Project Report

David Truong, Dan Le, Phuong Huynh

December 5, 2023

Which possible factors influence K-12 students' performance on The California Assessment of Student Performance and Progress (CAASPP) System?

1 Introduction

Adverse Childhood Experiences (ACEs) are traumatic events that occur during childhood and can have lasting effects on an individual's physical and mental health. As of 2017, the National Resilience Institute reported that by the age of 18, 72% of children are anticipated to have encountered one or more ACEs [1]. The substantial impact of ACEs on children is evident across multiple dimensions, significantly influencing cognitive development, emotional well-being, and overall health. Particularly noteworthy is the influence on educational outcomes, as these experiences play a critical role in shaping students' academic trajectories and molding.

The purpose of this analysis is to identify potential factors that may impact the performance of K-12 students. This will be achieved by examining the relationship between students' scores that do not meet the standard and various factors that could influence these outcomes across schools in the state of California.

2 Dataset and Statistical Methodology

2.1 Data Collection and Processing

Our research aims to investigate the impact of environmental factors, such as the characteristics of a school's neighborhood (including diversity rate and crime rate), in conjunction with student self-factors, on the performance of students in the 2022 CAASPP tests. The metric used to gauge performance is the average percentage of students falling below the standard, sourced from school-level data on the CAASPP website. Within this dataset, we calculate the student "give up rate" from the number of students enrolled and the number of students with scores, meaning that the rate will be the percentage of students from each school not taking any CAASPP test. Additionally, other predictors including the Chronic Absenteeism Rate and the Percentage of students eligible for Free or Reduced-Price Meals (K-12), aggregated from student poverty data available on the California Department of Education webpage. Beyond these student-related factors, we consider surrounding environmental factors, including the Diversity Index, Violent Crime, and Property Crime for each county or city in California, obtained from the Government Census webpage.

Since we have a large data source, we first used Python and Pandas to process and aggregate them into a complete dataset for analysis. We merged and cleaned the dataset from CAASPP data and California Department of Education data on the "school code" variable since they share this same variable. Then, we used the complete table to add variables from the Government Census dataset on the "school code", "county code", or "zip code" variables based on the aggregated data. It's important to recognize that aggregating data from different sources has the potential to reduce the accuracy and importance of the model. We attentively consider the potential loss of data that

could influence the final outcome. We noted the potential dropped observations resulting from aggregating the data. We also pay close attention when dropping null values or observations that do not provide numerical values. Since the dropping rate is very low, around 0.01%, it is reasonable to use our complete table for this regression analysis. We, then, use R to build regression models and perform regression analysis.

2.2 Method

To investigate potential influences on the academic performance of K-12 students, a multiple linear regression approach was employed. The Percentage Standard Not Met was chosen as the response and examined to determine if it significantly depended on any of the predictors. Our quantitative predictors included Give Up Rate, Chronic Absenteeism Rate, FRPM (Percentage of students eligible for Free or Reduced-Price Meals), Student Enrolled (number of students enrolled in each school), Diversity Index, Violent Crime, and Property Crime and our categorical predictors included County Name and Test ID.

In the regression analysis, we employed variance inflation factor (VIF) to assess the multicollinearity among the selected variables. Scatter plots were used to inspect the need for transformations on the independent variables (x), while the box-cox method was applied to assess the necessity of transformations on the dependent variable (y). Variable selection was carried out through forward, backward, and stepwise methods to identify the most influential predictors for the model. The predictors with the greatest significance according to these methods were FRPM, Chronic Absenteeism Rate, and Test ID. The predictors with the least significance seem to be Violent Crime, Property Crime, and Diversity Index. After applying these variable selections, we agreed upon a reduced model that dropped the predictor Violent Crime because it did not significantly contribute to the model and the p-value did not meet our threshold. Furthermore, we tested for interaction terms within the model.

At each stage, we conducted model adequacy checks and residual analyses to ensure adherence to residual assumptions and performed statistical tests to assess the significance of slopes. A significance level of $\alpha = 0.001$ was set as the threshold for the rejection region. To evaluate the accuracy of our model in predicting new observations, we utilized metrics such as R^2 and $R^2_{Prediction}$. We also considered MS_{Res} values and examined residual plots. The data, then, was split into training and testing sets, and through 1000 simulations, we derived 95% confidence intervals for $R^2_{Prediction}$ resulted values to estimate the model's accuracy.

2.3 Results

We find a positive relationship between Test ID, Student Enrolled, Chronic Absenteeism Rate, Give Up Rate, FRPM, and Property Crime predictors and a slightly negative relationship between Diversity Index, with the Percentage Standard Not Met for schools. The p-values for all slopes are remarkably low, under 0.001. Initially, we see that the number of students enrolled and Give Up Rate do show a clear linear connection with the response variable. This is confirmed by a model using only Student Enrolled and a model only using Give Up Rate as the predictors, resulting in a very low R^2 of 0.003. Nevertheless, we choose to keep these variables based on the low p-value from variable selection, confirming the significance of the variable for our regression model.

