Evaluating the Impact of School Environment on Student Academic Success

Introduction

Disparities in academic success levels between schools are viewed in public and policy discourse through district-level socioeconomic inequalities and funding disparities. These aspects tend to dominate discussions about educational inequities. However, while they are critical issues, they can overshadow more localized, medium-term factors that might influence academic outcomes within individual schools.

In this study, we aim to shift the focus toward these school-level factors, exploring how the internal dynamics of schools might correlate with student academic success.

How much of an effect can school environment alone have on student academic success, regardless of other factors (e.g. socio-economic profile)?

The importance of the question arises from the fact that it focuses on what has the potential to be a more malleable condition to the schools and still have a significant effect on students' performance. It focuses on localized factors within the school's control which could be more easily targeted for improvement than broader systemic issues facing the education sector.

Objective

To evaluate the extent to which the school environment influences student academic success, isolating its impact from external factors such as socio-economic background.

Sub-objective

To identify the specific aspects or elements of the school environment and dynamics that play the most significant role in shaping student academic outcomes.

Material and Methods

Datasets

Our focus will be on data of New York City Public Elementary Schools for the two academic years 2018-2019, 2019-2020. We investigate the relationship between school environment metrics, derived from survey responses, and academic performance, proxied by mathematics test scores.

The datasets used are all extracted from the NYC Open Data platform and the official website of the NYC Department of Education. They are as follows:

- 2017-2018 2021-2022 Demographic Snapshot
- Math Test Results 2013-2023
- NYC School Survey for the academic years 2017-2018, 2018-2019, 2019-2020.

The Demographic Snapshot includes school level data about the number of students enrolled in the school by grade (from 3k up to 12th grade), as well as proportions of different

ethnic groups, gender, disability, and economic groups among students for each academic year.[1]

The Math Test Results dataset includes school-level results for the New York State Math exams for each academic year across 4 levels of performance for the exam.[2]

The New York City Survey data includes parents, students, and teacher's responses at the school level to the survey questions, in addition to the 6 key elements used by the department to score schools on a scale of 1.00 to 4.99[3]:

- Rigorous Instruction;
- Collaborative Teachers;
- Supportive Environment;
- Effective School Leadership;
- Strong Family-Community Ties, and
- Trust.[4]

Limited Timeframe

The analysis is limited to the specified academic years due to the unavailability of consistent survey data and scoring methodology for a broader time range. On one hand, the newly developed "Framework for Great Schools" had been adopted as recently as 2016, making the first available rating for the 2017-2018 school year. On the other hand, the COVID-19 pandemic has led to a pause in these ratings from the 2020-2021 school year because of the inadequacy of the "elements" under the conditions of teaching brought about by the pandemic (e.g. measuring absenteeism).

We then discount the 2017-2018 school due to the difference in scoring for the New York State Math exams compared to 2018-2019 and 2019-2020 academic school years.

Our analysis is based on the data of 475 schools across two academic years. Given the limited timeframe of only two years, we chose not to consider our data as time series data.

Statistical Methods

We used the following statistical models to analyze the data:

Ordinary Least Squares Regression: We used OLS as our initial model to examine the relationship of school environment variables and socioeconomic variables vs academic performance. We also validated our assumptions (stated below), and based on violations to our assumptions, we used additional models that could serve as a better fit for the data.

---

[1] https://data.cityofnewyork.us/Education/2017-18-2021-22-Demographic-Snapshot/c7ru-d68s/about_data

[2] https://data.cityofnewyork.us/Education/Math-Test-Results-2013-2023/74kb-55u9/about_data

[3] https://infohub.nyced.org/reports/students-and-schools/school-quality/nyc-school-survey/survey-archives

[4] More details on the elements are available in Appendix A.

Robust Regression: We used this as it addresses outliers and influential observations to validate results.

Ridge Regression: To address concerns with multicollinearity and model selection, we implemented a ridge regression model (Appendix E).

Generalized Least Squares: We used it to address multicollinearity concerns in our OLS model.

Variable Transformation: We transformed the data using polynomial, rather than linear regression to check for non-linear effects of the variables. We also tested for the significance of certain interaction terms.

Model Evaluation Metrics: We also used the AIC, BIC, R-squared and MSE values to evaluate model performance.

Assumptions

We made the following assumptions for our models:

Linearity: Assessed the randomness of the Residuals vs Fitted plot.

Normality of Residuals: We checked to see if the residuals of our data were normally distributed using measures of skewness and kurtosis. We also used a Normal Q-Q Plot and Density Plot of Residuals to validate this assumption.

Homoscedasticity: We checked to see if the variance of residuals was constant. We used the Breusch-Pagan test, which was supplemented by a Residuals vs Fitted plot and a Scale-Location plot.

Autocorrelation: We checked to see if the residuals were correlated using the Durbin-Watson test.

Influential Observations: We used Cook's Distance to identify influential observations.

Multicollinearity: We checked for multicollinearity assumptions by calculating the VIF across all variables.

Results

The results of this analysis identified significant relationships between school environment factors and academic success (measured by mathematics exam scores). Interestingly, the results indicated that while school-associated factors play a role in academic performance, socioeconomic factors have an overwhelmingly greater impact.

We started by calculating some residual statistics metrics based on a full model of the data, regressing Mean.Scale.Score against all other variables[5], which showed :

[5] Details of the variables used in the final dataset in Appendix B

- Mean: ≈ 0 (-1.606e-16)
- Standard deviation: 5.89
- Range: -30.17 to 21.84

Using the full model, we get statistically significant coefficients, at the 0.05 significance level, for all variables included except female_per, economic.index, Total.Teacher.Response.Rate, Effective.School.Leadership.Score and Strong.Family.Community.Ties.Score. While these findings, on face value, would lead us to start making conclusions around the importance of certain school-level dynamics over others based on statistical significance, we do note that the OLS model assumptions are not met. So, we start by evaluating our model assumptions.

Evaluating Model Assumptions

Linearity: Our data validates this assumption by showing a relatively random Residuals vs Fitted plot (Figure 3 in the Appendix D).

Normality of Residuals: Our data violated this assumption as the distribution of the residuals was symmetrical (skewness = .051), and had heavier tails than a normal distribution (kurtosis = 3.602), suggesting that the data had more extreme values than would be expected in a normal distribution. The Kolmogorov-Smirnov Test had a test statistic of .344, with a p-value < 2.2e-16. The violation of assumptions is further supported by the Normal Q-Q Plot (Figure 1 in Appendix D) and the Density Plot of Residuals (Figure 2 in the Appendix D).

Homoscedasticity: Our data violated this assumption as the Breusch-Pagan test resulted in a BP statistic of 432.93, with a p-value < 2.2e-16. This highly significant result indicates heteroscedasticity which means that the variance of residuals is not constant. This violation is further proven by the Residuals vs Fitted plot (See Figure 3 in Appendix D), which did not show an equal spread of the residuals. The Scale-Location plot (See Figure 4 in Appendix D) also shows a trend in the residuals, which indicates non-constant variance. To remediate, we used robust standard errors and variable transformation.

Autocorrelation: Our data violated this assumption as the Durbin-Watson test resulted in a DW statistic of 0.98, with a p-value < 2.2e-16. This highly significant result indicates strong positive autocorrelation, which is a violation of independence assumption. We figured this could be due to spatial correlation in the data or time-series effects.

Influential Observations: Our threshold for Cook's Distance was 0.00047 and we identified 538 influential observations (See Figure 5 in Appendix D). This represents about 6.3% of the data which is higher than typically expected. Due to this, we had to check for data entry errors to understand if they represented valid but unusual cases and consider their impact on the model.

Multicollinearity: Our data violated multicollinearity assumptions which was evident from the results of calculating the VIF. We observed that severe multicollinearity exists among racial demographic variables, specifically:

    - black_per (VIF = 315.67)
    - hispanic_per (VIF = 305.91)

- white_per (VIF = 154.94)
- asian_per (VIF = 144.51)

Moderate multicollinearity was observed among the following socioeconomic variables:

- poverty_per (VIF = 17.94)
- economic.index (VIF = 17.12)

All other variables showed acceptable VIF values (< 5). To address the multicollinearity observed, we perform elastic net regularization.

Addressing Assumption Violations

To address the violations to assumptions mentioned above, we use multiple statistical methods and compare the performance of the newly developed models.

To address *multicollinearity,* we use our earlier findings in addition to ridge regression in order to recommend variable selection for the next models[6].

Confounders: We observed evidence of confounding when we controlled for the poverty percentage. While the correlation matrix between the variables indicate that there are strong associations between racial composition and poverty (See Figure 6 in Appendix D), the large changes shown below validate that poverty is a major confounder in the relationships between the racial composition variables and Mean Scale Score:

- hispanic_per: 103% change when controlling for poverty
- swd_per: 40% change
- ell_per: 98% change
- asian_per: 33% change
- black_per: 41% change

Poverty Percentage x Students with Disabilities Percentage Interaction: We observed that the interaction between these two variables is significant ($p < 0.005$) with an interaction coefficient of .288. This suggests that the positive effect of the Students with Disabilities Percentage variable on academic success increases with the poverty percentage. (See Figure 7 in Appendix D).

An example of an insignificant interaction that we tested is the interaction between the Economic Index and Rigorous Instruction and the lack of significance is evident in Figure 8 in Appendix D.

Robust Regression: To mitigate the *violation of heteroskedasticity and the presence of influential point*, we used robust regression[7].

