James Griffin

Stat 536

Elementary Education Report

**Introduction**

Elementary education has become an area of increasing study among social scientists. Indeed, a US News and World Report states, “The elementary years from kindergarten through third grade are particularly important ones for children’s schooling.” Harvard Economist Roland Fryer Jr. and others have found surprising links between early elementary educational performance, and lifetime incomes, and that much of the wage gap between races in the country can be traced back to very early beginnings in the difference in testing.

Given the social and political climate to try and make the world a more equitable place, it is important that we more fully understand how to help children succeed in the classroom early in life.

For this analysis we will use data from various school districts throughout California. The data is advantageous in that it is very likely to generalize to the entire country very well, there are very wealthy neighborhoods, neighborhoods famous for crime, middle class neighborhoods, rural counties, big cities, and California is also one of the most racially diverse states in the country. Below is a heatmap of the variables within the data. The outcome variable for this analysis will be “Score”, being defined as the average cumulative score on Stanford 9 standardized test, a score from 0 to 1600.

A screenshot of a computer

Description automatically generated with medium confidence

This data set did have lots of features that were highly correlated positively or negatively with the Score. Not surprisingly, income is very highly correlated with scores. Wealthy families and neighborhoods fund schools with their taxes and can largely afford for remedial tutoring should their child require extra help for a given subject. Inversely, having higher percentage of the student body on government subsidized lunch is very negatively correlated with test scores. With less public funding, less access to outside help, and parents who, on average will probably be less educated themselves, children in these neighborhoods not only tend to be of a lower social demographic, but also have a higher likelihood to be non-native English speakers or have parents who are not. All of these are, unfortunately, correlated with lower incomes in the United States. An additional compounding effect of this lack of funding and income to these areas, is the lack of the number of teachers. It has long been stipulated that the more children per classroom, the harder it is for a teacher to cater to the individual needs of a student. While we do not dive into this question in this analysis, it is a feature which is negatively correlated with test scores, and expenditure, but negatively correlated with income, and slightly positively correlated with subsidized lunch.

**Methodology**

Having established the co-linear nature of the data, my initial reaction was to choose a nonlinear model, however, for the questions we wanted to answer in this analysis:

1. “Income” is generally a measure of how much money a school has to spend on extracurricular activities (as opposed to expenditures which is how much spent per student in the class room). Is there evidence of diminishing returns on extracurricular activities in terms of student learning?
2. Is English as a second language a barrier to student learning?
3. In your opinion and based on the data, what can be done to increase student learning?

We decided that we did not need to use nonlinear techniques and could use linear models. We approached the first question by creating a scatter plot with test scores and income.

Chart, scatter chart

Description automatically generated

And saw a nonlinear, decreasing marginal benefit to of income with respect to test scores. To ensure that we could do this, we needed to satisfy the assumptions for a linear model. In the appendix, we snow the normality, of our scores, the normality of log(income)[[1]](#footnote-1). The last assumption we need to satisfy for a linear model is homoscedastic errors, here we show the relative normality of those errors.

Chart, histogram

Description automatically generated

**Models and Assumptions**

Having satisfied with the assumptions we need for a linear model, we proceeded by binning the observations on what level of log(income) they were associated with: (0, log(9.99)), (log(10), log(19.99)), (log(20), log(29.99)), (log(30), log(39.99)), (log(40), log(50+)). We wanted to see if the different levels of income had different effects on testing, so we ran a simple linear regression with only the logged income levels as our columns. The resulting equation was as follows:

Where:

The objective of the linear model is to minimize the sum of squared residuals equation:

Our results are presented in the following section.

We also wanted to investigate what if English as a second language was a barrier to learning. We wanted to take two approaches to this problem, first we wanted to compare groups schools with large amounts of non-English speakers with schools with lower amounts of non-English speakers.

Chart, scatter chart

Description automatically generated

We See a rather natural divide around 25%. Because the percentage of English-speaking students is colinear with the percentage of students on subsidized lunch and income[[2]](#footnote-2), We wanted to create some interaction terms to be able to tease out the height of the language barrier. We took two approaches to this problem, the first was a t-test difference in means, where we used the following equation:

Our second approach to the height of the language barrier, should one exist, was to use again, a linear model. Because we have discussed the correlation between number of English speaking students and the number of students on subsidized lunch, we needed to interact lunch with English. The resulting equation was as follows:

We used the negative coefficients on lunch and English to ensure that the negative correlation was preserved. The results of both tests are discussed in the following section.

**Results**

Below is the output table from the first linear regression described above. When determining how much of the effect was decreasing, we took the coefficients for each of the bins, and found the sequential difference for each bin. We find that the first bin, from 0-$10,000 has the coefficient of 1265, meaning that this group can attribute that many points to their score from that expenditure. The next group, from $10,000 - $20,000 also spends the $0 - $10,000 so in order to find the marginal benefit we take the coefficient from 1313.1669 – 1265.5592 = 47.60, and we did this iteratively, finding 40.39, and 35.66. We see clearly here that the marginal return to scores in terms of extracurricular expenditure, is clearly negative. It is also worth noting that in this very simple regression, that our R-squared is 0.975, so we can be confident this model, the coefficients and their decreasing effect.

A picture containing text, receipt, screenshot

Description automatically generated

Examining our second equation and approach we find that in our t test, for difference of means, we found a t statistic of 10.584 and a p value of essentially 0, so we know that these two groups do not have the same testing score means, and when we ran the regression with the second equation, we find a statistically significant negative coefficient on English. These two data points do strongly suggest that English as a second language, is a barrier to learning.

A picture containing text, receipt, screenshot

Description automatically generated  
 When we look at the proposition of improving elementary education, the most important variable we found in this data set was income. When we only included our binned income, our R-squared was 0.975. Essentially explaining all of the variability with this single feature. It is this feature, that is highly correlated with expenditure, student to teacher ratio, subsidized lunch, computers per student, and others we have discussed. Thus, to improve education, the data seem to indicate that a greater access to funds, would be the primary way to help underprivileged students with numerous obstacles impleading learning. It is worth noting, that for an additional reference, I ran this data against a random forest that I had already coded up, not having to worry about distributional assumptions made it easy to just throw the data at the model, and he most important feature by a country mile was indeed income[[3]](#footnote-3).

**Appendix**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Score, English, Income, Lunch, Expenditure, STratio plot matrixA picture containing window, surrounded, middle

Description automatically generated

Chart, scatter chart

Description automatically generated

A picture containing histogram

Description automatically generated

1. See Appendix [↑](#footnote-ref-1)
2. See Appendix [↑](#footnote-ref-2)
3. See Appendix for details. [↑](#footnote-ref-3)