FINAL PAPER

CAP4773

MARCH 31, 2021

SUMMER POISSONNIER
Florida Atlantic University

In this paper I will be discussing my findings from assignment 2. In assignment 2 I used the College dataset and the package this dataset came from is the ISLR package. The college dataset contains quantitative and qualitative variables. These variables represent important information that is typical to any university or college. This dataset only contains one qualitative variable: Private (measures yes or no indicating a private or public university). The remaining 17 quantitative variables in this dataset are called: Top10perc (representing the students in the top 10 percent of their high school class), Top25perc (representing the students in the top 25 percent of their high school class), Outstate (Out-of-state tuition), Apps (number of applications), Accept (number of accepted applications), Enroll (number of enrolled students), F.undergrad (number of fulltime undergraduate students), P.undergrad (number of part-time undergraduate students), Room.Board (cost of the room and board), Books (cost of books), Personal (personal spending), PhD (Percent of faculty with PhD's), Terminal (Percent of faculty with terminal degree), S.F.Ratio (Student and faculty ratio), perc.alumni (Percent of alumni who donate), Expend (instructional expenses per student), Grad.Rate (graduation rate).

The main problem addressed in assignment 2 was predicting a college's out-of-state tuition (the response or Y variable) from the percentage of students from the top 10% of their high school class (the feature or X variable). In mathematical terms, we are predicting Y from X. The variables used from the college data set to determine the solution to this problem were Outstate and Top10perc. The approach taken to address this problem was linear regression. I used linear regression to fit a model for the data being studied, I completed a hypothesis test to evaluate a linear relationship, and I examined the fit of least squares linear model.

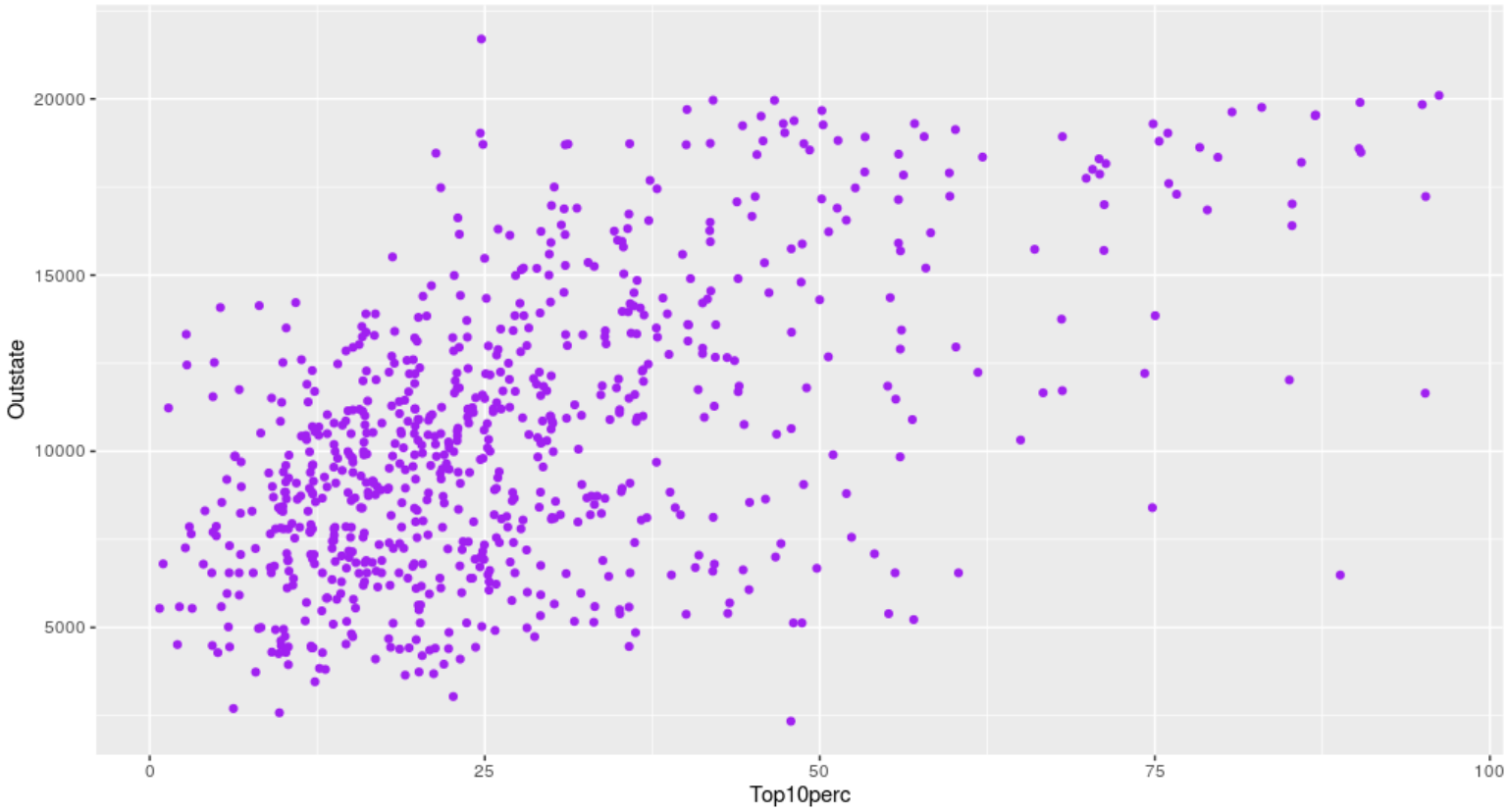
In part one of the assignment, I used simple linear regression to fit a model for the data. I coded for the model using the R language in R studio. From this model I found the y intercept and the slope of the least squares regression line. The y-intercept was 6906.5 and the slope of the least squares regression line was 128.2. I was then able to write a formula for the least

