Warwick Economics Summer School
Topics in Microeconometrics
Instrumental Variables Estimation

Michele Aquaro

University of Warwick

This version: July 21, 2016

Reading material

• Textbook: *Introductory Econometrics: A Modern Approach*, 5th edition, J.M. Wooldridge, Chapter 15.

• These slides. (When the symbol  appears on the top-right of the page, it means that the material presented is slightly more advanced.)

• If we think an explanatory variable in a model is endogenous, and we have cross-sectional data, we have basically two choices.
(1) Collect good controls in the hope that the endogenous explanatory variable becomes exogenous.
(2) Find one or more instrumental variables for the endogenous explanatory variable.
• Coming up with convincing instruments is difficult.

EXAMPLE: Estimating the return to education for married women.

• Consider a model in the population:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

where we think *educ* is endogenous.

• We can solve the endogeneity problem if we can collect data on a variable, z, that satisfies two restrictions.

$$y = \beta_0 + \beta_1 x + u$$

1. z is **exogenous** to the equation:

$$Cov(z, u) = 0$$

2. z is **relevant** for explaining x:

$$Cov(z, x) \neq 0$$

• Important difference between these two requirements: We cannot test (1), but we can determine (usually with high confidence) whether (2) is true.
• How can we use a variable z satisfying these two requirements?
• Take the covariance of z with both sides of the equation

$$y = \beta_0 + \beta_1 x + u$$

to get

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u).$$

• Now use $Cov(z, u) = 0$ (exogeneity) to get

$$Cov(z, y) = \beta_1 Cov(z, x).$$

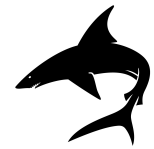
• Next, use $Cov(z, x) \neq 0$ (relevance) to get

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)}$$

• We have written β_1 as two population moments in observable variables.
• Given a random sample $\{(y_i, x_i, z_i) : i = 1, ..., n\}$, use the sample covariances to estimate the population covariances. (Method of moments estimation).
• This gives the IV estimator of β_1.

$$\hat{\beta}_{1,IV} = \frac{n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}$$

where $\bar{z} = n^{-1} \sum_{i=1}^{n} z_i$, similarly for $\bar{x}$ and $\bar{y}$.



• The variance of the IV estimator can be large. Under homoskedasticity of u,

$$Var(\hat{\beta}_{1,IV}) \approx \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2}$$

with $\sigma_u^2 = Var(u)$, $\sigma_x^2 = Var(x)$ and $\rho_{x,z} = Corr(x,z)$.
• Comparable formula for OLS (when OLS is consistent):

$$Var(\hat{\beta}_{1,OLS}) \approx \frac{\sigma_u^2}{n\sigma_x^2}$$

• So, as a rough rule of thumb, the standard error of the IV estimator is about

$$\frac{1}{r_{xz}}$$

larger than that for OLS, where r_{xz} is the sample correlation between x_i and z_i.
• Can think of this factor as the cost of doing IV when we could be doing OLS. (If OLS is inconsistent, the variance comparison makes little sense.)
• Often r_{xz} is small, so IV standard error is "large.".

• Can compute a heteroskedasticity robust or nonrobust standard error and conduct large-sample inference using t statistics and confidence intervals.
• No restrictions on the nature of x_i or z_i. For example, each could be binary, or just one of them.
• In Stata the command is `ivregress`:

```
ivregress 2sls y (x = z)
```

• Notice: The Stata command is `ivreg` is obsolete.

• To even proceed with IV, we need to first demonstrate that z_i helps to predict x_i. Easiest way is to just regress x_i on z_i and do a t test.
• Actually, recent research (Staiger and Stock, 1997) on so-called "weak instruments" says that, in this simple case, the t statistic from this regression should be at least $3.2 \approx \sqrt{10}$– much higher than just a rejection at the standard 5% level.

Example (MROZ.DTA)

- Estimating the return to education for married women (Example 15.1).

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- We use father's education ($fatheduc$) as an instrument for $educ$.

