SGPE Econometrics Lab 7:
(1) White’s test for heteroskedasticity
(2) OLS, IV and MSE

ASSIGNMENT
DUE DATE: MIDNIGHT, SUNDAY 18 NOVEMBER

Mark Schaffer
version of 12.11.2012

Introduction

This assignment has two parts.

In Assignment 1, you extend the results from Lab 7 for White’s general test for heteroskedasticity to other specifications.

In Assignment 2, you extend the Monte Carlo results from Lab 7 that compare OLS and IV to other specifications.

Assignment 1

The original version of White's general test can suffer from a loss of power because it is looking in too many different directions for heteroskedasticity. This will show up in a large number of degrees of freedom. We can increase the power and reduce the degrees of freedom by decreasing the dimension of the vector psi in our vector of contrasts test. The tradeoff here is that because these tests look in fewer directions for heteroskedasticity, the might be looking in the “wrong” directions and miss heteroskedasticity that is actually there.

You will calculate and report three different versions of White’s test:

1. Using only the levels of regressors as elements in psi.
2. Using the predicted value of the dependent variable as the sole element in psi. The predicted value is (of course) a linear combination of all the regressors.
3. Using the predicted value of the dependent variable and its square as the two elements of psi

In all cases you will use the MRW dataset and a simple specification for the OLS estimation with two explanatory variables, log GDP per capita in 1960 and the I/Y ratio. The squared residuals in the White test are the same for all 3 tests above.

```
reg growth lngdp60 igdp
capture drop ehat
predict double ehat, res
gen double ehatsq=ehat^2
```

The core code for the assignment are in the do file Lab7_Assignment1_het.do. Your assignment:

1. Insert the code that executes the 3 versions of White's test above into the do file.
2. For versions 1 and 2, verify that your code reproduces the results of Stata's official postestimation command for OLS, **`estat hettest`**.
3. Submit the following:
    a. Your modified do file.
    b. A log file of the results.
    c. A short discussion of the results of the testing.

Assignment 2

In Lab 7, we examined the relative performance of the OLS and IV estimators using a Monte Carlo approach. The criteria for comparing the estimators was MSE (mean squared error). MSE is a loss function that puts equal weight on variance and squared bias. A biased and inconsistent estimator such as OLS can, in some circumstances, out-perform a consistent estimator such as IV according to the MSE criterion if the advantages of the smaller OLS variance outweigh the disadvantages of the larger OLS bias.

In the lab, we looked at two specifications:

Specification 1: $cov(x_i, \varepsilon_i) = 0.3$
$cov(x_i, z_i) = 0.5$
n=100

Specification 2: $cov(x_i, \varepsilon_i) = 0.1$
$cov(x_i, z_i) = 0.3$
n=100

In this task, you will look at two more specifications:

Specification 3: $cov(x_i, \varepsilon_i) = 0.3$
$cov(x_i, z_i) = 0.5$
n=1,000

Specification 4: $cov(x_i, \varepsilon_i) = 0.1$
$cov(x_i, z_i) = 0.3$
n=1,000

The only difference between Specifications 1 and 2 (lab) and Specifications 3 and 4 (assignment) is that the lab specifications set the sample size n=100; the assignment specifications set n=1,000.

The code to run all four specifications is in the do file Lab7_Assignment2_endog.do. The do file also has code to create graphs and combine them.

Your assignment:

1. Run all 4 specifications.
2. Create any graphs you will need for your writeup.
3. Write a short analysis of the results and your conclusions.

CODE FOR ASSIGNMENT 1 (Lab7_Assignment1_het.do)

```
capture cd M:\Econometrics\Lab7

capture log close
log using Lab7_Assignment1, replace text

use mrw1992, clear

* Keep non-oil-exporters
keep if nonoil

* Keep only the variables we need
keep c_name gdp60 gdp85 pop igdp school

gen lngdp60 = ln(gdp60)
gen lngdp85 = ln(gdp85)

gen growth = (lngdp85 - lngdp60) / 25
label variable growth "average growth rate"

*************************************************************************
*
* Assignment 1: Replicate Stata's built-in estat hettest

reg growth lngdp60 igdp
capture drop ehat
predict double ehat, res
gen double ehatsq=ehat^2

* The original version of White's general test can suffer from a loss of
* power because it is looking in too many different directions for
* heteroskedasticity.  This will show up in a large number of degrees
* of freedom.  We can increase the power and reduce the degrees of
* freedom by decreasing the dimension of the vector psi in our
* vector of contrasts test.

* Simplest version - use only the levels of the regressors, and
* forget about the squares and cross-products.

*** INSERT YOUR CODE HERE ***
di e(N)*e(r2)

* Confirm it matches Stata's official estat hettest with the
* rhs and iid option.
* NB: omitting the iid option means hettest reports the default
* Breusch-Pagan/Godfrey version, which assumes normality.  The
* assumption of normality is often too strong in economic applications,
* so the iid option (the White version) is preferable.

reg growth lngdp60 igdp
estat hettest, rhs iid

* An alternative is, instead of using the levels of the regressors,
* use a linear combination of the levels of the regressors.  The
* obvious linear combination is the predicted values (yhat) from the
* regression.  Do the following:
* 1. Estimate the original equation.
* 2. Generate the predicted values.
* 3.  Run the White-style artificial regression and report the
*     NR2 test statistic.

*** INSERT YOUR CODE HERE ***
di e(N)*e(r2)

* Confirm it matches Stata's official estat hettest with the iid option.

reg growth lngdp60 igdp
estat hettest, iid
```

```
* A third alternative is to use a summary combination of levels
* and squares of the regressors, namely the predicted values
* (yhat) and the squares of the predicted values (yhat^2).

*** INSERT YOUR CODE HERE ***
di e(N)*e(r2)

* You can confirm this vs. the output of the Stata add-in ivhettest.
* ivhettest performs tests for heteroskedasticity for OLS and IV
* estimations, and includes a wide range of options.  See help ivhettest
* for a discussion.  You should use the fitsq option.

* To install ivhettest (you need do this only once):
ssc install ivhettest

* Confirm your test statistic matches that from ihvettest.

reg growth lngdp60 igdp
ivhettest, fitsq

capture log close
```

