Shruti Alavala
2/28/22

Final Project

Part One – Research Scenario and Question

The SAT is a standardized test that is widely used in the college admissions process. The test is typically taken by high school juniors and seniors and assesses the students' skills in three areas: critical reading, writing, and math. Each area is scored individually and then the three scores are combined for the final score. It is reasonable to think that a student's score in the critical reading and writing sections would be correlated as both of these sections test language skills. But how does the math score relate to the other two scores? The research question of interest is: Are SAT Critical Reading and SAT Writing scores, when considered together, predictors of the SAT Math score?

Part Two – Dataset Description

In 2013, New York City's Department of Education published the average SAT scores per school of students who graduated in 2012. There are 478 records in this dataset where each row represents a different school in New York City. The variables in this dataset are:

- DBN – a unique identifier for schools in the city
- School Name – the name of the school
- Num of SAT Test Takers – the number of students graduating in 2012 who took the SAT in that school
- SAT Critical Reading Avg. Score – the average critical reading score for the students who took the SAT in that school
- SAT Writing Avg. Score – the average writing score for the students who took the SAT in that school
- SAT Math Avg. Score – the average math score for the students who took the SAT in that school

The following data validation and cleaning was performed on this dataset:

1. Some of the rows in the dataset contained the value "s" in the "Num of SAT Test Takers", "SAT Critical Reading Avg. Score", "SAT Writing Avg. Score", and "SAT Math Avg. Score" columns instead of an actual number or score. It is unclear why these records contain the value "s" but since that is not a valid value for those columns, those records were removed from the dataset. This left 421 schools remaining in the dataset.
2. The data type of the "Num of SAT Test Takers", "SAT Critical Reading Avg. Score", "SAT Writing Avg. Score", and "SAT Math Avg. Score" columns were converted from character to numeric so that statistical analysis could be performed.
3. The possible range of scores in each section of the SAT is 200 to 800 so the minimum and maximum of "SAT Critical Reading Avg. Score", "SAT Writing Avg. Score", and "SAT Math Avg. Score" columns was verified to be in that range using the summary() function.

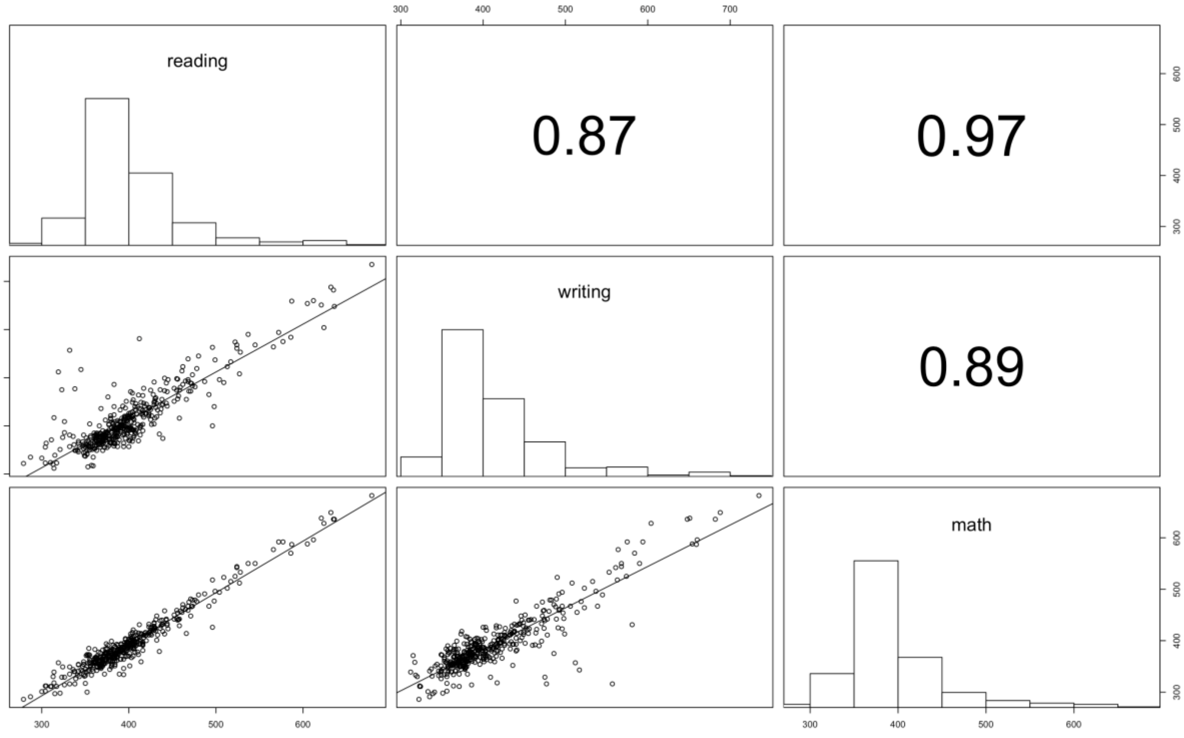
Link to Dataset - https://data.cityofnewyork.us/Education/2012-SAT-Results/f9bf-2cp4

