Do different properties of Massachusetts high schools impact the SAT scores of their students.

Introduction

During the pursuit of higher academic achievement, the SAT is the most critical moment where the future of students are decided. As an important factor in deciding university and college admissions in the United States, there have been assumptions regarding how students gain higher SAT scores. In an effort to quantify these assumptions, numerous studies were conducted to find relationships between SAT scores and affecting factors. In previous studies, a relationship was found between the density of competition and SAT performance (Garcia, Tor, 2009). As the number of competitors in a test taking environment increased, the performance level of students increased. Moreover, evidence regarding factors affecting student's study environment such as academic expenditure by schools (Smith, Eccles, 1998) and socioeconomic status of students (Zwick, Greif, 2007) suggests that as the environment improved, so did the SAT results.

However, these studies were limited to relating SAT scores with one specific part of a student's learning environment. Furthermore, some variables used such as morale (Smith, Eccles, 1998) are hard to quantify properly. Therefore, the purpose of this research is to find relations between SAT scores and other factors. By comparing variables affecting a student's academic environment in multiple different aspects, the effect these variables have on SAT scores will be explored.

Methods

This study was conducted using a dataset containing information regarding SAT scores and the student's learning environment (Dalziel, 2017). Through a data cleaning process, schools only containing grades 9 to 12 were extracted and individual SAT subjects were combined into an average SAT score. The predictors selected were average class size, average teacher salary, percentage of non-English speakers, percentage of students with disabilities and percentage of economically disadvantaged students. To find if a linear relationship exists between SAT scores and the predictors, a linear regression model was found to be suitable to see what the important predictors are if there is a linear relationship. However, before actual analysis may begin, assumptions had to be verified using residual plots. After the verification process, the simple linear regression model was fitted for individual predictors to investigate if a linear relationship

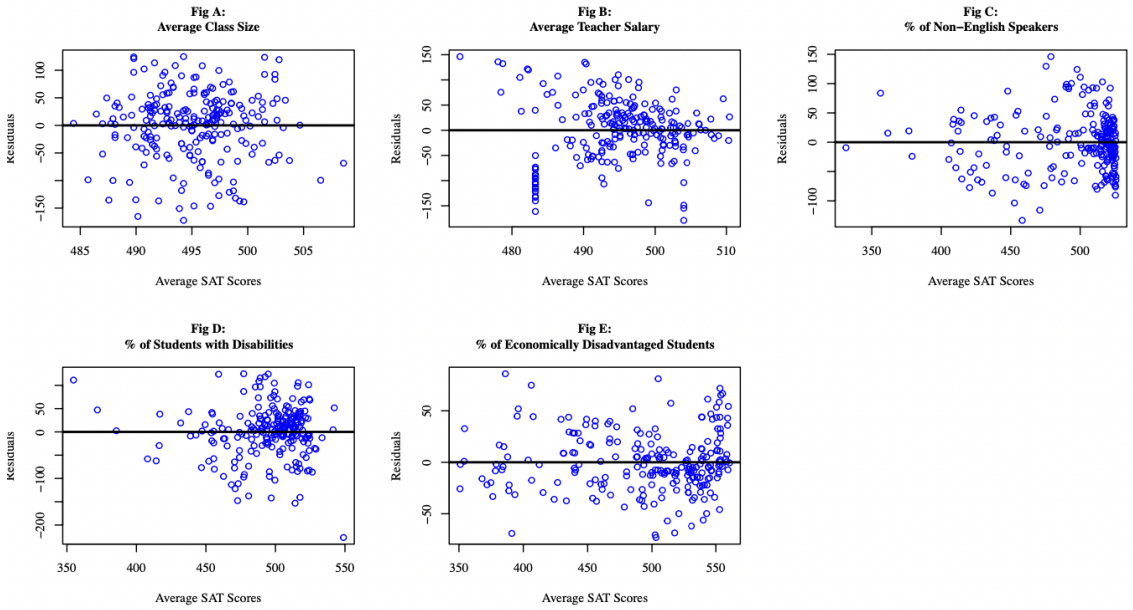
exists between the predictors and the response variable. If stabilization of the model was required, influential observations were identified based on leverage points and removed if required and then re-fitted.

