Chicago Public Schools Mathematics and Literacy Skills

Austen Lowitz, Annie DeForge, William Teng

Contents

Introduction

In an effort to enhance education in Chicago, we aim to gain an understanding of the factors that contribute to a positive learning experience for students. We seek to identify evidence-based actionable insights for schools and policymakers throughout the Chicago Public School District to make informed decisions so that students can thrive academically. This can range from staffing and leadership decisions, school security code changes, teaching staff rostering, and discipline policies.

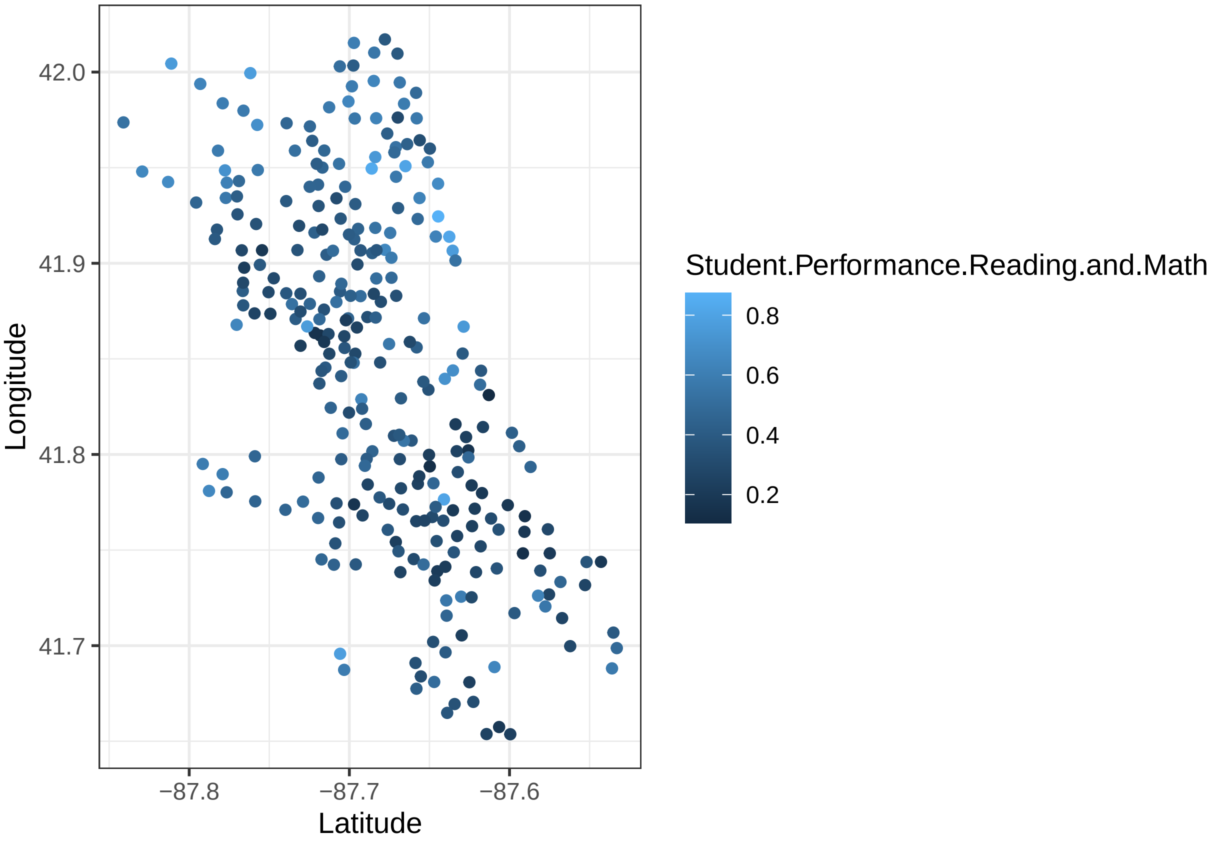
Data

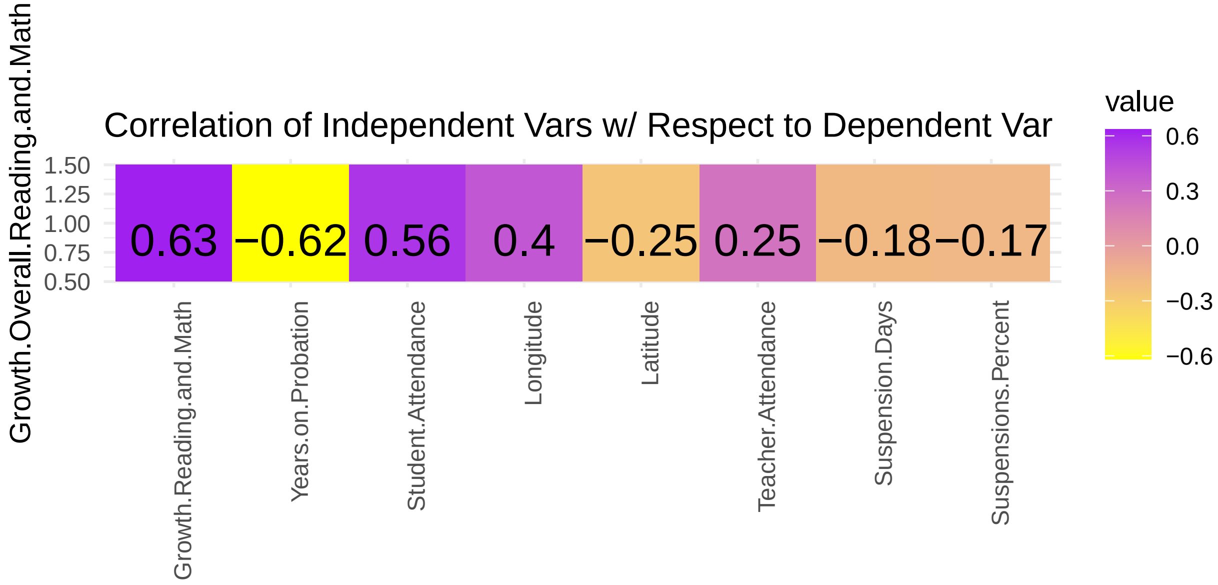
The 2012 School Progress Report Card data from the City of Chicago data portal provides us with a view of the education landscape. It includes the academic performance growth metric ($Growth.Reading.and.Math$), school culture insights ($Suspensions.Percent$,$Suspension.Days$ $Student.Attendance$, $Teacher.Attendance$), and environment variables for which the school was given a grade ($Involved.Families$, $Supportive.Environment$,$Safety$,$Effective.Leaders$, $Ambitious.Instruction$,$Collaborative.Teachers$).

