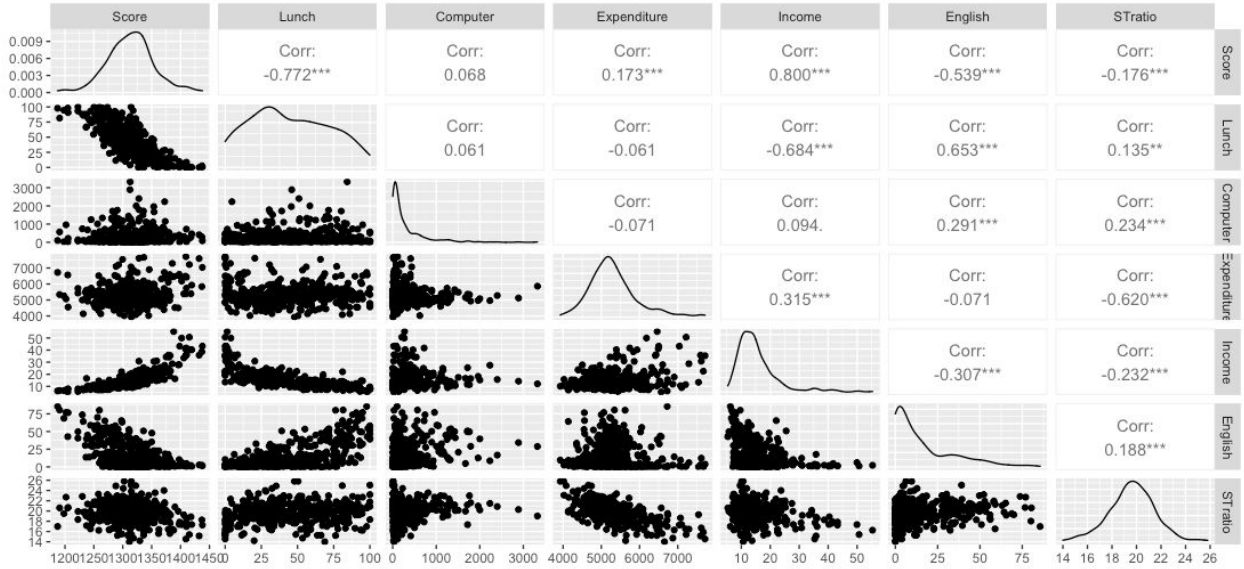
Camilla Handley
Jared Clark

Elementary Education Analysis

The early elementary school years have been cited as some of the most important in a person's education. It is during these years that students develop the skills that will carry them through the rest of their schooling. It has also been noted that children learn much more between Kindergarten and second grade, than they do in subsequent years.

Due to its high level of importance, educators would like to maximize the learning that happens during the early elementary years. The Stanford 9 is a standardized test that attempts to quantify the learning of young students. In this analysis, we will focus on the average Stanford 9 test score for 402 elementary schools. For each school, we have information about the percentage of students qualifying for reduced-price lunch, the number of computers, the expenditure per student, the average income for the district, the percentage of students learning English and finally, the student-to-teacher ratio.



The scatterplots above give an idea about the relationships between the different variables in our data set. Notice that Score appears to have a strong relationship with both the proportion of students qualifying for reduced-price lunch and the average district income. Due to the curvature in the income graph we suspect that there is an issue with nonlinearity, however due to the effects of the other covariates, this cannot be appropriately determined until we fit a preliminary linear model.

