Westminster Business School
School of Finance and Accounting
7FNCE044 Predictive Analysis for Decision Making
Semesters 1 and 2
Weeks 1 and 2
Dr. Issam Malki

Seminar Weeks 1 and 2

Linear Regression: Extension I

Computer Based Questions:

Question 0

Use R studio for this question. Suppose you have the following matrices:

y is (10×1) uniformly distributed random variable and X is (10×3) matrix, which contain a vector of ones and two vectors of normally distributed random variables.

Suppose you are interested in estimating a model of the following form:

$$y = X\beta + u \quad (1)$$

where β is (3×1) vector of unknown parameters and u is (10×1) vector of unknown error term, assumed to be normally distributed.

Your aim is to estimate the vector β using OLS defined – in matrix form – as:

$$\hat{\beta} = (X'X)^{-1}X'y \quad (2)$$

Write an R Studio script to (i) generate the random variables y and X and (ii) estimate β using OLS.

Question 1

The datafile **nls80.xls** contains labour market information for a sample 935 working males. The file contains the following variables:

Wage	monthly earnings
Hours	average weekly hours
iq	IQ score
kww	knowledge of world work score
educ	years of education
exper	years of work experience

tenure	years with current employer
age	age in years
married	=1 if married
black	=1 if black
south	=1 if live in south
urban	=1 if live in SMSA (SMSA: Standard Metropolitan Statistical Area)
sibs	number of siblings
brthord	birth order
meduc	mother's education in years
feduc	father's education in years
lwage	natural logarithm (ln) of wage

Use R studio (or any of the following: SPSS, Python, Eviews) to answer the following questions:

1. Consider two alternative specifications to explain wage using the number of hours worked and education:

$$wage_i = \gamma_1 + \gamma_2 educ_i + \gamma_3 hours_i + u_i \quad (1)$$

$$\ln(wage_i) = \beta_1 + \beta_2 educ_{i2} + \beta_3 hours_{i3} + v_i \quad (2)$$

Explain the differences between Equations (1) and (2). Do they measure the same quantities?

2. Estimate the two equations using OLS. Answer the following questions. Explain your answers.

i. Interpret the coefficients of both equations.
ii. Are the estimated effects statistically significant?
iii. What are the statistical properties of the residuals?

3. Test for the presence of heteroskedasticity in the two models. What are your conclusions?

4. Test for the presence of serial correlation in the two models. Are the residuals serially correlated?

5. How would your findings in (4) and (5) change your conclusions? Provide evidence to your claim.

Question 2 (attempt this by yourself. Try to write your own Python code):

Use hedonic.xls, file available from blackboard. The data provides a data set to analyse the housing market in Stockton California during the year 2002. The data are cross sections. The description of the variables in the file is givenbelow:

SP=	House selling price.
SFLA =	size of living area (in square feet).
BEDS =	Number of bedrooms.
BATHS =	Number of bathrooms.
STORIES =	Number of stories.
VACANT=	Vacancy status (1 if vacant, 0 if not at the time of the sale)
AGE=	Age of the house in years

Refer to the multiple regression model:

$$\ln(SP)_i = \beta_0 + \beta_1 SFLA_i + \beta_2 BEDS_i + \beta_3 BATHS_i + \beta_4 STORIES_i + \beta_5 VACANT_i + \beta_6 AGE_i + u_i$$

where u_i is the error term and *i* is a subscript referring to the house id.

1. Estimate the model. Save the residuals. Plot the distribution of the residual series. Does it behave like a normally distributed variable? Confirm the normality assumption using Jarque-Bera test. What do you conclude?

2. Use graphical approach to detect heteroscedasticity in the data. Use variables SFLA and Age. Are there any evidence suggesting the presence of heteroscedasticity? [**Hint**: use scatter plot]

3. Test for heteroscedasticity using White test. What are conclusions? Correct for heteroscedasticity using White heteroscedasticity consistent standard errors. How does this change your conclusions about the relationship between selling price and the other independent variables?
4. Use graphical approach to detect any potential serial correlation in the data. What do you notice?
5. Use Durbin-Watson test to test for the presence of autocorrelation. What are your conclusions?

6. Do your conclusions change when using the LM test?

Theoretical Questions

1. Some linear Algebra

Suppose we have the following, particularly simple, linear model:

$$y_i = \beta_1 + \beta_2 x_i + u_i,$$

which, as usual, can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \mathbf{u}$$

where **1** is a vector containing only ones.

a) Consider the following matrix:

$$\mathbf{X'X} = \begin{bmatrix} A & B \\ B & \sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

Using summation ($\sum$) notation, write down expressions for *A* and B.

b) Consider the following matrix:

$$\mathbf{X'y} = \begin{bmatrix} C \\ D \end{bmatrix}.$$

Using summation notation, write down expressions for *C* and *D*.

c) Using your answer to part (a) and the rule for the inverse of a 2×2 matrix, write down the matrix $\mathbf{X'X}^{-1}$.

d) Find expressions for the OLS estimates, $\hat{\beta}_1$ and $\hat{\beta}_2$.

e) In order for $\mathbf{X'X}^{-1}$ to exist, what must the rank of **X** be? What restrictions must the x_i satisfy in order for **X** to have this rank?

2. OLS properties:

Suppose you have the following model:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}.$$

We assume the following is satisfied:

1. **X** is a fixed (non-stochastic) matrix with rank *k*;

2. $\mathbf{u}$ is a random vector with $E(\mathbf{u}) = \mathbf{0}$ and $\text{var}(\mathbf{u}) = E(\mathbf{uu}') = \sigma^2\mathbf{I}$.

The OLS estimator is given as:

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

1. Show that $E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$.

2. Derive the distribution of the OLS estimator, $\widehat{\boldsymbol{\beta}}_{OLS}$ assuming Normality.