Data Analysis on tigerstats (SAT Dataset)

Sachin Wadhwa

Table of Contents

[1 Background 2](#_Toc84071860)

[2 Data Description 2](#_Toc84071861)

[3 Methods 3](#_Toc84071862)

[3.1 Salary and SAT Score 3](#_Toc84071863)

[3.2 Correlation Test - Salary and SAT Score 4](#_Toc84071864)

[3.3 Expenditure and SAT Score 4](#_Toc84071865)

[3.4 Correlation Test - Expenditue and SAT Score 4](#_Toc84071866)

[3.5 Student-Teacher Ratio and SAT Score 5](#_Toc84071867)

[3.6 Correlation Test - Student Ratio and SAT Score 5](#_Toc84071868)

[4 Results 6](#_Toc84071869)

[4.1 Salary and SAT Score 6](#_Toc84071870)

[4.2 Correlation Test Resullts - Salary and SAT Score 7](#_Toc84071871)

[4.3 Expenditure and SAT Score 7](#_Toc84071872)

[4.4 Correlation Test Results - Expenditure and SAT Score 9](#_Toc84071873)

[4.5 Student-Teacher Ratio and SAT Score 10](#_Toc84071874)

[4.6 Correlation Test - Student Teacher Ratio and SAT Score 11](#_Toc84071875)

[5 Conclusion 12](#_Toc84071876)

[6 References 12](#_Toc84071877)

# 1 Background

Does increased public investment in secondary education lead to better educational outcomes for students? We certainly hope so. Indeed when state officials are looking to determine the how much of tax revenue to set aside for public schooling, it can be helpful to know which type of investment —- increasing teacher salaries, spending more to educate each student, arranging for smaller classrooms—— are the most effective in helping students do better in school

We can investigate such question using the sat dataframe from the tigerstats contributed in R package. The sat dataframe consists of the information on all the fifty states in the United states. For each state, we are given the following variables to investigate

# 2 Data Description

* **expend** : Mean annual expenditure per student (in 1000$).
* **ratio** : Mean student-teacher ratio.
* **salary** : Mean annual teacher salary.
* **frac** : Percentage of students in the US state who take the SAT.
* **verbal** : Mean SAT Verbal score for the state.
* **math** : Mean SAT Math score for the state.
* **sat** : Sum of mean Verbal and mean Math.

| Table : SAT Dataset | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| state | expend | ratio | salary | frac | verbal | math | sat |
| Alabama | 4.405 | 17.2 | 31.144 | 8 | 491 | 538 | 1,029 |
| Alaska | 8.963 | 17.6 | 47.951 | 47 | 445 | 489 | 934 |
| Arizona | 4.778 | 19.3 | 32.175 | 27 | 448 | 496 | 944 |
| Arkansas | 4.459 | 17.1 | 28.934 | 6 | 482 | 523 | 1,005 |
| California | 4.992 | 24.0 | 41.078 | 45 | 417 | 485 | 902 |
| Colorado | 5.443 | 18.4 | 34.571 | 29 | 462 | 518 | 980 |

We can think of first three variables as explanatory variables,whereas seventh variable **sat** is clearly the response variable. Our aim is to study the relation ship between each of these explanatory and **sat** variables. That is we will look in to the following three questions?

* Is the state with higher expenditure liable to have higher mean SAT Score ?
* Is the state with higher salaries liable to have higher mean SAT Score ?
* Is the state with lower student ratio liable to have higher mean SAT score ?

# 3 Methods

At the first glance it would appear that all we have to do is to examine a few scatter plots: after all, in each of the three questions above, we are interested in the relationship between the two numerical variables. For example, In order to the examine the relationship between teacher salary and mean SAT

## 3.1 Salary and SAT Score

Below code will plot the mean salary for teachers in the US State on the horizontal axis and the vertical axis gives the mean SAT Score for school students

xyplot(sat~salary,  
 data=sat,  
 type=c("p","smooth"),  
 pch=19,  
 main="Salary and SAT Scores",  
 xlab="mean teacher salary (1000$)",  
 ylab="mean SAT score"  
 )

From the above code and Fig 1 in Results section we see that states with higher teacher salaries tend to have lower Sat scores. This does not seem right.

