NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

Department of Economics

ECONOMETRICS I

Assignment 2

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-90
Home page: people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: people.stern.nyu.edu/wgreene/Econometrics/Econometrics.htm

Note: The following exercises request you to compute numerous regressions with your favorite software and then "report your results." To report results, please extract from the computer output the specific results you wish to report and paste these into your submission. Please do not submit long unannotated scripts that contain buried deep within them the results that you would like to provide for the problems.

Part I. Basic Theory

1. Prove $\mathbf{e}'\mathbf{e}$ increases when a restriction, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, is imposed on the regression coefficients

2. Show how to do constrained least squares regression entirely using matrix algebra, not your software's built-in regression procedures. That is, what computations would you do to compute a constrained least squares estimator.

3. Write a paragraph that will show a beginning econometrics student who knows how to use a (some) regression package what computations they should do to test a hypothesis about the coefficients in a linear regression model. Include in your description a prescription for what should be reported to a reader of the study.

Part II. Applications of Hypothesis Testing

The data sets for these applications are posted on the course home page, in the Problems page, as .csv files.

1. This part of the exercise is based on the gasoline data, which you can download as

http://people.stern.nyu.edu/wgreene/Econometrics/gasoline.csv

Note, there is some ambiguity as to how to obtain the dependent variable in this data set. Use the following as a guide to setting up the first regression. The second and third will be straightforward.

```
create  ; g=((gasexp/gasp)/pop)*1000000 $
create  ; logg=log(g);logpg=log(gasp);logy=log(pcincome)$
create  ; upto73=year < 1974 ; after73=1-upto73 $
create  ; t=year-1952$
regress ; lhs=logg;rhs=one,logpg*upto73,logpg*after73,logy,t $
```

a. Estimate by least squares a version of the regression model of which looks as follows:

$$\log G = \beta_1 + \beta_2(\log \mathit{GasP} \text{ up to 1973, 0 else}) + \beta_3(\log \mathit{GasP} \text{ after 1973, 0 else})$$
$$+ \beta_4 \log \mathit{PCIncome} + \beta_5(\mathit{YEAR} - 1952) + \varepsilon.$$

b. Now, use least squares to fit the coefficients of the model

$$\log G = \beta_1 + \beta_2 \log \mathit{GasP} + \beta_3(\log \mathit{GasP} \text{ after 1973, 0 else})$$
$$+ \beta_4 \log \mathit{PCIncome} + \beta_5(\mathit{YEAR} - 1952) + \varepsilon$$

Report the least squares coefficients for both cases. Could you have computed the second least squares regression from the first one? If so, show how, algebraically. If not, why not? Describe this "model" in terms of the relationship between price and quantity that it implies.

c. We now examine whether the three aggregate price indexes, PD = durables price, PN = nondurbles price, PS = services price, are significant explanatory variables in the equation. Add logPD, logPN and logPS to your regression in part b. Report the results. Now, test the hypothesis that the coefficients on the three variables are all zero. Use an F test and a Wald test.

2. Using the health care data,

http://people.stern.nyu.edu/wgreene/Econometrics/healthcare.csv

carry out a Chow test for whether different regressions apply for men and women, using this model for log of income,

$$\log(\text{hhninc}) = \beta_1 + \beta_2 \text{ Age} + \beta_3 \text{ Educ} + \beta_4 \text{ Married} + \beta_5 \text{ Hhkids} + \varepsilon.$$

Carry out the chow test using the standard F test. Now, add at least one variable to the model and carry out the test again with your expanded model. Report all relevant results.

3. This exercise is based on the Spanish dairy data,

http://people.stern.nyu.edu/wgreene/dairy.csv

The data in the dairy data file are already in logs, so the regression model is

$$\text{yit} = \beta_1 + \beta_2 \text{ x1} + \beta_3\text{x2} + \beta_4\text{x3} + \beta_5\text{x4} + \varepsilon.$$

a. Compute the coefficients of this regression and report your results

b. The constant returns to scale hypothesis is $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 1$. Carry out a test of the hypothesis and report all results

c. These data have been used in many studies to study functional form in production. In part a, you fit a Cobb-Douglas model. A translog model would include all unique squares and cross products, x1*x1, x1*x2, etc. Fit a translog model, and test the hypothesis of the Cobb-Douglas model as a restriction on the translog model.

d. These data are a panel spanning 6 years. There might have been technological change in those 6 years. There are time dummy variables in the data set, YEAR93,…,YEAR98. Add the time effects (dropping one of the dummy variables, of course) to your regression, and examine the results to see if there is evidence that the production shifted over time. Test the joint hypothesis that the time effects are all zero in the context of your translog model of part c.

