)

∴) Wald type Hausman test; see example

ivreg y1 x1 (y2=z2)  
est store tsl  
reg y1 x1 y2  
hausman tsl.

⇐ ( reg y1 x1 y2  
est store ols  
ivreg y1 x1 (y2=z2)  
hausman ols.

$H_0$ : OLS is valid  
 $H_a$ : OLS is invalid  
If  $H_0$  is rejected  
then we'd use  
IV results.

(2) Are IVs valid?

i) Are IVs weak?

check the F-stat on the sig of IVs  
in the reduced form equations.

( reg y2 x1 z2  
test z2

.. Do this for each of the  
endogenous regressors

!!) Are IVs exogenous?

$$cov(z_2, e) = 0 ?$$

- If  $L = k$  (exactly identified), we cannot test this hypothesis.

- If  $L > k$  (over-identified), we can test this hypothesis

$$\# \text{ restriction} = L - k = \text{degree of freedom.}$$

( $\overset{\text{ivreg}}{\text{overid}}$   $y_1 \quad x_1 \quad (y_2 = z_2)$ )

If  $H_0$  is rejected, IVs are not valid.

( $H_0$ : IVs are valid)

$\Rightarrow$  Sargan's test, J-test (GMM)

(Basmann test)

Note If IVs are weak or not exogenous, the Hausman test for endogeneity is not trustful. Thus, it's important to check these two required conditions; not weak, exogenous.

TABLE VI  
THE IMPACT OF PRISON POPULATIONS ON AGGREGATE CRIME CATEGORIES

Variable	Δln Violent crime			Δln Property crime		
	OLS (1)	IV (2)	IV (3)	OLS (4)	IV (5)	IV (6)
Δln Prison population(t-1)	-.099 (.033)	-.424 (.201)	-.379 (.180)	-.071 (.019)	-.321 (.138)	-.261 (.117)
Δln Income per capita	.485 (.117)	.384 (.127)	.410 (.127)	.014 (.066)	.076 (.072)	.055 (.070)
Δ Unemployment rate	.564 (.333)	.411 (.301)	.451 (.302)	1.032 (.186)	1.138 (.188)	1.063 (.181)
Δln Police	.026 (.059)	.054 (.048)	.063 (.048)	-.004 (.033)	.012 (.030)	.002 (.029)
Δ % Black	-.015 (.029)	-.018 (.025)	.007 (.058)	-.043 (.016)	-.038 (.016)	.000 (.035)
Δ % Metro	.013 (.011)	.006 (.012)	.027 (.021)	.006 (.006)	-.000 (.006)	.005 (.011)
Δ % Age 0-14	-.287 (.412)	-.075 (.393)	-.127 (.447)	.220 (.230)	.121 (.234)	.399 (.257)
Δ % Age 15-17	-.041 (.213)	.169 (.205)	.180 (.226)	.351 (.119)	.320 (.121)	.390 (.127)
Δ % Age 18-24	.320 (.253)	.282 (.235)	.286 (.253)	.277 (.141)	.079 (.139)	.126 (.144)
Δ Age 25-34	.648 (.335)	.748 (.329)	.828 (.350)	.384 (.187)	.354 (.195)	.436 (.202)
Year controls?	Yes	Yes	Yes	Yes	Yes	Yes
State controls?	No	No	Yes	No	No	Yes
Instrument?	No	Yes	Yes	No	Yes	Yes
R <sup>2</sup>	.247	—	—	.606	—	—
P-value overidentifying restrictions	—	.369	.424	—	.416	.164

(y<sub>2</sub>) endo :

(x<sub>1</sub>)

Σ<sub>2</sub> ←  
validity  
of IVs

The dependent variable is Δln Violent crime rate or Δln Property crime. The data set is comprised of annual state level data from 1972-1993 (implying observations on changes for the years 1973-1993). Number of observations is equal to 1063 in all columns. Prison population data correspond to December 31 of the year. Consequently, the once-lagged value is used as an explanatory variable. In all cases, estimation allows for heteroskedasticity across states. In instrumental variables specifications, ten indicator variables corresponding to changes in prison overcrowding litigation status in the current year/two preceding years are used as instruments for the percent change in the prison population. In all columns using IV, the test of overidentifying restrictions is computed using an N × R<sup>2</sup> test, where N is the number of observations and R<sup>2</sup> is the R<sup>2</sup> from a regression of the residuals from the second-stage regression on all of the exogenous variables and the instruments. This test statistic is distributed χ<sup>2</sup> with degrees of freedom equal to the number of overidentifying restrictions (in this case nine). Overcrowding litigation status refers only to states whose entire prison system is under court control. For the definitions of status categories, see text.

Steven Levitt, "The Effects of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation", The Quarterly Journal of Economics, Vol III, May 1996, 319-351.

# IV Estimator (Algebra)

25

$$y_1 = (x_1, y_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + e$$

$$= X\beta + e \quad \text{where } X = (x_1, y_2), \quad y_2: n \times k_2$$

$$(IV = Z) \quad Z = (x_1, z_2), \quad z_2: n \times L_2$$

i)  $K = L$  (ie.  $k_2 = L_2$ ) : Exactly identified. IV

required condition:  $z'e = 0$

$$z'(y_1 - X\beta) = 0 \Rightarrow z'y_1 - z'X\beta = 0$$

H/w show these  $\Rightarrow \hat{\beta}_{IV} = (z'X)^{-1} z'y_1, \text{Var}(\hat{\beta}_{IV}) = \sigma^2 (z'X)^{-1} (z'z)^{-1} (z'X)^{-1}$

ii)  $L \geq k$  : Over-identified ( $L > k$ ) 2SLS

1st stage: reduced form

$$X = Z\gamma + e_{uvr}$$

$$\Rightarrow \hat{X} = Z\hat{\gamma} \quad \text{where } \hat{\gamma} = (z'z)^{-1} z'X$$

$$= P_Z X \quad \text{where } P_Z = z(z'z)^{-1} z'$$

2nd stage:

$$y_1 = \hat{X}\beta + e$$

$$\Rightarrow \hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1} \hat{X}'y_1 \quad \text{with } \hat{X} = P_Z X$$

$$= (X'P_Z X)^{-1} X'P_Z y_1, \text{Var}(\hat{\beta}_{2SLS}) = \sigma^2 (X'P_Z X)^{-1}$$

iii)  $L < k$  : Not-identified (underidentified)

No solution.

Distribution of IV estimator

Assume  $\therefore \text{plim} \frac{1}{n} z'x$  exists and is non-singular  $\Rightarrow$  (Note this is the rank condition for identification if  $L > K$ )  
 $\therefore \text{plim} \frac{1}{n} z'e = 0$

then

a)  $\frac{1}{\sqrt{n}} z'e \rightarrow N(0, \sigma^2 \text{plim} \frac{1}{n} z'z)$

b)  $\sqrt{n} (\hat{\beta}_{IV} - \beta) \rightarrow N(0, \sigma^2 \text{plim} (\frac{1}{n} z'x) \left[ \frac{1}{n} z'z \right]^{-1} (\frac{1}{n} z'x)' )$

$\therefore \text{Var}(\hat{\beta}_{IV}) = \sigma^2 (z'x)^{-1} z'z (z'x)^{-1}$

c)  $\sqrt{n} (\hat{\beta}_{2SLS} - \beta) \rightarrow N(0, \sigma^2 \text{plim} (\frac{1}{n} x'P_z x)^{-1})$

proof)  $\sqrt{n} (\hat{\beta}_{2SLS} - \beta) = (\frac{1}{n} x'P_z x)^{-1} \frac{1}{\sqrt{n}} x'P_z e$

i)  $\frac{1}{\sqrt{n}} x'P_z e = (\frac{1}{n} x'z) (\frac{1}{n} z'z)^{-1} \frac{1}{\sqrt{n}} z'e$

$\rightarrow N(0, \text{plim} (\frac{1}{n} x'z) (\frac{1}{n} z'z)^{-1} \sigma^2 \frac{1}{n} z'z (\frac{1}{n} z'z)^{-1} (\frac{1}{n} z'x))$

$= N(0, \sigma^2 \text{plim} \frac{1}{n} x'P_z x)$

$\therefore \text{Var}(\frac{1}{\sqrt{n}} x'P_z e) = \frac{1}{n} \sigma^2 (x'P_z x)$

$\therefore \text{Var} \left[ \underbrace{(\frac{1}{n} x'P_z x)^{-1}}_A \cdot \underbrace{\frac{1}{\sqrt{n}} x'P_z e}_B \right] = A \cdot \text{Var}(B) \cdot A'$

$= (\frac{1}{n} x'P_z x)^{-1} \cdot \frac{1}{n} \sigma^2 (x'P_z x) \cdot (\frac{1}{n} x'P_z x)^{-1}$

$= \sigma^2 (\frac{1}{n} x'P_z x)^{-1}$

then  $\text{Var}(\hat{\beta}_{2SLS}) = \sigma^2 (x'P_z x)^{-1}$

"Using more IVs leads to more efficient estimators"

Exercise

H/W

Let  $\hat{\beta}$  be 2SLS using  $z = [z_1, z_2]$  for an IV for  $x$ .

Let  $\tilde{\beta}$  be 2SLS using  $z_1$  for an IV for  $x$ .

Show that  $\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$  is p.s.d.

Note  $P_z = P_{z_1} + P_{M_1 z_2}$

## Robust std. error of IV estimators

27

$$\text{Var}(\hat{\beta}_{2SLS}) = \hat{\sigma}^2 (\hat{X}'\hat{X})^{-1} = \hat{\sigma}^2 (X'P_Z X)^{-1}$$

under homoskedasticity

where  $\hat{\sigma}^2 = \frac{1}{n-k} \sum \hat{e}_i^2$ ,  $\hat{e}_i$  is the residual from the 2nd stage regression (2SLS residuals).

$$\text{Var}(\hat{\beta}_{2SLS}) = (\hat{X}'\hat{X})^{-1} (\sum \hat{e}_i^2 X_i'X_i) (\hat{X}'\hat{X})^{-1} \quad \text{robust variance}$$

where  $\hat{X}$  is used (not  $X$ ).

(Note on the tricks on computing this)  
: Wooldridge, p. 100

Usual F-tests (Wald tests..) can be used

as long as 2SLS residuals ( $\hat{e}_i$ ) are used.

## Potential Pitfalls of 2SLS

i) IV estimators are never unbiased; Wooldridge, p. 101  
Kruel (1980)

:  $\text{Cov}(Z_2, e) = 0$  is hardly satisfied.

ii) Weak IVs

: Even large samples cannot help. Always check F-stat.

iii) 2SLS standard errors have a tendency to be "large".

: t-stat gets smaller, insignificant coeff.

## Exercises

Wooldridge 2

Ex 5.1 (p. 107), Ex 5.3 (p. 109, Empirical est.)

Ex 5.5 (p. 110)

# Inconsistency of IV estimates

$$(1) \frac{\hat{\beta}_{ZIV} - \beta}{\hat{\beta}_{OLS} - \beta} = \frac{(Z'X)^{-1} Z'u}{(X'X)^{-1} X'u} = \frac{\frac{Z'u}{(Z'Z)^{\frac{1}{2}}(u'u)^{\frac{1}{2}}} \frac{1}{(Z'X)}}{\frac{X'u}{(X'X)^{\frac{1}{2}}(u'u)^{\frac{1}{2}}} \frac{1}{(X'X)^{\frac{1}{2}}}}$$

$$= \frac{\text{Cor}(Z, u)}{\text{Cor}(X, u)} \frac{1}{\text{Cor}(X, Z)}$$

If  $\text{Cor}(X, Z) = 1$ , both are equally consistent.