Examining only two variables in isolation can be dangerous. In particular , we need to be on the watch for confounding variables. Recall that in a study of the relationship between an explanatory variable and a response variable , a third variable is said to be “confounding” variables if:-

* is related to the explanatory variable and
* helps to cause the response variable

Including the confounding variable, The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT

Percentage <- equal.count(sat$frac,number=3,overlap=0)  
xyplot(sat~salary | Percentage,  
 data=sat,  
 layout=c(3,1),  
 type=c("p","r"),  
 pch=19,  
 main="Salary and SAT Scores",  
 xlab="mean teacher salary (1000$)",  
 ylab="mean SAT score"  
 )

## 3.2 Correlation Test - Salary and SAT Score

Below code tells us the correlation between Salary and SAT Score by using the Pearson Method

df=sat  
cor.test(df$sat,df$salary,method=c("pearson"))

## 3.3 Expenditure and SAT Score

In order to the examine the relationship between Annual Expenditure and mean SAT, we could do it using below code :-

xyplot(sat~expend,  
 data=sat,  
 type=c("p","smooth"),  
 pch=8,  
 main="Annual Expenditure and SAT Scores",  
 xlab="Mean annual expenditure per student (in 1000$)",  
 ylab="mean SAT score"  
 )

If we include the confounding variable. The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT

Percentage <- equal.count(sat$frac,number=3,overlap=0)  
xyplot(sat~expend | Percentage,  
 data=sat,  
 layout=c(3,1),  
 type=c("p","r"),  
 pch=8,  
 main="Annual Expenditure and SAT Scores",  
 xlab="Mean annual expenditure per student (in 1000$)",  
 ylab="mean SAT score"  
 )

## 3.4 Correlation Test - Expenditue and SAT Score

Below code tells us the correlation between Expenditure and SAT Score by using the Pearson Method

df=sat  
cor.test(df$sat,df$expend,method=c("pearson"))

## 3.5 Student-Teacher Ratio and SAT Score

In order to determine the relationship of SAT Score and Student Teacher ratio , we can find the results using the below code

xyplot(sat~ratio,  
 data=sat,  
 type=c("p","smooth"),  
 pch=18,  
 main="Student-Teacher Ratio and SAT Scores",  
 xlab="mean student-teacher ratio",  
 ylab="mean SAT score"  
 )

Including the confounding variable, the mean student-teacher ratio and mean SAT score for all fifty states.The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT

Percentage <- equal.count(sat$frac,number=3,overlap=0)  
xyplot(sat~ratio | Percentage,  
 data=sat,  
 layout=c(3,1),  
 type=c("p","r"),  
 pch=18,  
 main="Student-Teacher Ratio and SAT Scores",  
 xlab="mean student-teacher ratio",  
 ylab="mean SAT score"  
 )

## 3.6 Correlation Test - Student Ratio and SAT Score

Below code tells us the correlation between Student Ratio and SAT Score by using the Pearson Method

df=sat  
cor.test(df$sat,df$ratio,method=c("pearson"))

# 4 Results

## 4.1 Salary and SAT Score

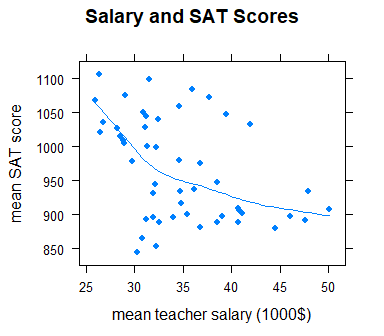


Fig 1 Each point on the plot represents one of the fifty states in the United States. The horizontal axis gives the mean salary for teachers in the US State, in the thousand of dollar. The vertical axis gives the mean SAT Score for school students, A loess curve is shown to indicate the negative relation between teacher salary and student SAT score

From figure 1 we see that states with higher teacher salaries tend to have lower Sat scores. This does not seem right.

