SGPE Econometrics Lab 7:
(1) Testing for heteroskedasticity
(2) OLS vs. IV – bias, variance, and MSE
(3) IV/GMM estimation

Mark Schaffer
version of 12.11.2012

Introduction

This lab has three parts. In the first part, you perform White's general test for heteroskedasticity using the MRW dataset. In the second part, you explore the finite-sample performance of OLS and IV using Monte Carlos. In the third part, you perform IV/GMM estimation using a real dataset, replicating some of the results in Hayashi chapter 3.

The post-lab assignment has two parts. In the first part, you extend the heteroskedasticity tests to other specifications. In the second part, you extend the Monte Carlos to other specifications and assess the results.

Task 1: Preparation

In the first part of the lab we will be working again with the Mankiw-Romer-Weil (1992) dataset. In this lab we use the same subset of the dataset as MRW: non-oil-exporting countries. For simplicity, we will use only two regressors: initial GDP per capita and the I/Y ratio.

Open the Lab7_Tasks1_2_3 do file and execute the code at the top: load the data, drop the unwanted observations, and create variables.

Next, estimate the basic model and then use the FWL (Frisch-Waugh-Lovell) theorem to examine the separate contribution of the different explanatory variables. We've seen earlier how to do this variable-by-variable using the **`avplot`** command multiple times to create added-variable plots, and also how to do it for all the regressors in a single command by using the **`avplots`** command (note the "s").

```
reg growth lngdp60 igdp
avplots
```

Discuss the results amongst yourselves.

Next, create the residuals and inspect them visually to see if it is plausible that they are heteroskedastic. The code below from the do file creates one graph that displays all the scatterplots. Note that we use the **`nodraw`** option to stop the individual graphs from being separately displayed as well (since we need only the combined graph).

```
capture drop ehat
predict ehat, res
scatter ehat lngdp60, nodraw name(GDP, replace)
scatter ehat igdp,    nodraw name(IY, replace)
graph combine GDP IY
```

What do you think? Can you see patterns that suggest conditional heteroskedasticity?

Lastly, perform the intra-ocular version of White's test (does it "hit you between the eyes?"). Estimate the model with classical and with heteroskedastic-consistent (HC) ("robust") standard errors:

```
reg growth lngdp60 igdp
reg growth lngdp60 igdp, rob
```

Are the SEs different? In what way? What does this suggest?

Task 2: White's general test for heteroskedasticity

Reminder: White's general test is based on a vector of contrasts between the classical **S** matrix and the heteroskedastic-robust **S** matrix. The **S** matrix is the "filling" in the sandwich variance formula for the OLS estimator, and the elements of the vector of contrasts correspond to the elements of the covariance matrix: levels, squares and cross-products of regressors. (The levels are the cross-products of the constant with the other regressors.) We use only the non-constant elements (so we exclude the constant). The test is the sample size ***n*** times the R^2 from a regression of the squared residuals on this vector.

We also drop any redundant elements in the vector. In Task 3 you will see an example of how redundant elements can arise.

To perform the test, first create the squared residuals. The lines for this are in the do file:

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

Next, create the required variables for the vector of contrasts. You should insert code in the do file that does this. They are:

(1) Levels: `lngdp60` and `igdp`. These variables already exist.
(2) Squares: `lngdp60^2` and `igdp^2`. Create these variables.
(3) Cross-products: `lngdp60` * `igdp`. Create this variable.

Then run the artificial regression with `ehatsq` as the dependent variable and the variables above as the regressors.

Finally, report the ***n***R^2 statistic:

```
di e(N)*e(r2)
```

And confirm that you have performed the test properly by replicating it using Stata's built-in postestimation command for the "information matrix" (IM) test, **`estat imtest`** with the **`white`** option:

```
reg growth lngdp60 igdp
estat imtest, white
```

How do you interpret the results?

