

Factors Associated with a State's Standardized Test Scores

How does the state budget influence test scores?

January 11, 2017

Kimia Mavon, Nicholas Hoernle, Franklin Wolfe, Amy Gu

Keywords: NAEP, USA, State, Education, Spending

1. Introduction

1.1. Education Scores and State Spending

In the 2016 November elections, the Massachusetts voter ballot asked voters whether the state should expand charter schools. This prompted the question of whether the amount that a state directly spends on education has a marked association with the average education scores that a state achieves. In this report we will analyze the association between state spending per pupil on education and on the standardized test scores that the pupils of that state achieve. Our aim is to determine the association between these two metrics, if any, and thereafter to analyze other potential contributing factors to a state's general education test scores.

We have used outcome data from The National Assessment of Educational Progress (NAEP) as the metric for evaluating educational performance.

1.2. Hypotheses of Interest

Using the 2009 grade 8 NAEP results, we are specifically aiming to determine the association between state spending and the education performance of a state; however, we will still be including a number of other predictors of interest in the assessment to test for confounding variables when determining how spending affects performance.

We therefore have the specific hypotheses under study:

- H_0 : 2009 state spending per pupil is not associated with 2009 student performance on NEAP.
- H_A : 2009 state spending per pupil is associated with 2009 student performance on NEAP.

2. Methods

2.1. Data Summary & Selection

Educational information from states and districts comes from the National Center for Educational Statistics (NCES), which is a federal entity for collecting and analyzing data related to private and non-private education in the U.S. and other nations (NCES, 2016). It operates under the U.S. Department of Education and fulfills a Congressional mandate to collect, analyze, and report on statistics that help decision-makers to act towards the nation's best educational interests. The National Assessment of Educational Progress (NAEP) is the NCES's primary assessment of the state of elementary and secondary students' knowledge and is the nations largest representative assessment of students. The program has conducted periodic tests in reading, mathematics, science, writing, and other subjects since 1969. Each state and jurisdiction that participates in the program receives a report on key findings and trends in a condensed format, including overall student performance and demographic performance information. A typical assessment includes approximately 500,000-1,000,000 student scores from a randomly sampled, representative population of students at large (NCES, 2016).

We chose to analyze 2009 educational statistics and test scores from the NAEP Data Explorer Tool because this was one of the few years in the past decade in which the test was administered nationwide to all 50 states. We picked math and reading scores because we felt these two metrics were the most representative of a broad, basic level of education. These two tests are also the most widely administered of the academic subjects, and thus would give us the greatest representative sample of state scores. The test is conducted at grades 4 and 8 at the state level to public schools, private schools, Bureau of Indian Education

schools, and Department of Defense Schools. We picked grade 8 for our analysis because we felt that it would be more representative than grade 4 for effects of school funding on academic achievement, whereas grade 4 would be affected more by confounding factors, such as access to pre-schooling, family income, early childhood development centers, etc.

Demographic information (poverty, income, Gini coefficient, etc.) primarily came from data files of the U.S. Census Bureau. The Bureau's goal is to serve as the leading source of quality data about the nation's people and economy (U.S. Census Bureau, 2016). Data from their archives was collected for the 2008-2009 year, to match the time when the test was taken (i.e. 2009 test covers 2008-2009 school year). We picked metrics, such as poverty and race, which we believed could be strong confounding influences on the level of academic achievement of students. Additional information came from The Tax Foundation and National Conference of State Legislatures for vouchers, and the National Education Association (number of schools and teachers). Appendix A contains a complete description and source listing for data included in this study, along with the motivation behind our choices of predictor variables.

2.2. Data Cleaning

The data were obtained in raw format and required a minimal amount of preprocessing before being made usable. Most notably, while the NAEP dataset contained scores for all 50 states as well as scores for District of Columbia, many of the other predictors were limited to only including state level data. We made the decision to remove District of Columbia from the dataset as it is expected to not be representative of states in a state level comparison. Further potential predictors were included in the source data files, but were excluded as they were deemed to be irrelevant for grade 8 level NAEP scores. These predictors included data such as 'Freshmen Graduation Rates', 'Grade 9 through 12 Demographics' and 'Private School' specific data. Some of these excluded data had missing variables; however, they were not removed for this reason.

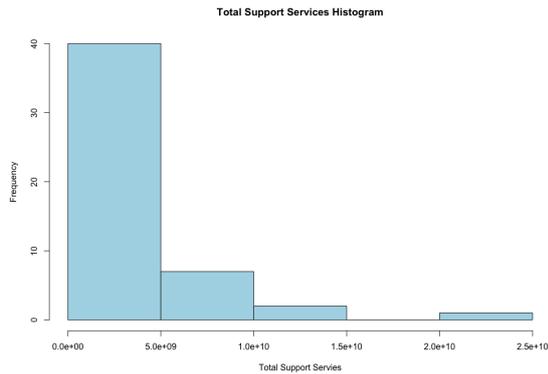


Figure 1: Total Support Services Raw Data: This histogram demonstrates the variable "Total Support Services" prior to transformation. The data is highly skewed.

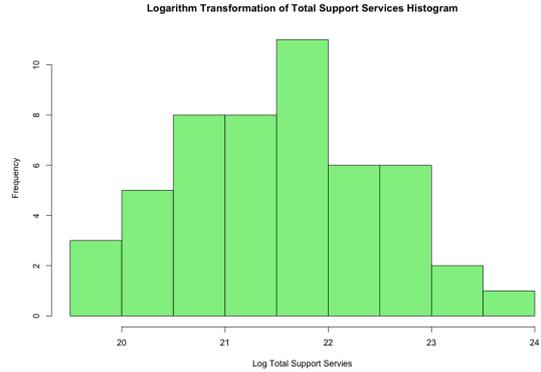


Figure 2: Total Support Services Log Transformed: This histogram demonstrates the variable "Total Support Services" after transformation. The data is now approximately normally distributed.

2.2.1. Transformations

To ensure the assumptions of linear regression are met, we assessed the relationship between the input and output variables.

Because much of the data was right skewed, logarithmic transformations were applied to the following variables: the state's per pupil expenditure in dollars, total employee salaries in dollars, total support services in dollars, mean number of people per household, the Gini coefficient, and number of full-time equivalent teachers, a qualification for a specific amount of hours taught in a given-year. A square root transformation was applied to the percentage of persons 18 and younger in poverty in the state. Figures 1 and 2 provides an example of such a transformation.

2.3. Model Selection

In terms of comparative models, we trained a model on all of the available main terms. We further trained a model on all of the significant predictors that were noticed in the fully trained model. In this model, we included the interaction terms among these predictors. We were unable to build a full model of all the main predictors and all of their interaction terms as this model had more predictors than there are datapoints (thereby using all of the available degrees of freedom). To attempt to identify possibly important interaction terms,

we executed a Lasso Regression to predict the outcome variable from all the main predictors with their two way interaction terms. Figure 3 shows the non-zero coefficients after Lasso was run with a regularization parameter of 0.79. The choice of this parameter was made purely for the use of coefficient selection. While there are 14 non-zero predictors shown in the plot, the magnitudes of the top 5 predictors are substantially larger (on the magnitude of at least 10 times larger) than the remaining 9 and thus we simply selected these predictors.