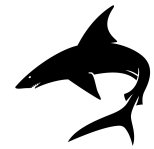
```
clear all
capture log close
set more off
set linesize 82
qui log using lg_wooldridge2013_chapter15_mroz_example1501.txt, text replace

/*
  Example 15.1
*/

use MROZ.DTA
des inlf wage educ fatheduc exper
sum inlf wage educ fatheduc exper

reg lwage educ
reg educ fatheduc if inlf == 1
ivregress 2sls lwage (educ = fatheduc)

qui log close
```

```
.
. /*
>   Example 15.1
> */
.
. use MROZ.DTA

. des inlf wage educ fatheduc exper

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
inlf            byte   %9.0g                  =1 if in lab frce, 1975
wage            float  %9.0g                  est. wage from earn, hrs
educ            byte   %9.0g                  years of schooling
fatheduc        byte   %9.0g                  father's years of schooling
exper           byte   %9.0g                  actual labor mkt exper

. sum inlf wage educ fatheduc exper

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        inlf |       753    .5683931    .4956295          0          1
        wage |       428    4.177682    3.310282      .1282         25
        educ |       753    12.28685    2.280246          5         17
    fatheduc |       753    8.808765     3.57229          0         17
       exper |       753    10.63081     8.06913          0         45

.
. reg lwage educ

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =   56.93
       Model |  26.3264193     1  26.3264193           Prob > F      =  0.0000
    Residual |  197.001022   426  .462443713           R-squared     =  0.1179
-------------+------------------------------           Adj R-squared =  0.1158
       Total |  223.327441   427  .523015084           Root MSE      =  .68003

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1086487   .0143998     7.55   0.000     .0803451    .1369523
```

```
       _cons |  -.1851968   .1852259    -1.00   0.318    -.5492673    .1788736
------------------------------------------------------------------------------

. reg educ fatheduc if inlf == 1

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =   88.84
       Model |  384.841983     1  384.841983           Prob > F      =  0.0000
    Residual |  1845.35428   426  4.33181756           R-squared     =  0.1726
-------------+------------------------------           Adj R-squared =  0.1706
       Total |  2230.19626   427  5.22294206           Root MSE      =  2.0813

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   .2694416   .0285863     9.43   0.000     .2132538    .3256295
       _cons |   10.23705   .2759363    37.10   0.000     9.694685    10.77942
------------------------------------------------------------------------------

. ivregress 2sls lwage (educ = fatheduc)

Instrumental variables (2SLS) regression              Number of obs =     428
                                                      Wald chi2(1)  =    2.85
                                                      Prob > chi2   =  0.0914
                                                      R-squared     =  0.0934
                                                      Root MSE      =  .68778

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0591735   .0350596     1.69   0.091     -.009542     .127889
       _cons |   .4411034   .4450583     0.99   0.322    -.4311947    1.313402
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   fatheduc

.
. qui log close
```

Example (MROZ.DTA)

- The IV estimate of *educ* is 5.9%, which is barely more than one-half of the OLS estimate (this suggests that OLS is too high and there may be omitted variable bias).



• Should not ignore the possibility that the instrument is not exogenous. From Wooldridge (5e, Chapter 15):

$$
\begin{aligned}
plim(\hat{\beta}_{1,OLS}) &= \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot Corr(x,u) \\
plim(\hat{\beta}_{1,IV}) &= \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot \frac{Corr(z,u)}{Corr(z,x)}
\end{aligned}
$$

• So even if $Corr(z,u)$ is smaller than $Corr(x,u)$, the bias in IV can be much larger because $Corr(z,u)$ is blown up by

$$\frac{1}{Corr(z,x)}$$

• Having $Corr(z,x)$ on the order or .10 or smaller is not unusual.