---

[6] Appendix E - Variable Selection

[7] Following rigorous model selection processes, the variables used in the robust regression are: Total.Enrollment, female_per, asian_per, black_per, hispanic_per, white_per, swd_per, ell_per, poverty_per, economic.index, Collaborative.Teachers.Score, Effective.School.Leadership.Score, Rigorous.Instruction.Score, Strong.Family.Community.Ties.Score.

When applying robust regression on the full model, coefficient estimates remain relatively the same.

With this model we observed a similar pattern of significant predictors in comparison to the full model OLS, but it provided more conservative estimates than the OLS. The robust regression confirmed the importance of demographic composition variables with statistically significant coefficients for the predictors.

- asian_per: β = 13.9
- ell_per: β = -17.48
- swd_per: β = -29.27[8].

The biggest coefficient estimates are for multirace_per (β = 65.56) and other_per (β = 193.65). However, this is due to the small range of percentages for these variables, so, a 1 percent increase is associated with a larger jump in mean score.

The robust regression also highlights the importance of the NYC DOE measures, albeit at small changes in score for every 1 point increase in the corresponding measure. The biggest coefficient is associated with the Rigorous Instruction Score. It is also the only one of the 5 measures in the model that has a small standard deviation for the coefficient (β = 3.16, StDev = 0.30). The rest of the measures have coefficients below 1 and standard deviation up to 0.39, making these estimates unreliable, especially if we're looking to effect policy changes.

Generalized Least Squares: To mitigate the *autocorrelation found between residuals*, we use a Generalized Linear Regression model.

Using GLS on the full model, we note that the coefficient estimates are more conservative under the GLS model. Further than that, we find changes in the statistical significance of coefficient estimates for variables we expected to be key in determining academic success. For instance, the coefficients for female_per and hispanic_per are not statistically significant under GLS.

Using GLS on the reduced number of variables, the statistical significance of coefficient estimates remains consistent. Coefficient estimates are also close.

Polynomial Model: In an attempt to mitigate *the violation of the normality assumption* in the full model, we use a polynomial model to the 2nd degree for each of the continuous variables in our model.

We find statistically significant coefficient estimates in both the full version and the reduced version of the polynomial model. For instance, the hispanic_per coefficient is not statistically significant in the OLS model with a reduced number of variables, however, the second degree coefficient for hispanic_per is statistically significant. Of the 17 continuous variables in the model, 6 are statistically significant at the 2nd degree.

[8] The coefficients represent the change in mean score for every one percent increase in the corresponding demographic.

Even following this correction, the residuals' distribution still does not meet the normality assumption, indicating that there might be correlation at a higher degree polynomial for some of these variables.

Model Comparison

We have compiled 12 different models using the same data:

- OLS model
- Robust regression model
- GLS model
- Polynomial regression model

Each of these models is fitted using all variables, then variables after fixing multicollinearity with VIF, then the reduced model following model selection with ridge regression.

To compare models, we use AIC, BIC, AICC, R-squared, and prediction error to quantify the improvement in the model. (All tables can be found in Appendix F).

The AIC (Table 1), BIC (Table 2) and AICC (Table 3) metrics point to the GLS model as being the best option with the lowest AIC, BIC and AICC measures for the GLS model using the full-data variables. While we would have expected the models using a reduced number of variables to have lower metrics, we note that the variables omitted following the Ridge Regression are dummy variables, and therefore get reduced to 0 despite bringing useful information to the model.

The R-squared (Table 4) and prediction errors (Table 5), on the other hand, point to the Polynomial regression model as being the best fit using the full-data variables. Given that our main objective is to assess the extent to which school level dynamics impact academic success, we are more concerned with prediction ability and therefore will use this model, without the variables eliminated using VIF, to guide the rest of our analysis.

First, we note that our model of choice's assumptions are not fully met either, for instance, the standard error of residuals is 5.732, not much different from the initial full OLS model. However, given our previous analysis, we choose to proceed.

Interpreting Full Polynomial Model

Using our model of choice, we then found the most significant predictors which were:

Poverty Percentage: We observed a strong negative association where $\beta = -69.69$ for the linear component with a p-value 3.42*10^-5 and $\beta = 28.17$ for the quadratic component of the variable, with a p-value of 3.5*10^-4. In both cases, this indicates a significant association to the response variable.

The discrepancy in the direction of the coefficient estimates for two components of the same variable indicate that, while an increase in the poverty percentage in a school's student body reduces the mean score obtained in NYS Math Exams, as the poverty percentage increases, the trend gets curved and the rate of reduction decreases.

Racial Composition: We observed a significant impact of racial identity on academic performance (category_black, category_white …). In fact, all the category variables have statistically significant coefficient estimates at the 0.05 significance level. Given these are dummy variables, they do not have a quadratic component.

The coefficient estimates indicate a negative relationship in maths exam performance for students who identify with the following groups:

- Current ELL
- African American
- Economically Disadvantaged
- Female
- Hispanic
- Native American
- Students with Disabilities

The direction of the estimates is in-line with the direction of the linear component for the percentage of each of these groups. The largest coefficient estimates and, perhaps, most dramatic of effects, are in the linear component of black_per where $\beta = -172.74$, a statistically significant coefficient (ie. unlikely to be 0), suggesting that every 1 percent increase in the percentage of black students among the student population is associated with a 172.74 point decrease in the mean math exam score for the school. Similarly, the coefficient for percentage of students with disabilities (swd_per) is $\beta = -112.74$ and for ell_per is $\beta = -138.09$.

In the cases of black_per and swd_per, the quadratic component of each of these variables has a positive coefficient estimate (43.82 for black_per and 60.97 for swd_per), indicating that, at as these percentages increase, the rate of decrease in the mean math exam score diminishes and the trend might even reverse after a certain point.

School Environment: The focus of this project is on the effect of school level dynamics, using The Great Schools framework's measures. Our model of choice yields statistically significant coefficient estimates for:

- Total Student Response Rate linear and quadratic components
- Collaborative Teachers' Score quadratic component
- Effective School Leadership Score's linear component
- Rigorous Instruction Score's linear component
- Strong Family Community Ties Score's quadratic component
- Trust score's linear and quadratic components

All these coefficient estimates are positive, indicating a positive relationship between higher scores and response rates with higher mean math exam results, except for the Effective School Leadership Score's linear component with $\beta = -32.97$. The strongest effect appears to be from the Rigorous Instruction Score's linear component where a 1 point increase in the score is associated with a 110 points increase in the mean math exam score for the school.

A discussion of the implication of these findings is featured in the next section.

Discussion

Our analysis highlights that while school-associated factors have an impact on academic performance, this is widely overshadowed by external factors. For example, we saw that the poverty percentage and racial demographics were some of the strongest predictors of academic success, regardless of the model we used. Specifically, it is evident that poverty rates have a negative impact on the mean NYS math exam score.

We also observed that the interaction effects, especially factors like the diminishing positive impact of the Asian student population in schools with a higher percentage of poverty, play a part in the impact of demographic and socioeconomic factors. The results ultimately show that while improving school-associated factors may improve academic performance, the greatest impact on academic performance would stem from addressing socioeconomic inequities between schools, such as poverty. This explains the predominance in discussion of such factors when there is intent to improve students' academic performance.

However, the objective of our analysis is to shift away from that viewpoint and explore other possible factors. We were able to establish the significance of in-school dynamics, with the strongest associations made to the Rigorous Instruction Score and Trust Score.

On the seemingly negative association to the Effective School Leadership score's quadratic component in our polynomial model, it is possible that it is not representative of the true nature of the relationship between inclusivity of the administration and their involvement and students' academic performance. After all, our model still violates key assumptions, and there are ways for a more acute analysis, for instance the introduction of interaction terms. On the other hand, it is possible that it is indicative of a dynamic we are not knowledgeable about. A helpful follow up would be the deaggregation of the components of the Effective School Leadership score for a more detailed examination of this relationship.

The significant impact of a Rigorous Instruction Score on student academic performance is something that we could have foreseen. This metric primarily measures the quality of teaching, the quality of the curriculum and, and perhaps as important, students and parents' perception of the quality of teaching and curriculum, given the survey basis for the metric. This yields questions about what came first: good grades or a good perception of the teaching? While it is undoubtable that quality of teaching has a significant impact on students' academic performance, a survey measure might not be the best means to prove that. Our analysis might benefit from introducing the Quality Review metrics used (as mentioned in Appendix A) given the independence of the reviewers.

From our findings, we can conclude that there exists a duality in addressing academic success concerns. Primarily, there should likely be a focus on public policies that could aid in reducing poverty and providing necessary resources to students who may not be able to afford them. In tandem, schools can make improvements within the classroom, such as adjusting training for teachers to address areas of improvement, fostering a more inclusive environment for students, and ensuring that students are engaged and are receiving the specific aid that they may need.

To address some of the limitations of our analysis, we recognize that the data we used is limited in its focus to only NYC public elementary schools. Thus, our findings are limited to this population and the results could potentially differ if we were to widen to include middle schools, high schools, private schools, etc. Our analysis is also limited to three academic years due to the unavailability of consistent survey data and scoring methodology for a broader time range. If we were to redo this analysis, we could attempt to use other datasets with more consistent data across time periods, especially if it had more current data (post 2020) where we could potentially analyze the difference in results pre and post the COVID-19 pandemic. In addition, there is an inherent bias in using survey-based studies which brings forth concerns of non-response bias and limits to the variables that were used in the analysis. Finally, in this analysis we used math test scores as a measure of academic success but this limits the scope of academic success which could be measured by other subject tests, homework, in-class assignments, etc.

