SGPE QM Lab 5: Endogeneity and Instrumental Variables

PRE-LAB ASSIGNMENT
DUE DATE: MIDNIGHT, SUNDAY 23 OCTOBER

Mark Schaffer
version of 23.10.2011

Introduction

Starting in the next lab, we consider various violations of the assumptions of the classical model. We begin with the assumption of exogeneity, i.e., E(Xu)=0, and how this problem is usually addressed, i.e., the method of instrumental variables (IV) and its extension the generalized method of moments (GMM).

In this pre-lab assignment you will begin this analysis. The assignment has three parts. In the first two parts, you look at two simple Monte Carlo exercises and interpret the results. In the third part, you are asked to read a well-known article using IV methods and briefly summarize the authors' estimation strategy.

When you have completed the assignment you should **submit the work to be marked**. This week, the only thing you need to submit is the writeup of your results; no do files or log files are necessary.

Each group is required to submit a single, jointly completed assignment. We strongly recommend, however, that **everyone** works through the material below prior to the lab.

Preparation

You may want to review what we covered in Lab 3 on Monte Carlos in Stata.

Created a folder on your network drive called "M:\QM\Lab5". Go to the coursework folder where we keep central copies of QM files and copy the files `lab5assignment1.do` and `lab5assignment2.do` to your M:\QM\Lab5 folder.

Assignment 1: Monte Carlo 1

The first Monte Carlo is in **lab5assignment1.do**. It has the same structure as the do files we used in Lab 3. The first part defines the program **mysim**:

```
program define mysim, rclass
      drop _all
      set obs 100
      gen t = _n
      tsset t

* Generate independent variable x and error u
      gen x = rnormal()
      gen u = rnormal()

* Generate the dependent variable y according to the
* following true model: y = b0 + b1*x + u
* b0 (the constant)   = 1
* b1 (the coeff on x) = 1
      gen y = 1 + 1*x + u

* Estimate the model using OLS
      reg y x

* Return the estimated b1 and the estimated SE(b1).
      return scalar b1=_coef[x]
      return scalar se1=_se[x]
end
```

The program **mysim** will be called by the Stata **simulate** command. Each time **mysim** is called, it generates a dataset with 100 observations and estimates an OLS regression using the **regress** command. It then reports back to **simulate** what the estimate coefficient **b1** is, and what the estimated standard error for **b1** is.

The **simulate** command calls **mysim**, collecting the estimated **b1** (coefficients) and **se1** (standard error of **b1**). It runs 5,000 times, so when it is done running there will be 5,000 different coefficients and SEs, representing 5,000 different uses of OLS on 5,000 different datasets:

```
simulate b1 = r(b1) se1 = r(se1), reps(5000): mysim
```

The data generation process (DGP) is as follows:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$
$$x \sim N(0,1)$$
$$u \sim N(0,1)$$
$$\beta_0 = 1$$
$$\beta_1 = 1$$

You will investigate the possible bias in estimates of β_1 (the coefficient on x), and whether there are signs of any size distortions in tests of the null hypothesis of H_0: $\beta_1=1$. (Recall the definition of a Type I error: incorrectly rejecting the null when it is actually true. A correctly-sized test will incorrectly reject the null 5% of the time when the chosen significance level is 5%, and similarly for other significance levels.)

At the bottom of the **lab5assignment1.do** are some code and questions relating to possible bias in the estimate of **b1** and to possible size distortions of test

Questions for Monte Carlo 1:

What is the evidence that the OLS estimate of β_1 is biased or unbiased?

The following code in the do file will generate output that lets you answer this question:

```
sum b1, detail
twoway (kdensity b1), xlabel(0.5 1 1.5) xline(1)          ///
      title(Distribution of b) name(bias, replace)
```

What is the evidence that 2-tailed tests of the null hypothesis H_0: β_1=1 at the 5% significance level are correctly sized, i.e., that they wrongly reject the null 5% of the time?

You should use the CLT and the Normal distribution, so the critical values are -1.96 and 1.96. The following code in the do file will generate the test necessary statistic:

```
gen t=(b1-1)/se1
```

[Hint: the **count** command is the fastest way to get the numbers you want. See Lab 3.]

What is the evidence that 2-tailed tests of the null are correctly sized across the full range of possible significance levels?

Here, use the approach that we used in Lab 3, namely that if the test is correctly sized, the distribution of p-values should have the uniform distribution. (Remember: the estimated coefficient is a random variable because it depends on the dataset at hand; the estimated SE is a random variable as well; so is the test statistic; and so, therefore, is the calculation of the p-value for the test statistic.)

A p-value is the probability of observing (if the null is true) a test statistic at least as extreme as the one we have, and "extreme" means either very large (positive) or very small (negative). For a 2-tailed test we work with the absolute value of the test statistic. The distribution of the absolute value of a Normal variable is called the *Half-Normal*. The CDF of the Half-Normal is simply F=2Φ(a)-1 where Φ(a) is the CDF for the Normal.

