Problem set 1: understanding estimators and getting to know Gretl

September 11, 2013

1 Introduction

This problem set is meant to allow students to practice the econometric theory which is learnt in the first 30 videos or so of the 'undergraduate econometric course' on youtube. The emphasis is on practical application of the subject, although there will also be a theory section for students wishing to polish up on this aspect.

2 Setting up Gretl

Throughout this course we shall be using Gretl; a nice statistical program which (importantly) is available free of charge. Of course, feel welcome to use other tools: Stata, EViews, R, SPSS, Matlab, Octave etc. will all be (more than) capable of handling the data analysis encountered in this course. However, in order to make this course available to as wide an audience as possible, Gretl will be the primary tool used for analysis. As such, some of the answers will be Gretl-specific, but I shall try to limit this specificity where possible.

In order to get Gretl set up on your computer go to:

- For Windows machines: http://gretl.sourceforge.net/win32
- For Mac machines: http://gretl.sourceforge.net/osx.html
- For Linux machines: http://gretl.sourceforge.net

From then on, I recommend using the self-installer if available for your machine. This will guide you through installation, and (in my experience) allow you to get setup as quickly as possible. Alternatively, follow the download instructions to get Gretl setup.

3 Crime and Unemployment - practical

In this question we are going to examine the relationship between rates of violent crime in the US, and the rate of unemployment. In order to download the data for this problem set go to:

http://www.oxbridge-tutor.co.uk/#!datasets/culy

and download the sample dataset for problem set 1 - it should be in a Excel 1997-2003 format. Don't worry if you don't have Excel on your computer, Gretl is still able to read the file. Start up Gretl. In order to import the data into Gretl navigate to File → Open data → Import → Excel. Then select the file from its location where it was downloaded to, and click 'open'. Just click 'ok' of a box opens asking about the row and column number to start from. Then click 'no' when asked about whether you want to give the data a time series or panel interpretation. If the import has worked you should see three variable names: const, Violence & Unemployment listed in the Gretl variable window.

The violent crime data comes from the FBI, who have data going back to 1960 by State, and is defined as, 'the number of victims of violent crime per 100,000 persons per year.' This particular data comes from 2010. The unemployment rate comes from the United States Department of Labor, and represents the preliminary estimates of the % of labor force who were unemployed in 2013.

1. Firstly let's look at our data. This is the most important part of econometrics, and it is often forgotten. Let's draw a histogram of the Violence data. Left click on the 'Violence' data to select it, then Variable → Frequency distribution, at the top of the Gretl GUI. Select the number of bins equal to 19, and select the then click 'ok'. A nice histogram should pop up as a figure.

2. Draw a similar histogram (with 19 bins) of the Unemployment data, and report the unemployment rate bin which has the highest frequency.

3. Can you find out which State has the highest rate of violent crime reported? To do this you just need to click on the 'Violence' variable, then look for the state with the highest violence rate.

4. Another way of understanding a dataset it to look at its summary statistics. Gretl provides a nice, and simple way of doing this. In order to view this information for a given variable, just click 'Variable' → 'Summary Statistics'. This will provide a statistical summary of a given variable. Why not have a look at the Unemployment dataset's summary statistics?

5. Let's look at whether it we can visibly see if there is any relationship between unemployment and violent crime rates by drawing a scatterplot. To do this go to View → Graph specified vars → X-Y scatter. Then select 'Violence' as a Y-axis variable variable, and 'Unemployment' as an X-axis variable. This should produce a scatterplot with a fitted regression line. From this, there appears to be some sort of positive relationship between 'Violence' and 'Unemployment'.

6. One way of quantifying the relationship between two variables is via their correlation coefficient. You can find this on Gretl by going to 'View' → 'Correlation matrix. If you then select both 'Violence' and 'Unemployment' you should see an outputted correlation (along with associated p values etc.) of around 0.42. What does this mean?

7. Since we've inspected our variables sufficiently, it is now time to run our first regression. Let's run an ordinary least squares regression with 'Violence' as a dependent variable, and 'Unemployment' (and a constant) as an independent variable. To do this go to 'Model' → 'Ordinary Least Squares'. Then select 'Violence' as a dependent variable, and 'Unemployment' as an independent (a constant should already be in the list of independent variables) and click 'ok'. This should provide a read-out of the results from your first OLS regression!

