The Relationship Between Percent of Students Tested and SAT Scores

Zoë Schopick

April 2023

1 Introduction

The SAT is an exam used by colleges and universities to make admissions decisions for future students, but the objectivity and purpose of this test has been questioned. The test was designed to measure a high school student's readiness for college and to help colleges easily compare all applicants [9]. Some schools, however, have been moving away from considering a student's SAT scores when reviewing their application. There have been claims that the SAT is not an objective test of intelligence and scores are unfairly affected by variables outside of students' controls. Schools moving away from the SAT say that high school grade point average or a written essay is a better indication of college success and dropping these test requirements have led to more diverse incoming classes [10]. There has been research that shows a positive correlation between household income and SAT scores which has led many critics to doubt the objectivity of the test and thus its usefulness of predicting college success [12].

The SAT[1] has 154 multiple choice questions split into three sections, math, writing, and verbal. Each section is scored between 200 and 800 points with a 2400 being a perfect SAT score[2]. The SAT is a relatively recent invention. In the early 1900s each college had its own individual admissions exam. In 1905 the first IQ test was invented and given to soldiers during World War I. That IQ test was developed into the SAT by College Board and, in 1926, the first SAT was administered [1]. The test has adapted and changed in various ways to become the modern test that over 1.5 million students take each year [8].

Multiple factors, such as race, socioeconomic status, extracurricular activities, and family background, have been studied in their impact on SAT scores

[1] SAT originally stood for the Scholastic Aptitude Test and then Scholastic Assessment Test. SAT no longer stands for anything and is just the actual name of the test [3]

[2] This is the scoring for the SAT before March of 2016. The modern SAT has two sections scored out of 800 each for a total score of 1600. Since the data in this analysis is from before 2016, only the previous method of scoring will be used [9].

[6]. There has been no research done as to whether the percentage of the high school that takes the SAT impacts the average SAT score for that school. High schools have a variety of requirements for the SAT. Some schools require every student to take the test and the schools pay for it. Some schools only offer the test on a Saturday and the cost is left up to the students. As the SAT costs $55 per student per test, this cost can be a barrier to some students taking the test [9]. This research project seeks to answer the question: can the number of test takers at a school predict the average SAT score?

2 Materials and Methods

The data set used for this project was created from two data sets merged together. One data set had high school information and SAT scores and the other data set had census information for each California county. The high school and SAT scores data set was from the California Department of Education and College Board. The data set was downloaded from DataWorld. The SAT report contains data from all California public schools with at least one test taker or schools with zero test takers, but at least eleven students enrolled in 11th or 12th grade. Enrollment information is based on data submitted by local educational agencies. This data includes county name, district name, school name, number of grade twelve students, number of students tested, percent of students tested, as well as average SAT scores for the total SAT, the math portion of the SAT, the writing portion of the SAT, and the verbal portion of the SAT [2]. To account for factors that influenced SAT scores discussed in other research such as household income and race [12], the researcher created another data set with information from the US census. The US census data set was created from compiling data found on the US census bureau website. This data was found per California county. Each school within the county has the same US census data. The US census information in the data set is median household income, percent of the county that is white, percent of households with computers, average number of persons per household, percent of the county that are high school graduates (over the age of 25), and percent of households that speak a language other than English [11].

Before the data was analyzed it was cleaned. Rows with missing data were removed as well as rows with schools that had a percentage of test takers over 100% (as this is not a possible value). District name, county name, school name, number of grade 12 students, and number of students tested were removed from the data set before analysis.

A multiple linear regression was performed on the data four times to create four different models. Models were created for the responses of total mean SAT scores, mean math SAT scores, mean writing SAT scores, and mean verbal SAT scores. All other variables (percent tested, median income, percent white, percent with computers, average persons per household, percent high school graduates, and percent speaking a language other than English) were used as predictors. A multiple linear regression was chosen because the goal was to be able to predict the mean SAT scores based on multiple quantitative, independent factors.

To create the model, the cleaned data was split into training and test data sets with a 70:30 ratio. Linear regression models were created for the responses of total mean SAT score, mean math SAT score, mean writing SAT score, and mean verbal SAT score. All models were created using the training data set. The R^2 values were found for each linear model. Predictions were made for each linear model using the test data set. The root mean squared error (RMSE) values were then found for each model to find the accuracy of the model.