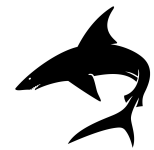
• Sometimes a potential instrument is exogenous only when other factors are controlled for.
• Let us consider a model with two explanatory variables:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

We use a new notation to distinguish endogenous from exogenous variables. The variables y_2 and z_1 are the explanatory variables and u_1 is the error term. The variable y_2 is suspected to be endogenous (correlated with u_1). An example can be

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1$$

where $y_1 = \log(wage)$, $y_2 = educ$, and $z_1 = exper$: we assume that $exper$ is exogenous but we allow that $educ$ is correlated with u_1.
• We need another exogenous variable (z_2) that does not appear in the equation above, and for which $Cov(y_2, z_2) \neq 0$ *and* $Cov(z_2, u_1) = 0$.



Given

$$E(u_1) = 0, \qquad Cov(z_1, u_1) = 0, \qquad Cov(z_2, u_1) = 0$$

we can estimate β_0, β_1, and β_2 by solving the sample counterparts of the three moment conditions above:

$$\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i1} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i1} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i1} - \hat{\beta}_2 z_{i1}) = 0$$

We still need the instrumental variable z_2 to be correlated with y_2, but this is now complicated by the presence of z_1 (we need to state the assumption in terms of *partial correlation*):

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2,$$

The key identification condition is that

$$\pi_2 \neq 0$$

In other words, after partialling out z_1, y_2 and z_2 are still correlated.

Example (CARD.DTA)