Examining only two variables in isolation can be dangerous. In particular , we need to be on the watch for confounding variables. Recall that in a study of the relationship between an explanatory variable and a response variable , a third variable is said to be “confounding” variables if:-

* is related to the explanatory variable and
* helps to cause the response variable

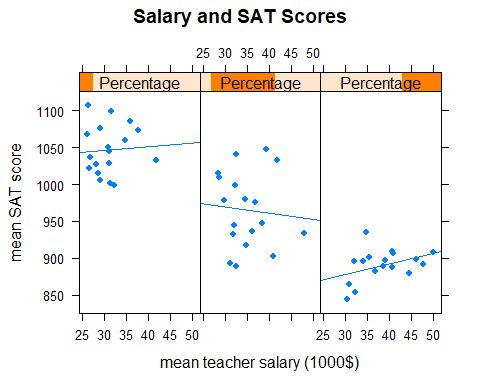


Fig 2 Scatterplot of mean salary for teachers and mean SAT score for all fifty states.The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT.The group with the lowest percentage is on the left. Regression lines are inlcuded in order to give an idea of trend in each group

## 4.2 Correlation Test Resullts - Salary and SAT Score

Below code tells us the correlation between Salary and SAT Score by using the Pearson Method. It shows that there is a negative correlation with moderate magnitude

##   
## Pearson's product-moment correlation  
##   
## data: x and y  
## t = -3.3936, df = 48, p-value = 0.001391  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.6398834 -0.1840738  
## sample estimates:  
## cor   
## -0.4398834

## 4.3 Expenditure and SAT Score

In order to the examine the relationship between Annual Expenditure and mean SAT , we could examine in Fig 3

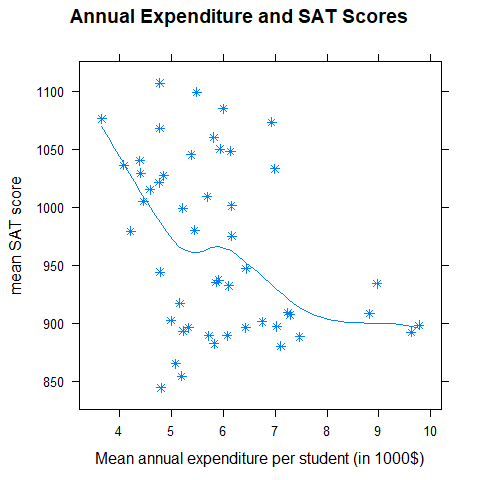


Fig 3 Each point on the plot represents one of the fifty states in the United States. The horizontal axis gives the mean expenditure per student, in the thousand of dollar. The vertical axis gives the mean SAT Score for school students

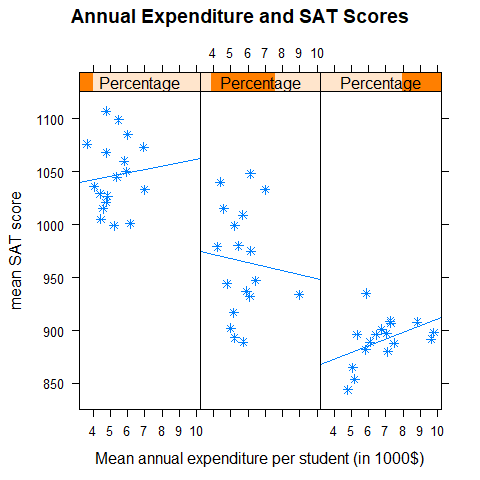


Fig 4 Scatterplot of mean expenditure per student, in the thousand of dollar and mean SAT score for all fifty states.The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT.The group with the lowest percentage is on the left. Regression lines are inlcuded in order to give an idea of trend in each group

## 4.4 Correlation Test Results - Expenditure and SAT Score

Below code tells us the correlation between Expenditure and SAT Score by using the Pearson Method.It shows that there is a negative correlation with moderate magnitude

##   
## Pearson's product-moment correlation  
##   
## data: x and y  
## t = -2.8509, df = 48, p-value = 0.006408  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.5957789 -0.1142957  
## sample estimates:  
## cor   
## -0.380537

