An Evaluation of Factors Influencing the Percentage of High School Graduates from Massachusetts Public School Districts Attending Institutions of Higher Education

May Xia

ABSTRACT

The COVID-19 pandemic led to an unprecedented economic crisis all over the world. As the job market became increasingly competitive, earning a post-secondary degree would vastly enhance a candidate's chance of acquiring a job and obtaining higher income in the long term. In this study, I ran a multiple linear regression model on a cleaned dataset with the most updated information about 288 public school districts in Massachusetts in order to investigate the potential factors associated with the percentage of high school graduates pursuing post-secondary education from each school district. My final model was built using backward stepwise procedure under BIC criteria, and it revealed that college attendance rate is associated with high school type (traditional vs. vocational/technical), class attendance rate, the enrollment rates of African American (or Black), Multi-race and Non-Hispanic, male students, and English language learners, Grade 9 course pass rate, and average SAT math score.

1. Background and Significance

There are numerous benefits of attending institutions of higher education, including but not limited to higher income, lower unemployment rate, more career opportunities and expanded professional networks [1]. Nevertheless, not every high school graduate has the privilege of pursuing post-secondary education. Possible factors that either directly or indirectly impact students' decisions about their education plan include the socioeconomic status of a student's family, the overall teaching quality of a high school, the financial resources available to a school district, etc. My main research question is: what are the most influential factors that are associated with the percentages of Massachusetts public high school graduates attending institutions of higher education?

2. Data

2.1 Data Description

For my study, I compiled 12 most recent statewide reports from the "School and District Profiles" section on the official website of the Massachusetts Department of Elementary and Secondary Education [2]. Using the "tidyverse" package in R, I selected and merged data from the 12 reports into one single dataset for subsequent analysis. The dataset contains information about 295 public school districts in Massachusetts. It has 26 predictor variables and one quantitative response variable (percentage of 2018 -19 high school graduates attending college by the March following their high school graduation year; this percentage takes into account all the college types including public and private four-year colleges as well as public and private two-year colleges). A brief summary of all the predictor variables is shown in Table 1. Variable type is either binary categorical (C) or quantitative (Q).

2.2 Data Cleaning

Predictor Variable	Type
School type: traditional (coded as 0) or vocational/technical (coded as 1)	C (binary)
Attendance rate: average class attendance rate in grades PK-12	Q
Art course: percentage of students in grades K-12 taking at least one arts course	Q
Average class size: average class size across all subjects and grades	Q
Expenditure: in-district expenditure per student	Q
Grade 9 course pass rate: percentage of students who completed and passed all courses in Grade 9	Q
Mass Core complete rate: the percentage of high school graduates who completed MassCore curriculum.	Q
Percentage of enrollment by race: African American (or Black), Asian, Hispanic (or Latino), Native American, Native Hawaiian or Other Pacific Islander, Multi-race and Non-Hispanic, White	Q
Percentage of enrollment by gender: female, male, non-binary (individuals who don't identify as just female or male)	Q
Percentage of enrollment by selected population: English language learner*, high needs students*, non-native speaker, students with disability, students from economically disadvantaged backgrounds	Q
Percentage of teachers retained: percentage of teachers who remain working in the same position from one year to the next	Q
Percentage of teachers licensed: percentage of teachers who have obtained teaching license	Q
SAT performance: average SAT verbal score, average SAT math score	Q
Special notes: *English language learners* are defined as students whose first language is not English and who are unable to perform ordinary classroom work in English. *High needs students* are defined as students who are economically disadvantaged, English language learners (either currently or formerly), or are living with disabilities.	

Table 1: Descriptions of Predictor Variables

There were 2 school districts that did not display the response variable (percentage of high school graduates attending college). Based on the information provided on the official website, results were not reported for higher education enrollments of fewer than 15. I removed these 2 school districts from further analysis. In addition, I found missing average SAT verbal and math scores from another 5 school districts. According to the website, the average scores from schools with fewer than 10 test takers were not reported. Since the values were missing not at random, there was no appropriate imputation method that I have learnt so far. Therefore, I decided to remove these 5 school districts

from subsequent analysis as well. The cleaned dataset has 288 observations in total. Furthermore, I created a binary categorical predictor "school_type" that codes 0 for traditional public schools (including charter schools) and 1 for vocational/technical schools. I computed the ratio of standard deviations between the college attendance rates from the two types of schools to check whether the constant variance assumption for the response variable had been satisfied or not. The ratio was 1.08, which fell between 0.5 and 2, and thus the assumption was satisfied. Hence, I could include the binary categorical predictor in the multiple linear regression model.

