"Multiple Regression Analysis: Prediction of Average Overall Achievement Score of schools in North Carolina"

By: Shcherbachuk Viktoriia, Sun Boru, Wong Jason, Li Angela, Olea Joshua Marc
DATA 603 - Project
Assistant Professor: Dr. Thuntida Ngamkham
University of Calgary
Calgary, Alberta
December 5, 2022

Contents

Introduction

In our research for this project, we discovered that many different factors contribute to an individual's success in terms of their education. The goal of this project is to explore these factors, and analyze how each factor negatively, or positively impacts an individual's success in the education system in order to be able to make informed decisions on the overall quality of education these students are receiving, as well as predict school achievement success.

According to Crossley (2003) poverty is associated with adverse childhood outcomes, one of which is described as 'inequality of opportunity'. We will look to explore topics such as poverty level, funding, resources available per student (books and devices), and criminal acts per student, to determine how each of these factors that are associated with poverty, lead to an inequality of learning opportunities for the students in each school.

Furthermore, Lieberman (2022) outlines that hiring enough qualified teachers has been a pervasive challenge for many schools, and that these shortages have a direct impact on a student's ability to learn in an ideal environment. Therefore, we will also look at factors such as number of teachers, total enrollment of students, and the overall quality of teachers in order to determine how they contribute to an individual's success in school.

Lastly, Long, Conger, & Latarola (2012) examined the associations between student's high school course-taking in various subjects and found that those who take rigorous courses are more likely to enroll in postsecondary studies when compared to high school students who did not take rigorous courses. The competitive nature of post-secondary enrollment leads us to assume that taking rigorous courses in high school lead to a higher chance of enrolling in postsecondary, and leads to a higher overall achievement score for the student's studies. It will be interesting to see if this holds any statistical significance. Thus, the last set of factors we will look to explore are courses that contain arts components such as music offered at a particular school, will impact the overall achievement score of the school.

Methodology

Data Source

In order to see the role that school facilities play in affecting the overall achievement score of students in a school, the open source dataset "School Report Cards (SRC)" was taken from: the North Carolina Department of Public Instruction.

The information in the datasets are aggregated together for all districts, charter, and alternative schools operating during the school year in different categories such as school grades for different subjects, level of performance for each teacher, distribution of various courses provided by school, financial funds statistics by state and local level per pupil for schools, media equipment presence by school, security level and rate of health atmosphere by school.

The North Carolina Department of Public Instruction provided the dictionary and zip file with multiple datasets in CSV format. All datasets have a key column "agency_code" which includes various information regarding school and staff performance together with aggregated scores for each school in every subject (Math, English and etc.) for each year.

The final research dataset was combined together in CSV format by the "agency_code" column, which represents the individual code for each school in the U.S state of North Carolina. Elementary, middle and high schools were combined together in order to increase the sample size, and all student subgroups were chosen for data collection. Race, gender, disability, immigration status and economic level of the students were not taken into consideration in that research. The data was collected in 2019 for this particular study.

Variables were selected based on the literature review of factors that could affect student performance as well as the number of accessible information in the datasets. In total, the sample size of 2190 schools was collected after missing values were deleted and the rest of the dataset was cleaned.

Variable explanation

Dependent variable:

Our *dependent variable is the overall achievement score*[score]. Based on the datasets dictionary, we can conclude that *overall achievement score* was collected based on the average scores from all subjects of all students in each school per year.

It is a decimal variable which has a range from the minimum value of (17.1) to the maximum value of (99.8). We will eliminate 134 data points where the overall achievement score was equal to zero. As our dataset is real-world data, we assumed that this was a mistake, as it doesn't make logical sense that a school would receive a score of zero at the end of the educational year. Below is a graph that shows the distribution of our dependent variable. After removing values of zero, we can see that our dependent variable is normally distributed. This tells us that we are able to move forward and build a multi-linear regression model.

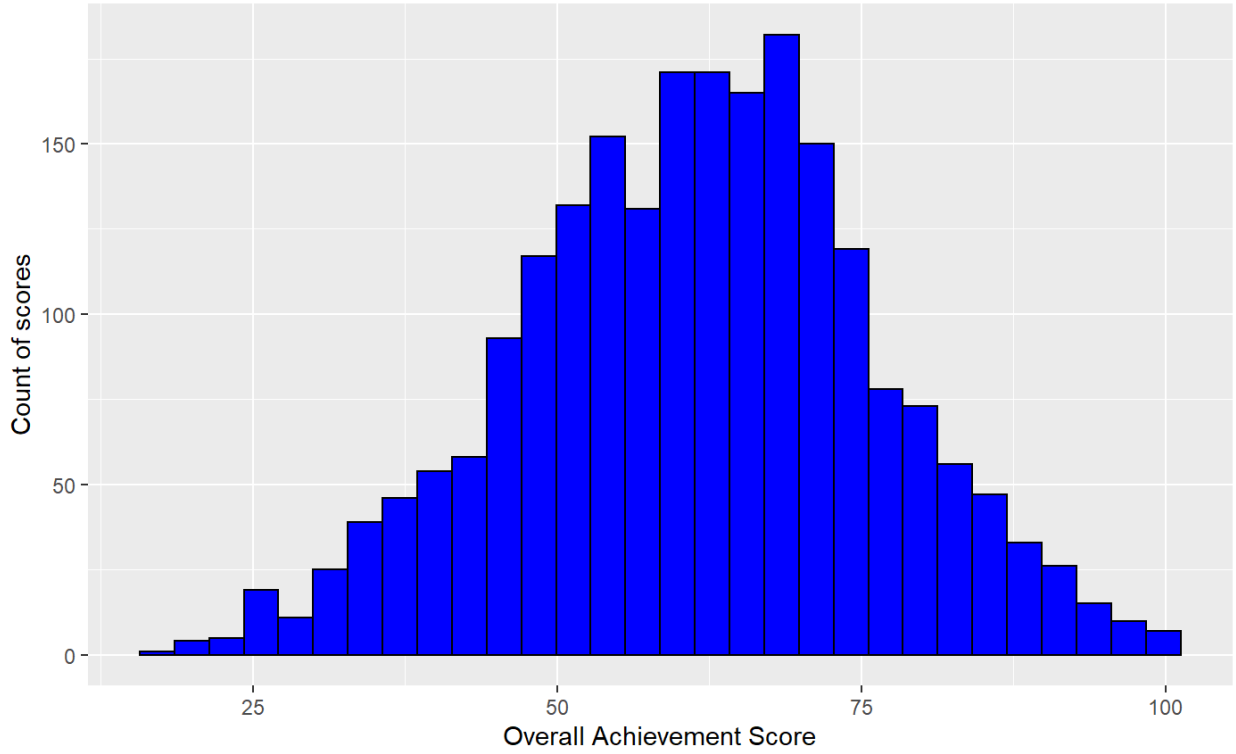

Figure 1: Histogram of Overall Achievement Score distribution

Independent variables:

Legend: - We used brackets to indicate the actual variable code that used to do our regression analysis, ie. [variable_name]

1. [growth_score] - a decimal numerical variable distributed having a range from 0 to 100, representing the EVAAS (Education Value-Added Assessment System) Growth Score. EVAAS provides North Carolina's educators with tools to improve student learning and to reflect and improve on their own effectiveness. EVAAS plays a valuable role in the success of North Carolina's schools and students.