If  $|\text{Cor}(X, Z)| < 1$ ,  $\hat{\beta}_{ZIV}$  is inconsistent.  
(Ratio > 1) ;  $\text{Cor}(Z, u) > \text{Cor}(X, u)$

$$(2) \text{Var}(\hat{\beta}_{ZIV}) = \sigma^2 (X'Z)^{-1} Z'Z (Z'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1} \cdot \left( \frac{(Z'Z)^2}{(Z'Z)(X'X)} \right)^{-1} \quad \text{under homoskedasticity for simplicity}$$

$$= \text{Var}(\hat{\beta}_{OLS}) \cdot \frac{1}{\text{Cor}(X, Z)^2}$$

If  $\text{Cor}(X, Z) = 1$ , both are equally efficient

If  $|\text{Cor}(X, Z)| < 1$ ,  $\text{Var}(\hat{\beta}_{ZIV}) > \text{Var}(\hat{\beta}_{OLS})$

thus  $\hat{\beta}_{ZIV}$  is less efficient.

Alternatively

$$\text{Var}(\hat{\beta}_{ZIV}) - \text{Var}(\hat{\beta}_{OLS}) = \sigma^2 (X'Z)^{-1} Z'Z (Z'X)^{-1} - \sigma^2 (X'X)^{-1} \quad \text{assuming } \sigma^2 \text{ is known}$$

$\Rightarrow$  this is psd since  $(X'X) - (X'Z)(Z'Z)^{-1}(Z'X)$  is psd  
[ =  $X'[I - Z(Z'Z)^{-1}Z']X = X'P_Z X$  is psd ]

# Rank Condition (If we have a fully specified system)

When the order condition ( $R \geq G-1$ ) is met, we can continue to check the rank condition for each equation

$R = \#$  of restrictions,  $G = \#$  of endogenous variables

$$\text{rank}(\phi\beta) \geq G-1 \quad \begin{matrix} = \text{exact} & > \text{over-identified} \\ & < \text{under(not) identified} \end{matrix}$$

where  $\phi$  is the restriction matrix whose element is 1 if the corresponding parameter is zero.

Ex) (1)  $y_1 = \gamma_{12} y_2 + \beta_{11} x_1 + e_1$   $G=3$  ( $y_1, y_2, y_3$ )  
 (2)  $y_2 = \gamma_{21} y_1 + \gamma_{23} y_3 + \beta_{21} x_1 + \beta_{23} x_3 + e_2$   $G=3$  (exog:  $x_1, x_2, x_3$ )  
 (3)  $y_3 = \gamma_{32} y_2 + \beta_{32} x_2 + \beta_{33} x_3 + e_3$

- i) order condition
- 1st  $R=3$  ( $y_3, x_1, x_3$ )  $R \geq G-1$  ok
  - 2nd  $R=1$  ( $x_2$ ) Not identified
  - 3rd  $R=2$  ( $y_1, x_2$ )  $R \geq G-1$  ok

ii) Rank condition

1st  $\phi\beta = \begin{bmatrix} 0 & \gamma_{23} & -1 \\ 0 & 0 & \beta_{32} \\ 0 & \beta_{23} & \beta_{33} \end{bmatrix}$   $(y_3)$  rank = 2 =  $G-1$   
 $(x_2)$  ok  
 $(x_3)$

(1)    (2)    (3)

2nd Not identified (order condition failed!)

3rd  $\phi\beta = \begin{bmatrix} -1 & \gamma_{21} & 0 \\ \beta_{11} & \beta_{21} & 0 \end{bmatrix}$   $(y_1)$   
 $(x_1)$

(1)    (2)    (3)

rank = 2 unless  $\beta_{21} + \beta_{11}\gamma_{21} = 0$   
(ok)

point to be able to estimate the whole system using 3SLS all equations should be identified (order & rank conditions should be satisfied)

### Example of IV estimation

- Romer (1993, Quarterly Journal of Economics) proposed theoretical models of inflation which imply that more "open" countries should have lower inflation rates (*Inf*). His empirical analysis explains average annual inflation rates in terms of the average share of imports in gross domestic product, which is his measure of openness. While Romer did not specify explicitly, he must have had the following models in mind.

$$(A) \quad \text{Inf} = b_0 + b_1 \text{Open} + b_2 \log(\text{PINC}) + b_3 \log(\text{population}) + u$$

$$(B) \quad \text{Open} = c_0 + c_1 \text{Inf} + c_2 \log(\text{PINC}) + c_3 \log(\text{land}) + e$$

where *PINC* is per capita income (assumed to be exogenous), *land* is the land area of the country in square miles (also assumed to be exogenous) and *population* is the population of the country.

- What are endogenous variables? What are exogenous variables?
  - Is each of the above equations considered as a reduced form model?
  - Which equation(s) is identified? Which equation(s) can you estimate?
  - What kind of problem you will have if you estimate the identified equation(s) by OLS? What classical assumption is violated?
  - How can you detect the problem you answered in part (d)? Explain the Hausman testing procedure for each of the equations that are identified.
  - What will happen if in the Hausman test for endogeneity, you omit the original endogenous regressor but include both predicted value and residual variables from the first stage regression?
  - What is your proposed estimation method as a solution to the problem you answered in part (e), for each of the equations that are identified? Discuss the procedures briefly, but as specifically as possible.
- Consider the three-equation model system, where  $X_1$  and  $X_2$  are exogenous.

$$(A) \quad Y_1 = a_1 + a_2 Y_2 + a_3 X_1 + a_4 X_2 + e_1$$

$$(B) \quad Y_2 = b_1 + b_2 Y_3 + b_3 X_2 + e_2$$

$$(C) \quad Y_3 = c_1 + c_2 Y_2 + e_3$$

- (a) Check if each of the three equations is identified.
- (b) For each of the equations which are identified, propose your suggested estimation method. Briefly explain the estimation procedure. Be specific on the instrumental variables that you will be using for each.

3. Explain why  $y_{t-1}$  is endogenous for an AR(1) model with AR(1) error.

$$y_t = \alpha + \beta x_t + \gamma y_{t-1} + e_t$$

$$e_t = \rho e_{t-1} + u_t$$

What will be your suggested IVs?

## Empirical Exercises

### (IV Estimation)

We will estimate the type of the wage equation estimated by Grilliches using an extract from the NLS-Y used by Blackburn and Neumark (1992). The NLS-Y is panel data, with the same set of young men surveyed at several points in time (we will not exploit the panel feature of the data set in this exercise, though). The extract contains information about those individuals at two points in time: first, the earliest year in which wages and other variables are available, and second, in 1980. In a data file "return\_to-school\_data\_grilic.xls", data are provided on *RNS*, *RNS80*, *MRT*, *MRT80*, *SMSA*, *SMSA80*, *MED*, *IQ*, *KWW*, *YEAR*, *AGE*, *AGE80*, *S*, *S80*, *EXPR*, *EXPR80*, *TENURE*, *TENURE80*, *LW*, and *LW80* (in this order, with columns corresponding to the variables, as usual). The variable *YEAR* is the year of the first point time. Variables without "80" are for the first point, and those with "80" are for 1980. The definition of the variables for the first point is:

*RNS* = dummy for residency in the southern states  
*MRT* = dummy for marital status (1 if married)  
*SMSA* = dummy for residency in metropolitan areas  
*MED* = mother's education in years  
*KWW* = score on the "Knowledge of the World of Work" test  
*IQ* = IQ score  
*AGE* = age of the individual  
*S* = completed years of **schooling**  
*EXPR* = experience in years  
*TENURE* = tenure in year  
*LW* = log wage

Since the year the wage rate observed differs across individuals, the wage rate will have the year effect. Generate eight year dummies for *YEAR*= 66, . . . 73. (Note: There is observation for 1972.) The year dummies will be included in the log wage equation to control for the year effect.

Hint: generate  $d66 = (year == 66)$

generate  $d73 = (year == 73)$

Now, consider the wage equation (dropping the individual subscript  $i$ )

$$LW = \alpha + \beta_1 S + \gamma IQ + \delta' \mathbf{h} + \varepsilon \quad \dots (1)$$

where *LW* is log wages, *S* is schooling and

$$\mathbf{h} \equiv (\text{EXPR}, \text{TENURE}, \text{RNS}, \text{SMSA}, \text{year dummies})'$$

(You may drop one time dummy variable, say  $d66$ , when you include a constant term.)

In the above, IQ is used as a proxy variable for *Ability*. If IQ is a perfect measure of ability then the wage equation can be estimated consistently. If not, IQ is measured with error.

- (a) Estimate the model (1) by OLS, when IQ is used as a proxy variable for *Ability*.
- (b) The IQ measure may not be an error-free measure of ability. What happens then?
- (c) Now, consider KWW as an instrumental variable for IQ. Test if there is an error-in-variables problem.
- (d) Test if the coefficient of KWW is significant in the first stage reduced form equation.
- (e) Using KWW as an instrumental variable for IQ, estimate the model (1) by an IV estimation.
- (f) Test if there is an error-in-variables problem, using a set of instruments (*MED, KWW, MRT, AGE*).
- (g) Test if the coefficients of these IVs are significant in the first stage reduced form equation.
- (h) Using a set of instruments (*MED, KWW, MRT, AGE*) for IQ, estimate the model (1) by 2SLS. *Provide three statistics to justify your use of 2SLS.*
- (i) Test if the coefficients of time dummy variables are significant.
- (j) Grilliches mentions that *S* (schooling), too, may be endogenous. What is his argument?
- (k) Estimate by 2SLS the wage equation (1), treating both *IQ* and *S* as endogenous. What happens to the schooling coefficient? How would you explain the difference between your 2SLS estimate of the schooling coefficient here and your 2SLS estimate in (h)?  
*Provide three statistics to justify your use of 2SLS as well.*

# Additional Empirical Exercises

34

HW

- Wooldridge 2, ch 6, Ex 6.7 (p. 136)

"HPRICE.RAW" by Kiel & McClain (1995) JCEM

a) "reg lprice ldist if y81"

... Regression using the 1981 cross-section data. (# obs = 142)

b) "gen y81ldist = y81 \* ldist"

... creating an interaction variable.

c) # of obs = 321

- Wooldridge 2, ch 6, Ex 6.1 (p. 135)

"CARD.RAW" by Card (1995) Endogeneity test

Note Use two different tests for endogeneity

① Hausman test

② LM test

Note Calculating LM stat

"display 3010 \* 0.0004"

ans = 1.204

"display chiprob(1, 1.204)"

ans = 0.2733

... This is the p-value for the overidentifying restriction test (df = 1) why!

More on Simultaneous equations Models

(1) Can we use nonlinear function of z as additional IV in estimation?  $x_1^2, x_1^3, \dots$

Wooldridge 2  
p 229

$$y_1 = x_1 \beta_1 + y_2 \beta_2 + x_2 \beta_3 + u_1 \quad - (A)$$

$$y_2 = x_1 c_1 + y_1 c_2 + u_2 \quad - (B)$$

which equation is identified?

- consider eq (A).

Can we use  $x_1^2$  as an IV for  $y_2$ ?

$\left\{ \begin{array}{l} x_1^2 \text{ and } y_2 \text{ may be correlated} \\ x_1^2 \text{ could be exogenous since } x_1 \text{ is exog.} \end{array} \right.$

The answer is no!

Why?

the reduced form for  $y_2$  does not contain  $x_1^2$ .  
the partial correlation between  $y_2$  and  $x_1^2$  could be weak if  $x_1$  &  $x_2$  are included.

$$y_2 = d_1 x_1 + d_2 x_2 + d_3 x_1^2 + \text{err}$$

$\downarrow$   
 $\Rightarrow d_3 = 0$  is not rejected

How about using  $x_2^2$  as an IV?

$\Rightarrow$  the same problem.

