

Using "Proxy"

In the classic errors-in-variables problem I told you that LS is biased and inconsistent if the endogenous variable is measured with error.

$$y = \beta_1 + \beta_2 x_2^* + u.$$

x_2^* = unobserved, replaced
by x_2

$$= \beta_1 + \beta_2 x_2 + \beta_2 x_2^* - \beta_2 x_2 + u.$$

$$= \beta_1 + \beta_2 x_2 + \beta_2 (x_2^* - x_2) + u.$$

Since $\text{Cov}(x_2, x_2^*) \neq 0$ LS inconsistent
for β_2 .

Suppose

$$y = \beta_1 + \beta_2 \gamma_2^* + \beta_3 \gamma_3 + u.$$

and interest is in estimating
 β_3 . Again γ_2^* not observed.

Suppose I have a proxy for γ_2^*

$$\tilde{\gamma}_2^* = s_1 + s_2 \gamma_2 + v$$

which is correlated with γ_2^* .

s_1 is included because there
 is a difference in scale.

v is included because proxy is NOT
 perfect.

$$y = \beta_1 + \beta_2 (s_1 + s_2 \gamma_2 + v) + \beta_3 \gamma_3 + u.$$

$$= (\beta_1 + \beta_2 s_1) + \beta_2 s_2 \gamma_2 + \beta_3 \gamma_3 + (u + \beta_2 v)$$

As long as v is not correlated
 with γ_3 then proxy is
 "good" and LS is consistent
 for β_3 . (NOT β_2)

$$\ln(\text{wage}) = \beta_1 + \beta_2 \text{education} + \beta_3 \text{experience} + \beta_4 \text{IQ} + u.$$

IQ is proxy for Ability.

If errors for IQ are not correlated with educ or exper

then this "controls" for Ability in the regression.

$$E(\text{ability} | \text{IQ}) = \delta_1 + \delta_2 \text{IQ}$$

Average Ability, given IQ
does not depend on educ
or exper.

Another words, "Ability" does
not change with educ or exper.

Possible to include more Proxys. Bottom line.

Proxys are useful control variables
as long as their Avg's do not depend
on any of the other ~~dependent~~ Indep vars.

TSLS

The Two stage Least Squares Estimate is an instrumental variable estimate. It is best to think of it this way.

Although we often rationalize its use as being computed in Two-stage - it should not be - at least if you want to get correct* std. errors.

But, for pedagogical reasons it's common to talk about the estimate as TSLS.

* Consistent: required for valid tests
in confidence intervals.

$$y_i = \beta_1 + \beta_2 \chi_{i2} + s_1 z_{i1} + \dots + s_s z_{is} + u_i$$

z_{ij} $j = 1, 2, \dots, s$ are exogenous

χ_{i2} is endogenous.

$w_{i1}, w_{i2}, \dots, w_{ik}$ instruments.

note: w 's do not appear in eq for y . These are correlated with χ_{i2} but not with u .

Stage 1:

$$\begin{aligned} \chi_{i2} = & \pi_0 + \pi_1 w_{i1} + \pi_2 w_{i2} + \dots + \pi_k w_{ik} \\ & + \pi_{k+1} z_{i1} + \dots + \pi_{k+s} z_{is} + \text{Res}_i \end{aligned}$$

notice that this eq contains

All instruments

AND

All exog. vars

This eq is often referred to
as the "Reduced Form".

Reduced Form equations express
each endog. Variable as a
function of All exog. Variables
(Z 's and W 's)

Stage 2:

Take Predictions from Reduced
Form and Replace Z_{i2} with
 \hat{Z}_{i2} . Est. using LS.

$$Y_i = \beta_1 + \beta_2 \hat{Z}_{i2} + s_1 Z_{i1} + \dots + s_s Z_{is} + \text{Res.}$$

LS of stage 2 is consistent
for β 's and s 's. But
Their est. std. errors are
not est. consistently.

This is equiv. to IV est with inst. W 's
and Z .

Instrument Strength

Two major problems with Causality

By having weak instruments

(1) The IV estimates will
Be BADLY Biased.

