



Authors	Jennifer Bar (barrij12@students.ecu.edu) 800 Heath St. Apt. 64, Greenville, NC 27858 Phone: (252)320-2525
Title	Analysis of 2011 Average SAT Data for the United States: Can a state's average SAT score be predicted using certain data?
Institution	Department of Mathematics, East Carolina University
Instructor Sponsor	Said Said (said@ecu.edu)

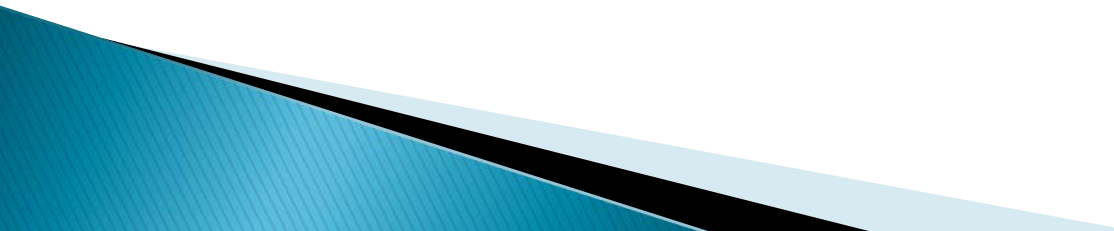
Analysis of 2011 Average SAT Data for the United States

Can a state's average SAT score be predicted using certain data?

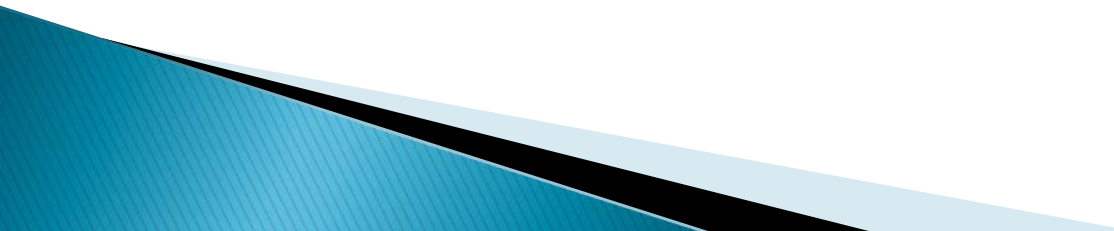
Introduction

- ▶ The purpose of this analysis is to see if it is possible to predict the average SAT score of a state based on a series of economic and educational factors:
 - ❖ Pupil/Teacher Ratio
 - ❖ Expenditures Per Student (in Thousands)
 - ❖ Grade 8 Math Scale Scores for 2011
 - ❖ Grade 8 Reading Scale Scores for 2011
 - ❖ SAT Participation Rate (as an integer)
- ▶ The data was further analyzed by looking at the region for which each state belongs:
 - ❖ Northeast, Midwest, South, and West

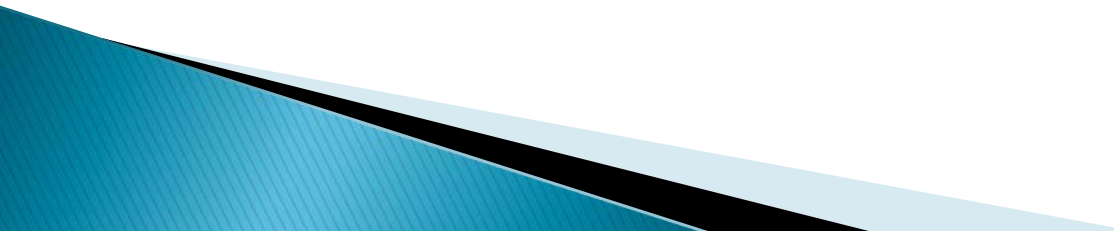
Background

- ▶ The SAT, or SAT Reasoning Test, as it was renamed in 2004, is a test which aims to assess a student's analytical and problem solving skills by forcing students to work through a series of verbal and mathematical questions. The test is scored in three parts, each worth a possible 800 points. Thus, the highest score achievable is 2400. It generally taken by high school students who wish to proceed to college after graduation.
- 

