ECON 407/PUBP 615: Problem Set 1   1/23/2014

Due date to be determined.

Linear Algebra

1. (6 points) For this question please perform calculations by hand and in the stata mata environment. Consider a very small sample of individuals with one categorical variable (having three values: $1, 2, 3$). Suppose that these categories represent educational attainment, 1=Less than high school, 2=at least high school and less than a 4 year degree, and 3= 4 year degree or more. For a sample of 5 individuals, these values are

$$\mathbf{educ} = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \\ 1 \end{bmatrix} \tag{1}$$

   (a) By hand, convert this vector into a matrix of dummy variables representing educational attainment. Add a vector of ones for the model constant. Carefully describe the meaning of each column and how it would help you identify model coefficients. Enter the matrix into stata (in the `mata` environment) and check that it is of full rank. If it isn't, why not?

2. (6 points) For this question please perform calculations by hand and in the stata mata environment. Consider the two small matrices:

$$\mathbf{z} = \begin{bmatrix} 2 & 6 \\ 9 & 2 \end{bmatrix} \qquad \mathbf{m} = \begin{bmatrix} 4 \\ 8 \end{bmatrix} \tag{2}$$

   (a) Calculate $\mathbf{z}^{-1}$.

   (b) Show that $\mathbf{z}'\mathbf{z}$ is symmetric and square (# rows = #columns).

   (c) Show that $(\mathbf{zm})' = \mathbf{m}'\mathbf{z}'$

The Ordinary Least Squares Model

3. (18 points) We have collected data on an dependent variable ($\mathbf{y}$) and an independent variables ($\mathbf{x_1}$) This data, organized in matrix/vector form is

$$\mathbf{y} = \begin{bmatrix} 0.5377 \\ 2.8339 \\ -0.2588 \end{bmatrix} \qquad \mathbf{x_1} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \tag{3}$$

The population regression function looks like this:

$$\mathbf{y} = \beta_0 + \beta_1 * \mathbf{x_1} + \epsilon \tag{4}$$

(a) Calculate the OLS estimator for β_0 and β_1 for this problem, using the formula we derived in class ($\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$). Verify that this is identical to the scalar summation version of this estimate for β_0 and β_1 (denoted by b_0 and b_1 in the equation below) that you learned in earlier econometrics courses:

$$b_0 = \sum_{i=1}^{N} \frac{y_i - b_1 x_i}{N} \qquad b_1 = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}(\sum_{i=1}^{N} x_i)^2} \tag{5}$$

(b) Show that the OLS estimator $\mathbf{b}$ ensures that $\mathbf{x_1'e} = 0$. Discuss this finding in the context of one of the key assumptions we made in class: $E[\mathbf{x}'\epsilon] = 0$.

(c) Given your estimates for β in the previous part, show that $\sum_{i=1}^{N}(e_i)^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y}-\mathbf{xb})'(\mathbf{y}-\mathbf{xb})$.

(d) Contrast this with $\mathbf{ee}'$. What dimension is this product and show that the information (relative to the sum square errors ($\mathbf{e}'\mathbf{e}$)) on the

  (a) diagonal elements provides information about the variance of the error (Var(e_i))

  (b) off-diagonal elements provides information about the covariance of the errors (Cov(e_i,e_j))

4. (14 points) The OLS estimate of $\mathbf{y} = \mathbf{x}\beta + \epsilon$ was shown to be $\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$.

   (a) Show that one gets the same parameter estimates if a regression is run on $\mathbf{y}^* = \mathbf{y}\alpha$ and $\mathbf{x}^* = \mathbf{x}\alpha$ where α is some real number scalar value. Does the OLS variance/covariance matrix for the estimated parameters change?

   (b) Now let $\mathbf{y}^* = \mathbf{y}$ and $\mathbf{x}^* = \mathbf{x}\alpha$ (only $\mathbf{x}$ is scaled by α). Write down the formula for the ols estimate of β in terms of $\mathbf{x}$ and $\mathbf{y}$.

5. (16 points) Show that the OLS estimate of β is unbiased in the presence of heteroskedasticity. Provide an intuitive argument for why the OLS Model in this case in inefficient.

6. (40 points) Find a spreadsheet version of the data used in Tom Mrozs 1987 Econometrica paper ("The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions", Volume 55, [July], pages 765-799), on women's labor choices. The paper can be found here and the data as described in the data appendix of Greene is available here. As usual, I have put the data on my website so you can access it. In stata, issue the commands `webuse set "http://rlhick.people.wm.edu/econ407/data"` and then `webuse mroz` to access the data

   (a) Fully summarize the data (stata `sum` command) and pay close attention to means, minimums, and maximum values. Does this database have the same sample size as that used by mroz? Look both at the subsample of women working for pay and those that are not. Plot the distribution of each variable using the stata `hist` command. While you don't need to report all of these plots in your homework writeup, take care to demonstrate that the data you choose in the next part satisfy the assumptions required for an OLS model.

   (b) Using the Mroz data, estimate a labor supply model similar to his equation (1) using OLS model (using stata's `reg` command). Also estimate the model using robust standard errors (using the `robust` option). Explain your parameter estimates and defend your model specification.

   (c) Using the stata `mata` environment for linear algebra, do the following.

1. Verify that you can replicate stata's "canned" regression package results for **b**, root mean square error (RMSE), standard errors, and t-statistics using the formulas we have developed in class. Interpret your results.
2. Calculate the model R^2, and interpret.
3. Calculate the robust standard errors. Interpret your results especially taking note of how the model predicted error **e** is used to "fix" the standard errors. You may find the mata command `diagonal` useful for this problem. **a** =`diagonal(A)` will create a vector **a** comprised of the diagonal elements of the matrix A.

(d) Test the model for heteroskedasticity using both the stata `estat hettest` and the "manual" method outlined in chapter 1, in the section on heteroskedasticity. The key thing you will need to ask stata to do is to save your residuals from the first step regression. To do this, use the command `predict res, r` will save the predicted model errors in a variable called `res`. You might get some minor differences in the test statistic using these two methods since the hettest command uses a non-parametric form of the test (making no assumptions about the distribution of the errors). What is the intuition of this test?