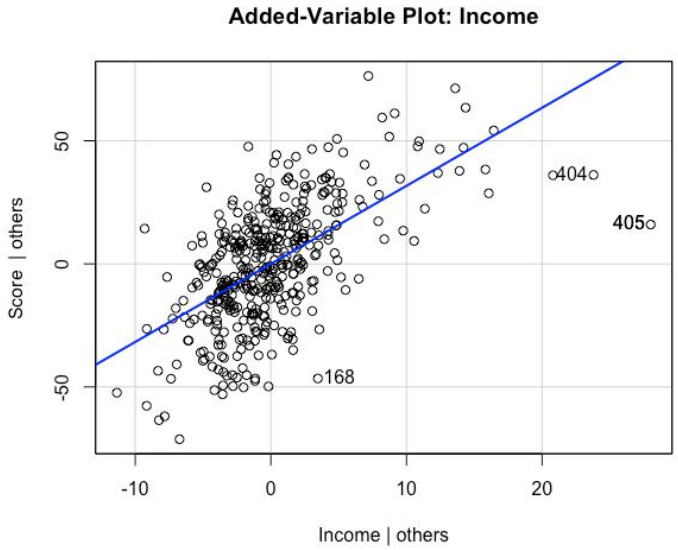
If nonlinearity is unaccounted for, the results of our analysis would be highly misleading. It doesn't make sense to fit a line to the data if we don't believe that there is a linear relationship between the response and explanatory variables. Furthermore, if we were to try and use the model for prediction, we would have difficulties since we would be forced to predict along the fitted line, even if the observations clearly deviate from that line and our results would not be accurate.

In conducting this analysis we hope to understand which factors influence student learning the most in the early elementary years and determine how best to facilitate students. Though prediction of a school's average Stanford 9 score might be interesting, it is not the primary goal with this analysis. Rather we would like to suggest ways that student learning could be improved. Note that the validity of this analysis relies on the notion that Stanford 9 scores are an accurate measure of a student's learning.

We are also interested in learning if there are diminishing returns from the district average income (which determines the school's extracurricular budget) on student learning. That is to say, we are interested in whether the benefit of increasing income decreases at higher income levels. Additionally, we want to know if the need to learn English negatively impacts a student's learning.

Model Specification

We started the analysis by fitting a naive linear model. The added-variable plot below provides evidence of an issue with nonlinearity. At the upper end of the incomes, there seems to be some clear deviation from the fitted line and general curvature.



In order to remedy this issue we employed both polynomial regression and natural splines. Both of these techniques allowed us to fit a linear model with the addition of new covariates (generated from our original data set), which essentially fixed our issues with nonlinearity.

Polynomial regression takes a covariate and adds a quadratic (and possibly cubic, quartic, quintic, …) term to the model. The curvature in scores at the upper end of the recorded incomes suggests that a quadratic term for `Income` might greatly improve the model. Polynomial regression is often useful as polynomials can be extremely flexible when provided with enough

terms. However, there is a notion of diminishing returns as high order terms tend to be fairly collinear. This means that extra care is needed in fitting the model. In some cases restricting the higher order terms to be orthogonal will solve the collinearity issue. Furthermore, polynomial regression often has extreme behavior in the extremes of the data.

In fitting a polynomial regression we added a quadratic term for `Income`. Adding a cubic term for this same variable created an increase in AIC which caused us to settle on only including a quadratic term. We recognize that there are other metrics for assessing model fit but for this analysis, we will utilize AIC. We also saw an improvement in the model by adding a quadratic term for `English`.

Our other method for remedying the nonlinearity issue was natural splines. Natural splines work by fitting cubic functions to different subsets of the data. In this case we were looking at fitting the based on different intervals of the `Income` variable. Cubic splines have the advantage of only trying to fit a curve to a single subset of the data at any given time. This means that it becomes easier to capture local behavior. Natural splines are an extension of cubic splines that provide better behavior in the extremes of the data.

In using natural splines to analyze this data set, by way of cross validation, we found that it was best to use five knots. This essentially means that we fit six cubic functions, broken up by the district average incomes. In assessing this model, we found a fairly good model fit (determined by AIC), however there was still a slight issue with nonlinearity. This led to the model that we actually settled on which was a combination of both the polynomial regression and the natural splines. Although this model may not have the best interpretability (in terms of the `Income` variable), as soon as we started trying to correct for the issues with nonlinearity, we lost some interpretability of the model. When combining the two methods, we found it best to exclude the quadratic term for the `English` variable.