We also acknowledge the limit that is in using school-level data across multiple grade levels and groups, as it can mask important variations and trends that may exist within specific grade bands or subject areas. School-wide data often aggregates results in a way that overlooks the diverse needs of different student groups, such as those in English as a Second Language programs, or Students with Disabilities. This broad approach can obscure differences in performance or growth across grade levels, making it difficult to pinpoint areas of concern or success at a more granular level. Moreover, the teaching strategies and curricular focus can vary significantly across grades, so using data at the school level risks drawing generalized conclusions that do not accurately reflect the unique dynamics of each grade. As a result, this approach may lead to misguided decisions or interventions that fail to address the specific needs of certain student demographics.

Finally, our model doesn't account for interactions between the different variables used. While we attempt to meet a number of model assumptions, our project would have benefited from further sector knowledge in the selection of variables. Unfortunately, the contacts we were able to reach at the NYC DOE had no current needs for research support.

If we were to improve upon the limitations of our analysis, we would broaden the scope of the types of studies used and include other measures of academic success to provide a more holistic view on the impact of school-based factors and socioeconomic factors on academic performance.

However, overall, we can conclude from our analysis that poverty and other systemic inequities have a significant impact on academic performance, but, school-level dynamics do appear to have an effect on academic performance, or at least be correlated to it. We've established, at this stage, that teaching quality and community trust are significant factors. Further analysis, both quantitative and qualitative, is required to truly establish the nature of this relationship.

Appendix

Appendix A - Understanding NYC DOE School Survey Elements

The NYC School Survey evaluates public schools across six key elements that are critical to fostering effective school environments.

These elements are **Rigorous Instruction**, which assesses the quality and challenge of academic programming; **Supportive Environment**, focusing on the emotional, physical, and academic support provided to students; **Collaborative Teachers**, measuring staff collaboration and professional growth opportunities; **Effective School Leadership**, which gauges the administration's ability to inspire and manage the school community; **Strong Family-Community Ties**, reflecting the engagement and partnership between schools, families, and the local community; and **Trust**, which aims at measuring the level of trust and respect among students, teachers, parents, and school leadership.

These elements are from the new Framework for Great Schools, aiming to promote student success and equity. We must note that the scoring is not dependent on survey responses alone. It includes assessments from Quality Reviews which are conducted by trained reviewers and designed to evaluate the effectiveness of schools in promoting student achievement and meeting the diverse needs of their communities. It also includes additional metrics, such as chronic absenteeism.

The NYC Department of Education provides a more detailed explanation and breakdown of the measures in its yearly scoring technical guides.[9]

Appendix B - Full Model Variables

The final dataset contains data for 475 elementary schools in New York City, for 2 academic years, under the following variables:

Demographic Information

- **Total Enrollment:** the total number of students enrolled at the school
- **Female_per:** percentage of female students
- **Asian_per:** percentage of asian students
- **Black_per:** percentage of african-american students
- **Hispanic_per:** percentage of hispanic students
- **White_per:** percentage of white students
- **Multirace_per:** percentage of multiracial students
- **Other_per:** percentage of other racial backgrounds not listed above
- **Swd_per:** percentage of students with disabilities
- **Ell_per:** percentage of students with English as a second language
- **Poverty_per:** percentage of students considered to be living in poverty

[9] 2018-19 Framework and School Survey Scoring Technical Guide, https://infohub.nyced.org/docs/default-source/default-document-library/framework-school-survey-scoring-technical-guide.pdf

- **Economic.Index:** refers to the Economic Need Index and estimates the percentage of students facing economic hardship[10]

Survey Data

- **Total.Teacher.Response.Rate:** rate of teachers at the school who completed the survey
- **Total.Student.Response.Rate:** rate of students at the school who completed the survey
- **Collaborative.Teachers.Score:** the school's score under the Collaborative Teachers element of the Great Schools Framework
- **Effective.School.Leadership.Score:** the school's score under the Effective School Leadership element of the Great Schools Framework
- **Rigorous.Instruction.Score:** the school's score under the Rigorous Instruction element of the Great Schools Framework
- **Strong.Family.Community.Ties.Score:** the school's score under the Strong Family Community Ties element of the Great Schools Framework
- **Trust.Score:** the school's score under the Trust element of the Great Schools Framework

New York State Math Exam Data

- **Category_Asian:** T/F variable for whether the math results data is based only on the Asian student population of the school (all false in our dataset)
- **Category_Black:** T/F variable for whether the math results data is based only on the Black student population of the school (all false in our dataset)
- **Category_Hispanic:** T/F variable for whether the math results data is based only on the Hispanic student population of the school (all false in our dataset)
- **Category_Multi.racial:** T/F variable for whether the math results data is based only on the multiracial student population of the school (all false in our dataset)
- **Category_Native.American:** T/F variable for whether the math results data is based only on the native-american student population of the school (all false in our dataset)
- **Category_White:** T/F variable for whether the math results data is based only on the White student population of the school (all false in our dataset)
- **Category_Current.ELL:** T/F variable for whether the math results data is based only on the student population that has english as a second language at the time of taking the test (all false in our dataset)
- **Category_Ever.ELL:** T/F variable for whether the math results data is based only on the student population that has, at some point in their education, taken english as a second language (all false in our dataset)
- **Category_Econ.Disadv:** T/F variable for whether the math results data is based only on the student population of the school considered to be economically disadvantaged (all false in our dataset)

---

10 https://data.cccnewyork.org/data/bar/1371/student-economic-need-index#:~:text=The%20Economic%20Need%20Index%20%28ENI%29%20estimates%20the%20percentage,the%20first%20time%20within%20the%20last%20four%20years.

- **Category_Female:** T/F variable for whether the math results data is based only on the female student population of the school (all false in our dataset)
- **Category_SWD:** T/F variable for whether the math results data is based only on students with disabilities at the school (all false in our dataset)
- **Mean.Scale.Score:** Mean score for all students at the school (or with specified group) at the New York State Math Exam.

Appendix C - Exploratory Data Analysis

After performing exploratory data analysis on the dataset, specifically on the continuous variables, we saw that outliers existed for all predictors except for the hispanic_per variable. Based on our objective, some notable outliers that we observed were from:

- **Total.Enrollment (**Figure 9 in Appendix D): outliers indicate schools with either higher than or lower than median enrollment, potentially skewing results
- **Poverty_per** (Figure 10 in Appendix D): outliers represent schools with extreme levels of poverty which may exacerbate the impact of poverty on academic success
- **Economic.Index** (Figure 11 in Appendix D): outliers reflect extreme economic conditions, either extreme poverty or extreme wealth, which can similarly exacerbate the impact on academic success
- **Total.Teacher.Response.Rate** (Figures 12 in Appendix D): outliers reflect anomalies in the response rate which skews the data if there's higher than or lower than the median engagement

These outliers indicated that we may have needed to apply robust regression or transformation to remediate their impact on the data, and we ultimately performed these methods in the analysis.

We also calculated measures of location of the continuous variables where we observed that slightly less than half of the variables seemed to exhibit a normal distribution, while the rest were skewed. Some key observations were made in:

- **Total.Enrollment** (Figure 13 in Appendix D): the distribution of this group was skewed left indicating that there was a concentration in responses from smaller schools in comparison to larger schools
- **Poverty_per** (Figure 14 in Appendix D): the distribution of this group was skewed right, indicating that a larger concentration of schools had higher poverty levels
- **Other_per** (Figure 15 in Appendix D): the distribution of this group and the multirace_per variable were very narrow as most of the data was clustered in one area. This indicates low variation in these groups, with most of the data concentrated near 0.
- **Rigorous.Instruction.Score** (Figures 16 in Appendix D): the distribution of this group and and **Effective.School.Leadership.Score** displayed signs of bimodality, which indicates that there may be significant difference in schools with better resources and instruction compared to those who lack in these areas

Finally, we looked at some measures of dispersion where we saw that **"Total.Enrollment"**, **"poverty_per"**, and **"Total.Teacher.Response.Rate"** showed high variance, indicating high variability across the size of schools, socioeconomic status of students, and engagement levels from teachers. Variables like **"multirace_per"** and **"other_per"** had lower variance, as expected from the histograms, indicating that there's not much variability in these groups

across schools. When we normalize the standard deviation to get the coefficient of variation, we observed that **"poverty_per"**, **"asian_per"**, and **"other_per"** had higher CVs indicating more variability relative to the mean across socioeconomic status and demographic groups. **"Rigorous.Instruction.Score"** and **"Trust.Score"** had lower CVs indicating that these variables were fairly consistent across schools.

