NYU | Wagner
Advanced Empirical Methods
Rajeev Dehejia

Problem Set 2: Instrumental Variables

Part I

1. Consider the following instruemental variable regression output from Stata.

```
. sysuse auto
. ivreg price (mpg = displacement),first

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

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     71.41
       Model |  1216.67534         1  1216.67534   Prob > F        =    0.0000
    Residual |  1226.78412        72  17.0386683   R-squared       =    0.4979
-------------+----------------------------------   Adj R-squared   =    0.4910
       Total |  2443.45946        73  33.4720474   Root MSE        =    4.1278

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
displacement |  -.0444536   .0052606    -8.45   0.000    -.0549405   -.0339668
       _cons |   30.06788   1.143462    26.30   0.000     27.78843    32.34733
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     21.13
       Model |   105028215         1   105028215   Prob > F        =    0.0000
    Residual |   530037181        72  7361627.51   R-squared       =    0.1654
-------------+----------------------------------   Adj R-squared   =    0.1538
       Total |   635065396        73  8699525.97   Root MSE        =    2713.2

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -357.5834   77.78566    -4.60   0.000    -512.6463   -202.5205
       _cons |   13780.82   1686.382     8.17   0.000     10419.08    17142.56
------------------------------------------------------------------------------
Instrumented:  mpg
Instruments:   displacement
------------------------------------------------------------------------------
```

This is a standard "test" dataset often used in Stata. You can learn a bit it by loading using the first command above. Interpret the above regression output: the coefficients' economic and statistical meaning, and the accompanying statistical output. Discuss the plausibility of the instrument.

Part II. Stata Problem. In this problem, you will use data and replicate some of the results from Ashenfelter, Orley and Alan Krueger (1994), "Estimates of the Economic Returns to Schooling from a New Sample of Twins." *American Economic Review* 84(5): December, p. 1157-73. If you are interested, you could read this study; a summary of the relevant issues appears in chapter 6 of the Angrist and Pischke textbook.

Background. Ashenfelter and Kreuger, two Princeton economists, descended on Twinsburg, Ohio's annual Twin Festival. They collected data on the education levels and earnings of a large sample of twins to see if they could use a panel of twin families to get rid of bias in the OLS estimates of the returns to schooling.

Data source: pubtwins.dta
To do these exercises, you will use the Stata data set pubtwins.dta, which is available on our course web directory. You will have to create additional variables from pubtwins.dta to conduct the analysis.

Data notes:

- The data set contains 680 observations on individuals ordered by twin pairs. There are therefore 340 twin pairs, with the first two observations representing the first twin pair, the next two observations representing the second twin pair, etc.
- Key variables:

  hrwage = the self-reported hourly wage of the individual (in dollars)

  lwage = the natural log (ln) of the hourly wage

  age and age2 = the age of an individual and its square (age2)

  female = an indicator variable equal to one if the person is female, zero otherwise

  white = an indicator variable equal to one if the person is white, zero otherwise

  educ = the educational attainment of the individual

  educt_t = the other twin's report of the individual's education

  first = an indicator equal to one if the twin was the first-born (equal to . otherwise)

  dlwage = the difference in the log wages of twins

  deduc = the difference in twins' education based on their self-reports

  deduct = the difference in twins' education based on each twin's report of the other twin's education

Note: As always, you should do this problem set by writing a program (a *.do file in Stata). There is a template program on our site. In your solution packet, include a well-annotated log file program, as well as relevant STATA output. And for this problem set, **I will also require you also to turn in an "outreg2" table** (worth one point of your problem set.) Recall that you must first type

ssc install outreg2

on STATA's command line to get the command to work (you only ever have to do this once). I have put the first two outreg2 commands into the template program to help get you started.

1. Cross-section regressions

a. Run the bivariate regression of log wages on a constant and education and show the scatter plot.

b. Regress log wages on a constant, education, age, age2, and white and female. Compare the estimated return to education from the multivariate regression to the one from the bivariate regression in (a). Are they different?

c. Regress education on the other controls in (b). Is it significantly related to these other controls? Can you think of variables that we have not controlled for that may be related to both educational attainment and earnings? What does this imply about how we should interpret the least squares estimate of the relation between log wages and education?

Make sure all of your regressions include the robust standard error correction for heteroskedasticity!

2. Standard Errors

a. There is no variable which identifies twin pairs in the data, and we will need such a variable in order to use our panel data techniques. So use the following STATA commands to create a variable that separately identifies each twin pair in the data set (Note: the data must be in its original order for this to work):

```
gen pairno = round(_n/2,1)
```

Type "browse pairno" and see what this command did. (By the way, "_n" is the "observation #." No further answer is required to this question!

b. Run the regression of log wages on education, age, age2, female, and white applying the cluster option in STATA to this "pairno" variable to correct the estimated standard errors for error correlation between twins. Explain why the standard errors on the estimated return to education are higher (and t-ratio lower) than when clustering is not corrected for.

3. Measurement Error in the Cross Section

a. Suppose that a twin's self-report of education is an imperfect measure of the twin's actual educational attainment due to misreporting. In addition, suppose that this measurement error is "classical" (or, loosely speaking, "just noise" uncorrelated with anything.) In this case you have a second (independent) report of years of education: each twin was asked about the education of the other twin. This is recorded

in the "educt_t" variable. Regress educt_t on educ. What would be the coefficient estimate if there were no measurement error? What does this slope coefficient instead represent?

b. If you have two independent reports you can also correct estimates for measurement error using instrumental variables regressions. So now run the following STATA commands: `ivreg lwage (educ = educt_t)` AND `ivreg lwage (educ = educt_t) age age2 female white`. This performs two-stage least squares estimation of the return to education using the other twin's report of the individual's education level as an instrument for the individual's self-reported education. How do these estimates compare to the ones you got in 1(a) and 1(b). Are the results consistent with your estimates of the attenuation factor?