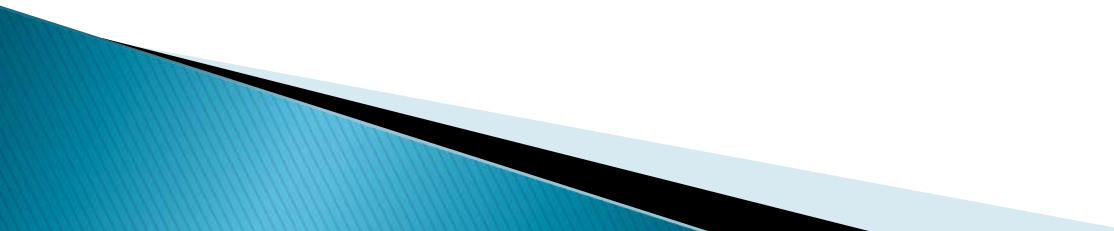
Methodology

- ▶ The data that was analyzed was gathered directly from CollegeBoard.org (the organization that offers the SAT), the National Center for Education Statistics, and the U.S. Census Bureau. After being collected, the data was analyzed in the software program Minitab. The data was selected based on availability and proposed relevance. All the data analyzed is from 2011.
- 

Methodology

- ▶ Most of the data was collected and used without modification. The exception was the Expenditures Per Pupil data. This data was calculated by taking the Total Expenditures for Education and dividing it by the Total Students for each state. This was done in order to accommodate the expenditure differences for more states with larger populations.
- 

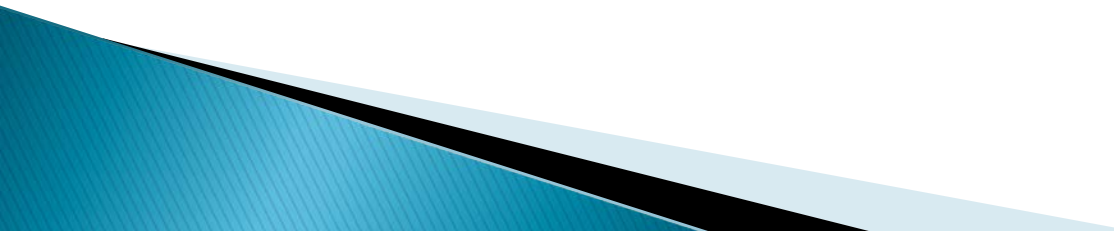
Methodology

- ▶ The determination to include data for the Grade 8 Scale Scores in Reading and Math was based on the fact that the SAT covers both of these areas. The Scale Scores come from tests administered to students in the eighth grade in each state. Including these scores allows for determining if there is a relationship between SAT scores and a state's public education system preparation.
- 

