Computer Exercise 7 – Answers

Good practice to summarise the data before you do any regressions. Can see average age of sample is 28. 73% live in urban areas, 44% near a 2 year college. Avererage years of mother's education is 10.6 years. Avereage hourly wage is 588 (this is US data so the hourly wage variable is measured in cents and averages $5.88 an hour)

```
. su

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          id |      1900    2500.921    1491.346          3       5218
      nearc2 |      1900         .44    .4965176          0          1
      nearc4 |      1900    .6878947    .4634745          0          1
        educ |      1900    13.62895    2.592716          1         18
         age |      1900    27.92211    3.087326         24         34
-------------+--------------------------------------------------------
    fatheduc |      1900    10.11368    3.691888          0         18
    motheduc |      1900    10.61526    3.021056          0         18
      ethnic |      1900    .1589474    .3657233          0          1
       urban |      1900    .7284211     .444891          0          1
       south |      1900    .3784211     .485121          0          1
-------------+--------------------------------------------------------
       hwage |      1900    588.2095    262.1315        100       2404

. g lhwage=log(hwage)
```

OLS estimates produce the following

```
. reg lhwage educ

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  1,  1898) =  161.61
       Model |  28.6408137     1  28.6408137           Prob > F      =  0.0000
    Residual |  336.358312  1898  .177217235           R-squared     =  0.0785
-------------+------------------------------           Adj R-squared =  0.0780
       Total |  364.999126  1899  .192205964           Root MSE      =  .42097
------------------------------------------------------------------------------
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0473669   .0037259    12.71   0.000     .0400596    .0546743
       _cons |   5.638038   .0516908   109.07   0.000     5.536662    5.739415
------------------------------------------------------------------------------
```

The OLS estimate suggests that an extra year of education is associated with a 4.7% increase in wages, (this is a log-lin regression so the coefficients are interpreted as semi-elasticities, so dLwage/dEduc = .047 = % change in wage /100 with respect to a unit change in education). t value suggests that this effect is statistically significantly different from zero.

Concerns over the possible endogeneity of the education variable (correlated with missing variables like ability or motivation and therefore correlated with the error term) mean that IV estimation may be more appropriate (assuming a good instrument can be found)

Using residence near a 2 year college (nearc2) when a teenager produces the following

```
. ivreg lhwage (educ=nearc2)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  1,  1898) =    4.83
       Model | -1352.91536     1 -1352.91536           Prob > F      =  0.0281
    Residual |  1717.91448  1898  .905118273           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  364.999126  1899  .192205964           Root MSE      =  .95138

------------------------------------------------------------------------------
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ |    .376345   .1712558     2.20   0.028     .0404756    .7122144
       _cons |   1.154414   2.334139     0.49   0.621    -3.423333    5.732161
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   nearc2
------------------------------------------------------------------------------
```

Now the coefficient on the instrumented endogenous education variable is around 8 times larger (as is the standard error) than in the OLS regression.

Remember the IV estimator is consistent, so that it performs much better in larger samples than small samples (where the IV estimator is biased)

However IV estimator will also be biased even in quite large samples if the correlation between the instrument and the endogenous variable is very small (even if the correlation between the instrument and the error term is low) – see lecture notes.

In the 2 variable model, the relative efficiency of OLS to IV is given by the square of the correlation coefficient between the endogenous variable and the instrument, (see problem set 6, question 7)

The correlation coefficient is

```
. corr nearc2 educ
(obs=1900)

             |   nearc2     educ
-------------+------------------
      nearc2 |   1.0000
        educ |   0.0492   1.0000
```

so relative efficiency $Var(b_{OLS})/Var(b_{IV}) = (.0492)^2 = .0024$

ie variance of OLS estimator is some 400 times smaller than IV using nearc2
Hence may weel be sensible to trade of bias in OLS for efficiency loss from IV in this case.

Check to see if heteroskedasticty is a problem

```
. ivreg lhwage (educ=nearc2), robust

Instrumental variables (2SLS) regression              Number of obs =    1900
                                                      F(  1,  1898) =    4.83
                                                      Prob > F      =  0.0281
                                                      R-squared     =       .
```

```
                                                      Root MSE      =  .95138
------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .376345   .1712928     2.20   0.028      .040403     .712287
       _cons |   1.154414   2.334604     0.49   0.621    -3.424246    5.733073
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   nearc2
------------------------------------------------------------------------------
```

It isn't. Standard errors almost unaffected by correction for heteroskedasticity (of unknown form)