How Key Concepts are Operationalized

EDA

By examining the data, we saw that the likert variables had a normal spread centered around a neutral rating as the most likely ranking for all of the variables. Student and teacher attendance by percentage was also mostly normally distributed. The percentage of misconducts resulting in suspensions was skewed towards the left and the average days of suspension was skewed towards the right. From examining the plots of the continuous variables vs the outcome variable, there were no significant curves in the data that would require transformations.





Assumptions

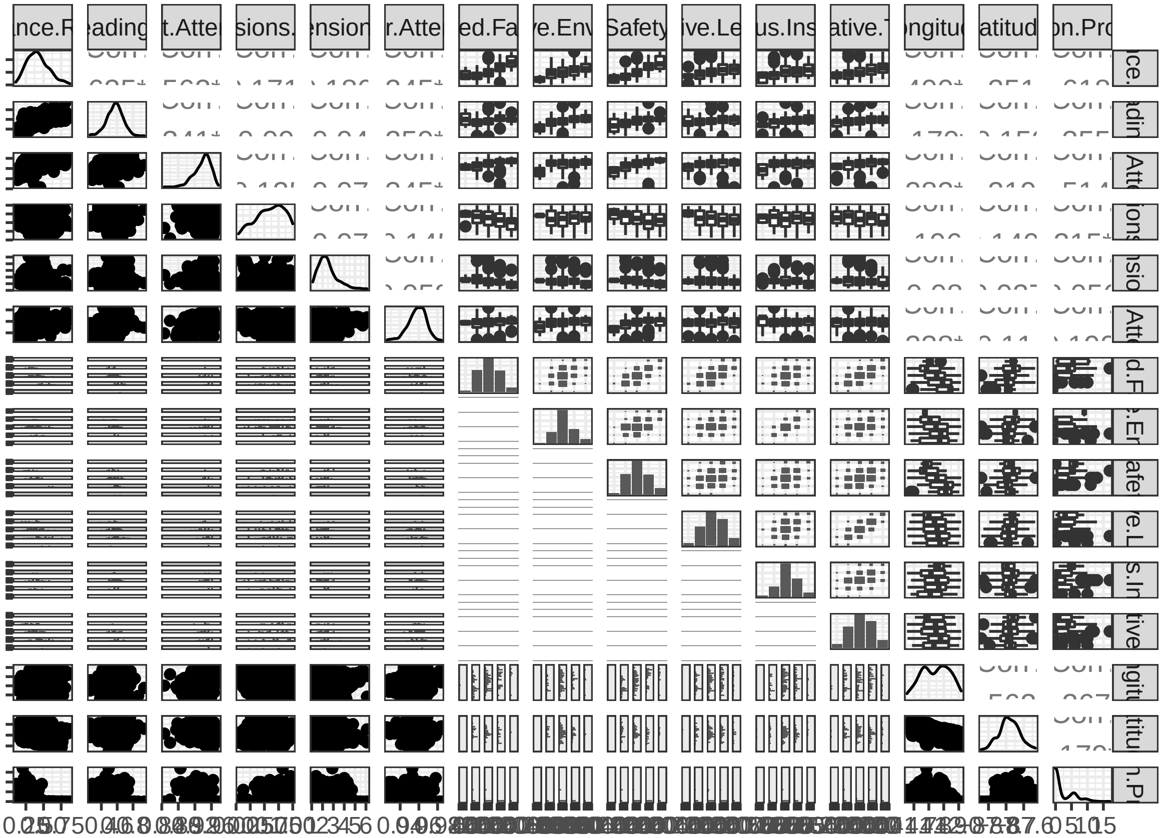
Given our large sample size (218 observations), we can utilize a multiple regression model to assess the key drivers on student success, assuming that our data meets the following criteria: The data must be Independent and Identically Distributed (IID), there must be no perfect collinearity, and appropriate transformations need to be executed (if necessary) to ensure we meet the linear conditional expectation assumption.