We understand that including Property Crime, Diversity Index, County Name, and Violent Crime will create multicollinearity among these variables, but we would like to use these predictors and explore which of these will affect the students' performance the most. After fitting a general model that includes all of the predictors, we noted a very high multicollinearity in the model which gives the null estimated slope for Diversity Index. Dropping either County Name or Diversity Index could solve the problem, so we decided to fit two separate models with either County Name or Diversity Index. We performed the same task for Property Crime and Violent Crime after VIF calculation that yields the high multicollinearity between these predictors. County Name and Violent Crime are omitted after the considerations. We, then, confirmed with the variable selection method and concluded that no further variable omission was needed.

Once we identified the crucial predictors for our model, we applied the Boxcox method to discover the optimal transformation for the response variable, setting λ to 0.5. Better constant variance and normal distribution also suggested that the model with transformed response fitted the data better. In addition, we explored interaction terms while constructing our predictive model. While the F-test statistics demonstrated the significance of the model with interaction terms compared to the one without them, with a p-value significantly lower than 0.001, the inclusion of interaction terms resulted in a more notable violation of residual assumptions. As a result, we chose to stick with our original model.

We conducted a thorough examination of outliers, leverage points, and influential observations, utilizing values derived from the H-matrix and Cook's Distance to identify these data points. Subsequently, we re-fitted the model, temporarily excluding the 20, 200, and 1000 most influential and high-leverage observations. Notably, we observed only a marginal increase in the R^2 values.

After a careful examination of various potential models, our concluding model reasonably satisfies the residual assumptions, ensuring normality and constant variance. The model can explain around 68% of the variation in the response variable based on the R-squared value. The accuracy also is confirmed after simulating splitting the data, fitting the same model for the training data, and using the model the predict the testing data 1000 times. The 95% CI of $R^2_{Prediction}$ values

between 0.615 and 0.65 was obtained indicating a high efficiency of our model in predicting new incoming observations.

3 Discussion

In summary, a meaningful relationship exists between self-factors and certain environmental factors influencing students' academic performance. Educators and policymakers are increasingly alarmed about student absenteeism, particularly focusing on the emerging concern of chronic absenteeism. Chronic absenteeism is associated with lower levels of school readiness, a higher likelihood of not reading at grade level by the third grade, diminished social engagement, an increased risk of dropping out of school, and a lower likelihood of graduating from high school or attending college [2]. Our model strongly indicates that as the average absentee rate of students increases, there is a notable correlation with academic performance falling below the standard on the CAASPP tests. Schools with higher absence rates tend to have a greater likelihood of students not meeting the expected standards on these tests. In association with the number of students enrolled in each school, both this factor and the Give Up Rate predictor demonstrate a linear relationship with students' performance. There is also a significant difference between test type one and test type two. Using a 95% confidence interval, we can confidently say that students performed slightly better on the English exam than on the Math exam, indicating that students were more likely to struggle with Math problems.

In addition, national assessments have highlighted a concerning statistic almost two-thirds of all fourth graders in the United States do not read on grade level, and the situation is even more alarming for students from low socioeconomic backgrounds [3]. Two variables, violent crime and property crime, were considered as indicators of low socioeconomic status and investigated as potential factors influencing students' performance. However, the analysis revealed that violent crime does not significantly explain students' performance. In contrast, property crime exhibits a negative relationship with students' performance. This suggests that the average property crime within counties does not have a significant impact on average low students' performance. While the Average Diversity Index within each county may not directly signify a student's low socioeconomic background, it remains a factor worthy of consideration in the comprehensive assessment. Notably, a positive relationship was identified, but its impact appears to be less pronounced compared to other predictors. In our data, the Percentage of students eligible for Free or Reduced-Price Meal serves as another indicator of students' low socioeconomic backgrounds. Eligibility for Free or Reduced-Price Meals is determined by the income level of students' households. Our model indicates a strong correlation between this factor and the Standard Not Met, suggesting that as the percentage of Eligibility for Free or Reduced-Price Meals increases there is an increase in low

academic performance by students. Similarly, as the Chronic Absenteeism Rate increased there was a significant increase in Standard Not Met, suggesting that schools with high absent rates were also likely to have low academic performance.

Despite the limitation of accessing a dataset with an appropriate number of predictors, aggregating data from different sources gives us more information that could be proven to be the factors impacting the students' low CAASPP test scores. While the model may not be deemed final solely based on this limitation, future investigations could enhance this work by acquiring a more robust data source.

References

[1] Snodgrass, J. (2021). The Impact of Targeted Social and Emotional Learning Strategies on Middle School Students' Academic Achievement in Mathematics: A Quantitative Study. ProQuest Dissertations Publishing.

