ECON452 : Homework

Pier-André Bouchard St-Amant *

1. This homework is due on March 9th. Drop it in the homework box on the second floor of Dunning Hall.
2. Working in pairs is allowed. If so, both students will receive the same grade.
3. Append your do-file as well as the relevant stata output when necessary.
4. Some questions have length limitations. Do not try to get around them by adding a lot of comments in your do-files. The purpose of these limitations is to force you to choose what you think is relevant and what is not.

1 Review of OLS (50 points)

This question explores the properties of Ordinary Least Squares. Throughout the question, work with the dataset regress.dta available directly from Stata's webpage (you can obtain it with the command *webuse regress*).

a)

Inspect graphically the relationship between the dependent variable y and the three variables x_1, x_2, x_3 and answer the following questions :

1. What kind of variable is x_2 ? What will be the interpretation for $\hat{\beta}_2$?
2. Although the relationship between y and x_1 seems linear, it does not seem to be the case for x_3. Can you think of a transformation of x_3 or a new variable based on x_3 that you can create to represent a better relationship between y and x_3 ?

*pabsta@econ.queensu.ca.

b)

Stata has a built-in matrix language called mata which allows to compute directly in matrix algebra. Some code is showed below.

```
gen ones = 1
mkmat ones x1 x2 x3, matrix("X") //Creates a matrix of regressors
mkmat y //Transforms the variable y into a vector
matrix b = inv(X'*X)*X'*y
matrix list b //displays the matrix b
matrix av = inv(X'*X)
matrix list av //displays the matrix av
```

Run this code in Stata. What does the matrix b contains and how it relates to the command below ?

```
reg y x1 x2 x3
```

c)

If you ran the mata code correctly in the last question, the matrix av was displayed. What is the relationship between this matrix, the root of the mean square errors and the columns of standard errors ? Show with calculations the relationship between the standard error of $\hat{\beta}_1$ and one element of av.

d)

Using the command « predict », generate the residuals of the last regression (see the help file). Use the command *corr* to compute the correlation between the residuals and each dependent variable. What is the correlation between each variable and the residuals ? How does this relate to theoretical properties of OLS ?

e)

Using the command *di*, the command *corr* and the values for each $\hat{\beta}$, verify for $\hat{\beta}_1$ and $\hat{\beta}_2$ that this formula holds :

$$\hat{\beta}_j = \frac{\text{cov}(x_j, y) - \sum_{m \neq j} \hat{\beta}_m \text{cov}(x_j, x_m)}{\text{var}(x_j)}.$$

Using a diagram, explain intuitively what this formula means in terms of available information contained in each variable.

f)

Still on the regression above, perform the following test :

$$H_0 : \hat{\beta}_1 \leq 0,$$
$$H_1 : \hat{\beta}_1 > 0$$

with a confidence level of 95%. Should you reject H_0 or not ? Can the result of this test be seen directly from Stata's output ?

g)

Perform the following test :

$$H_0 : \hat{\beta}_1 + \hat{\beta}_2 = 3.5,$$
$$H_1 : \hat{\beta}_1 + \hat{\beta}_2 \neq 3.5$$

with a confidence level of 95%. Should you reject H_0 or not ? Given the nature of x_2, does this tests make sense ?

h)

Based on your answers in question a.2), test two alternative specifications of regression and argue which model you should keep. Keep your explanations and results under half a page.

j)

A regression has been performed and the output is shown in appendix A. Unfortunately, some numbers are missing (and replaced with Xs). Use your knowledge of t-statistics and Stata's output to find the missing numbers (no marks without explanations).

(*Hint* : the command di ttail(n, t) might be useful.)

2 A Primer On Time Series (50 points)

Although the problem is simpler, this question is close to the type of work I expect from you in your final report. In the dataset someARp.dta available on the website, there is some data drawn from a *stationary* autoregressive process of degree p (with $p \leq 8$).

Using various diagnostic tools explained in class, find the best representation of this time series.

Note also that the sample size is quite large (a thousand observations). Compare the OLS estimators of your chosen model with the ARIMA estimators if you drop the last 900 observations. What happens if you drop fewer and fewer observations (say, only the last 500, 250 and 100) ? How does this relate to the properties of the MLE estimator ?

Some remarks about this question :

1. No more than two pages to explain your strategy (excluding the relevant Stata output and do-files).
2. No marks will be awarded for a choice strategy based on a complete enumeration of all possible AR(p) with p≤8. Use the tools seen in class.

A Output for question 1 j)

```
      Source |       SS       df       MS              Number of obs =     313
-------------+------------------------------           F(  3,   309) =  213.41
       Model |  2634884.26     3  878294.754           Prob > F      =  0.0000
    Residual |  1271713.21   309  4115.57673           R-squared     =  XXXXXX
-------------+------------------------------           Adj R-squared =  0.6713
       Total |  3906597.47   312  12521.1457           Root MSE      =  XXXXXX

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -2.681508   1.393991    -1.92   XXXXX     XXXXXXXX    XXXXXXXX
          x2 |  -3.702419   XXXXXXXX   -24.04   0.000    -4.005491   -3.399348
          x3 |   .1086104    XXXXXXX     XXXX   0.232    XXXXXXXXX   XXXXXXXXX
       _cons |   906.7392   28.26505    32.08   0.000     851.1228    962.3555
------------------------------------------------------------------------------
```