The final model is as follows:

$$Score_i = \beta_0 + \beta_1 Income_i^2 + \beta_{2j} Income_i^* + \beta_3 Lunch + \beta_4 Computer_i + \beta_5 Expenditure_i + \beta_6 English_i + \epsilon_i$$

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

In this model, β_0 is the intercept term, i.e. the expected average Stanford 9 score for a school where all of the explanatory variables are zero. β_1 is the coefficient for the quadratic `Income` term, meaning that this is the expected increase in school average score for an increase of one in `Income` squared. The parameters β_3, β_4, β_5 and β_6 are coefficients associated with the other explanatory variables. For an increase of one in the percentage of students qualifying for reduced-price lunch, we would expect an increase of β_3 in the average Stanford 9 score for the school, holding all else constant.

What has been identified as β_{2j} is actually a collection of six coefficients. The value of this term will change depending on the value of `Income`. `Income*` is a cubic recentering of the `Income` variable which is the application of natural splines in our model. Note that ε_i refers to a residual error for the ith observation.

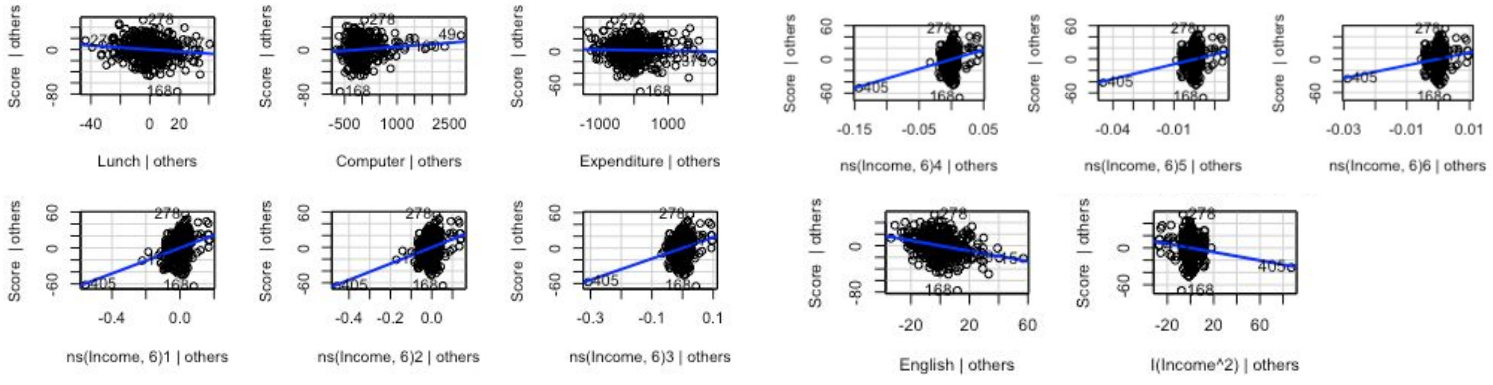
In this model, it is assumed that the residuals are independent and identically distributed. Furthermore, we are assuming that the residuals follow a normal distribution with a common variance, σ^2. These assumptions are important for the validity of the inference that we plan to make. If these conditions do not hold, we can expect standard errors to be incorrect. The issue would extend to both hypothesis testing and the construction of confidence intervals.

We are also assuming that the average Stanford 9 score for a school is a linear combination of our explanatory variables. As mentioned in the introduction, if linearity is not a valid assumption, the accuracy of our analysis will suffer.

