

## Homework 1 – Getting Started with OLS

Due Feb 8<sup>th</sup>, 2007

- 1) Keep your answer short and to the point.
- 2) Turn in a hard copy (electronic copy is NOT preferred).
- 3) Discuss the homework with your classmates, partner or pet, but turn in your own copy.
- 4) Don't copy and paste the data you've used or numbers you've generated in the Matlab and EViews exercises. I just need your results, m-files and interpretations!
- 5) Read the textbooks and the papers carefully before you work on the homework.

### QUESTION 1

- 1) Finish exercises 1, 2, 3, and 6 from Chapter 3 of Greene (6<sup>th</sup> edition).
- 2) Finish exercises 1, 2, 3, 4, and 5 from Chapter 4 of Greene (6<sup>th</sup> edition).

### QUESTION 2

Consider the bivariate normal density function for the random vector

$$\mathbf{y} = [y_1 \quad y_2]' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}):$$

$$f(\mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)$$

where

$$\boldsymbol{\mu} = [\mu_1 \quad \mu_2]' \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Derive  $g(y_1 | y_2)$ , the conditional distribution of  $y_1$  given  $y_2$ . Show that the mean of

that distribution is  $\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(y_2 - \mu_2)$  and the variance is  $\sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}$ .

### QUESTION 3 (The source of this question will be revealed in the solution)

- 1) Let  $x$  and  $y$  be two scalar random variables with zero mean. Define

$$u = y - \frac{\text{cov}(x, y)}{\text{var}(x)}x. \quad \text{Prove that } E(u | x) = 0. \quad \text{Are } u \text{ and } x \text{ independent?}$$

2) Let  $y$  be a scalar random variable and  $\mathbf{x}$  be a vector of random variables. Prove

$$E[y - E(y | \mathbf{x})]^2 \leq E[y - g(\mathbf{x})]^2 \text{ for any function } g.$$

### QUESTION 3

Consider the multiple regression model:

$$[1] \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i, \text{ where } U_i \sim N(0,1) \text{ for observations } i = 1, 2, \dots, 1000$$

In words, we have a model with two explanatory variables and a constant. You are asked to play with this model using artificial data.

[As in all the following assignments, please include the relevant EViews outputs in your write-up. By “relevant” I mean something like the regression results, but not every single value of your data.]

#### EViews Problems

- a) Create an undated workfile of 1000 observations. Generate a series of standard normal random numbers called  $u$ , a series of normal random numbers (with mean = 2 and standard deviation = 2) called  $x1$ , and a series of normal random numbers (with mean = 1 and standard deviation = 0.5) called  $x2$ . Finally generate our left-hand-side  $y$  according to the model [1] by setting  $\beta_0 = 1$ ,  $\beta_1 = 2$ , and  $\beta_2 = 3$ . [Hint: You just need **genr** and **nrnd** to do the above.]
- b) In a) you have generated some data, based on some arbitrary “truth” we made up (the parameter values, and the distribution of the error term). Now we want to see if the least squares method gives you back the truth. Run OLS for  $y$  on a constant,  $x1$  and  $x2$ . Do you get estimates close to the true values?
- c) Now run OLS for  $y$  on a constant and  $x2$ . Compare the results with those in b).
- d) Generate another series of normal random numbers called  $n$ . Add it to  $x1$  and create a “noisy”  $x1$  called  $x1\_noisy$ . Run OLS for  $y$  on a constant,  $x1\_noisy$  and  $x2$ . Compare the results with those in b).

#### Matlab Problems

- e) Write a short program as an m-file to do the following. Five numbers  $x1$ ,  $x2$ ,  $x3$ ,  $x4$  and  $x5$  are drawn from a normal distribution with mean 3 and standard deviation 1.

Mr. Tsang, not knowing how to estimate the mean of the distribution using the five numbers, come up with three estimators:

$$E1 = (x1+x2+x3+x4+x5)/5$$

$$E2 = 0.2*x1+0.3*x2+0.3*x3+0.1*x4+0.1*x5$$

$$E3 = 0.6*x1+0.1*x2+0.1*x3+0.1*x4+0.1*x5$$

Now repeat the drawing 10,000 times, so that each estimator produces 10,000 estimates of the mean. Do the estimators produce estimates close to the truth on average? Are the variances of the estimates the same for the three estimators? Which estimator would you recommend? [Notice: Please turn in an m-file. Also, for each estimator, include the mean, variance and a histogram for the 10,000 estimates.]

- f) Redo a) to d) 10,000 times, save the parameter estimates in each run (three for b), two for c) and three for d)). After the 10,000 runs, you should have 10,000 estimates for each of the eight parameters. For each of the eight parameters, create a *histogram* for the 10,000 estimates. [Hint: You need to write a for-loop. In each run, you ask Matlab to do what you did with EViews for a) to d). Before you start the loop, create some vectors/matrices to store the many numbers that you will generate with the loop.]
- g) For each of the eight parameters, find the mean, median and standard deviation of the 10,000 estimates.

#### **QUESTION 4**

Download the 2007 CPS data from the class website:

[http://www.econ.washington.edu/user/startz/482/482\\_data/cpsmar2007.wf1](http://www.econ.washington.edu/user/startz/482/482_data/cpsmar2007.wf1)

Though you see a bunch of variables, we only run:

$$[1] \quad \ln wage_i = \beta_0 + \beta_1 age_i + \beta_2 education_i + \beta_3 education_i^2 + u_i$$

- 1) Using the whole sample, estimate [1]. What is the percentage change in wage for an extra year of education for someone with 12 years of education? What is the percentage change in wage for getting a year older? According to the results, what is the wage in \$ (not log) for a 50-year old person with 12 years of education?