Appendix D - Figures

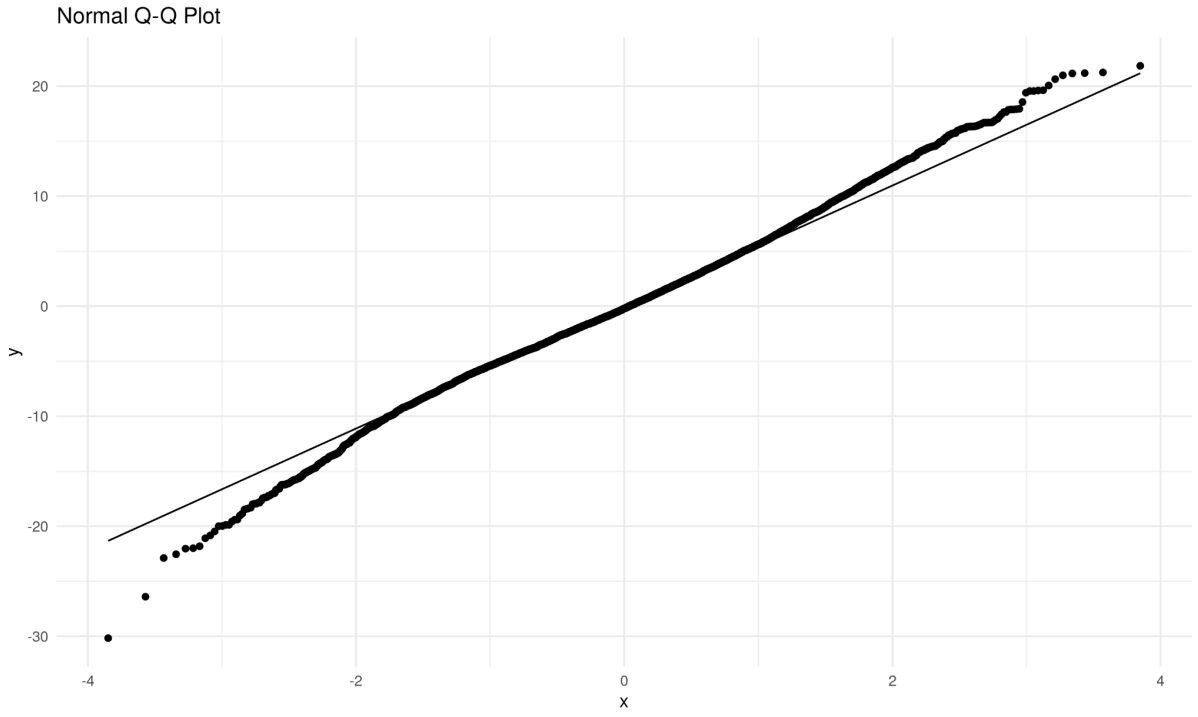

Figure 1 - Q-Q Plot of full OLS model

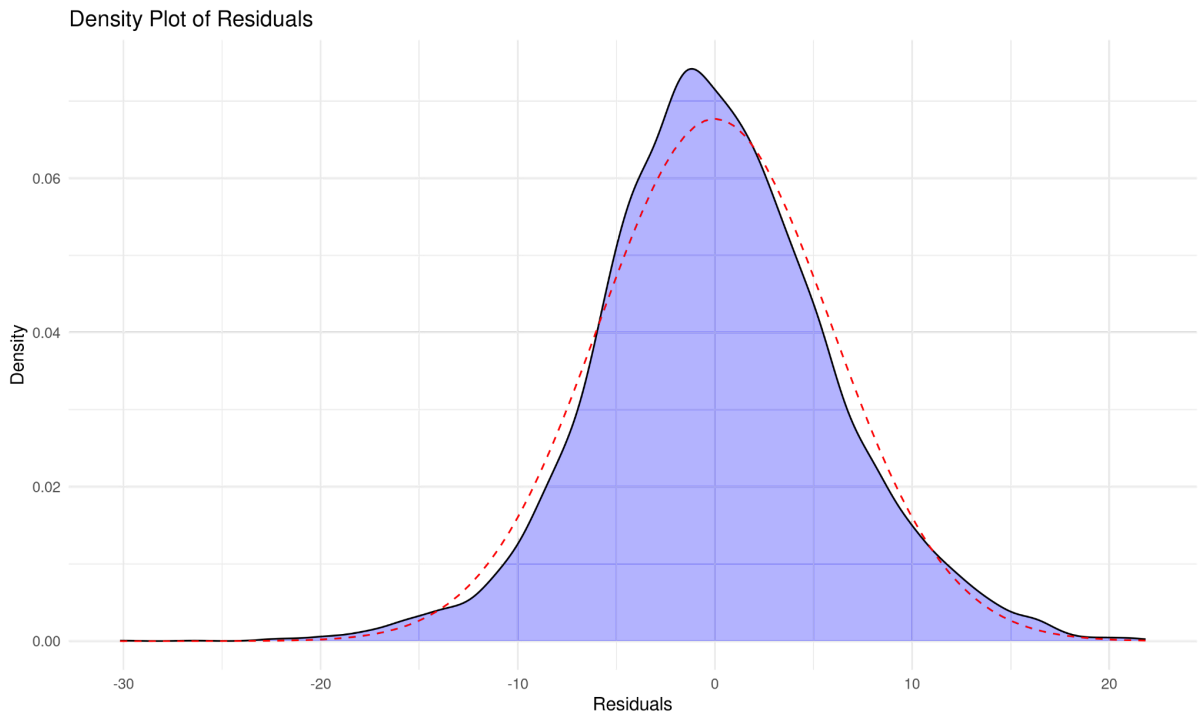

Figure 2 - Residuals Density plot for Full OLS model

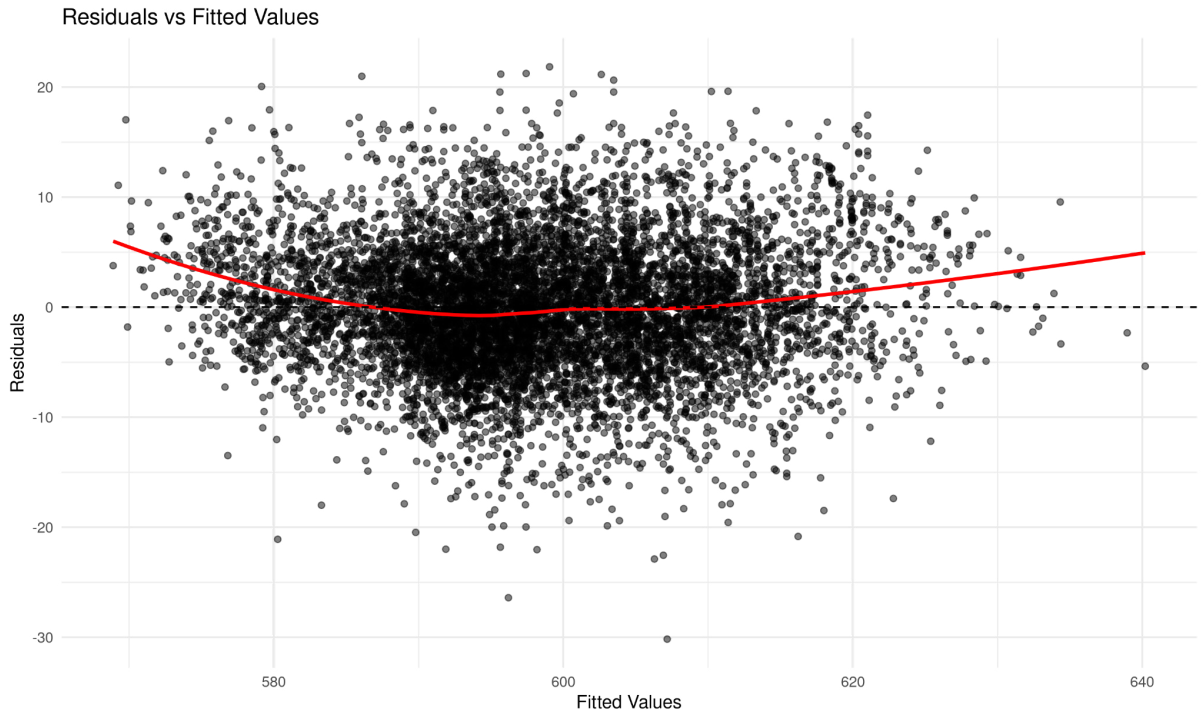

Figure 3 - Residuals plot for Full OLS model

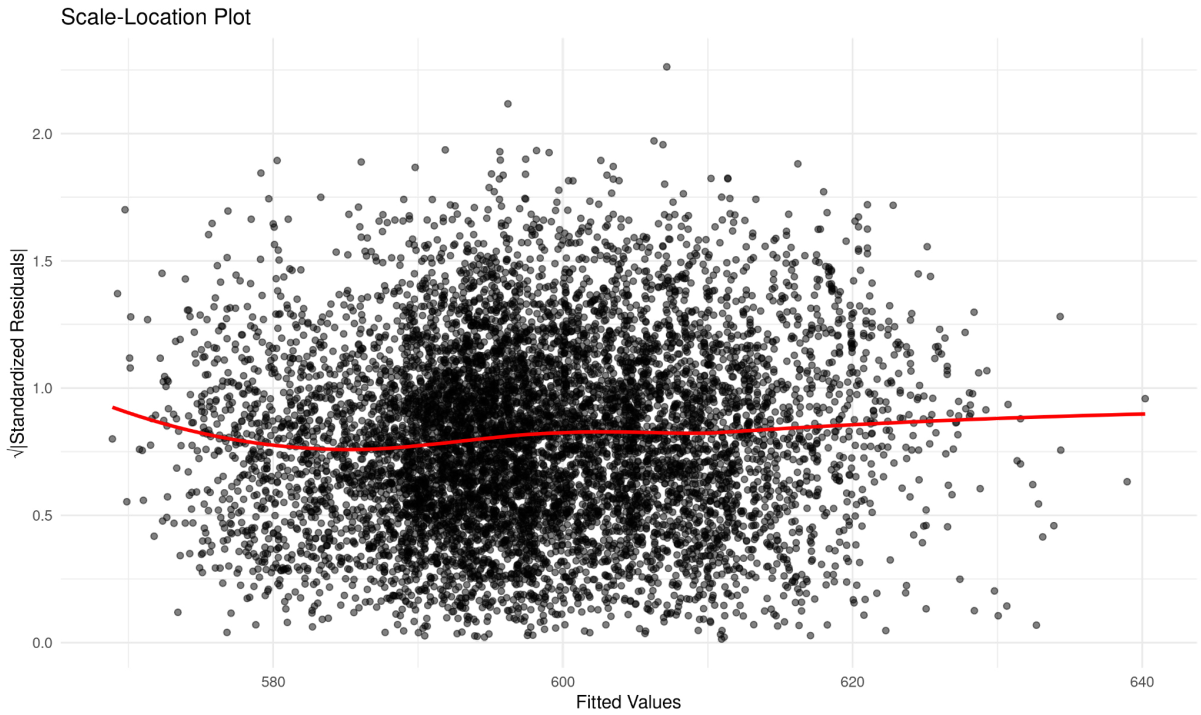

Figure 4 - Scale location plot of full OLS model

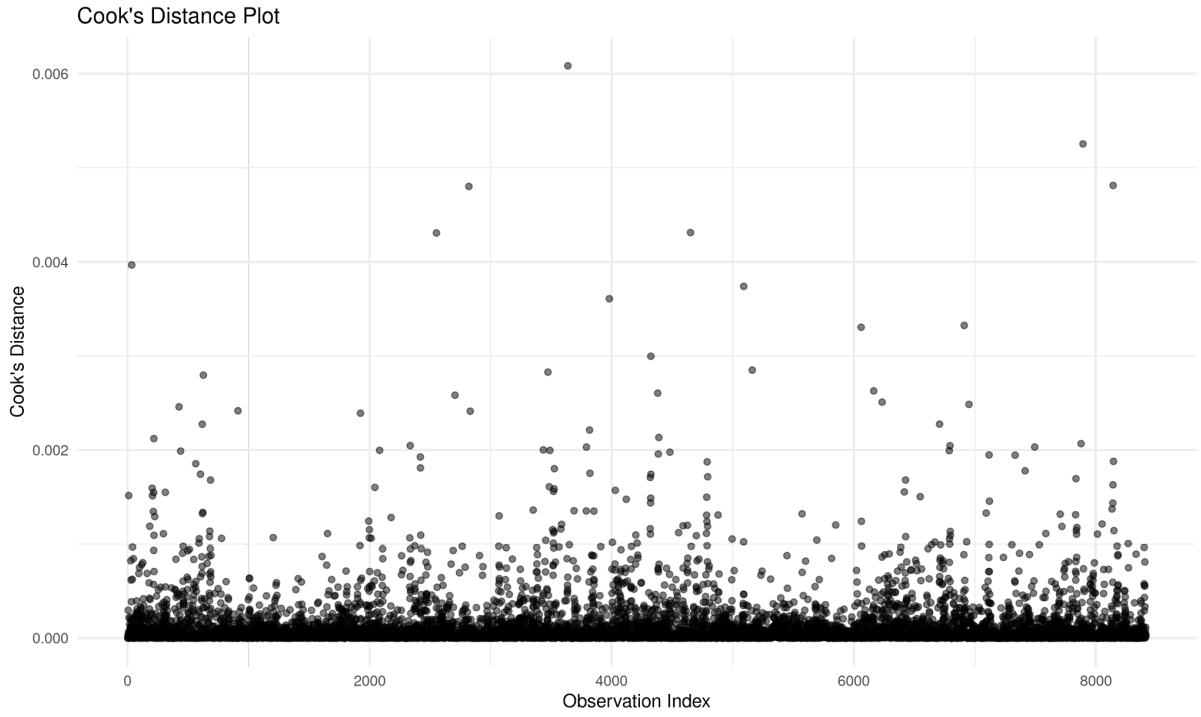

Figure 5 - Cook’s Distance Plot

Figure 6 - Correlation Matrix for Explanatory Variables

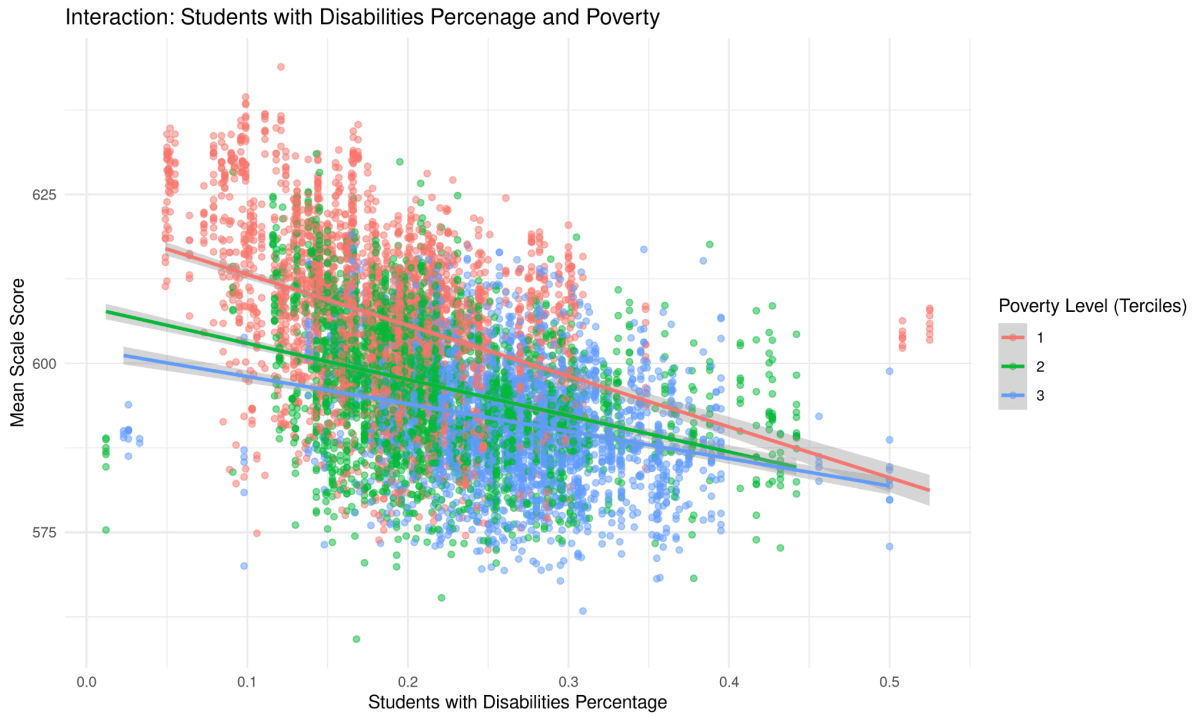

Figure 7 - Interaction between swd_per and poverty levels