2. [book] - a decimal numerical variable representing the number of book per student.

3. [criminal_act] - a decimal numerical variable representing criminal acts per 1000 students per school school.

4. [bullying] - a decimal numerical variable representing bullying and harassment accidents per 1000 students per school.

5. [staff] - a decimal numerical variable representing the percentage of staff with growth scores with a rating of "Highly Effective".

6. [devices] - a decimal numerical variable representing number of students per device.

7. [media] - a quantitative variable representing media collection age (year).

8. [salary_funds] - a numerical variable representing total funds at all levels on salaries in USD.

9. [pupil_funds] - a quantitative variable representing total per pupil expenditures across all funding sources in USD.

10. [n_teachers] - a quantitative variable representing number of teachers per school.

11. [goal_met] - a quantitative variable representing percentage of long term goal targets met.

12. [device_access] - a qualitative variable (Levels: Yes/No) representing whether school issues devices to each student.

13. [poverty_level] - a qualitative variable (Levels: High/Neither/Low) representing school poverty level.

14. [music] - a qualitative variable (Levels: Yes/No) representing whether music courses are offered at the school.

15. [theater] - a qualitative variable (Levels: Yes/No) representing whether theater courses are offered at the school.

Modeling plan

Our plan is to implement knowledge of statistical concepts from the DATA 603 "Statistical Modeling with Data" course offered by the Master's of Data Science & Analytics Program at the University of Calgary, and build the best multi-linear regression model for the prediction of overall achievement score of students at school. This will be based on educational and economical factors together with security and development level of schools.

The first step is to include all independent variables in a model and check the *multicollinearity assumption*. Then, we will conduct *the individual t-test* together with the *stepwise regression procedure* for finalizing the best first order model. It is important to say that we will test all our hypotheses with a significance level of less than $\alpha = 0.05$.

Secondly, we will check *the interaction terms* between all significant predictors of the overall achievement score and run *an Anova test* to indicate which model is better. As a final step, we will check model assumptions such as linearity, normality, equal variance assumption (heteroscedasticity) as well as check for any outliers. We will then see if our best model follows these assumptions or not. Based on the result, we will analyse how we can improve our model in order to satisfy all assumptions and implement them if possible.

Workload Distribution

Jason: Assist with overall multiple regression analysis, providing final edits to the report and ensuring that all information is accurate.
Viktoriia: Assist with overall multiple regression analysis, creating and formatting LaTex file, and discovering the best model for our regression.
Serena: Assist with overall multiple regression analysis, interpreting coefficients including interaction terms, and discovering the best model for our regression.
Josh: Assist with overall multiple regression analysis, testing the model to ensure assumptions are met, and discovering the best model for our regression.
Angela: Describing the dataset and identifying variables.

Results

Variable Selection Procedures

Before finding the best first order model to predict the overall achievement score of schools, we will check for any multi-collinearity between our independent variables in order to satisfy this assumption for our model. We will do this by calculating the variance inflation factor (VIF) value, which measures the strength of correlation between the predictor variables in a regression model. A value of 1 indicates that there is no correlation between a given predictor variable and any other predictor variables in the model. Values that are between 1 and 5 indicate moderate collinearity, but not enough to warrant any corrections. Values that are greater than 5 indicate potentially severe collinearity between the given predictor variables in the model, and will require correction if possible.

Below you can see our full model with all independent predictors which might affect the overall achievement score in schools.

Full model:

$$\begin{aligned}\widehat{Achievement}Score = \beta_0 + \beta_1 GrowthScore + \beta_2 Books + \beta_3 CriminalAct + \beta_4 Bullying + \\ \beta_5 HighlyEffectiveStaff + \beta_6 Devices + \beta_7 MediaYear + \beta_8 SalaryFunds + \\ \beta_9 PupilFunds + \beta_{10} NTeachers + \beta_{11} GoalMet + \beta_{12} DeviceAccess + \\ \beta_{13} PovertyLevel + \beta_{14} Music + \beta_{15} Theater\end{aligned}$$

Based on the VIF values table, we can eliminate one of the predictors (number of teachers and salary funding) from the model as their VIF values are considerably higher than 5. This means that these variables are highly correlated between each other (0.98 correlation value). To confirm this, we plotted these two variables and saw that there is a the strong pattern between number of teachers at school and school salary funding.

In the end, we decided to keep the "Salary funds" variable and eliminate "Number of teachers" from the model as the VIF value for "Salary funds" is higher (22.93) compared with the VIF value (22.26) for "Number of teachers". After removing the "Number of teachers" variable, we ran the VIF test again to make sure that other variables are not correlated with each other.

Hypothesis Statement for Individual T-tests:

After conducting the multi-collinearity test, we will proceed with the individual t-test to check which predictors are significant to keep in our model.

$$\begin{aligned}&H_0 : \beta_i = 0 \\ &H_a : \beta_i \neq 0 \text{ (i=14th independent predictors)}\end{aligned}$$

Predictors	VIF values
growth score	1.135110
book	1.531893
criminal act	1.288046
bullying	1.094324
staff	1.193074
devices	1.188783
Media	1.444687
salary funds	22.931199
pupil funds	1.540833
Number of teachers	22.260022
goal met	1.133859
factor(device access)	1.141918
factor(poverty level)	1.431744
factor(music)	1.154746
factor(theater)	1.519626

