Nonparametric Methods — Final Project Report

Tyler Maule

June 7th, 2022

Introduction & Project Goals

For policymakers aiming to improve the American education system, good policy development requires a nuanced understanding of which schools require more investment. As the College Board's SAT remains a significant determinant in how high school students gain admission to colleges and universities, SAT results still tie into school-level disparities in educational opportunities.

For this project, I aim to aid policymakers by identifying the best predictor of SAT critical reading scores at the school level (hereafter sometimes referred to as "SAT scores" for short). I will leverage 2012 demographic and SAT score data from New York City, NY schools to create multiple nonparametric regression models and compare their properties.

Data

To compile the complete data set of interest, I drew school-level demographic data on NYC schools from 2012. The City of New York provides this data at the NYC Open Data site through the 2006 - 2012 School Demographics and Accountability Snapshot. Their data set includes a school database number identifier, school names, and demographic indicators including the proportion of students at each school who are white, who are female, whose families live below the poverty line, who are reported to have a disability, and who do not speak English as a first language. Simple enrollment counts for each grade are also included. While the data provided includes elementary and middle schools, only schools that teach the 12th grade are relevant in this case.

From the same website I downloaded aggregate school-level SAT data on 12th grade students who took the SAT in 2012, listed as 2012 SAT Results. This data set includes the number of students who took the SAT at each school, their average critical reading score, their average writing score, and their average mathematics score. Note that average writing scores and average mathematics scores had a high positive Pearson correlation with average critical reading score (of 0.88 and 0.97, respectively).

The two datasets were joined on school database number such that each school with SAT data had that data augmented with demographic data from the year 2012. This joined dataset formed the basis for analysis in this report.

Results

Understanding the distribution of average SAT critical reading scores allows for a better approach to this prediction problem. Using the `GoFKernel` package, I estimated the probability density function of the critical reading scores. Below, see the KDE result when using four different methods for selecting the kernel bandwidth: `nrd` (based on the estimated standard deviation of the distribution using the empirical IQR), `nrd0` (an adjusted standard deviation estimate), `SJ` (the Sheather-Jones method), and `ucv` (the cross-validation method). Selected bandwidths are 12.25896, 10.40855, 9.882592, and 10.16906, respectively; as the underlying data has a range of 400, there appears to be little difference in the KDE using each bandwidth selection method. The fifth image plots three KDEs with bandwidth-selection methods `nrd0`, `SJ`, and `ucv` atop one another.