3. Methods and Results

3.1 Model Fitting

I first examined if multicollinearity existed among my predictor variables by employing the variance inflation factor (VIF) test (threshold was set to 5). The test result showed that 6 variables (enrollment rates of White students, female students, high need students, non-native speakers, economically disadvantaged students, and average SAT verbal score) were highly correlated with other variables, and thus they were removed from the dataset. Next, I used the backward stepwise procedure under AIC and BIC criteria to find two candidate models. The model built under AIC criteria (Model 1) has 9 predictor variables whereas the model built under BIC criteria (Model 2) has 8 predictor variables.

3.2 Model Evaluation

We would prefer a model with higher adjusted R squared value, lower AIC and lower BIC, which reflect better trade-off between model fit and model complexity. However, it is worth noting that BIC tends to impose a larger penalty on the number of parameters in the model than the other two measures do. Based on Table 2 below, we could see that Model 1 and Model 2 have the same adjusted R squared value, but Model 1 has lower AIC whereas Model 2 has lower BIC. I performed 5-fold cross validation (CV) and calculated average prediction error for each model. Model 2 has lower average prediction error and it is more parsimonious than Model 1, and thus I chose Model 2 to be my final model.

	Adjusted R^2	AIC	BIC	Cross-Validated Average Prediction Error
Model 1	0.75	1953	1994	53.88
Model 2	0.75	1956	1992	53.47

Table 2: Performance Metrics for Model 1 and Model 2

3.3 Model Diagnostics and Final Model Interpretation

In order to ensure that the final model has fulfilled all the linear regression assumptions, I examined the studentized deleted residual plot and the normal probability plot (Figure 1). Firstly, there is no special pattern present among the residuals, and thus the linearity assumption is satisfied. Moreover, the residuals seem to have similar variations across different levels of the fitted values, and thus the constant variance assumption for the random error terms is also satisfied. Furthermore, based on the normal probability plot,

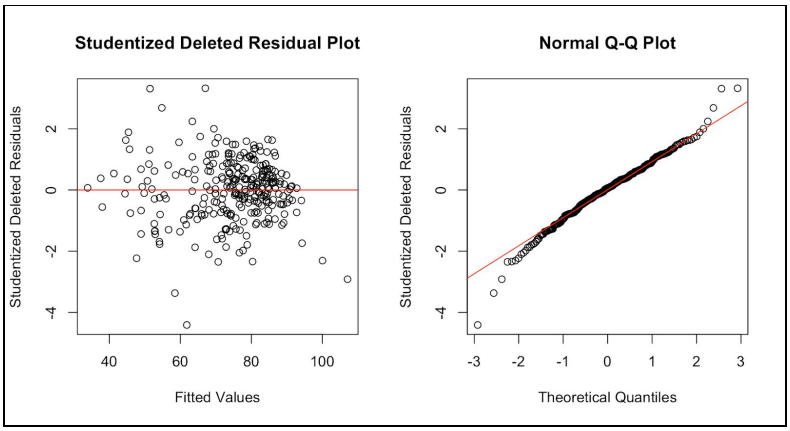


Figure 1: Model Diagnostics Plots

most of the points are close to the straight line, and thus the normality of errors assumption is satisfied. Nevertheless, I could observe a few potential outliers on the residual plot. In order to determine whether the potential outliers are influential, I compared Cook's distance of the predicted responses to the F distribution (df1=8, df2=280), and I found that the Cook's distance of 1 observation (row 189) fell above the 50th percentile of the F distribution. Hence, I removed row 189 from the dataset. Afterwards, I refit the model and found that the adjusted R squared value became 0.76, and both AIC and BIC measures decreased to 1939 and 1975 respectively. My new final model is shown below:

$$\widehat{percent\ attend\ college} = -78.4 - 21.1 \times school\ type + 1.26 \times class\ attendance\ rate$$
$$+ 0.315 \times African\ American - 1.08 \times multi\ race - 0.660 \times males$$
$$+ 0.337 \times Grade\ 9\ pass\ rate - 0.228 \times English\ learner + 0.0804 \times SAT\ math$$

My final model suggests that the predicted percentage of high school graduates from Massachusetts public school districts attending institutions of higher education is associated with the type of high school (traditional vs. vocational/technical), class attendance rate, the enrollment rates of African American (or Black), Multi-race and Non-Hispanic, male students, and English language learners, Grade 9 course pass rate, and average SAT math score. The most influential variable is the school type, which is indicated by the greatest negative slope (-21.1). The predicted college attendance rate for vocational/technical school graduates is significantly lower than the one for traditional high school graduates if we hold other variables constant.

