Assignment 03

Simple Linear Regression: Inference

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 14 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 for Part I. Fit a regression model using teacher salaries to predict SAT scores.

1. Using symbols, write the null hypothesis that is tested by the F-statistic in this analysis.
2. Write no more than three sentences (to be included in a publication) that summarizes the results of the omnibus analysis. A summarization of the results includes a written description of what is being tested by the F-test and the statistical results. At a minimum report the F-statistic, df, and p-value. A summary should also indicate what the statistical results suggest about the tenability of the null hypothesis and what this means about the potential relationship between age and book length.
3. Using symbols, write the null hypothesis that is tested by the t-statistic for the slope.
4. Based on the results of the t-test, what do the data suggest about the tenability of the null hypothesis for the slope? Explain.
5. Compute and interpret the confidence interval for the slope.

Part II

6. Examine the structure and formatting of Table 1 in the article: Garcia, D. R., McIlroy, L., & Barber, R. T. (2008). Starting behind: A comparative analysis of the academic standing of students entering charter schools. *Social Science Quarterly, 89*(1), 199–216. Notice that models are presented in columns (in the article 6 models are presented). Predictors used in the models are presented in rows, and so are the model-level summaries (e.g., N, F, R^2). Also note that the intercept ('Constant') is the last term presented in the table, generally because it is the least important coefficient. Blank cells indicate that the model does not include that particular predictor. Mimic the format and structure of this table to create a table to present the numerical information from the model you fitted in Part I of this assignment. Re-create the formatting of Table 1 as closely as you can. Instead of giving the adjusted R^2 value, provide the unadjusted R^2 value. Make sure the table you create also has an appropriate caption.

7. Create a plot that displays the regression line from the analysis in Part I. This plot should also include a scatterplot of the observed data. The data should be semi-transparent, and the regression line should be completely opaque (non-transparent). Also plot the point that represents the mean salary and mean SAT score. Make this point larger so it can easily be seen on the plot. Give your plot an appropriate caption.

Part III

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. Regress the SAT scores on the centered salaries.

8. The results of the F-test for this analysis are identical to the results of the F-test for the analysis in Part I. Explain why this is expected by referring to and comparing what is being tested in the hypothesis in both sets of analyses.

9. The results of the t-test for the intercept in this analysis are different than the results of the t-test for the intercept in the analysis in Part I. Explain why this is expected by referring to and comparing what is being tested in the hypothesis in both sets of analyses.

10. Create a plot that displays the regression line from the analysis in Part III. This plot should also include a scatterplot of the observed data. The data should be semi-transparent, and the regression line should be completely opaque (non-transparent). Also plot the point that represents the mean salary and mean SAT score. Make this point larger so it can easily be seen on the plot. Give your plot an appropriate caption.

11. Compare and contrast the plots from the two analyses. What is the same? What is different?

Part IV

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`. Regress the SAT z-scores on the salary z-scores.

12. Create a plot that displays the regression line from the analysis in Part IV. This plot should also include a scatterplot of the observed data. The data should be semi-transparent, and the regression line should be completely opaque (non-transparent). Also plot the point that represents the mean salary and mean SAT score. Make this point larger so it can easily be seen on the plot. Give your plot an appropriate caption.

13. The p-value of the t-test for the intercept in this analysis is one. Explain why this is expected by referring to what is being tested in the hypothesis in this analysis. (Hint: Think about what the intercept is and how that relates to what is being tested.)

14. The test of the slope (regardless of analysis) suggests that teacher salaries are significantly related to SAT scores. Unfortunately this relationship is negative, indicating that higher teacher salaries are associated with lower SAT scores. A public-policy wonk wants to use this data to support the de-funding of public schools. Write a couple sentences that explain to this person why your analysis does not support this conclusion based on the study design.