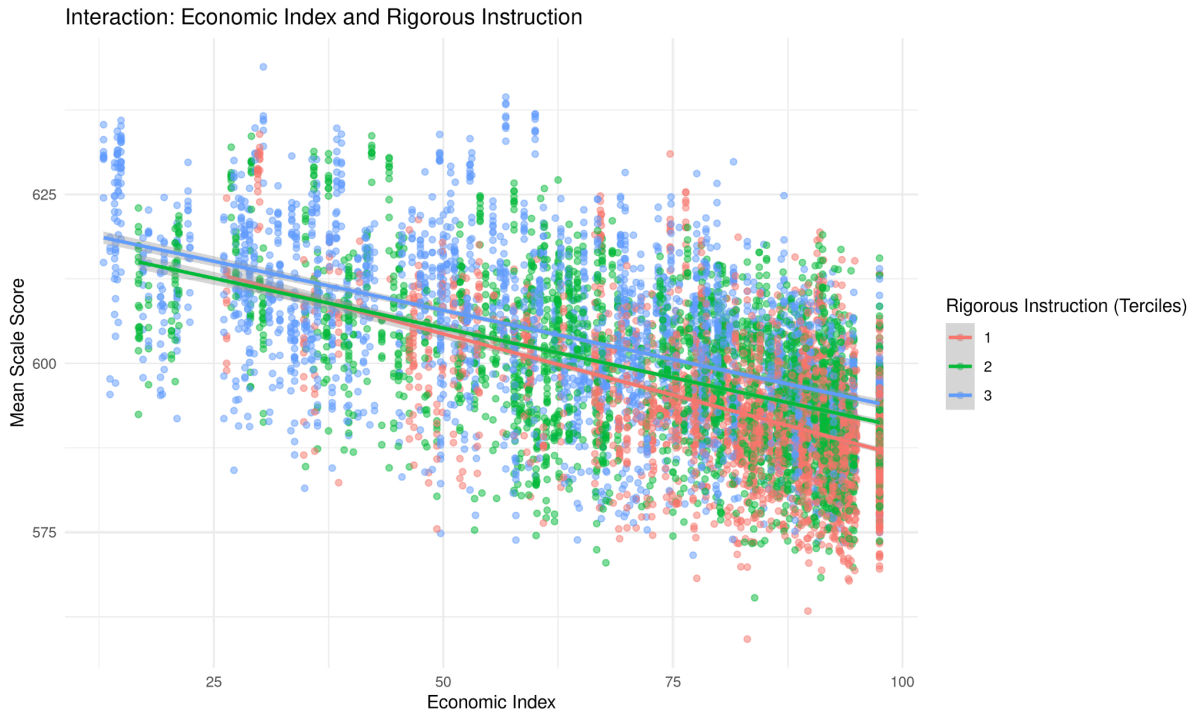

Figure 8 - Interaction between Economic Index and Rigorous Instruction Score

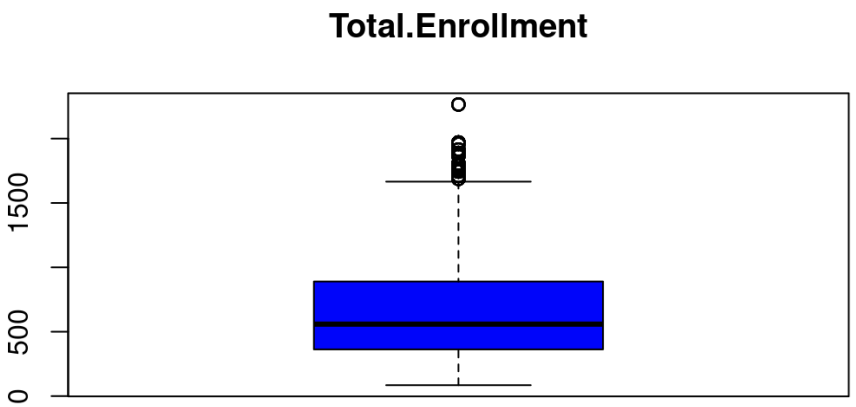

Figure 9 - Boxplot of Total.Enrollment variable with outliers

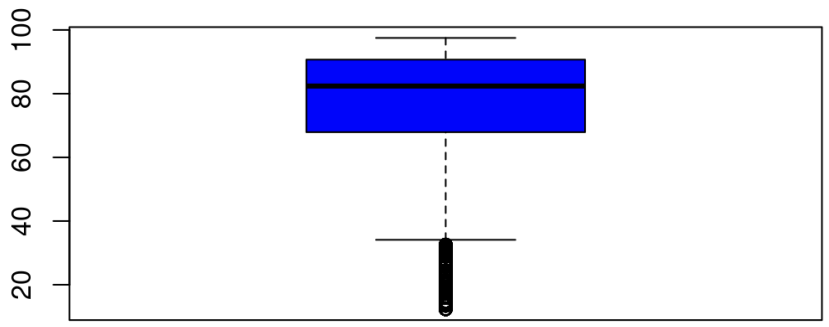

Figure 10 - Boxplot of poverty_per variable with outliers

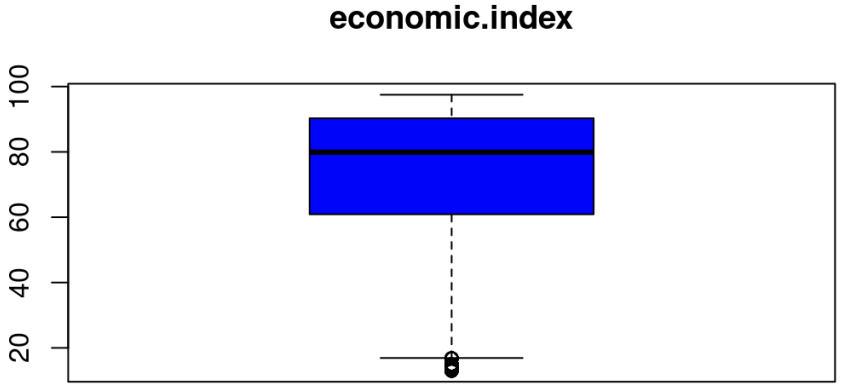

Figure 11 - Boxplot of economic.index variable with outliers

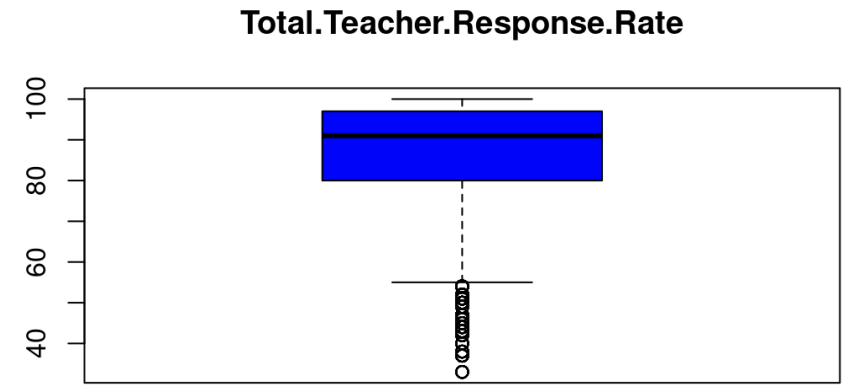

Figure 12 - Boxplot of Total.Teacher.Response.Rate variable with outliers

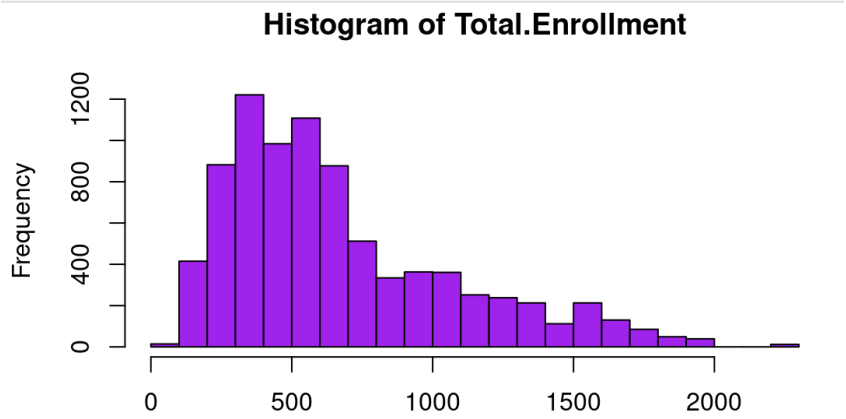

Figure 13: Histogram of Total.Enrollment (right skew)