(2) Tests about parameters
will have "large size
distortion" - That means
wrong size

If W_i & Z_i are highly correlated
These problems diminish.

$$y_i = \beta_1 + \beta_2 x_{i2} + \delta_1 z_{i1} + u_i$$

x_{i2} endog

z_{i1} exog.

w_{i1} is instrument, which is
correlated with x_{i2} but not
 u_i

Stage 1

$$Y_2 = \pi_1 + \pi_2 \cdot Z_2 + \pi_3 W_1 + \text{residual}.$$

includes all exogenous variables
including instruments Z_i & W_i

Test:

$$H_0: \pi_3 = 0$$

$$H_A: \pi_3 \neq 0$$

(Coef on The instruments = 0.)

use F-STAT (t^2) > 10

\Rightarrow instruments are "strong enough"
to get bias within 10%

$\approx \frac{1}{F}$ see Staiger & Stock 1997

- * The precise tables can be found in Stock & Yogo 2005
And the ones produced by Stata use these.

If you have extra instruments
Then still do joint test. on
First stage regression.

$$X_2 = \beta_1 + \beta_2 Z_2 + \beta_3 W_1 + \beta_4 W_2 + \text{res.}$$

$$H_0: \beta_3 = \beta_4 = 0$$

$$H_A: \text{NOT } H_0:$$

Rule of thumb: $F > 10 \Rightarrow$ reject weak instruments.

Instrument Validity

When Model is overidentified,
you can indirectly test the
Validity of instruments.

To Be Valid

$$\text{Cov}(W_i, u_i) = 0$$

i.e., instruments are exog.

Use The SARGAN TEST.

~~steps~~

$$H_0: E(w_i; u_i) = 0 \quad j = i, \dots, K$$

$$H_A: E[w_i; u_i] \neq 0$$

$$R > S$$

$$\# \text{inst} > \# \text{RHS endog vars}$$

Steps

(1) ~~con~~ Estimate your Model using Instrumental Variables (TSLS): use ALL instruments

(2) get residuals, \hat{u}_i
(3) Regress \hat{u}_i on all available instruments from (1), including the exog. z^i 's.

(4) $NR^2 \sim \chi_{R-S}^2$ if H_0 True

$n = \# \text{ instruments}$

$s = \# \text{ endog. Regressors.}$

Rejection implies that at least
1 of your instruments is
not exogenous

(or, that it has a direct
effect)

STATA

ivregress 2sls y ($Z_2 = w_1 \ w_2$) $z_1 \ z_2$,
small

estat firststage (strength test)

estat overid (Sargan)

predict ehat, res

reg ehat $w_1 \ w_2 \ z_1 \ z_2$

$N R^2 \sim \chi^2$, if inst. exog.

One More Possibility

Maybe the x 's are exogenous and you should have used least squares rather than IV.

$$H_0: \text{Cov}(x, u) = 0 \quad \text{use LS}$$
$$H_A: \text{Cov}(x, u) \neq 0 \quad \text{use IV}$$

Hausman Test

$$y = \beta_1 + \beta_2 x + \beta_3 z + u$$

and we have instruments
 w_1, w_2 . x may be endog.

(1) First stage

$$x = \pi_1 + \pi_2 w_1 + \pi_3 w_2 + \pi_4 z + \text{residual.}$$

include on R.h.s. ALL instruments and all z 's that are not expected to be exogenous.

(2) get residuals from (1)
⇒ \hat{u} and run
regression

$$y = \beta_1 + \beta_2 x + \beta_3 z + \gamma \hat{u} + \text{res.}$$

$$H_0: \gamma = 0$$

$$H_A: \gamma \neq 0$$

use t-test. $\tilde{\gamma} \sim N(0,1)$ under H_0 .

Rejection implies x is
endog. and you should use
IV estimator.

