**KYIV SCHOOL OF ECONOMICS**

**Statistics and Econometrics for Business II, Fall 2014**

Instructor: Maksym Obrizan

HOMEWORK 1 by ________________________________(First and Last Name)

**Due:** At 9 am on Monday, November 10th (or earlier). Late homeworks will lose one letter grade per day.

**Instructions:**

Failure to follow these instructions will result in losing 2 % points for each!

1. Create one .doc, docx or .txt file for the **entire** homework and one .do file for the **entire** homework. The document should include answers to the questions while do-file should show how you got the results.

2. Name all your files starting with your last name (i.e. Pjatochkyn_HW1.do).

3. DO NOT PRINT THE HOMEWORK - submit files in an electronic form to Econometrics TA at sbsuleimanov@kse.org.ua

Please DO NOT SEND the files to me!-:) but you can copy yourself to avoid disputes

4. Please include both files as separate attachments and not as a single zipped folder.

I. You have been hired as a consultant to design a salary scheme in a large corporation.

a. Use function *use* to open file "wage1.dta" in Stata. Count the number $N$ of letters in the English version of your first and last name. Use function

*drop in N/N+9*

where $N = 13$ for a person named *Maksym Obrizan.*[1] Use function *describe* to find out the number of observations and number of explanatory variables.

b. Provide the descriptive statistics for variable *educ* using function *summarize.*

c. What is the correlation between *educ, exper* and *wage*? Use function *corr.*

d. Use function *tabulate* to find out the number of race "white" respondents

e. Use function *ta nonwhite female* to find out the percent of nonwhite males in the sample.

f. Use function *hist* to plot the histogram of wages in the sample. Does the wage distribution look symmetric? Explain.

g. Use function *scatter wage educ* to plot the figure of wage as a function of education.[2] Does wage seem to depend on education? Explain.

h. Use function *reg* to run regression for log wage *lwage* on slide 18 of the Lecture notes I. Do you

[1] This is done to ensure that everybody uses slightly different data set.

[2] Notice that Stata puts the first variable on the vertical axis and the second on the horizontal axis.

get exactly the same results? Are results sufficiently close? Explain.

i. Which of the variables in part h are statistically significant based on t-statistics?

j. Now include variable $female$ into your specification. Is there evidence of discrimination against women? Compare this model with part h. Which model has higher R-squared? Explain whether your result is surprising. *Use this model for the rest of your assignment.*

k. Take observation number 50 in your data set and predict their salary based on your model in part j. Take the actual *lwage* for this person. Did your model over- or underpredicted the actual salary?

l. Use function

*predict Mod_Res, r*

to predict residuals from your model. Plot a histogram of residuals. Do residuals look symmetric?

m. Suppose you want to test whether *exper* and *tenure* are jointly equal to zero. Use slides 28-29 from Lecture notes II to construct F statistic for this test. Use function *disp invFtail(df1,df2,.05)* to find out the critical value of F distribution.

n. Compute the F statistics for the overall significance of a regression. Based on the test what can you say about overall significance of our regression?

o. What other variables may affect *lwage*? Build a better model using your data. Explain your results.

II. The goal of this exercise is to replicate the results in the working paper by Obrizan and Wehby (2012). Using function *insheet* in Stata open the file "HE_LE_Data_Lagged.csv". Count the number $N$ of letters in the English version of your first and last name. Use function

*drop in N/N+9*

where $N = 13$ for a person named *Maksym Obrizan.*

a. Plot the scatter plot of female life expectancy ($fle$) as a function of health expenditure per capita (*he_lagged*) for year 2008. Explain if you find any relationship between two variables.

b. Generate variable he_lagged2

*gen he_lagged2 = he_lagged*he_lagged*

and provide descriptive statistics for it.

c. List countries for which fle is recorded only once but not twice.

d. Run a regression

*regress fle he_lagged he_lagged2 r_**

and use the original paper to describe variables that are included in the regression.[3]

e. Explain your main findings: what variables affect female life expectancy and in what way.

f. What is the difference between two variables $r\_fem\_lfpr$ and $fem\_lfpr$? Explore the data set.

[3] For example, Table 1 gives descriptive statistics which may help you to identify variables.