Significantly, the predictors from this regression can be used (including all of the main predictor terms to ensure that the results are interpretable) to train a new linear regression model.

Each of the above models was then further run through a step wise ‘backward’ model selection with the lower bound being an empty model. The models were selected based on their AIC scores and the summaries of these results can be found in Table 1.

3. Results

3.1. Assumptions

The modeling method used for this analysis required the following assumptions be met:

- Linearity:

The four models’ Residuals vs Fitted plots (Figures 4, 8, 12, 16) demonstrated equally spread residuals around a horizontal line without distinct patterns, indicating linear relationships.

- Normality:

Our Q-Q plots demonstrate that the residuals in models 1 (Figure 5), 2 (Figure 9), and 3 (Figure 13) are normally distributed. Model 4’s Q-Q plot (Figure 17) demonstrates this dataset contains fatter tails than typically found in normal distributions.

- Multicollinearity:

Some pairs of predictor variables were correlated with each other with a coefficient greater than 0.9 in models 1 through 4, so this assumption was violated. For example, people per household and gini coefficient were almost perfectly correlated ($r = 0.992$),

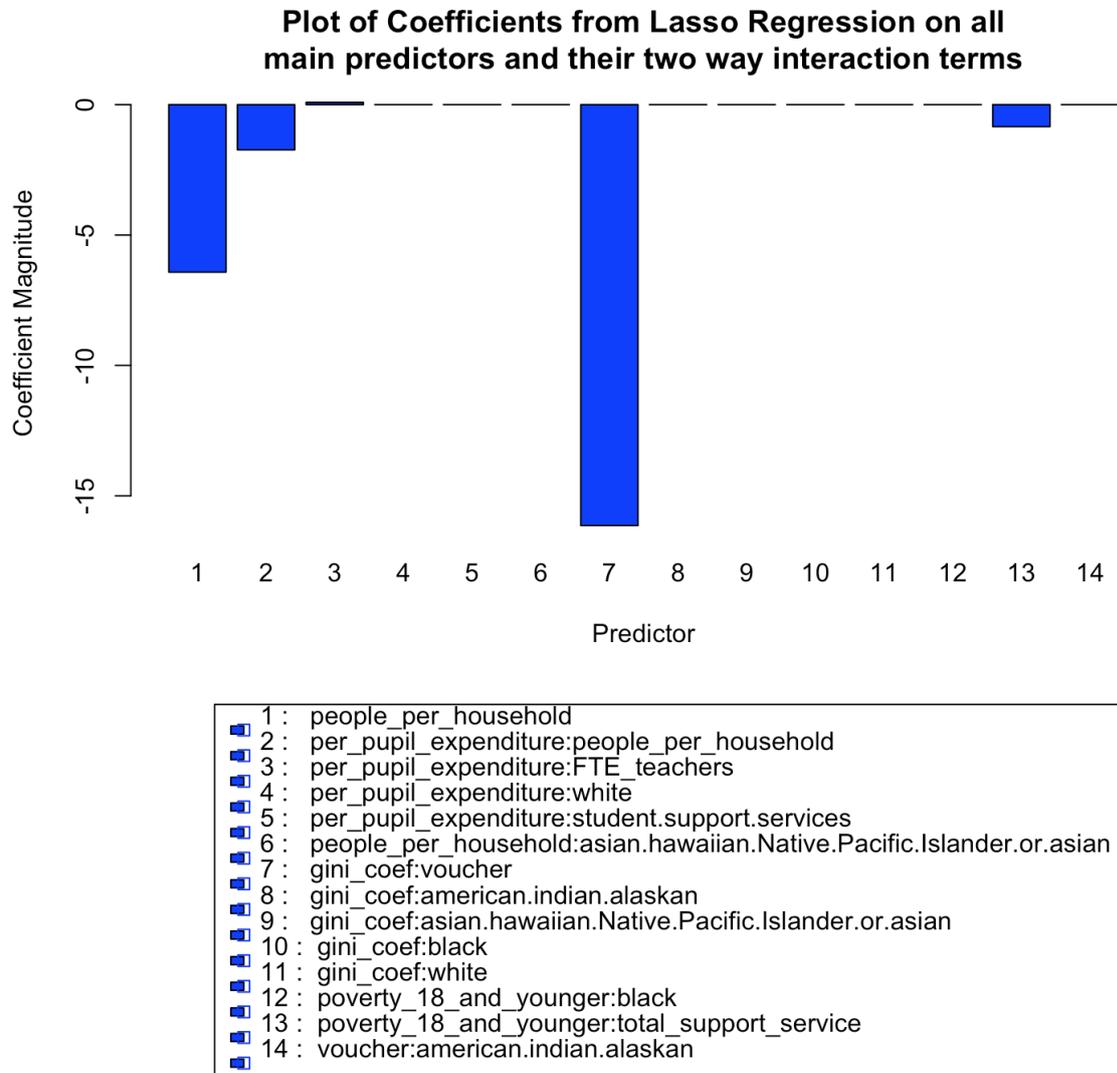


Figure 3: Plot of the magnitudes of the coefficients after Lasso Regression was run on all main predictors and their two way interaction terms

as well as employee salaries and total support services ($r = 0.991$). The multicollinearities signify that the models are more prone to estimation error and the regression coefficients are more volatile.

- Independence:

States tend to be correlated by region, so independence is violated. For example, figure 20 shows that state spending in the Northeast is higher than the national average, while the South spends less than average.

- Homoscedasticity:

Overall, we find that the residuals in all four models are roughly centered on zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values. The Scale-Location plots for all four models (Figures 6, 10, 14, 18) demonstrated residuals spread equally along the ranges of the predictors, supporting our assumption of equal variance (homoscedasticity).

3.2. Influential Points

The residuals vs leverage plots demonstrate a handful of states that may have a relatively large influence on the regression line. In models 1 and 2 (Figures 7, 11), California (state 5) is located on the borderline of Cooks distance. In model 3 (Figure 15), California is borderline outside of Cook's distance. California may have high leverage since it has a large population and therefore has high values for explanatory variables like Asian/Pacific Islander population and total support services. However, it does not appear to be an outlier in the response variable, so we will proceed with the model.

In Model 4, no state is particularly influential. Overall, the analysis of residuals suggests that our models represent the data well.

4. Discussion

Each model received a statistically significant F score and p value (Table 1). Balancing the explanatory power of regression models and the numbers of predictors, we chose model 1 to best fit our regression line. Model 1 provided the lowest AIC score (350.9), the highest

adjusted R squared term (0.7048), and an F statistic of 12.7. An F score of 12.7 with 43 and 6 degrees of freedom respectively results in a p value of 1.922×10^{-9} . Additionally, model 1 best met the regression assumptions, as discussed in the results section.

