Instrumental-variables estimation

CONTENTS:

- **IV estimators: IV, 2SLS and GMM**
- **Testing for endogeneity and overidentifying restrictions**
- **Weak instruments**

IV estimators: IV, 2SLS and GMM

Instrumental variables

To motivate the need for the implementation of an instrumental variables (IV) approach, consider the following linear population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u \quad (1)$$

$$E(u) = 0, \quad Cov(x_j, u) = 0, \;\; j = 1,2, \ldots K-1 \quad (2)$$

where x_K might be correlated with u. That is x_1, x_2,..., x_{K-1} are **exogenous**, but x_K is potentially **endogenous** in equation (1). Equation (1) is known as the **structural equation.**

Endogeneity may result from many sources such as:

- **Omitted variables**: it appears when the specified model incorrectly leaves out one or more important casual factors. A good example of it is omitted ability in a wage equation, where an individual's years of schooling are likely to be correlated with unobserved ability.
- **Measurement errors**: it occurs when we can only observe an imperfect measure of one of the variables we want to include in the model. An example of measurement error is found when we want to estimate a savings function with permanent income as a regressor. Since we do not observe permanent income we use current income (observable) as an imperfect measure of the permanent income.
- **Simultaneity**: it arises when at least one of the explanatory variables is determined simultaneously along with y. We can find an example of simultaneity in looking at the effect of alcohol consumption on worker productivity (as typically measured by wages), as alcohol demand would usually depend on income which is largely determined by wage.

OLS estimation of equation (1) will result in inconsistent estimates of all β_j if $Cov(x_K, u) \neq 0$ and the method of instrumental variables provides a solution to the problem of an endogenous explanatory variable.

Instrumental variables (IV)

To use the IV approach with x_K endogenous, we need an observable variable, z_1, not in equation (1) that satisfies two conditions:

- **IV1**: $cov(z_1, u) = 0$, that is, $\boldsymbol{z_1}$ **is uncorrelated with** $\boldsymbol{u}$

Consider the linear projection of x_K on all the exogenous variables (this is the so called **reduced form equation**):

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K \quad (3)$$

The key assumption on this linear projection is that the coefficient of z_1 is nonzero:

- **IV2**: $\theta_1 \neq 0$; that is, z_1 is *partially* correlated with x_K after accounting for all other exogenous variables x_1, x_2,…, x_{K-1}. Loosely speaking we can describe this condition as $\boldsymbol{z_1}$ **is correlated with** $\boldsymbol{x_K}$

Note here an important difference between condition IV1 and condition IV2: the first one cannot be tested (because it involves the unobservable) while the second one can be tested. When z_1 satisfies the two conditions above then it is said to be a **(valid) instrument** or **instrumental variable** for x_K.
Because x_1, x_2,…, x_{K-1} are already uncorrelated with u, they serve as their own instruments in equation (1).

The key to derive the IV estimator comes from the condition IV1 which implies that E($\mathbf{z}_i u_i$)=0 and hence the **moment condition** $\mathrm{E}\{\mathbf{z}_i'(y_i - x_i'\,\beta)\} = 0$. Using the sample analog of the moment condition we can solve for β and find the IV estimator. When the number of instruments is equal to the number of regressors (**just-identified case**), the instrumental variables (IV) estimator is defined as:

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \quad (4)$$

where **Z** is an *N* x *K* matrix of exogenous variables (instruments)[1], **X** is the *N* x *K* matrix of regressors and **y** is an *N* x 1 vector of the dependent variable.