8. What is the coefficient on 'Unemployment'? What is the interpretation of this value?

9. What would this model predict would be the increase in the rate of violent crime for a 1 standard deviation increase in unemployment? What is this increase in terms of standard deviations of the rate of violence?

10. What does a regression of the rate of unemployment on violent crime rates (the other way round to that in the last part) suggest would be the increase in the unemployment rate for a 1 standard deviation increase in the rate of violent crime?

11. Can you use this regression to uncover what it suggests the increase in unemployment associated with a 1 standard deviation increase in violent crime? Is this the same as we found previously? Why is this the same/different?

12. What can you conclude about the causal mechanism between violent crime and unemployment based on the two regressions you have run? Does violent crime cause unemployment or vice versa?

13. Why might it be incorrect to conclude that increases in unemployment lead to increases in rates of violent crime?

4 Theory

This section aims to build up your theoretical knowledge of econometrics, and should cover the first 30 videos or so of material from the 'undergraduate econometrics course'. This section will complement the practical part of the problem set, but is not a required part of the course.

1. For a pupil, i, selected at random from a school, the number of years of education of their parents, X_i, is given by:

$$X_i = \mu + \varepsilon_i$$

$\varepsilon_i \sim iid(0, \sigma^2)$. Here μ is the mean number of years of education completed by parents. For a sample of N students selected independently from the population:

   (a) What is the expected value of the sample mean?

   (b) Calculate the variance of the sample mean. What happens to the variance as the sample size increases?

   (c) Is the sample mean consistent?

   (d) Prove that the sample mean is a least-squares estimator for the population mean.

   (e) Is the sample mean BLUE? Either way, prove it.

2. For each of the following state whether or not the estimator is biased, consistent, both or neither, when used to estimate the population mean:

   (a) $\tilde{X} = \frac{1}{N-1}\sum_{i=1}^{N} X_i$ ○ Unbiased ○ Consistent ○ Both ○ Neither

   (b) $\hat{X} = \frac{2}{N}\sum_{i=1}^{N/2} X_i$ ○ Unbiased ○ Consistent ○ Both ○ Neither

   (c) Assuming N is even. $\bar{X} = \frac{2}{N}\sum_{i=1}^{N/2}(X_i + \mu) + \frac{2}{N}\sum_{i=N/2+1}^{N}(X_i - \mu)$ ○ Unbiased ○ Consistent ○ Both ○ Neither

   (d) $Y \sim N(\mu, \sigma^2)$ ○ Unbiased ○ Consistent ○ Both ○ Neither

   (e) $Z = \frac{1}{N}\sum_{i=1}^{N} w_i X_i$ Where: $\sum_{i=1}^{N} w_i = 1$ ○ Unbiased ○ Consistent ○ Both ○ Neither

3. Examine the following economic model

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

   (a) Derive the formula for the sample least squares estimator for the parameters α and β.

   (b) In the regression of X on Y (the reverse of the above), what is the formula for the least squares estimator for the slope parameter on Y?

   (c) If the slope parameter for the reverse regression is δ. Is the value of $\delta \times \beta = 1$? Explain your reasoning.

   (d) Show that the geometric mean of δ and β is equal to the correlation coefficient.

4. There are two populations of individuals called *samies varies* respectively. The height of individuals in the *samies* is given by:

$$X_i \sim \mu + \varepsilon_i$$

   And the height of individuals in the *varies* is given by:

$$Y_i \sim \mu + \epsilon_i$$

   Where $\varepsilon_i \sim iid(0, \sigma^2)$ and $\epsilon_i \sim iid(0, 4\sigma^2)$.

   (a) Is the sample mean from the population of *samies* an unbiased estimator of μ?

   (b) Is the sample mean from the population of *varies* an unbiased estimator of μ and consistent?

   (c) Which of the two previous estimators is most efficient, and why?

   (d) You have a sample of N individuals from each population. Is the average of the two sample means unbiased? Is this the best estimator you can construct?

   (e) For the previous example, if relevant construct a BLUE estimator. If not, prove why the mean of the sample means is best.