Model Justification and Performance Evaluation

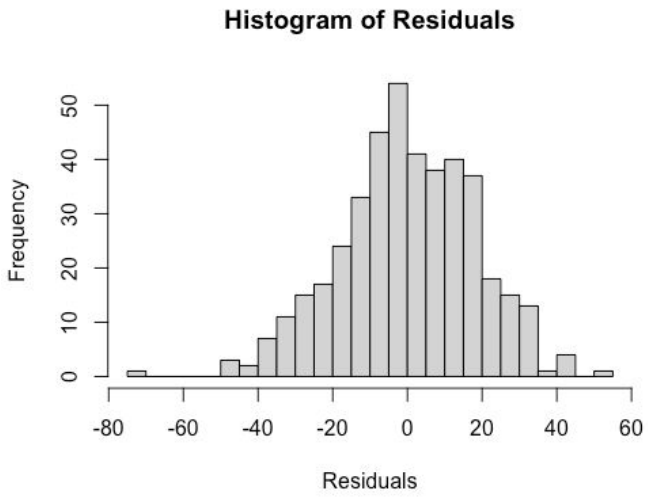
In order to use this model for prediction and/or inference, it needs to meet certain assumptions for linear modeling.

Linearity: The linearity condition applies to the quantitative explanatory variables in the model and their relation to the response variable. To justify this, we will look at the added variable plots seen below and see that there is no obvious curvature. It is important to note that none of the plots for Income have the curvature that was seen before with the naive model.

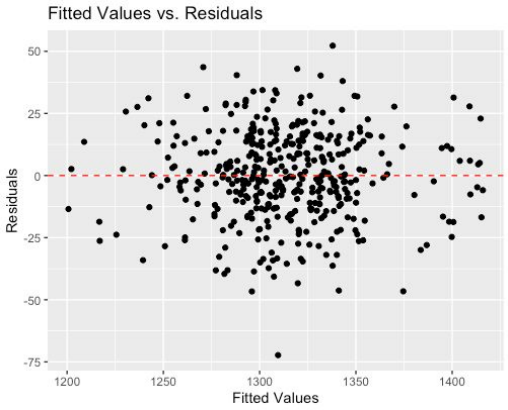


Independence: We assume that the observations are independent from one another. We are told that this sample is from various school districts in California and will assume that they were randomly sampled.

Normality: The distribution of the model residuals must be approximately normal. Below is a histogram of the residuals showing approximate normality.



Equal Variance: The spread of the residuals must be approximately constant across its fitted values, or predictions (also known as homoscedasticity). The plot of fitted values vs. residuals below demonstrates this. Since we are looking at averages, the variance could be changing with class size. For our analysis we will assume this is not an issue, but this could be an interesting point for future research.

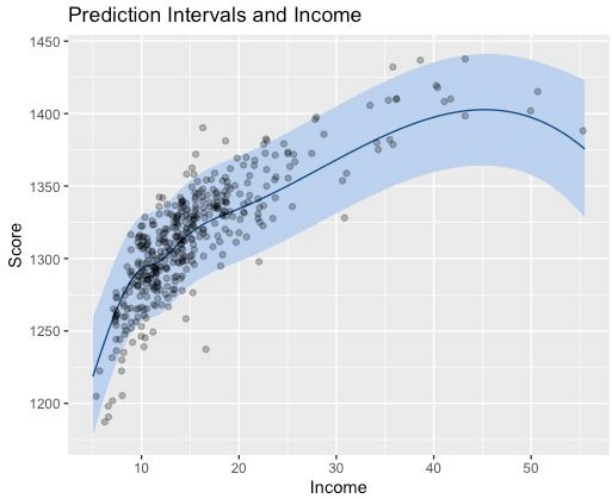


With all the assumptions met, we will continue with our analysis using this final model. We utilized the best subset selection with AIC as the criteria and decided to take out the Student-Teacher Ratio variable from the models because it does not improve the fit when included. Our model has an AIC of 3647.919, the smallest out of all the models that we fit. It has an R-squared of 0.8019, indicating that 80% of the variation in `Score` is explained by the covariates in the model. We see this as an adequate model fit. Using leave-one-out cross validation, we also evaluated the predictive capability of the model. This resulted in an average Root Mean Square Error of 14.661, which is very good compared to the standard deviation of `Score`, which is 40.58. Thus we are generally predicting well within a standard deviation. The CV also provided a very low bias of -0.0067 and a coverage of 95.7%. Thus our model can

accurately predict. While prediction isn't our primary concern, it is interesting to note that this model has some level of predictive capability.