2) Estimate [1] separately for **men and women**. Plot in the same graph (easier in Excel) the lnwage-education profile (lnwage on y-axis, education on x-axis) for men and women who are 25 years old. Any economic explanation for the results?

3) Estimate [1] separately for **married and unmarried people**. Plot in the same graph (easier in Excel) the lnwage-education profile (lnwage on y-axis, education on x-axis) for married and unmarried people who are 25 years old. Any economic explanation for the results?

### **QUESTION 5**

**Simple Monte Carlo Experiments** - We work with the true model:

$$y_i = 1 + 2x_{1i} + 3x_{2i} + e_i, \quad e_i \sim N(0,1)$$

Of course, this is not enough for a DGP, and we need more assumptions to generate data.

1) Add the assumptions  $x_{1i} \sim N(2,1)$  and  $x_{2i} = 0.6x_{1i} + u_i$ ,  $u_i \sim N(0,1)$ , and

$x_{1i}$ ,  $e_i$  and  $u_i$  are uncorrelated. Now you have enough assumptions to generate the data.

Repeat the following 10,000 times: run a regression for  $y_i$  on a constant,  $x_{1i}$  and  $x_{2i}$ , and run another regression for  $y_i$  on a constant and  $x_{1i}$  only. Summarize the parameter estimates for  $x_{1i}$  (mean, histogram...). Is any least squares assumption violated in each regression?

2) Drop the assumptions in 1), and add the assumptions  $x_{1i} \sim N(2,1)$  and  $x_{2i} = 2 + 0.5e_i + v_i$ ,

where  $v_i \sim N(0,1)$ , and  $x_{1i}$ ,  $e_i$ ,  $v_i$  and  $u_i$  are uncorrelated. Repeat the following 10,000

times: run a regression for  $y_i$  on a constant,  $x_{1i}$  and  $x_{2i}$ , and run another regression

for  $y_i$  on a constant and  $x_{1i}$  only. Summarize all parameter estimates (mean, histogram...).

Is any least squares assumption violated in each regression?

### **QUESTION 6** – Barro’s 1991 QJE Paper

1) At the beginning of page 414, the author claims that “...Heteroskedasticity could be important across countries...” Do you agree? Why?

- 2) Look at Figure II on page 415. What is shown in the figure? What does the author mean by “partial association” based on Regression 1? How does it differ from simply plotting the data for capita growth and 1960 GDP per capita?
- 3) In Regression 2 the author adds the RHS variable GDP60SQ. How should we interpret the coefficient on this variable? Do you expect the coefficient estimate to be positive or negative? Why?
- 4) There is a typo on page 416. Find it.
- 5) On page 419 the author mentions the measurement error in GDP. Suggest a reason for why GDP is measured incorrectly.
- 6) Let’s say I want to extend this paper and throw in another 100 RHS variables (e.g. annual rainfall) in Regression 1 for the same countries. Will I succeed? Why or why not?
- 6) Besides all the variables the author puts in the regression at the end of the paper, suggest another variable that you think can explain economic growth (your choice of variable does not need to be available in reality).

**QUESTION 7 – Mankiw, Romer and Weil’s 1992 QJE Paper**

- 1) How can the authors treat equation (7) as a linear regression?
- 2) Suggest one reason why there is an error term in equation (7).
- 3) Explain intuitively the meaning of “restricted regression” for equation (7).
- 4) On page 419, explain why “...if SCHOOL is proportional to  $s_h$ , then we can use it to estimate equations (11); the factor of proportionality will affect only the constant term.”
- 5) How can the authors treat equation (11) as a linear regression?
- 6) In Figure I, how are panels B and C different from simply plotting the data (panel A)?

**QUESTION 8 – Hall and Jones’s 1999 QJE Paper**

- 1) Why is the RHS variable  $\tilde{S}$  in equation 7) correlated to the error term  $\tilde{\epsilon}$ ?
- 2) Explain why the authors’ measure of social infrastructure contains measurement error (i.e.  $\tilde{S} = S + v$ ).

**QUESTION 9 – Fama’s 1975 AER Paper**

- 1) On page 269, explain why “Fisher’s empirical evidence, and that of the most other researchers, is in fact inconsistent with a well-functioning or “efficient” market.”
- 2) Look at equation (6). What does it mean intuitively for the two density functions to be the same? What does it mean when the two are not the same?
- 3) In the section “A Simple Model of Market Equilibrium”, why does the author argue that any test of efficiency is simultaneously a test of efficiency *and* of the assumed model of equilibrium? Why are we not able to test just one of the hypotheses?
- 4) In equation (11), can I replace the information set  $(r_{t-1}, r_{t-2}, \dots)$  with  $(R_{t-1}, R_{t-2}, \dots)$ ?  
What about  $(r_{t+1}, r_{t+2}, \dots)$ ?
- 5) Does the regression equation (19) satisfy the four assumptions of the classical linear regression model? (Be very careful on this).
- 6) On page 275, what does the author mean by “[t]he hypothesis is only meaningful, however, if past rates of change in purchasing power do indeed have information about the expected future rate of change”?

**QUESTION 10 – Hall’s 1978 JPE Paper**

- 1) How does the author justify using equation 1.3 in Table 1 instead of the other two?
- 2) For equation 3, do the four assumptions of the classical linear regression model hold?
- 3) Let’s say the real rate of interest is a random variable that varies over time. Describe in words how will it change the way we test the hypothesis.
- 4) Do you think data on consumption of nondurables and services are accurate? What can go wrong when compiling the data?
- 5) The author drops durable consumption from the analysis. Suggest one reason why it is appropriate, and suggest one reason why it is inappropriate.