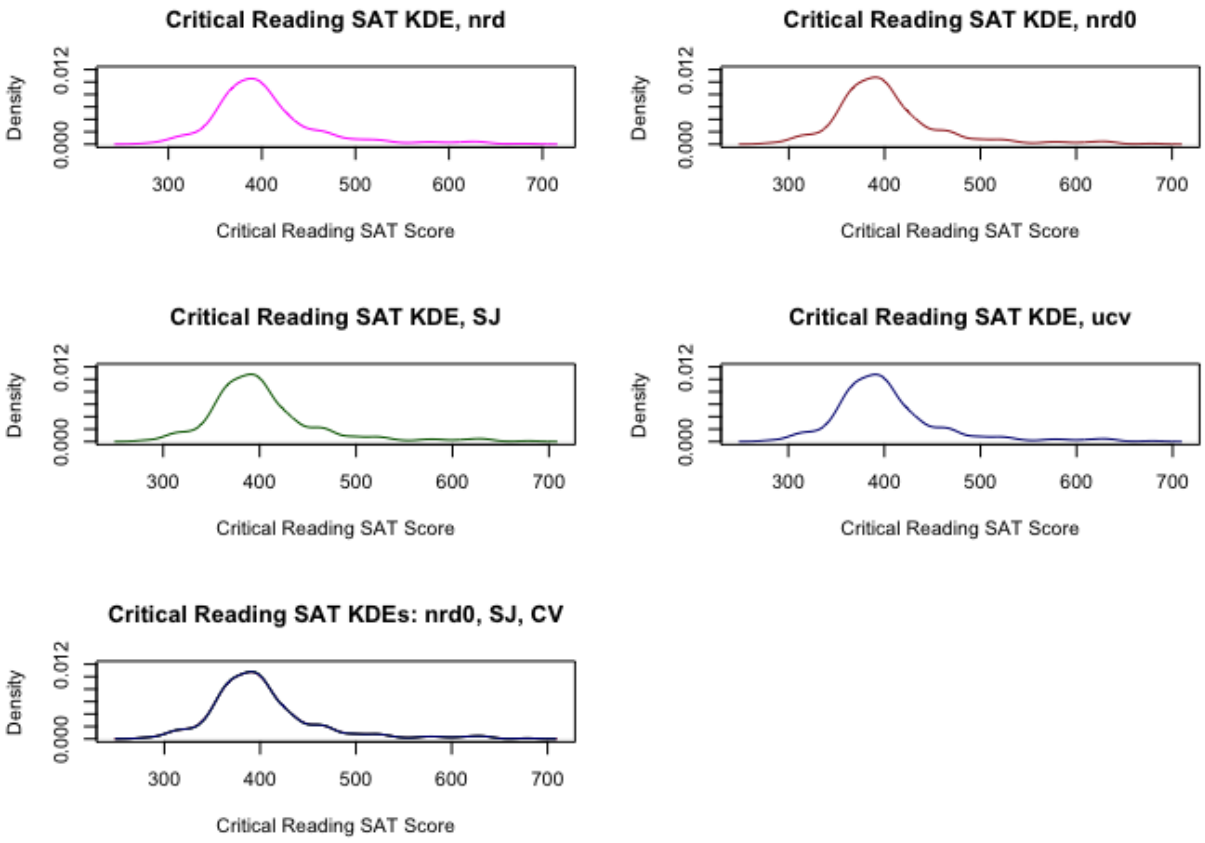


Figure 1: "Four KDEs for Average SAT Critical Reading Score"

As evidenced by the fifth image in Figure 1, the kernel density estimates based on different methods do not differ significantly. In each case, we note a mode around a score of 390.7632, with a shorter left tail extending below the observed minimum average score of 279 to around 220 and a long right tail that extends beyond the maximum observed mean score of 679 to a little over 700. While the distribution generally looks asymetrically bell-shaped, there are small increases and decreases in density around 320, 460, 520, 580, 620, and 680. Thus while the density is quite smooth overall, there are a few "bumps" beyond the overall mode.

To determine which covariates are of interest and serve as good predictors of average SAT critical reading scores within schools, we examined a simple pairplot and correlation matrix, excluded here for the sake of consision. Based on this exploratory data analysis, we selected three predictors of interest: the percent of students at the given school who are recorded as white (not hispanic or latino), the percent of students who are English language learners, and the percentage of students who come from families that fall below the poverty line.

We gain a greater depth of understanding by developing bivariate KDEs of the predictors of interest alongside the response variable. See below subplots of perspective plots for each predictor, with optimal bandwidth determined by the bivariate normal reference approximation method (plots in the left column) and the cross-validation method (plots in the right column).

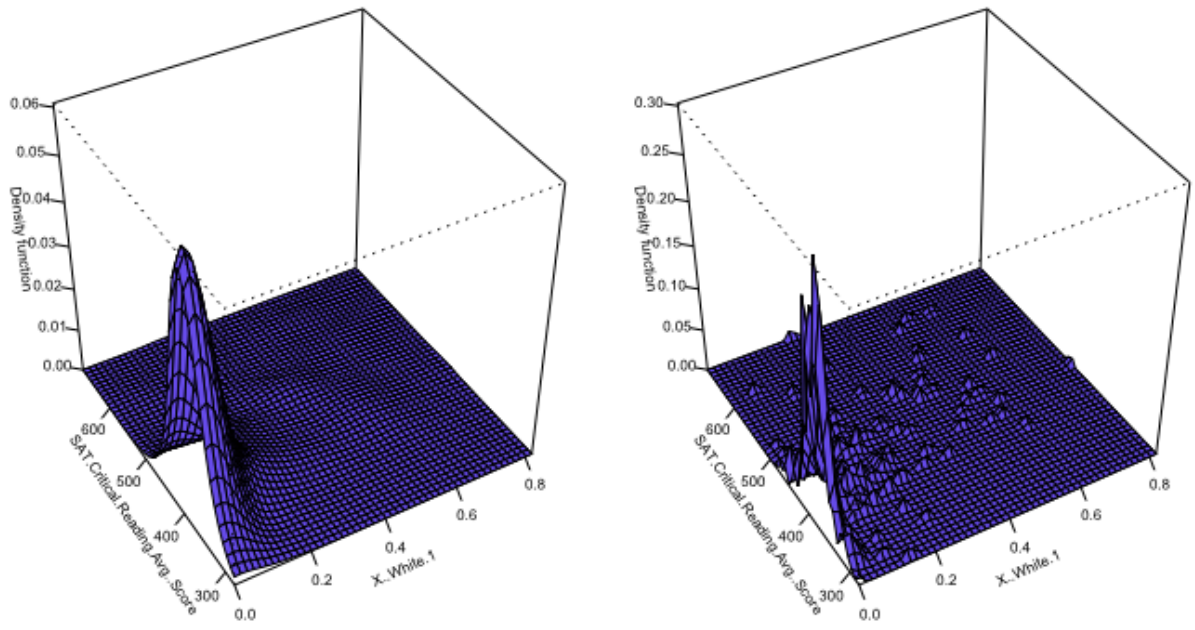


Figure 2: Bivariate KDEs with SAT Scores and Percent White

In the above bivariate KDE of percent of students who are white versus the average SAT critical reading score, the cross-validation method seems more illuminating—while many schools have a modest proportion of students who are white and the SAT score density appears the same as in the univariate case of viewing only the SAT score KDE, note that there are a fair number of "spikes" in density where the percentage of students who are white is higher. In these cases, there appears to be a positive quadratic relationship between the

percentage of students who are white and the mode of the SAT score density at that given demographic percentage.

