Assignment 03

Simple Linear Regression: Inference

This assignment is intended to give you more experience fitting and interpreting regression models. Submit your responses to each of the questions below in a printed document. All graphics should be resized so that they do not take up more room than necessary and also should have an appropriate caption. This assignment is worth 13 points. (Each question is worth 1 point unless otherwise noted.)

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.csv* (see the data codebook).

Part I: Unstandardized Regression

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 for the slope, are the empirical data consistent with the null hypothesis that the sample slope is entirely due to sampling error? Explain.
5. Compute and interpret the confidence interval for the slope.
6. Create a plot that displays the regression line from the unstandardized regression analysis. 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 II: Centering a Predictor

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.

7. 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 the hypothesis being tested in both sets of analyses.

8. 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 the hypothesis being tested (and what that means) in both sets of analyses.

9. Create a plot that displays the regression line from the centered-predictor regression analysis. 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.

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

Part III: Standardized Regression

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.

11. Create a plot that displays the regression line from the standardized regression analysis. 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.

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

13. The test of the slope (regardless of analysis) suggests that teacher salaries seem to be 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.