3 Results

In creating multiple linear regression models to predict mean total SAT scores, mean math SAT scores, mean verbal SAT scores, and mean writing SAT scores the same predictor variables were found to be significant for each of them. Those predictor variables are: percent of students tested, percent of the county that is white, median household income, and average persons per household. Percent of households with computers, percent of high school graduates over 25, and percent of households that speak a language other than English had no significance for any of the four response variables. The models created included all predictor variables. Other models were tested which included only the significant predictors, but those models did not perform as well.

Based on the root mean squared error (RMSE) values for each test (Table 1), the linear model for predicting mean math SAT scores performed the best. The RMSE values for mean math SAT scores, mean writing SAT scores, and mean verbal SAT scores were all relatively close. Based on the R^2 values (Table 2), the linear model predicting total SAT scores performed the best.

RMSE Values			
Mean Total SAT Scores	Mean Math SAT Scores	Mean Writing SAT Scores	Mean Verbal SAT Scores
15.49	4.97	5.50	5.04

Table 1: RMSE Values for Each Linear Model

R^2 Values			
Mean Total SAT Scores	Mean Math SAT Scores	Mean Writing SAT Scores	Mean Verbal SAT Scores
0.2744	0.2739	0.2675	0.2654

Table 2: R^2 Values for Each Linear Model

4 Discussion

Four predictor variables were found to have significant impacts on the mean SAT scores for the data provided. Those four variables were: percent of students tested, percent of the county that is white, median household income, and average persons per household. This means the new information being studied here, the percent of students tested, did have a significant impact on mean SAT scores. There was a positive correlation between percent of students tested and mean SAT scores meaning as the percentage of students tested increased the mean SAT scores increased as well. This factor has not been investigated in previous research. There are many factors that could lead to this correlation. Some schools may push students to take the SAT more than others. Some schools may teach SAT test-taking skills in the classroom. Some schools pay for their students to take the SAT which could lead to more students in those schools taking the test or those students having additional money to spend on study materials [4]. The reason for the correlation between percent of students taking the SAT and mean SAT scores should be investigated further in a future study.

The current study aligned with information found in previous studies on factors influencing SAT scores. The significance of median household income as it correlates with mean test score is in line with previous research which finds a positive correlation between median household income and SAT scores [12]. The percent of the county that is white as it correlates with average SAT score is also in line with the majority of previous studies. There is a positive correlation between the percent of the county that is white and mean SAT scores [5]. There has been no research on the correlation between average persons per household and SAT scores. This study found a negative correlation between this variable and average SAT score suggesting that as the average number of persons in a household decreases, the average SAT scores increase. This variable was not the subject of the study, but would be an interesting future project as it has not been studied.

The model predicting mean math SAT scores performed the best on the test data set with a root mean squared error (RMSE) of 4.97. The models to predict mean verbal and mean writing SAT scores had very similar RMSE values (5.50 and 5.04 respectively). These RMSE values are very low and suggest well-fitting models. They say that on average, the model predicts a score that is only around five points off from the actual score. The individual tests have total scores of 800 each. In that context, a five point error is very low. The model for mean total SAT score is 15.49 which is also low because the total SAT score is out of 2400 points. The RMSE value for total mean SAT score was expected to be higher than those of the individual subjects because the total number of possible points is larger. Dividing the total RMSE score by three (to account for the three sections) we get 5.163 which is very close to the RMSE scores of the math, writing, and verbal tests. This means all of the models created were fairly accurate in being able to predict SAT scores based on the variables provided.

Looking at the R^2 values for the linear models, the model for mean total SAT scores did the best at explaining the variation in the response variable with a value of 0.2744. The other models had slightly lower, but very similar R^2 values (0.2739 for math, 0.2675 for writing, and 0.2654 for verbal). These R^2 values seem low. For the total mean SAT scores, the predictors in the linear model only explain about 27% of the variation seen in the response variable. In the context of the analysis, the low R^2 value makes sense. The models only take into account variables that are outside of each student's control. There are no variables that account for study time, grade point average, or overall intelligence. Even though research has shown SAT scores are impacted by variables outside of students' controls, the test still does measure intelligence and academic ability to some level [7]. The test was designed to purely measure the intelligence and college readiness of a student [1] so 27% is a relatively large number when we are only accounting for outside factors.