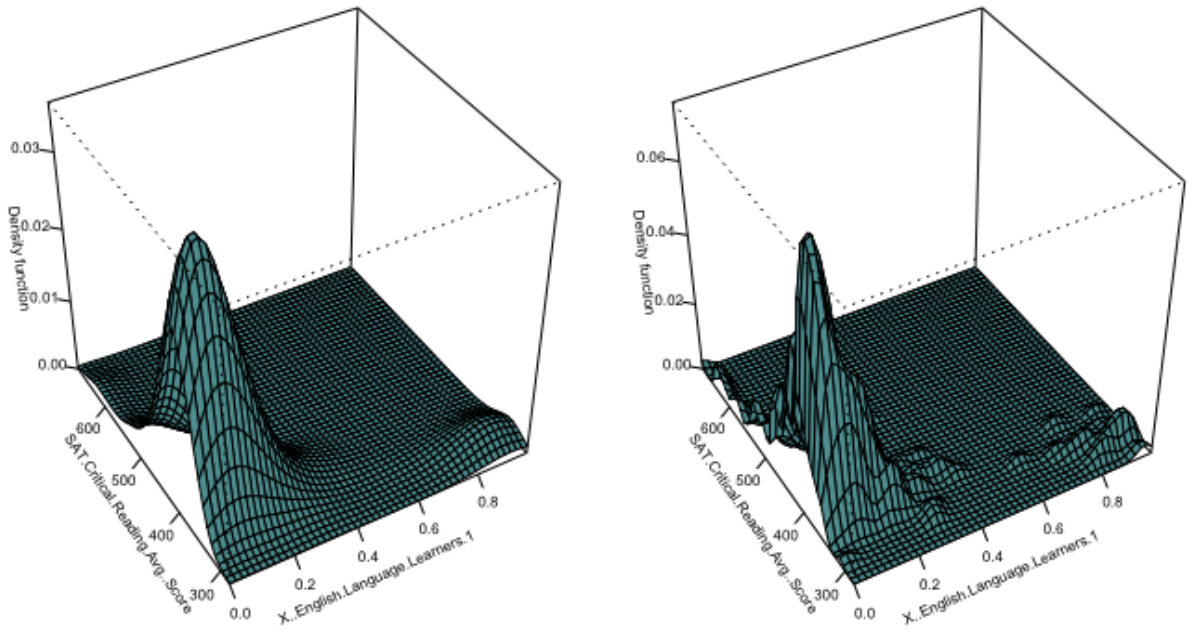


Figure 3: Bivariate KDEs with SAT Scores and Percent English Learners

For the case of proportion of students who are English language learners, both the normal reference approximation method and cross-validation method yield interesting results. It seems like although the majority of schools have a low proportion of English language learners, as that proportion increases the modal SAT score decreases at the given proportion. In addition, there is a moderately significant density of schools with over 80% of students who are English language learners, where the modal SAT score is close to the minimum average SAT score.