Task 3: White's general test with redundancies

Redundancies in the variables used in the vector of contrasts can arise for various reasons. One obvious one is if you are using a quadratic specification.

Estimate the MRW model as a quadratic in initial GDP per capita:

```
gen lngdp60sq=lngdp60^2

reg growth lngdp60 lngdp60sq
```

Note that, as in Task 2, there are two explanatory variables. However, because of redundancies the test will **not** have the same degrees of freedom this time.

Perform White's general test as before, assembling the levels, squares and cross-products of the two regressors. When you do this, do you find you are creating a variable that already exists?

Confirm that your test statistic matches that reported by Stata's official **`imtest`** with the **`white`** option:

```
reg growth lngdp60 lngdp60sq
estat imtest, white
```

Task 4: OLS vs. IV– bias, variance and MSE

Say a regressor x_i is correlated with the error, i.e., the assumption of weak exogeneity fails. Then OLS estimation will give you a **biased** (finite sample perspective) and **inconsistent** (large sample perspective) estimate β_{OLS} of the true slope parameter β.

This problem can be addressed by the method of instrumental variables (IV) or its generalization the Generalized Method of Moments (GMM). If we have a variable z_i that is both weakly exogenous (uncorrelated with the error) and correlated with x_i, we can obtain a **consistent** estimate $\hat{\beta}_{IV}$ of the true β.

IV/GMM estimators have a large-sample justification, i.e., consistency. This is their key advantage over OLS. In finite samples, IV/GMM estimators may be biased. (Or, even worse, their mean may not even exist! More about this below.)

Thus one practical issue is the finite-sample performance of IV/GMM estimators in terms of bias.

But unbiasedness/consistency is not the only thing we care about in an estimator. We also want our estimator to be precise – to have a small variance.

This leads to an important practical question: might we be willing to sacrifice unbiasedness/consistency in exchange for a big improvement in variance? Remember the Hiawatha poem from earlier in the semester: at the end, Hiawatha, the mighty hunter and archer, sits alone in the forest thinking about shooting arrows and

> Wondering in idle moments
> Whether an increased precision
> Might perhaps be rather better,
> Even at the risk of bias,
> If thereby one, now and then, could
> Register upon the target.

The tradeoff between bias/consistency and variance is the subject of this task.

The OLS estimator $\hat{\beta}_{OLS}$ has a smaller variance than the IV estimator $\hat{\beta}_{IV}$, but it may be biased/inconsistent. In what circumstances might we prefer to use the OLS estimator even though it's biased?

To answer this question, we need a **loss function**: a way of weighting the payoff to smaller bias vs. the payoff to smaller variance. There is no "right" loss function: some researcher or applications may use different weights for various reasons. But some loss functions are more popular than others, and one commonly used loss function is **mean squared error** (MSE).

In our application, we want to know the MSE of an estimator $\hat{\beta}$:

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]$$

Define the bias of the estimator $\hat{\beta}$ as:

$$Bias(\hat{\beta}) = E[\hat{\beta} - \beta]$$

It can be shown that the MSE is the sum of the squared bias of estimator and the variance of the estimator:

$$MSE(\hat{\beta}) = E\left[(\hat{\beta} - \beta)^2\right] = \left[E[\hat{\beta} - \beta]\right]^2 + Var(\hat{\beta})$$

Note that if the estimator is unbiased, the MSE criterion amounts to choosing an estimator with a small variance, i.e., one that is more efficient (a criterion we've seen before – remember BLUE).

MSE is a very natural loss function for evaluating $\hat{\beta}$, because it weights the squared bias and the variance equally. Using the MSE criterion, we might be willing to accept a larger bias in exchange for a big decrease in the variance.

We will investigate the circumstances where OLS might be preferable to IV, even though the OLS estimator is inconsistent (because of endogeneity) and the IV estimator is consistent.