Methodology

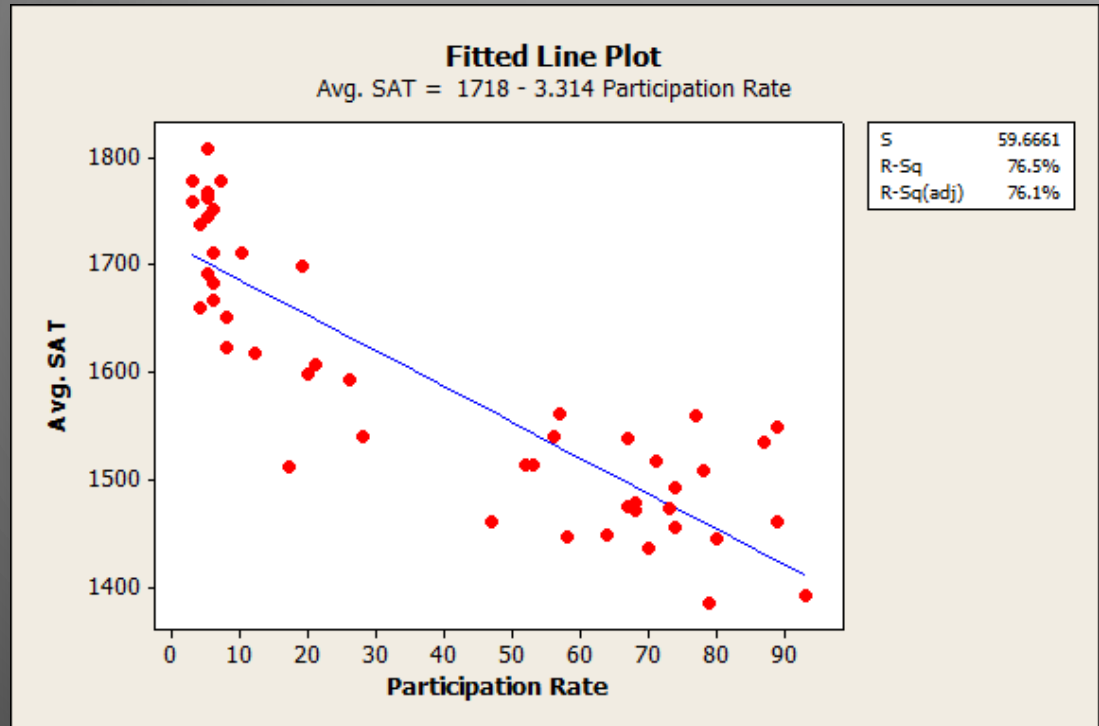
- ▶ The U.S. Census Bureau groups states according to four geographic regions: Northeast, Midwest, South, and West. To consider whether a state's geographic location effected SAT scores, three indicator variables were created. The data were given a value of 1, or 0, assigned as follows:
 - ❖ Z1 – 1 if located in the Northeast, 0 if otherwise;
 - ❖ Z2 – 1 if located in the Midwest, 0 if otherwise;
 - ❖ Z3 – 1 if located in the South, 0 if otherwise.
 - ❖ Any state located in the West was assigned a value of 0 for all three levels, thus accounting for all regions.

Process

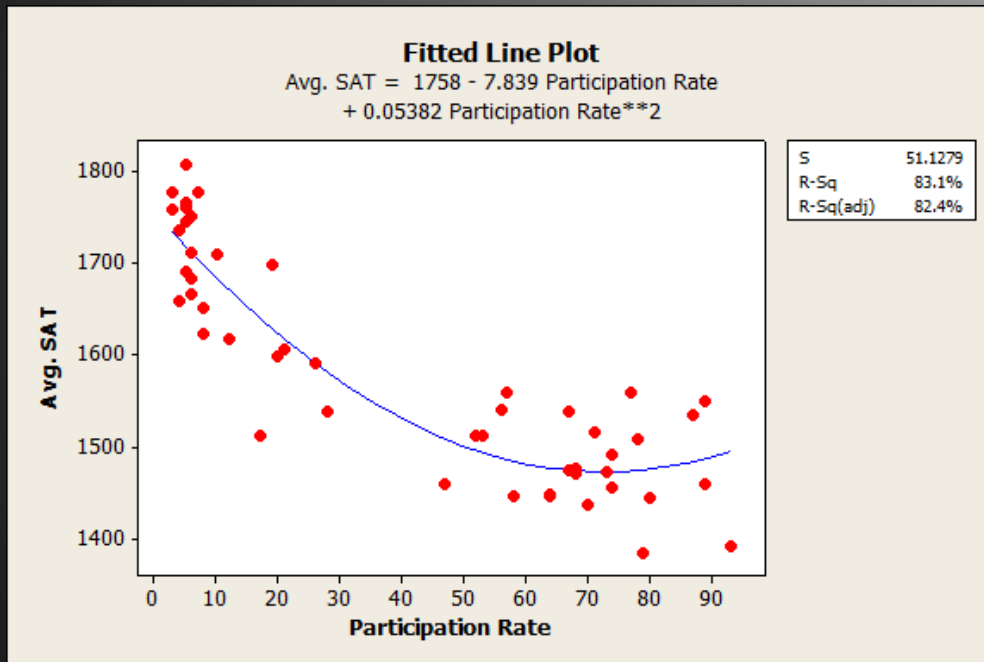
- ▶ The first step in analyzing the data was to look at the fitted line plot of each single predictor variable against the response variable. Then, a multiple regression model was considered, with each of the potential predictor variables included. A number of tools were utilized to consider the ways that the variables worked together or independently to predict the average SAT score, until a final model was determined.
- 

Finding #1

- ▶ The first important finding was that the single best predictor of the average SAT score was a state's Participation Rate. This variable alone explained 76.5% of the unadjusted variation in the scores.