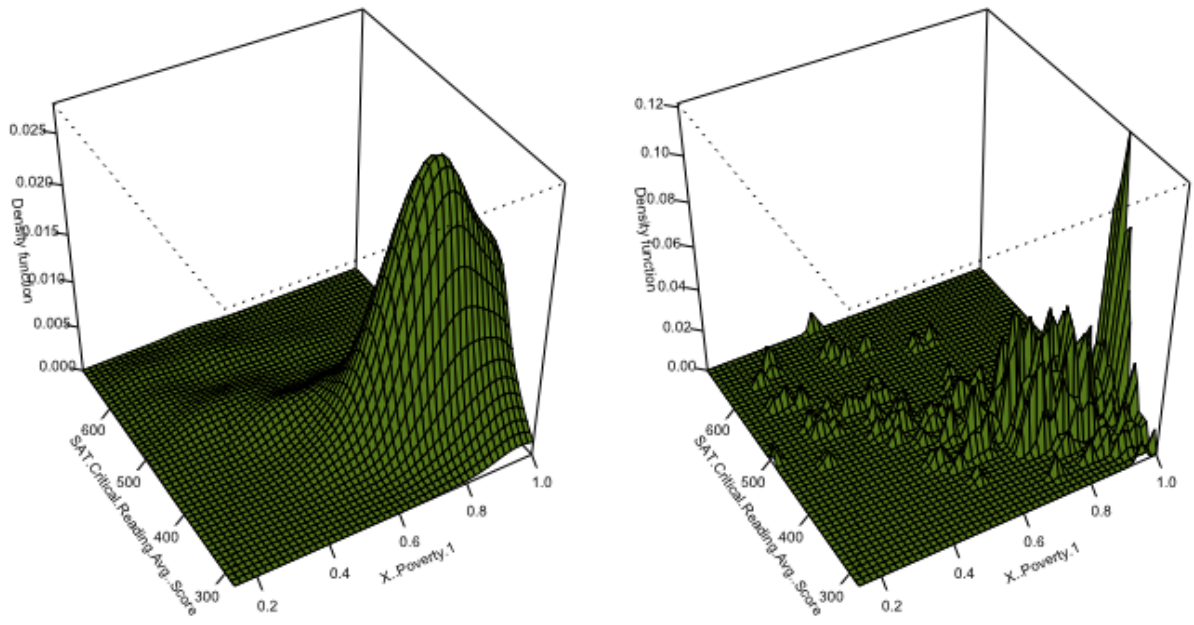


Figure 4: Bivariate KDEs with SAT Scores and Percent in Poverty

The bivariate KDE of percentage of students from families in poverty and school average SAT critical reading scores tells a different story. In this case, the bandwidth chosen by cross-validation appears to overfit the data, so we interpret the bivariate KDE with bandwidth chosen by the normal reference method instead. The majority of schools have more than 60% of students hailing from families below the poverty line, with a mode of around 0.8. While there are some schools with 40%-80% of students below the poverty line that have modal SAT scores above 550, in the majority of cases there seems to be a decreasing quadratic relationship between percentage of students from families in poverty and the average SAT score.

With exploratory data analysis complete, we turn to estimating the models of interest with nonparametric regression. Since each model includes a single covariate, both local linear regression and local constant regression are employed in every case.

See below the plot associated with local linear regression of the percentage of students who are white versus the average SAT score. The selected bandwidth is 0.1652871, the associated degrees of freedom is 3.56, and the second-differences estimate of standard deviation is 47.49857. We also evaluate a hypothesis test that a linear function explains the relationship between the variables, yielding a p-value of 0. Thus at the $\alpha = 0.05$ level, we reject null hypothesis that a linear relationship is appropriate. The reference bands in the visualization support this conclusion. In addition, we evaluate the predicted average critical reading SAT score for schools with 0%, 25%, 50%, 75% and 100% white students—the predictions are 376.3653, 460.0364, 506.8689, 504.8451, and 436.4142, respectively.

For local constant regression with the same variables, the selected bandwidth is 0.02941858, the associated degrees of freedom is 11.5, and the second-differences estimate of standard deviation is 47.49857. Here too, we evaluate a hypothesis test that a linear function explains the relationship between the variables, yielding a p-value of 0. Again, at the $\alpha = 0.05$ level, we reject null hypothesis that a linear relationship is appropriate.

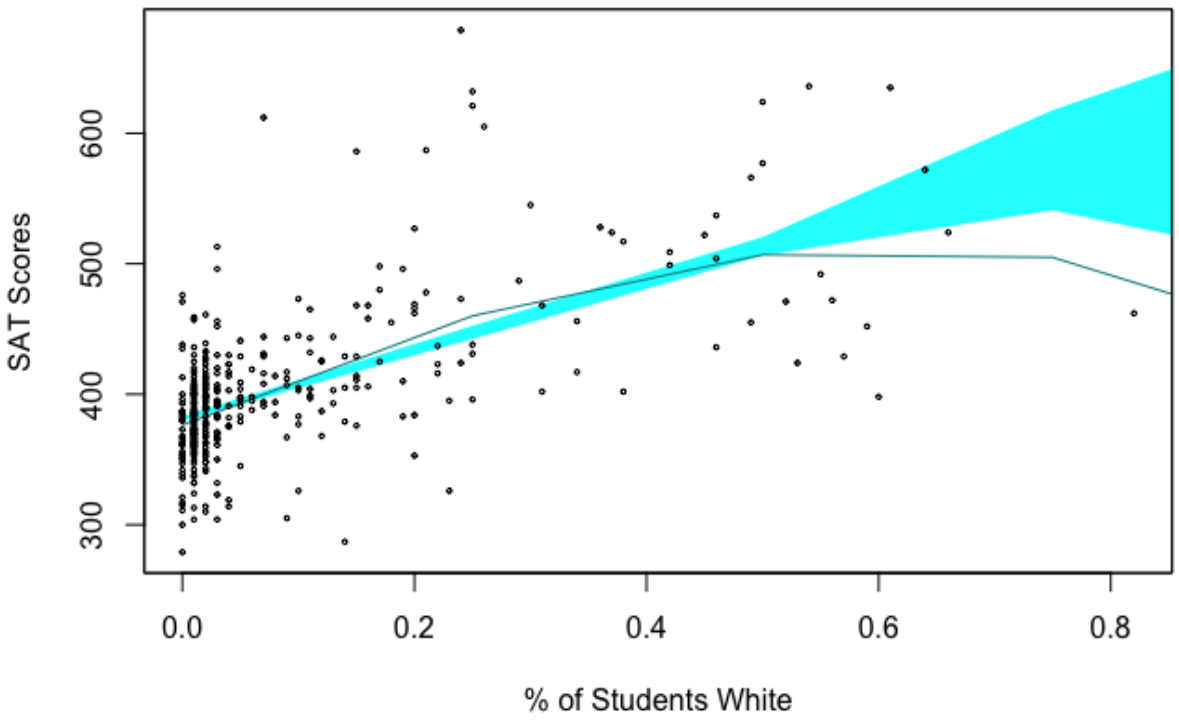


Figure 5: Local Linear Regression with SAT Scores and Percent White

See below the relevant plot with reference bands associated with the null hypothesis of a linear relationship. Once more, we evaluate the predicted average critical reading SAT score for schools with 0%, 25%, 50%, 75% and 100% white students—the predictions are 380.8931, 483.1420, 525.4038, 471.8111, and 462.0000, respectively.