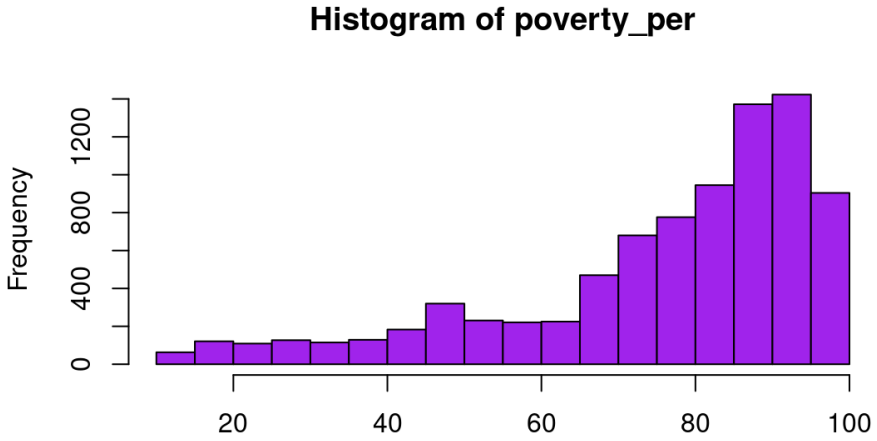

Figure 14: Histogram of poverty_per (left skew)

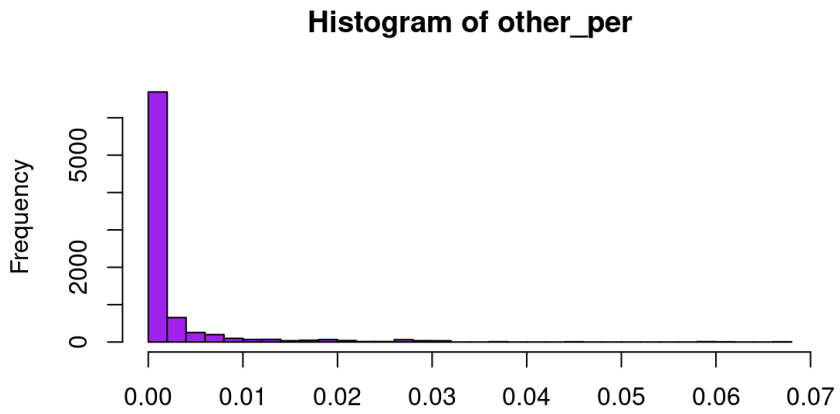

Figure 15: Histogram of other_per (narrow distribution clustered around 0)

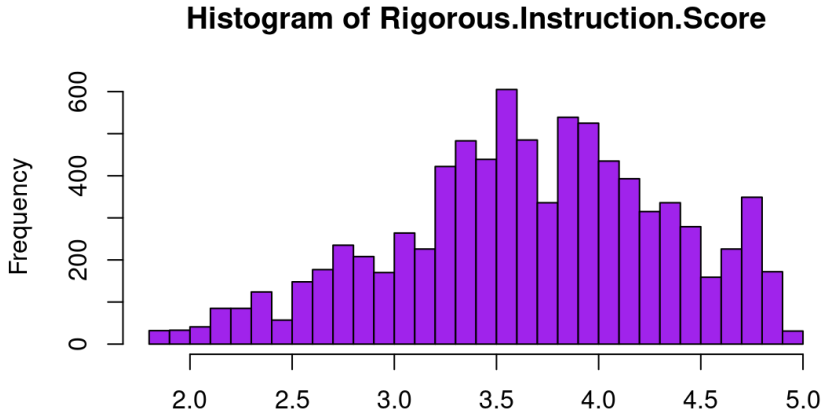

Figure 16: Histogram of Rigorous.Instruction.Score (bimodal distribution)

Appendix E - Variable Selection

VIF and Correlation Matrix

Based on the results of VIF on the variables of the full model, as well as the output of the correlation matrix for the predictor variables, we have chosen to remove the poverty_index variable and the white_per variable.