Figure 2: Table of VIF values for Full model

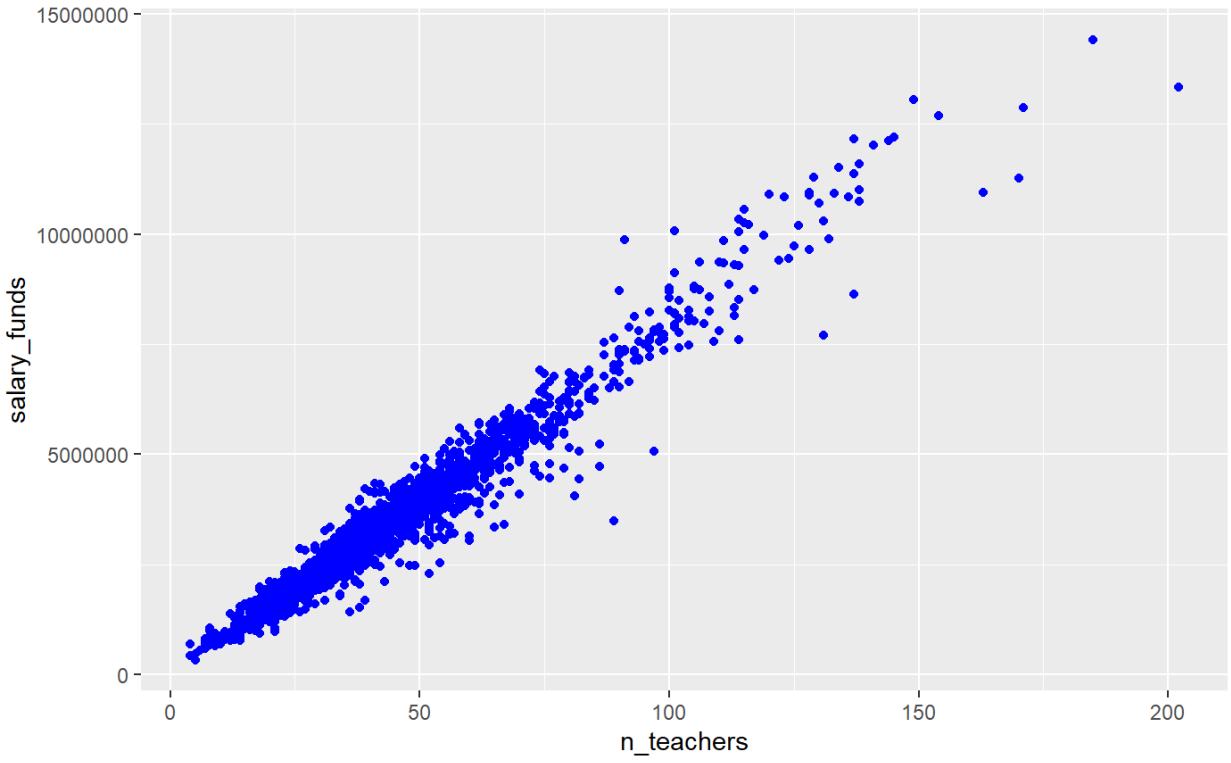

Figure 3: Correlation plot between "Number of teachers" and "Salary funds" variables

Predictors in a model:	P-value
(Intercept)	< 2e-16 ***
growth score	< 2e-16 ***
book	0.55541
criminal act	1.33e-06 ***
bullying	1.45e-11 ***
staff	< 2e-16 ***
devices	0.81842
media	3.10e-10 ***
salary funds	0.00115 **
pupil funds	1.67e-09 ***
goal met	< 2e-16 ***
factor(device access)Yes	0.29291
factor(poverty level)Low	< 2e-16 ***
factor(poverty level)Neither	< 2e-16 ***
factor(music)Yes	0.06918 .
factor(theater)Yes	2.62e-09 ***

Residual standard error: 9.246 on 2174 degrees of freedom, Adjusted R-squared: 0.6025

Figure 4: Individual T-test Summary. Model 1

By looking at the summary of the individual t-test of our model, we can conclude that these four highlighted independent variables (books per student, number of students per device, device access for students, music lessons at school) should be removed from the model as their p-values are greater than our $\alpha = 0.05$ for those variables. The other predictors are significant to include in our model to predict the overall achievement score of students per school. Therefore, we reject the null hypothesis in favour of the alternative.

The categorical variable "Music" has a p-value of (0.06918) which means that this variable is located in the "grey" zone. We decided to remove this variable after taking into consideration that we still have 10 other significant predictor variables.

The result of our individual T-test, allowed us to build the best first order regression model to predict the overall achievement score:

$$\widehat{AchievementScore} = \beta_0 + \beta_1 GrowthScore + \beta_2 CriminalAct + \beta_3 Bullying + \beta_4 HighlyEffectiveStaff + \beta_5 MediaYear + \beta_6 SalaryFunds + \beta_7 PupilFunds + \beta_8 GoalMet + \beta_9 PovertyLevel + \beta_{10} Theater$$

Predictors in a model:	Estimated Beta's coefficients	P-value
(Intercept)	46.18	< 2e-16 ***
growth score	0.2269	< 2e-16 ***
criminal act	-0.1396	7.27e-07 ***
bullying	-0.06835	9.26e-12 ***
staff	18.32	< 2e-16 ***
media	-0.005354	6.87e-13 ***
salary funds	-0.0000005052	0.000436 ***
pupil funds	-0.001164	1.31e-09 ***
goal met	0.09337	< 2e-16 ***
factor(poverty level)Low	22.95	< 2e-16 ***
factor(poverty level)Neither	10.25	< 2e-16 ***
factor(theater)Yes	3.629	2.58e-09 ***

Residual standard error: 9.248 on 2178 degrees of freedom, Adjusted R-squared: 0.6023

Figure 5: Individual T-test Summary. The best 1st order model.

By looking at the slope coefficients, we can give an interpretation:

1. $0.2269 * GrowthScore$: increasing 1 point of the Education Value-Added Assessment System growth score at school leads to an increase in the overall achievement score of students of that school on average by 0.2269 points when other predictors stay constant.

2. $-0.1396 * CriminalAct$: increasing the number of criminal acts per 1,000 students at school leads to a decrease in the overall achievement score of students of that school on average by 0.1396 points when other predictors stay constant.

3. $-0.06835 * Bullying$: increasing 1 case of bullying and harassment incidents per 1,000 students leads to a decrease in the overall achievement score of students of that school on average by 0.06835 points when other predictors stay constant.

4. $18.32 * HighlyEffectiveStaff$: increasing the number of teachers with a "Highly Effective" rating score at school leads to an increase in the overall achievement score of students of that school on

average by 18.32 points when other predictors stay constant.

5. $-0.005354 * Media$: increasing the age of media equipment of the school by 1 year leads to a decrease in the overall achievement score of students of that school on average by 0.005354 points when other predictors stay constant.

6. $-0.0000005052 * SalaryFunds$: increasing the total funds of teacher's salaries by \$1 leads to to a decrease the overall achievement score of students of that school on average by 0.0000005052 points when other predictors stay constant.

7. $-0.001164 * PupilFunds$: increasing pupil expenditures across all funding sources by \$1 leads to to a decrease in the overall achievement score of students of that school by 0.001164 points when other predictors stay constant.

8. $0.09337 * GoalMet$: increasing the long term goal targets of the school by 1% leads to an increase in the overall achievement score of students of that school on average by 0.09337 points when other predictors stay constant.

9. $22.95 * PovertylevelLow$: the average difference of the overall achievement score of students in schools with low poverty level compared with the school with high poverty level is 22.95 points.

10. $10.25 * PovertylevelNeither$: the average difference of the overall achievement score of students in schools with medium poverty level compared with the school with high poverty level is 10.25 points.

11. $3.629 * TheaterLessons$: the average difference of the overall achievement score of students in schools with theater courses provided compared with the schools where there is no theater courses is 3.629 points.

After summarizing our interpretations, we can conclude that the school development factor, decent security level, great staff performance, the financial assets of a school and whether or not a school has an art component are positively correlated to the overall achievement score of students.

In order to confirm our best first order model, we will proceed with the Stepwise Modeling procedure.

Stepwise Modeling procedure

One of the strategies we explored to find the best fit model for our data is the stepwise regression selection. The reason we chose this regression method over forward and backwards regression is because of the multitude of significant independent variables for our response variable (overall achievement score). Stepwise being the strictest of the 3 methods will allow us to narrow down our model to only a few variables as this method chooses which predictors are the most important, and replaces them if at any point they become insignificant by checking them twice with p-ent, and p-rem.

We performed stepwise modeling with the given parameters, *pent = 0.05* and *prem = 0.1*. As the goal of performing stepwise was to narrow down our significant variables, we decided to implement stricter conditions, rather than the default for stepwise. This way, we will increase our chances of picking the 'best' predictors.

After running the stepwise procedure, we ended up with the same model we had after doing our individual t-test.

Our best first order model has an adjusted R-squared of 0.6023. Considering that this is "real world" data, an adjusted R-squared value of 0.6023 is considered decent when we consider the fact that values for the individual student level is not present in our original dataset, and we are predicting the average overall achievement score of students in a particular school. Despite this, we will try to improve our adjusted R-squared by including any applicable interaction terms.