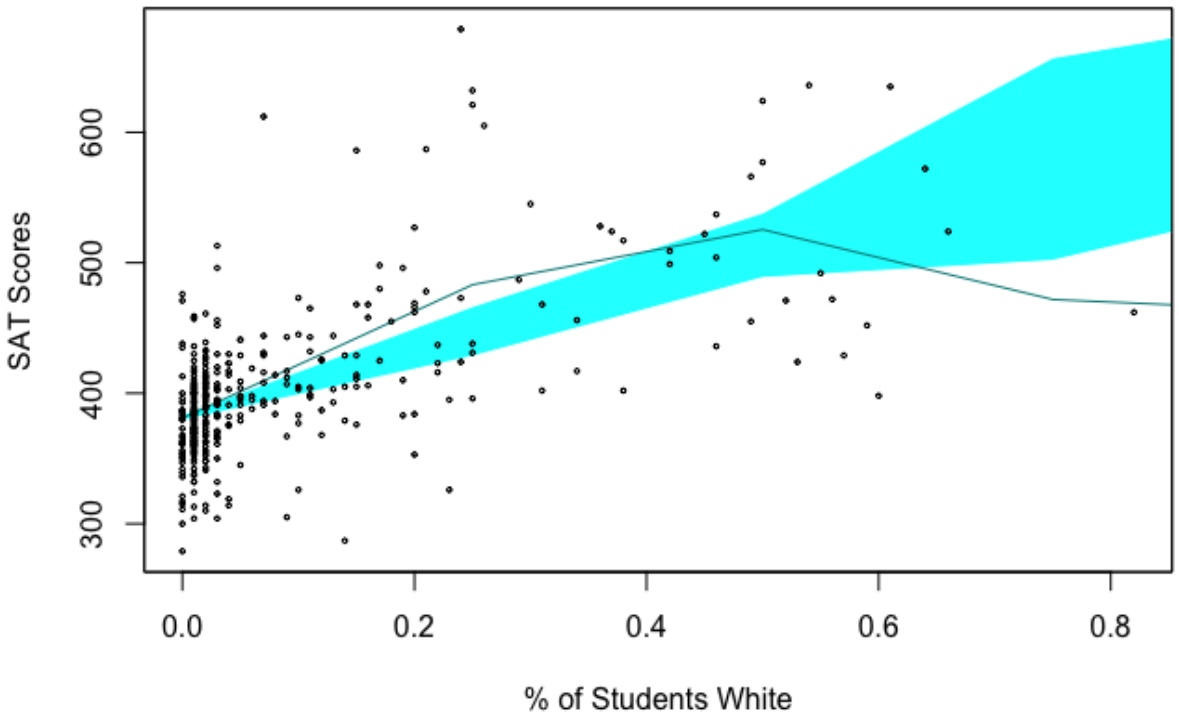


Figure 6: Local Constant Regression with SAT Scores and Percent White

While both nonparametric regression methods result in estimates that could not be adequately represented by simple linear regression, it appears that a second-order or third-order linear regression model might be appropriate in this case. The estimated average SAT critical reading score reaches its peak at schools where around 50% of students are white, and gradually decreases when moving away from that peak. Unsurprisingly, local constant regression seems to overfit the data at schools with higher proportions of students who are white, so I'd stick with the smoother fit given by local linear regression.

Next, we inspect a local linear regression model of the percentage of students who are English language learners versus the mean SAT critical reading score, with reference bands pertaining to a linear relationship. Here our selected bandwidth is 0.01348127, the associated degrees of freedom is 22.69, and the second-differences estimate of standard deviation is 35.17191. The hypothesis test of a linear relationship leads to a p-value of 0. Once more, at the $\alpha = 0.05$ level, we reject null hypothesis that a linear relationship is appropriate. We also calculate the predicted average critical reading SAT score for schools with 0%, 25%, 50%, 75% and 100% students who are English language learners—the predictions are 520.274, 376.3010, 341.0008, 331.9267, and 370.3901, respectively.