Povery_index has a 0.9 Pearson correlation coefficient with poverty_per, and white_per has a -0.81 Pearson correlation coefficient with poverty_per;

Following these adjustments, we note the following changes in the Variance Inflation Factor compared to the full model:

- black_per VIF drops from 315.67 to 7.82.
- hispanic_per VIF drops from 305.91 to 7.96.

- asian_per VIF drops from 144.51 to 4.36.
- poverty_per VIF drops from 17.94 to 7.42.

Ridge Regression

To further reduce the number of parameters in our model, we perform a ridge regression on the newly selected models. The choice for ridge regression specifically is to mitigate any lingering multicollinearity despite the reduction in Variance Inflation Factor among the predictor variables.

As a result, the process reduces the coefficients of the Category variables to 0. So, we remove the following variables:

- Category_Asian
- Category_Black
- Category_Hispanic
- Category_Multi.racial
- Category_Native.American
- Category_White
- Category_Current.ELL
- Category_Ever.ELL
- Category_Econ.Disadv
- Category_Female
- Category_SWD

Appendix F - Model Evaluation

To evaluate the models, we use different measures, namely:

- AIC
- BIC
- AICC
- R-squared
- MSE

The following tables indicate our findings such that “full model”, “vif reduction” and "ridge reduction” refer to the variables used, first the full variables in our data, then the variables without those we removed after examining VIF, then the variables without those removed by the ridge regression.

Table 1 - AIC

	full model <dbl>	vif reduction <dbl>	ridge reduction <dbl>
OLS	27072.19	27072.19	29852.14
Robust Regression	27078.95	27078.95	29865.28
GLS	25956.70	25956.70	29692.29
Polynomial	26606.67	26709.63	29659.71

Table 2 - BIC

	full model <dbl>	vif reduction <dbl>	ridge reduction <dbl>
OLS	27262.52	27262.52	29972.68
Robust Regression	27269.28	27269.28	29985.82
GLS	26153.16	26153.16	29819.09
Polynomial	26930.23	27007.81	29888.10

Table 3 - AICC

	full model <dbl>	vif reduction <dbl>	ridge reduction <dbl>
OLS	27072.64	27072.64	29852.32
Robust Regression	27079.40	27079.40	29865.46
GLS	25957.18	25957.18	29692.49
Polynomial	26607.94	26710.72	29660.35

Table 4 - R-Squared

	full model <dbl>	vif reduction <dbl>	ridge reduction <dbl>
OLS	0.7590536	0.7590536	0.5321616
Robust Regression	0.7602960	0.7602960	0.5332613
GLS	0.7593524	0.7593524	0.5339292
Polynomial	0.7853630	0.7798369	0.5548748

Table 5 - Prediction Error (MSE)

	full model <dbl>	vif reduction <dbl>	ridge reduction <dbl>
OLS	38.77496	38.77496	72.01127
Robust Regression	38.92602	38.92602	71.96035
GLS	38.77848	38.77848	72.02971
Polynomial	34.05463	34.86225	68.15545

Appendix G - Code

```
###### EDA #######

#install.packages("dplyr")
#install.packages("ggplot2")
library(dplyr)
library(ggplot2)

data <- read.csv("master_data_cleaned.csv")

#data for 2018 and 2019
```

```
data <- data %>% filter(Year %in% c(2018, 2019))

#continuous predictors
continuous_predictors <- c(
  "Total.Enrollment", "female_per", "asian_per", "black_per",
  "hispanic_per", "white_per", "swd_per", "ell_per", "poverty_per",
  "multirace_per", "other_per", "economic.index",
  "Total.Teacher.Response.Rate", "Total.Student.Response.Rate",
  "Collaborative.Teachers.Score",
"Effective.School.Leadership.Score",
  "Rigorous.Instruction.Score",
"Strong.Family.Community.Ties.Score",
  "Trust.Score"
)

#boxplots for continuous variables
for (col in continuous_predictors) {
  boxplot(data[[col]], main = col, col = "blue")
}

#IQR and outlier function
iqr_outliers <- function(data, cols) {
  results <- list()
  for (col in cols) {
    if (col %in% colnames(data)) {
      q1 <- quantile(data[[col]], 0.25, na.rm = TRUE)
      q3 <- quantile(data[[col]], 0.75, na.rm = TRUE)
      iqr <- q3 - q1
      lower_bound <- q1 - 1.5 * iqr
      upper_bound <- q3 + 1.5 * iqr
      outliers <- data[[col]][data[[col]] < lower_bound |
data[[col]] > upper_bound]
      results[[col]] <- list(
        Q1 = q1,
        Q3 = q3,
        IQR = iqr,
        Lower_Bound = lower_bound,
        Upper_Bound = upper_bound,
        Outlier_Count = length(outliers),
        Outliers = outliers
      )
    }
```

```
  }
  return(results)
}

#calculate outliers
outlier_results <- iqr_outliers(data, continuous_predictors)

#summary
outlier_summary <- lapply(outlier_results, function(x) {
  list(Outlier_Count = x$Outlier_Count)})
print(outlier_summary)

#categorical predictors
categorical_predictors <- c(
  "Category_Asian", "Category_Black", "Category_Current.ELL",
  "Category_Econ.Disadv", "Category_Ever.ELL", "Category_Female",
  "Category_Hispanic", "Category_Multi.Racial",
"Category_Native.American",
  "Category_SWD", "Category_White"
)

#frequency
for (col in categorical_predictors) {
  if (col %in% colnames(data)) {
    cat("Frequency of", col, ":\n")
    print(table(data[[col]], useNA = "ifany"))
    cat("\n")
  }
}

#barcharts
for (col in categorical_predictors) {
  barplot(
    table(data[[col]]),
    main = paste("Frequency of", col),
    col = "pink",
    border = "black",
    las = 2
  )
}
```

```r
#mean and median
location_summary <- data.frame(
  variables = continuous_predictors,
  mean = sapply(data[continuous_predictors], mean, na.rm = TRUE),
  median = sapply(data[continuous_predictors], median, na.rm = TRUE)
)
print(location_summary)

#histograms
for (col in continuous_predictors) {
  hist(data[[col]],
    main = paste("Histogram of", col),
    xlab = col,
    col = "purple",
    border = "black",
    breaks = 30
  )
}



#coefficient of variation
cv <- function(x) {sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE)}

#variance and cv
variance_cv <- data.frame(
  variable = continuous_predictors,
  variance = sapply(data[continuous_predictors], var, na.rm = TRUE),
  cv = sapply(data[continuous_predictors], cv)
)

print(variance_cv)

##### ASSUMPTIONS #####

# Load required packages
if (!require("car")) install.packages("car")
if (!require("lmtest")) install.packages("lmtest")
if (!require("nortest")) install.packages("nortest")
if (!require("ggplot2")) install.packages("ggplot2")

library(car)
library(lmtest)
```

```
library(nortest)
library(ggplot2)

# 1. Linearity Test
# Create residual vs. fitted plot
residual_plot <- ggplot(data.frame(
  fitted = fitted(full_model),
  residuals = residuals(full_model)
), aes(x = fitted, y = residuals)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()

# 2. Normality Tests
# QQ Plot
qq_plot <- ggplot(data.frame(residuals = residuals(full_model)),
                  aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Normal Q-Q Plot") +
  theme_minimal()

# Shapiro-Wilk test
shapiro_test <- shapiro.test(residuals(full_model))

# Anderson-Darling test
ad_test <- ad.test(residuals(full_model))

# 3. Homoscedasticity Tests
# Breusch-Pagan test
bp_test <- bptest(full_model)

# Scale-Location Plot
scale_loc_plot <- ggplot(data.frame(
  fitted = fitted(full_model),
  std_residuals = sqrt(abs(rstandard(full_model)))
), aes(x = fitted, y = std_residuals)) +
  geom_point(alpha = 0.5) +
```

```
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Scale-Location Plot",
       x = "Fitted Values",
       y = "√|Standardized Residuals|") +
  theme_minimal()

# 4. Independence of Observations
# Durbin-Watson test for autocorrelation
dw_test <- dwtest(full_model)

# 5. Influence Diagnostics
# Cook's distance plot
cooks_plot <- ggplot(data.frame(
  index = 1:length(cooks.distance(full_model)),
  cooks = cooks.distance(full_model)
), aes(x = index, y = cooks)) +
  geom_point(alpha = 0.5) +
  labs(title = "Cook's Distance Plot",
       x = "Observation Index",
       y = "Cook's Distance") +
  theme_minimal()

# Save plots
ggsave("residual_plot.png", residual_plot, width = 10, height = 6)
ggsave("qq_plot.png", qq_plot, width = 10, height = 6)
ggsave("scale_loc_plot.png", scale_loc_plot, width = 10, height = 6)
ggsave("cooks_plot.png", cooks_plot, width = 10, height = 6)

# Print test results
cat("\nNormality Tests:\n")
cat("Shapiro-Wilk test:\n")
print(shapiro_test)
cat("\nAnderson-Darling test:\n")
print(ad_test)

cat("\nHomoscedasticity Test:\n")
cat("Breusch-Pagan test:\n")
print(bp_test)

cat("\nIndependence Test:\n")
cat("Durbin-Watson test:\n")
print(dw_test)
```

```
# Check for influential observations
influential_obs <- which(cooks.distance(full_model) > 4/nrow(data))
cat("\nNumber of potentially influential observations:",
length(influential_obs))


##### MODELLING #####

library(car)
library(corrplot)
library(ggplot2)
library(dplyr)
library(lmtest)
library(nortest)
library(moments)
library(sandwich)
library(nlme)
library(MASS)
library(glmnet)


# Data
data = read.csv("master_data_cleaned.csv")
data = data[data$Year == c(2018,2019),]

key_predictors <- c(
  "Total.Enrollment",
  "female_per", "asian_per", "black_per",
  "hispanic_per", "white_per", "swd_per", "ell_per", "poverty_per",
  "multirace_per", "other_per", "economic.index",
  "Total.Teacher.Response.Rate",
  "Total.Student.Response.Rate", "Collaborative.Teachers.Score",
  "Effective.School.Leadership.Score", "Rigorous.Instruction.Score",
  "Strong.Family.Community.Ties.Score", "Trust.Score",
  "Category_Asian", "Category_Black", "Category_Current.ELL",
  "Category_Econ.Disadv", "Category_Ever.ELL", "Category_Female",
  "Category_Hispanic", "Category_Multi.Racial",
"Category_Native.American",
  "Category_SWD", "Category_White"
)
```

```
# Variable Selection
# Remove economic.index, because almost 1 to 1 correlation with
poverty_per.
cor.test(data$poverty_per, data$economic.index)

# Remove white_per because extremely high correlation with
poverty_per
# Other race variables will be measure for diversity
cor.test(data$poverty_per, data$white_per)

# New predictors list
vif_predictors <- c(
  "Total.Enrollment",
  "female_per", "asian_per", "black_per",
  "hispanic_per", "swd_per", "ell_per", "poverty_per",
  "multirace_per", "other_per",
  "Total.Teacher.Response.Rate",
  "Total.Student.Response.Rate", "Collaborative.Teachers.Score",
  "Effective.School.Leadership.Score", "Rigorous.Instruction.Score",
  "Strong.Family.Community.Ties.Score", "Trust.Score",
  "Category_Asian", "Category_Black", "Category_Current.ELL",
  "Category_Econ.Disadv", "Category_Ever.ELL", "Category_Female",
  "Category_Hispanic", "Category_Multi.Racial",
"Category_Native.American",
  "Category_SWD", "Category_White"
)

# Variable Selection with Ridge Regression
Prepare data for glmnet
for(i in 55:69){
  temp = data[,i]
  temp = ifelse(temp == "False", 0, 1)
  data[,i] = temp
}

X <- data[, key_predictors]
X = as.matrix(X)
y <- data$Mean.Scale.Score

# Ridge Regression
cv_ridge <- cv.glmnet(X, y, alpha = 0)  # Alpha = 1 for LASSO
lambda = cv_ridge$lambda.min
```

```
final_ridge <- glmnet(X, y, alpha = 0, lambda = lambda)
coef(final_ridge)

ridge_predictors = c(
  "Total.Enrollment",
  "female_per", "asian_per", "black_per",
  "hispanic_per", "swd_per", "ell_per", "poverty_per",
  "multirace_per", "other_per",
  "Total.Teacher.Response.Rate",
  "Total.Student.Response.Rate", "Collaborative.Teachers.Score",
  "Effective.School.Leadership.Score", "Rigorous.Instruction.Score",
  "Strong.Family.Community.Ties.Score", "Trust.Score"
)

# ALL MODELS
```{r}
# Formulas
full_formula = as.formula(paste("Mean.Scale.Score ~",
                                paste(key_predictors, collapse = " +
")))
vif_formula = as.formula(paste("Mean.Scale.Score ~",
                               paste(vif_predictors, collapse = " +
")))
ridge_formula = as.formula(paste("Mean.Scale.Score ~",
                                 paste(ridge_predictors, collapse =
" + ")))

