Final Project

Camille Starck

5/6/2021

## Introduction

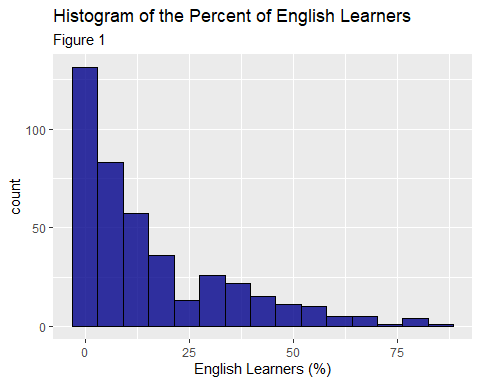
In this report, we will perform an analysis on the CASchools dataset to find factors that affect test scores. The CASchools dataset contains data on test performance, school characteristics and student demographic backgrounds for school districts in California. The data used here are from all 420 K-6 and K-8 districts in California with data available for 1998 and 1999. School characteristics and demographic variables are averaged across their respective districts. The following are the variables in the dataset.

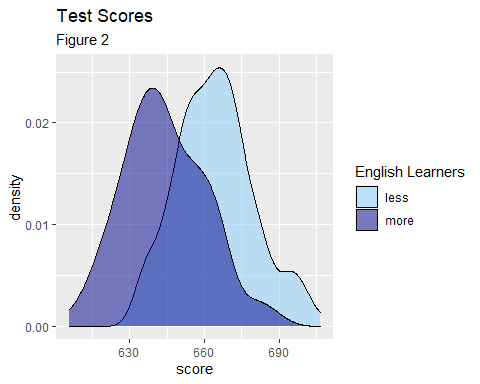
district: District code  
school: School name  
county: Indicating county  
grades: Indicating grade span of district  
students: Total enrollment  
teachers: Number of teachers  
calworks: Percent qualifying for CalWorks (income assistance)  
lunch: Percent qualifying for reduced-price lunch  
computer: Number of computers  
expenditure: Expenditure per student  
income: District average income (in USD 1,000)  
english: Percent of English learners (that is, students for whom English is a second language)  
read: Average reading score  
math: Average math score

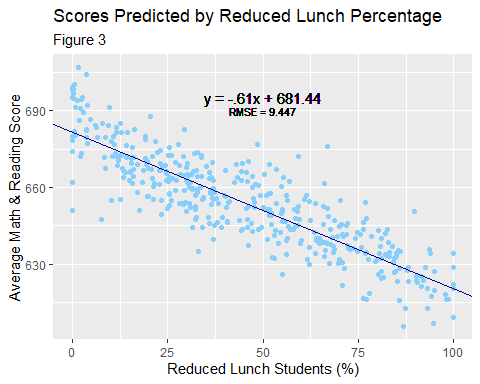
Here is a sample of the dataset. Note that only a few variables have been selected and some of the original data has been rounded for the display.

## district students teachers lunch computer income english read math  
## 1 75119 195 10.9 2.04 67 22.69 0.00 692 690  
## 2 61499 240 11.1 47.92 101 9.82 4.58 660 662  
## 3 61549 1550 82.9 76.32 169 8.98 30.00 636 651  
## 4 61457 243 14.0 77.05 85 8.98 0.00 652 644  
## 5 61523 1335 71.5 78.43 171 9.08 13.86 642 640  
## 6 62042 137 6.4 86.96 25 10.41 12.41 606 605

## Visualizations

Here are visualizations of the data to which the Analysis and Conclusion sections will refer. 





## Analysis

For the first analysis, we conduct a hypothesis test to answer the question- **Do schools with a higher percentage of English learners have lower test scores than other schools?** The test score is calculated by averaging the math and reading scores. Also, recall that an English learner is one for whom English is a second language. From Figure 1, we can see that the distribution of English learners is right-skewed, so we use the median (instead of the mean) as the defining distinction between more English learners versus less. Schools labeled *more* have a percentage greater than the median of English learners, and otherwise schools are labeled *less*. Here are the average scores of the two categories.

## eng\_learn avg\_test\_score  
## 1 less 664.354  
## 2 more 643.959

We can see from this summarized data that schools with more English learners have a lower test average. When we use a t-test to test the claim that schools with a larger percentage of English learners have a lower test score average than schools with a smaller percentage of English learners, R yields this p-value.

## [1] 7.302066e-33

Next, we try to model this data. Here we look at the correlation coefficients of all the variables with *score*.

## score  
## students -0.15  
## teachers -0.14  
## calworks -0.63  
## lunch -0.87  
## expenditure 0.19  
## income 0.71  
## computer -0.07  
## english -0.64  
## score 1.00

We see that *score* has the highest correlation with *lunch*, so we make a simple linear model using these variables. This regression yields an RMSE = 9.447, illustrated in Figure 3. Further, a multiple linear regression of *score* predicted by *english*, *lunch*, *expenditure*, and *income* yields an RMSE = 8.424.

## Conclusion

From the t-test that yields a very small p-value, we can conclude that there is evidence to support the claim that schools with more students whose first language is not English tend to perform worse on these tests than schools with more native English speakers. Figure 2 shows this gap. Also, from the simple regression model and Figure 3, we can conclude that as the percent qualifying for reduced lunch increases, the test scores decreases. Further, from the multiple linear regression, we can see that *english*, *lunch*, *expenditure*, and *income* all affect the test scores as well. Because of a smaller RMSE value, this model is slightly better.

In conclusion, this report shows that four important factors affecting test scores are the percent of English learners, the percent qualifying for reduced lunches, the expenditure per student, and the district average income. These findings are consistent with our previous knowledge. First, it is reasonable that children learning in a language that is not as familiar to them will not perform as well on tests. Also, schools with more students from poverty tend to perform worse than other schools. This is supported by a plethora of research on poverty and education. And finally, the expenditure per student represents a portion of how much each school is investing in its students, which will understandably affect test scores. Decision-makers should consider the research so they can provide the resources to students who may suffer from a language barrier or financial reasons.