## 4.5 Student-Teacher Ratio and SAT Score

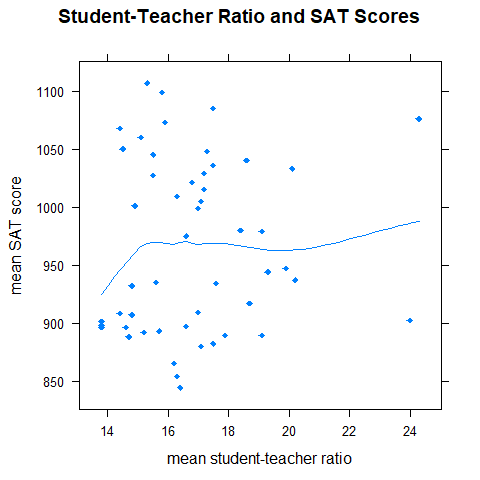


Fig 5 Scatterplot determines the percentage of students in the state US that take SAT.The group with the lowest percentage is on the left. Regression lines are inlcuded in order to give an idea of trend in each group

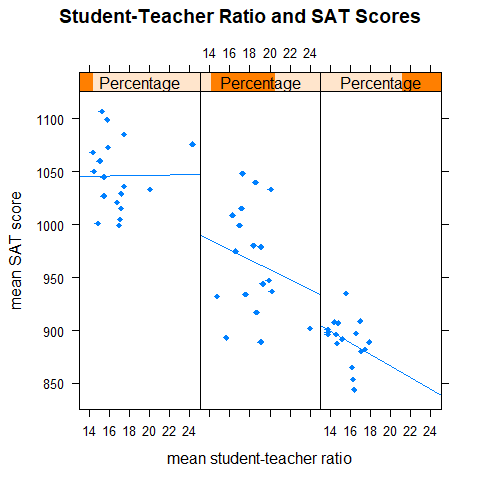


Fig 6 Scatterplot of mean student-teacher ratio and mean SAT score for all fifty states.The states are broken down in to 3 groups of about equal size, determined by the percentage of students in the state US that take SAT.The group with the lowest percentage is on the left. Regression lines are inlcuded in order to give an idea of trend in each group

## 4.6 Correlation Test - Student Teacher Ratio and SAT Score

Below code tells us the correlation between Student Teacher Ratio and SAT Score by using the Pearson Method.It shows that there is a positive correlation with weak magnitude

##   
## Pearson's product-moment correlation  
##   
## data: x and y  
## t = 0.56481, df = 48, p-value = 0.5748  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2016547 0.3516484  
## sample estimates:  
## cor   
## 0.08125382

# 5 Conclusion

For each of our three research questions, we found that after we corrected for the **frac** variable there was not a particularly strong relationship between the explanatory variable and mean SAT Score , In some case though , the relationship appear to have sort of direction that intuition would have led us to expect : a positive relationship between salary/expenditure and SAT, and a negative relationship between student-teacher ratio and SAT

There is a serious challenge , however, to the approach of correcting for a confounding variable by breaking down one’s data in to small groups that are like , within each group , in their values of the confounded.When the data is already small, it was in our study where there are only 50 states, then the separate groups become too small to pick up on definite relationship that might exist between the two variable

Another problem with our approach is that it relied entirely on graphs . SOmetimes a relationship s difficult to find visually, but appear more obviously in a numerical analysis

We need to find the numerical a method of analysis that keeps all the data together but still accounts for confounding variables, such methods do exist. Later on in this course on perhaps in future courses you take in stats, you will meet linear models that allow for many explanatory variables. Such models allows you to perform multiple regression and they shed a great deal more light we brought to sat data

# 6 References

* “Getting What You Pay For : The Debate Over Equity in Public School Expenditures” by Deborah Lynn Gruber *Journal of Statistics Education*, volume 7, Issue 2, 1999. Weblink: <http://www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm>.
* <https://rdrr.io/cran/tigerstats/man/sat.html>
* <https://www.rdocumentation.org/packages/tigerstats/versions/0.3.2>
* <https://cran.r-project.org/web/packages/tigerstats/tigerstats.pdf>
* <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test>