Econ 322 - Fall 2025

Problem Set 3

Due October 22nd, 2025, 8 pm

Please answer each of the questions below. Note that writing only a numeric answer to the question is *not* enough to receive full credit unless otherwise stated. Please submit: (1) a PDF/Word Document with your answers (including any plots/tables), and (2) the `R` (or other language) code you used to answer the questions. *You need to submit both files for your problem set to be graded.*

Total: 60 points.

Question 1: Determinants of Educational Outcomes (32 Points)

This question studies the impact of environmental conditions, specifically high temperatures, on student learning outcomes. To do this, we will examine the relationship between mathematics scores from the Programme for International Student Assessment (PISA) and the average temperature across various countries. PISA is a triennial worldwide study conducted by the Organisation for Economic Co-operation and Development (OECD). It evaluates 15-year-old students' performance in key subjects, including mathematics, reading, and science, aiming to assess the extent to which students near the end of compulsory education have acquired the knowledge and skills essential for full participation in modern societies.

For your analysis, you will use the attached dataset "`grades_and_temps.csv`". This dataset contains observed PISA scores and corresponding average temperature data for a collection of countries across selected years between 2000 and 2012. The list of variables is as follows:

- `cname`: Country name
- `year`: Year
- `read_score`: Average reading score, 0 to 1000
- `math_score`: Average math score, 0 to 1000
- `sci_score`: Average science score, 0 to 1000
- `avg_temp`: Average yearly temperature in a given year, in Celsius degrees
- `gdppc`: GDP per capita, in 2012 dollars
- `income_group`: Income group of the country

Answer the following questions (please round all numerical answers to *three* decimal places):

1. (4 points) Construct a variable that *you will use throughout this question*: the average yearly temperature in a given year, expressed *in Farenheit degrees*. Call it `avg_temp_f`. Generate a scatter plot with the average yearly temperature (`avg_temp_f`) on the x-axis and the average math score (`math_score`) on the y-axis. Using visual inspection, do these variables seem to be positively correlated, negatively correlated, or not correlated at all?

2. (6 points) Consider the following equation (i represents countries):

$$\texttt{math\_score}_i = \alpha + \beta \texttt{avg\_temp\_f}_i + \varepsilon_i \tag{1}$$

   Regress the average math score on the average yearly temperature. Report the estimated intercept ($\hat{\alpha}$) and the estimated slope ($\hat{\beta}$). Interpret both coefficients. Does it make sense to interpret $\hat{\alpha}$?

3. (6 points) Based on visual inspection of the plot in question 1, do you think the errors are homoskedastic or heteroskedastic? Why? Compare the standard errors for $\hat{\alpha}$ and $\hat{\beta}$ both under the homoskedasticity and the heteroskedasticty asssumptions. *Hint: Real data is messy, so the scatter plot in part 1 may not look like the textbook case of homoskedasticty/heteroskedasticity. Use the definitions we learned in class to support your argument!*

4. (9 points) Using the heteroskedasticity-robust standard errors: (a) test the null hypothesis that $H_0 : \beta = 0$ at the 5% significance level, and (b) test the null hypothesis that $H_0 : \beta = -1.85$ at both the 5% and the 10% significance levels. Write out the t-statistic formulas to perform these two-sided tests.

5. (4 points) Suppose the OECD is seriously considering implementing stricter climate change agreements that are projected to decrease global average yearly temperatures by 2.35°F. Using the econometric model in Equation 1 and your estimated coefficients, compute by how much you would expect the average math score to change as a result of this projected temperature decrease.

6. (3 points) Do you think your estimate $\hat{\beta}$ is causal (e.g., does the answer in the previous part make sense to you)? Explain your answer.

Question 2: Why is Housing Construction so Expensive? (28 Points)

In recent years, housing prices have increased substantially. While this trend can be attributed to various factors (e.g., constraints in housing supply, increased demand), this problem set will focus on examining the extent to which local permit fees contribute to higher housing construction costs. These fees include various categories such as building, electrical, plumbing, fire, state surcharges, and certificate fees, and their amounts can vary significantly by municipality.

To investigate this relationship, consider the following regression equation:

$$\texttt{construction\_cost}_i = \alpha_1 + \beta_1 \texttt{fees}_i + \varepsilon_i \tag{2}$$

You will use the attached dataset, "`middlesex_permits.csv`", for your analysis. This dataset comprises real-world, publicly available data obtained from the New Jersey Department of Community Affairs. For the purpose of this question, the dataset has been restricted to include only new residential construction permits issued in Middlesex County since 2020. The variable definitions are as follows:

- `record_id`: Permit identifier
- `municipality_name`: Municipality
- `construction_cost`: Construction cost of the building, in dollars
- `units`: Number of units in the building
- `fees`: Total sum of all fees charged for the building, in dollars
- `square_feet`: Total square feet area of the building
- `volume`: Total cubic volume of the building

Answer the following questions (please round all numerical answers to *three* decimal places):

1. (6 points) Estimate Equation 2. Report your estimate for $\hat{\beta}_1$ and its respective heteroskedasticity-robust standard error. Interpret the coefficient.

2. (6 points) Since you took Econ 322, you suspect that Equation 2 may suffer from omitted variable bias (OVB). But you also have data on other permit characteristics, so you can investigate whether OVB is a concern! You start by exploring whether the number of units in the building can be a source of OVB. Compute the correlation between units and fees. Compute the correlation between units and construction cost. Based on these results, do you think your estimate of $\hat{\beta}_1$ is biased? If so, is it upward or downward biased? Provide an intuitive explanation.

3. (6 points) You decide to estimate the following equation:

$$\texttt{construction\_cost}_i = \alpha_2 + \beta_2 \texttt{fees}_i + \theta_2 \texttt{units} + \varepsilon_i \tag{3}$$

   Report the estimated $\hat{\beta}_2$ and $\hat{\theta}_2$ coefficients, along with their respective heteroskedasticity-robust standard errors. Interpret the coefficients. How does $\hat{\beta}_2$ compare with your estimate $\hat{\beta}_1$ from Equation 2? Relate the answer to this question to your answer in the previous part. *Hint: You can easily include more variables in your regression using a plus sign. E.g.,* `lm(y~x+z, data=df)` *regresses* `y` *on both* `x` *and* `z`*.*

4. (6 points) Still concerned about OVB coming from other variables, you decide to "throw the kitchen sink" and include all available variables in the regression:

$$\begin{aligned}\texttt{construction\_cost}_i = \quad & \alpha_3 + \beta_3 \texttt{fees}_i + \theta_3 \texttt{units} + \kappa_3 \texttt{square\_feet} + \nu_3 \texttt{volume} \\ & + \varphi_3 \texttt{new\_brunswick} + \varphi_3 \texttt{edison} + \varepsilon_i\end{aligned} \tag{4}$$

   Note that you will need to construct two new variables from the variable `municipality_name`: `new_brunswick` is a dummy variable that takes value 1 if the permit was issued in New Brunswick, 0 otherwise; and `edison` is a dummy variable that takes value 1 if the permit was issued in Edison, 0 otherwise. Report all the estimated coefficients in this regression, along with their heteroskedastic-robust standard errors. How does $\hat{\beta}_3$ compare with your estimate $\hat{\beta}_2$ from Equation 3? Do you think that your estimate of β_3 is causal? Explain. *Hint: you can check the exact values of* `municipality_name` *and their frequency using* `table(mydata$municipality_name)`*.*

5. (4 points) In the previous regression, (i) How do you interpret α_3? Does this interpretation make sense? and (ii) How do you interpret the coefficient on `new_brunswick`? Use your estimates to answer these questions.