# OLS
full_ols = lm(full_formula, data = data)
vif_ols = lm(vif_formula, data = data)
ridge_ols = lm(ridge_formula, data = data)

# Robust Regression
full_robust = rlm(full_formula, data = data)
vif_robust = rlm(vif_formula, data = data)
ridge_robust = rlm(ridge_formula, data = data)

# GLS
full_gls = gls(full_formula, data = data, correlation = corAR1())
vif_gls = gls(vif_formula, data = data, correlation = corAR1())
ridge_gls = gls(ridge_formula, data = data, correlation = corAR1())
```
```

```
# Polynomial
full_poly = lm(as.formula(paste("Mean.Scale.Score ~",
                                paste(c(trans_vars, other_vars),
                                      collapse = " + "))), data =
data)
vif_poly = lm(as.formula(paste("Mean.Scale.Score ~",
                               paste(c(trans_vif, vif_other),
                                     collapse = " + "))), data =
data)
ridge_poly = lm(as.formula(paste("Mean.Scale.Score ~",
                                 paste(c(trans_reduced,
reduced_other),
                                       collapse = " + "))), data =
data)

models = list(full_ols, vif_ols, ridge_ols,
              full_robust, vif_robust, ridge_robust,
              full_gls, vif_gls, ridge_gls,
              full_poly, vif_poly, ridge_poly)

# Model Evaluation
column_names = c("full model", "vif reduction", "ridge reduction")
row_names = c("OLS", "Robust Regression", "GLS", "Polynomial")

# AIC
aic_values = lapply(models, AIC)
aic_values = as.vector(unlist(aic_values))

aic_df = as.data.frame(matrix(aic_values, ncol = 3, byrow = T))
colnames(aic_df) = column_names
rownames(aic_df) = row_names
aic_df

# BIC
bic_values = lapply(models, BIC)
bic_values = as.vector(unlist(bic_values))

bic_df = as.data.frame(matrix(bic_values, ncol = 3, byrow = T))
colnames(bic_df) = column_names
rownames(bic_df) = row_names
bic_df
```

```
# AICC
library(MuMIn)
aicc_values = lapply(models, AICc)
aicc_values = as.vector(unlist(aicc_values))

aicc_df = as.data.frame(matrix(aicc_values, ncol = 3, byrow = T))
colnames(aicc_df) = column_names
rownames(aicc_df) = row_names
aicc_df

# R²
tot_var = var(data$Mean.Scale.Score)
r2_values = c(summary(full_ols)$adj.r.squared,
              summary(vif_ols)$adj.r.squared,
              summary(ridge_ols)$adj.r.squared,
              1 - (var(residuals(full_robust)) / tot_var),
              1 - (var(residuals(vif_robust)) / tot_var),
              1 - (var(residuals(ridge_robust)) / tot_var),
              cor(predict(full_gls), data$Mean.Scale.Score)^2,
              cor(predict(vif_gls), data$Mean.Scale.Score)^2,
              cor(predict(ridge_gls), data$Mean.Scale.Score)^2,
              summary(full_poly)$adj.r.squared,
              summary(vif_poly)$adj.r.squared,
              summary(ridge_poly)$adj.r.squared)

r2_df = as.data.frame(matrix(r2_values, ncol = 3, byrow = T))
colnames(r2_df) = column_names
rownames(r2_df) = row_names
r2_df

# Prediction Error
# Split the data into training (80%) and testing (20%)
set.seed(123)
train_indices <- sample(seq_len(nrow(data)), size = 0.8 *
nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Define a function to calculate prediction error
calculate_prediction_error <- function(model, test_data,
response_var) {
  predictions <- predict(model, newdata = test_data)
```

```
  true_values <- test_data[[response_var]]
  mean((predictions - true_values)^2) # Mean Squared Error (MSE)
}

# Fit models on training data
full_ols <- lm(full_formula, data = train_data)
vif_ols <- lm(vif_formula, data = train_data)
ridge_ols <- lm(ridge_formula, data = train_data)

full_robust <- rlm(full_formula, data = train_data)
vif_robust <- rlm(vif_formula, data = train_data)
ridge_robust <- rlm(ridge_formula, data = train_data)

full_gls <- gls(full_formula, data = train_data, correlation =
corAR1())
vif_gls <- gls(vif_formula, data = train_data, correlation =
corAR1())
ridge_gls <- gls(ridge_formula, data = train_data, correlation =
corAR1())

full_poly <- lm(as.formula(paste("Mean.Scale.Score ~",
                                 paste(c(trans_vars, other_vars),
                                       collapse = " + "))), data =
train_data)
vif_poly <- lm(as.formula(paste("Mean.Scale.Score ~",
                                paste(c(trans_vif, vif_other),
                                      collapse = " + "))), data =
train_data)
ridge_poly <- lm(as.formula(paste("Mean.Scale.Score ~",
                                  paste(c(trans_reduced,
reduced_other),
                                        collapse = " + "))), data =
train_data)

# Calculate prediction errors
models <- list(full_ols, vif_ols, ridge_ols,
               full_robust, vif_robust, ridge_robust,
               full_gls, vif_gls, ridge_gls,
               full_poly, vif_poly, ridge_poly)

model_names <- c("OLS Full", "OLS VIF", "OLS Ridge",
                 "Robust Full", "Robust VIF", "Robust Ridge",
```

```
                "GLS Full", "GLS VIF", "GLS Ridge",
                "Poly Full", "Poly VIF", "Poly Ridge")

response_var <- "Mean.Scale.Score"

prediction_errors <- sapply(models, calculate_prediction_error,
test_data = test_data, response_var = response_var)

# Combine prediction errors into a dataframe
prediction_error_df <- data.frame(
  Model = model_names,
  Prediction_Error = prediction_errors
)

MSE_df = as.data.frame(matrix(prediction_errors, ncol = 3, byrow =
T))
colnames(MSE_df) = column_names
rownames(MSE_df) = row_names
```