This study had a few limitations that could be improved upon in future research. The current study was limited to only California students in the 2012-2013 school year. The SAT has drastically changed since then [1], so it is unclear if results of the same study on the modern SAT would be the same. The individual predictors were also assumed to be independent of one another, but there are probably some small dependencies between them. There are links between race and income that were not accounted for here [1]. Another limitation was that United States Census data was only found per county, while the SAT data was found per school. The median household income, racial makeup, and various other statistics can be vastly different for different schools within the same county. The data was not available for each individual school. If this data becomes available, the same study could be done with more individualized data for each school. This study was only done with data from California, but having data from other states, and comparing them would also be a great way to extend this research further.

5 Conclusion

The current study found that accurate models could be made to predict average SAT scores based on multiple predictors. Some of the significant predictors, median household income and percent of the county that is white, had been studied before [12]. Two significant predictors, the percent of students tested and the average number of persons per household, had not been studied. Four models were created that were able to predict mean SAT scores within just a few points of the actual scores and were able to account for over 26% of the variability of the data. These two new factors that have not yet been studied help to provide more of an insight into the SAT and its purpose. Many colleges and universities say that the SAT helps to determine college preparedness and intelligence, but with so many factors outside of a student's control significantly impacting scores, these assumptions may have to be reexamined.

References

[1] Bhutta N., Chang, A., Dettling, L., & Hsu, J. (2020, September 28). *Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances.* Board of Governors of the Federal Reserve System. Retrieved April 26, 2023, from https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.html

[1] Cairns, H. (2022, December 22). *History of the SAT* CollegeRaptor. Retrieved April 25, 2023, from https://www.collegeraptor.com/getting-in/articles/act-sat/history-of-the-sat/

[2] California Department of Education & Data.World. (2017, March 14). California SAT Report 2012-2013. California. https://data.world/education/california-sat-report-2012-2013

[3] Cheney, M. (2023) *What Does "SAT" Stand For?* History and Social Justice. Retrieved April 25, 2023, from https://justice.tougaloo.edu/standardized-testing/what-does-sat-stand-for/

[4] College Raptor Staff. (2022, December 22). *States That Provide A Free ACT/ SAT Test.* CollegeRaptor. Retrieved April 26, 2023, from https://www.collegeraptor.com/getting-in/articles/act-sat/states-act-sat-given-free/

[5] Ford, J., & Triplett, N. (2019, August 28). *E(race)ing Inequities — The influence of race on SAT scores.* EdNC. Retrieved April 24, 2023, from https://www.ednc.org/eraceing-inequities-the-influence-of-race-on-sat-scores/

[6] Howard T. Everson & Roger E. Millsap (2004). *Beyond Individual Differences: Exploring School Effects on SAT Scores.* Educational Psychologist, 39:3, 157-172, DOI: 10.1207/s15326985ep3903_2

[7] Lindsay, S. (2019, August 4). *What Do SAT Scores Measure? IQ? Income?* Prep Scholar. Retrieved April 26, 2023, from https://blog.prepscholar.com/what-do-sat-scores-measure-iq-income

[8] Nam, J. (2023). *Average SAT Score: Full Statistics.* Best Colleges. Retrieved April 25, 2023, from https:www.bestcolleges.com/research/average-sat-score-full-statistics/#text=Note%20Reference-,More%20than%201.7%20million%20high%20school%20students%20took%20the%20SAT,of%20them%20juniors%20and%20seniors.&text=By%20comparison%2C%201.3%20million%20students,)%2C%20in%20the%20same%20year.

[9] The Princeton Review. (2023). *What is the SAT?* The Princeton Review. Retrieved April 18, 2023, from https://www.princetonreview.com/college/sat-information

[10] Rooney, C.J., & Schaeffer, B. (1998).*Test Scores Do Not Equal Merit: Enhancing Equity & Excellence in College Admissions by Deemphasizing SAT and ACT Results.* ERIC. https://eric.ed.gov/?id=ED426107

[11] US Census Bureau. (2010). *Quick Facts.* US Census Bureau. Retrieved March 10, 2023, from https://www.census.gov/quickfacts/fact/table/yubacountycalifornia/INC110221

[12] Wade, L. (2012, August 29). *The Correlation Between Income and SAT Scores.* The Society Pages. Retrieved April 24, 2023,

from https://thesocietypages.org/socimages/2012/08/29/the-correlation-between-income-and-sat-scores/