IID

Each row in our dataframe represents a unique elementary school in Chicago. Because the data includes the entire population of elementary schools in Chicago rather than a specific sample, we have reduced concerns related to sampling bias since every school is represented in the dataset. Since we are analyzing the entire population of elementary schools in Chicago, we assume that the schools operate under similar conditions and follow similar educational standards and policies. This supports the notion that our observations are drawn from the same distribution, therefore meeting the IID assumption. That said, our scope is limited towards only being able to make generalizations about Chicago public schools.

No Perfect Collinearity

Another key assumption in using large linear models is that the predictor variables must have no perfect collinearity. As shown in the correlation matrix below, the strongest correlation between independent variables is between $Collaborative.Teachers$ and $Effective.Leaders$, with a Pearson's Correlation Coefficient of r = 0.78. Because this relationship is still far away from 1, even the strongest relationship between independent variables does not result in near perfect collinearity.



Additionally, we can run Variance Inflation Factor (VIF) on our complex model to better understand how much the variance of the various beta coefficients increases due to multicollinearity. A VIF of less than 5 is indicative of no perfect collineaity. As we can see from the VIF results above, none of our independent variabels have a VIF value of >= 5. Between the correlation matrix and VIF test, we can conclude our model passes the "no perfect collinearity" assumption.

```
##                                GVIF Df GVIF^(1/(2*Df))
## Growth.Reading.and.Math   1.436470  1        1.198528
## Student.Attendance        1.809489  1        1.345172
## Suspensions.Percent       1.189002  1        1.090413
## Suspension.Days           1.167955  1        1.080720
## Teacher.Attendance        1.323450  1        1.150413
## Involved.Families         8.685160  4        1.310229
## Supportive.Environment   12.302150  4        1.368509
## Safety                    6.388265  4        1.260877
## Effective.Leaders         9.967373  4        1.332977
## Ambitious.Instruction     7.658465  4        1.289786
## Collaborative.Teachers   10.156843  4        1.336118
## Longitude                 1.835610  1        1.354847
## Latitude                  1.612757  1        1.269944
## Years.on.Probation        1.721942  1        1.312228
```

As we can see from the VIF results above, none of our independent variables have a VIF value of >= 5. Between the correlation matrix and VIF test, we can conclude our model passes the "no perfect collinearity" assumption.

Since our data is IID and has no perfect collinearity, we pass all of our large linear model assumptions and can therefore proceed with the use of a multiple regression model.

Key Modeling Decisions

Regression Table

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sat, Aug 05, 2023 - 1:16:31 PM

Table 1: Model Comparison

	Regression Table			
	Naive	Complex	Controllable	Controllable Geo
	(1)	(2)	(3)	(4)
Growth.Reading.and.Math		0.58*** (0.05)		
Student.Attendance	4.14*** (0.36)	1.12*** (0.29)	1.77*** (0.32)	1.69*** (0.32)
Suspensions.Percent		0.03 (0.02)		
Suspension.Days		−0.01* (0.01)		
Teacher.Attendance		−0.47 (0.54)		
Longitude		0.18*** (0.07)		0.15** (0.07)
Latitude		0.03 (0.10)		
Years.on.Probation		−0.01*** (0.002)	−0.01*** (0.002)	−0.01*** (0.002)
Constant	−3.46*** (0.34)	−5.57 (7.62)	−1.23*** (0.30)	−7.55*** (2.78)
Involved Families		✓	✓	✓
Supportive Environment		✓		
Safety		✓	✓	✓
Effective Leaders		✓		
Ambitious Instuction		✓		
Collaborative Teachers		✓		
Observations	281	281	281	281
R^2	0.32	0.79	0.65	0.66
Adjusted R^2	0.31	0.76	0.64	0.64
Residual Std. Error	0.13 (df = 279)	0.07 (df = 248)	0.09 (df = 270)	0.09 (df = 269)
F Statistic	129.04*** (df = 1; 279)	29.16*** (df = 32; 248)	49.97*** (df = 10; 270)	46.61*** (df = 11; 269)
Note:				*p<0.1; **p<0.05; ***p<0.01

Discussion of Results

Discussion of Limitations (Statistical Limitations, Structural Limitations)

Conclusion