Predictors in a model:	Estimated Beta's coefficients	t-values	P-value
(Intercept)	46.18	21.534	0.000 ***
factor(poverty level)Low	22.95	33.790	0.000 ***
factor(poverty level)Neither	10.25	19.429	0.000 ***
growth score	0.23	13.015	0.000 ***
media	-0.005354	-7.226	0.000 ***
staff	18.32	11.553	0.000 ***
goal met	0.0934	10.039	0.000 ***
bullying	-0.0684	-6.855	0.000 ***
pupil funds	-0.00116	-6.092	0.000 ***
criminal act	-0.1396	-4.968	0.000 ***
factor(theater)Yes	3.6285	5.981	0.000 ***
salary funds	-0.0000005052	-3.523	0.000436 ***

Residual standard error: 9.248 on 2178 degrees of freedom, Adjusted R-squared: 0.6023

Figure 6: The Stepwise regression procedure "best" model summary

Interaction Terms

Starting from the reduced model we have in step 1, in order to improve it and have a better model that explains a larger portion of the variance in the response variable, we will consider any interaction terms. To test all interaction terms in our regression model, we will apply the Individual Coefficient Test method:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0 \text{ (i=interaction terms)}$$

After conducting all possible interactions between 10 independent predictors, we ended up with 8 significant interaction terms. Below is the output and summary of the final interaction model.

$$\begin{aligned}\widehat{AchievementScore} = \beta_0 + GrowthScore(\beta_1 + \beta_{15}GoalMet) + CriminalAct(\beta_2 + \beta_{16}PupilFunds) + \\ Bullying(\beta_3 + \beta_{17}Staff) + \\ Staff(\beta_4 + \beta_{12}PupilFunds) + Media(\beta_5 + \beta_{13}factor(PovertyLevel)) + \\ \beta_6 SalaryFunds + \beta_7 * PupilFunds + \\ GoalMet(\beta_8 + \beta_{18} * factor(PovertyLevel) + \\ \beta_{19}factor(PovertyLevel)) + \beta_9 factor(Theater) + \\ \beta_{10}factor(PovertyLevel) + \beta_{11}factor(PovertyLevel)\end{aligned}$$

We finalized the best model with interaction terms through a few steps. First, we will check all possible interactions and eliminate interaction terms where the p-value was greater than $\alpha = 0.05$. After checking the second interaction model, we eliminated [salary_funds:goal_met], [criminal_act:factor(poverty_level)] and [growth_score:salary_funds] as these interactions also had a p-value greater than 0.05. Lastly, we deleted the last insignificant interaction [growth_score:media] as this p-value was equal to $0.128554 > 0.05$. To provide a better understanding of our study, we will elaborate 3 interaction terms that we believe are the most significant.

Coefficient Interpretation of Interaction Terms: 1. poverty_level and goal_met

Interaction terms:	Estimated Beta's coefficients	P-value
(Intercept)	69.8197	0.000000000000383 ***
Goal met: factor(poverty level)LOW	-0.0931	0.000877 ***
Goal met: factor(poverty level)NEITHER	-0.0275	0.235215
media: factor(poverty level)LOW	0.0129	0.005995 **
media: factor(poverty level)NEITHER	0.0075	0.108684
staff:pupil_funds	0.00438	0.000070973476043 ***
bullying: Staff	-0.1996	0.007151 **
Criminal act: pupil funds	0.000072	0.000286 ***
Growth score: goal met	-0.00248	0.000039768384464 ***

Residual standard error: 9.083 on 2170 degrees of freedom, Adjusted R-squared: 0.6164

Figure 7: The best interaction model summary

The effect of increasing the percentage of long-term goals met will depend on the poverty level of a school. For example, if the school's poverty level is HIGH, then increasing the percentage of long-term goals met will decrease the school's overall achievement score. In other words, this relationship depends on what level of poverty the school has; if the poverty level is LOW, improving the percentage of long term goal targets met for a given subject area will increase the overall achievement score of a school by (0.3249(goal_met) + 0.7953* factor(poverty_level)LOW). If poverty level is NEITHER, improving the percentage of long term goal targets met for a given subject area will increase the value of overall achievement score of school by (0.3249(goal_met) + 4.4561* factor(poverty_level)NEITHER). Therefore, the effect of whether or not long-term goals are met will have minimal impact on a school's overall achievement score when the poverty level of the school is high.

Coefficient Interpretations of Interaction Terms: 2. criminal_act and pupil_funds

When looking to see the interaction between the number of criminal acts per 1000 students and the per-pupil expenditures, we discovered that if pupil funding is zero, then an increase in criminal acts per 1000 will decrease the school's overall achievement score by (0.6037(criminal_acts) + 0.00007238(pupil funds)). Therefore, the more per-pupil expenditure a school has, the higher the school's overall achievement score will be, which also lowers the effect that criminal acts will have.

Coefficient Interpretations of Interaction Terms: 3. bullying and number of staff

The interaction between the rate of bullying and the percentage of staff at a particular school indicates that an increase in bullying and harassment will depend on the percentage of staff available. In this case, if the percentage of highly effective staff at a school is zero, then an increase in bullying and harassment will decrease the school's overall achievement score by (0.02166(bullying) + 0.1996(staff)). Therefore, the greater the effect bullying and harassment has on the school's overall achievement score, the less impact that highly effective staff will have on increasing the achievement score, and vice-versa.

When comparing the adjusted R-squared from our best first order model (60.23%) with that of our interaction model (61.64%), we find an improvement of 1.41%. The residual standard error was reduced by 0.165 when compared with our best first order model. Therefore, we will use the final interaction model as our final model for predicting the overall achievement score of schools. As a final confirmation, we will run an ANOVA test and check that the interaction model truly is better at explaining the overall achievement score than our best first order model.

ANOVA table: First Order model vs. Interaction model

$$H_0 : \beta_i = 0 \text{ (i=interaction terms are not significant)}$$
$$H_a : \beta_i \neq 0 \text{ (i=interaction terms are significant)}$$

Source of variation	Df	Sum of squares	Mean squares	F-statistics	P-value
Regression	8	7,272	909	11.019	0.000000000000002352 ***
Residual	2170	179,016	82.49585		
Total	2178	186,288			

Figure 8: ANOVA table

The ANOVA test confirms that the interaction terms are significant enough to include in the final model for predicting the overall achievement score of schools. We reject the null hypothesis and conclude that the model with interaction terms is significantly better at predicting the overall achievement score of schools at a p-value of less than $\alpha = 0.05$.

Multiple Regression Assumptions

In the sections below, we will discuss all the tests that we performed on our model to ensure the assumptions of multiple linear regression are met and to address any issues that may arise.

Linearity Assumption

The linear regression model assumes that the relationship between the predictors and the response variable is linear. Using residual plots, we were able to check if our assumptions were met by analyzing any patterns or trends that may be present, the presence of which would indicate non-linearity. From the plot below, we see that there are no prominent patterns, therefore we can conclude that our model passes the linearity assumption.

Equal Variance Assumption

$$H_0 : \text{ Heteroscedasticity is not present (homoscedasticity)}$$
$$H_a : \text{ Heteroscedasticity is present}$$

Another assumption that a linear regression model must meet is that of homoscedasticity, meaning that the error terms have constant variance. The opposite of which is heteroscedasticity. In order to test our model for heteroscedasticity, we performed the Breusch-Pagan test shown below. The results indicate that our error terms do not have a constant variance and that the assumption for homoscedasticity is not met.