Also  $(x_1 x_2)$  and other nonlinear functions of  $x_1$  and  $x_2$  (say  $\log(x_1), \exp(x_2), \dots$ )

(2) nonlinear endogenous regressors

How about the following model where  $y_2^2$  appears?

$$y_1 = x_1\beta_1 + (y_2^2)\beta_2 + x_2\beta_3 + u_1 \quad - (A)$$

$$y_2 = x_1\alpha_1 + y_1\alpha_2 + u_2 \quad - (B)$$

- One may be tempted to use  $(\hat{y}_2)^2$ , where  $\hat{y}_2$  is the predicted value of  $y_2$  in the 1st stage reduced form, to estimate (A);  $\hat{y}_2 = x_1\hat{\alpha}_1 + x_2\hat{\alpha}_2 \Rightarrow (\hat{y}_2)^2$ .

$$y_1 = x_1\beta_1 + (\hat{y}_2)^2\beta_2 + x_2\beta_3 + u_1 \quad - (A)^\dagger$$

This is the example of a forbidden regression.

point  $E(y_2^2 | x_1, x_2) \neq [E(y_2 | x_1, x_2)]^2$

$$\begin{matrix} \downarrow & & \downarrow \\ \widehat{(y_2^2)} & & (\widehat{y_2})^2 \end{matrix}$$

[ say,  $y_3 = y_2^2$   
this is  $\hat{y}_3$ . ]

[Another example of a forbidden regression]

$$y_1 = \exp(x_1\beta_1 + y_2\beta_2 + x_2\beta_3) + u_1$$

IV for  $y_2 = z_2$

One may be tempted to use  $\hat{y}_2 = x_1\hat{\alpha}_1 + x_2\hat{\alpha}_2 + z_2\hat{\alpha}_3$  to replace  $y_2$ .

$$y_1 = \exp(x_1\beta_1 + (\hat{y}_2)\beta_2 + x_2\beta_3) + u_1$$

$\Rightarrow$  we can't do this. why?

the reduced form for  $y_2$  is not a linear function of  $x_1, x_2$  and  $(z_2)$ .

• But we can still estimate eq (A).

Eq (A) is still identified due to a nonlinearity.

• Eqs (A) & (B) make a nonlinear system  
(even though we have a system of models  
nonlinear in variables)

• the identification is weak, though,  
if there are no additional IVs, say  $z_2$ .

[similar example: selection bias model]

$$y = x\beta + e\hat{x} + dD + u; \hat{x} = \text{inverse Mills ratio}$$

$$P(D=1) = \Phi(Z\gamma)$$

If  $X=Z$ , there is no IV for  $D$ .

But the system is still (weakly) identified  
due to a nonlinearity.

• Back to the nonlinear endogenous regressors

$$y_1 = x_1\beta_1 + y_2^2\beta_2 + x_2\beta_3 + u_1 \quad \text{--- (A)}$$

the solution is to use the nonlinear functions  
of exogenous variables for  $y_2^2$ .

$$\text{IVs} = x_1^2, x_2^2, (x_1x_2) \text{ in addition to } x_1 \text{ \& } x_2$$

Note It is easier to think that we have three  
endogenous variables:  $y_1$ ,  $y_2$  and  $y_3 \equiv y_2^2$ .  
The 3rd equation for  $y_3 \equiv y_2^2$  can be seen as

$$y_3 = x_1d_1 + x_2d_2 + x_1^2d_3 + x_2^2d_4 + (x_1x_2)d_5 + e_m$$

(3) F-test in 2SLS

$$y_1 = x_1\beta_1 + y_2\beta_2 + u \Rightarrow \hat{\beta}_{2SLS} = (\hat{\beta}_{1,2SLS}, \hat{\beta}_{2,2SLS})$$

let  $b$  be a subset of  $(\beta_1, \beta_2)$ .

$H_0: b = 0$  (m restrictions)

the usual f-stat is,  $F = \frac{(SSR_R - SSR_u) / m}{SSR_u / (N-k)}$

question How can we obtain  $SSR = \sum \hat{u}_i^2$  ?

let  $\hat{u}_i = y_{1i} - x_{1i}\hat{\beta}_{1,2SLS} - y_{2i}\hat{\beta}_{2,2SLS}$  : 2SLS residuals  
using  $x_1$  &  $y_2$

$\tilde{u}_i = y_{1i} - x_{1i}\hat{\beta}_{1,2SLS} - \hat{y}_{2i}\hat{\beta}_{2,2SLS}$  : 2nd stage residuals  
using  $x_1$  &  $\hat{y}_2$

$$\text{where } \hat{y}_{2i} = x_{2i}b_1 + z_{2i}b_2$$

in the 2nd stage regression  
using IVs  $z_2$  for  $y_2$ .

$$\Rightarrow \underbrace{\hat{SSR}_R, \hat{SSR}_u}_{\text{using } \hat{u}_i} \text{ and } \underbrace{\tilde{SSR}_R, \tilde{SSR}_u}_{\text{using } \tilde{u}_i}$$

Correct

$$F = \frac{(\tilde{SSR}_R - \tilde{SSR}_u) / m}{\hat{SSR}_u / (N-k)} : \text{2SLS F-stat}$$

incorrect

$$F = \frac{(\hat{SSR}_R - \hat{SSR}_u) / m}{\hat{SSR}_u / (N-k)}$$

the numerator & the denominator are Not independent! Not valid F-stat.

Note Do NOT use F-stat based on

$R^2 \Rightarrow$  the  $R^2$  from 2SLS can be unstable.

#### (4) Simultaneous equation or 3SLS?

39

- Simultaneous equation : ONE equation, 2SLS

$$y_1 = x_1\beta_1 + \gamma_2 y_2 + u$$

$$y_2 : \text{exog.} \quad IV = z_2$$

we do not care for the structural form for  $y_2$

But, we use the reduced form for  $y_2$

$$y_2 = x_2 b_2 + z_2 \varepsilon \Rightarrow \hat{y}_2 \text{ or } \hat{\varepsilon}$$

- System of equations with endogeneity : Many eqs, 3SLS

$$\begin{cases} y_1 = x_1\beta_1 + \gamma_2 y_2 + u_1 & y_2 : \text{endo} \Rightarrow z_2 \\ y_2 = x_2\gamma_1 + \gamma_1 y_1 + u_2 & y_1 : \text{endo} \Rightarrow z_3 \end{cases}$$

$$: u_1 \text{ \& } u_2 \text{ are correlated} \Rightarrow \text{Cov} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \Omega$$

$$\begin{cases} \text{Do 2SLS on each and obtain } \hat{u}_1, \hat{u}_2 \text{ to construct } \hat{\Omega} \\ \text{Do GLS using } \hat{\Omega} \Rightarrow \text{3SLS (} \approx \text{2SLS + SUR)} \end{cases}$$

In the 3SLS, endogeneity correction can be done

$$\text{via } P = \text{Cov}(u_1, u_2).$$

Question : which one is preferred?

i) 3SLS can be more efficient than 2SLS.

ii) However, 3SLS can lead to inconsistent estimates if any equation in the system is mis-specified.

iii) In many cases, we do not care for the structural form for  $y_2$ . If so, forget 3SLS. Moreover, 3SLS can be unstable in many cases.

(5) Generated Regressors (Problem associated with Hausman tests  
But this is an important issue.)

$$y = X\beta + \gamma q + u \quad \dots (1)$$

$q$  is unobserved. Thus we estimate  $q$  with  $\hat{q}$

$$q = z\delta + \varepsilon \quad \dots (2)$$

$$y = X\beta + \gamma \hat{q} + u \quad \dots (3) \quad \hat{q} \text{ is a generated regressor}$$

point i) the estimator  $\hat{\beta}$  from (3) is consistent  
as long as  $\hat{q}$  and  $u$  are uncorrelated!!

ii) But the standard errors of  $\hat{\beta}$  from (3) are NOT valid (thus, t-stat on  $\beta$  is also invalid),  
since  $s(\hat{\beta})$  was obtained by not taking care of correlation between  $\varepsilon$  and  $u$ .

Also  $\rightarrow$   
 $s(\hat{\beta})$  is  
invalid.

$$H_0: \gamma = 0 \quad H_a: \gamma \neq 0$$

usual t-tests are invalid. (also F-test, LR, Wald & LM tests are invalid.)  
- Pagan (1984)

when does the problem occur?

i) Barro's unexpected monetary shocks

$$\text{Inflation} = \dots + \gamma \tilde{M} + u$$

where  $\tilde{M}$  is the residual from the money function (demand)

$$M = z\delta + \varepsilon \Rightarrow \tilde{M} = \hat{\varepsilon} \text{ (unexpected) (residuals)}$$

ii) Testing for endogeneity (Hausman test)

ii) Selection Bias model, Treatment effect models.

41

$$y_i = x_i \beta + d D_i + u_i$$

where  $D_i = \begin{cases} 1 & \text{if treated (treated)} \\ 0 & \text{if not treated (control)} \end{cases}$

Due to selection bias,  $D_i$  is endogenous.

$\therefore D_i$  &  $u_i$  are correlated.

We estimate a probit model

$$D_i = z_i \alpha + \varepsilon_i$$

to obtain  $\hat{\lambda}_i$  (inverse Mills ratio)

$$\text{where } \hat{\lambda}_i = \frac{\phi(z_i \hat{\alpha}) (D_i=1)}{\Phi(z_i \hat{\alpha})} \quad \text{or} \quad \frac{-\phi(z_i \hat{\alpha})}{1 - \Phi(z_i \hat{\alpha})} \quad \text{for } (D_i=0)$$

and add  $\hat{\lambda}_i$  to the main equation

$$y_i = x_i \beta + d D_i + \gamma \hat{\lambda}_i + u_i^*$$

$\Rightarrow$  Standard errors of the estimated parameters should be corrected by taking into account of  $\text{corr}(u_i^*, \varepsilon_i)$ .

Solution Murphy & Topel (1985)'s correction

... Greene, p30, Wooldridge p140

(1)  $y = h(x, \beta; z, \alpha) + u$  ... nonlinear function but can include a linear function

(2)  $z = f(w, \alpha) + \varepsilon$

eg)  $y = x\beta_1 + z\beta_2 + u$ ,  $\beta = (\beta_1, \beta_2)'$  ... (1)'

$z = w\alpha + \varepsilon$  ... (2)'

Let  $V_\beta = \text{Var}(\hat{\beta})$  from (1)

$V_\alpha = \text{Var}(\hat{\alpha})$  from (2)

then, the corrected variance of  $\hat{\beta}$  from (1) is given as

$$V_\beta^* = \sigma^2 V_\beta + V_\beta [C V_\alpha C' - C V_\alpha R' - R V_\alpha C'] V_\beta$$

where  $C = n \text{plim} \frac{1}{n} \sum_{i=1}^n X_i^0 \hat{u}_i^2 \left( \frac{\partial h_i}{\partial \alpha} \right)$

$X_0$  is the matrix of regressors in (1) evaluated at the true parameter values  $X_0 = \frac{\partial h}{\partial \beta}$

$$R = n \text{plim} \frac{1}{n} \sum X_i^0 \hat{u}_i \left( \frac{\partial \hat{\epsilon}_i}{\partial \alpha} \right)$$

$$\frac{\partial \hat{\epsilon}_i}{\partial \alpha} = \text{derivative of SSR in (2)} = \hat{\epsilon}_i \frac{\partial f}{\partial \alpha}$$

eg) linear model (previous page)

$$C = n \text{plim} \frac{1}{n} \sum X_i^0 \hat{u}_i^2 \frac{\partial h_i}{\partial \alpha} = \hat{\beta}_2 \sum \hat{u}_i^2 X_i W_i'$$

$$\left( \frac{\partial h_i}{\partial \alpha} = \hat{\beta}_2 \frac{\partial f}{\partial \alpha} = \hat{\beta}_2 \cdot W_i' \right)$$

$$X_{0i} = \frac{\partial h}{\partial \beta_1} = X_i$$

$$R = n \text{plim} \frac{1}{n} \sum X_{0i} \hat{u}_i (\hat{\epsilon}_i W_i) = \sum (\hat{u}_i \hat{\epsilon}_i) X_i W_i'$$

Note Additional examples: Heckmit (Heckman's selection models)  
(Green, p. 785)

Alternative (Stata journal  
by Harding).