Part Three – Statistical Methods

The main statistical method I will use in my analysis of the research question above is multiple linear regression. This is the appropriate choice because we are interested in understanding the relationship between a response variable, SAT Math score, and two or more explanatory variables, SAT Critical Reading and SAT Writing scores, where all the variables are quantitative and continuous. A multiple linear regression will not only allow us to evaluate whether both explanatory variables together have a relationship with the response variable but also whether each individual explanatory variable has a relationship with the response variable, after accounting for the other explanatory variable. In the context of our research question, if there is significant evidence that SAT Critical Reading and Writing scores are predictors of SAT Math scores, we can also determine the individual effect of SAT Critical Reading and Writing scores on the SAT Math score. Multiple linear regression also results in an equation that quantifies the relationship between the response and explanatory variables and allows us to predict a new observation of the response variable, given values for the explanatory variables.

Part Four - Results

First, in order to get a sense of the data, we generate a graphic that displays the scatterplots, distributions, and correlations between the three variables.



All three scatterplots show linear form, positive direction, and a strong association between the two variables being considered. The scatterplot between the math and reading score has the strongest relationship, as proven by the very strong correlation of 0.97. The three distributions are also similar in that they are all skewed right, meaning there are more low scores than high scores.

Next, we will run the multiple linear regression and produce the least square regression equation. The summary of this regression will allow us to determine whether SAT Critical Reading and SAT Writing scores are predictors of SAT Math score.

```
m <-lm(math~reading+writing)
m

Call:
lm(formula = math ~ reading + writing)

Coefficients:
(Intercept)      reading      writing
    -9.9308       0.8436       0.1591
```

Least Squares Regression Equation:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_{reading}} x_{reading} + \widehat{\beta_{writing}} x_{writing}$$

$$\hat{y} = -6.9308 + 0.8436 x_{reading} + 0.1591 x_{writing}$$

```
summary(m)
```

Residual Standard Error	13.28 on 418 df
Multiple R-squared	0.9489
Adjusted R-squared	0.9487
F-statistic	3883 on 2 and 418 DF
p-value	< 2.2 x 10^{-16}

With an alpha level of 0.05, we have significant evidence that SAT Critical Reading and SAT writing scores, when taken together, are predictors of a school's average SAT Math score, since the p-value is less than alpha. The R-squared value tells us that 94.89% of the variability in average SAT Math score can be explained by average SAT Critical Reading and SAT Writing scores together.

```
summary(m)
```

	Estimate	**Standard Error**	**t value**	**p value**
Intercept	-9.93082	4.63017	-2.145	0.0325
SAT Critical Reading	0.84363	0.02335	36.135	< 2 x 10^{-16}
SAT Writing	0.15906	0.02050	7.758	6.67 x 10^{-14}

```
confint(m, level = 0.95)
```

	2.5%	**97.5%**
Intercept	-19.0321446	-0.8295026
SAT Critical Reading	0.7977350	0.8895178
SAT Writing	0.1187564	0.1993553