Results

In performing this analysis, we specifically wanted to know if there is evidence of diminishing returns on extracurricular activities in terms of student learning, given that district average income is generally a measure of the school's extracurricular activity budget. To answer this question, we used our model to predict `Score` holding all covariates constant at their average except for `Income`, which we tested at equally-spaced amounts within the range of the data. The resulting graph can be seen below with the prediction intervals and estimates in blue and the black points representing the observed data.



This graph implies that there is evidence of diminishing returns because after ~45 (thousand) `Income`, the predicted average Stanford 9 score starts to decrease rather than increase. We also wanted to know if English as a second language is a barrier to learning. To answer this question we will look at the estimated coefficients from the model. These are shown in the table below.

Covariate	**95% Confidence Interval (final model)**
Lunch	(-.319, -0.0446)
Computer	(0.00036, 0.00927)
Expenditure	(-0.0041, 0.00209)
English	(-0.6001, -0.2975)

ns(Income)1	(0.7454, 0.01388)
ns(Income)2	(0.9631, 0.0179)
ns(Income)3	(0.01163, 0.02429)
ns(Income)4	(0.0187, 0.04935)
ns(Income)5	(0.04187, 0.00134)
ns(Income)6	(0.04371, 0.00185)
Income^2	(-0.5817, -0.0995)

We are 95% confident that as the percent of English learners in a class goes up by one, we expect `Score` will go down by between -0.6 and -0.3 on average, holding all else constant. This implies that English as a second language does create some kind of barrier to learning, as measured by the Stanford 9 standardized test. All other coefficients can be interpreted in the same way as `English` except `Income`, which we interpreted previously with the graph. For example, as the number of computers in a class goes up by one, we expect the mean score to increase by between 0.0004 and 0.009 on average, holding all else constant. Because this interval is strictly positive, we would suggest that increasing the number of computers in a classroom is one thing that can be done to increase student learning. It appears that the percent of students that qualify for reduced price lunch has a negative effect on learning. This is not an aspect that is easily changed or controlled because it has to do more with the socio-economic status of the students, which is controlled by factors outside of the classroom. However, if there is a way to improve the financial situations of the students' families with some kind of employment or self-reliance program, this would be a possible way to increase student learning.

For sake of comparison, we will also list the confidence intervals from the polynomial model in the table below. This model produces similar results, with `Lunch` and `English` having generally negative effects on learning and `Computer` and `Income` with generally positive effects. Also similar to our final model, the significance of `Expenditure` is questionable because the interval contains zero.

Covariate	**95% Confidence Intervals (polynomial model)**
Lunch	(-0.7585, -0.5343)
Computer	(0.0064, 0.01597)
Expenditure	(-0.00239, 0.0046)
English^2	(-0.0104, -0.0058)

Income^2	(0.03713, 0.0525)

Conclusion

Using the linear model previously described, we were able to adequately meet the goals of our study. We determined that there is evidence of diminishing returns of Income on student learning and that being an English learner is a barrier to learning. Based on our results we would suggest increasing the number of computers in a classroom and improving the socio-economic situation of the students' families in order to increase student learning. We acknowledge that there are other things that could help improve student learning that were not measured in this data. One variable that might be interesting to include would be school size or class size. As mentioned before, the school size might also influence the variability of average test scores.

This model was a good fit for the data and was able to accurately explain the variability in average Learning Score. However, since this approach to modeling is very subjective, there are many other combinations of methods that could be used to model this data. We also lost some interpretability of the effect of `Income` due to its non-linear nature. Our model was satisfactory for this analysis, but we acknowledge that it may not be the best fit possible.

Teamwork

Jared did the coding for the polynomial regression as well as the writing of the Introduction and Method sections.

Camilla coded the natural splines model. She also wrote the Model Justification, Results and Conclusion sections.

We both co-edited the paper and chose our final model together.