# More on IV estimator (Wald estimator)

43

$$y_1 = c + \beta y_2 + e, \quad \text{IV for } y_2 = z.$$

$$\hat{\beta}_{IV} = (z'y_2)^{-1} z'y_1 = \frac{\text{Cov}(y_1, z)}{\text{Cov}(y_2, z)} \quad \text{if } K=L=1$$

$$= \frac{\text{Cov}(y_1, z) / \text{Var}(z)}{\text{Cov}(y_2, z) / \text{Var}(z)} = \frac{\text{Coeff of } z \text{ from LS of } y_1 \text{ on } z}{\text{Coeff of } z \text{ from LS of } y_2 \text{ on } z}$$

$$= \frac{\hat{\beta}_1}{\hat{\beta}_2} \quad \begin{cases} \text{from } y_1 = z\beta_1 + e_{1w} \\ y_2 = z\beta_2 + e_{2w} \end{cases} \quad ; \text{ratio of two coefficients.}$$

Remark a) If there are other regressors

$$y_1 = c + \beta y_2 + \gamma X_1 + e$$

$$\hat{\beta}_{IV} = \frac{\text{Cov}(y_1, \tilde{z})}{\text{Cov}(y_2, \tilde{z})}$$

where  $\tilde{z}$  is the residuals from LS of  $z$  on  $X_1$

b) Suppose that  $z$  is a dummy variable (0, 1).

$$y_1 = c_1 + D y_1 + e_{1w} \Rightarrow \hat{\beta}_1 = \text{diff in } y_1 \text{ bet 2 groups}$$

$$y_2 = c_2 + D y_2 + e_{2w} \Rightarrow \hat{\beta}_2 = \text{" } y_2 \text{"}$$

thus

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\beta}_2} = \frac{E(y_1 | D=1) - E(y_1 | D=0)}{E(y_2 | D=1) - E(y_2 | D=0)}$$

= "Wald estimator" Wald (1940)

$$Eg) \text{ Salary} = \alpha + \beta \begin{pmatrix} \text{Veteran} \\ \text{status} \end{pmatrix} + u$$

$$IV = D = \begin{cases} 1 & \text{lottery numbers to be drafted} \\ 0 & \text{Not to be drafted} \end{cases}$$

$$\hat{\beta}_1 = -\$435.80$$

$$\hat{\beta}_2 = +\$0.159$$

$\Rightarrow D$  is purely random (exog).

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_1}{\hat{\beta}_2} = -\$2,741 !$$

# Dummy endogenous variables models

$$y_i = \alpha + \beta_1 D + \beta_2 X_i + e$$

$D = \text{endogenous}$

if  $D$  is "RANDOM",  
 OLS is fine.  
 $\therefore D$  is exogenous  $\text{cov}(D, e) = 0$

(a) If there are IVs, we may use 2SLS.

$$D = \alpha_0 + \alpha_1 Z_2 + \alpha_2 X_i + u \quad \begin{array}{l} \text{LPM as 1st stage} \\ \text{(Dummy dependent} \\ \text{variable)} \end{array}$$

$Z_2$  is IV for  $D$

$\Rightarrow$  Better way: Heckman's treatment effect models

(b) what if we do not have IVs? (no  $Z_2$  available)

then we may use a probit model.

to estimate  $P(D=1) = \Phi(X_i \hat{\gamma})$ . then

$\Leftarrow$  use  $\Phi(X_i \hat{\gamma})$  as an IV for  $D$ , and do 2SLS.

1st stage:  $P(D=1) = \Phi(X_i \hat{\gamma})$

$$D = \alpha_0 + \alpha_1 [\Phi(X_i \hat{\gamma})] + \alpha_2 X_i + u \Rightarrow \hat{D}$$

2nd stage:  $y_i = \alpha + \beta_1 \hat{D} + \beta_2 X_i + e$  gives 2SLS.

Note the following method is invalid (inconsistent)

$$y_i = \alpha + \beta_1 [\Phi(Z_i \hat{\gamma})] + \beta_2 X_i + e$$

$\hookrightarrow D$  is replaced with  $\Phi(Z_i \hat{\gamma})$

i) This assumes that the correct model does not require IVs; which is not justified.

ii) It is a generated regressor, thus std. errors need to be adjusted.

why?  
 i)  $D$  and  $\Phi(X_i \hat{\gamma})$  are highly correlated  
 ii) Identified due to nonlinearity (nonlinear solution possibly exists!)

$P(D=1 | X_i, Z)$   
 $= P(D=1 | X_i)$   
 $\therefore Z$  is not needed

Note the above procedure (b) can be possibly applied when  $y_i$  is a dummy variable.

(c) Heckman's treatment effect models

$$y_1 = \alpha + \beta_1 D + \beta_2 X_1 + e$$

$\downarrow$   
 endog:  $IV = z_2$

let  $z = (z_2, X_1)$ .

1st stage probit  $P(D=1) = \Phi(z\hat{\beta})$

2nd stage add the inverse Mills ratio

$$\hat{\lambda} = \begin{cases} \phi(z\hat{\beta}) / \Phi(z\hat{\beta}) & \text{if } D=1 \\ -\phi(z\hat{\beta}) / (1 - \Phi(z\hat{\beta})) & \text{if } D=0 \end{cases}$$

$$y_1 = \alpha + \beta_1 D + \beta_2 X_1 + c \hat{\lambda} + e$$

$H_0: c=0$  (no selection bias)

If  $c=0$  is rejected, std. errors need to be adjusted.

Note treatment effect =  $\hat{\beta}_1 + \hat{c} \frac{\phi(z\hat{\beta})}{\Phi(1-\Phi)}$

Dummy  $\downarrow$       Dummy  $\downarrow$

$$y_1 = \alpha + \beta_1 y_2 + \beta_2 X_1 + e$$

(d)  $y_1$  is a dummy var and  $y_2$  is also a dummy endogenous regressor. How about using this?

$\left. \begin{matrix} 1 \\ 0 \end{matrix} \right\} y_1 = \alpha + \beta_1 \Phi(z\hat{\beta}) + \beta_2 X_1 + e$   
 with  $z = (X_1, z_2)$

$\Rightarrow$  This is invalid: forbidden regression.

MLE is possible and efficient.

(kind of a system of equations of probit/logit models)

## Part 2

### IV Estimation

(continued)

- Other Estimators
  - GMM-IV, CUE, LIML
- 3 Results to show
  - exogeneity of ZUS
  - weak or not
  - evidence of endogeneity

# IV Estimation

"ivreg2" \*\*  
 "ivregress" (new)

$$y = X\beta + e$$

$$IV = Z$$

OLD: ivreg cstate)

## 1. Estimation methods

### ① 2SLS (IV)

$$\hat{\beta}_{2SLS} = (X'P_Z X)^{-1} X'P_Z y$$

$$(\hat{\beta} = (Z'X)^{-1} Z'y \text{ if } L=K)$$

### ② GMM-IV

$$E(g_i(\beta)) = 0 \quad \text{where } g_i(\beta) = z_i' e_i = z_i'(y_i - x_i \beta)$$

(  $Z \perp e$  )

$$\Rightarrow \bar{g}(\beta) = \frac{1}{n} Z'e$$

Find  $\beta$  to min  $n \bar{g}(\beta)' W \bar{g}(\beta)$   $W = L \times L$  weighting matrix

$$\Rightarrow \hat{\beta}_{GMM} = (X'Z W Z'X)^{-1} X'Z W Z'y \quad \dots \quad (*)$$

- what is  $\text{var}(\hat{\beta}_{GMM})$ ? what is  $W$ ?

Define  $S = \text{var}(\sqrt{n} \bar{g}(\beta)) = \frac{1}{n} E(Z'e e' Z)$

then since  $\hat{\beta}_{GMM} = \beta + (X'Z W Z'X)^{-1} X'Z W Z'e$

$$\text{var}(\sqrt{n} \hat{\beta}_{GMM}) = ( \dots )^{-1} X'Z W \cdot S \cdot W Z'X ( \dots )^{-1} \dots \quad (**)$$

... sandwich form (robust variance)

we can use

$$\hat{S} = \frac{1}{n} \sum \hat{e}_i^2 z_i' z_i \quad \hat{W} = \frac{1}{n} Z' \hat{\Omega} Z \quad \text{with } \hat{\Omega} = \begin{bmatrix} \hat{e}_1^2 & 0 \\ \vdots & \vdots \\ 0 & \hat{e}_n^2 \end{bmatrix}$$

- thus GMM estimator naturally uses hetero-skedasticity consistent variance.

- cluster-robust  $\hat{S} = \sum_{j=1}^M \hat{u}_j \hat{u}_j'$  where  $\hat{u}_j = (y_j - x_j \hat{\beta}) X' Z (Z' Z)^{-1} Z_j$

- efficient GMM estimator (it uses  $W = S^{-1}$ )

$$\hat{\beta}_{\text{EGMM}} = (X'ZS^{-1}Z'X)^{-1} X'ZS^{-1}Z'y \quad \text{using } (*)$$

$$\text{Var}(\hat{\beta}_{\text{EGMM}}) = (X'ZS^{-1}Z'X)^{-1} X'ZS^{-1}Z' \cdot S \cdot S^{-1}Z'X (X'ZS^{-1}Z'X)^{-1} \text{ using } (*)$$
$$= (X'ZS^{-1}Z'X)^{-1} \quad \text{Note if } S = \sigma^2 Z'Z,$$

we can use  $\hat{S}$  for  $S$ .

$$\hat{\beta}_{\text{EGMM}} = \hat{\beta}_{\text{OLS}}$$

- 2-step GMM estimator (W and S, which comes first?)

• To obtain  $\hat{\beta}_{\text{GMM}}$  or  $\hat{\beta}_{\text{EGMM}}$ , we need W or  $S^{-1}$ . Since it is not given until we estimate  $\hat{\beta}$  and  $\hat{e}$ , we start with using  $W = I$ . then obtain  $\hat{e}$  and compute  $\hat{S}$ . then this is the 1st-step GMM estimator. (consistent but inefficient)

• Then using  $W = \hat{S}^{-1}$ , we repeat the procedure to obtain  $\hat{\beta}$ . It is the 2nd-step GMM estimator (more efficient).  $\hat{S}$  is treated as given (since it is based on the residuals from the 1st-step estimator, but the residuals in the orthogonality condition  $g(\hat{\beta}) = \frac{1}{n} Z'\hat{e}$  are the 2nd stage residuals).

- continuously updated GMM estimator (CUE)

• Can we use the same residuals for both  $\hat{\beta}$  and  $\hat{S}$ ?

Min  $n \bar{g}(\hat{\beta})' S(\hat{\beta})^{-1} \bar{g}(\hat{\beta})$  using the same residuals  $\hat{e}$ .

$$\begin{cases} \bar{g}(\hat{\beta}) = Z'\hat{e} = Z'(y - X\hat{\beta}) \\ S(\hat{\beta}) = \frac{1}{n} Z\hat{e}Z'\hat{e} \end{cases}$$

• However, this requires numerical optimization methods (iterations like MLE..)

• Under homoskedasticity, CUE  $\approx$  2step GMM.