This is confirmed by the scale-location plot between the fitted values and standardized residuals of our best model with interaction terms. In our case, the red line indicates a downward trend rather than a horizontal plane, and the spread of the residuals is unequal along the range of predictors.

In order to pass the assumption of equal variance, we will try multiple methods to improve our model in the hopes that our model will show homoscedasticity.

The first transformation method we tried was by raising the power of some of our predictors to the second order. To help us figure out which variables we should raise to the second order, we used a ggpairs plot to show the most highly correlated variables in our model. The most highly correlated variables will be raised

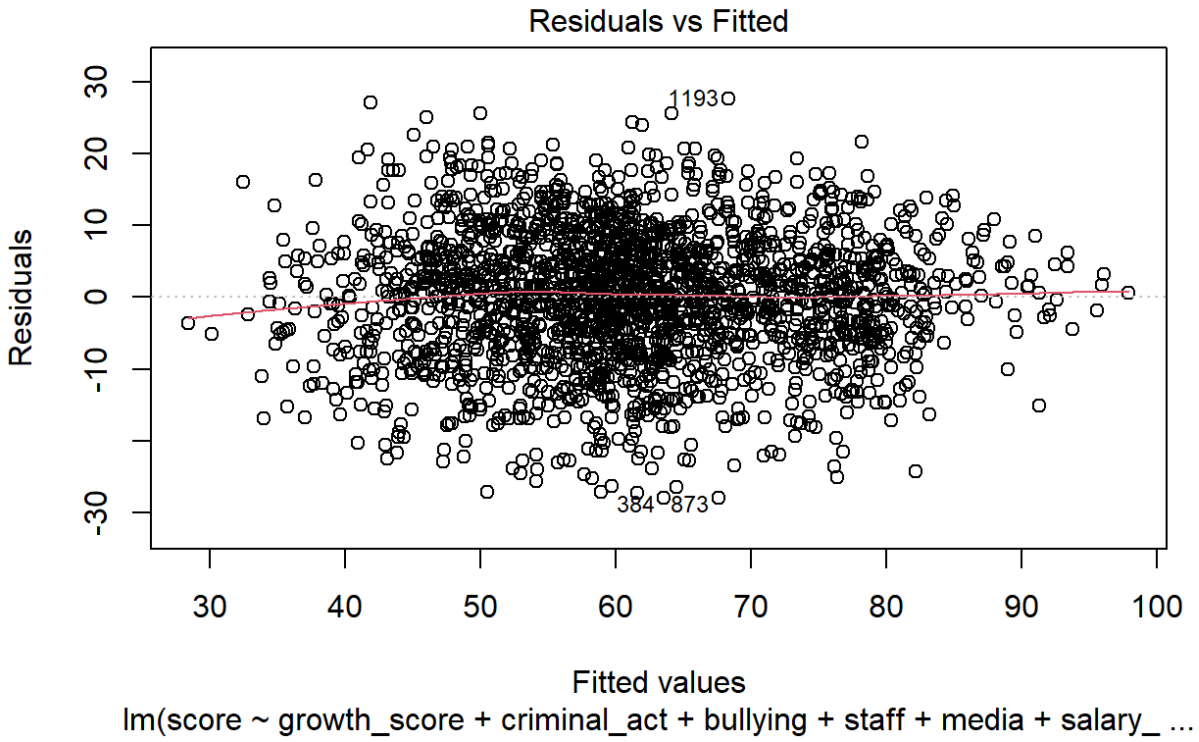

Figure 9: Residual plot

Breusch-Pagan Test:	
BP	92.911
DF	19
P-Value	0.00000000001007

Figure 10: The Breusch-Pagan test for the best model with interaction terms

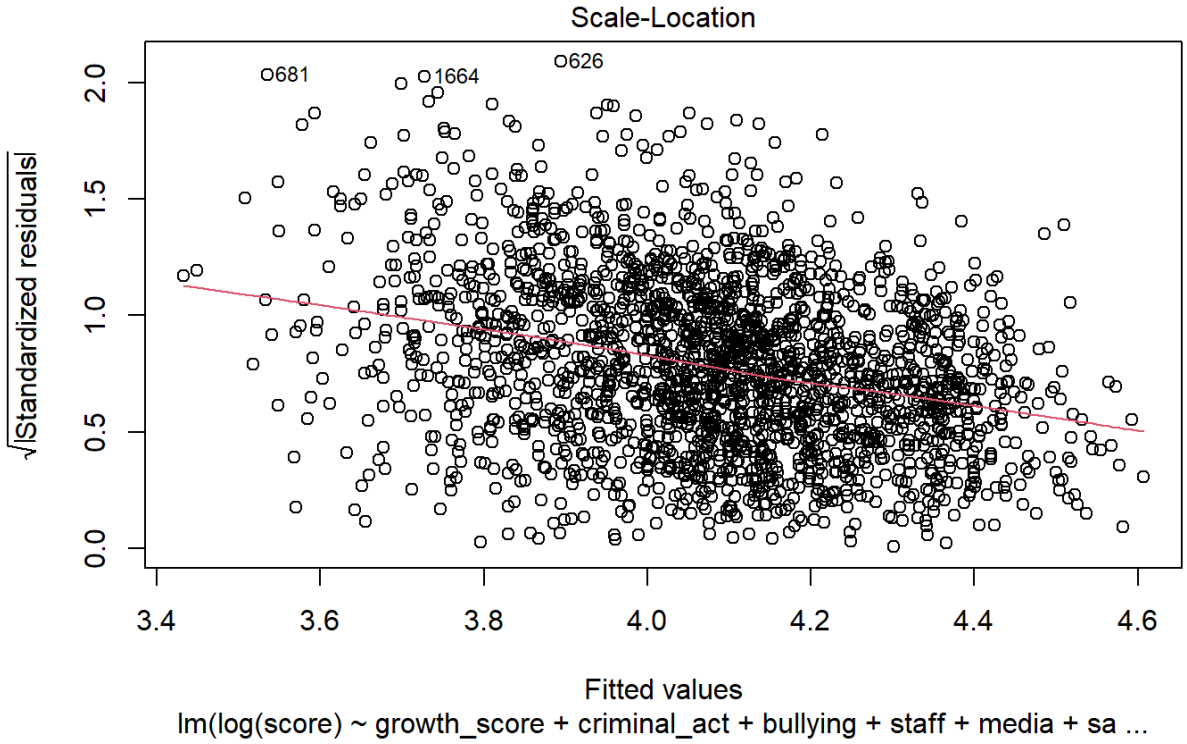

Figure 11: Spread-Location plot of the best model with interaction terms

to the second order, and we will test the model again for any signs of heteroscedasticity. The idea is to go one-by-one and check p-values of the second order variable.

After testing several trials of possible variations of second order models, we concluded that second-order transformations will not help our model meet the assumptions of homoscedasticity, as all predictors still had p-values of less than $\alpha = 0.05$.