- Using college proximity as an IV for education (Example 15.4).
- Sample of man in 1976.
- $y_1 = \log(wage)$. A binary indicator, $nearc4$, equal to one if the man was near a four-year college in high school can be used as an IV. Would expect $y_2 = educ$ and $nearc4$ to be positively related.
- Card argues that, while $nearc4$ may be correlated with ability, it is uncorrelated after controlling for region of the U.S. where the man lived at age 16. He also includes a race indicator, living in an SMSA (both currently and at age 16), living in the south (currently), and experience.

```
clear all
capture log close
set more off
set linesize 82
qui log using lg_wooldridge2013_chapter15_card.txt, text replace

/*
  Example 15.4
*/

use CARD.DTA
des wage educ exper black smsa south smsa66 nearc4 reg662

* OLS
reg lwage educ exper expersq black smsa south smsa66 reg662-reg669

* IV
reg lwage educ exper expersq black smsa south smsa66 reg662-reg669
ivregress 2sls lwage exper expersq black smsa south smsa66 reg662-reg669 ///
                     (educ = nearc4),

qui log close
```

```
.
. /*
>   Example 15.4
> */
.
. use CARD.DTA

. des wage educ exper black smsa south smsa66 nearc4 reg662

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
wage            int    %9.0g                  hourly wage in cents, 1976
educ            byte   %9.0g                  years of schooling, 1976
exper           byte   %9.0g                  age - educ - 6
black           byte   %9.0g                  =1 if black
smsa            byte   %9.0g                  =1 in in SMSA, 1976
south           byte   %9.0g                  =1 if in south, 1976
smsa66          byte   %9.0g                  =1 if in SMSA, 1966
nearc4          byte   %9.0g                  =1 if near 4 yr college, 1966
reg662          byte   %9.0g                  =1 for region 2, 1966

.
. * OLS
. reg lwage educ exper expersq black smsa south smsa66 reg662-reg669

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 15,  2994) =   85.48
       Model |  177.695591    15  11.8463727           Prob > F      =  0.0000
    Residual |  414.946054  2994  .138592536           R-squared     =  0.2998
-------------+------------------------------           Adj R-squared =  0.2963
       Total |  592.641645  3009  .196956346           Root MSE      =  .37228

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0746933   .0034983    21.35   0.000     .0678339    .0815527
       exper |    .084832   .0066242    12.81   0.000     .0718435    .0978205
     expersq |   -.002287   .0003166    -7.22   0.000    -.0029079   -.0016662
       black |  -.1990123   .0182483   -10.91   0.000    -.2347927   -.1632318
        smsa |   .1363845   .0201005     6.79   0.000     .0969724    .1757967
       south |   -.147955   .0259799    -5.69   0.000    -.1988952   -.0970148
```

```
      smsa66 |   .0262417   .0194477     1.35   0.177    -.0118905    .0643739
      reg662 |   .0963672   .0358979     2.68   0.007     .0259801    .1667542
      reg663 |     .14454   .0351244     4.12   0.000     .0756696    .2134105
      reg664 |   .0550756   .0416573     1.32   0.186    -.0266043    .1367554
      reg665 |   .1280248   .0418395     3.06   0.002     .0459878    .2100618
      reg666 |   .1405174   .0452469     3.11   0.002     .0517992    .2292356
      reg667 |    .117981   .0448025     2.63   0.008     .0301343    .2058277
      reg668 |  -.0564361   .0512579    -1.10   0.271    -.1569404    .0440682
      reg669 |   .1185698   .0388301     3.05   0.002     .0424335     .194706
       _cons |   4.620807   .0742327    62.25   0.000     4.475254    4.766359
------------------------------------------------------------------------------


.
. * IV
. reg lwage educ exper expersq black smsa south smsa66 reg662-reg669

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F( 15,  2994) =   85.48
       Model |  177.695591    15  11.8463727           Prob > F      =  0.0000
    Residual |  414.946054  2994  .138592536           R-squared     =  0.2998
-------------+------------------------------           Adj R-squared =  0.2963
       Total |  592.641645  3009  .196956346           Root MSE      =  .37228

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0746933   .0034983    21.35   0.000     .0678339    .0815527
       exper |    .084832   .0066242    12.81   0.000     .0718435    .0978205
     expersq |   -.002287   .0003166    -7.22   0.000    -.0029079   -.0016662
       black |  -.1990123   .0182483   -10.91   0.000    -.2347927   -.1632318
        smsa |   .1363845   .0201005     6.79   0.000     .0969724    .1757967
       south |   -.147955   .0259799    -5.69   0.000    -.1988952   -.0970148
      smsa66 |   .0262417   .0194477     1.35   0.177    -.0118905    .0643739
      reg662 |   .0963672   .0358979     2.68   0.007     .0259801    .1667542
      reg663 |     .14454   .0351244     4.12   0.000     .0756696    .2134105
      reg664 |   .0550756   .0416573     1.32   0.186    -.0266043    .1367554
      reg665 |   .1280248   .0418395     3.06   0.002     .0459878    .2100618
      reg666 |   .1405174   .0452469     3.11   0.002     .0517992    .2292356
      reg667 |    .117981   .0448025     2.63   0.008     .0301343    .2058277
      reg668 |  -.0564361   .0512579    -1.10   0.271    -.1569404    .0440682
      reg669 |   .1185698   .0388301     3.05   0.002     .0424335     .194706
       _cons |   4.620807   .0742327    62.25   0.000     4.475254    4.766359
------------------------------------------------------------------------------
```

```
. ivregress 2sls lwage exper expersq black smsa south smsa66 reg662-reg669 ///
>                     (educ = nearc4),

Instrumental variables (2SLS) regression              Number of obs =    3010
                                                      Wald chi2(15) =  769.20
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.2382
                                                      Root MSE      =   .3873

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1315038   .0548174     2.40   0.016     .0240637     .238944
       exper |   .1082711   .0235956     4.59   0.000     .0620246    .1545176
     expersq |  -.0023349   .0003326    -7.02   0.000    -.0029868    -.001683
       black |  -.1467757   .0537564    -2.73   0.006    -.2521364   -.0414151
        smsa |   .1118083   .0315777     3.54   0.000     .0499171    .1736995
       south |  -.1446715    .027212    -5.32   0.000    -.1980061   -.0913369
      smsa66 |   .0185311   .0215511     0.86   0.390    -.0237082    .0607704
      reg662 |   .1007678   .0375854     2.68   0.007     .0271017    .1744339
      reg663 |   .1482588   .0367162     4.04   0.000     .0762964    .2202211
      reg664 |   .0498971   .0436234     1.14   0.253    -.0356032    .1353974
      reg665 |   .1462719   .0469387     3.12   0.002     .0542738    .2382701
      reg666 |   .1629029   .0517714     3.15   0.002     .0614328    .2643731
      reg667 |   .1345722   .0492708     2.73   0.006     .0380032    .2311413
      reg668 |   -.083077   .0591735    -1.40   0.160    -.1990548    .0329008
      reg669 |   .1078142   .0417024     2.59   0.010      .026079    .1895494
       _cons |   3.666151   .9223682     3.97   0.000     1.858342    5.473959
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq black smsa south smsa66 reg662 reg663 reg664
               reg665 reg666 reg667 reg668 reg669 nearc4

.
. qui log close
```