With local constant regression of the percentage of students who are English language learners versus the mean SAT critical reading score, we find that a bandwidth of 0.01283536 is selected, degrees of freedom are 386, and the second-differences estimate of standard deviation is 35.17191. Predictably, with a p-value of

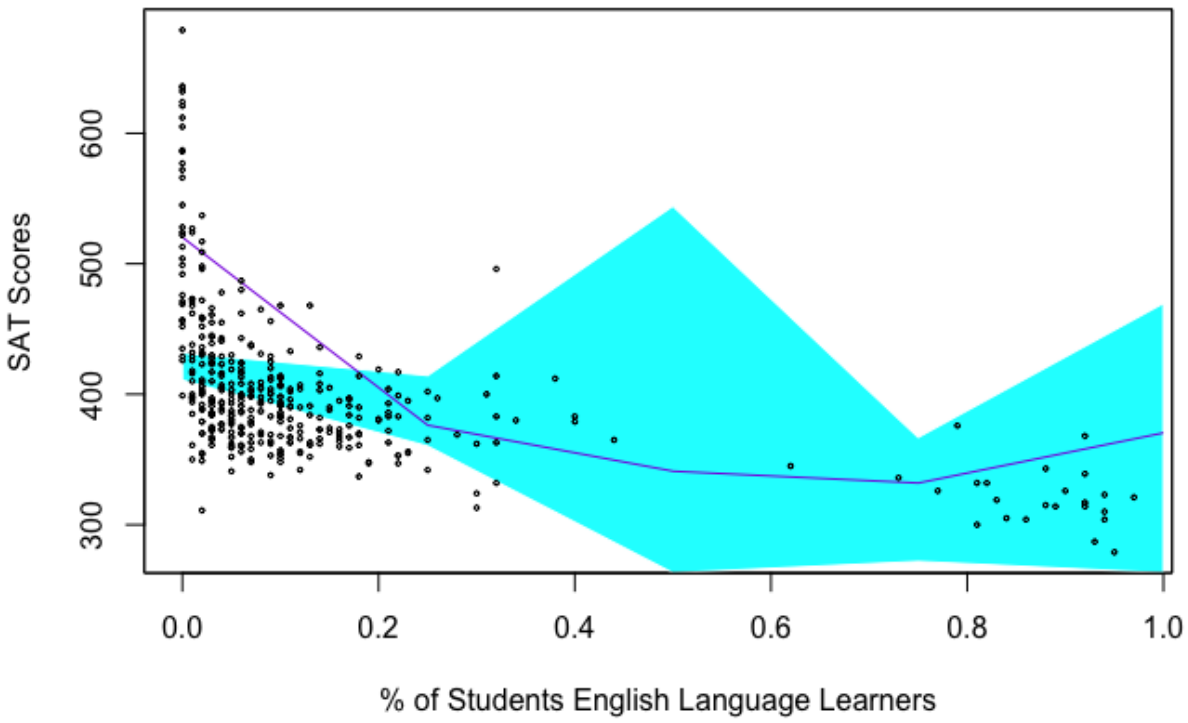
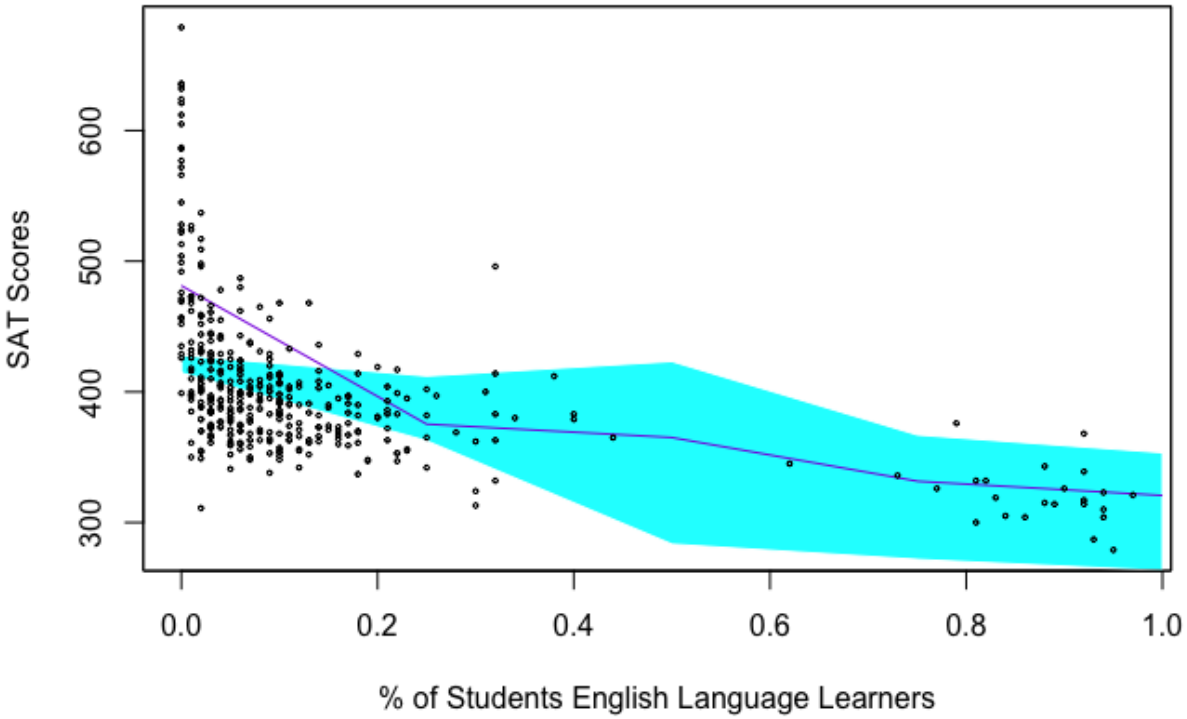


Figure 7: Local Linear Regression with SAT Scores and Percent English Learners

0, the null hypothesis of a linear relationship is again rejected at the $\alpha = 0.05$ level. Now we calculate the predicted average critical reading SAT score for schools with 0%, 25%, 50%, 75% and 100% students who are English language learners—the predictions here are 481.2353, 375.2767, 365.0000, 331.5810, and 320.6687, respectively.