Other methods that we tried were log-transformations and Box-Cox transformations of our dependent variable, but neither of these were able to help our model. The last method we tried was to implement Weighted Least Squares (WLS) onto our model where the observations with lower variance were given more weight and is used to correct for unequal variability or precision in observations. Since by using the method of WLS, each weight is inversely proportional to the error variance, therefore an observation with a small error variance has a large weight since it contains relatively more information than an observation with a large error variance (ie. an observation that has a small weight). Incorporating Weighted Least Squares (WLS) proved to be successful, as our p-values from the Breusch-Pagan test were now greater than $\alpha = 0.05$, and our model is finally able to satisfy the assumptions of homoscedasticity.

$$\begin{aligned}\widehat{Achievement}Score = \beta_0 + \beta_1 GrowthScore + \beta_2 CriminalAct + \beta_3 Bullying + \\ \beta_4 HighlyEffectiveStaff + \beta_5 MediaYear + \beta_6 SalaryFunds + \\ \beta_7 PupilFunds + \beta_8 GoalMet + \beta_9 PovertyLevel + \beta_{10} Theater + \\ \beta_{11} GoalMet * PovertyLevel + \beta_{12} Media * PovertyLevel + \\ \beta_{13} Staff * PupilFunds + \beta_{14} Bullying * Staff + \\ \beta_{15} CriminalAct * PupilFunds + \beta_{16} GrowthScore * GoalMet \\ weight = 1/lm(abs(modelResiduals)\ modelFitted.values) fitted.values^2\end{aligned}$$

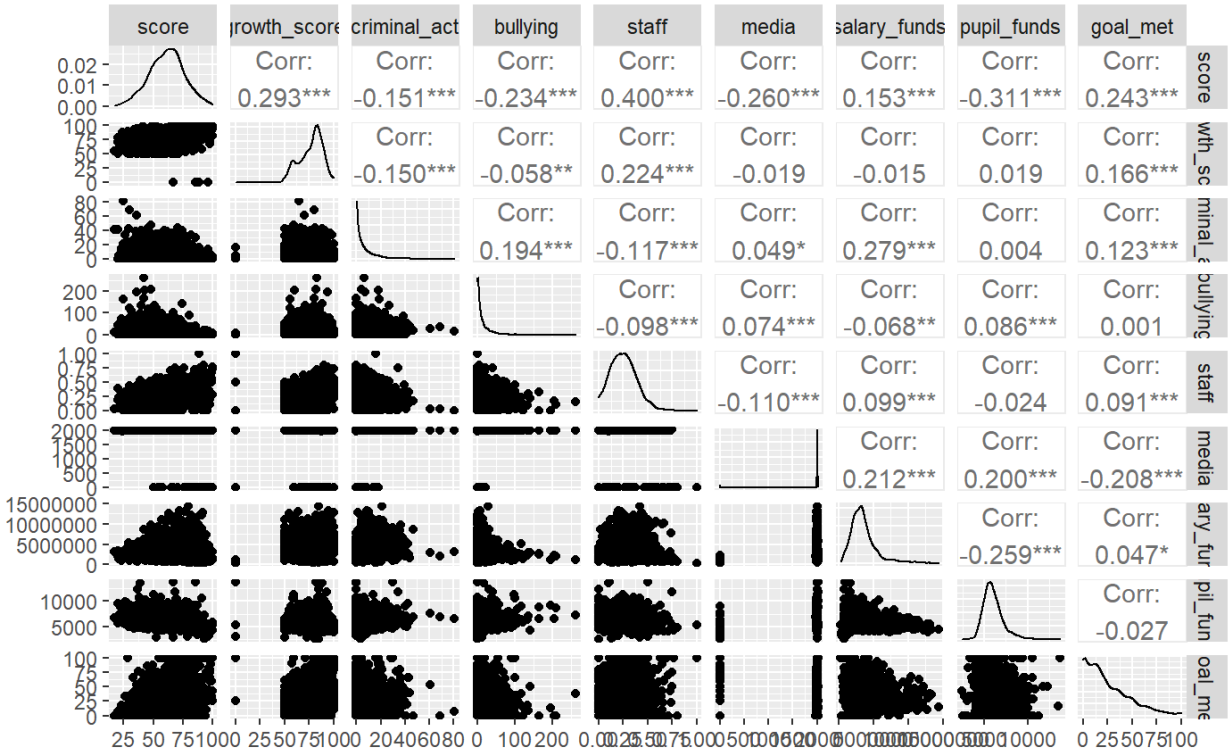

Figure 12: The ggpairs plot for the numerical variables in a model

Breusch-Pagan Test:	
BP:	14.75
DF:	19
P-value:	0.7384

Figure 13: The Breusch-Pagan test for WLS transformed the best model with interaction terms

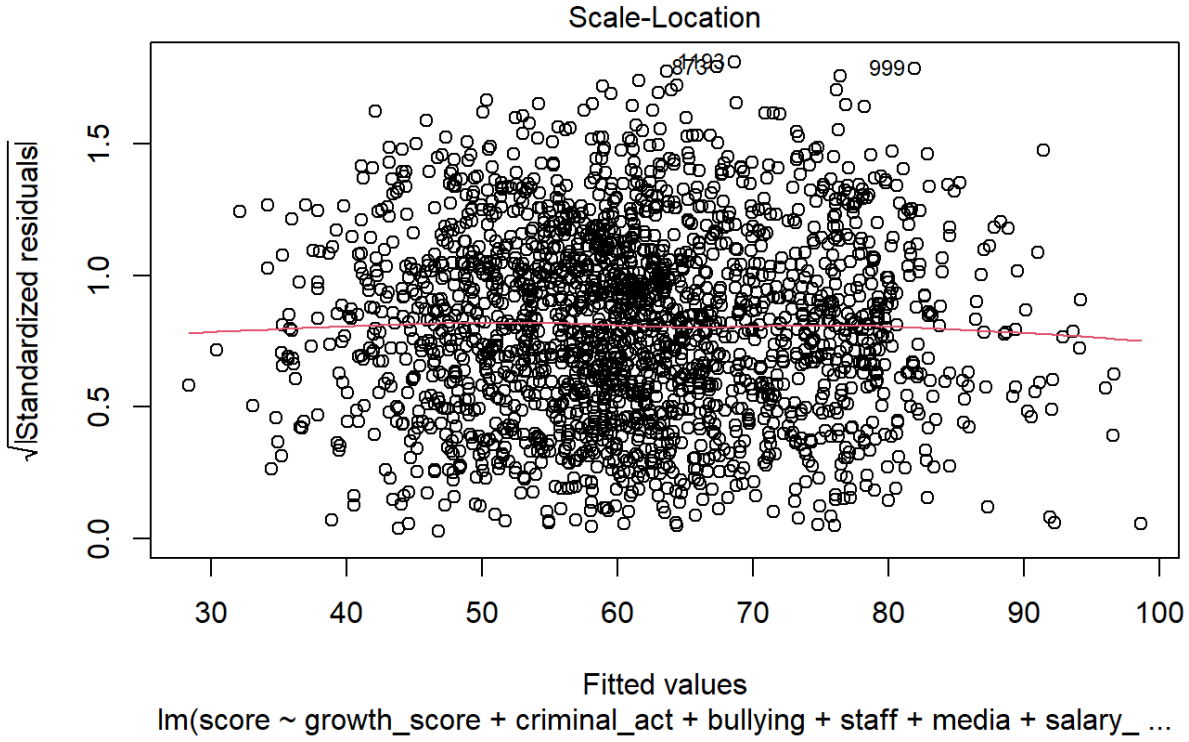

Figure 14: Spread-Location plot of the WLS transformed the best model with interaction terms

Normality Assumption

In order to satisfy the assumption of normality, the errors between the observed and predicted values (i.e. the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram, a normal probability plot or a QQ-Plot and the Shapiro-Wilk test. Our hypotheses are as follows:

$$H_0: \quad \text{The sample is normally distributed}$$
$$H_a: \quad \text{The sample is not normally distributed}$$

According to the QQ-Plot below, the residuals of the model seem to follow a normal distribution as the points fall along the diagonal reference line. However, after conducting the Shapiro-Wilk test, we ended up with a p-value $< \alpha = 0.05$. Therefore, we reject our null hypothesis and conclude that our data is not normally distributed, and some transformations must be done in order to satisfy this requirement.