It is important to note in Table 1, the slope coefficients include both positive and negative values, suggesting that certain predictors (employee salary, total support services etc) are positively associated with the outcome variable while others (people per household, poverty 18 and younger) have a negative association. Furthermore, while the per pupil expenditure predictor has a positive slope coefficient in models 3 and 4, suggesting that it is indeed associated with increased test scores, these values are far from significant with p-values on the order of 0.5.

4.1. Simpson's Paradox

Thus far, we have been careful not to make inferences about individuals, schools, or counties within states since we only have looked at state-level data. This was to avoid ecological fallacy and Simpson's paradox. Associations that were drawn from aggregated data do not necessarily hold for the individual data, because there may be confounding factors such as the distribution of the predictor variable being very different between groups than the distribution within a group.

For example, in the case of education spending, funding is actually controlled at the district level rather than at the state level, so each school district decides the amount of funds it allocates to education spending. The distribution of per pupil expenditure across states could differ from the distribution among the districts within a state.

NAEP also released scores for 18 districts that participated in 2009. To investigate whether results at a more individual level line up with the state-level results, we shall look at these district NAEP scores and fit models to them with the following factors: per pupil expenditure, percentage white, percentage black, mean people per household, and poverty rate for person 5 to 17 years old. Per pupil expenditure and poverty rate were slightly right-skewed, so a log-transformation and a square-root transformation were applied, respectively.

The backward stepwise model generated from the district data included 2 significant predictors: percentage of white people and poverty rate (Figure 21). In contrast, the cor-

Table 1: Table showing the different models, their predictors, the associated β values and the standard errors associated with those values.

Term	Model1		Model2		Model3		Model4	
	Beta	SE	Beta	SE	Beta	SE	Beta	SE
employee_salaries	1.9e+01	8.2e+00	1.3e+01	8.6e+00	-1.1e+03	3.8e+02		
total_support_services	2.6e-09	1.4e-09	3.3e-09	1e-09	-4.4e-07	1.8e-07		
people_per_household	-1.2e+02	7.2e+01					3.3e+02	2.7e+02
gini_coef	2.6e+02	1.9e+02						
poverty_18_and_younger	-1.8e+01	2.3e+00	-1.8e+01	2.3e+00	1.2e+02	8.2e+01	-1.8e+01	3e+00
asian.hawaiian.Native.Pacific.Islander.or.asian	-1.4e-04	2.7e-05	-1.2e-04	2.6e-05	-2e-04	5e-05		
hispanic	-1.4e-05	9.8e-06						
black	-4.5e-05	1.4e-05						
Total.number.of.students	1.8e-05	8.5e-06						
total_support_service	-2.4e+01	8.9e+00	-1.7e+01	9.1e+00	-4.7e+01	3.7e+02		
per_pupil_expenditure					6.4e+01	1.3e+02	5.8e+01	8.7e+01
employee_salaries:per_pupil_expenditure					5.1e+01	3.3e+01		
employee_salaries:total_support_services					-8.7e-09	3.3e-09		
employee_salaries:total_support_service					2.9e+01	9.8e+00		
per_pupil_expenditure:total_support_service					-5.4e+01	3.5e+01		
total_support_services:total_support_service					2.6e-08	1e-08		
poverty_18_and_younger:total_support_service					-6.3e+00	3.9e+00		
FTE_teachers							-4.7e+01	3.6e+01
people_per_household:per_pupil_expenditure							-4e+01	2.9e+01
per_pupil_expenditure:FTE_teachers							5.2e+00	3.9e+00
AIC Score	350.88		362.46		356.51		367.30	

responding backward stepwise model from state-level data (Figure 22) included household size in addition to white population and poverty rate. Thus, although household size has a significant association with NAEP scores for states, this association disappears at the district level, illustrating Simpson’s paradox.

5. Conclusion

This study aimed to determine the association between the education performance of a state and a number of predictors, including state spending per pupil. We trained a number of linear models aiming to provide an indication of the relationship between certain predictors and the education scores of a state. We are able to conclude that state spending per pupil is not significantly associated with these scores (and thus are unable to reject the null hypothesis introduced in Section 1.2), but predictors such as employee salaries, total support services, people per household, Gini coefficient, child poverty (18 and younger), Asian population, Hispanic population, black population, total number of students, and the total support services are. We used AIC as a metric for comparing these models and ultimately focused on the specific relationship that the predictors had with the education score outcome variable. Our conclusion is that policy regarding education funding should rather focus on general community services than directly on increasing the education budget.

The random sampling of the NAEP data collection allows us to conclude that the study may be generalized to the greater American 8th grader population’s scores in math and reading. Future studies will aim to build models on district data, rather than state, as this will provide more granular results, allow for more predictors, and increase the number of observations.

6. References

Analysis of K-12 Educational Spending, Ballotpedia (2012). Analysis of Spending in America’s Largest School Districts. [Webpage]. Available from https://ballotpedia.org/Analysis_of_spending_in_America%27s_largest_school_districts

Mathematics and Reading Results, The National Report Card (2015). District Results Overview. [Webpage]. Available from <https://www.nationsreportcard.gov/reading>

[_math_2015/#mathematics/district?grade=4.](#)

National Center for Educational Statistics (2010). Number of private schools, students, full-time equivalent (FTE) teachers, and 2008-09 high school graduates, by state: United States, 2009-10. [Data File]. Retrieved from http://nces.ed.gov/surveys/PSS/tables/table_2009_15.asp.

National Conference of State Legislatures (2016). School Voucher Laws: State by State Comparison. [Webpage]. Available from <http://www.ncsl.org/research/education/voucher-law-comparison.aspx>.

National Education Association (2010). Ranking of the States 2009 and Estimates of School Statistics 2010. [Report]. Available from <http://www.nea.org/assets/docs/010rankings.pdf>.

National Public Education Financial Survey Data, Common Core of Data (2013). Fiscal Year 2009. [Data File]. Retrieved from <http://nces.ed.gov/ccd/stfis.asp>.

National State Comparisons, National Center for Educational Statistics (2016). Grade 8 Reading and Mathematics Scores, National Assessment of Educational Progress for 20w09 school year. [Data File]. Available from <http://nces.ed.gov/nationsreportcard/statecomparisons/>

Quick Facts, United States Census Bureau (2010). People. [Data File]. Available from <http://www.census.gov/quickfacts/table/PST045215/00>.

Small Area Income and Poverty Estimates, United States Census Bureau (2009). [Data File]. Available from <https://www.census.gov/did/www/saipe/data/schools/data/2009.html>.

The Tax Foundation (2016). People per Household by State 2009-2010. [Data File]. Retrieved from http://taxfoundation.org/sites/taxfoundation.org/files/docs/state_people_per_household_2009-10-20120216.pdf.

