A Multiple Variable Regression Model to predict public high schools graduation rates in NYC (2014) based on NYC Open Data

Jay KIM and Fabrice MESIDOR • 08.23.2019

Overview

Objective:

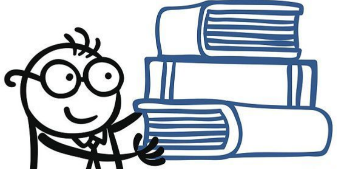
Identify factors that can predict NYC public high school graduation rates.

Questions:

- Do demographics of a school impact student graduation rates? If so, which attributes (gender, ethnicity, new English Learner, etc.) ?
- Similarly, can factors pertaining to school quality or individual student achievements help predict graduation rates?

Hypothesis:

- The high rate of graduation in a school is not dependent of any factors related to the school or the students.



Literature Review

Definition

- **Graduation rate** is the percentage of a school's first-time, first-year undergraduate students who complete their program within 150% of the published time for the program. For example, for a four-year degree program, entering students who complete within six years are counted as **graduates**.
- School graduation is an important outcome measure for the school systems. Hence, determining its different factors can help improving the school systems.

Existing works

- Rumberger and Larson, 1998; Ensminger and Slusarcick, 1992).
- Ensminger and Slusarcick predicted using background and demographic variables (low grades, aggressive behavior, student poverty, and parents' education level.)
- Rumberger and Larson predicted using low grades, misbehavior, and high absenteeism.

Dataset and Variables

Sources

- Data extracted from NYC Open Data (https://opendata.cityofnewyork.us/) in csv format, cleaned and treated using python (pandas)
- Disclaimer: NYC Open Data format has inconsistencies (i.e., results are only as good as its source data)

Data

- 420 observations - each observation represents a school. (3 primary tables)

Target Variable

- High school graduation rate (% - continuous)

Independent variables (40 after filtering)

- Classified between ethnicity - language proficiency - gender - SAT - achievement scores

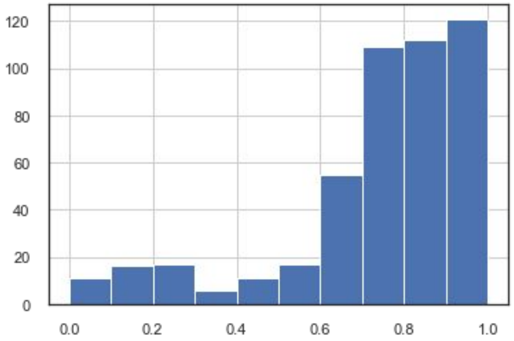
Empirical Approach

- Transformation of target variables to consider success (graduation rate ≥ 50%)

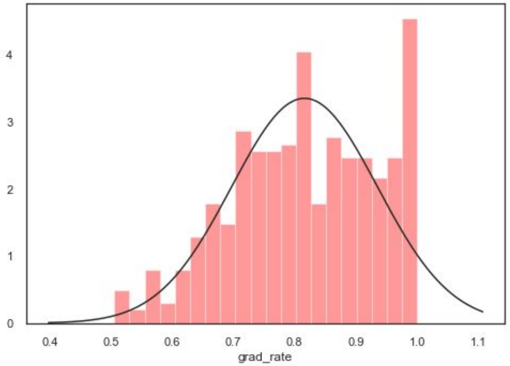
- Distribution approximates a normal distribution after change

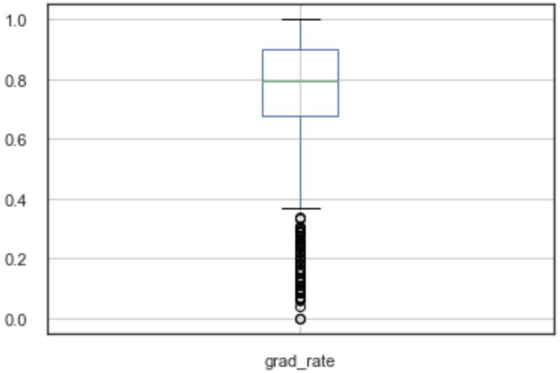
- It also helped to remove data that would create biases in our models

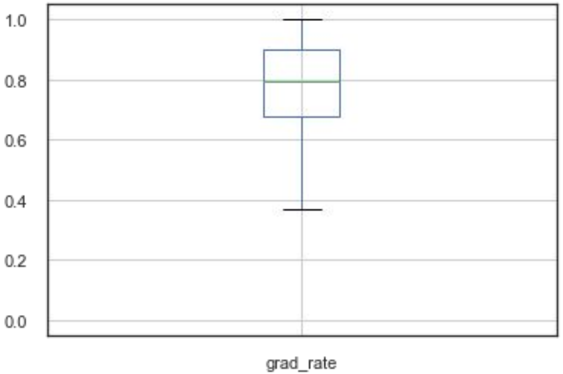
Distribution of school by graduation rates