③ LIML (limited information maximum likelihood) estimator  
 ; used to be popular, but has not been used much.  
 Recent research suggests that the finite-sample performance  
 is better than 2SLS or GMM.

$$\hat{\beta}_{LIML} = [X'(I - \hat{k}M_Z)X]^{-1} X'(I - \hat{k}M_Z)y = \text{least variance ratio estimator}$$

where  $\hat{k}$  is the eigenvalue (min) of  
 $(Y'M_Z Y)^{-1/2} (Y'M_1 Y) (Y'M_Z Y)^{-1/2}$

Equivalently,  $\hat{k}$  minimizes

$$k = \frac{\beta' Y' M_1 Y \beta}{\beta' Y' M_Z Y \beta}$$

where  $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$   
 $M_Z = I - Z(Z'Z)^{-1}Z'$   
 $Z = (X_1, Z_2)$   
 $Z_2 \uparrow$   
 $ZV \text{ for } y_2$   
 $Y = (y_1, y_2)$

Note  $y_1 = y_2\beta_1 + X_1\beta_2 + e$   
 $\downarrow$   
 $ZV = X_2$

Note It is a special case of the k-class estimator

$$\hat{\beta}_{k\text{-class}} = (X'(I - kM_Z)X)^{-1} X'(I - kM_Z)y$$

- if  $k=1$ , it's 2SLS.
- if  $k=0$ , it's OLS.
- if  $k=\hat{k}$ , it's LIML
- if  $k = \hat{k} - d/df$ , it's Fuller's modified LIML

Note why LIML?

(\*)  $\begin{cases} y_1 - y_2\beta_2 = X_1\beta_2 + e & \dots \text{main eq for } y_1 \\ y_2 = X_1\gamma_1 + X_2\gamma_2 + v & \dots \text{reduced form for } y_2 \end{cases}$

Suppose we have  $Y = (y_1, y_2, y_3)$  where  $y_3$  are other endogenous variables. But we do not need the inf. on  $y_3$  for LIML.  
 Moreover, we do not need to have a structural form for  $y_2$ .

Note

Why minimize the variance ratio?

We may consider a FIML like estimation using MLE.  
But, it's a recursive system: see LHS of (\*)

$$(y_1, y_2) \begin{bmatrix} 1 & 0 \\ -\beta_2 & I \end{bmatrix}$$

Then, we can use SUR estimation. But we rarely use SUR or FIML estimation.

Instead, we can minimize the determinant

$$L = (2\pi)^{n/2} |\Sigma|^{-1/2} \exp[-\frac{1}{2} u' \Sigma^{-1} u]$$

where

$$|\Sigma| \equiv k |Y' M_Z Y|, \quad k = \frac{\beta' Y' M_1 Y \beta}{\beta' Y' M_Z Y \beta} \text{ as given before.}$$

Note Why eigen value?

We wish to find  $\beta$  st we minimize  $k$ .

( $Y' M_Z Y$  does not contain  $\beta$ , thus we may ignore it)

FOC

$$2 Y' M_1 Y \beta (\beta' Y' M_Z Y \beta) - 2 Y' M_Z Y \beta (\beta' Y' M_1 Y \beta) = 0$$

Dividing this by  $2 \beta' Y' M_Z Y \beta$  gives

$$Y' M_1 Y \beta - k Y' M_Z Y \beta = 0$$

Rearranging this after multiplying  $(Y' M_Z Y)^{-1/2}$  gives

$$\left( (Y' M_Z Y)^{-1/2} (Y' M_1 Y) (Y' M_Z Y)^{-1/2} - k I \right) (Y' M_Z Y)^{1/2} \beta = 0$$

$\Rightarrow$  thus  $\beta$  is the eigen vector,

$k$  is the eigen value (choose the min.)

Note However,  $\hat{\beta}_{KLM}$  and  $k$ -class estimator ASSUME homoskedasticity.

$$\text{Var}(\hat{\beta}_{KLM}) = \frac{\sigma^2}{\sigma^2} (X' (I - k M_Z) X)^{-1}$$

But a robust estimator is still available.

\* Optimization problem

One wishes to max  $\lambda = \frac{x' H x}{x' E x}$

FOC

$$\frac{\partial \lambda}{\partial x} = \frac{2 H x (x' E x) - (x' H x) 2 E x}{(x' E x)^2} = 0, \quad x' E x \neq 0$$

Dividing the numerator by  $2 x' E x$  gives

$$H x - \underbrace{\left( \frac{x' H x}{x' E x} \right)}_{\lambda} E x = 0$$

$$(H - \lambda E) x = 0$$

$$(E^{-1} H - \lambda I) x = 0$$

Then  $\lambda =$  eigenvalue of  $E^{-1} H$ ,  $x =$  eigenvector of  $E^{-1} H$

Note Canonical Correlation: Choose  $H = R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx}$ ,  $E = Z$

Note MANOVA test (G groups)

Choose  $E = \sum_{i=1}^G \sum_{t=1}^T (y_{it} - \bar{y}_{i.}) (y_{it} - \bar{y}_{i.})'$  : within variations  
(GxGT)

$H = G \sum_{i=1}^G (\bar{y}_{i.} - \bar{y}_{..}) (\bar{y}_{i.} - \bar{y}_{..})'$  : between variations  
(GxGT)

$\Rightarrow$  Wilk's test  $= \frac{|S|}{|E+H|} = \prod_{i=1}^{G \times T} \frac{1}{1 + \lambda_i}$ ,  $\lambda_i =$  eigenvalues...

$= \prod_{i=1}^{G \times T} (1 - r_i^2)$ ,  $r_i^2 = \frac{\lambda_i}{1 + \lambda_i}$  : squared canonical correlation

Note Principal Component, Discriminant Analysis

Find the max of  $\lambda = \frac{x' S x}{x' x} \Rightarrow$  FOC  $(S - \lambda I) x = 0$

P.C.:  $z_i = x_i' y$

where  $\lambda =$  sample variance ...

STATA commands

ivregress 2sls y1 (y2 = z2) x1, first  
 , first -- 2-step gmm  
 gmm , first  
 liml , first  
 gmm , igmm -- iterative gmm

Note "robust" is default. 2-step gmm is default in gmm.  
 (in 2SLS, gmm)

ivreg2 y1 (y2 = z2) x1, first .. 2sls is default  
 , first gmm2s  
 , cue robust  
 , klass(1.2)  
 , fuller(2)

Diagnostic check-up of IV results

- next →
- 2. exogeneity of IVs  $z_2 \perp u$  ( $x_2 \perp u$ )
  - 3. weak IV tests & identification test ( $\gamma_2 \neq 0$ )
  - 4. Evidence of endogeneity: Hausman test & others

$$\begin{cases} y_1 = y_2 \beta_1 + x_1 \beta_2 + u \\ y_2 = x_1 \gamma_1 + x_2 \gamma_2 + v \end{cases}$$

$\downarrow$   
 IV for  $y_2 = z_2$  ( $EX_2$ )

Note There can be more than one endogenous regressors.

eg)  $y_1 = \underbrace{y_{21} \beta_{11}}_{z_1 = x_{21}} + \underbrace{y_{22} \beta_{12}}_{z_2 = x_{22}} + x_1 \beta_2 + u$ ,  $\beta_1 = (\beta_{11}, \beta_{12})$

two reduced form equations (2 first stage)

$$\begin{cases} y_{21} = x_{11} \gamma_1 + x_{21} \gamma_{21} + x_{22} \gamma_{22} + v_1 \\ y_{22} = x_{11} \gamma_1^* + x_{21} \gamma_{21}^* + x_{22} \gamma_{22}^* + v_2 \end{cases}$$

\*\*\* IV estimation (2sls, gmm, liml, k-class)

\* ivregress

```

ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq, first
ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first
ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first igmm
ivregress liml lwage (educ = motheduc fatheduc huseduc) exper expersq, first

```

\* ivreg2

```

ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first endog(educ)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first robust
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, gmm2s first robust
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, cue first robust
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, kclass(1.2) first
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, fuller(2) first

```

Example: 2sls result (not using a robust option)

```

ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first endog(educ)

```

First-stage regression of educ:

OLS estimation

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

Report 1st stage results  
 Provide the test results for endogeneity of educ.

	Number of obs =	428
	F( 5, 422) =	63.30
	Prob > F =	0.0000
	Centered R2 =	0.4286
	Uncentered R2 =	0.9820
	Root MSE =	1.738
Total (centered) SS =	2230.196262	
Total (uncentered) SS =	70816	
Residual SS =	1274.365654	

educ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exper	.0374977	.0343102	1.09	0.275	-.0299424 .1049379
expersq	-.0006002	.0010261	-0.58	0.559	-.0026171 .0014167
motheduc	.1141532	.0307835	3.71	0.000	.0536452 .1746613
fatheduc	.1060801	.0295153	3.59	0.000	.0480648 .1640955
huseduc	.3752548	.0296347	12.66	0.000	.3170049 .4335048
_cons	5.538311	.4597824	12.05	0.000	4.634562 6.44206

Included instruments: exper expersq motheduc fatheduc huseduc

Partial R-squared of excluded instruments: 0.4258

Test of excluded instruments:

F( 3, 422) = 104.29

Note

(3-0) →  
 (3-3) →

Prob > F = 0.0000

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F( 3, 422)	P-value
educ	0.4258	0.4258	104.29	0.0000

(3-0)  
(3-3)

Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)  
 Ha: matrix has rank=K1 (identified)  
 Anderson canon. corr. N\*CCEV LM statistic Chi-sq(3)=182.22 P-val=0.0000  
 Cragg-Donald N\*CDEV Wald statistic Chi-sq(3)=317.33 P-val=0.0000

(3-2)

Weak identification test

Ho: equation is weakly identified  
 Cragg-Donald Wald F-statistic 104.29  
 See main output for Cragg-Donald weak id test critical values

(3-3)

Weak-instrument-robust inference

Tests of joint significance of endogenous regressors B1 in main equation  
 Ho: B1=0 and overidentifying restrictions are valid  
 Anderson-Rubin Wald test F(3,422)= 4.48 P-val=0.0041  
 Anderson-Rubin Wald test Chi-sq(3)=13.63 P-val=0.0035  
 Stock-Wright LM S statistic Chi-sq(3)=13.21 P-val=0.0042

(4)

Number of observations N = 428  
 Number of regressors K = 4  
 Number of instruments L = 6  
 Number of excluded instruments L1 = 3

IV (2SLS) estimation

Estimates efficient for homoskedasticity only  
 Statistics consistent for homoskedasticity only

		Number of obs =	428
		F( 3, 424) =	11.52
		Prob > F =	0.0000
Total (centered) SS =	223.3274513	Centered R2 =	0.1495
Total (uncentered) SS =	829.594813	Uncentered R2 =	0.7711
Residual SS =	189.9347086	Root MSE =	.6662

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
educ	.0803918	.021672	3.71	0.000	.0379155	.1228681
exper	.0430973	.0132027	3.26	0.001	.0172204	.0689742
expersq	-.0008628	.0003943	-2.19	0.029	-.0016357	-.0000899
_cons	-.1868574	.2840591	-0.66	0.511	-.743603	.3698883

(3-2) Underidentification test (Anderson canon. corr. LM statistic): 182.225  
 Chi-sq(3) P-val = 0.0000

(3-3) Weak identification test (Cragg-Donald Wald F statistic): 104.294  
 Stock-Yogo weak ID test critical values: 5% maximal IV relative bias 13.91  
 10% maximal IV relative bias 9.08  
 20% maximal IV relative bias 6.46  
 30% maximal IV relative bias 5.39  
 10% maximal IV size 22.30  
 15% maximal IV size 12.83  
 20% maximal IV size 9.54  
 25% maximal IV size 7.80

Source: Stock-Yogo (2005). Reproduced by permission.