EXAMPLE 1 (Instrumental variables for Education in a wage equation)

Consider the following equation:

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 educ + \beta_3 age + \beta_4 married + u \quad (5)$$

In this case, u can be thought of being correlated with *educ* because of omitted unobserved ability and other factors such as quality of education and family background that can be determining your wage as well as the level of education attained. We can use the mother's education (*meduc*) as an instrument for education. For *meduc* to be a **valid instrument** for *educ* we must assume that *meduc is* uncorrelated with u and that $\theta_1 \neq 0$ in the reduced form equation. Using the WAGE2.RAW again we find the following results:

[1] Note that any row vector of **Z** is a 1 x K vector of the form $\mathbf{z} \equiv (1, x_2, x_3, \dots, x_{K-1}, z_1)$

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/wage2

. *testing if mother's education is correlated with education

. regress educ exper age married meduc

      Source |       SS       df       MS              Number of obs =     857
-------------+------------------------------           F(  4,   852) =  116.72
       Model |  1463.23508     4  365.808771           Prob > F      =  0.0000
    Residual |  2670.16048   852  3.13399118           R-squared     =  0.3540
-------------+------------------------------           Adj R-squared =  0.3510
       Total |  4133.39557   856  4.82873314           Root MSE      =  1.7703

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |  -.2822407   .0166502   -16.95   0.000    -.3149209   -.2495606
         age |   .2111836   .0225492     9.37   0.000      .166925    .2554422
     married |  -.1273378   .1967901    -0.65   0.518    -.5135881    .2589125
       meduc |   .2087239   .0216673     9.63   0.000     .1661965    .2512514
       _cons |    7.72743   .7174672    10.77   0.000      6.31922     9.13564
------------------------------------------------------------------------------
```

The results suggest that the education of the mother is partially correlated with the education of the individual, as condition IV2 requires.

```
. ivreg lwage exper age married ( educ = meduc )

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     857
-------------+------------------------------           F(  4,   852) =   22.14
       Model |  7.5680422     4  1.89201055           Prob > F      =  0.0000
    Residual |  141.793009   852  .166423719           R-squared     =  0.0507
-------------+------------------------------           Adj R-squared =  0.0462
       Total |  149.361051   856  .174487209           Root MSE      =  .40795

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1520837   .0239216     6.36   0.000     .1051315    .1990359
       exper |   .0399734   .0083948     4.76   0.000     .0234964    .0564503
         age |  -.0068149   .0075032    -0.91   0.364    -.0215419    .0079121
     married |   .2035129   .0454691     4.48   0.000     .1142683    .2927574
       _cons |   4.314075   .2827491    15.26   0.000     3.759109    4.869041
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper age married meduc
------------------------------------------------------------------------------
```

All the parameter estimates changed from the previous estimation without instrumenting (in the linear models handout). Now the results suggest that one additional year of education generates an expected percentage change of 15.2% in monthly earnings at a 1% significance level.

Two-stage least squares

Now, consider the case where there is more than one instrumental variable for x_K (**over-identified case**):

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_M z_M + r_K \quad (6)$$

Let $z_1, z_2, \dots, z_M$ be variables such that $cov(z_h, u) = 0,\ h = 1, \dots, M$ so each variable is exogenous in equation (1). The moment condition presented above has no solution for β because it is a system with more equations than unknowns. One possible solution is to arbitrarily drop instruments to get to the just-identified case but there are more efficient estimators. One estimator is the two-stage least squares (2SLS) estimator:

$$\hat{\beta}_{2SLS} = \{\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\}^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \quad (7)$$

This estimator equals the $\hat{\beta}_{IV}$ in the just-identified case. The term 2SLS arises because the estimator can be computed in two steps. First, estimate by OLS the first-stage regression given by the reduced form equation in (3) and second, estimate by OLS the structural equation (1) with endogenous regressors replaced by their predictions from the first step.

EXAMPLE 1 (2SLS for Education in a wage equation)

We use data in the example above to perform a two-stage least squares estimation. Now, we can take advantage of the fact that we also have data on father's education (*feduc*) and use it as an instrument for *educ* with the same argument as above. Assuming that *meduc* and *feduc* are exogenous in the log (*wage*) equation we can check that the coefficients for *meduc* and *feduc* are statistically different from zero in the reduced form equation to proceed with the 2SLS estimation.