To see if a multiple linear relationship exists between the predictors and the response variable, a multiple regression model was created similar to the simple regression model. The normality and linearity assumptions were checked using residual plots and the multiple regression line was created using a scatter plot. Similar to the simple regression model, if the fit could have been improved with a transformation, the process was repeated and a final multiple regression line was created. Afterwards, the variance inflation factor was found to ensure that there is no multicollinearity so that the predictors are not correlated with each other. Finally, the predictors that were to be included in the model were selected.

Results

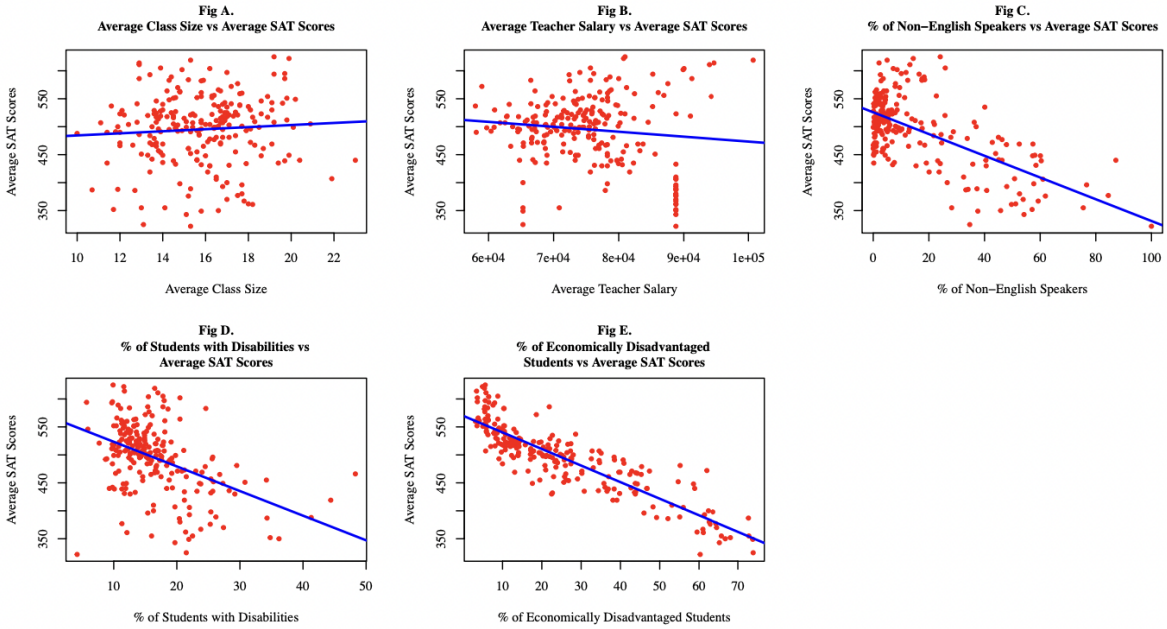
In order to verify the assumptions required before carrying on the analysis, a residual vs. fitted plot was created as shown in Figure 1. This was to find that no evident patterns are seen in the residual plots and that the residuals are uniformly distributed around zero for the full range of predictor values. The residual vs. fitted plot showed no cluster of residuals or systematic patterns are found. Moreover, the homoscedasticity assumption is verified as there was no fanning pattern where residuals gradually become more spread out. However, some predictors, especially Figure 1.B, and 1.C, exhibit unusual characteristics which may require re-fitting using transformation to relax the issue.

Figure 1. Residual vs Fitted Plot for each predictor



A Q-Q plot was used to verify the normality assumption and even though there were some points that deviate, all observations lie approximately on a straight line. After the verification process, it was found that through general observation it was difficult to confirm that the required assumptions were all verified. Therefore, before further analysis scatterplots for the predictor variables were plotted to see whether the predictors can actually be fitted into linear models. When the scatterplot was created as shown in Figure 2, it shows that the observations clearly do not fit into a linear model and therefore a stabilization method was required.