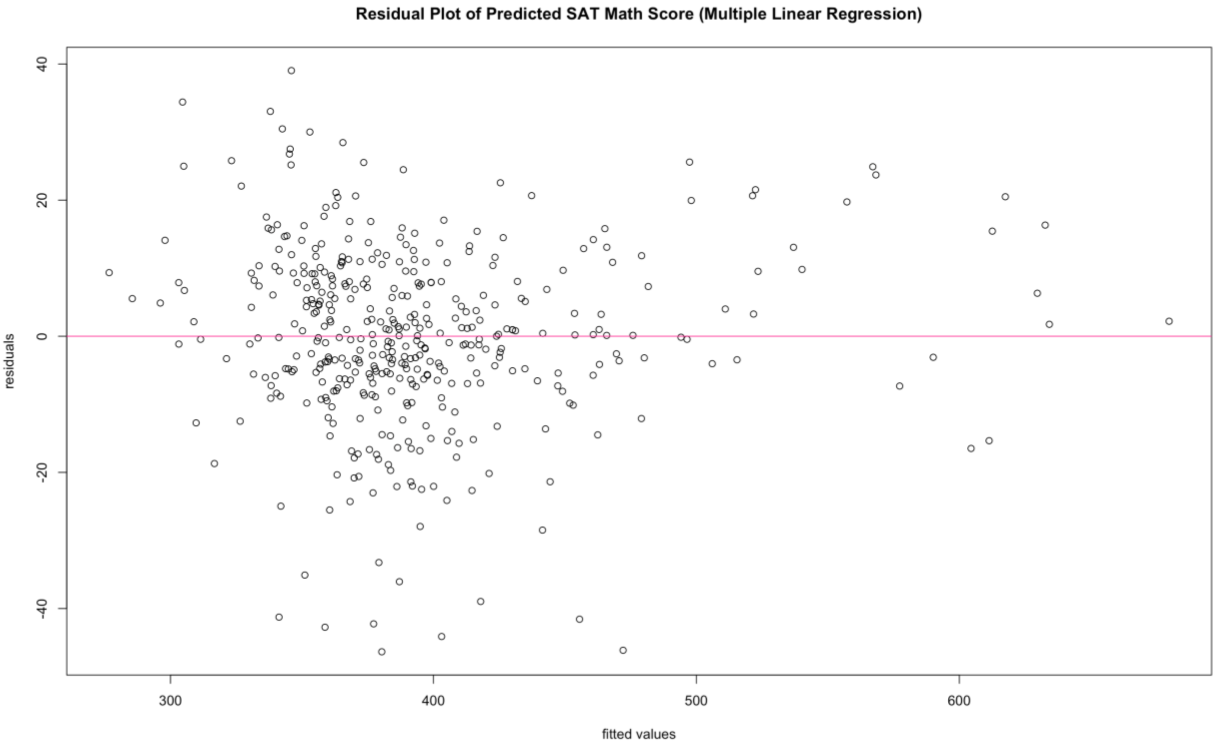
SAT Critical Reading score

After controlling for the SAT Writing score, we can say that the SAT Critical Reading score is a significant predictor of a school's average SAT Math score because the p-value (< 2 x 10^{-16}) for this variable is less than the significance level of α = 0.05. Additionally, the slope for this variable can be interpreted as an increase in average SAT Math score by 0.84 points on average for every additional point in SAT Critical Reading. We are 95% confident that for every additional point in SAT Critical Reading, the SAT Math score will be between 0.80 and 0.89 points higher.