Part III. Model building

Your professor (once) drove to work each morning on the dreaded Long Island Distressway, from a faraway suburb (that sits high on a hill, overlooking things green and beautiful…..) The amount of time the drive took each day was relatively constant, but it did vary systematically with some obvious factors and with some less obvious factors. These included:

a. Time of departure from home – there was a distinct peaking effect.
b. Day of the week. Monday seemed to be the peak day, Friday was less crowded.
c. Holiday. Holidays were special.
d. Rain. Actual occurrence of rain had an obvious effect. 99.9% of New Yorkers cannot drive in the rain. Your professor could. So could all drivers of SUV's. They were always eager to demonstrate this, at high speed.
e. Snow. Snow on the road had a surprising effect on drive time. Many people stayed home. New Yorkers are also unable to drive in snow. However, unlike rain, they do not like even to try to drive in snow.
e. The idiot effect. An idiot in an SUV who weaves their way into (up) a tree or into some other car affects everyone's time for most of the morning.

The data set *commute.csv* contains a sample of observations on these effects and the drive time. You can download them from the course home page at

http://people.stern.nyu.edu/wgreene/Econometrics/commute.csv

Your assignment is to build a regression model for drive time which incorporates these observed effects. Some notes about the data set:

1. Drive time is coded in minutes
2. Exit time is coded in minutes past 6:30 in the morning. E.g., 5.0 means 6:35.
3. Day is coded 1 for Mon., 2 for Tues, etc.
4. Holiday, Rain, Snow, and Idiot are ordinary dummy variables.

A hint: There is a strong peaking effect with respect to time of day. The peak comes at 7:00 AM. After that, the expected drive time goes down a little. Before 7:00, expected drive time goes up steadily. You will have to think about this one a bit, and decide how you want departure time to enter your equation. Note, it is not a simple linear term.

a. Estimate the coefficients of your model.
b. Identify which are the most significant effects on drive time, and which are the least important.
c. Compute the predicted drive times based on your equation.
d. Note that only one day was affected by the idiot effect. Compare the actual and fitted values for this day, and comment.
e. Compare the day of the week effects. Is the drive typically worse on Tuesday than on Thursday?
f. Test the hypothesis that my guess about weekday effects is incorrect. That is, test the hypothesis that variation in drive time is not explained by day of the week.
g. My worst day took a miserable 195 minutes to get to work. What were the conditions that day. What drive time would your regression model predict for that day? Compute a 95% prediction interval for the drive time for this day. Does the interval contain the actual outcome?

NOTE: There is no single right answer for this exercise. Your assignment is to (be a bit creative) and build a reasonable model for the outcome (drive time) based on the information you have been given and the data in the data file.

Part II. Suggested Scripts

1.

```
create  ; g=gasexp/gasp/pop*1000000 $
create  ; logg=log(g);logpg=log(gasp);logy=log(pcincome)$
create  ; upto73=year < 1974 ; after73=1-upto73 $
create  ; t=year-1952$
regress ; lhs=logg;rhs=one,logpg*upto73,logpg*after73,logy,t $
regress ; lhs=logg;rhs=one,logpg       ,logpg*after73,logy,t $
create  ; logpd=log(pd);logpn=log(pn);logps=log(ps)$
? (1) Short model
regress ; lhs=logg;rhs=one,logpg       ,logpg*after73,logy,t $
calc ; ss0 = sumsqdev $
regress ; lhs=logg;rhs=one,logpg       ,logpg*after73,logy,t,logpd,logpn,logps$
calc ; ss1 = sumsqdev $
calc ; list ; F = ((ss0 - ss1)/3) / (ss1/(n-8)) $
regress ; lhs=logg;rhs=one,logpg       ,logpg*after73,logy,t, logpd,logpn,logps
;test: logpd=0,logpn=0,logps=0$$
matrix ; bp=b(6:8);vp=varb(6:8,6:8);list;wald=bp'<vp>bp $
```

2.

```
namelist ; x = one,age,educ,married,hhkids$
regress  ; for[female=0] ; lhs=log(hhninc) ; rhs = x $
calc ; ssmale = sumsqdev $
regress  ; for[female=1] ; lhs=log(hhninc) ; rhs = x $
calc ; ssfemale = sumsqdev $
regress  ; lhs=log(hhninc) ; rhs = x $
calc ; sspool = sumsqdev $
calc ; list ; f = ((sspool - (ssmale+ssfemale)) / kreg ) /
                  ((ssmale+ssfemale)/(n-2*kreg) ) $
? This can be done using a Wald test with a built in procedure
regress ; for[female=*,0,1] ; Lhs = log(hhninc) ; Rhs = x $
decompose $
```

3.

```
namelist ; x = one,x1,x2,x3,x4 $
regress ; lhs = yit ; Rhs = x $
calc ; ssu = sumsqdev $
regress ; lhs = yit ; Rhs = x
        ; cls: x1 + x2 + x3 + x4 = 1 $
calc ; ssr = sumsqdev $
calc ; list ; f = ((ssr-ssu)/1)/(ssu/(n-5))$
? Note that the F statistic is actually reported with
? the regression results
? This command carries out the test without computing the
? restricted regression
regress ; lhs = yit ; Rhs = x
        ; test: x1 + x2 + x3 + x4 = 1 $
```