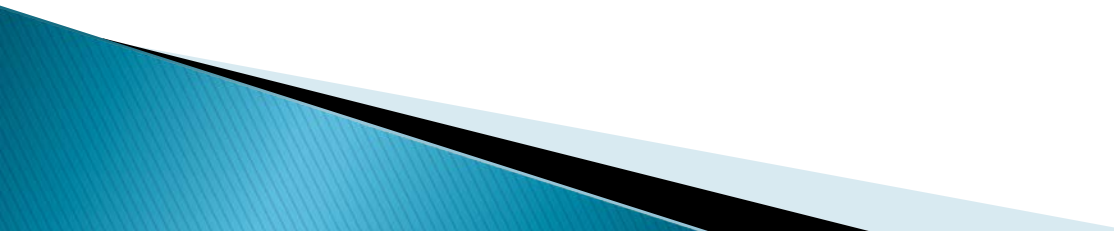


Finding #2



- ▶ The second important finding was that, when the graph of the Participation Rate was plotted against the Average SAT Score, the data took on a curved shape rather than a straight line. This indicated a potential quadratic relationship, and so a fitted quadratic model was examined. This model accounted for 83.1% of the unadjusted variation in the average SAT scores. Because of this, a quadratic value of this variable was added to the full model.

Finding #3

- ▶ The next finding was that there was no significant interaction between any of the variables. This was concluded based on analysis of models which included interaction terms for all the variables under consideration. The p-values for the interaction terms ranged from 0.368 to 0.989, which were all larger than 0.05.
- 

Finding #4

- ▶ Because there was no interaction, the data was examined to determine if any confounding was present. There were several variable relationships which showed potential confounding, so all the variables were included in the full model.

One example:

Avg. SAT versus Pupil/Teacher Ratio

The regression equation:

$$\text{Avg. SAT} = 1596 - 0.49 \text{ Pupil/Teacher Ratio}$$

Avg. SAT versus Pupil/Teacher Ratio, Expenditures Per Pupil

The regression equation:

$$\text{Avg. SAT} = 1970 - 11.6 \text{ Pupil/Teacher Ratio} - 16.2 \text{ Expenditures Per Pupil}$$

Finding #5

- ▶ At this point, a Stepwise Regression and a Best Subsets Regression were performed to determine which of the predictor variables should be included in the final model. All of the original five predictor variables were included, as well as a quadratic value for the Participation Rate, due to earlier findings. No interaction variables were included. The conclusion of this analysis was that the best predictors for the Average SAT Score of a particular state would be that state's Grade 8 Reading Scale Score, Participation Rate, and the squared value of the Participation Rate.

Stepwise Regression: Avg. SAT versus Pupil/Teache, Expenditures, ...

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.05

Response is Avg. SAT on 5 predictors, with N = 51

Step	1	2	3
Constant	1718.3	116.8	381.7
Participation Rate	-3.31	-3.47	-6.77
T-Value	-12.64	-17.31	-7.99
P-Value	0.000	0.000	0.000
Grade 8 Reading SScore		6.07	5.18
T-Value		6.09	5.76
P-Value		0.000	0.000
Participation Rate^2			395
T-Value			3.98
P-Value			0.000
S	59.7	45.3	39.6
R-Sq	76.54	86.77	90.11
R-Sq(adj)	76.06	86.22	89.47
Mallows Cp	65.3	18.3	4.4

Best Subsets Regression: Avg. SAT versus Pupil/Teacher

Response is Avg. SAT

Vars	R-Sq	R-Sq(adj)	Cp	S	o	l	e	e	e	2
1	76.5	76.1	66.3	59.666						X
1	64.2	63.5	125.8	73.670						X
2	86.8	86.2	18.9	45.268					X	X
2	84.5	83.9	29.9	49.000			X		X	
3	90.1	89.5	4.8	39.564				X	X	X
3	89.8	89.2	6.2	40.143			X		X	X
4	90.3	89.5	5.7	39.523		X		X	X	X
4	90.3	89.5	5.9	39.601			X	X	X	X
5	90.6	89.6	6.4	39.411	X	X		X	X	X
5	90.5	89.5	6.8	39.566		X	X	X	X	X
6	90.9	89.7	7.0	39.231	X	X	X	X	X	X

