SAT SCORE PREDICTOR

Anita Ahmed, github handle: ama908, NYU ID - N11949714

Abstract:

This study examined the relationship between Attendance Rate, Pupil Teacher Ratio and overall SAT performance among college-bound seniors in 2012 in the 32 school districts of New York City. Correlation between SAT performances is examined with individual variables, and multiple regression was used to examine if significant moderator effects existed between variables. The Results indicated that multiple regression model was the strongest predictor of SAT Score with the R-Squared value of 0.553, compared to the separate main effects of Attendance Rate, and Pupil Teacher Ratio. In addition, Correlation between Student performance in each district for Math and English schoolwide test in year 2006 and district wide average of 2012 SAT Math and English score (Reading and writing) respectively. Both showed a strong relationship with R squared values of 0.639 and 0.526 respectively.

Introduction:

SAT scores is the most accepted standardized test by colleges to measure the college readiness of students in the United States. It is a strong determinant of student's acceptance in good colleges and scholarship amount. As it is known, admission good colleges and getting scholarship both contributes to student's overall success and therefore performing well in SAT is very important for student. Therefore, it is an interesting problem to determine what factor can affect the students SAT performance.

Researchers in the field of education have been constantly studying what are some meaningful factors that can affect Standardized test performance like SAT. St. Rose (2009) Gender, Race/Ethnicity, Socioeconomic status and SAT performance in 2004 using regression analysis and found Race as the strongest predictor of SAT performance compared to other variable. [1] Dawkin (2010) studied ACT scores of 250 students in Marion County, Tennessee against students high school GPA, socioeconomic status and disability status and conducted a multivariate regression. [2] Dawkin found strong correlation exist between all these 3 variables and ACT Score.

This study examined the relationship between Attendance Rate, Pupil Teacher Ratio, 2006 Grade 6 Mathematics and English scores with and SAT performance among college-bound seniors in 2012 in the 32 school districts of New York City. Single variable and multivariable regression analysis was conducted to find correlation between SAT scores and stated variable.

Data:

The following data were available and suitable to answer the research question:

1. 'SAT Results' data set contained mean SAT subject scores for 2012 college-bound seniors at individual school level. SAT scores range from 200-800, and the subjects tested are Mathematics, Reading and Writing. Data set had missing data for few schools; those schools were removed from the analysis. The data was then aggregated at a school district level. Mean SAT scored for Reading, Writing and Math Scores were added to get the total SAT score for all subject, the range of score here was 600-2400. Mean SAT scored for Reading, Writing and Math Scores were added to get the total SAT score for English; the range of score here was 400-1600. NYC has 32 regular school district and 2 special school district. For this study data for special school districts were dropped. [3]
2. 'School Attendance and Enrollment' dataset contained average attendance rate & enrollment of schools in 32 regular school district and 2 special school district for year 2010 -2011. The average attendance rates for only 32 regular school district were used. [4]
3. Pupil teacher ratio was retrieved from '2010-2011 Class Size - School-level detail Average class sizes for each school, by grade and program type' dataset for the 32 regular school district. This data was collected in January 28, 2011. To perform analysis Pupil teacher ratio was aggregated at a district level. [5]
4. 2006 Math results were taken from 'NYS Math Test Results by Grade for year 2006-2011'. The dataset contains math test results between years 2006-2011 for grades 3 -12 of 32 school districts. Data was sorted for grade 6 and year 2006. [6]
5. 2006 English Language Arts (ELA) results were taken from 'NYS ELA Test Results by Grade for year 2006-2011'. The dataset contains ELA test results between years 2006-2011 for grades 3 -12 for 32 school districts. Data was sorted for grade 6 and year 2006. [7]

Methodology:

In order to find correlation between Attendance Rate, Pupil-Teacher Ratio, 2006 Grade 6 Mathematics and English scores with SAT performance among college-bound seniors in 2012, 5 hypothesis test were set up and the 1st and 2nd Degree OLS Regressions were performed to verify correlation between variables.

Test 1:

Problem Statement: School with higher attendance rate SHOULD have a higher SAT Score

NULL HYPOTHESIS: In NYC School District (1-32), SAT Scores of Students has NO Correlation to Attendance % of Student.

ALTERNATE HYPOTHESIS: In NYC School District (1-32), SAT Scores of Students has Correlation to Attendance % of Student.