(2) Sargan statistic (overidentification test of all instruments): 1.115  
 Chi-sq(2) P-val = 0.5726

(4) -endog- option:  
 Endogeneity test of endogenous regressors: 2.746  
 Chi-sq(1) P-val = 0.0975

Regressors tested: educ  
 Instrumented: educ  
 Included instruments: exper expersq  
 Excluded instruments: motheduc fatheduc huseduc

. estimates store ivreg2

overid, all // Sargan, Hansen and others

(2) Tests of overidentifying restrictions:  
 Sargan N\*R-sq test 1.115 Chi-sq(2) P-value = 0.5726  
 Sargan (N-L)\*R-sq test 1.105 Chi-sq(2) P-value = 0.5756  
 Basman test 1.102 Chi-sq(2) P-value = 0.5763  
 Sargan pseudo-F test 0.552 F(2,424) P-value = 0.5760  
 Basman pseudo-F test 0.551 F(2,422) P-value = 0.5767

(We will discuss about each of these results below.)

## 2. Exogeneity of IVs

Non-rejection is preferred.

(1) Over-identification restriction test ( $H_0$ : IVs are valid (exogenous))

- Sargan's  $NR^2$  test (when 2SLS is used)

Regress 2SLS residuals on all exog variables ( $x_1$  &  $x_2$ ) and compute  $NR^2 \sim K^2_{L-K}$  ;  $\hat{u} = z\gamma + e \Rightarrow NR^2$

$L = \#$  of IVs

$K = \#$  of regressors

$L-K = \#$  of extra IVs

$H_0$ : IVs are valid ( $z_2 \perp u$ ) ; IVs are exogenous.

Note If  $L=K$  (exactly identified), this test is not given (not possible to test exogeneity of IVs if  $L=K$ )

Note Basman test

$$\text{Basman} = NR^2 \cdot \left( \frac{N - K_2}{N - NR^2} \right) \sim K^2_{L-K}$$

↑ correction term in finite samples

Note Both Sargan's and Basman's test assume homoskedasticity. If robust option is used, these cannot be used.

However, Hansen's J-statistic can be used if gmm is used. It is robust to heteroskedasticity.

- Wooldridge's robust score test (when 2SLS is used) ... robust to heteroskedasticity

Its details are skipped, but this can be used after robust option. Score test  $\sim K^2_{L-K}$

eg) regress 2SLS  $y_1$  ( $y_2 = z_2$ )  $x_1$ , robust estat overid

If  $H_0$  is rejected, IVs are not exogenous (not valid)

- Hausen's J-statistic (gmm is used)

$$J = \left( \frac{1}{N} \sum z_i \hat{u}_i \right)' W \left( \frac{1}{N} \sum z_i \hat{u}_i \right) \quad \text{with } z = (z_1, z_2)$$

$$\sim \chi^2_{L-k} \quad (W = \hat{J}^{-1} \text{ can be used...})$$

Note if  $L=k$ ,  $J=0$  (thus, not possible to test exogeneity of ZVs if  $L=k$ )

- Anderson-Rubin test of over-identification restrictions (LIML is used)

$$\text{Anderson-Rubin} = N(\hat{\lambda} - 1) \quad \text{where } \hat{\lambda} = \text{min. eigenvalue}$$

$$\sim \chi^2_{L-k}$$

Note  $N(\hat{\lambda} - 1) \approx N \ln(\hat{\lambda})$  if  $\hat{\lambda}$  is near 1

Basmann's F-stat (if LIML is used)

$$B_F = (\hat{\lambda} - 1)(N - k_2) / (L - k) \sim F(m, N - k_2)$$

(2) Over-identification restrictions test on a subset of IVs

C-test = difference in Sargan tests =  $J_u - J_R \sim \chi^2_m$   
( $m = \#$  of suspect IVs)

eg) IVs = mother-edu, father-edu, hus-edu

suppose: mother-edu & father-edu are questionable.

[ ivreg2 lwage (educ = mother-edu . father-edu hus-edu)  
expvar expvars, ortho(mother-edu father-edu)

$H_0$ : suspect IVs are valid.

$\Rightarrow$  If  $H_0$  is rejected, suspect IVs are not valid.

Note this also reports J-stat for which suspect ZVs are dropped.

If  $H_0$  is rejected, ZVs are not exogenous (not valid)



\*\*\* Overidentification restriction tests

```
. *2sls
. ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. estat overid // Sargan's test and Basman's test
```

Tests of overidentifying restrictions:

Sargan (score) chi2(2) = 1.11504 (p = 0.5726)  
 Basman chi2(2) = 1.10228 (p = 0.5763)

```
. *robust option
. ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq,
vce(robust)
. estat overid // Woodridge robust score test
```

Test of overidentifying restrictions:

Score chi2(2) = 1.04213 (p = 0.5939)

```
. *2-step gmm
. ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. estat overid // Hansen's J-statistic
```

Test of overidentifying restriction:

Hansen's J chi2(2) = 1.04213 (p = 0.5939)

```
. *iterative gmm
. ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first igmm
. estat overid // Hansen's J-statistic
```

Test of overidentifying restriction:

Hansen's J chi2(2) = 1.04124 (p = 0.5942)

```
. *liml
. ivregress liml lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. estat overid // Anderson-Rubin test and Basman's F-test
```

Tests of overidentifying restrictions:

Anderson-Rubin chi2(2) = 1.1179 (p = 0.5718)  
 Basman F(2, 428) = .558948 (p = 0.5722)

\*\*\* over-identification restriction test on a subset of IVs

```
. *testing exogeneity of excluded instruments
. ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(fatheduc)
```

-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions): 1.111
Chi-sq(1) P-val = 0.2918
C statistic (exogeneity/orthogonality of suspect instruments): 0.004
Chi-sq(1) P-val = 0.9515
Instruments tested: fatheduc

-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions): 0.305
Chi-sq(1) P-val = 0.5808
C statistic (exogeneity/orthogonality of suspect instruments): 0.810
Chi-sq(1) P-val = 0.3681
Instruments tested: motheduc

-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions): 0.384
Chi-sq(1) P-val = 0.5354
C statistic (exogeneity/orthogonality of suspect instruments): 0.731
Chi-sq(1) P-val = 0.3926
Instruments tested: huseduc

-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions): 0.000
Chi-sq(0) P-val = .
C statistic (exogeneity/orthogonality of suspect instruments): 1.115
Chi-sq(2) P-val = 0.5726
Instruments tested: motheduc fatheduc

-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions): 0.000
Chi-sq(0) P-val = .
C statistic (exogeneity/orthogonality of suspect instruments): 1.115
Chi-sq(2) P-val = 0.5726
Instruments tested: fatheduc huseduc

-----
Instrumented: educ
Included instruments: exper expersq
Excluded instruments: motheduc fatheduc huseduc
-----

### 3. "Weak IV" tests

(also identification tests)

Consider

$$y_1 = \gamma_2 \beta_1 + x_1 \beta_2 + u$$

endo:  $IV = z_2$       Thus  $z = (x_1, z_2)$

[  $x_1$ : included exogenous regressors

$z_2$ : excluded exogenous (IV)

(Thus called, "exclusion restrictions")

:  $z_2$  does not belong in the structural eq. for  $y_1$

The first stage regression implies

$$y_2 = x_1 \gamma_1 + z_2 \gamma_2 + v \quad (z_2 = \text{excluded instruments})$$

If  $H_0: \gamma_2 = 0$  is rejected,  $z_2$  is not weak.  
(F-test)

Note

:) The distribution of the F-test can be non-standard  
: Stock & Yogo (2002) suggest that  $F^* > 10$   
is a guideline, if there is only 1  
endogenous regressor.

(regardless of # of IVs?)

:) What if there are more than 1 endo. regressor?

$$y_2 = (y_{21}, y_{22})', \quad IVs = (z_{21}, z_{22})' \text{ for mem.}$$

then, there are two 1st-stage regressions.

$$\begin{cases} y_{21} = x_1 \gamma_{11} + z_{21} \gamma_{21} + z_{22} \gamma_{31} + v_1 \Rightarrow H_0: \gamma_{21} = \gamma_{31} = 0 \\ y_{22} = x_2 \gamma_{12} + z_{21} \gamma_{22} + z_{22} \gamma_{32} + v_2 \Rightarrow H_0: \gamma_{22} = \gamma_{32} = 0 \end{cases}$$

$\Rightarrow$  there can be two F-tests; not convenient.

∴) Suppose that  $z_{21}$  is not weak, but  $z_{22}$  is weak (not strongly correlated with  $y_{22}$ )  
If so, F-test in  $H_0: \alpha_2 = \alpha_3 = 0$  will be still rejected due to  $\alpha_2 \neq 0$  ( $z_{21}$  is not weak).

But, we have only 1 valid IV ( $z_{21}$ ), while we have 2 endogenous regressors ( $y_{21}, y_{22}$ ) if  $z_{22}$  is weak.

⇒ thus, F-tests can be less informative if there are more than 1 endogenous regressor.

There are two sets of tests on the validity of IVs in the 1st stage regression.

∴) (under)identification tests :  $H_0$ : not-identified (under)

∴) weak IV tests :  $H_0$ : IVs are weak

⇒ thus, if  $H_0$  is not rejected, IVs are not valid in the sense of the 1st stage predictability.

[Overview on the tests in the 1st stage regression]

- Some statistics on the explanatory power ( $R^2$ ) of  $z_2$

- Under-identification tests

• Assuming homo-skedasticity

↳ Anderson's canonical correlation LM tests

↳ Cragg-Donald Wald tests

• Robust to Heteroskedasticity

Kleibergen-Paap rank test

- Weak IV tests & Redundancy tests

⇒ use Stock-Jogo's crit. values

- Weak IV robust inference

(1) Some statistics on the explanatory power of  $Z_2$

- Partial  $R^2 = R^2$  bet  $y_2$  &  $Z_2$  after controlling the effect of  $X_1 = R^2$  bet  $\tilde{y}_2$  and  $\tilde{Z}_2$   
 where  $\tilde{y}_2 = M_1 y_2, \tilde{Z}_2 = M_1 Z_2, M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
- F-test in  $\gamma_2 = 0$  (use robust variance, if available)  
 (Stock & Yogo suggested that  $F^* > 10$ )  
 (But, as noted previously, some limitations exist.)
- Shea's partial  $R^2 = R^2$  from the regression of  $\tilde{y}_1$  on  $\tilde{Z}_1$   
 where  $\tilde{y}_1 =$  residuals from LS of  $y_1$  on  $X_1$  and other endogenous regressors (say  $y_3$ )  
 $\tilde{Z}_1 =$  residuals from LS of  $\hat{y}_1$  (1st stage fitted value) on  $X_1$  and  $\hat{y}_3$  (1st stage fitted value for  $y_3$ )

(2) Under-identification tests

- Assume iid errors (homoskedasticity)
- Assume non-iid " (hetero, cluster, ...)
- (:) LM test of Anderson's canonical correlation

this uses the minimum eigenvalue of  $Q^*$  (below)  
 (canonical correlation coeff =  $\sqrt{\text{eigenvalue}}$ )

1st stage regression is

$$y_2 = X_1 \gamma_1 + Z_2 \gamma_2 + v \quad \text{or} \quad \tilde{y}_2 = \tilde{Z}_2 \gamma_2 + \tilde{v}$$

$$\text{or } \tilde{y}_2 = \tilde{Z}_2 \gamma_2 + v \quad \text{where } \tilde{y}_2 = M_1 y_2, \tilde{Z}_2 = M_1 Z_2$$

(residuals eliminating the effect of  $X_1$ )

Let  $R_{yy} = \hat{y}_2' \hat{y}_2$  ,  $R_{zz} = \hat{z}_2' \hat{z}_2$  ,  $R_{yz} = \hat{y}_2' \hat{z}_2$  ,  $R_{zy} = R_{yz}'$

$Q^* = R_{zz}^{-1} R_{zy} R_{yy}^{-1} R_{yz}$

$\Rightarrow$  Find eigen values. Minimum eigen values =  $\lambda$  let  $\lambda$

$LM = N \hat{\lambda} \sim \chi^2_{L-k+1}$  (  $L = \#$  of IVs ,  $L_2 = \#$  of  $z_2$   
(  $k = \#$  of regressors ,  $k_2 = \#$  of  $y_2$  )

(Note canonical correlation coeff =  $\sqrt{\hat{\lambda}}$  )  
 ; cov. coeff among multiple variables.