squares regression line using these variables. The formula for the purpose of this model is: Outstate = 6906.5 + 128.2Top10perc. I then had to find 95% confidence intervals for the slope, B_1, and the y-intercept, B_0 . The 95% confidence interval for B_0 was: [6471.424, 7341.493] and the 95% confidence interval for B_1 was: [114.946, 141.541]. I then completed a point estimate to obtain a 95% confidence interval for the predicted out-of-state tuition when 33% of the students are from the top 10% of their high school class. The point estimate I obtained was: 11138.5 and the confidence interval was: [10893.16, 11383.84]. The point estimate I obtained was in between the confidence interval, proving the accuracy of the result. For the final part of fitting the data with the simple linear regression model, I created a scatterplot in R with the feature on the x axis and the response on the y axis. To avoid overplotting I had to include a position called 'jitter' in my code. The scatterplot contained the data from the regression model and a least squares regression line with a 95% confidence band.

In part two of the assignment, I conducted a hypothesis test to evaluate the linear relationship between Outstate and Top10perc. I predicted that the null hypothesis would be H_0: $\beta_1 = 0$ and the alternative hypothesis would be H_A: $\beta_1 \neq 0$. Upon running my R code, I was able to find the test statistic, 18.93, and the p-value which was less than .05. Since the p value was very small and below a certain threshold, I was able to reject the null hypothesis and conclude that $\beta_1 \neq 0$. The data in this finding supports the prediction that there will be a linear relationship between the college's out-of-state tuition and the top 10 percent of students from their high school class. I was able to conclude that the Outstate and Top10perc were directly related.

In the final part of the assignment, I examined the fit of the least squares linear model. In this part I created a plot that showed the data points, the least squares regression line, and vertical line segments that connected the data points to the least squares regression line. I also made the data points red to help visualize them. The vertical line segments in the plot represented the standard deviation of residuals. I then provided the RSE and R^2 values. The RSE was 3329 and represented the average deviation between

observed and predicted out-of-state tuition, in dollars $3329. The R^2, I found, was 0.3162. The R^2 value represents the proportion of variation in a college's out-of-state-tuition that can be explained by the percentage of students from the top 10 percent. There was a 31.62% variation.



The scatter plot I created above provides a simple look at the data. The response, Top10perc is on the X axis and the feature, Outstate is on the y axis. This simple scatter plot shows the data for the colleges represented in the color purple. Upon glancing at this scatterplot without completing any simple linear regression tests, one can tell that the data is positively and linearly related.

In the first part of the assignment, using a linear regression model, I found that the results proved that there is a linear relationship between Outstate and Top10perc. The data proving this was the least squares regression line: Outstate = 6906.5 + 128.2Top10perc. This line proves that there is a

positive linear relationship as there is only one X variable and the slope is positive. I was also able to conclude that the scatter plot created in part one proved the linear relationship between the variables as the least squares regression line was a positive linear line that fit the data. In the second part of the assignment, I used the hypothesis test to prove that the data was linearly related. Since I was able to reject the null hypothesis and support the alternative hypothesis through a very small p value, this result showed that the data is linearly related. In the final part of the assignment, examining the fit of the least squares regression line, the variation of 31.62% demonstrates that there is a variation between Outstate and Top10perc, thus further proving the linear relationship between the response and feature. The R^2 and RSE measure the goodness of fit values. The RSE measures the lack of fit and it is better for the model to have a smaller RSE; in assignment two, the RSE value was quite large. For the R^2 value of .3162, this is a direct estimate of the goodness of fit of the model that ranges from 0-1 (with 1 meaning there is a perfect fit for the model). Since the value in assignment two is closer to 0, the model does not have the best fit.

In conclusion, the methods used to predict the linear relationship between a college's out of state tuition and the top ten percent of students from the high school class were average to say the least. There are better methods of examining this data such as through multiple linear regression in place of simple linear regression. In multiple linear regression, there are multiple features compared to one response. The overall fit of the model can be improved through multiple linear regression since there are more features which can improve the accuracy and results of the regression model. In this method, there are more variables to study and more data provided to prove the linear relationship of Outstate compared to the other features in the college dataset.