[2] Lara, J., Noble, K., Pelika, S., & Coons, A. (2018). Chronic Absenteeism. NEA Research Brief. NBI No. 57. National Education Association.

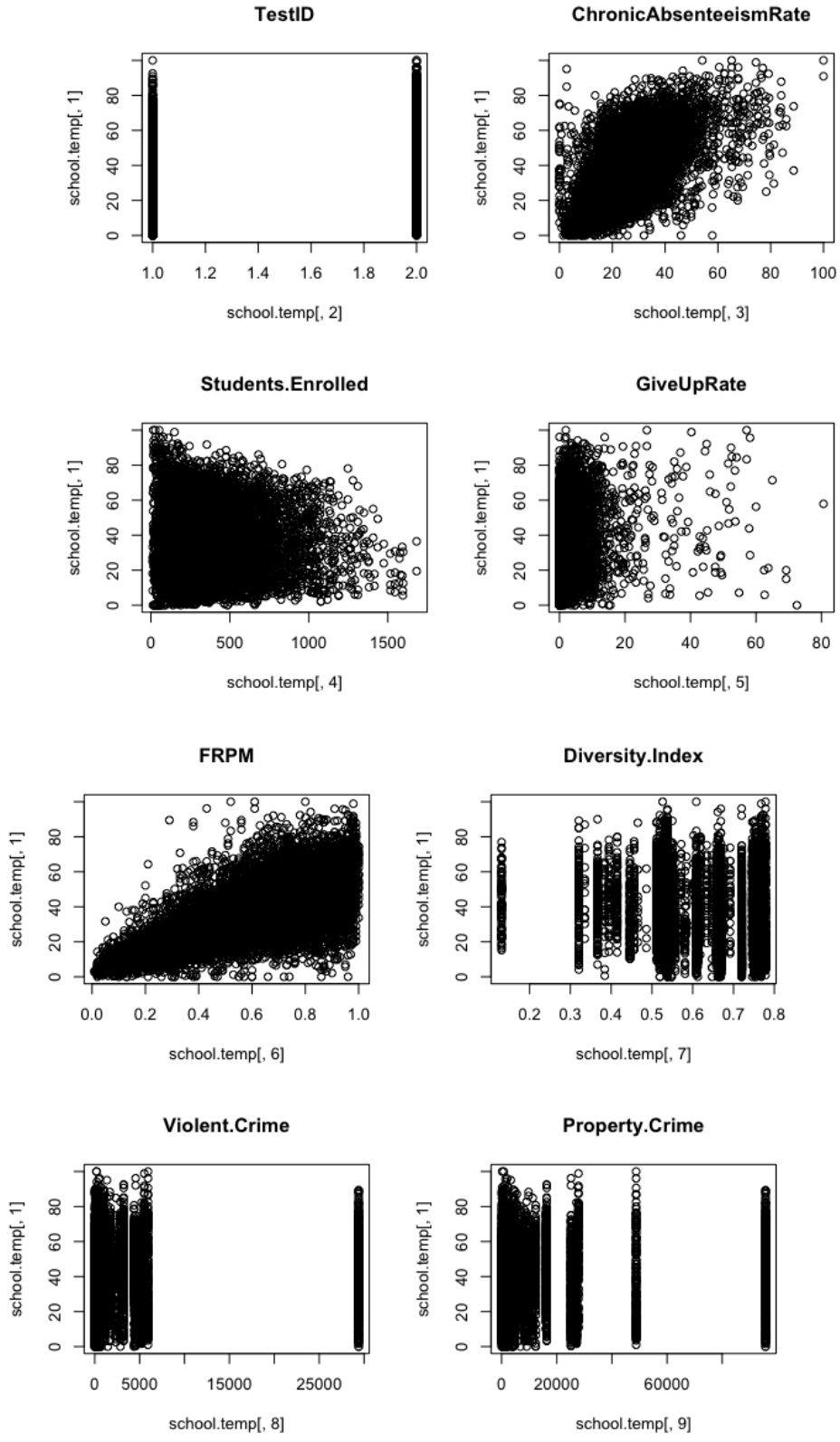
[3] Glaesmann, C. D. (2020). The Effects of Chronic Absenteeism in Kindergarten on Third-Grade STAAR Scores. ProQuest Dissertations Publishing.