$H_0$ : under-identified system

(if  $z_2$  is not good, rank of  $Q^*$  matrix will be less than full rank; singular matrix if one eigen value is close to 0)

$\Rightarrow$  If  $H_0$  is not rejected, the system is under-identified (not-identified), and IVs are not valid.

eg) endog regressor: educ ( $k_2=1$ )  
 excluded IVs: mother-edu, father-edu, hwr-edu ( $L_2=3$ )

Anderson LM stat = 182.22  $\sim \chi^2_{3-1+1}$  (df=3)  
 p-value = 0.000

thus, IVs are valid (not under-identified).

(::) Crags-Donald Wald stat

CD Wald =  $N \cdot \left( \frac{\hat{\lambda}}{1-\hat{\lambda}} \right) \sim \chi^2_{L-k+1}$

eg) CD Wald = 317.33  $\sim \chi^2_{3-1+1}$  p-value = 0.000

thus, not under-identified.

Note Anderson's LM and Cragg-Donald tests 13  
 Assume iid errors, they are not valid  
 under heteroskedasticity (also cluster, ...)

(iii) Kleibergen - Paap (2006) rank statistic

: robust to non-iid error (thus these are computed when robust is used)

"rk" statistic is based on "ranktest" module in stata

there are two different versions

$$\begin{cases} \text{KP-LM} \sim \chi^2_{L-k+1} \\ \text{KP-Wald} \sim \chi^2_{L-k+1} \end{cases}$$

: Interpretation of the results would be the same as before

eg) KP-LM stat = 106.70  $\sim \chi^2_{3-1+1}$  p-value = 0.000

KP-Wald stat = 324.42  $\sim \chi^2_3$  p-value < 0.000

thus the equation is not under-identified.  
 (IVs are valid).

(3) Weak IV tests

(i) Cragg-Donald (Wald) F-statistic

$$\text{CD F-stat} = \frac{N-L}{L_1} \cdot \frac{\hat{\lambda}}{1-\hat{\lambda}} \Leftrightarrow \left( \frac{N-L}{N} \right) \left( \frac{1}{L_1} \right) \cdot N \cdot \frac{\hat{\lambda}}{1-\hat{\lambda}}$$

$\uparrow$  df adjustment       $\uparrow$  CD-wald  
 by dividing by  $L_1$   
 it becomes  
 F-stat.

$$\begin{cases} L = \# \text{ of IVs } (X_1, Z_2) \\ L_1 = \# \text{ of excluded IVs } (Z_2) \\ N = \# \text{ of obs} \end{cases}$$

But the distribution of CDF-stat is non-standard. 14

Stock & Yogo (2005) provided crit. values

for IV, LIML and Fuller-LIML estimators

; STATA reports these crit. values, assuming iid error  
(ivreg2)

i) 2SLS relative bias critical values (5, 10, 20%...)

$m = \#$  of endog. regressors  
 $L_2 = \#$  of excluded IVs

For each pair of  $(m, L_2)$ ,  $m < L_2$

eg) 2SLS relative bias (compared to OLS)

5% crit value = 13.91 if  $m=1, L_2=3$

Min eigen value stat = 104.294

F-stat (robust) = 106.623

F-stat (non-robust) = 104.294 = min eigen stat

since all of these stats  $> 13.91$  (crit. value) (if  $m=1$ )

we reject  $H_0$ : IVs are weak.

$\Rightarrow$  IVs are valid (not weak).

Note when  $m=1$ , min eigen value stat = F-stat

Note when robust option is used, the crit. values of Stock & Yogo (2005) are not used. But they are still reported with forcenonrobust option, and they can be used regardless..

ii) 2SLS size of nominal 5% Wald test

; size distribution of Wald tests

$(m, L_2)$  pairs,  $m \leq L_2$

crit. values.



## (4) Redundancy tests

check if some IVs are redundant (weak subsets)

$$y_1 = y_2 \beta_1 + X_1 \beta_2 + u$$

$$\downarrow$$

$$\rightarrow IV = (z_{21}, z_{22})$$

$$y_2 = z_{21} \gamma_{11} + \underbrace{z_{22} \gamma_{22}} + X_1 \gamma_3 + v \quad \text{1st stage}$$

wish to check if  $H_0: \gamma_{22} = 0$  holds.

If  $H_0$  is not rejected,  $z_{22}$  is redundant.

$\Rightarrow$  Hall & Peixe (2003) suggested the LR version (LM) test  $\sim \chi^2_{\dim \gamma_{22}}$

(g)  $IV = \text{moth-edu}, \text{father-edu}, \text{hus-edu}$

redundancy	LM stat (p-val)
moth-edu	12.941 (0.0003)
father-edu	12.771 (0.0004)
hus-edu	68.874 (0.000)
moth & father-edu	42.572 (0.000)

$\Rightarrow$  Since  $H_0$  is rejected in all cases, IVs are valid.

\*\*\* First Stage, Identification and WEAK IV tests

(Robust 2sls)

\*\* ivregress

```
. ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq,
vce(robust)
. estat firststage, forcenonrobust
```

(3-1)

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(3,422)	Prob > F
educ	0.4286	0.4218	0.4258	106.623	0.0000

The critical values of Stock & Yogo are valid only with iid errors, but this command forces to use them despite of the robust option.

(3-2)

Minimum eigenvalue statistic = 104.294

Critical Values # of endogenous regressors: 1 ←  $H_0$   
 Ho: Instruments are weak # of excluded instruments: 3 ←  $H_2$

	5%	10%	20%	30%
2SLS relative bias	13.91	9.08	6.46	5.39
2SLS Size of nominal 5% Wald test	22.30	12.83	9.54	7.80
LIML Size of nominal 5% Wald test	6.46	4.36	3.69	3.32

← We reject  $H_0$  of under-identification since 104.294 is greater than crit. values.

```
. ivregress 2sls lwage (educ = motheduc) exper expersq, first robust
. estat firststage, forcenonrobust
```

(3-1)

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,424)	Prob > F
educ	0.1527	0.1467	0.1485	71.2531	0.0000

Minimum eigenvalue statistic = 73.9459

(3-2)

Critical Values # of endogenous regressors: 1  
 Ho: Instruments are weak # of excluded instruments: 1

	5%	10%	20%	30%
2SLS relative bias	(not available)			
	10%	15%	20%	25%

```

2SLS Size of nominal 5% Wald test | 16.38    8.96    6.66    5.53
LIML Size of nominal 5% Wald test | 16.38    8.96    6.66    5.53
-----

```

**\*\* ivreg2**

```

. ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first robust
endog(educ)

```

First-stage regressions

First-stage regression of educ:

OLS estimation

Estimates efficient for homoskedasticity only  
 Statistics robust to heteroskedasticity

*(But still we use the crit. values  
 which may be valid only with  
 iid errors)*

Total (centered) SS = 2230.196262  
 Total (uncentered) SS = 70816  
 Residual SS = 1274.365654

Number of obs = 428  
 F( 5, 422) = 64.93  
 Prob > F = 0.0000  
 Centered R2 = 0.4286  
 Uncentered R2 = 0.9820  
 Root MSE = 1.738

	educ	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
exper		.0374977	.0345099	1.09	0.278	-.030335	.1053305
expersq		-.0006002	.0010514	-0.57	0.568	-.0026669	.0014665
motheduc		.1141532	.0300915	3.79	0.000	.0550053	.1733011
fatheduc		.1060801	.0284825	3.72	0.000	.0500948	.1620654
huseduc		.3752548	.0344659	10.89	0.000	.3075087	.443001
_cons		5.538311	.4617617	11.99	0.000	4.630672	6.44595

Included instruments: exper expersq motheduc fatheduc huseduc

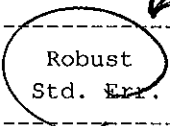
Partial R-squared of excluded instruments: 0.4258

Test of excluded instruments:

F( 3, 422) = 106.62  
 Prob > F = 0.0000

Summary results for first-stage regressions

Variable	Shea Partial R2	Partial R2	F( 3, 422)	P-value
educ	0.4258	0.4258	106.62	0.0000



(3-1)

NB: first-stage F-stat heteroskedasticity-robust

When Robust variance is used (hetero, cluster, autocorrelation,,)

(3-2) Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)

Ha: matrix has rank=K1 (identified)

Kleibergen-Paap rk LM statistic Chi-sq(3)=106.70 P-val=0.0000

Kleibergen-Paap rk Wald statistic Chi-sq(3)=324.42 P-val=0.0000

Note: When Robust variance is not used (iid error); copied from the previous results

(3-2) Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)

Ha: matrix has rank=K1 (identified)

Anderson canon. corr. N\*CDEV LM statistic Chi-sq(3)=182.22 P-val=0.0000

Cragg-Donald N\*CDEV Wald statistic Chi-sq(3)=317.33 P-val=0.0000

(3-3) Weak identification test

Ho: equation is weakly identified

Kleibergen-Paap Wald rk F statistic 106.62

See main output for Cragg-Donald weak id test critical values

(4) Weak-instrument-robust inference

Tests of joint significance of endogenous regressors B1 in main equation

Ho: B1=0 and overidentifying restrictions are valid

Anderson-Rubin Wald test F(3,422)= 4.53 P-val=0.0039

Anderson-Rubin Wald test Chi-sq(3)=13.79 P-val=0.0032

Stock-Wright LM S statistic Chi-sq(3)=12.62 P-val=0.0055

NB: Underidentification, weak identification and weak-identification-robust test statistics heteroskedasticity-robust

Number of observations	N =	428
Number of regressors	K =	4
Number of instruments	L =	6
Number of excluded instruments	L1 =	3

IV (2SLS) estimation

Estimates efficient for homoskedasticity only

Statistics robust to heteroskedasticity

		Number of obs =	428	
		F( 3, 424) =	9.19	
		Prob > F =	0.0000	
Total (centered) SS	=	223.3274513	Centered R2 =	0.1495
Total (uncentered) SS	=	829.594813	Uncentered R2 =	0.7711
Residual SS	=	189.9347086	Root MSE =	.6662

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
lwage	.0803918	.0216016	3.72	0.000	.0380533	.1227302
exper	.0430973	.0152347	2.83	0.005	.0132378	.0729568
expersq	-.0008628	.0004197	-2.06	0.040	-.0016854	-.0000402
_cons	-.1868574	.2998514	-0.62	0.533	-.7745554	.4008407

(3-2) Underidentification test (Kleibergen-Paap rk LM statistic): 106.698  
 Chi-sq(3) P-val = 0.0000

(3-3) Weak identification test (Kleibergen-Paap rk Wald F statistic): 106.623  
 Stock-Yogo weak ID test critical values: 5% maximal IV relative bias 13.91  
 10% maximal IV relative bias 9.08  
 20% maximal IV relative bias 6.46  
 30% maximal IV relative bias 5.39  
 10% maximal IV size 22.30  
 15% maximal IV size 12.83  
 20% maximal IV size 9.54  
 25% maximal IV size 7.80  
 Source: Stock-Yogo (2005). Reproduced by permission.  
 NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.