The Stata function **normal(t)** is the Normal CDF, i.e., the area under the curve and to the LEFT of **t**. The area under the Half-Normal CDF to the LEFT of abs(t) is **2*normal(abs(t))-1**. The p-value for **t** is therefore the area under the Half-Normal CDF to the RIGHT of abs(t) = **1-(2*normal(abs(t))-1)** = **2*(1-normal(abs(t)))**. The last command graphs the data.

```
gen p = 2*(1-normal(abs(t)))
hist p, bin(40) percent yline(2.5) title("Size (40 bins, 2.5%)")  ///
      name(size, replace)
```

Based on the above, in your view is OLS a good estimator to use for such a DGP?

Assignment 2: Monte Carlo 2

The second Monte Carlo is in **`lab5assignment2.do`**. It is almost identical to the first MC, but with a slightly different DGP. The motivation is that the explanatory variable in the DGP is x^* - the "true" x – but researcher can't observe the true x. Instead, the researcher has to work with x measured with error. This is the case of *measurement error* in the explanatory variable. The DGP is:

$y_i = \beta_0 + \beta_1 x_i^* + u_i$ (True DGP)

$x_i = x_i^* + e_i$ (x_i is x_i^* observed with error)

$y_i = \beta_0 + \beta_1 x_i + \eta_i$ (Researcher estimates using observed x_i, not true x_i^*)

$x \sim N(0,1)$

$u \sim N(0,1)$

$e \sim N(0, 0.4)$ (SD of measurement error=0.4)

$\beta_0 = 1$

$\beta_1 = 1$

And this is implemented in the **`mysim`** program by:

```
* Generate independent variable xstar and error u
     gen xstar = rnormal()
     gen u = rnormal()

* Generate the dependent variable y according to the
* following true model: y = b0 + b1*xstar + u
* b0 (the constant)   = 1
* b1 (the coeff on x) = 1
     gen y = 1 + 1*xstar + u

* But we don't observe xstar, we observe x.
* x is xstar but measured with error.
* Generate measurement error e and observed x.
     gen e = 0.4*rnormal()
     gen x = xstar + e

* Estimate the model using OLS
     reg y x
```

The questions for MC 2 are the same as for MC 1:

Questions for Monte Carlo 2:

What is the evidence that the OLS estimate of β_1 is biased or unbiased?

What is the evidence that 2-tailed tests of the null hypothesis H_0: $\beta_1=1$ at the 5% significance level are correctly sized, i.e., that they wrongly reject the null 5% of the time?

What is the evidence that 2-tailed tests of the null are correctly sized across the full range of possible significance levels?

Based on the above, in your view is OLS a good estimator to use for such a DGP?

Assignment 3: Background work for Acemoglu-Johnson-Robinson (2001)

"The Colonial Origins of Comparative Development" is a 2001 AER paper by Daron Acemoglu, Simon Johnson and James Robinson (hereafter AJR). The AJR paper is an attempt to tease out the strength of the relationship between the log of GDP per capita (our dependent variable) and the quality of institutions (our independent variable of interest). Economists have long believed that there should be a positive relationship between institutional quality and GDP per capita, and the main purpose of the paper is to quantify that relationship.

The AJR paper can be found at http://economics.mit.edu/files/4123, and we will also put a copy in the Lab5 coursework folder (where the do files are).

The three main variables of interest are:

y_i	**logpgp95**	Log PPP GDP per capita in 1995, World Bank (note: PPP denotes that the measure is based on purchasing power parity rather than nominal exchange rates)
x_i	**avexpr**	Average protection against expropriation risk. This is measured on a 1 to 10 scale, where a low score denotes a bad institutions (the government is likely to steal your property) and a high score indicates good institutions (the government is *not* likely to steal your property).
z_i	**logem4**	Log settler mortality. This is based on the earliest available record of the mortality rate of European settlers living in the country of interest, measured as deaths per thousand per year

The model to be estimated is (ignoring all other explanatory variables):

$$\boldsymbol{y_i = \beta x_i + u_i}$$

You should read the AJR paper and then answer the following questions.

Questions for Assignment 3:

Why is it a problem for OLS estimation if x is correlated with u? [Hint: which of the classical OLS assumptions is violated?]

Why are AJR worried that this could be the case here? [Hint: see AJR pp. 1369 & 1379-80.]

What is the AJR argument for why z (death rates of European settlers in their colonies in the 19th century and earlier) should be correlated with x (quality of institutions today)? [Hint: see AJR pp. 1370 & 1374-76.]

What is the AJR argument for why z should NOT be correlated with u, i.e., why death rates of European settlers should have no effect living standards (GDP per capita) today other than x? [Hint: see AJR pp. 1371-2.]