Assignment 02

Simple Linear Regression: Description

Should more money be spent on public schools or should that money be spent elsewhere? Both sides of this ongoing public debate have been argued passionately, using a multitude of anecdotal evidence. Although we will not settle this debate, we will examine data akin to the types of data that policy makers use to make funding decisions. Specifically, we will examine whether teacher salaries are related to SAT scores at the state level. For this assignment, you will use the file *state-education-data.csv*. This file contains state-level aggregate data. The variables are:

- `state`: State name
- `postal`: State postal code
- `region`: Region of the country (Midwest, Northeast, South, West)
- `salary`: Average teacher salary in the state
- `sat`: Average SAT score in the state
- `participation`: Percentage of students in the state who took the SAT

This assignment is worth 16 points. Each question is worth 1 point unless otherwise noted.

Please submit your responses to each of the questions below in a printed document. Also, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be resized so that they do not take up more room than necessary and should have an appropriate **caption** and **labels**.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document using Equation Editor, Markdown, or LaTeX.

Part I

Before carrying out any analyses, create a predictor called `salary_thousand` that indicates the average state salary in thousands of dollars (e.g., salary = 52143; salary_thousand = 52.143). This variable (not `salary`) should be used in all analyses in Part I.

1. Create a plot of the marginal distribution of SAT scores. Make sure your plot has a caption.
2. Examine the structure and formatting of Table 1 in the article: Snedker, K. A., Herting, J. R., & Watson, E. (2009). Contextual effects and adolescent substance use: Exploring the role of neighborhoods. *Social Science Quarterly, 90*(5), 1272—1296. Notice that variables are presented in rows and summary statistics are presented in columns. Mimic the format and structure of this table to create a table to present the numerical summary information for the marginal distributions of SAT scores and salaries. Provide the same measures for these variables as is given in Table 1 in the article. Re-create the formatting of Table 1 as closely as you can. Finally, make sure the table you create also has an appropriate caption.
3. Create a plot of the distribution of SAT score *conditioned on* teacher salary (i.e., a scatterplot). Make sure your plot has a caption.
4. Describe the relationship between SAT scores and teacher salaries. Be sure to comment on the structural form, direction and strength of the relationship. Also comment on any potential observations that deviate from following this relationship (unusual observations or clusters of observations).
5. Compute and report the Pearson correlation coefficient between SAT scores and teacher salaries.

6. Based on your answer to the Question #4, is the Pearson correlation coefficient an appropriate summary measure of the relationship? Explain. (Hint: Pay attention to the structural form!)
7. Regress SAT scores on teacher salaries. Write the *fitted equation* using Equation Editor (or some other program that correctly types mathematical expressions).
8. Interpret the value of the intercept and the slope from the regression equation using the context of the data.
9. Compute, report, and interpret the value for R^2 based on values from the ANOVA decomposition.

Part II

Center the `salary_thousand` predictor by subtracting the mean teacher salary from each value. Call this new variable `center_salary_thousand` This variable should be used in all analyses in Part II.

10. Report and interpret the value of Minnesota's centered teacher salary.
11. Compute and report (a) the mean centered teacher salary, (b) the standard deviation of centered teacher salaries, and (c) the Pearson correlation coefficient between SAT scores and centered teacher salaries. How do these values compare to the values you computed earlier using the uncentered teacher salaries?
12. Explain how the slope of the regression you found in Question #7 will compare to the slope of the regression if we regressed SAT scores on the centered teacher salaries by making reference to the values in the mathematical formula for slope, which is:

$$r \times \left(\frac{SD_{\text{Outcome}}}{SD_{\text{Predictor}}} \right)$$

13. Regress SAT scores on the centered teacher salaries. Write the *fitted equation* using Equation Editor (or some other program that correctly types mathematical expressions). (You should be able to now check your previous response as well.)
14. Interpret the value of the intercept from the regression equation using the context of the data. (Hint: Think about what the predictor value of zero represents.)

Part III

Convert the uncentered teacher salaries (`salary_thousand`) into z-scores by subtracting the mean salary and dividing by the standard deviation. (See here if you need a refresher on z-scores.) Call this new variable `z_salary`. Also convert the SAT scores into z-scores and call that variable `z_sat`.

15. Regress the `z_sat` variable on the `z_salary` variable. Interpret the value of the intercept from the regression equation using the context of the data. (Hint: To help with the intercept interpretation, think about what the predictor value of zero represents. To help with the slope interpretation, consider what a one-unit difference represents on the z-scale.)
16. Using the mathematical formula for slope from Question #12, explain why the slope from regressing z_{Outcome} on $z_{\text{Predictor}}$ will be the correlation coefficient between the predictor and outcome.