CODE FOR ASSIGNMENT 2 (Lab7_Assignment2_endog.do)

```
capture cd M:\Econometrics\Lab7

capture program drop mysimendog
program define mysimendog, rclass
      drop _all
      set obs $n

* x = regressor, possibly endogenous = correlated with e
* e = error
* z = instrument, correlated with x but not with e

* xe_cov = corr(x,e)
* xe_cov = 0 => x is exogenous
* xe_cov /ne 0 => x is ENDOGENOUS

* xz_cov = corr(x,z)
* xz_cov = 0 => instrument is useless
* xz_cov /ne 0 => instrument is "relevant"
* Bigger xz_cov, "stronger" instrument
* By assumption:
* corr(z,0) = 0

* Create x, z and e using the drawnorm command.
* The covariance matrix is V.
      drawnorm x z e, cov(V)

* Create the dependent variable.
* In the true model, b1 (the constant) =0 and b (the coeff on x) =1.
      gen y = x + e

* Run the regression using OLS
      reg y x
      return scalar b_ols=_coef[x]

* Run the regression using IV
* Use the simple, old and now undocumented "ivreg" command - very fast.
      ivreg y (x = z)
      return scalar b_iv=_coef[x]

end
```

```
***********************************************************************
******************** Simulate *****************************************
***********************************************************************

* Create the covariance matrix V(x,z,e):

* V =
*   |    1       xz_cov    xe_cov  |
*   |  xz_cov      1         0     |
*   |  xe_cov      0         1     |

* Column/row 1: x
* Column/row 2: z
* Column/row 3: e
* The ones are the variances of x, z and e.
* The zeros are the covariances of z with u.
* A zero covariance with e makes z an exogenous instrument.
* A nonzero covariance with x makes z a "relevant" instrument.
* Note that be using var(x)=var(z)=var(e)=1, the covariances
* can also be interpreted as correlations.

* We do this using Stata matrix mini-language (not Mata!).
* This matrix will be used by simulate and mysimendog.

* To run a simulation, remove the comment delimiters /* and */
* (and put them back around the other ones).

/*
* Specification 1:
global simnumber=1
global simname "Cov(x,e)=0.3; Cov(x,z)=0.5; n=100"
global xe_cov=0.3
global xz_cov=0.5
global n=100
*/

/*
* Specification 2:
global simnumber=2
global simname "Cov(x,e)=0.1; Cov(x,z)=0.3; n=100"
global xe_cov=0.1
global xz_cov=0.3
global n=100
*/

/*
* Specification 3:
global simnumber=3
global simname "Cov(x,e)=0.3; Cov(x,z)=0.5; n=1,000"
global xe_cov=0.3
global xz_cov=0.5
global n=1000
*/

/*
* Specification 4:
global simnumber=4
global simname "Cov(x,e)=0.1; Cov(x,z)=0.3; n=1,000"
global xe_cov=0.1
global xz_cov=0.3
global n=1000
*/

mat V = (1, $xz_cov , $xe_cov \ $xz_cov, 1, 0 \ $xe_cov, 0, 1)
mat list V
```

```
* Simulate.  Obtain coefficient estimates and standard errors.
set more off
* Control the random number seed so that the results are replicatable.
set seed 1
simulate                                              ///
                  b_ols=r(b_ols)                      ///
                  b_iv=r(b_iv)                        ///
                  , reps(10000): mysimendog
***********************************************************************
***********************************************************************


**************** Address the "No Moments" Problem *********************
* The IV estimator b_iv has moments up to the degree of
* overidentification, L-K.  Here, the equation is exactly
* identified, L=K, so the IV estimator has no moments at all.
* Thus an attempt to estimate the population mean of b_iv - the first
* moment - with the population mean will fail, because we the sample
* mean can't converge to the population mean if the population mean
* doesn't exist!  The same applies to the second moment: the variance
* of b_iv also doesn't exist in the L=K case.
* If the mean of b_iv doesn't exist, we can't talk about bias or
* the MSE criterion.
* We address this problem by imposing the assumption for b_iv that
* we know the true beta lies in the interval [0,2].  This will give
* our new IV estimator a mean and a variance.  We do the same for OLS.
replace b_iv=0  if b_iv<0
replace b_iv=2  if b_iv>2
replace b_ols=0 if b_ols<0
replace b_ols=2 if b_ols>2
***********************************************************************


* Are the estimates for b biased or unbiased?
twoway
      ///
            (kdensity b_ols if b_ols>0 & b_ols<2)           ///
            (kdensity b_iv if b_iv>0 & b_iv<2)              ///
            , xlabel(0 0.5 1 1.5 2) xline(1)                ///
            title(Distribution of b)                        ///
            subtitle("$simname")                            ///
            name(bias$simnumber, replace)

* Which estimator performs better in terms of bias?
* Which estimator performs better in terms of variance?
* (Just look at the standard deviations.)
sum b_ols b_iv


* Which estimator performs better in terms of MSE (mean squared error)?
gen mse_ols=(b_ols-1)^2
gen mse_iv =(b_iv -1)^2
sum mse_ols mse_iv

* Do this after all simulations have been run:
* graph combine bias1 bias2 bias3 bias4
* graph export bias1_2_3_4.emf, replace
```