Unsurprisingly, both estimated models predict that maximum average scores will occur at schools with a very small number of English language learners (<5%). Local linear regression predicts a slight upswing in average SAT score in schools when >80% of students are English language learners, whereas local constant regression largely predicts a decreasing relationship overall. Again, to minimize overfitting I'd move forward with the local linear regression model.

Finally, we inspect a local linear regression model of the percentage of students from families below the poverty line as a predictor for the mean SAT score. In this case, the bandwidth determined by cross-validation is $h = 0.1677681$, the corresponding degrees of freedom is 3.58, and the standard deviation as determined by second differences estimation is $\sigma = 39.12624$. Yet again, a p-value of 0 leads us to reject the null hypothesis of a linear relationship at the $\alpha = 0.05$ level. Once more we calculate the predicted average critical reading SAT score for schools with 0%, 25%, 50%, 75% and 100% students who are from families that fall below the poverty line—the predictions in this case are 549.1568, 561.8644, 485.5269, 406.7600, and 370.3611, respectively.

Our last model uses a local constant regression of the percentage of students from families below the poverty line to predict schools' mean SAT score. For this final model, the bandwidth determined by cross-validation is $h = 0.01608273$, the corresponding degrees of freedom is 23.9, and the standard deviation as determined by second differences estimation is $\sigma = 57.43459$. Yet again, with a p-value of 0, we reject the null hypothesis of a linear relationship at the $\alpha = 0.05$ level. For the final time we calculate predicted average critical reading SAT scores, this time for schools with 0%, 25%, 50%, 75% and 100% students who are from families that fall below the poverty line—the predictions in this case are 504.8525, 553.5376, 478.0418, 398.9329, and 375.1774,

