*Department of Economics* *UN3412*
*Columbia University* *Spring 2026*

**Problem Set 3**
**Introduction to Econometrics**
**for Section 1**
**(Erden)**

---

Instructions: Please upload your solutions to GRADESCOPE as a single pdf. If you have several documents you need to combine them to a single pdf (You can use sites like https://www.ilovepdf.com/merge_pdf ) Please specify the page number of each solution while uploading your document.

**1.** (40p) Use Table 2 and **GPA4.dta** data file to answer the following questions.

**(a)** (8p) Table 2 presents the results of three regressions, one in each column. Estimate the indicated regressions and fill in the values (you may either handwrite or type the entries in, at you convenience). For example, to fill in column (1), estimate the regression with *colGPA* as the dependent variable and *hsGPA* and *skipped* as the independent variables, using the "robust" option, and fill in the estimated coefficients

**Table 2**
**College GPA Results**
Dependent variable: *colGPA*

| Regressor | (1) | (2) | (3) |
|---|---|---|---|
| *hsgpa* | ( ) | ( ) | ( ) |
| *skipped* | ( ) | ( ) | ( ) |
| *PC* | — | ( ) | ( ) |
| *bgfriend* | — | — | ( ) |
| *campus* | — | — | ( ) |

| | | | |
|---|---|---|---|
| *Intercept* | ( ) | ( ) | ( ) |
| ***Regression summary statistics*** | | | |
| $R^2$ | | | |
| *Regression RMSE* | | | |
| *n* | | | |

**(b)** (6p) Write the regression in column (1) in "equation form," with the standard error below the respective regression coefficient.

**(c)** (5p) Explain in words what the coefficient on *hsGPA* means in regression (1), holding *skipped* unchanged.

**(d)** (5p) Using regression (1), test the hypothesis that the coefficient on *skipped* is zero, against the alternative that it is nonzero, at the 5% significance level. In everyday words (not statistical terms), what precisely is the hypothesis that you are testing?

**(e)** (6p) Test the hypothesis that the coefficient on *skipped* is zero in regressions (1), (2), and (3) at 1% **using the p value**, does your answer change depending on what other variables are included in the regression?

**(f)** (5p) Using regression (3), consider the coefficient on *campus*. Does the sign and magnitude make sense? Explain.

**(g)** (5p) Using regression (3), consider the coefficient on *bgfriend*. Does the sign and magnitude make sense? Explain In regression (3), is the coefficient on *campus* statistically significant at the 1% significance level? Is the coefficient on *bgfriend* statistically significant at the 1% significance level?

**2.** (25p) **Nursing Home Utilization**. This question considers nursing home data provided by the Wisconsin Department of Health and Family Services (DHFS) and its description is in posted dataset.

**2.1.** (20p) **Part 1: Use cost-report year 2000 data, and do the following analysis.**

**a.** (6p) **Correlations**

i. (3p) Calculate the correlation between TPY and LOGTPY. Comment on your result.

ii. (3p) Calculate the correlation among TPY, NUMBED, and SQRFOOT. Do these variables appear highly correlated?

b. (4p) **Scatter plots.** Plot TPY versus NUMBED and TPY versus SQRFOOT. Comment on the plots.

**c.** (10p) **Basic linear regression, homoskedactic errors are fine here.**

i. (2p) Fit a basic linear regression model using TPY as the outcome variable and NUMBED as the explanatory variable. Summarize the fit by quoting the coefficient of determination, $R^2$, and the t-statistic for NUMBED.

ii. (3p) Repeat c(i), using SQRFOOT instead of NUMBED. In terms of $R^2$, which model fits better?

iii. (2p) Repeat c(i), using LOGTPY for the outcome variable and LOG(NUMBED) as the explanatory variable.

iv. (3p) Repeat c(iii) using LOGTPY for the outcome variable and LOG(SQRFOOT) as the explanatory variable.

**2.2.** (5p) **Part 2: Run the same regression as in Part 1.c(i) using 2001 data. Are the patterns stable over time?**

**3.** (35p) Consider the following Population Linear Regression Function (PLRF):

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

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 WAGE data set. 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. Var # 21

| | |
|---|---|
| 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

(a) (6p) 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?

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

(c) (6p) 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) (6p) 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 $\beta_1$(before and after) and why it happened.

(e) (5p) Now run the full model (1), using both homoskedasticy-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) (6p) 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 $i = 1, 2, \ldots, 5$

ii. $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ $vs$ $H_1$: $At\ least\ one\ \beta_i \neq 0$

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

1. SW Empirical Exercises 6.1
2. SW Empirical Exercises 7.1
3. SW Exercises 6.1 – 6.4.
4. SW Exercises 7.1 – 7.6.