Department of Economics *UN3412*
Columbia University *Spring 2021*

Problem Set 4
Introduction to Econometrics
for both sections
Seyhan Erden and Tamrat Gashaw

1.(30p) Use the data in **hprice1.dta**. to estimate the following model (description of the variables in the data set is listed below in Table 1:

$$price = \beta 0 + \beta 1 sqrft + \beta 2 bdrms + u$$

where price = the (selling) price of the house (in 1000 dollars), sqrft = size of house (square feet) and bdrms = number of bedrooms in the house.

- **(a)** (3p) Write out the estimation result in equation form.
- **(b)** (3p) What is the estimated increase in price for a house with one more bedroom keeping square footage constant?
- **(c)** (6p) What is the estimated increase in price for a house with an additional 1400-square-foot bedroom added? Compare this to your answer in (b).
- **(d)** (6p) What percentage of the variation in price is explained by square footage and number of bedrooms? Compare your answer to the adjusted R^2. Explain the difference.
- **(e)** (6p) Consider the first house in the sample. Report the square footage and number of bedrooms for this house. Find the predicted selling price for this house from the OLS regression line.
- **(f)** (6p) What is the actual selling price of the first house in the sample? Find the residual of this house. Does it suggest that the buyer underpaid or overpaid for the house? Explain.

Table 1: DATA DESCRIPTION, FILE: hprice1.dta

Variable	**Definition**
price	House price, in $1000.
Assess	Assessed value in $1000.
bdrms	Average number bedrooms.
Lotsize	Size of lot in square feet.
Sqft	Size of house in square feet
colonial	= 1 if house is in Colonial style. = 0 otherwise.
Lprice	Log(price)
lassess	Log(assess)
llotsize	Log(lotsize)
lsqft	Log(sqft)

2. (40p) Use **cps92_12.dta** to answer following questions. You can find the description of the variables in the file named **cps92_12.description.pdf**.

(a) (5p) Generate natural logarithm of *ahe*, call it *lahe*. Regress *lahe* on *age, female, bachelor* and interaction of female and bachelor (call this variable *femxbac*) this is your regression 1. Report (copy/paste) your results here title it **Regression 1**

(b) (5p) Regress *lahe* on *female*, *age*, *bachelor* and interaction between female and age (call this variable *femxage*). Report (copy/paste) your results here title it **Regression 2**

Using the appropriate regression please answer the following questions

(c) (5p) What is the estimated average hourly earnings difference between females with bachelor degrees and males with bachelor degree? State which regression (Regression1 or Regression2) do you need to use to answer this question and explain your answer.

(d) (5p) What is the estimated average hourly earnings difference between females with bachelor degrees and females without bachelor degree? State which regression (Regression1 or Regression2) do you need to use to answer this question and explain your answer.

(e) (5p) How would you test if there is a significance difference in average hourly earnings of females with bachelor degrees and males with bachelor degrees? Please write commands necessary, you can use any method you want, but you must write the null hypothesis first. State which regression (Regression1 or Regression2) do you need to use to answer this question and explain your answer. Commands do not need to be implemented in Stata.

(f) (5p) How would you test if there is an intercept and if there is a slope difference in the two estimated regression lines for males and females if your y-axis is average hourly earnings and x-axis is *age*? Which regression (Regression1 or Regression2) do you need to use to answer this question? Write the null hypothesis for each test. Commands do not need to be implemented in Stata.

(g) (5p) As people get older, does the wage gender gap widen? Draw a sketch graph (with average hourly earnings on vertical axis and age on horizontal axis) of what this result tells you. Which regression (Regression1 or Regression2) do you need to use to answer this question and explain your answer?

(h) (5p) Is the coefficient of *female* in Regression 2, statistically significant at 1% significance level? Please write the null hypothesis and calculate the test statistic before you answer this question. Also make sure to give your reason for your answer.

3. (30p) This question is based on the posted article by David L. Sjoquist and John V. Winters (2015) entitled "*State Merit Aid Programs and College Major: A Focus on STEM*". Journal of Labor Economics, Vol. 33, No 4. You are required to read this paper thoroughly and assess this paper based on:

(a) (6p) The employed research methodology (i.e., What type of research method is used in the paper?)

(b) (6p) The data set used (i.e., Is it cross-section, panel, or time series data?) and related data gathering techniques and issues.

(c) (6p) Alternative estimation techniques (i.e., Do you think of other estimation techniques that can be used compared to what is used by the authors)

(d) (6p) Internal and external validity of the paper's findings (i.e., Do you see any internal and external validity issues with the paper before the findings of the paper is implemented?)

(e) (6p) Identifying limitations of the paper and suggesting improvements. That is, can you find at least one limitation of this paper and suggest fixes?

<u>Following questions will not be graded, they are for you to practice and will be discussed at the recitation:</u>

1. US states differ in the generosity of their welfare programs. We here wish to analyze which factors play a role in the level of benefits across different states. The data set TANF2.dta contains data from each of 49 states. The variables in the data set are given in the following table:

Table 3
DATA DESCRIPTION, FILE: TANF2.dta

Variable	Definition
tanfreal	State's real maximum benefit for single parent with three kids.
black	Percentage of state's population who are African Americans.
blue	Dummy variable, equals 1 if state voted Democratic in 2004 presidential election.
mdinc	State's median income.
west	= 1 if state is in West = 0 otherwise
south	= 1 if state is in South. = 0 otherwise.
midwest	= 1 if state is in Midwest = 0 otherwise
northeast	= 1 if state is in Northeast =0 otherwise

Use data set TANF2.dta to examine whether Midwest states differ in their welfare programs from other states. To do this, we will use the following regression model:

$$tanfreal = \beta_0 + \beta_1\, black + \beta_2\, blue + \beta_3\, midwest + \beta_4\, (black*midwest) + \beta_5\, (blue*midwest) + u$$

Here, *black*midwest* is the product of the regressors *black* and *midwest* and so forth.

(a) Write the null hypothesis to test whether there is a difference between the welfare programs of Midwest states and all other states, explain.

(b) Construct new set of interaction regressors in STATA. Estimate the model above. Write your answer as a regression equation with standard errors in parenthesis underneath each coefficient. Perform the test for the null hypothesis in part (a) with a robust F-test. What is your conclusion?

(c) Introduce a new variable *nonmidwest* = 1 – *midwest.* That is *nonmidwest* = 1 if a state is not in the Midwest and zero otherwise. Consider the following alternative regression model:

$$tanfreal = \gamma_1\, nonmidwest + \gamma_2\, (black*nonmidwest) + \gamma_3\, (blue*nonmidwest) + \gamma_4\, midwest$$
$$+ \gamma_5\, (black*midwest) + \gamma_6\, (blue*midwest\ ) + u$$

Write up the hypothesis of no differences in welfare programs in terms of $\gamma_1 .... \gamma_6$
What is the relationship between the parameters $\gamma_1 .... \gamma_6$ in this new model and $\beta_1 .... \beta_6$

in the previous model? Estimate the model in STATA and write the result in usual regression equation form with standard errors in parentheses underneath coefficients.

(d) What happens if you include an intercept γ_0 in the model in part (c)? Explain.

2. [Practice question, not graded] SW Empirical Exercise E8.1 (use lead_mortality.dta)

3. [Practice question, not graded] SW Empirical Exercise E8.2 (use CSP2015.dta)

4. [Practice question, not graded] SW Empirical Exercises 9.1 (use CSP2015.dta)