```
. ivreg lwage exper age married ( educ = meduc feduc )

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =     722
-------------+------------------------------           F(  4,   717) =   23.60
       Model |  7.62622488     4  1.90655622           Prob > F      =  0.0000
    Residual |  119.185706   717  .166228321           R-squared     =  0.0601
-------------+------------------------------           Adj R-squared =  0.0549
       Total |  126.811931   721    .1758834           Root MSE      =  .40771

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1448957   .0203597     7.12   0.000     .1049238    .1848675
       exper |   .0391087   .0076469     5.11   0.000     .0240956    .0541218
         age |   -.007241   .0074878    -0.97   0.334    -.0219417    .0074596
     married |   .2027841   .0484987     4.18   0.000     .1075677    .2980006
       _cons |   4.435099   .2561893    17.31   0.000     3.932128     4.93807
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper age married meduc feduc
------------------------------------------------------------------------------
```

The 2SLS estimate of the returns to education is about 14.5% and it is statistically significant.

Generalized Method of Moments (GMM)

The generalized method of moments is a generalization of the OLS and IV estimators. GMM is based on moment functions that depend on observable random variables and unknown parameters, and that have zero expectation in the population when evaluated at the true parameters. Its general expression is

$$\hat{\beta}_{GMM} = (\mathbf{X}'\mathbf{Z}\mathbf{X}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y} \tag{8}$$

where **W** is any full-rank symmetric-weighting matrix.[2] In general, the weights in **W** will depend on data, on unknown parameters and on the shape on the moment function.

Testing for endogeneity and overidentifying restriction

Testing for endogeneity

In the previous examples we treated the variable e*duc* as an endogenous variable but if instead, the variable is exogenous, the IV estimators (IV, 2SLS and GMM) are still consistent but they can be much less efficient than the OLS estimator. For this reason, it is important to test for endogeneity.

The **Hausman test** provides a way to test whether a regressor is endogenous. If there is little difference between OLS and 2SLS estimators, then there is no need to instrument and we conclude that the regressor is exogenous. If instead, there is considerable difference, then we need to instrument and the regressor is endogenous. In the case of just one potentially endogenous regressor with a coefficient denoted by β, the Hausman test statistic

$$T_H = \frac{(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})^2}{\hat{V}(\hat{\beta}_{2SLS}) - \hat{V}(\hat{\beta}_{OLS})} \tag{9}$$

is $\chi^2(1)$ distributed under the null hypothesis that the regressor is exogenous. Note that $\hat{V}(\hat{\beta}_{2SLS})$ and $\hat{V}(\hat{\beta}_{OLS})$ are the estimated variances of the 2SLS and OLS estimates respectively. When $\hat{V}(\hat{\beta}_{2SLS}) < \hat{V}(\hat{\beta}_{OLS})$ the results are hard to interpret.

Another convenient way of testing the same hypothesis is to estimate the following regression by OLS:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \delta\hat{r} + u \tag{10}$$

where $\hat{r}$ is the residual from the reduced form equation for x_K and do a simple t-test to see whether the estimate of δ is significantly different from zero. If $\hat{\delta}$ is significantly different from zero then x_K is endogenous. We can always use this second approach.

When we have **two potential endogenous regressors** $(x_K,\ x_{K+1})$ we can test for endogeneity in a similar way as above estimating the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \beta_{K+1} x_{K+1} + \delta_1\hat{r}_1 + \delta_2\hat{r}_2 + u \tag{11}$$

where $\hat{r}_1$ is the residual from the reduced form equation for x_K and $\hat{r}_2$ is the residual from the reduced form equation for x_{K+1}. Now, we can compute the joint F-test. If $\hat{\delta}_1$ and $\hat{\delta}_2$ are jointly and significantly different from zero then x_K and x_{K+1} are endogenous,

[2] A matrix is **full rank** if all its rows are linearly independent and all its columns are linearly independent.

Note here that endogeneity tests are based on the assumption that the instruments, z_1, z_2,… are valid instruments for the endogenous regressors.

EXAMPLE 1 (Testing the endogeneity of the variable education in a wage equation)

Using the previous estimations we can proceed with the **Hausman test**:

$$T_H = \frac{(0.1448-0.07395)^2}{0.0203597^2-0.0067421^2} = 13.602$$

The result suggests that we reject the null hypothesis that the regressor is exogenous at 1% significance level.[3]

Also, we can apply the alternative test for endogeneity:

```
. quietly regress educ meduc feduc exper age married

. predict  e, residual
(213 missing values generated)

. regress lwage educ exper age married e

      Source |       SS       df       MS              Number of obs =     722
-------------+------------------------------           F(  5,   716) =   32.67
       Model |  23.5591297     5  4.71182594           Prob > F      =  0.0000
    Residual |  103.252802   716  .144207823           R-squared     =  0.1858
-------------+------------------------------           Adj R-squared =  0.1801
       Total |  126.811931   721    .1758834           Root MSE      =  .37975

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1448957   .0189633     7.64   0.000     .1076653     .182126
       exper |   .0391087   .0071225     5.49   0.000     .0251253    .0530921
         age |   -.007241   .0069742    -1.04   0.300    -.0209335    .0064514
     married |   .2027841   .0451723     4.49   0.000     .1140981    .2914701
           e |  -.0850928    .020619    -4.13   0.000    -.1255738   -.0446119
       _cons |   4.435099   .2386178    18.59   0.000     3.966625    4.903573
------------------------------------------------------------------------------
```

The result gives us the same conclusion as before, education is an endogenous variable in the wage equation. Hence, we need to implement an instrumental variables estimator.

Testing for overidentifying restrictions

When we have more instruments than we need to identify an equation, we can test whether the instruments are valid in the sense that they are uncorrelated with u in equation (1). To perform this test we estimate equation (1) by 2SLS or IV and obtain the estimated residuals $\hat{u}$. We then regress $\hat{u}$ on all the exogenous variables (including the instruments) and obtain the R-squared of the regression. Under the null hypothesis that the instruments are uncorrelated with u in which case they are valid instruments and the statistic N x R-

[3] $\chi^2(1)$=6.635 at 0.01 probability.

squared follows a $\chi^2(\mathrm{r})$ distribution.[4] Stata performs this test directly with the post/estimation command *estat overid*.

EXAMPLE 1 (Testing overidentifying restrictions in a wage equation)

```
. quietly ivreg lwage exper age married married ( educ = meduc feduc )

. predict u, residuals

. regress u exper age married meduc feduc

      Source |       SS       df       MS              Number of obs =     722
-------------+------------------------------           F(  5,   716) =    0.02
       Model |  .013944313     5  .002788863           Prob > F      =  0.9999
    Residual |  119.171762   716  .166441008           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0069
       Total |  119.185706   721  .165306111           Root MSE      =  .40797

------------------------------------------------------------------------------
           u |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |  -.0000421   .0042455    -0.01   0.992    -.0083771     .008293
         age |  -.0000527   .0056688    -0.01   0.993    -.0111822    .0110768
     married |   .0001829   .0485038     0.00   0.997    -.0950438    .0954097
       meduc |   .0017707   .0065915     0.27   0.788    -.0111703    .0147117
       feduc |  -.0013684   .0057241    -0.24   0.811    -.0126065    .0098697
       _cons |   -.003053   .1803577    -0.02   0.986    -.3571461    .3510402
------------------------------------------------------------------------------

. display 0.0001*722
.0722
```

We will not reject the null hypothesis that the instruments are valid since $\chi^2(1)$ =6.635 at 0.01 probability

When performing the test directly in Stata the results suggest exactly the same, which confirm the validity of the two instruments used.