The DGP (data generation process) is as follows. We generate n observations according to the following specification:

$$n = 100$$

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$\beta_1 = 0$$
$$\beta_2 = 1$$

$$\begin{bmatrix} x_i \\ z_i \\ \varepsilon_i \end{bmatrix} \sim N(0, V) \qquad V = \begin{bmatrix} 1 & & \\ cov(x_i, z_i) & 1 & \\ cov(x_i, \varepsilon_i) & 0 & 1 \end{bmatrix}$$

where we report only the lower-triangle of V because it's symmetric.

The three variables – the regressor x_i, the instrument z_i, and the error ε_i – are distributed as multivariate normal with zero mean and variances=1. Note that this means we can interpret the covariances below as correlations.

$cov(x_i, \varepsilon_i) \neq 0$: this is the **endogeneity problem**. Our regressor x_i is not weakly exogenous. The bigger this number, the bigger the endogeneity problem and the bigger the bias in $\hat{\beta}_{OLS}$. $cov(x_i, \varepsilon_i)$ is one of the parameters we will vary in the Monte Carlo.

$cov(z_i, \varepsilon_i) = 0$: this is the 0 in the [3,2] entry. This is the requirement that our instrument z_i is **weakly exogenous**.

$cov(x_i, z_i) \neq 0$: this is the **rank condition**, also known as the requirement of **instrument relevance**. Our instrument z_i must be correlated with our regressor x_i. The bigger this number, the stronger the correlation – the stronger our instrument – and the better our IV estimator will perform. $cov(x_i, z_i)$ is one of the parameters we will vary in the Monte Carlo.

In this task, we run two different specifications:

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

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

The simulations are in the do file Lab7_Task4_endog.do.

The variables are created in the program `mysimendog` by the lines:

```
drawnorm x z e, cov(V)
gen y = x + e
```

The `drawnorm` command takes the specified covariance matrix `V` and draws from the multivariate normal (the default mean is 0). The matrix V is constructed using Stata's mini-matrix language (not Mata!) prior to calling `simulate`. Thus Specification 1 is:

```
/*
* 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
*/
```

To run Specification 1, just remove the comment commands `/*` and `*/`. (Everything between `/*` and `*/` is treated by Stata as a text comment and ignored.)

The line

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

creates the matrix `V` that `mysemendog` and `drawnorm` use. After the matrix is created, for reference it is displayed on the screen, e.g.,

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

. mat list V

symmetric V[3,3]
    c1  c2  c3
r1   1
r2  .5   1
r3  .3   0   1
```

The estimation lines use Stata's `regress` command and an undocumented command for IV estimation called `ivreg`.

```
reg y x
return scalar b_ols=_coef[x]
ivreg y (x = z)
return scalar b_iv=_coef[x]
```

The syntax of ivreg is