United States Census Bureau (2013). State and County Estimates for 2009. [Data File], Retrieved from <https://www.census.gov/did/www/saipe/data/statecounty/data/2009.html>.

United States Census Bureau (2010). Household Income for States: 2008 and 2009. [Report]. Retrieved from <https://www.census.gov/prod/2010pubs/acsbr09-2.pdf>.

Vintage 2009: State Tables, United States Census Bureau (2012). Population by Selected

Age Groups and Population by Race and Hispanic Origin. [Data Files (2)]. Retrieved from https://www.census.gov/popest/data/historical/2000s/vintage_2009/state.html.

Why America's Schools have a money problem (2016). National Public Radio. [Web-page]. Retrieved from <http://www.npr.org/2016/04/18/474256366/why-americas-schools-have-a>

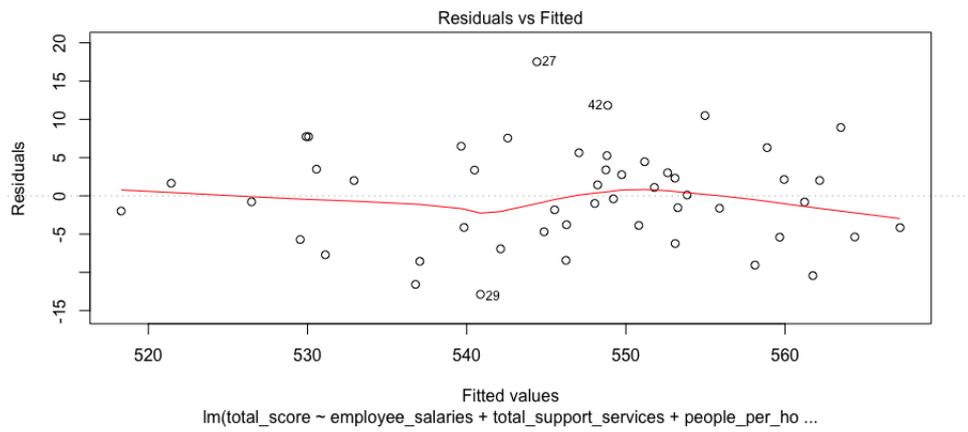


Figure 4: This plot demonstrates the Residual v Fitted values in Model 1.

7. Appendix

Table 2: State data, sources, descriptions, and motivations for including them in the study. The html links for these data sources can be found in the references section of the report.

Variable Name	State Data	Source	Description	Motivation
total_score	National Assessment of Educational Progress Scores (Math & Reading) for 8th Graders in 2009 (NAEP)	National Center for Educational Statistics (NCES) (2008-2009)	The NAEP Reading scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.	NAEP is the largest nationally representative assessment of students in the United States. NAEP results are based on representative samples of students at grades 4, 8, and 12.
per_pupil_expenditure	Per Pupil Expenditure	National Center for Educational Statistics (NCES) (2008-2009)	The number of dollars spent by the state per student. Calculated by using average attendance numbers for each state.	It is thought that spending more money on education (better salaries, better equipment, better teaching materials, will lead to higher performing students.
total_support_services	Total Support Services	National Center for Educational Statistics (NCES) (2008-2009)	The number of dollars spent on student support, instructional staff support, pupil transportation, and administration, among other support services.	It is thought that spending more money for support for students will give them more opportunity to perform highly with extra help.
employee_salaries	Employee Salaries	National Center for Educational Statistics (NCES) (2008-2009)	The number of dollars spent on teacher salaries.	It is thought that higher salaries may attract better teachers or motivate teachers to work at a higher level, thus potentially leading to higher performing students.
voucher	School Voucher Program	National Conference of State Legislators	A binary predictor variable that is 1 if they have a state voucher program and 0 if they do not. The voucher program is state funding for qualified students (often with disabilities or from low income families) to attend private or charter schools.	Thirteen states and the District of Columbia provide state-funded school vouchers to qualifying students. Many have argued that school choice and competition for students will foster higher performance, though this is debated.
people_per_household	Average People Per Household	The Tax Foundation	The average number of people that live in each household	It will be interesting to see how this may effect student performance. It is thought that perhaps smaller households will have higher performing students because the parents have less of a burden; however, smaller households could mean single parent families, which may lead to lower performance.
poverty_18_and_younger	Poverty	U.S. Census Bureau	The percentage of persons less than 18 years of age living in families below the poverty line.	It is thought that higher poverty rates will lead to lower performance as there are less home support service for students.
median_income	Income	U.S. Census Bureau	The median income in dollars of residents living in the state.	It is thought that higher income will lead to higher performance as there will be better living conditions for students.
gini_coef	Gini Coefficient	U.S. Census Bureau	The gini coefficient is a measure of statistical dispersion intended to represent the income distribution of a nation's residents, and is the most commonly used measure of inequality.	It will be interesting to see how income inequality may effect student performance.
FTE_teachers	Teachers	National Education Association	The number of K-12 teachers	It is thought that more teachers will lead to higher performance.
schools	Schools	National Education Association	The number of schools in the state	It is thought that more schools will lead to higher performance.
white_black_asian_hawaiian_native_pacific_islander_or_asian_hispanic_two_or_more_races_american_indian_alaskan	Demographic Data	U.S. Census Bureau	The number of persons who identify by each race (asian/hawaiian native/Pacific Islander or Asian, hispanic, black, white two or more races)	It is thought that more diversity may lead to lower performance.
Total.number.of.students	Students	U.S. Census Bureau	The total number of students measured by attendance at schools	It is thought that perhaps more students will lead to higher performance.
student.support.services	Students Support Services	U.S. Census Bureau	The total number of employees that work as a support for students (administration, transportation, etc.)	It is thought that more support will lead to higher performance.

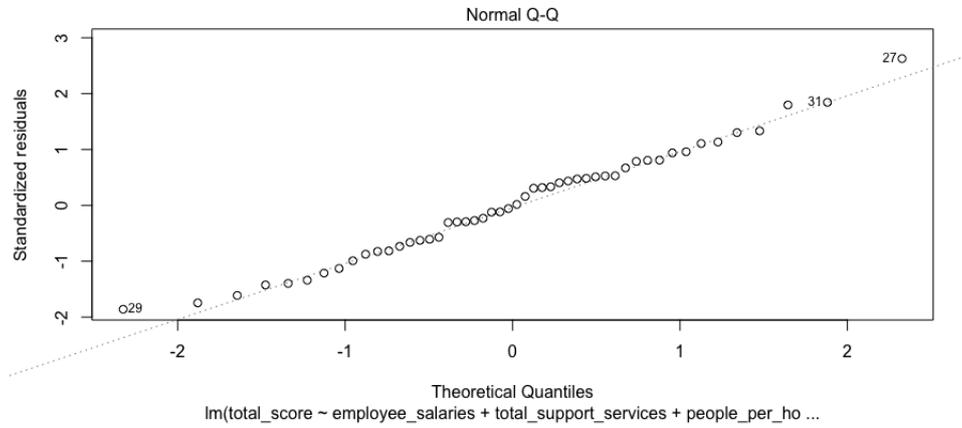


Figure 5: This QQ plot demonstrates normality in Model 1.