Figure 2. Scatterplot for each predictor.

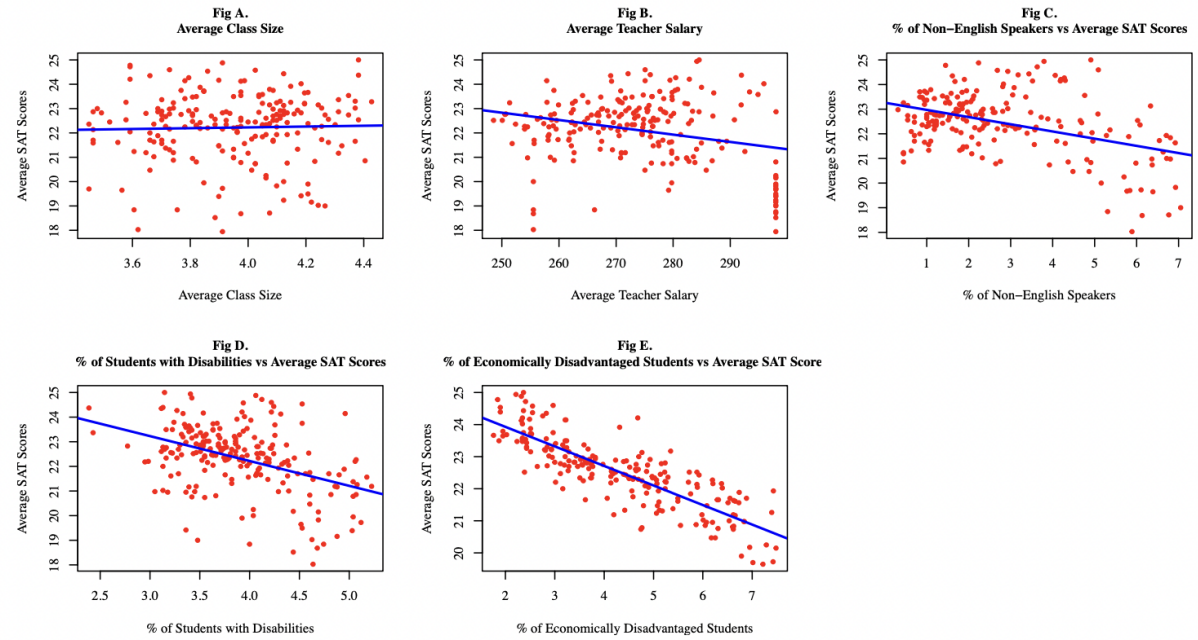


To stabilize the model, influential points were first identified based on leverage points using standardized residuals as shown in Appendix 1. The models were refitted after the influential points were removed yet the linear regression continuously did not fit the data. After this point, the data utilized for analysis was the one used for refitting with the influential points removed. Continuing on with the analysis, the homoscedasticity assumption needed to be re-verified and a standardized residual plot was recreated with the new data which concluded that the homoscedasticity assumption is violated as the residuals gradually become more spread out. This meant that a variance stabilizing transformation should be used.

A square-root transformation and a log transformation were used to compare the results and identify which transformation fit the data best. As it can be seen in Appendix 2, the square root transformation showed a better stabilized variance as the plots are randomly distributed