We implemented a Box-Cox transformation on our WLS transformed model using $\lambda = 1.393939$. Upon conducting the Shapiro-Wilk test on our further transformed model, we ended up with a p-value of 0.255 $> \alpha = 0.05$, therefore we fail to reject our null hypothesis, and conclude that our transformed model is normally distributed.

Thus, our Final Model for predicting overall achievement score in schools is shown below:

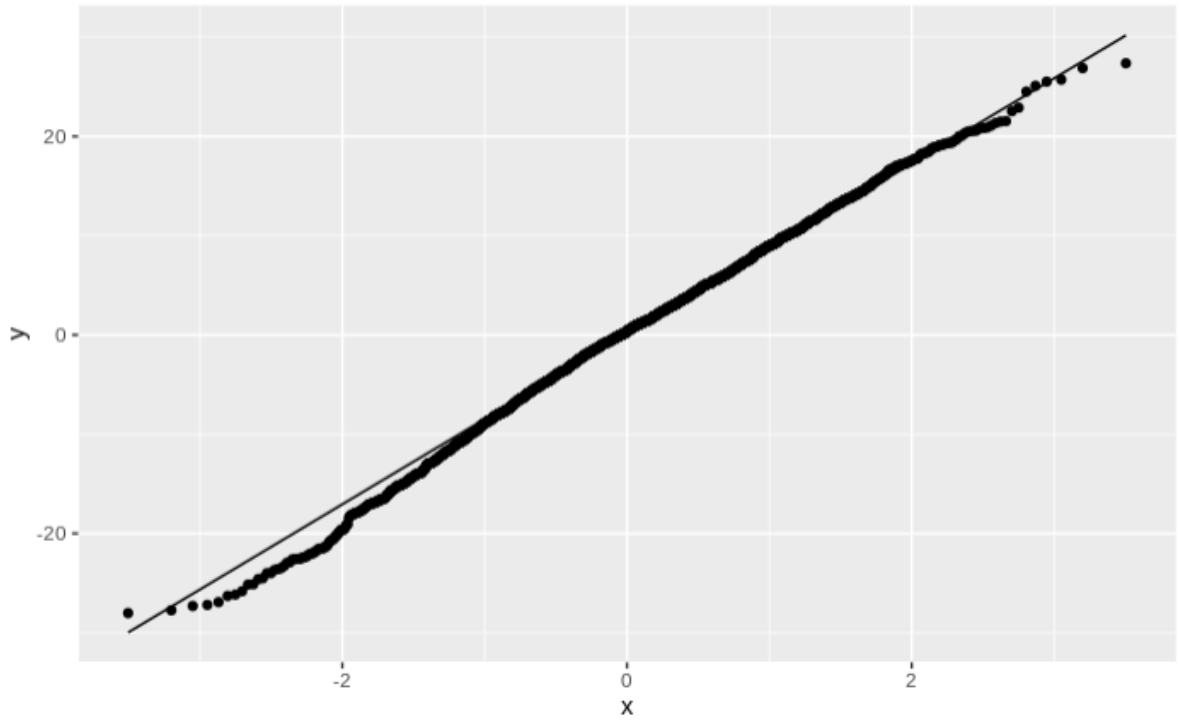

Figure 15: Probability Plot of the WLS transformed the best model with interaction terms

Shapiro-Wilk Test:	
W	0.99693
P-Value	0.0002238

Figure 16: The Shapiro-Wilk test of the WLS transformed model

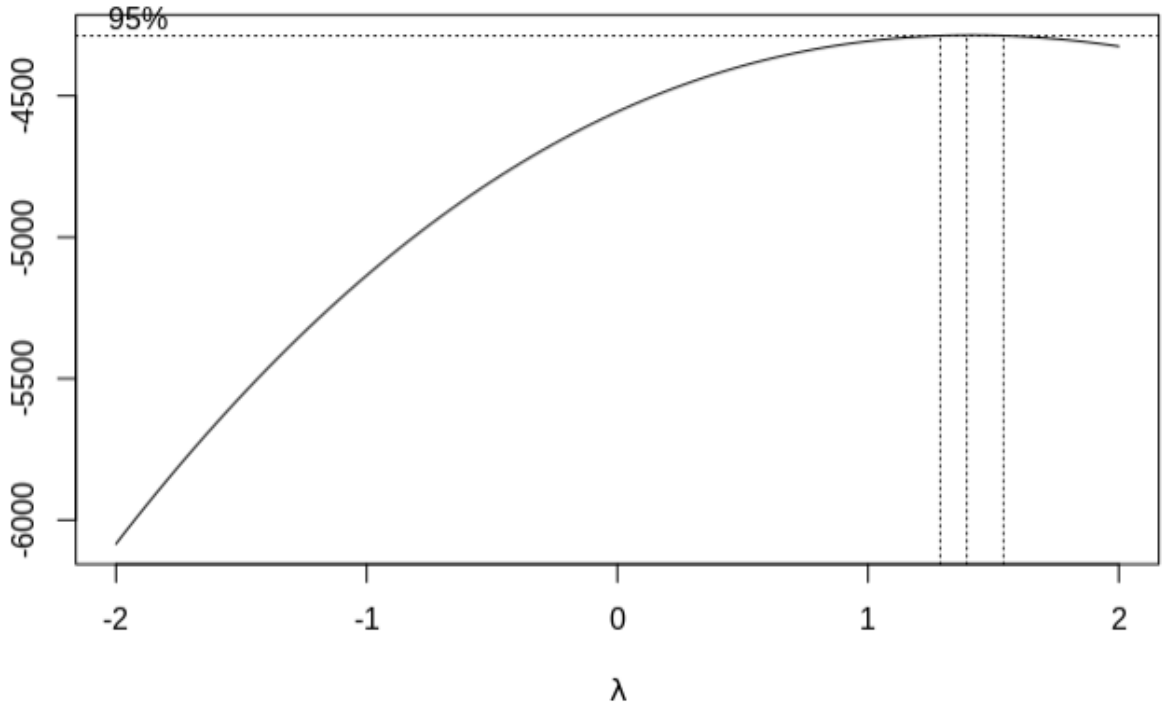

Figure 17: Finding lambda for the Box Cox transformation of our WLS transformed model

Shapiro-Wilk Test:	
W:	0.998
P-value:	0.255

Figure 18: The Shapiro-Wilk test of the Box Cox transformed WLS model

$$\widehat{AchievementScore}^{1.393939} = \beta_0 + \beta_1 GrowthScore + \beta_2 CriminalAct + \beta_3 Bullying + \beta_4 HighlyEffectiveStaff + \beta_5 MediaYear + \beta_6 SalaryFunds + \beta_7 PupilFunds + \beta_8 GoalMet + \beta_9 PovertyLevel + \beta_{10} Theater + \beta_{11} GoalMet * PovertyLevel + \beta_{12} Media * PovertyLevel + \beta_{13} Staff * PupilFunds + \beta_{14} Bullying * Staff + \beta_{15} CriminalAct * PupilFunds + \beta_{16} GrowthScore * GoalMet$$

$$weight = 1/lm(abs(modelResiduals)\ modelFitted.values) fitted.values^2$$

Outliers and Influential Points

Influential cases and Outliers can sometimes have effects that significantly impact our model. To check for any outliers in our data, we will plot the values against Cook's Distance, with the criteria that values having a Cook's Distance of greater than 0.5 as being outliers. After plotting the values in the graph shown below, we found one outlier in our data which was observation 74.