Table 3: District data, sources, descriptions, and motivations for including them in the study. The html links for these data sources can be found in the references section of the report.

Variable Name	District Data	Source	Description	Motivation
naep_total	National Assessment of Educational Progress Scores (Math & Reading) for 8th Graders in 2009 (NAEP)	The Nations Report Card	The NAEP Reading scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant.	These estimates have the same motivations as before on the state level. We are also using these to assess Simpson's Paradox.
perpoverty	Poverty	U.S. Census Bureau	The percentage of the population in poverty per district.	
per_pupil_expenditure	Per Pupil Expenditure	Ballotpedia	The number of dollars spent by the state per student.	
	Income	U.S. Census Bureau	The median income in dollars of residents living in the district.	
census_black2010, census_white2010	Demographic Data	U.S. Census Bureau	The percentage of the population that identifies as black and white for race.	
household_size	Average People Per Household	U.S. Census Bureau	The average household size in the district.	

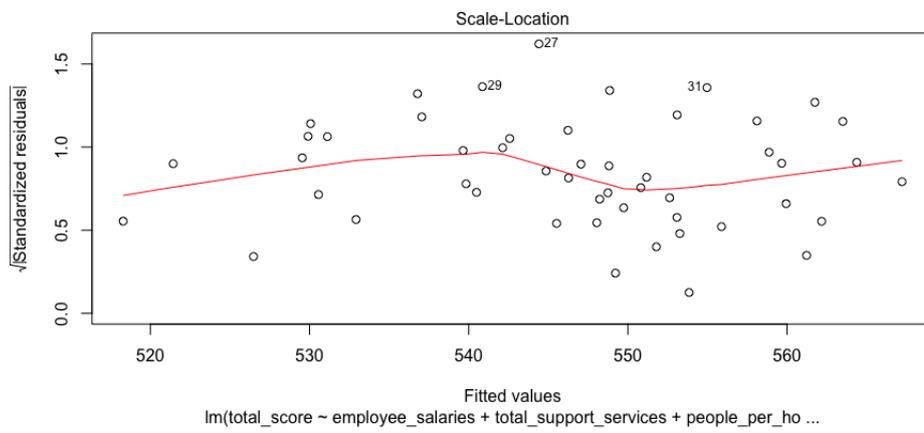


Figure 6: The The Spread-Location plot for all model 1.

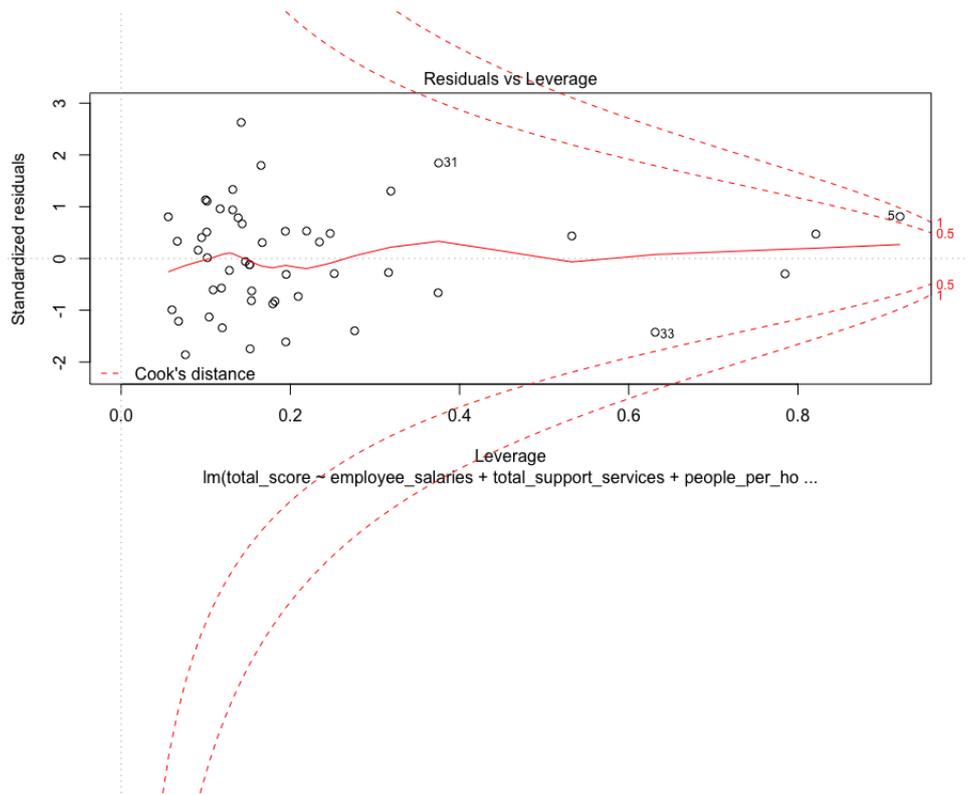


Figure 7: This plot demonstrates Residuals v Leverage in Model 1.

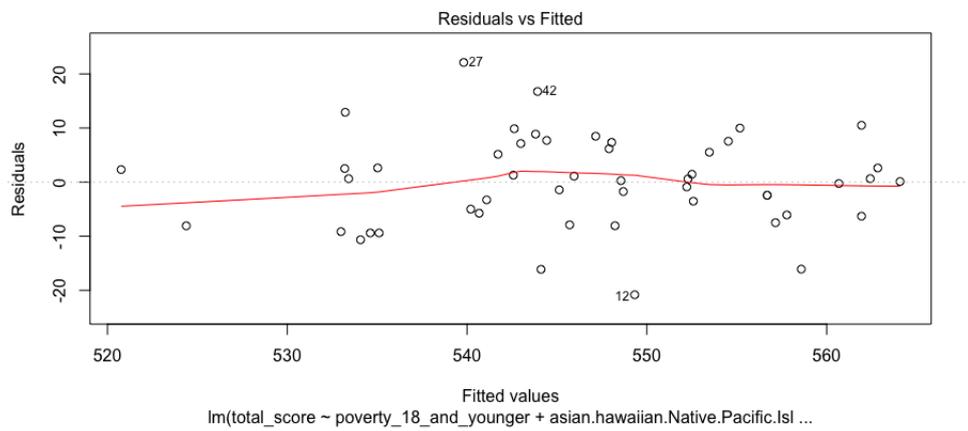


Figure 8: This plot demonstrates the Residual v Fitted values in Model 2.