```
ivreg <depvar> <exog vars> (<endog vars> = <instruments>)
```

The reason we use **`ivreg`** instead of the more powerful add-in **`ivreg2`** or Stata's official **`ivregress`** is that **`ivreg`** is much faster (because it reports only the IV estimates with no specification tests etc.).

The MC is run 10,000 times. At the end, you will have a dataset with 10,000 realizations of **`b_ols`** and **`b_iv`**.

You should: (1) examine the distribution of the two estimators graphically; (2) look at the means and assess the finite sample bias; (3) look at the variances (or standard deviations) and assess the precision of the estimators; (4) look at the MSE of the two estimators.

What do you conclude?

How would your results change if Specification 1

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

were instead

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

or instead

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

(You should be able to answer this without running Specifications 1' or 1'' – just the original Specification 1 plus some reasoning will tell you the answer.)

Addendum: The "No Moments" problem with IV

To analyse the finite sample performance of the IV estimator in this way means we looking at the sample mean and sample variance (across 10,000 MC realizations) of $\hat{\beta}_{IV}$. We are using these sample means and variances to estimate the population means and variances of of $\hat{\beta}_{IV}$.

What we are doing requires these population means and variances to exist. For example, you can't talk about the bias of the IV estimator,

$$Bias(\hat{\beta}_{IV}) = E[\hat{\beta}_{IV} - \beta] = E[\hat{\beta}_{IV}] - \beta$$

if $E[\hat{\beta}_{IV}]$ doesn't exist. And if it doesn't exist, what does the sample mean of 10,000 realizations of $\hat{\beta}_{IV}$ mean? It isn't converging to anything – the Law of Large Numbers doesn't apply. And the same applies to $Var(\hat{\beta}_{IV})$ – if it doesn't exist, then trying to estimate it with the sample variance doesn't make sense.

Unfortunately, this is **exactly** the problem we have with the IV estimator above.

It has been shown that IV estimator $\hat{\beta}_{IV}$ has moments up to the degree of overidentification, L-K, where L=number of instruments and K=number of regressors. In the exactly-identified case above, L-K=0 and the IV estimator $\hat{\beta}_{IV}$ has no moments at all. The integral $E[\hat{\beta}_{IV}]$ doesn't converge, and neither does $Var(\hat{\beta}_{IV})$.

This is a good example of the limitations of finite-sample theory and why we use large-sample theory so often in econometrics: although $E[\hat{\beta}_{IV}]$ does not exist in the exactly-identified case, $plim[\hat{\beta}_{IV}]$ certainly does!

So what we do is ad hoc: we say we estimate using IV and OLS with the prior knowledge that the true β lies on the interval [0,2]. This will give them finite means and variances. This is what the code in the do file does:

```
************** Address the "No Moments" Problem **************
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
**************************************************************
```

For a good discussion of the no-moments problem for IV and related estimators, including some Monte Carlos, see Russell Davidson and James G. MacKinnon, "Moments of IV and JIVE Estimators", June 2007, available at: http://russell-davidson.arts.mcgill.ca/articles/jive2.pdf.

Task 5: IV/GMM estimation of a cross-section model

In this task you will estimate a standard cross-section model of the return to education. The example is taken from Hayashi (2000), Chapter 3, pp. 236-244 and 250-255. The original dataset and paper are due to Griliches (1976), "Wages of Very Young Men", *Journal of Political Economy*, Vol. 84, pp. S69-S85. Further readings are cited in Hayashi.

A typical wage equation in this literature looks like this:

$$\ln(w_i) = \alpha + \beta S_i + x_i'\boldsymbol{\delta} + \varepsilon_i \qquad (1)$$

where we vary from Hayashi's notation slightly (the above is more consistent with the lectures).

$\ln(w_i)$ Log of wage received by individual i

S_i	Years of schooling of individual i
x_i'	Vector of other explanatory variables
ε_i	Error

The parameter of interest is β, the coefficient on years of schooling S. The interpretation of β is that it is an estimate of the "return to education". Estimates of β are typically in the neighbourhood of 10%, i.e., education is a good investment.

The standard problem is **endogeneity bias**, and the particular version of it that shows up here is **omitted variable bias**. The omitted variable is

A_i	"Ability" of individual i

The simplified version of the endogeneity problem:

1. Ability A_i is not observable. Thus A_i is "inside" the error ε_i.
2. High ability people have high productivity and therefore earn higher wages w_i.
3. Smart people have high ability A_i.
4. Smart people are likely to get more schooling S_i. They go to university etc. more often than not-so-smart people.
5. (4) and (5) imply schooling S_i is correlated with unobserved ability A_i.
6. (5) and (1) imply that schooling S_i is correlated with the error ε_i via ability A_i. S_i is **not** weakly exogenous. Hence we have an endogeneity problem.
7. (6) and (2) imply that if we use OLS, the coefficient β will be **biased upwards**. Intuitively, high ability explains part of high wages, but because we don't observe ability, this is being attributed to schooling instead.

Note that if (4) did not hold, we would have an omitted variable A_i but we would *not* have an omitted variable bias in β. For that we also need (observable) schooling S_i to be correlated with (unobservable) ability A_i.

The usual approach in the literature is to look for instruments that are (a) correlated with schooling S_i and (b) uncorrelated with the error ε_i (including the omitted variable ability A_i).

The Griliches approach is different. He includes a direct measure of ability, namely IQ score IQ_i:

$$\ln(w_i) = \alpha + \beta S_i + \gamma IQ_i + x_i'\boldsymbol{\delta} + \varepsilon_i \tag{2}$$

An IQ test is a kind of "intelligence test". If IQ_i were a perfect measure of ability A_i, that would be enough for OLS to give us a consistent estimate of the return to schooling β, because now S_i is not endogenous – A_i has been moved out of the error ε_i.

However, IQ_i is probably **not** a perfect measure of ability A_i. It is only a "proxy" for ability, and an imperfect one at that. For this reason, it is possible that S_i is still endogenous – still correlated with whatever is still in ε_i – and therefore OLS will generate a biased estimate of β.

Because ability is measured with error, the coefficient γ on IQ_i may **also** suffer from endogeneity bias. Note that **if** the measurement error is "classical" – uncorrelated with anything – then the impact on the OLS estimate of β is "attenuation bias", i.e., it will be biased downwards, towards zero. But this is a big "if"! The measurement error in IQ_i may not be classical – it might be correlated with the error ε_i, and that could generate endogeneity bias in the coefficient γ. Moreover, the endogeneity bias related to IQ_i could also spill over and affect the estimate of β (see the discussion in Hayashi).

We will estimate equation (2) using several different combinations of specifications. Our core dataset is 758 observations of young men in the late 1960s in the US.

In brief:

Regressors of interest, may be treated as endogenous or exogenous:

S_i	Schooling
IQ_i	Score on IQ ("intelligence") test

Exogenous regressors (controls):

$EXPR_i$	Work experience in years
$TENURE_i$	Job tenure in years
$SMSA_i$	Resident in a city
(year)	Dummies for year

Excluded instruments (exogenous):

MED_i	Mother's education
KWW_i	"Knowledge of the World of Work" (a test score)
AGE_i	Age of the individual
MRT_i	Marital status dummy, =1 if married, =0 if single

We consider two different kinds of estimation: Feasible Efficient GMM and inefficient GMM.

In a Feasible Efficient GMM estimation, the weighting matrix W_n is the inverse of the estimated covariance matrix of orthogonality conditions S. The formula for the asymptotic variance of the estimator has a non-sandwich form.

In an inefficient GMM estimation, the weighting matrix is *not* the inverse of S. The formula for the asymptotic variance of the estimator *does* have a sandwich form, and this formula *does* us S.

Whether a GMM estimation is efficient or inefficient depends S, and therefore on the assumptions we make about obtaining a consistent S. We will consider two possibilities:

$S_{classical}$	We assume independence and homoskedasticity.
S_{HC}	We assume independence but allow arbitrary heteroskedasticty.

We consider several different estimations. Our main focus in this task is on the orthogonality conditions and estimation methods used. In each case, you should look in particular at:

1. What is being treated as endogenous? Exogenous?
2. The coefficient estimates. Are they the same or different from other estimations?
3. The SEs. Are they the same or different from other estimations?
4. The overidentification statistic, and in particular: (a) How many degrees of freedom does it have? (b) What does the test statistic mean?

Preparation

First, you should install the **`ivreg2`** add-in. This is an IV/GMM estimator with a wide variety of options and specification tests. You can install it by typing

```
findit ivreg2
```

and clicking on the links, or by typing

```
ssc install ivreg2
```

The syntax of **`ivreg2`**:

```
ivreg2 <depvar> <exog vars> (<endog vars> = <instruments>), options
```

The main options:

`robust`	Use the heteroskedastic-consistent S_{HC} for the $AVar$
`gmm2s`	Use 2-step Feasible Efficient GMM
`small`	Use finite-sample formula for SEs (comparable to **`regress`**)

What the combinations mean:

<nothing>	OLS or IV as efficient GMM. $S_{classical}$ in W_n and in $AVar$.
`robust`	OLS or IV as inefficient GMM. $S_{classical}$ in W_n. S_{HC} in $AVar$.
`robust gmm2s`	2-step Feasible Efficient GMM. S_{HC} in W_n and in $AVar$.

Next, open the do file Lab7_Task5_griliches.do and execute the lines at the top that load the Griliches dataset and create the necessary year dummies:

```
use http://fmwww.bc.edu/ec-p/data/Hayashi/griliches76.dta, clear

* Create the year dummies.
* xi will use the prefix "_I" by default; "d" is more intuitive.
capture drop d*
xi i.year, prefix(d)
```

You are now ready to examine the results for 6 different specifications.

Specification 1:

S_i and IQ_i endogenous
Independence and conditional homoskedasticity
Efficient GMM
Weighting matrix is inverse of $S_{classical}$
AVar uses $S_{classical}$

In this case the IV estimator is the efficient GMM estimator. The **`ivreg2`** command line is:

```
ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt)
```

What is the interpretation of the overidentification statistic?

Specification 2:

S_i and IQ_i endogenous
Independence but *conditional heteroskedasticity*
Inefficient GMM
Weighting matrix is inverse of $S_{classical}$
AVar uses S_{HC}

Again we look at the IV estimator, but now IV is an inefficient GMM estimator – the weighting matrix is not the efficient one. However, if we use S_{HC} in the formula for *AVar* we can still get the right SEs and perform tests with the right size asymptotically.

```
ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt), robust
```

NB: Because IV is now inefficient GMM, the overidentification test statistic uses a special calculation and is not the value of the minimized GMM objective function.

Specification 3:

S_i and IQ_i endogenous
Independence but *conditional heteroskedasticity*
Efficient GMM
Weighting matrix is inverse of S_{HC}
AVar uses S_{HC}

We look at the 2-step Feasible Efficient GMM estimator.

```
ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt), robust gmm2s
```

Optional: an alternative FEGMM estimator is the CUE GMM estimator (Continuously Updated GMM). Replace **`gmm2s`** with **`cue`** and see what happens.

Specification 4:

S_i and IQ_i *exogenous*
Independence and conditional homoskedasticity
Efficient GMM
Weighting matrix is inverse of $S_{classical}$
AVar uses $S_{classical}$

In this case the OLS estimator is the efficient GMM estimator. Compare the estimations using **`regress`** and **`ivreg2`**:

```
regress lw s iq expr tenure rns smsa dyear*

ivreg2 lw s iq expr tenure rns smsa dyear*            ///
      ( = med kww age mrt), small
```

We are using the **`small`** option with **`ivreg2`** so that it uses the same finite-sample formula for the SEs. This also causes **`ivreg2`** to report t-stats instead of z-stats, like **`regress`**.

Are the coefficients the same? Are the SEs the same?

Because the OLS estimator is efficient, the additional orthogonality conditions specified by **`( = med kww age mrt)`** are *redundant*: they do not improve the efficiency of the estimator.

However, even though additional orthogonality conditions do not add to the efficiency of the estimator – the results are still OLS results – the overidentification test reported by **`ivreg2`** still has an interpretation. What is it?

Specification 5:

S_i and IQ_i *exogenous*
Independence but *conditional heteroskedasticity*
Inefficient GMM
Weighting matrix is inverse of $S_{classical}$
AVar uses S_{HC}

Now the OLS estimator is inefficient, but we can still obtain standard errors and test statistics that are robust to arbitrary heteorskedasticity.