Example (CARD.DTA)

- First we test if the coefficient on *nearc*4 in the regression of *educ* on *nearc*4 and the other exogenous variables is equal to zero (reject the null).
- The IV estimate of return to education is almost twice as large as the OLS estimate (13.2% for IV, 7.5% for OLS). However, the standard error of the IV estimate is over 18 times larger than the OLS standard error.

Two Stage Least Squares (2SLS)

• When we have more instruments than necessary, IV becomes **two stage least squares (2SLS)**.
• `ivreg` in Stata works the same. We list all outside instruments for the endogenous explanatory variable (such as *educ*).
• Trying to do 2SLS "by hand" can lead to mistakes – at a minimum, wrong standard errors.
• The Staiger-Stock rule-of-thumb for whether we have strong enough IVs is that, in the first stage regression, the joint F for significance of IVs should be above 10. (The t statistic rule is a special case: $\sqrt{10} \approx 3.2$.)

2SLS

Consider the following model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

which has two explanatory variables, one endogenous (y_2), one exogenous (z_1). Suppose we have *two* instruments for y_2; let us call them z_2 and z_3.
Suppose

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

where

$$E(v_2) = 0$$
$$Cov(z_j, v_2) = 0 \qquad j = 1, 2, 3$$

2SLS

To find the best IV, we choose the linear combination that is most highly correlated with y_2, which we call y_2^*

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

This is estimated as

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

and then test

$$H_0 \colon \pi_2 = 0 \text{ and } \pi_3 = 0$$

Two stage regression:

1. OLS regression of y_2 on z_1, z_2, and z_3;
2. OLS regression of y_1 on z_1, and $\hat{y}_2$.

Example (MROZ.DTA)

- In the Mroz data set, we can use *motheduc* as an additional instrument for *educ* (together with *fatheduc*).

```
clear all
capture log close
set more off
set linesize 82
qui log using lg_wooldridge2013_chapter15_mroz_example1505.txt, text replace

/*
  Example 15.5
*/

use MROZ.DTA
des inlf wage educ fatheduc motheduc exper
sum inlf wage educ fatheduc motheduc exper

reg educ exper expersq motheduc fatheduc if inlf == 1
test motheduc fatheduc
ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)

* Compute 2sls estimator "manually"
reg educ exper expersq motheduc fatheduc if inlf == 1
predict hat_educ, xb
reg lwage hat_educ exper expersq

qui log close
```