(2) Hansen J statistic (overidentification test of all instruments): 1.042  
 Chi-sq(2) P-val = 0.5939

(4) -endog- option:  
 Endogeneity test of endogenous regressors: 2.976  
 Chi-sq(1) P-val = 0.0845

Regressors tested: educ  
 Instrumented: educ  
 Included instruments: exper expersq  
 Excluded instruments: motheduc fatheduc huseduc

16-5

# (Redundancy tests)

```
. ** redundancy test
. ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust
redundant(fatheduc)

-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    12.771
                                                                    Chi-sq(1) P-val =    0.0004
Instruments tested:    fatheduc

-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    12.941
                                                                    Chi-sq(1) P-val =    0.0003
Instruments tested:    motheduc

-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    68.874
                                                                    Chi-sq(1) P-val =    0.0000
Instruments tested:    huseduc

-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    42.452
                                                                    Chi-sq(2) P-val =    0.0000
Instruments tested:    motheduc fatheduc

-redundant- option:
IV redundancy test (LM test of redundancy of specified instruments):    94.391
                                                                    Chi-sq(2) P-val =    0.0000
Instruments tested:    fatheduc huseduc
```

## 4. Testing for Endogeneity

17

i) Hausman test (Wald)

$$Wald = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})' [\text{Var}(\hat{\beta}_{2SLS}) - \text{Var}(\hat{\beta}_{OLS})]^{-1} (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})$$

: see earlier note (lecture note for EC670)

ii) Residual based test

$$y_1 = y_2 \beta_1 + x_1 \beta_2 + u$$

$$y_2 = z_2 \gamma_1 + x_1 \gamma_2 + v \Rightarrow \text{obtain the residuals } \hat{v}$$

$$y_1 = y_2 \beta_1 + x_1 \beta_1 + \rho \hat{v} + u$$

$H_0: \rho = 0$  (no endogeneity)  $\Rightarrow$  usual F-test (robust option is better)

$\Rightarrow$  If  $H_0$  is rejected, there is an issue of endogeneity.

Stata , endog (educ)  
, orthog (educ)

iii) "Weak-IV robust Inference"

Consider

$$y_1 = \gamma_2 \beta_1 + X_1 \beta_2 + u$$

$$\gamma_2 = z_2 \delta_1 + X_1 \delta_2 + v$$

$$\Rightarrow y_1 = z_2 \gamma_1 + X_1 (\delta_2 + \beta_2) + u$$
  
$$= z_2 \alpha_1 + X_1 \alpha_2 + u \quad \dots \text{(1st stage of } y_1 \text{ (not } \gamma_2 \text{))}$$

$$H_0: \alpha_1 = 0$$

"Tests of joint sig of endog regressors  $\beta_1$  in main equation"  $H_0: \beta_1 = 0$

3 tests

- Anderson-Rubin Wald stat  $\sim \chi^2_{L-k+1}$
- " " F-stat  $\sim F_{L, L-k+1}$
- Stock-Wright LM S stat  $\sim \chi^2_{L-k+1}$

$H_0: \beta_1 = 0$  (no endogeneity of  $\gamma_2$ )  
(but using the set of  $z_2$ )

$\Rightarrow$  If  $H_0$  is rejected,  $\gamma_2$  is endogenous.

5. OTHER TESTS

- Reset test : `ivreset`  $H_0$ : correctly specified.
- Heteroskedasticity tests in IV estimation : `ivhettest`  $H_0$ : no heteroskedasticity

## \*\*\* Endogeneity tests (Hausman type)

```
. *2sls
. ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. estat endogenous
```

## Tests of endogeneity

Ho: variables are exogenous

```
Durbin (score) chi2(1)          = 2.74613 (p = 0.0975)
Wu-Hausman F(1,423)            = 2.73157 (p = 0.0991)
```

```
. *robust option
. ivregress 2sls lwage (educ = motheduc fatheduc huseduc) exper expersq,
vce(robust)
. estat endogenous
```

## Tests of endogeneity

Ho: variables are exogenous

```
Robust score chi2(1)           = 3.13828 (p = 0.0765)
Robust regression F(1,423)     = 3.2177 (p = 0.0736)
```

```
. *2-step gmm
. ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. estat endogenous
```

## Test of endogeneity (orthogonality conditions)

Ho: variables are exogenous

GMM C statistic chi2(1) = 2.97627 (p = 0.0845)

```
.
. *iterative gmm
. ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first igmm
. estat endogenous
```

## Test of endogeneity (orthogonality conditions)

Ho: variables are exogenous

GMM C statistic chi2(1) = 2.97627 (p = 0.0845)

```
. *liml
. ivregress liml lwage (educ = motheduc fatheduc huseduc) exper expersq, first
. *estat endogenous // not available for liml
```

(RESET specification test)

```
. *iv RESET specification test
. ivreset
Ramsey/Pesaran-Taylor RESET test
Test uses square of fitted value of y (X-hat*beta-hat)
Ho: E(y|X) is linear in X
Wald test statistic:           Chi-sq(1) = 0.00   P-value = 0.9629
```

(IV heteroskedasticity test)

```
. *IV heteroskedasticity test(s) using fitted value
. ivhettest, fitlev
IV heteroskedasticity test(s) using fitted value (X-hat*beta-hat)
Ho: Disturbance is homoskedastic
Pagan-Hall general test statistic : 4.640 Chi-sq(1) P-value = 0.0312
```

capture log close  
clear all  
set more off

\* Reading the data  
log using 2s1s\_new.log, replace  
use 2s1s\_mroz\_428.dta

\* ----- Part 1 (your own and ivreg) -----

\*\* OLS (1)  
regress lwage educ exper expersq

\*\* 2SLS (2)  
ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq

\* My own 2SLS

\* 1st stage  
regress educ exper expersq motheduc fatheduc huseduc  
predict edu\_pre, xb  
predict edu\_res, res  
test motheduc fatheduc huseduc

\* 2nd stage (3)  
regress lwage edu\_pre exper expersq

\* Consider these, too  
\* (4)  
regress lwage edu\_pre exper expersq edu\_res

\* (5)  
regress lwage educ exper expersq edu\_res

\* (6)  
regress lwage educ exper expersq edu\_pre

\* testing for endogeneity (Hausman test)

\* (7)  
regress lwage educ exper expersq edu\_pre  
test edu\_pre

\* (8)  
regress lwage educ exper expersq edu\_res  
test edu\_res

2sls\_new.do

```

*** Hausman test (using Wald type test)
regress lwage educ exper expersq
est store ols

```

```

ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
hausman ols .

```

```

ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
est store keep

```

```

regress lwage educ exper expersq
hausman keep .

```

```

*** Hausman test (using IVENDOG)

```

```

ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
ivendog

```

```

ivreg lwage (educ exper = motheduc fatheduc huseduc)
ivendog

```

```

*** F-test for weak IVs

```

```

ivreg lwage (educ = motheduc fatheduc huseduc)
regress educ exper expersq motheduc fatheduc huseduc
test motheduc fatheduc huseduc

```

```

ivreg lwage (educ exper = motheduc fatheduc huseduc)
regress educ exper expersq motheduc fatheduc huseduc
test motheduc fatheduc huseduc
regress exper expersq motheduc fatheduc huseduc
test motheduc fatheduc huseduc

```

```

*** Overidentification restriction test

```

```

ivreg lwage (educ = motheduc fatheduc huseduc) exper expersq
overid

```

```

* ivreg lwage (educ exper = motheduc fatheduc huseduc) expersq
overid

```

```

* ----- Part 2 (ivregress; new command from STATA) -----

```

```

**** using ivregress (new command extended from ivreg)

```

N

```

*2s1s
ivregress 2s1s lwage (educ = motheduc fatheduc huseduc) exper expersq, first
estimates store endo_2s1s
estat overid // Sargan's test and Basmann's test
estat endogenous
estat firststage, forcenonrobust

*robust option
ivregress 2s1s lwage (educ = motheduc fatheduc huseduc) exper expersq, vce(robust)
estat overid // Woodridge robust score test
estat endogenous
estat firststage, forcenonrobust

*2-step gmm
ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first
estimates store endo_gmm
estat overid // Hansen's J-statistic
estat endogenous
estat firststage, forcenonrobust

*iterative gmm
ivregress gmm lwage (educ = motheduc fatheduc huseduc) exper expersq, first igmm
estimates store endo_igmm
estat overid // Hansen's J-statistic
estat endogenous
estat firststage, forcenonrobust

*liml
ivregress liml lwage (educ = motheduc fatheduc huseduc) exper expersq, first
estimates store endo_liml
estat overid // Anderson-Rubin test and Basmann's F-test
*estat endogenous // not available for liml
estat firststage, forcenonrobust

*making a summary table
estimates table endo_2s1s endo_gmm endo_igmm endo_liml, b(%7.3f) t(%6.2f) stat(N rmse r2 r2_a)

*Finding the min eigenvalue statistic for each of IVs
*2s1s
ivregress 2s1s lwage (educ = motheduc) exper expersq, first robust
estat firststage, forcenonrobust
ivregress 2s1s lwage (educ = fatheduc) exper expersq, first robust
estat firststage, forcenonrobust
ivregress 2s1s lwage (educ = huseduc) exper expersq, first robust
estat firststage, forcenonrobust
ivregress 2s1s lwage (educ = motheduc fatheduc) exper expersq, first robust
estat firststage, forcenonrobust

```

```

* ----- Part 3 (ivreg2, extra module) -----
**** using ivreg2 (new command extended from ivreg or ivregress)

** various estimation methods along with
* (i) first stage identification tests and
* (ii) over-identification tests

*ivreg2 (with weak IV tests) // Hansen's J-stat
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first endog(educ)
estimates store ivreg2
overid, all // Sargan, Hansen and others

ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, first robust endog(educ)
estimates store ivreg2_robust

*2-step gmm // Hansen's J-stat
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, gmm2s first robust endog(educ)
estimates store ivreg2_gmm2s

*cue (continuously updated gmm estimator) // Hansen's J-stat
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, cue first robust endog(educ)
estimates store ivreg2_cue

*kclass // Hansen's J-stat
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, kclass(1.2) first robust endog(educ)
estimates store ivreg2_kclass

*Fuller // Hansen's J-stat
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, fuller(2) first robust endog(educ)
estimates store ivreg2_fuller

**making a summary table
estimates table ivreg2 ivreg2_robust ivreg2_gmm2s ivreg2_cue ivreg2_kclass ivreg2_fuller, b(%7.3f) t(%6.2f)

** K-P test on each or subsets of three IVs (2sls)

ivreg2 lwage (educ = motheduc) exper expersq, first robust
ivreg2 lwage (educ = fatheduc) exper expersq, first robust
ivreg2 lwage (educ = huseduc) exper expersq, first robust
ivreg2 lwage (educ = motheduc) exper expersq, first robust
ivreg2 lwage (educ = fatheduc) exper expersq, first robust
ivreg2 lwage (educ = huseduc) exper expersq, first robust

** over-identification restriction test on a subset of IVs

```

3

```

**testing exogeneity of excluded instruments
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(fatheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(motheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(huseduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(motheduc fatheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(fatheduc huseduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, orthog(motheduc fatheduc huseduc)

** testing endogeneity (exogeneity) of a regressor, educ
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust endog(educ)
ivreg2 lwage educ ( = motheduc fatheduc huseduc) exper expersq, robust orthog(educ)

** redundancy test
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust redundant(fatheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust redundant(motheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust redundant(huseduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust redundant(motheduc fatheduc)
ivreg2 lwage (educ = motheduc fatheduc huseduc) exper expersq, robust redundant(fatheduc huseduc)

*iv RESET specification test
ivreset

*IV heteroskedasticity test(s) using fitted value
ivhettest, fitlev

```