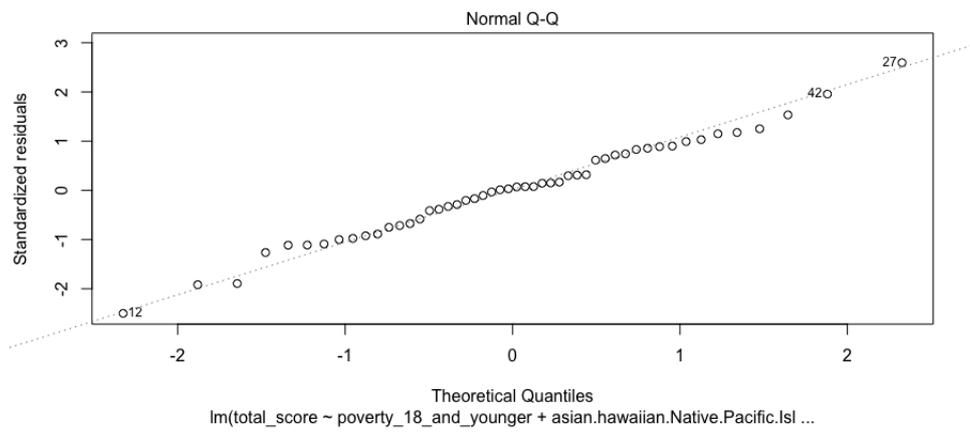


Figure 9: This QQ plot demonstrates normality in Model 2.

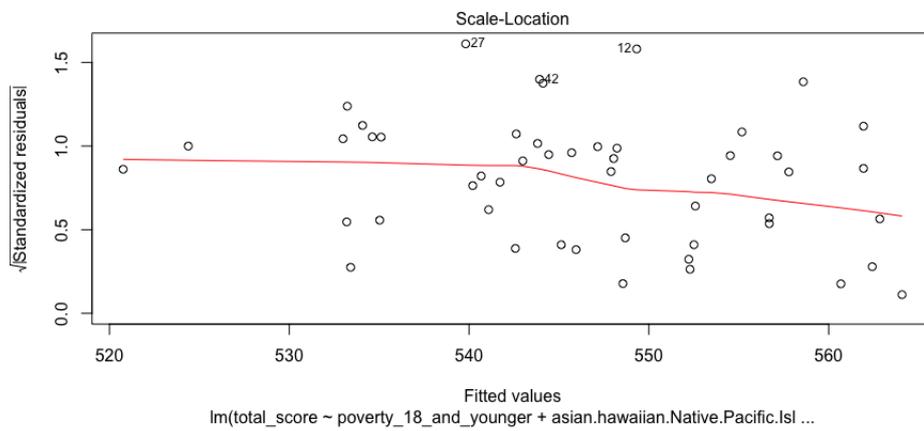


Figure 10: The The Spread-Location plot for all model 2.

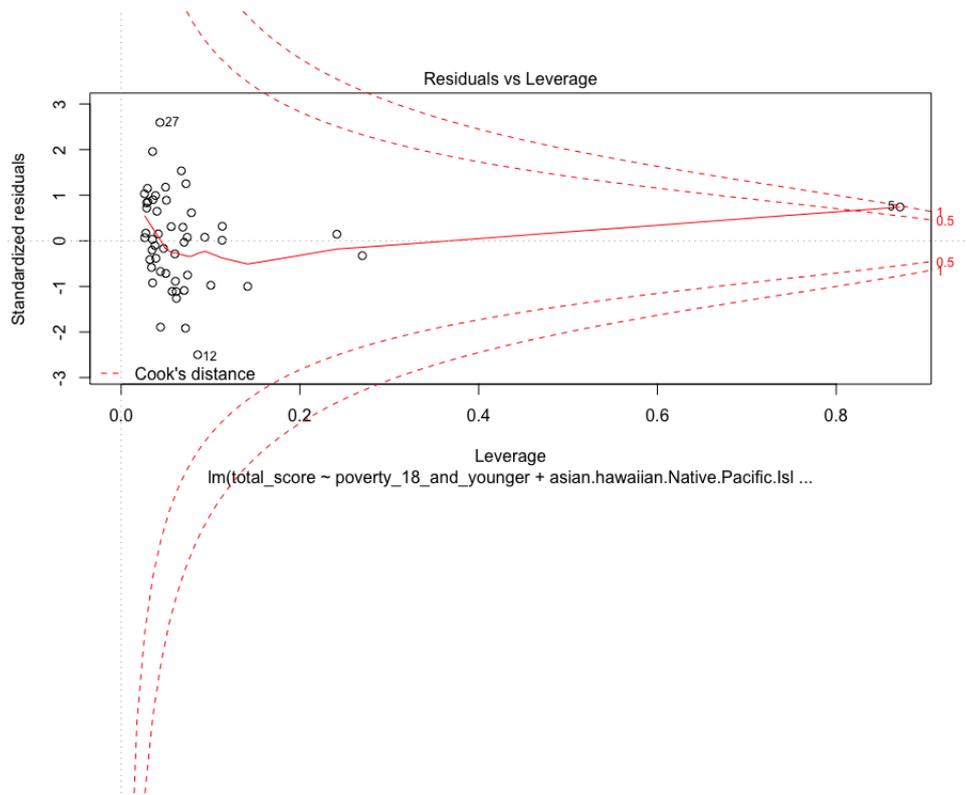


Figure 11: This plot demonstrates Residuals v Leverage in Model 2.

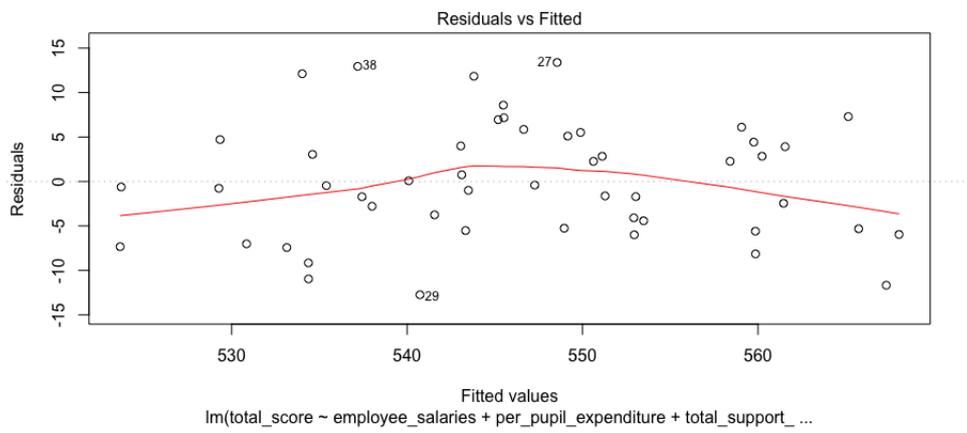


Figure 12: This plot demonstrates the Residual v Fitted values in Model 3.