```
regress lw s iq expr tenure rns smsa dyear*, robust

ivreg2 lw s iq expr tenure rns smsa dyear*            ///
      ( = med kww age mrt), small robust
```

NB: Because OLS is inefficient GMM, the overidentification test statistic uses a special calculation and is not the value of the minimized GMM objective function.

Specification 6:

S_i and IQ_i *exogenous*
Independence but *conditional heteroskedasticity*
Efficient GMM
Weighting matrix is inverse of S_{HC}
AVar uses S_{HC}

We look at the 2-step Feasible Efficient GMM estimator. This estimator is sometimes called HOLS (Heteroskedastic OLS) and was introduced by Cragg.

```
ivreg2 lw s iq expr tenure rns smsa dyear*      ///
      ( = med kww age mrt), small robust gmm2s
```

Compare the estimated coefficients with those of OLS. Are they the same or different?

The additional orthogonality conditions are now *not* redundant – they *do* increase the efficiency of the estimator.

How do you interpret the overidentification statistic?

CODE FOR TASKS 1-3 (Lab7_Task1_2_3_het.do)

```
capture cd M:\Econometrics\Lab7

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"

************************************************************************
*
* Task 1: The simplified MRW estimation for 105 non-oil-exporters
* Use only initial income per capita and I/Y as regressors

reg growth lngdp60 igdp
* Use FWL/added-variable plots to examine contributions of the Xs.
* avplots (note the "s") will do all these in one picture
avplots

* Use a simple visual inspection of the residuals to decide if it is
* plausible there is heteroskedasticity related to individual Xs.
capture drop ehat
predict ehat, res
scatter ehat lngdp60, nodraw name(GDP, replace)
scatter ehat igdp,    nodraw name(IY, replace)
graph combine GDP IY, name(ols_resid, replace)

* Next, perform the intraocular version of White's test:
* When you estimate using classical vs. robust SEs, are the SEs
* different?  Bigger?  Smaller?  What does this suggest?
reg growth lngdp60 igdp
reg growth lngdp60 igdp, rob

************************************************************************
*
* Task 2: White's general test for heteroskedasticity

* Estimate the equation and create the squared residuals
* We use double precision because we want the test stat to be precise.

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

* Assemble the levels, squares, and cross-products of the regressors.
* This means:
* 1. Levels: lngdp60 and igdp
* 2. Squares: lngdp60^2 and igdp^2
* 3. Cross-products: lngdp*igdp

* The variables for (1) already exist.
* Create the variables for (2) and (3).
* There won't be any redundancies (this time).

*** Your code here ***

* Perform the artificial regression.
* The dep var is ehatsq.
* The regressors are (1), (2) and (3) above.
```

```
*** Your code here ***

* Finally, display the NR2 test statistic
di e(N)*e(r2)

* Confirm you've done it correctly.  It should match the test stat
* reported by Stata's postestimation imtest with the white option.
* How many degrees of freedom does the test have?  Why?

reg growth lngdp60 igdp
estat imtest, white

* How do you interpet the result?

*************************************************************************
*
* Task 3: White's test with redundancies

* Use a specification that is quadratic in lngdp60:
* lngdp60 appears as a level and as a square.
* Create this new variable:
gen lngdp60sq=lngdp60^2

* There are no other regressors, so as in Task 2 we have 2 regressors
* plus a constant.

* Estimate and create the squared residuals.
reg growth lngdp60 lngdp60sq
capture drop ehat
predict double ehat, res
gen double ehatsq=ehat^2

* Assemble the levels, squares, and cross-products of the regressors.
* Look closely - where is the redundancy?  That is, where are you
* creating a variable that already exists?

*** Your code here ***

* Perform the artificial regression.
* The dep var is ehatsq.
* The regressors are the levels, squares and cross-products
* MINUS ANY REDUNDANCIES.