Finding #6

- ▶ It was now important to determine if the Average SAT Scores varied among the four geographic regions. To do this, an analysis of covariance was conducted. The analysis consisted of a hypothesis test to determine if the adjusted means for each location, accounting for each of the variables, were significantly different. The result of this test was that the averages did not change for each of the regions (P-value 0.07). Thus, the region that a state is located in does not affect Average SAT Scores, and the final model should not include the regional indicator variables.

Model A: Regression Analysis: Avg. SAT versus Z1, Z2, Z3

Avg. SAT = 1572 - 73.3 Z1 + 155 Z2 - 22.9 Z3

Predictor	Coef	SE Coef	T	P
Constant	1572.23	25.96	60.55	0.000
Z1	-73.34	40.59	-1.81	0.077
Z2	154.69	37.48	4.13	0.000
Z3	-22.88	34.49	-0.66	0.510

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	331665	110555	12.61	0.000
Residual Error	47	411898	8764		
Total	50	743563			

Model B: Regression Analysis: Avg. SAT versus Grade 8 Reading SScore, Participation...

$Y = 604 + 4.25 \text{ Grade 8 Reading SScore} - 604 \text{ Participation Rate} + 337 \text{ Participation Rate}^2 + 6.5 Z1 + 40.9 Z2 - 4.4 Z3$

Unadjusted Means:

Z1: 1498.7

Z2: 1727

Z3: 1549.1

Z4: 1572

Adjusted Means:

Z1: 1584.3079

Z2: 1618.7079

Z3: 1573.4079

Z4: 1577.8079

Hypothesis Test for Equivalent Adjusted Means:

Full Model B, from above

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0$$

H_1 : Not all β 's equal

$$\text{TS: } F = \frac{\frac{11016}{3}}{1422} = 2.5823$$

$$\text{RR: } F = F > F_{44}^3(0.05) = F > 2.82$$

$$\text{P-Value: } P(F > 2.5823) = 0.07$$

Conclusion: Do not reject H_0 .

In other words, at the 5% level of significance, there is no evidence that the adjusted means are significantly different from one another, and thus, we can conclude that a state's region will not affect its average SAT score.

Finding #7

- ▶ It was important at this point to examine the assumptions of the final model, to determine if any of the assumptions made during the process of regression were potentially violated. To this end, graphs of the deleted residuals were reviewed. The histogram appeared to be very slightly skewed to the right. Also, the Normal Probability Plot showed a very slight curvature at each end. These are indications that the assumption of normality may be violated for this data set. To address these issues, further examination of the data should be performed, which may include transformations to the data. Such processes are not within the scope of this class, however, and were not performed.

Finding #8

- ▶ There were several unusual observations in the data. They included the data for Delaware, which has a large leverage value, as well as Idaho, Louisiana, and Washington, which had large standardized residual values. The model that included these observations explained 90.1% of the variance, while the model with these observations removed explained 92.5% of the variance in Average SAT Scores. Because the difference was relatively small, and because each data point in the model represents real-world data, it was decided not to omit these data points from the final model.

Conclusions

- ▶ The final model for this data is:
 - $\text{Average SAT Score} = 382 + 5.18 (\text{Grade 8 Reading Scale Score}) - 677 (\text{Participation Rate}) + 395 (\text{Participation Rate Squared})$.
- ▶ The final model explains approximately 90.1% of the variance observed in the Average SAT Scores for a state.
- ▶ This model indicates that increase a state's Grade 8 Reading Scale Score will increase it's Average SAT Score, holding Participation Rate constant.
- ▶ This model also suggests that the more students that participate in the SAT, the lower a state's average SAT score will be, holding their Grade 8 Reading Scale Score constant. One possible explanation for this is that in states with lower participation, only the most motivated and prepared students are taking the test.