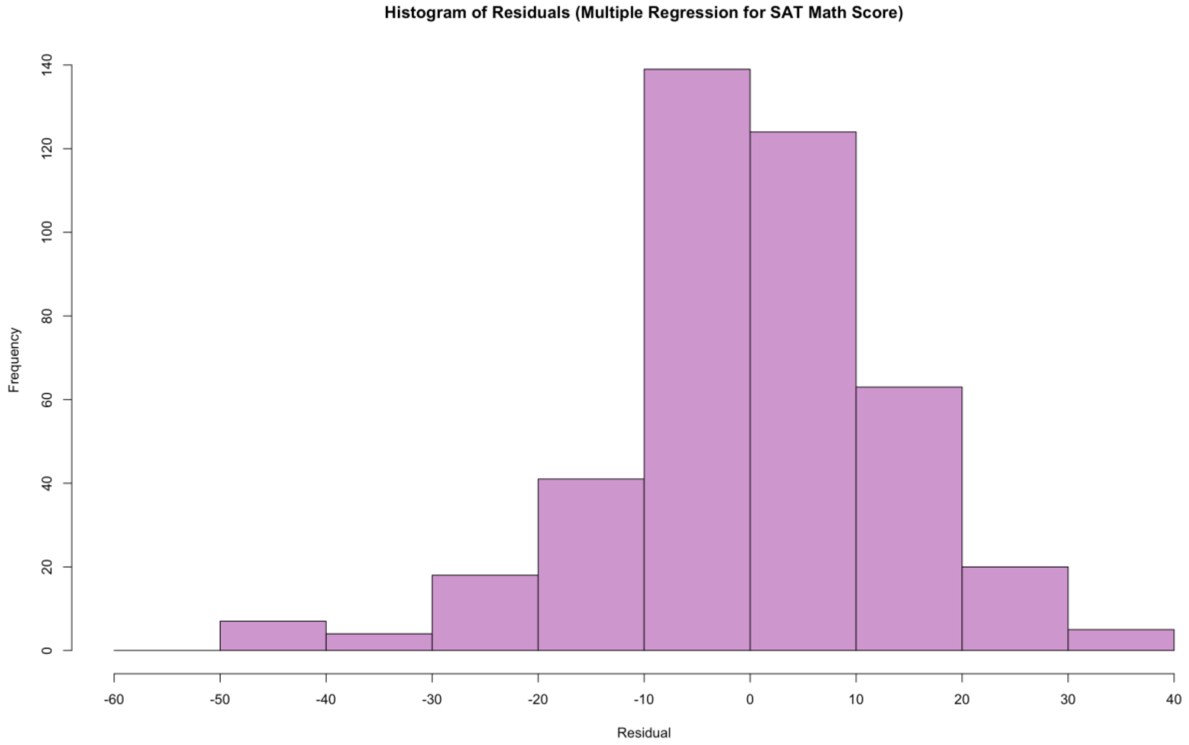
SAT Writing score

After controlling for the SAT Critical Reading score, we can say that the SAT Writing score is a significant predictor of a school's average SAT Math score because the p-value (6.67 x 10^{-14}) for this variable is less than the significance level of α = 0.05. Additionally, the slope for this variable can be interpreted as an increase in average SAT Math score by 0.16 points on average for every additional point in SAT Writing. We are 95% confident that for every additional point in SAT Writing the SAT Math score will be between 0.12 and 0.20 points higher.

Since the slope associated with SAT Critical Reading is larger than the slope associated with SAT Writing, we can say that the SAT Critical Reading score has more of an impact in predicting the SAT Math score than the SAT Writing score.



The residual plot above can be used to assess linearity and constant variance which are two of the conditions in deciding whether it is appropriate to use the least-squares regression equation to make inferences and predictions. The linearity assumption holds for the data because there the residual plot does not show a curved relationship, instead the points are randomly scattered. However, the constant variance assumption is violated because there is more scatter in the left side of the plot than the right side. The variability of the residual decreases as the fitted value (predicted SAT Math score) increases. The normality assumption can be assessed using the distribution of the residual values below. The distribution approximates a bell-shaped curve, even though it is not totally symmetric, so the normality assumption is not violated. The final condition that needs to be checked is independence. It is hard to say whether the data points are completely independent because while each point represents the average scores for students at different schools, all of the schools are from the same city and under the purview of the same Department of Education.



Part Five – Conclusions and Limitations

After running the multiple linear regression, we can conclude that SAT Critical Reading and SAT Writing scores, when considered together, are significant predictors of SAT Math score for students in New York City schools. We can also draw conclusions about the relationship between each individual explanatory variable and the response variable. SAT Critical Reading score, after controlling for SAT Writing score, is a significant predictor of SAT Math score. SAT Writing score, after controlling for SAT Critical Reading score, is a significant predictor of SAT Math score.

However, there are some important limitations to this analysis and dataset. We found that the correlation coefficient between the two explanatory variables (SAT Critical Reading score and SAT Writing score) is 0.87. This is a very high correlation coefficient and means that the two explanatory variables are highly correlated. This situation is called collinearity and can be problematic when included in a multiple linear regression because it can produce inaccurate results and conclusions. Another limitation found in this analysis is that the constant variance assumption was violated. Since this assumption was violated, it is not appropriate to make inferences or predictions using the regression equation because they could be inefficient or misleading.

In conclusion, while this dataset seemed to fit the bill for conducting a multiple linear regression analysis, it was not an appropriate analysis because of collinearity and the constant variance assumption being violated. In the future, to better answer the research question, we could collect data SAT scores from individual students who attend different schools from different school districts and perform the multiple linear regression. This would introduce more variability into the dataset and hopefully eliminate the collinearity and violated constant variance problems.