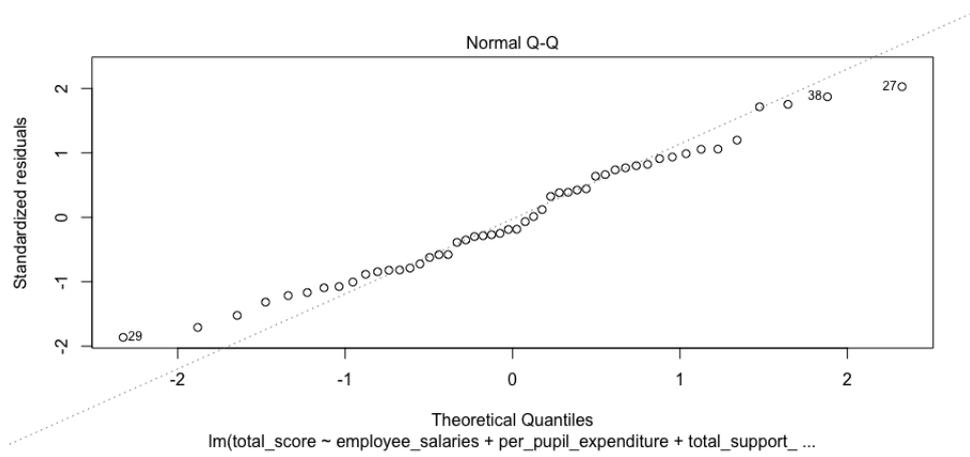


Figure 13: This QQ plot demonstrates normality in Model 3.

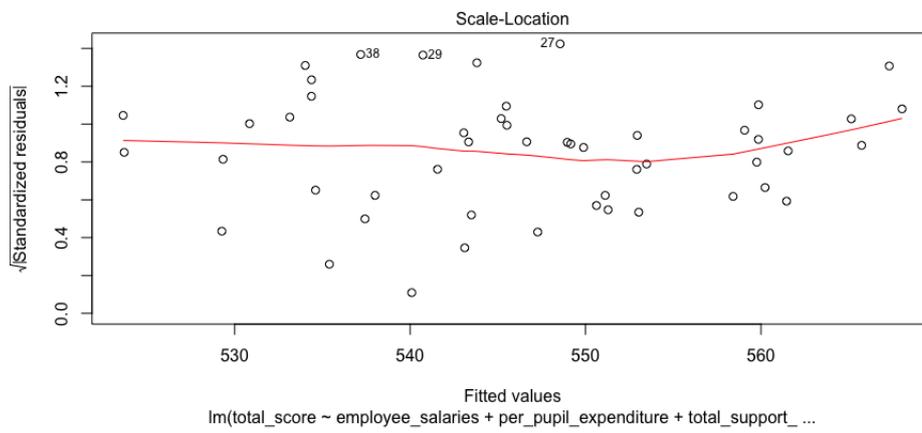


Figure 14: The The Spread-Location plot for all model 3.

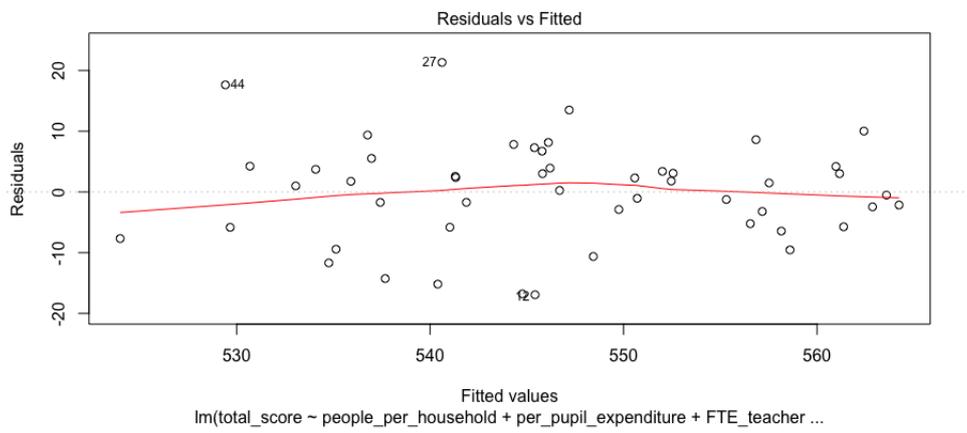


Figure 16: This plot demonstrates the Residual v Fitted values in Model 4.

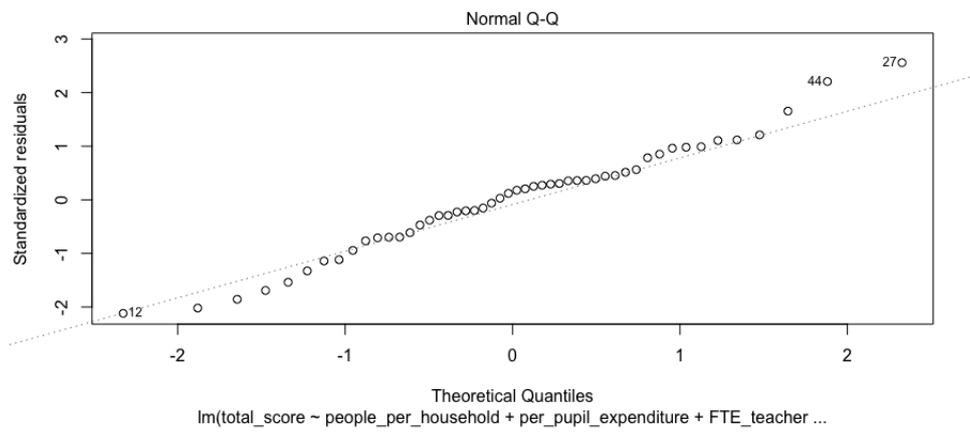


Figure 17: This QQ plot demonstrates normality in Model 4.

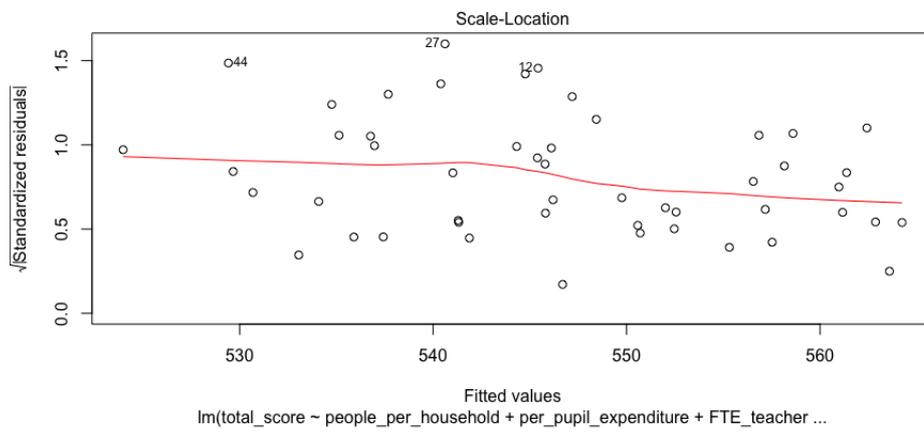


Figure 18: The The Spread-Location plot for all model 4.