4 Appendix

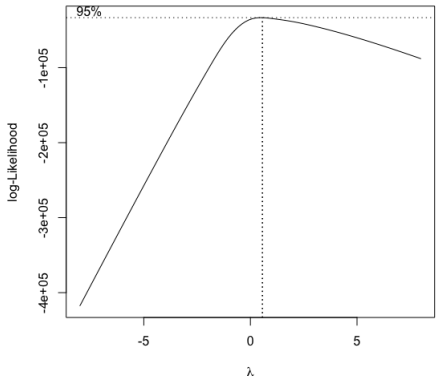
4.1 Final Model

```
Call:
lm(formula = I(sqrt(Percentage.Standard.Not.Met)) ~ Test.ID +
    Students.Enrolled + ChronicAbsenteeismRate + GiveUpRate +
    FRPM + Diversity.Index + Property.Crime, data = school)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2358 -0.6135  0.0264  0.6389  4.5198

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.467e+00  6.900e-02  21.263  < 2e-16 ***
Test.ID2              6.660e-01  1.930e-02  34.502  < 2e-16 ***
Students.Enrolled     4.635e-04  4.191e-05  11.060  < 2e-16 ***
ChronicAbsenteeismRate  4.470e-02  1.017e-03  43.953  < 2e-16 ***
GiveUpRate            1.855e-02  2.019e-03   9.190  < 2e-16 ***
FRPM                  3.862e+00  4.862e-02  79.418  < 2e-16 ***
Diversity.Index       4.857e-01  9.159e-02   5.303 1.16e-07 ***
Property.Crime       -6.323e-06  4.240e-07 -14.914  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9838 on 10415 degrees of freedom
Multiple R-squared:  0.676,     Adjusted R-squared:  0.6758
F-statistic:  3104 on 7 and 10415 DF,  p-value: < 2.2e-16
```

4.2 Individual predictor against response scatter plot

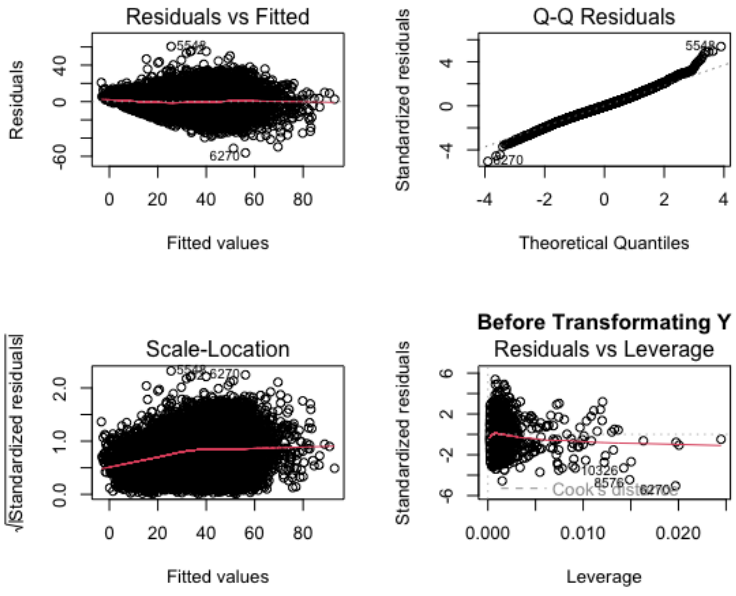


4.3 Boxcox

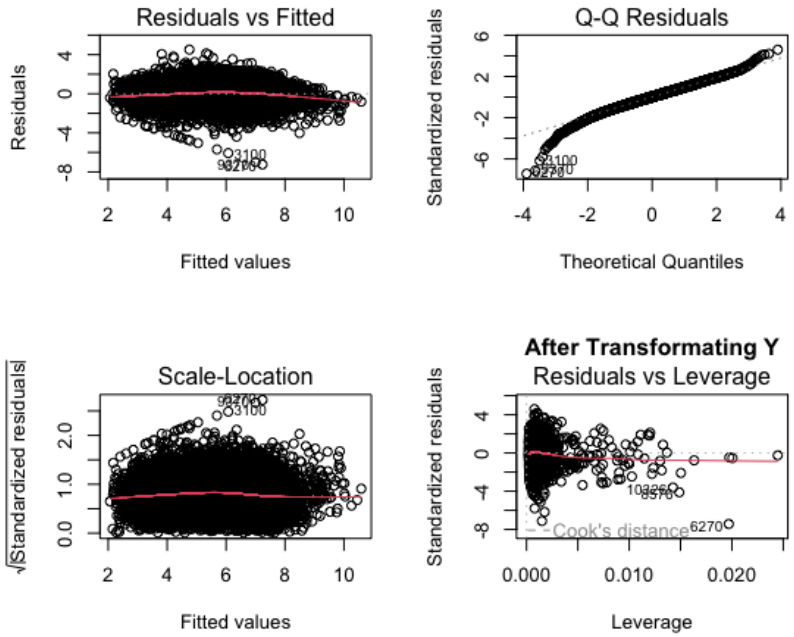


4.4 Residual Analysis

Before transforming the response (Final model)



After transforming the response



4.5 $R^2_{Prediction}$