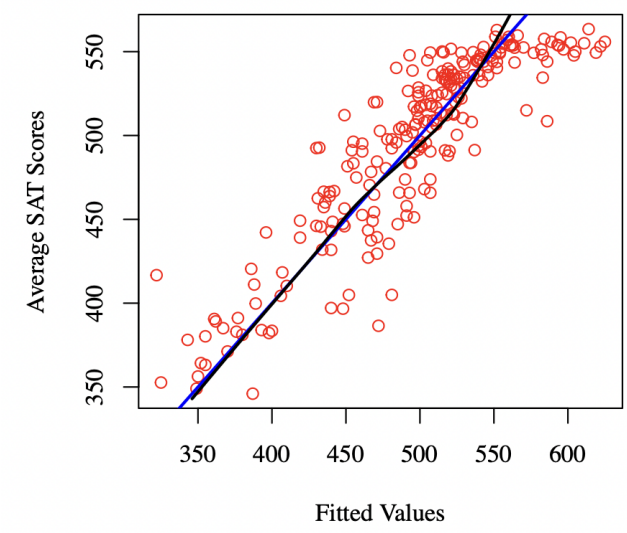
without any pattern. However, as the transformed data cannot interrupt the linearity of the model, it must be checked. The scatterplot and regression line shown in Figure 3 show that most of the variables did not display a linear relationship with the response variable of SAT scores. However, the variable percentage of economically disadvantaged students somewhat shows a linear relationship with the response variable as shown in Figure 3.E. The regression coefficient of this relationship shows that on average, as the percentage of economically disadvantaged students increase by 1%, then the predicted average SAT scores decrease by 2.91 as calculated in Appendix 3.

Figure 3. Square-root transformed scatterplot for each predictor



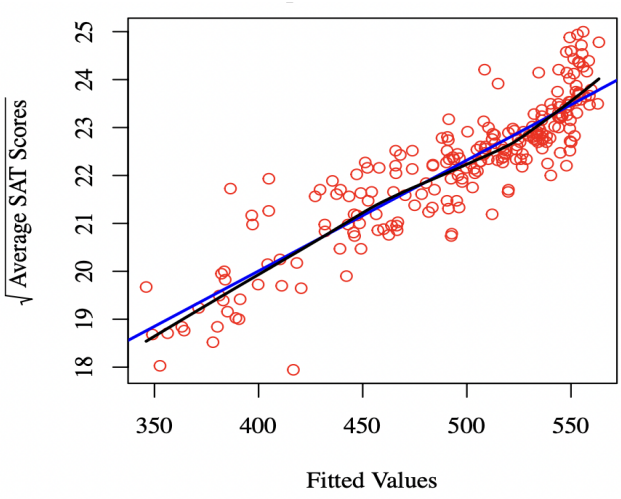
For the multiple regression model, the assumptions had to be verified as well. The normal Q-Q model and the residual vs. fitted plot indicated that the standard residuals deviate from normal and the errors have a slight non-linear pattern. Even though the deviation is not significantly large, this still indicates that there is some non-linearity between the predictors and the response. As the assumptions were verified, the data was fitted to a multiple regression model shown in Figure 4.

Figure 4. Response vs Fitted scatterplot for the multiple linear regression model



Although there was no clear evidence of non-linearity, square root transformation was used to see if there is improvement after the transformation and the improvement in the normality of the residuals was obtained. With the transformation, the more stable linearity and normality were obtained as shown in Figure 5 compared to Figure 4.

Figure 5. Response vs Fitted scatterplot after the square-root transformation



In order to check if multicollinearity exists, the variance inflation factor was checked. As a common cutoff for a large VIF is 5, multicollinearity was not an issue for the predictors used. Finally, to investigate which predictors should be included in the model, the stepwise variable selection was used. Using both the Akaike's Information Criterion and the Bayesian Information Criterion, the predictors that should remain are reduced to average class size, percentage of students with Disabilities and percentage of economically disadvantaged students.