4. Conclusions

In addition to the school type, my study showed that class attendance rate, enrollment rates of underrepresented groups, Grade 9 pass rate and average SAT math score would also influence the predicted college attendance rate. The school district administrators might consider reviewing these factors and designing certain measures to improve college attendance rate. For example, the school districts could implement more education support programs for the English language learners to improve their language skills. Furthermore, the data of college attendance rates only considered high school graduates attending college by the March following their high school graduation year. However, students might consider attending college a few years later after earning enough tuition. It would be interesting to investigate the factors that influence the percentage of high school graduates pursuing post-secondary degrees in their lifetime.

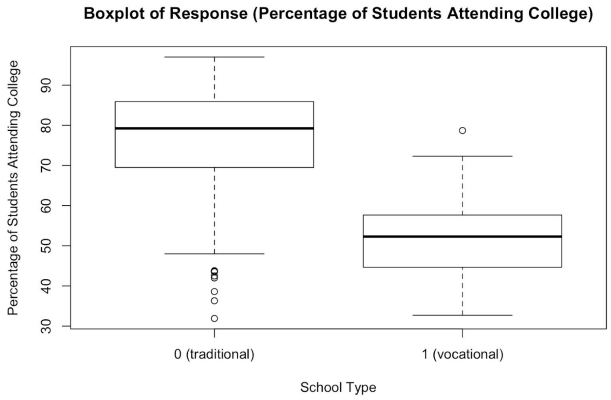
5. References

[1] “9 Benefits of Going to College.” Indeed, 10 Dec. 2020, https://www.indeed.com/career-advice/career-development/college-benefits.

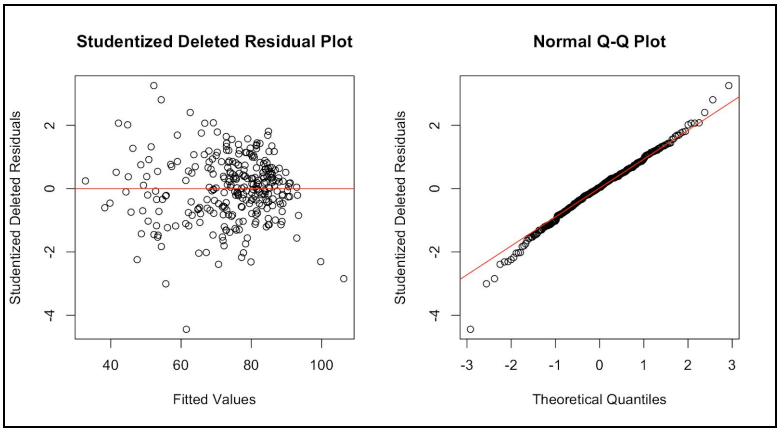
[2] “Statewide Reports.” Massachusetts Department of Elementary and Secondary Education, https://profiles.doe.mass.edu/state_report/.

6. Appendix

A. Data Exploration Graph



B. New Model Diagnostic Plots after Removing the Influential Outlier



C. Final Model Summary Output

```
Call:
lm(formula = attend_college ~ school_type + attendance_rate +
    African.American + multi_race + Males + grade9_passrate +
    englishlearner + sat_math, data = newma_data)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5981  -4.0782   0.2678   4.3518  21.1533

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -78.36078   14.91180  -5.255 2.95e-07 ***
school_type      -21.11712    1.69244 -12.477  < 2e-16 ***
attendance_rate    1.25957    0.15372   8.194 9.32e-15 ***
African.American   0.31532    0.03722   8.471 1.43e-15 ***
multi_race        -1.08151    0.22141  -4.885 1.75e-06 ***
Males             -0.65953    0.11191  -5.894 1.09e-08 ***
grade9_passrate    0.33688    0.04198   8.026 2.85e-14 ***
englishlearner    -0.22795    0.07511  -3.035  0.00263 **
sat_math           0.08035    0.01354   5.932 8.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.955 on 278 degrees of freedom
Multiple R-squared:  0.7634,    Adjusted R-squared:  0.7566
F-statistic: 112.1 on 8 and 278 DF,  p-value: < 2.2e-16
```