- Note: the option `inlf==1` is used to restrict the analysis to the sub-group of working women (i.e. those for which we observe wage).

```
.
. /*
>   Example 15.5
> */
.
. use MROZ.DTA

. des inlf wage educ fatheduc motheduc exper

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
inlf            byte   %9.0g                  =1 if in lab frce, 1975
wage            float  %9.0g                  est. wage from earn, hrs
educ            byte   %9.0g                  years of schooling
fatheduc        byte   %9.0g                  father's years of schooling
motheduc        byte   %9.0g                  mother's years of schooling
exper           byte   %9.0g                  actual labor mkt exper

. sum inlf wage educ fatheduc motheduc exper

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        inlf |       753    .5683931    .4956295          0          1
        wage |       428    4.177682    3.310282      .1282         25
        educ |       753    12.28685    2.280246          5         17
    fatheduc |       753    8.808765     3.57229          0         17
    motheduc |       753    9.250996    3.367468          0         17
-------------+--------------------------------------------------------
       exper |       753    10.63081     8.06913          0         45

.
. reg educ exper expersq motheduc fatheduc if inlf == 1

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   28.36
       Model |  471.620998     4   117.90525           Prob > F      =  0.0000
    Residual |  1758.57526   423  4.15738833           R-squared     =  0.2115
-------------+------------------------------           Adj R-squared =  0.2040
       Total |  2230.19626   427  5.22294206           Root MSE      =   2.039

------------------------------------------------------------------------------
```

```
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0452254   .0402507     1.12   0.262    -.0338909    .1243417
     expersq |  -.0010091   .0012033    -0.84   0.402    -.0033744    .0013562
    motheduc |    .157597   .0358941     4.39   0.000      .087044    .2281501
    fatheduc |   .1895484   .0337565     5.62   0.000     .1231971    .2558997
       _cons |    9.10264   .4265614    21.34   0.000     8.264196    9.941084
------------------------------------------------------------------------------

. test motheduc fatheduc

 ( 1)  motheduc = 0
 ( 2)  fatheduc = 0

       F(  2,   423) =   55.40
            Prob > F =    0.0000

. ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)

Instrumental variables (2SLS) regression              Number of obs =     428
                                                      Wald chi2(3)  =   24.65
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.1357
                                                      Root MSE      =  .67155

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0613966   .0312895     1.96   0.050     .0000704    .1227228
       exper |   .0441704   .0133696     3.30   0.001     .0179665    .0703742
     expersq |   -.000899   .0003998    -2.25   0.025    -.0016826   -.0001154
       _cons |   .0481003    .398453     0.12   0.904    -.7328532    .8290538
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc

.
. * Compute 2sls estimator "manually"
. reg educ exper expersq motheduc fatheduc if inlf == 1

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   28.36
       Model |  471.620998     4   117.90525           Prob > F      =  0.0000
```

```
    Residual |  1758.57526   423  4.15738833           R-squared     =  0.2115
-------------+----------------------------           Adj R-squared =  0.2040
       Total |  2230.19626   427  5.22294206           Root MSE      =   2.039

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0452254   .0402507     1.12   0.262    -.0338909    .1243417
     expersq |  -.0010091   .0012033    -0.84   0.402    -.0033744    .0013562
    motheduc |    .157597   .0358941     4.39   0.000      .087044    .2281501
    fatheduc |   .1895484   .0337565     5.62   0.000     .1231971    .2558997
       _cons |    9.10264   .4265614    21.34   0.000     8.264196    9.941084
------------------------------------------------------------------------------

. predict hat_educ, xb

. reg lwage hat_educ exper expersq

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =    7.40
       Model |   11.117828     3  3.70594266           Prob > F      =  0.0001
    Residual |  212.209613   424   .50049437           R-squared     =  0.0498
-------------+------------------------------           Adj R-squared =  0.0431
       Total |  223.327441   427  .523015084           Root MSE      =  .70746

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    hat_educ |   .0613966   .0329624     1.86   0.063    -.0033933    .1261866
       exper |   .0441704   .0140844     3.14   0.002     .0164865    .0718543
     expersq |   -.000899   .0004212    -2.13   0.033    -.0017268   -.0000711
       _cons |   .0481003   .4197565     0.11   0.909    -.7769624     .873163
------------------------------------------------------------------------------

.
. qui log close
```

Section 1

Testing Whether a Variable is Endogenous

Testing for Endogeneity

• To illustrate, consider we have a single suspected endogenous variable

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1,$$

where z_1 and z_2 are exogenous.
• We want to test whether y_2 and u_1 are uncorrelated – that is, the null hypothesis is that y_2 is exogenous and so we can use OLS rather than IV.

