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

Problem Set 4
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw

Q#1: [30 points] 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) [3 point]** Write out the estimation result in equation form.
- **(b) [3 point]** What is the estimated increase in price for a house with one more bedroom keeping square footage constant?
- **(c) [6 point]** 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) [6 point]** 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) [6 point]** 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) [6 point]** 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)

Q#2: [40 Points] Consider the following Population Linear Regression Function (PLRF):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i \quad (1)$$

where, Y_i = average hourly earnings/wage in \$, X_1= years of education, X_2 = years of potential experience, X_3 = years with current employer (tenure), X_4 = 1 if female, X_5 = 1 if nonwhite, and u_i= the usual error term of the model.

For this question, use the accompanying WAGE dataset with this problem set (i.e., a different dataset than you used in PS#2). Here is the description of the variables in the dataset for your consumption. We might be using this data set for the coming problem sets too.

```
Obs:   526

 1. wage                 average hourly earnings
 2. educ                 years of education
 3. exper                years potential experience
 4. tenure               years with current employer
 5. nonwhite             =1 if nonwhite
 6. female               =1 if female
 7. married              =1 if married
 8. numdep               number of dependents
 9. smsa                 =1 if live in SMSA
10. northcen             =1 if live in north central U.S
11. south                =1 if live in southern region
12. west                 =1 if live in western region
13. construc             =1 if work in construc. Indus.
14. ndurman              =1 if in nondur. Manuf. Indus.
15. trcommpu             =1 if in trans, commun, pub ut
16. trade                =1 if in wholesale or retail
17. services             =1 if in services indus.
18. profserv             =1 if in prof. serv. Indus.
19. profocc              =1 if in profess. Occupation
20. clerocc              =1 if in clerical occupation
21. servocc              =1 if in service occupation
22. lwage                log(wage)
23. expersq              exper^2
24. tenursq              tenure^2
```

(a) **[7 Points]** Consider the following restricted version of equation (1) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Suppose that X_2 is omitted from the model by the researcher. For X_2 to cause omitted variable bias (OVB), what conditions should it satisfy? Show mathematically that the OLS estimator β_1 is biased if X_2 is omitted from the model.

(b) **[6 Points]** Run a regression of $Y_i = \beta_0 + \beta_4 X_4 + u_i$ and interpret the slope coefficient β_4. (Hint: X_4 is a binary explanatory variable.)

(c) **[7 Points]** First generate a dummy variable D_i such that $D_i = 1$ if male and $D_i = 0$ if female. Then run a regression of $Y_i = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_6 D_i + u_i$. What do you notice in the result? Explain why? Show mathematically that if X_4 and D_i are related, this result is inevitable.

(d) **[7 Point]** Run, first, a simple regression of $Y_i = \beta_0 + \beta_1 X_1 + u_i$ then $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$. Explain what happened to β_1(before and after) and why it happened.

(e) **[7 Points]** Now run the full model (1), using both homoscedastic-only and heteroskedasticity-robust standard errors, and interpret and compare the results of both regressions. Why do we care about heteroskedasticity problem that might exist in the data?

(f) **[6 Point]** Based on the regression result of the later (i.e., heteroskedasticity-robust standard errors), conduct the following hypothesis testing:
   i. H_0: $\beta_i = 0$ vs H_1: $\beta_i \neq 0$ where $i = 1, 2, \dots, 5$
   ii. H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs H_1: $At\ least\ one\ \beta_i \neq 0$

Q#3: [10points] Suppose that you have been asked to assess a research paper that is based on your regression results in 2(e). Evaluate the external and internal validity issues of your regression in 2 (e) above.

Q#4. [20 Points] Consider the following model, known as the exponential regression model:

$$Y_i = \beta_0 e^{\beta_1 X_i} + u_i$$

(a) [**6 points**] Do you think that you can estimate the model parameters using OLS? Explain.
(b) [**7 points**] What do you suggest how to estimate the model parameters?
(c) [**7 points**] How do you think one can proceed estimating these by trial-and-error, or iterative, process?

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

1. SW Exercise 7.1
2. SW Exercise 7.4
3. SW Empirical Exercises 7.1
4. SW Exercise 8.2
5. SW Exercise 8.10
6. SW Exercise 9.6
7. SW Empirical Exercise 8.2.
8. SW Empirical Exercise 9.2