```
. quietly ivregress 2sls lwage exper age married ( educ = meduc feduc )

. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .084471  (p = 0.7713)
  Basmann chi2(1)        =  .083779  (p = 0.7722)
```

[4] r represents the degrees of freedom and equals the number of overidentifying restrictions.

Weak instruments

Recall the two conditions for the instrumental variables to be valid: (IV1) uncorrelated with u but (IV2) partially and sufficiently strongly correlated with x_K, once the other independent variables are controlled for. We already indicated that it is necessary to check the second condition to determine the validity of the instrument. Imagine we have now more than one instrumental variable as in equation (6). We can estimate this reduced form equation (6) by OLS and obtain the F-statistic on the estimators of the instrumental variables: $H_0 = \theta_1 = \cdots = \theta_M = 0$. If the F-statistic is small, then we conclude that the instrumental variables are **weak.** When the instrumental variables are weak, the IV or 2SLS estimators could be inconsistent or have large standard errors.

A **rule of thumb** to find weak instruments suggests that the F-statistic of the instrumental variables in (6) should be larger than 10 to ensure that the maximum bias in IV estimators be less than 10%.

EXAMPLE 1 (Testing for weak instruments in a wage equation)

```
. quietly regress educ exper age married meduc feduc

. test (meduc=0) (feduc=0)

 ( 1)  meduc = 0
 ( 2)  feduc = 0

       F(  2,   716) =   65.24
            Prob > F =    0.0000
```

The joint test on the instrumental variables *meduc* and *feduc* indicates that the instruments are not weak.

EXAMPLE 1 (Instrumental variables for Education in a wage equation using *ivreg2*)

The same exercise can be done using the *ivreg2* command. This command is similar to *ivregress* but provides additional estimators and statistics. When specifying the option "*first*", the first stage regressions are shown and some tests are performed directly. The first tests displayed are useful to determine the weakness of the instruments. The **partial R-squared** measures the squared-partial correlation between the excluded instruments and the endogenous regressor in question. As a rule of thumb, if the first-stage regression yields a large value of the standard R-squared and a small value of the partial R-squared, you should conclude that the instruments lack sufficient relevance to explain the endogenous regressor. In this case, the partial R-squared is 0.1542, which does not cast doubts about the strength of the instruments. This combined with an F-statistic higher than 10, allows us to conclude that the instruments are not weak.

Another test displayed with the "*first*" option is the **underidentification test.** The underidentification test is a test of whether the equation is identified, i.e., that the excluded instruments are "relevant", meaning correlated with the endogenous regressors. The test is essentially a test of the rank of a matrix: under the null hypothesis that the equation is underidentified, the matrix of reduced form coefficients on the L1 excluded instruments has rank=K1-1 where K1=number of endogenous regressors. Under the null, the statistic is distributed as a chi-squared with degrees of freedom=(L1-K1+1). A rejection of the null indicates that the matrix is full column rank, i.e., the model is identified. In this case, we reject the null hypothesis.

```
. ivreg2 lwage exper age married  ( educ = meduc feduc ),first

First-stage regressions
-----------------------

First-stage regression of educ:

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      722
                                                      F(  5,    716) =    91.55
                                                      Prob > F      =   0.0000
Total (centered) SS     =  3607.214681                Centered R2   =   0.3900
Total (uncentered) SS   =       138397                Uncentered R2 =   0.9841
Residual SS             =  2200.437288                Root MSE      =    1.753

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |  -.2643674   .0182429   -14.49   0.000    -.3001834   -.2285515
         age |   .2186147   .0243591     8.97   0.000      .170791    .2664385
     married |  -.0576974   .2084221    -0.28   0.782    -.4668889    .3514941
       meduc |   .1204721   .0283239     4.25   0.000     .0648643      .17608
       feduc |   .1584086   .0245967     6.44   0.000     .1101183    .2066988
       _cons |   6.592171   .7750015     8.51   0.000     5.070624    8.113718
------------------------------------------------------------------------------
Included instruments: exper age married meduc feduc
------------------------------------------------------------------------------
Partial R-squared of excluded instruments:   0.1542
Test of excluded instruments:
  F(  2,   716) =    65.24
  Prob > F      =   0.0000
```

```
Summary results for first-stage regressions
-------------------------------------------

Variable     | Shea Partial R2 |   Partial R2    |  F(  2,   716)    P-value
educ         |     0.1542      |     0.1542      |       65.24       0.0000