Testing for Endogeneity

- In addition to assuming z_1 and z_2 are exogenous – so they act as their own IVs – we need at least one outside exogenous variable. We might have more than one; suppose we have two, z_3 and z_4.
- Write the so-called *reduced form* for y_2:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2$$

- With y_2 written this way, it is endogenous if and only if

$$Cov(v_2, u_1) \neq 0.$$

Testing for Endogeneity

- To test the null that $Cov(v_2, u_1) = 0$, we can write

$$u_1 = \delta_1 v_2 + e_1$$

where the new error e_1 is uncorrelated with z_j, $j = 1, \ldots, 4$ and v_2, and therefore y_2.

- Plug in for u_1 into the original equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 v_2 + e_1$$

which is an equation that, if we observed v_2, could be estimated by OLS.

- We can estimate v_2 using the first-stage regression.

Testing for Endogeneity

• The two-step testing procedure is
(i) Regress y_{i2} on an intercept and all z_j, $j = 1, \ldots, 4$ to obtain the residuals, $\hat{v}_{i2}$ (one for each observation i).
(ii) Run the regression (using n observations)

$$y_{i1} \text{ on } y_{i2}, z_{i1}, z_{i2}, \hat{v}_{i2}$$

and use a (robust) t test on $\hat{v}_{i2}$.
• Alternatively, Hausman (1978) suggested directly comparing the OLS and 2SLS estimates and determining whether the differences are statistically significant.

```
clear all
capture log close
set more off
set linesize 82
qui log using lg_wooldridge2013_chapter15_mroz_example1507.txt, text replace

/*
  Example 15.7
*/

use MROZ.DTA
reg educ exper expersq motheduc fatheduc if inlf == 1
predict hat_v_2 if inlf == 1, res
reg lwage educ exper expersq hat_v_2

qui log close
```

```
.
. /*
>   Example 15.7
> */
.
. use MROZ.DTA

. reg educ exper expersq motheduc fatheduc if inlf == 1

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   28.36
       Model |  471.620998     4   117.90525            Prob > F      =  0.0000
    Residual |  1758.57526   423  4.15738833            R-squared     =  0.2115
-------------+------------------------------           Adj R-squared =  0.2040
       Total |  2230.19626   427  5.22294206            Root MSE      =   2.039

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0452254   .0402507     1.12   0.262    -.0338909    .1243417
     expersq |  -.0010091   .0012033    -0.84   0.402    -.0033744    .0013562
    motheduc |    .157597   .0358941     4.39   0.000      .087044    .2281501
    fatheduc |   .1895484   .0337565     5.62   0.000     .1231971    .2558997
       _cons |    9.10264   .4265614    21.34   0.000     8.264196    9.941084
------------------------------------------------------------------------------

. predict hat_v_2 if inlf == 1, res
(325 missing values generated)

. reg lwage educ exper expersq hat_v_2

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  4,   423) =   20.50
       Model |  36.2573098     4  9.06432745            Prob > F      =  0.0000
    Residual |  187.070131   423  .442246173            R-squared     =  0.1624
-------------+------------------------------           Adj R-squared =  0.1544
       Total |  223.327441   427  .523015084            Root MSE      =  .66502

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0613966   .0309849     1.98   0.048      .000493    .1223003
```

```
       exper |   .0441704   .0132394     3.34   0.001     .0181471    .0701937
     expersq |   -.000899   .0003959    -2.27   0.024    -.0016772   -.0001208
     hat_v_2 |   .0581666   .0348073     1.67   0.095    -.0102501    .1265834
       _cons |   .0481003   .3945753     0.12   0.903    -.7274721    .8236727
------------------------------------------------------------------------------


.
. qui log close
```

Example (MROZ.DTA)

- The coefficient of $\hat{v}_2$ is $\hat{\delta}_1 = 0.058$, and $t = 1.67$.
- According to these results, we fail to reject the null that *educ* is exogenous. However, the test relies on the assumption that the two instruments, $motheduc$ and $fatheduc$, are exogenous, which very likely are not.
- It is probably a good idea to report both estimates, because the 2SLS estimate of the return to education (6.1%) is well below the OLS estimate (10.8%).