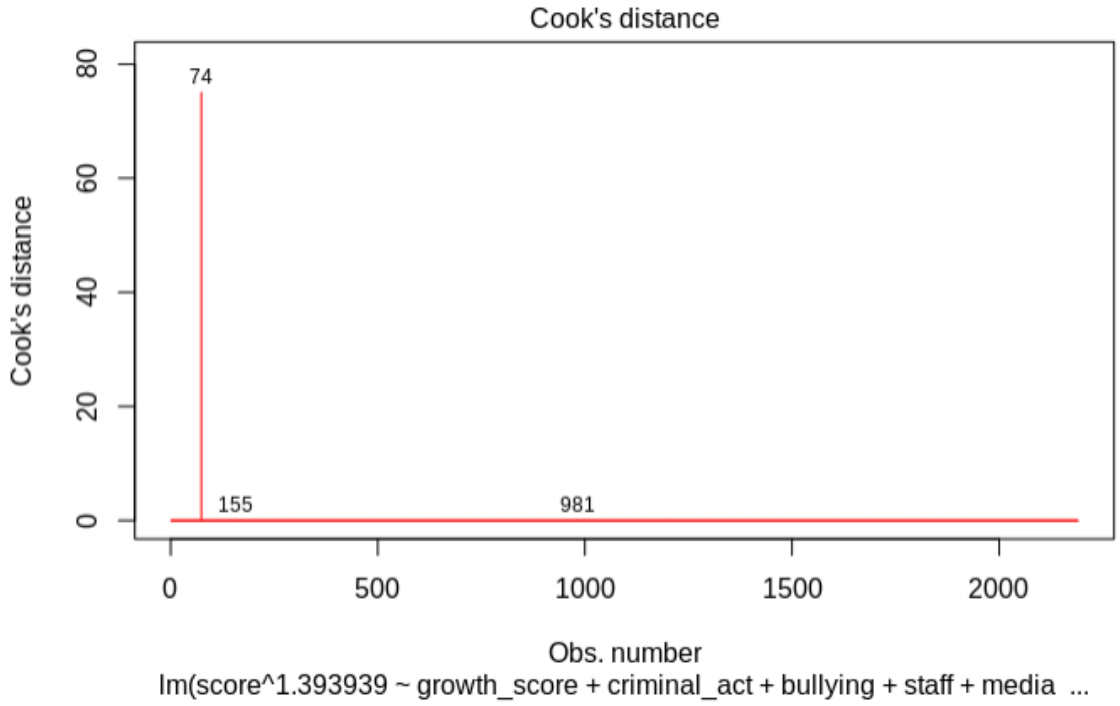

Figure 19: Leverage points of our final Box Cox transformed WLS model

To test the significance of this outlier, we will create a model using a subset of data with the outlier removed, and compared the adjusted R-squared and RMSE value to that of our model that includes the outlier. We discovered that the differences in the adjusted R-squared and RMSE values between the models were minimal, therefore, we decided to keep the outlier as removing it will not result in any meaningful changes.

Conclusion and Discussion

The main challenge that we faced was the issue of meeting the assumptions for homoscedasticity. This may be because our model was missing some key variables that were related to predicting the overall achievement

score of schools in North Carolina, and these factors may not have been present in the dataset to begin with. There were also a considerable amount of missing values for some of the variables, which did not allow us to include them in our model. Another factor that may have influenced this was that the nature of our data only allowed us to consider the achievement score of the school as a whole, rather than going into the individual student level. The large amount of nuance when dealing with human-focused research provides us with a level of uncertainty when it comes to ensuring that our model meets the appropriate assumptions.

Even though we were able to a build a model that can predict the overall achievement score of schools, the predictions from this model may not be very precise, as the adjusted R-squared value of 61.64% is lower than what we would like to see. Again, we point to the unpredictability of human-beings, as well as the nuances of human-based data, and conclude that these factors may be a key driver to the unexplained variability we see. Variables such as "Number of Highly Effective Staff" have no explanation on what criteria needed to be met for the staff member to be considered "Highly Effective", which further emphasizes the subjectiveness in parts of our data, as one school's criteria for rating staff members may not be the same as other schools. We draw comparisons to psychological studies, many of which have R-squared values of less than 50%. Despite our adjusted R-squared value being on the lower end, the low p-values we discovered still indicate a real relationship between the significant predictors and the response variable.

Overall, the approach we took yielded some meaningful results, and the adjusted R-squared we achieved was satisfactory considering the multitude of factors within our dataset that could influence our model. In the future, we would like to consider building a model containing only objective measurements and see if that improves our adjusted R-squared. We would also like to find some data that would allow us to dive into the individual level, and predict student achievement rather than school achievement as a whole.

References

Crossley, T. (2003). Child poverty in Canada, Canadian Electronic Library. Retrieved from https://canadacommons.ca/artifacts/1203557/child-poverty-in-canada/1756666/ on 16 Nov 2022. CID: 20.500.12592/pgj6d6

Lieberman, M. (2022, June 15). How school staffing shortages are hurting students. Education Week. Retrieved November 15, 2022, from https://www.edweek.org/leadership/whos-at-risk-when-schools-staffing-shortages-persist/2022/06

Long, M. C., Conger, D., & Latarola, P. (2012). Effects of High School Course-Taking on Secondary and Postsecondary Success. American Educational Research Journal, 49(2), 285–322. http://www.jstor.org/stable/41419458

Feldman, Marissa A., et al. (2014) "THE EFFECTS OF MIDDLE SCHOOL BULLYING AND VICTIMIZATION ON ADJUSTMENT THROUGH HIGH SCHOOL: GROWTH MODELING OF ACHIEVEMENT, SCHOOL ATTENDANCE, AND DISCIPLINARY TRAJECTORIES: Effects of Middle School Bullying." Psychology in the Schools, p. n/a-n/a. DOI.org (Crossref), https://doi.org/10.1002/pits.21799.

Polanin, Joshua R., et al. (2021) "A Meta-Analysis of Longitudinal Partial Correlations between School Violence and Mental Health, School Performance, and Criminal or Delinquent Acts." Psychological Bulletin, vol. 147, no. 2, pp. 115–33. DOI.org (Crossref), https://doi.org/10.1037/bul0000314.