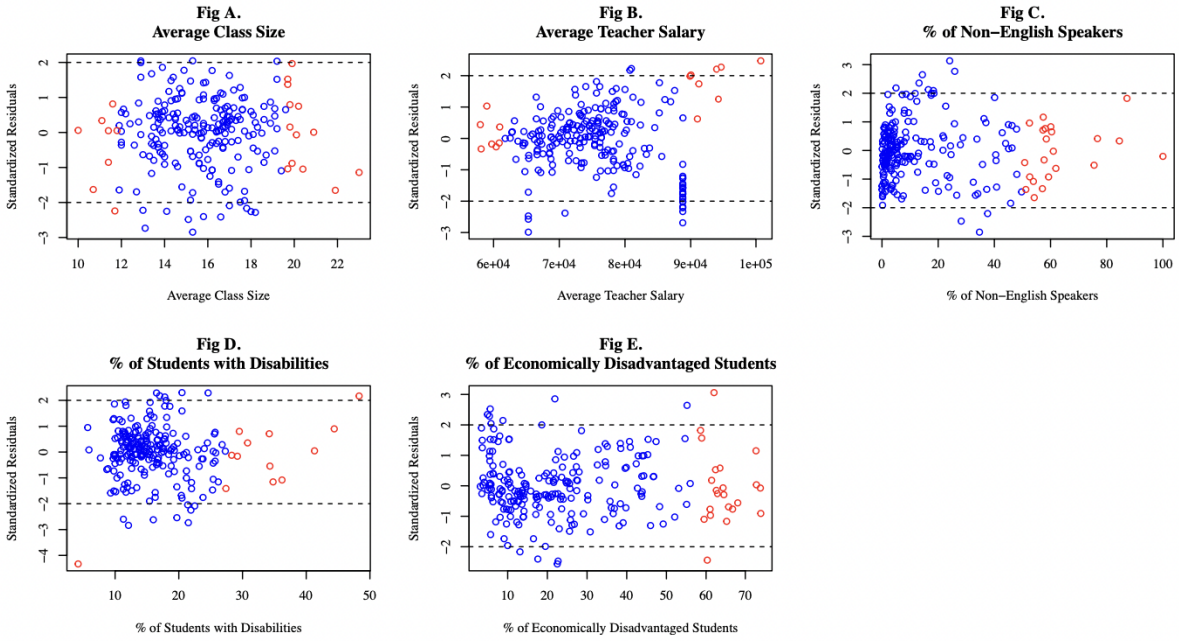
Discussion

The results found show that there is no strong linear relationship between the predictors and the response variable in a simple linear regression model. The only predictor which seemed to have a linear relationship was the percentage of economically disadvantaged students as mentioned above. However, the multiple regression model showed that there are other predictors with a linear relationship with the response variable. This showed that with the multiple regression model, more aspects of the student's learning environment are considered when looking for a relationship with the SAT scores. As students are not exposed to a single controlled environment during their academic career, the multiple regression model fits this narrative better. Therefore, the best model to predict a student's SAT score would be the multiple regression model with average class size, percentage of students with disabilities and percentage of economically disadvantaged students.