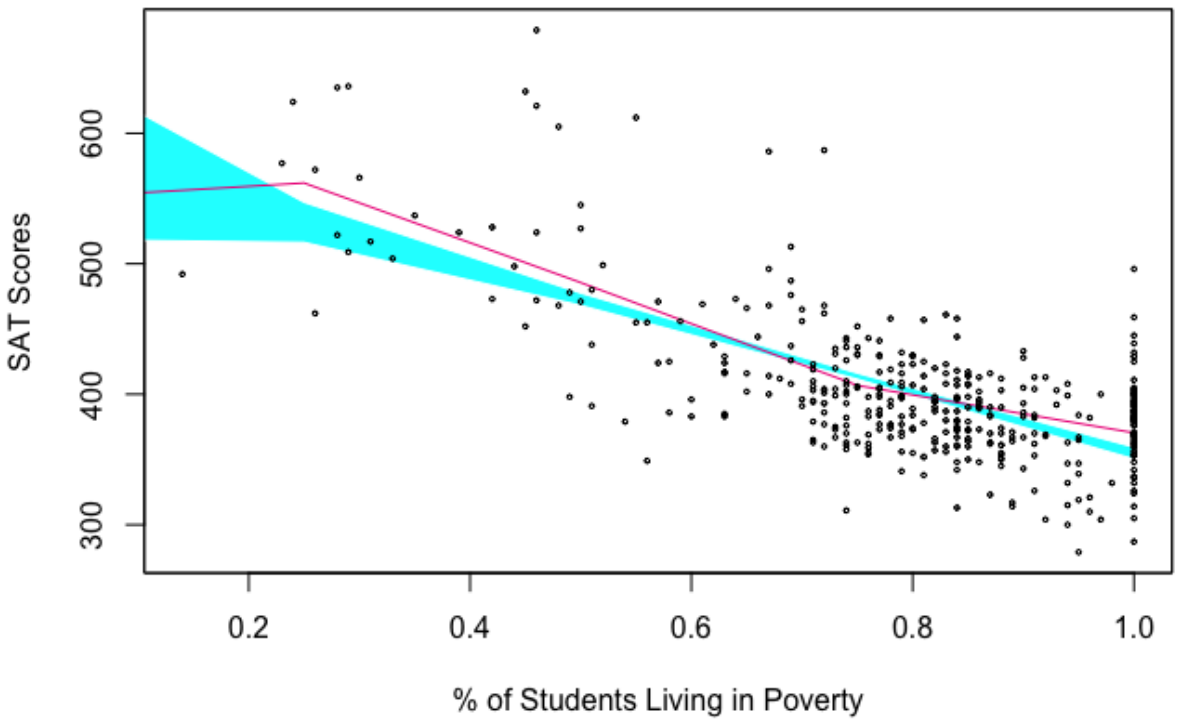
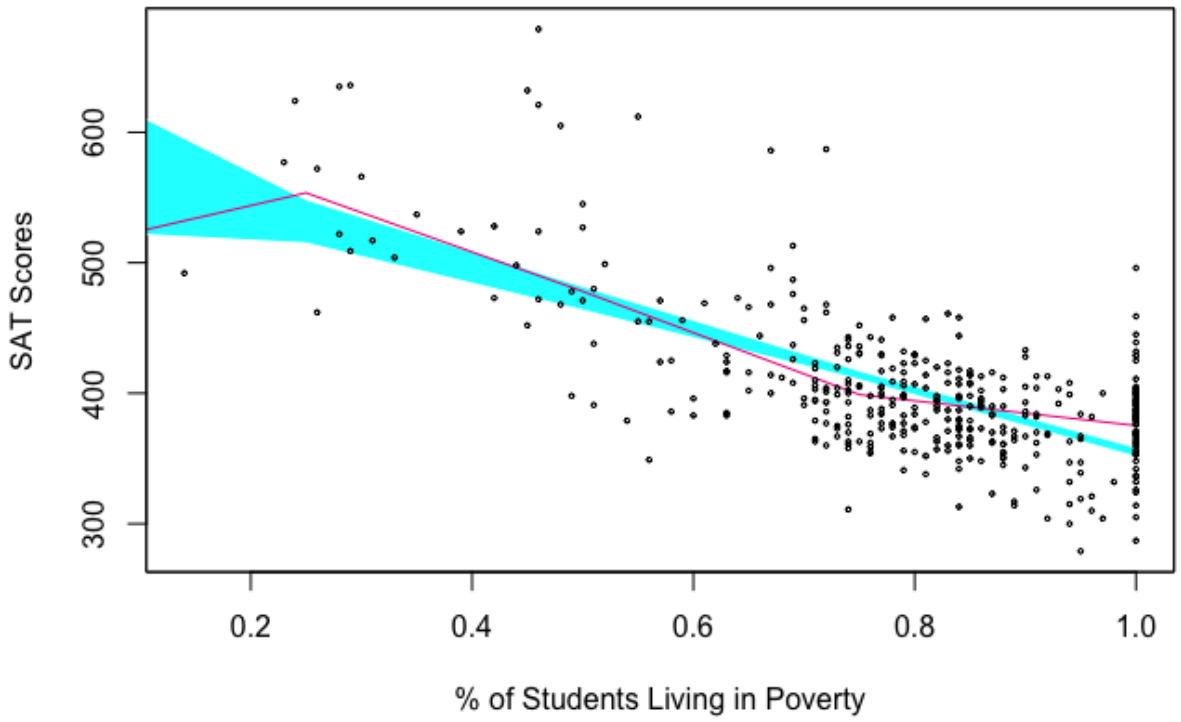


Figure 8: Local Linear Regression with SAT Scores and Percent in Poverty

respectively.



There seems to be little difference in the results when using local linear versus local constant regression to predict SAT score based on the proportion of students living in poverty. In both cases, there is a strictly decreasing relationship between proportion of students living in poverty and scores after the 30% poverty mark. There seems to be a slight leveling out in the rate of decline around the 75% poverty mark.

Discussion

To effectively compare the predictors' performance, the models' success, and the nonparametric methods' utility compared to parametric models, we include below a table of relevant metrics for each predictor.

	Prop. of Students White	Prop. of Students Learning English	Prop. of Students in Poverty
Local.Linear.h	0.165	0.013	0.168
Local.Linear.df	3.560	22.690	3.580
Local.Linear.sigma	47.499	35.172	39.126
Local.Linear.pval	0.000	0.000	0.000
Local.constant.h	0.029	0.013	0.078
Local.constant.df	11.500	386.000	6.070
Local.constant.sigma	47.499	35.172	39.126
Local.constant.pval	0.000	0.000	0.000

Note that in every case, we reject the null hypothesis that the predictor-score relationship is adequately represented by a simple linear regression model (at the $\alpha = 0.05$ level). Naturally, bandwidths are smaller

and degrees of freedom larger for a given predictor's local constant regression model relative to its local linear regression model. Overall, models have the highest bandwidths when using the proportion of students in poverty as the predictor, followed by the proportion of students who are white and the proportion of students who are English language learners. More importantly, see that the second-differences estimate of sigma is largest when using proportion of students who are white as the predictor, next largest with the proportion of students in poverty, and smallest when using the proportion of students learning English. Based on those results, if I needed to use a single predictor and model to predict schools' average SAT critical reading scores, I'd use a local linear regression model with the proportion of students learning English as the predictor.

So overall, proportion of students learning English best explains variation in the SAT critical reading score at the school level. While there are some exceptions, an increase in the proportion of students learning English at a given school is associated with a decrease in the SAT score. Based upon this limited analysis, I would advise policymakers to focus on English language learners at schools and to prioritize increasing funding for schools with high proportions of English language learners if they wish to maximize average SAT critical reading scores at the school level.