Since IV estimates are radically different and standard error around estimate is very large, should check for weakness of instrument. Can do this by looking at 1st stage of 2SLS estimation –the regression of the endogenous variable on the instrument(s).

```
. ivreg lhwage (educ=nearc2), robust first

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

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  1,  1898) =    4.60
       Model |  30.8613295     1  30.8613295           Prob > F      =  0.0321
    Residual |  12734.5466  1898  6.70945551           R-squared     =  0.0024
-------------+------------------------------           Adj R-squared =  0.0019
       Total |  12765.4079  1899  6.72217372           Root MSE      =  2.5903
------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc2 |   .2567498   .1197144     2.14   0.032     .0219641    .4915355
       _cons |   13.51598   .0794096   170.21   0.000     13.36024    13.67172
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression              Number of obs =    1900
                                                      F(  1,  1898) =    4.83
                                                      Prob > F      =  0.0281
                                                      R-squared     =       .
                                                      Root MSE      =  .95138
------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .376345   .1712928     2.20   0.028      .040403     .712287
       _cons |   1.154414   2.334604     0.49   0.621    -3.424246    5.733073
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   nearc2
------------------------------------------------------------------------------
```

The 1st stage of the 2SLS regression above indicates that nearc2 and educ are not that closely related. While the t value on nearc2 is greater than two, the F value in the 1st stage regression (4.6) is less than the rule of thumb cutoff point of 10.

Conclude that nearc2 is a weak instrument for educ and that the IV estimates in this case are no more reliable than the OLS ones.

(you should always check that your instrument is correlated with the endogenous variable using the 1st stage regression)

Using a different instrument, mother's education

```
. ivreg lhwage (educ=motheduc), robust first

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

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  1,  1898) =  457.41
       Model |  2478.96879     1  2478.96879           Prob > F      =  0.0000
    Residual |  10286.4391  1898  5.41962018           R-squared     =  0.1942
-------------+------------------------------           Adj R-squared =  0.1938
       Total |  12765.4079  1899  6.72217372           Root MSE      =   2.328
------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |   .3781936   .0176833    21.39   0.000     .3435129    .4128744
       _cons |   9.614322   .1951628    49.26   0.000     9.231566    9.997078
------------------------------------------------------------------------------
```

This time the correlation of instrument and endogenous variable as revealed by the 1st stage of the 2sls regression, is much stronger. The t value >21 and the F value=457 which is obviously far higher than the rule of thumb value of 10.

```
. corr motheduc educ
(obs=1900)

             | motheduc     educ
-------------+------------------
    motheduc |   1.0000
        educ |   0.4407   1.0000
```

so relative efficiency $Var(b_{OLS})/Var(b_{IV}) = (.4407)^2 = .17$
ie variance of OLS estimator is now just some 5 times smaller than IV using motheduc

As a result, the second stage of the IV is much closer to the original OLS estimate. The IV standard error on educ is still somewhat larger than in OLS, (which is again what you would expect – see lecture notes). If we take the point IV estimate as true, this suggests that the omitted variable causing endogeneity, (say ability), is negatively correlated with education – (see lecture notes on omitted variable bias). Netting out this effect should raise the (true) coefficient on education.

```
Instrumental variables (2SLS) regression              Number of obs =    1900
                                                      F(  1,  1898) =   69.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0523
                                                      Root MSE      =  .42691
------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0747341   .0089501     8.35   0.000      .057181    .0922873
       _cons |   5.265052   .1218374    43.21   0.000     5.026103    5.504002
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   motheduc
------------------------------------------------------------------------------
```

When searching for more instruments remember that, asymptotically, more instruments means a more efficient IV estimator, but that in small samples more instruments **increases** the bias of the IV estimator. Sample is probably large enough to follow asymptotic rule, so the next step is to add both instruments.

```
. ivreg lhwage (educ=nearc2 motheduc), robust first

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

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  2,  1897) =  228.88
       Model |  2481.58157     2  1240.79079           Prob > F      =  0.0000
    Residual |  10283.8263  1897   5.4210998           R-squared     =  0.1944
-------------+------------------------------           Adj R-squared =  0.1935
       Total |  12765.4079  1899  6.72217372           Root MSE      =  2.3283

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc2 |   .0749413   .1079477     0.69   0.488    -.1367673      .28665
    motheduc |    .377218   .0177415    21.26   0.000     .3424232    .4120128
       _cons |   9.591705   .1978896    48.47   0.000     9.203601    9.979809
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression              Number of obs =    1900
                                                      F(  1,  1898) =   71.78
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0501
                                                      Root MSE      =  .42739

------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .075829   .0089505     8.47   0.000     .0582751    .0933828
       _cons |   5.250131   .1218241    43.10   0.000     5.011208    5.489054
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   nearc2 motheduc
------------------------------------------------------------------------------
```

Note IV estimate is little changed compared to that using motheduc as sole instrument. Not surprising since 1st stage of 2slsl confirms again that nearc2 is a weak (insignificant) instrument, more so when motheduc is included than before.