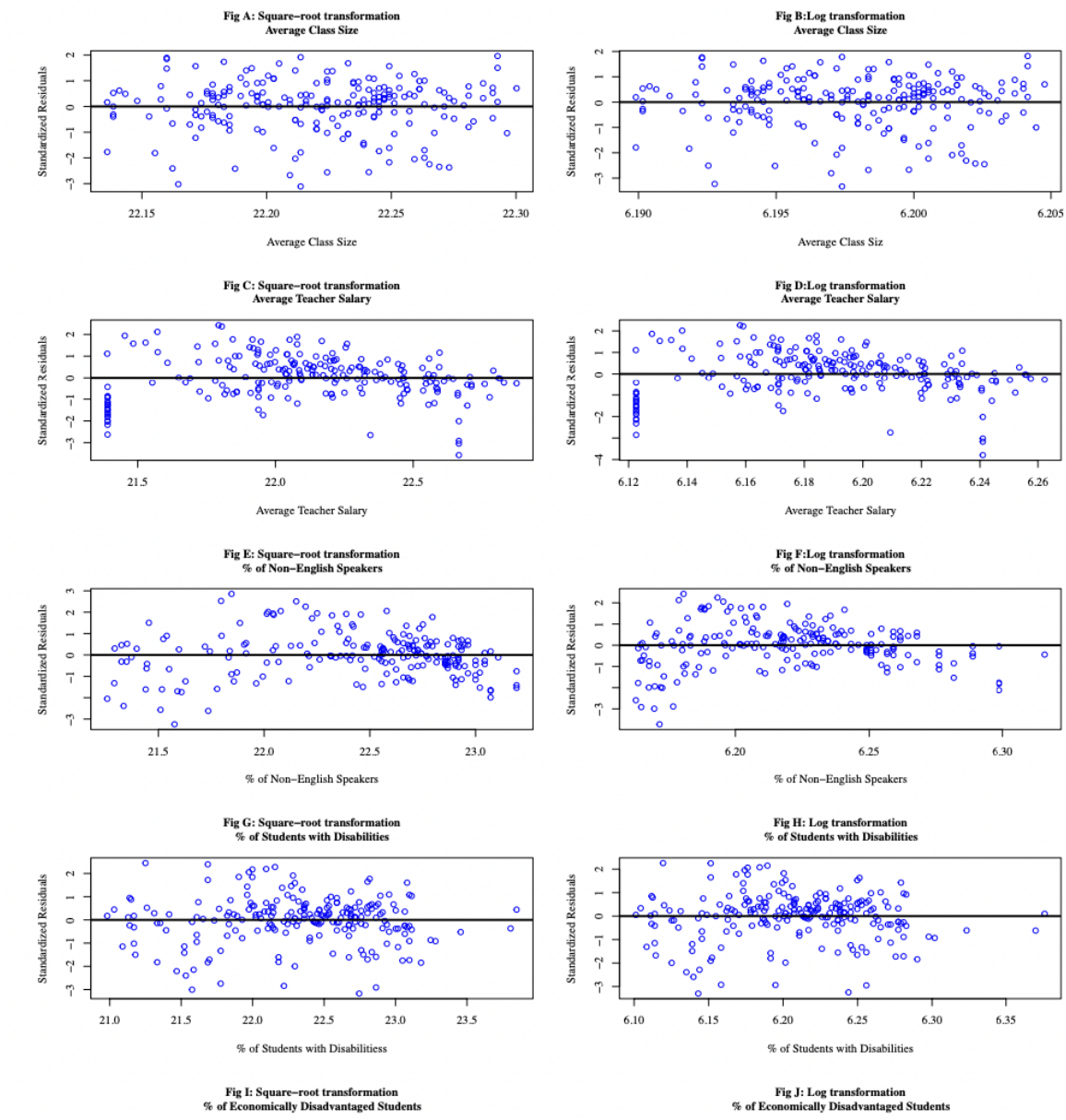
Understandably, there may be other variables not included in the dataset that can possibly be correlated to the current predictors such as number of extra-curricular activities and the number of different subjects being taught. In further research, more predictors regarding the learning environment should be included to determine if there are other variables which affect the SAT result that a student will receive.

Appendices

Appendix 1. Influential points based on leverage points



Appendix 2: Square root and log transformations



Appendix 3: Summary from the simple linear regression model

Term	estimate	std. error	statistic	p. value
(Intercept)	568.894422	3.6238197	156.98751	0
X.. Economically.Disadvantaged	-2.912148	0.1446092	-20.13806	0

References

1. Garcia, S. M., & Tor, A. (2009). The N-Effect: More Competitors, Less Competition. Psychological Science, 20(7), 871–877. https://doi.org/10.1111/j.1467-9280.2009.02385.x

2. Smith, K. B., & Eccles, J. (1998). Buying a Better SAT Score: A Renewed Search for the Elusive Link between Education Expenditures and Outcomes. State & Local Government Review, 30(1), 42–51. https://doi.org/10.1177/0160323X9803000104

3. Zwick, R., & Greif Green, J. (2007). New Perspectives on the Correlation of SAT Scores, High School Grades, and Socioeconomic Factors. Journal of Educational Measurement, 44(1), 23– 45. https://doi.org/10.1111/j.1745-3984.2007.00025.x

4. Dalziel, Nigel. (2017). Massachusetts Public School Data (Version 1) [Data file]. Retrieved from https://www.kaggle.com/datasets/ndalziel/massachusetts-public-schools-data