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(2)=111.30   P-val=0.0000
Cragg-Donald N*CDEV Wald statistic          Chi-sq(2)=131.58   P-val=0.0000

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F-statistic                     65.24
See main output for Cragg-Donald weak id test critical values

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(2,716)= 27.17     P-val=0.0000
Anderson-Rubin Wald test     Chi-sq(2)=54.80     P-val=0.0000
Stock-Wright LM S statistic  Chi-sq(2)=50.93     P-val=0.0000

Number of observations               N  =        722
Number of regressors                 K  =          5
Number of instruments                L  =          6
Number of excluded instruments       L1 =          2

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      722
                                                      F(  4,   717) =    23.60
                                                      Prob > F      =   0.0000
Total (centered) SS     =  126.8119312                Centered R2   =   0.0601
Total (uncentered) SS   =  33511.33726                Uncentered R2 =   0.9964
Residual SS             =  119.1857064                Root MSE      =    .4063

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1448957   .0202891     7.14   0.000     .1051297    .1846616
       exper |   .0391087   .0076204     5.13   0.000     .0241729    .0540444
         age |  -.007241    .0074618    -0.97   0.332    -.021866     .0073839
     married |   .2027841   .0483305     4.20   0.000     .1080581    .2975101
       _cons |   4.435099   .2553006    17.37   0.000     3.934719    4.935479
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):         111.297
                                                   Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               65.243
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.084
                                                   Chi-sq(1) P-val =    0.7713
------------------------------------------------------------------------------
Instrumented:         educ
Included instruments: exper age married
Excluded instruments: meduc feduc
------------------------------------------------------------------------------

.
```

EXERCISE 1

Consider estimating the effect of personal computer ownership, as represented by a binary variable, PC, on college GPA, *col*GPA. With data on SAT scores and high school GPA you postulate the model

$$colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 SAT + \beta_3 PC + u$$

a) Why might u and PC be positively correlated?
b) If the given equation is estimated by OLS using a random sample of college students, is $\hat{\beta}_3$ likely to have an upward or downward bias?
c) What are some variables that might be good proxies for unobservables in u that are correlated with PC?

EXERCISE 2

Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + u$$

where *male* is a binary variable indicator equal to one if the child is male; *parity* is the birth order of this child; *faminc* is family income; and *packs* is the average number of packs of cigarettes smoked per day during pregnancy.

a) Why might you expect *packs* to be correlated with u?
b) Suppose that you have data on average cigarette price in each woman's state of residence. Discuss whether this information is likely to satisfy the properties of a good instrumental variable for *packs*.
c) Use the data in BWGHT.RAW ("use http://fmwww.bc.edu/ec-p/data/wooldridge/bwght") to estimate the equation above. First use OLS. Then, use 2SLS, where *cigprice* is an instrument for packs. Discuss any important differences in the OLS and the 2SLS estimates.
d) Estimate the reduced form for *packs*. What do you conclude about identification of the equation above using *cigprice* as an instrument for *packs*? What bearing does this conclusion have on your answer from part c?

EXERCISE 3

Use the CARD.RAW ("use http://fmwww.bc.edu/ec-p/data/wooldridge/card") for this problem.

a) Estimate a log (*wage*) equation by OLS with *educ ,exper,* $exper^2$*, black, south, smsa, reg661* through *reg668* and *smsa66* as explanatory variables.
b) Estimate a reduced form equation for *educ* (years of education) containing all explanatory variables from part a and the dummy variable *nearc4* (if individual grew up in vicinity of 4-year college). Do *educ* and *nearc4* have a practically and statistically significant partial correlation?
c) Estimate the log (*wage*) equation by IV, using *nearc4* as an instrument for *educ*. Compare the 95% confidence interval for the return of education with that obtained in part a
d) Now use *nearc2* (if individual grew up in vicinity of 2-year college) along with *nearc4* as instruments for *educ*. First estimate the reduced form for *educ*, and comment on whether *nearc2* or *nearc4* is more strongly related to *educ*. How do the 2SLS estimates compare with the earlier estimates?