*** Your code here ***

* Finally, display the NR2 test statistic
di e(N)*e(r2)

* Confirm you've done it correctly.  It should match the test stat
* reported by Stata's postestimation imtest with the white option.

reg growth lngdp60 lngdp60sq
estat imtest, white

* How many degrees of freedom does the test have this time?
* How does this compare to Task 2, when you had the same number
* of regressors?
```

CODE FOR TASK 4 (Lab7_Task4_endog.do)

```
capture cd M:\Econometrics\Lab7

capture program drop mysimendog
program define mysimendog, rclass
      drop _all
      set obs 100

* 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 the 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 one).

/*
* 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
*/

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 both simulations have been run:
* graph combine bias1 bias2
* graph export bias1_2.emf, replace
```

CODE FOR TASK 5 (Lab7_Task5_griliches.do)

```
capture cd M:\Econometrics\Lab7

use http://fmwww.bc.edu/ec-p/data/Hayashi/griliches76.dta, clear

* Create the year dummies.
* xi will use the prefix "_I" by default; "d" is more intuitive.
capture drop d*
xi i.year, prefix(d)

**** Quick guide to ivreg2 ****
* Basic syntax:
*
* ivreg2 y x1 x2 x3 (v1 v2 = z1 z2 z3 z4), <options>
*    y  = dependent variable
*    xs = exogenous regressors
*    vs = endogenous regressors
*    zs = excluded instruments
*
* Covariance options:
*    nothing = classical S, homoskedasticity assumed
*    robust = robust S_HC, heteroskedasticity-consistent
*
* Estimator options:
*    nothing = OLS or IV
*    gmm2s = 2-step feasible efficient GMM


****** SPECIFICATION 1 ******
* Schooling and IQ endogenous
* IV as efficient GMM
* W = inverse of S_classical
* S_classical also used in AVar
* Replicates Hayashi Table 3.3 Line 3
* What does overid test mean?  How many degrees of freedom and why?

ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt)


****** SPECIFICATION 2 ******
* Schooling and IQ endogenous
* IV as inefficient GMM
* W = inverse of S_classical
* S_HC used in AVar
* What does overid test mean?  How many degrees of freedom and why?
* Compare with the previous estimation.  What is the same/different?
ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt), robust


****** SPECIFICATION 3 ******
* Schooling and IQ endogenous
* Replicates Hayashi Table 3.3 Line 5
* 2-step Feasible Efficient GMM
* W = inverse of S_HC
* S_HC also used in AVar
* What does overid test mean?  How many degrees of freedom and why?
* Compare with the previous estimations.  What is the same/different?
ivreg2 lw expr tenure rns smsa dyear*           ///
      (s iq = med kww age mrt), robust gmm2s
```

```
****** SPECIFICATION 4 ******
* Schooling and IQ endogenous
* Replicates Hayashi Table 3.3 Line 1
* OLS as efficient GMM
* W = inverse of S_classical
* S_classical also used in AVar
* Use ivreg2 small option to use same finite-sample formula as
* regress and to report t-stats instead of z-stats.
regress lw s iq expr tenure rns smsa dyear*
ivreg2 lw s expr tenure rns smsa dyear*         ///
      (s iq = med kww age mrt), small


****** SPECIFICATION 5 ******
* Schooling and IQ exogenous
* OLS as inefficient GMM
* W = inverse of S_classical
* S_HC used in AVar
* What does overid test mean?  How many degrees of freedom and why?
* Compare with the previous estimation.  What is the same/different?
regress lw s iq expr tenure rns smsa dyear*, robust
ivreg2 lw s iq expr tenure rns smsa dyear*            ///
      ( = med kww age mrt), small robust


****** SPECIFICATION 6 ******
* Schooling and IQ exogenous
* 2-step Feasible Efficient GMM, also known as HOLS
* W = inverse of S_HC
* S_HC also used in AVar
* What does overid test mean?  How many degrees of freedom and why?
* Compare with the previous estimations.  What is the same/different?
ivreg2 lw s iq expr tenure rns smsa dyear*            ///
      ( = med kww age mrt), small robust gmm2s
```