More than 1 endog regressor.

$$y = \beta_1 + \beta_2 \chi_2 + \beta_3 \chi_3 + \beta_4 z + u$$

First stage

$$\left. \begin{array}{l} (1) \quad \chi_2 = \pi_1 + \pi_2 z + \pi_3 w_1 + \pi_4 w_2 + \nu_1 \\ \Rightarrow \hat{u}_1 \\ (2) \quad \chi_3 = \pi_1 + \pi_2 z + \pi_3 w_1 + \pi_4 w_2 + \nu_2 \\ \Rightarrow \hat{u}_2 \end{array} \right\}$$

$$(3) \quad y = \beta_1 + \beta_2 \chi_2 + \beta_3 \chi_3 + \beta_4 z + \delta_1 \hat{u}_1 + \delta_2 \hat{u}_2 + \nu_3$$

$$H_0: \delta_1 = \delta_2 = 0$$

H_A not H_0 :

F-stat $\tilde{F}_{2, n-6}$ if H_0 True.

Rejection implies

χ_2 & χ_3 endog.
use IV.

Printed on Tuesday, March 29, 2011 at 09:13:48

```

01 * file chap10.do for Using Stata for Principles of Econometrics, 4e
02
03 * cd c:\data\poe4stata
04 webuse set http://www.principlesofeconometrics.com/poe4/data/stata/
05 * Stata do-file
06 * copyright C 2011 by Lee C. Adkins and R. Carter Hill
07 * webused for "Using Stata for Principles of Econometrics, 4e"
08 * by Lee C. Adkins and R. Carter Hill (2011)
09 * John Wiley and Sons, Inc.
10
11 * setup
12 version 11.1
13
14 * open data and examine
15 webuse mroz, clear
16 describe
17 summarize
18
19 * drop nonworking women and summarize
20 drop if lfp==0
21 summarize wage educ exper
22
23 * create variables
24 gen lwage = ln(wage)
25
26 * Least squares estimation
27 reg lwage educ exper c.exper#c.exper
28 estimates store ls
29 ivregress 2sls lwage (educ=mothereduc) exper c.exper#c.exper, small
30 estimates store iv1
31 ivregress 2sls lwage (educ=mothereduc fathereduc) exper c.exper#c.exper, small
32 estimates store iv2
33
34 esttab ls iv1 iv2, compress t(%12.2f) b(%12.5f) nostar ///
35     gaps scalars(rss F) title("LS vs IV")
36
37 * Manually using only mothereduc as IV (wrong standard errors)
38 * first stage regression
39 reg educ exper c.exper#c.exper mothereduc
40
41 * test IV strength
42 test mothereduc
43
44 * obtain predicted values
45 predict educhat
46
47 * 2sls using 2-stages
48 reg lwage educhat exper c.exper#c.exper
49
50 ivregress 2sls lwage (educ=mothereduc fathereduc) exper c.exper#c.exper, small
51 estat firststage
52 estat overid
53
54 ***** Hausman test
55
56 * reduced form
57 reg educ exper c.exper#c.exper mothereduc fathereduc
58 predict vhat, residuals
59
60 * augment wage equation with reduced form residuals
61 reg lwage exper c.exper#c.exper educ vhat
62 reg lwage exper c.exper#c.exper educ vhat, vce(robust)
63
64 * Hausman test automatic
65 hausman iv2 ls, constant sigmamore
66

```

Printed on Tuesday, March 29, 2011 at 09:13:48

```
67 ***** Testing surplus moment conditions
68
69 * obtain 2sls residuals
70 quietly ivregress 2sls lwage (educ=mothereduc fathereduc) exper c.exper#c.exper, small
71 predict ehat, residuals
72
73 * regress 2sls residuals on all IV
74 reg ehat exper c.exper#c.exper mothereduc fathereduc
75 ereturn list
76
77 * NR^2 test
78 scalar nr2 = e(N)*e(r2)
79 scalar chic = invchi2tail(1,.05)
80 scalar pvalue = chi2tail(1,nr2)
81 di "R^2 from artificial regression = " e(r2)
82 di "NR^2 test of overidentifying restriction = " nr2
83 di "Chi-square critical value 1 df, .05 level = " chic
84 di "p value for overidentifying test 1 df, .05 level = " pvalue
85
86 * Using estat
87 quietly ivregress 2sls lwage (educ=mothereduc fathereduc) exper c.exper#c.exper, small
88 estat overid
89
90
```