To do the **Wu-Hausman test** for endogeneity of the education variable

1. regress the potentially endogenous variable on *all* the **exogenous** variables in the *system* (ie those variables both in and outside the original equation)

```
. reg educ mothed nearc2

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  2,  1897) =  228.88
       Model |  2481.58157     2  1240.79079           Prob > F      =  0.0000
    Residual |  10283.8263  1897   5.4210998           R-squared     =  0.1944
-------------+------------------------------           Adj R-squared =  0.1935
       Total |  12765.4079  1899  6.72217372           Root MSE      =  2.3283

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |    .377218   .0177415    21.26   0.000     .3424232    .4120128
      nearc2 |   .0749413   .1079477     0.69   0.488    -.1367673      .28665
       _cons |   9.591705   .1978896    48.47   0.000     9.203601    9.979809
------------------------------------------------------------------------------
```

Save the residuals (or the predicted values) from this regression and include them as an extra variable in the original OLS specification
(if there were more than one endogenous right hand side variable, just repeat the above and include as may predicted residual values as there are endogenous variables)

```
. predict resaux, resid

. reg lhwage educ resaux

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  2,  1897) =   88.46
       Model |   31.136211     2  15.5681055           Prob > F      =  0.0000
    Residual |  333.862915  1897  .175995211           R-squared     =  0.0853
-------------+------------------------------           Adj R-squared =  0.0843
       Total |  364.999126  1899  .192205964           Root MSE      =  .41952

------------------------------------------------------------------------------
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .075829   .0084214     9.00   0.000     .0593127    .0923452
      resaux |  -.0353302   .0093827    -3.77   0.000    -.0537316   -.0169287
       _cons |   5.250131   .1151781    45.58   0.000     5.024242     5.47602
------------------------------------------------------------------------------
```

The t value on reshat (or phat) is above the critical value of statistical significance at the 5% level. (Note that the t values from using either the residuals or the predicted values are the same).

conclude that the null hypothesis of no endogeneity **can be rejected** at the 5% level.

As a check Stata will do the alternative form of the Hausman test –comparing the difference in OLS and IV coefficients net of sampling variation –If no endogeneity IV and OLS will be consistent but only OLS efficient. If endogeneity

present only IV is consistent and the IV and OLS coefficients will be a quite different.

```
quietly ivreg lhwage (educ=nearc2 motheduc), robust
hausman, save
quietly reg lhwage educ
hausman
```

```
                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |   Consistent   Efficient      Difference          S.E.
-------------+------------------------------------------------------------
        educ |     .075829     .0473669         .028462        .0081381
------------------------------------------------------------------------------
                          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

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       12.23
                Prob>chi2 =      0.0005
```

Again since the estimated chi-squared value exceeds the critical value at 1 degree of freedom (the number of potentially endogenous variables), the result is to reject the null of no endogeneity.

Can’t test exogeneity of all instruments, but if the equation is over-identified (more instructments than endogenous RHS variables can ask whether the additional instruments are exogenous).

first save the residuals from the original IV regression

```
quietly ivreg lhwage (educ=nearc2 educ), robust
predict resiv, resid
```

Then regress these residuals on **all** the **exogenous** variables in the **system** (inside and outside the original equation)

```
. reg resiv motheduc nearc2

      Source |       SS       df       MS              Number of obs =    1900
-------------+------------------------------           F(  2,  1897) =    7.78
       Model |  2.82218034     2  1.41109017           Prob > F      =  0.0004
    Residual |  343.877212  1897  .181274229           R-squared     =  0.0081
-------------+------------------------------           Adj R-squared =  0.0071
       Total |  346.699392  1899  .182569454           Root MSE      =  .42576

------------------------------------------------------------------------------
       resiv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0014275   .0032442    -0.44   0.660    -.0077902    .0049351
      nearc2 |   .0778455   .0197396     3.94   0.000     .0391319     .116559
       _cons |  -.0190985   .0361866    -0.53   0.598    -.0900682    .0518711
------------------------------------------------------------------------------
```

The test is an LM test of the form

$$N*R^2_{auxillary} = 1900*0.0081 = 15.4 \sim \chi^2(L-k)$$

Where the degrees of freedom equals the total number of exogenous variables in the system (9) minus the number of rhs variables in the original regression excluding the constant (6) = number of extra instruments at your disposal

So in this case we have used 2 instruments when we could have proceeded with just 1, so the degrees of freedom for the test are 1 = 4-3

From tables $\hat{\chi}^2 > \chi^2_{(1)}$critical= 3.84 so reject null that additional instruments are uncorrelated with the structural form residuals.

Note that this result is driven by the significance of nearc2 in the auxillary regression. So may be better to drop this variable.

Hausman version of this test given by

```
quietly ivreg lhwage (educ=nearc2), robust
hausman, save
quietly ivreg lhwage (educ=nearc2 educ), robust

hausman
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |   Consistent   Efficient      Difference          S.E.
-------------+----------------------------------------------------------------
        educ |     .376345     .0473669        .3289781        .1712502
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
            B = inconsistent under Ha, efficient under Ho; obtained from ivreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        3.69
                Prob>chi2 =      0.0547
```

Note test can vary (in finite samples) depending on order in which compare variables

```
quetly vreg lhwage (educ=mothed), robust
hausman, save
quietly hwage (educ=nearc2 educ), robust

hausman
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |   Consistent   Efficient      Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .0747341     .0473669        .0273672        .0080944
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
            B = inconsistent under Ha, efficient under Ho; obtained from ivreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       11.43
                Prob>chi2 =      0.0007
```