After conducting 1st degree linear regression R-squared value was 0.520, therefore the model captured 52% of the actual data. X- Coefficient was 32.95 for a p-value of 0.00. Therefore attendance rate is an influential factor for 1% increase in attendance rate the SAT score increases by 32 units. The 2nd degree linear regression was conducted and R-squared value

value was 0.531, the model captures 53.1% of the actual data. Since the R-squared value was very similar for 1st Degree and 2nd Degree the models were plotted on a scatter plotted visualize which model is a better fit.[Fig 1]. Better model could not be detected visually so a likelihood ratio test was conducted. For the likelihood ratio test, null hypothesis was that the linear model (Model 1) is a better fit than the curve model (Model 2). The likelihood ratio was 0.77 which is lower than the chi square value (3.84 for a significance level of 0.05) therefore the null hypothesis was accepted that Linear model is better fit.

Test 2:

Problem Statement: School with Lower Pupil to Teacher Ratio SHOULD have a Higher SAT Score, since it indicates teachers can give more attention to individual student.

NULL HYPOTHESIS: In NYC School District (1-32), SAT Scores of Students has NO Correlation to Pupil -Teacher Ratio.

ALTERNATE HYPOTHESIS: SAT In NYC School District (1-32), SAT Scores of Students has NEGATIVE Correlation to Pupil - Teacher Ratio.

After conducting 1st degree linear regression R-squared value is 0.349, therefore the model captured 34.9% of the actual data. X- Coefficient was 48.5 for a p-value of 0.00. Therefore this variable is an influential factor, and every unit increases the SAT score by 48 units. The 2nd degree linear regression was conducted and R-squared value was 0.392, the model captures 39.2% of the actual data. Since the R-squared value was very similar for 1st Degree and 2nd Degree the models were plotted on a scatter plotted visualize which model is a better fit.[Fig 2] Better model could not be detected visually so a likelihood ratio test was conducted. For the likelihood ratio test, null hypothesis was that the linear model (Model 3) is a better fit than the curve model (Model 4). The likelihood ratio was 2.17 which is lower than the chi square value (3.84 for a significance level of 0.05) therefore the null hypothesis was accepted that Linear model is better fit.

Test 3:

Problem Statement: **MULTIVARIATE ANALYSIS to predict Attendance Rate and Pupil to Teacher Ratio together affect SAT Score**

NULL HYPOTHESIS: In NYC School District (1-32), SAT Scores of Students has NO Correlation to Attendance Rate and Pupil - Teacher Ratio

ALTERNATE HYPOTHESIS: In NYC School District (1-32), SAT Scores of Students has Correlation to Attendance Rate and Pupil - Teacher Ratio

After conducting 1st degree linear regression R-squared value is 0.553, therefore the model captured 55.3% of the actual data. X- Coefficient for attendance was 23.32 for a p-value of 0.001. Therefore attendance rate is an influential factor for 1% increase in attendance rate the SAT score increase by 23 units. X- Coefficient for pupil-teacher ratio was 19.18 for a p-value of 0.150. Therefore pupil to teacher ratio is a non-influential factor to determine SAT Score.

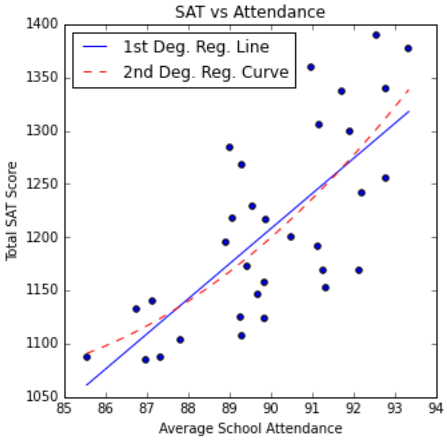

Figure 1 Test 1 Output

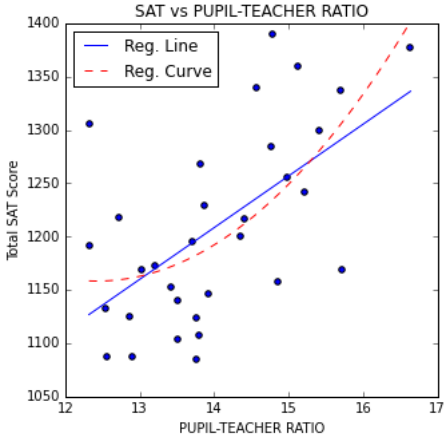

Figure 2 Test 2 Output

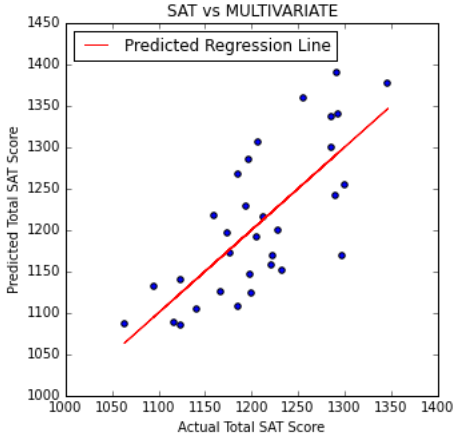

Figure 3 Test 3 Output

Test 4:

Problem Statement: School district with students who consistently do well in math even at younger age is more likely to do well in SAT Math Test. Here 2006 Grade 6 students Math score is the variable, it is assumed those students took SAT in 2012.

NULL HYPOTHESIS: In NYC School District (1-32), Avg. 2012 SAT Math Scores has NO Correlation to Avg. **Math Score in Grade 6 in 2006*.***

ALTERNATE HYPOTHESIS: In NYC School District (1-32), Avg. 2012 SAT Math Scores has Correlation to Avg. **Math Score in Grade 6 in 2006*.***

After conducting 1st degree linear regression R-squared value is 0.639, therefore the model captured 63.9% of the actual data. X- Coefficient was 2.13 for a p-value of 0.00. Therefore this variable is an influential factor, and every unit increases the SAT score by 2 units. The 2nd degree linear regression R-squared value was same as 1st degree R-squared value, and fitted the data points same way. [Fig. 4] Since both R-squared value were same, likelihood ratio test was conducted. For the likelihood

ratio test, null hypothesis was that the linear model (Model 1) is a better fit than the curve model (Model 2). The likelihood ratio was 0.00028 which is lower than the chi square value (3.84 for a significance level of 0.05) ~~therefore the null hypothesis was accepted that Linear model is better fit.~~ REJECTED!

Test 5:

Problem Statement: School district with students who consistently do well in English Language Arts(ELA) even at younger age is more likely to do well in SAT Reading and writing Test. Here 2006 Grade 6 students ELA score is the variable, it is assumed those students took SAT in 2012.

***NULL HYPOTHESIS: In NYC School District (1-32) , Avg. SAT Reading and writing Scores has NO Correlation to Avg* ELA Score in Grade 6 in 2006.**

***ALTERNATE HYPOTHESIS: In NYC School District (1-32) , Avg. SAT Reading and writing Scores has Correlation to Avg* ELA Score in Grade 6 in 2006.**

After conducting 1st degree linear regression R-squared value is 0.526, therefore the model captured 52.6% of the actual data. X- Coefficient was 3.13 for a p-value of 0.00. Therefore this variable is an influential factor, and every unit increases the SAT score by 3 units. The 2nd degree linear regression R-squared value was same as 1st degree R-squared value, and fitted the data points same way. [Fig. 4] Since both R-squared value were same, likelihood ratio test was conducted. For the likelihood ratio test, null hypothesis was that the linear model (Model 1) is a better fit than the curve model (Model 2). The likelihood ratio was 0.025 which is lower than the chi square value (3.84 for a significance level of 0.05) therefore the null hypothesis was accepted that Linear model is better fit.

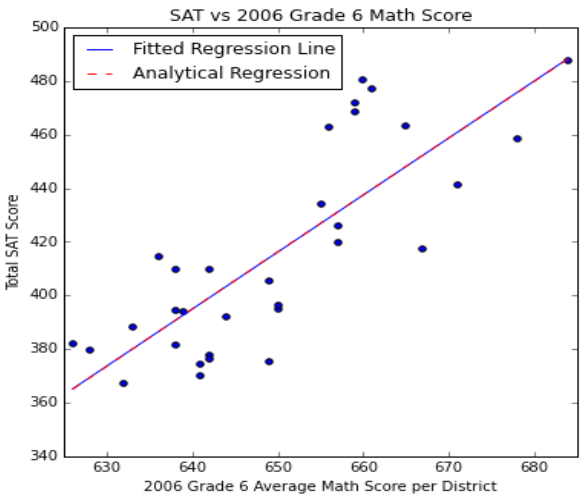

Figure 4 Test 4 Output

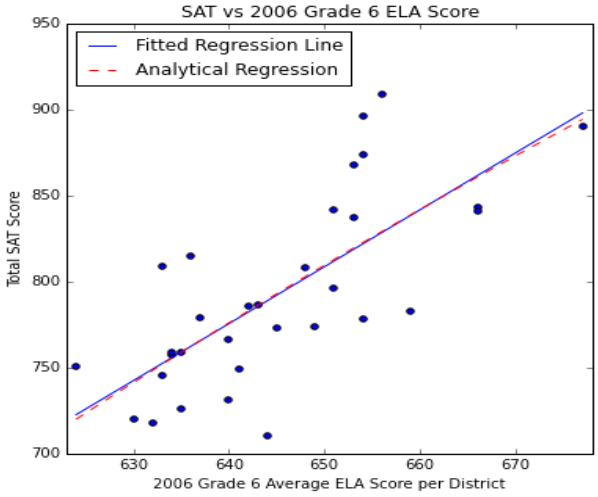

Figure 5 Test 5 Output

Conclusions:

Comparing the R-squared values for single variable models with multivariable model, for attendance rate and pupil-teacher ratio it is evident that the Multivariate model has the highest R-value of 0.553 and hence is the predictor of SAT scores. For the single variable model it is observe Class attendance has a stronger influence on SAT Score than pupil-teacher ratio. Between test 4 and 5, Grade 6 English scores seemed to be more influential when predicting SAT scores. Therefore it can be said that doing good in English at early ages increases chance to do well in SAT when compared to Mathematics. Also unlike expected, lower pupil to teacher ratio has a positive impact on SAT score, which can mean self-study impacts SAT more than what is taught in class. The only short coming of this test was missing SAT scores for some school. Also there were missing information for school district 75 and 79 which is why the model doesn't represent all of NYC schools.

Future work:

For future analysis, students high school grades will be compared to SAT score and then another multivariate model can be built with Attendance, Pupil-Teacher Ratio and Highschool Test Results. To be more thorough the missing data for schools could be gathered and the analysis would be re-run using a complete data set.

there is an important covariance that you are ignoring here: probably the english grade and the math grade are correlated.

Annotated Bibliography:

1. An examination of the relationship between gender, race/ethnicity, socioeconomic status, and SAT performance by St. Rose, Andresse This paper examined the relationship between gender, race/ethnicity, socioeconomic status, and SAT math and verbal performance among college-bound seniors in 2004. Using multiple regressions this study looks at all three variables simultaneously to identify possible interrelationships between them.
   Link: http://pqdtopen.proquest.com/doc/304645025.html?FMT=ABS
2. An Examination of the Relationship among Grade Point Average, Socioeconomic Status, Disability Status and ACT Performance by Tennessee Technological University. This study was to examine the relationship between students' Grade Point Average, socioeconomic status, and disability status and ACT composite scores of 250 students enrolled in public high schools located in Marion County, Tennessee. Data was analyzed and interpreted through a multiple correlation.
   Link: https://iweb.tntech.edu/jcbaker/Sample%20paper%201%20-%20FOED%206920.pdf
3. SAT Results-The most recent school level results for New York City on the SAT. Results are available at the school level for the graduating seniors of 2012. Records contain 2012 College-bound seniors mean SAT scores taken during SY 2012. Link: https://data.cityofnewyork.us/Education/SAT-Results/f9bf-2cp4
4. School Attendance and Enrollment By District - 2010-11 - 2010- 11 Attendance & Enrollment (Unaudited) by District as of December 31, 2010. Charter Schools, Community-Based Organizations (Pre-K), Home Instruction, and Hospital Schools are not included.
   Link: https://data.cityofnewyork.us/Education/School-Attendance-And-Enrollment-By- District2010/rfpq-hs49
5. 2010-2011 Class Size - School-level detail Average class sizes for each school, by grade and program type (General Education, Self-Contained Special Education, Collaborative Team Teaching (CTT)) for grades K-9 (where grade 9 is not reported by subject area), and for grades 5-9 (where available) and 9-12, aggregated by program type (General Education, CTT, and Self-Contained Special Education) and core course (e.g. English 9, Integrated Algebra, US History, etc.).
   Link : https://data.cityofnewyork.us/Education/2010-2011-Class-Size-School-level-detail/urz7pzb3
6. NYS Math Test Results By Grade 2006-2011 - District - All Students New York City Results on the New York State Mathematics Tests, Grades 3 – 8.
   Link: https://data.cityofnewyork.us/Education/NYS-Math-Test-Results-By-Grade-2006-2011District-/gyaz-82xj
7. English Language Arts (ELA) Test Results 2006-2012 - District - All Students Latest available data and trends in the state assessment results of English Language Arts for grades 3 through 8.
   Link: https://data.cityofnewyork.us/Education/English-Language-Arts-ELA-Test-Results-20062012-D/yhfh-vyns