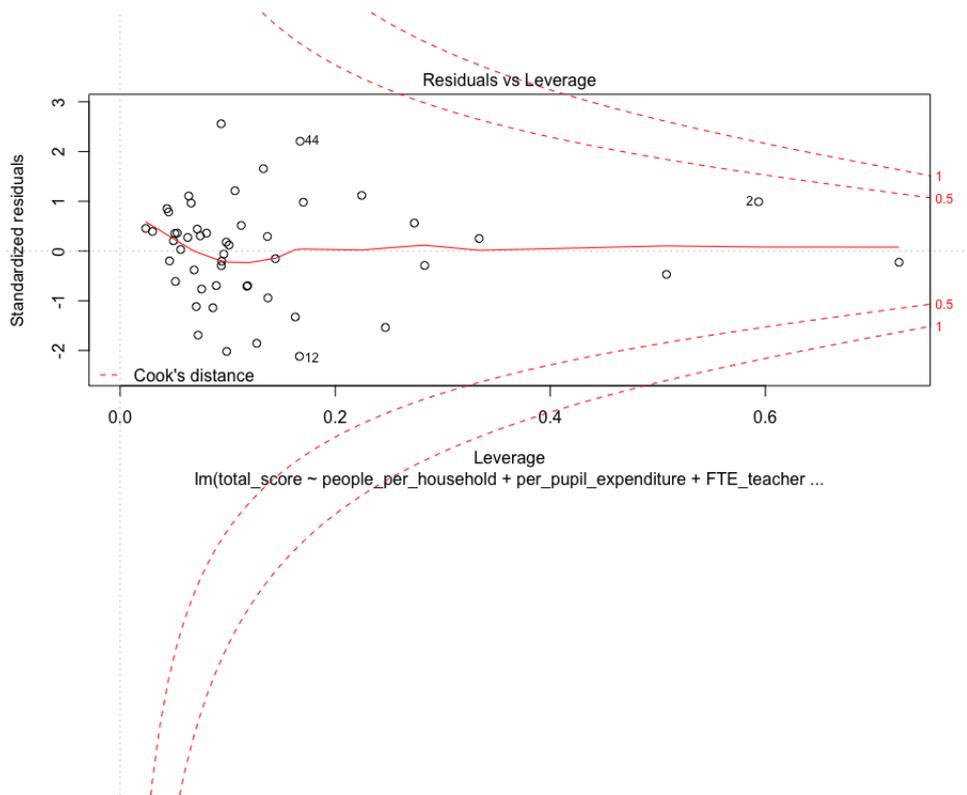


Figure 19: This plot demonstrates Residuals v Leverage in Model 4.

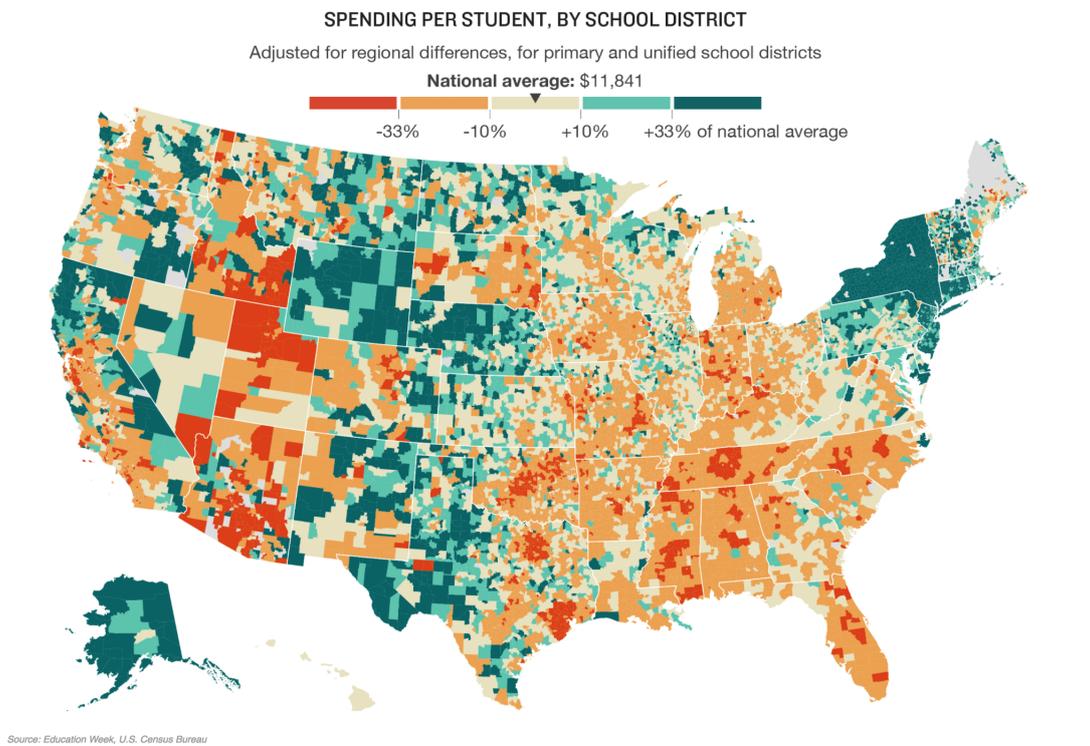


Figure 20: Per student spending by school district

```

> summary(districtstepdown)

Call:
lm(formula = naeptotal ~ percWhite + studentPoverty)

Residuals:
    Min       1Q   Median       3Q      Max
-19.241  -8.786   1.273   9.247  15.720

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  620.7878    28.5944   21.710 9.58e-13 ***
percWhite     0.3689     0.1836    2.009 0.062860 .
studentPoverty -228.9998    47.0723  -4.865 0.000206 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.52 on 15 degrees of freedom
Multiple R-squared:  0.7379, Adjusted R-squared:  0.7029
F-statistic: 21.11 on 2 and 15 DF, p-value: 4.351e-05

```

Figure 21: Per student spending by school district

```

Call:
lm(formula = df_without_state$total_score ~ df_without_state$white +
    df_without_state$people_per_household + df_without_state$poverty_18_and_younger)

Residuals:
    Min       1Q   Median       3Q      Max
-18.4631  -4.3016  -0.0324   4.6795  22.1456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.014e+02  2.245e+01  31.247 < 2e-16 ***
df_without_state$white  8.087e-06  2.965e-06  2.728 0.008996 **
df_without_state$people_per_household -3.116e+01  7.560e+00 -4.122 0.000155 ***
df_without_state$poverty_18_and_younger -1.831e+01  2.322e+00 -7.888 4.33e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.513 on 46 degrees of freedom
Multiple R-squared:  0.6125, Adjusted R-squared:  0.5872
F-statistic: 24.24 on 3 and 46 DF, p-value: 1.478e-09

```

Figure 22: Per student spending by school district