Distribution of schools with graduation rate more than 50%

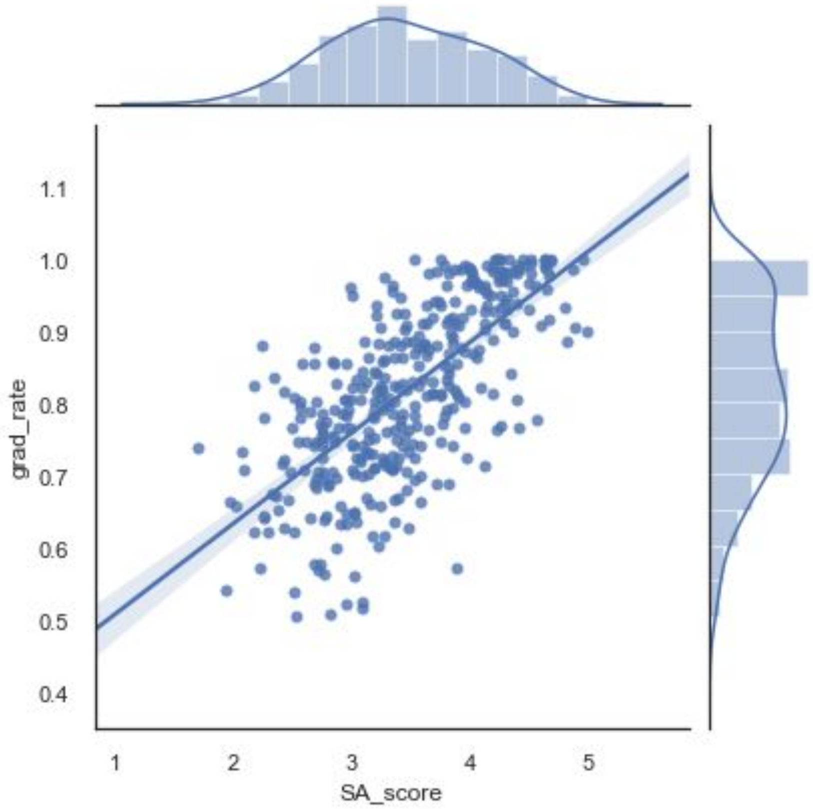


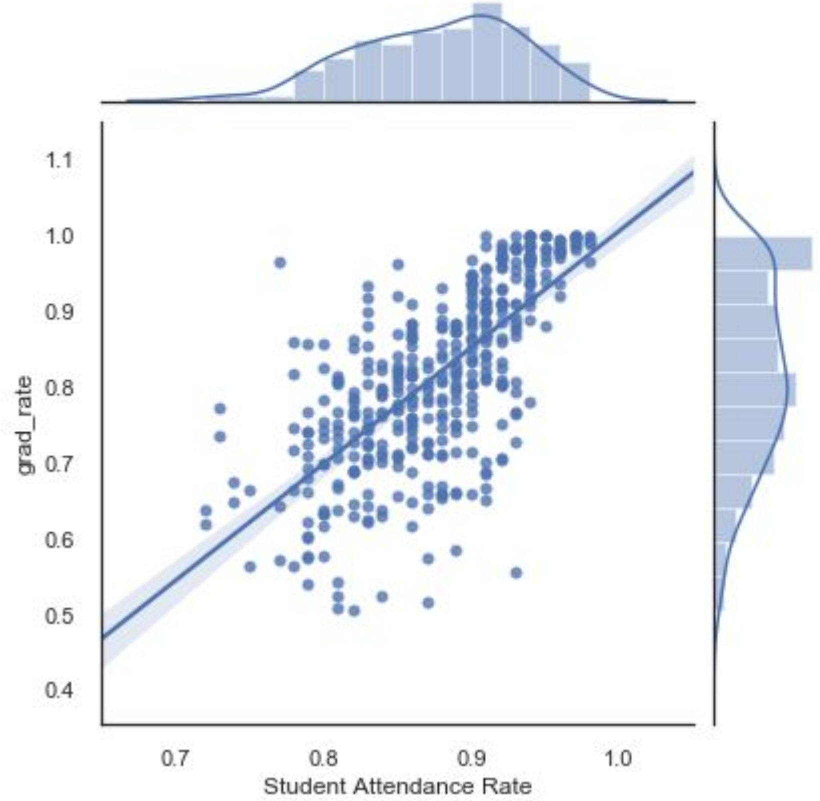




Empirical Approach (suite)

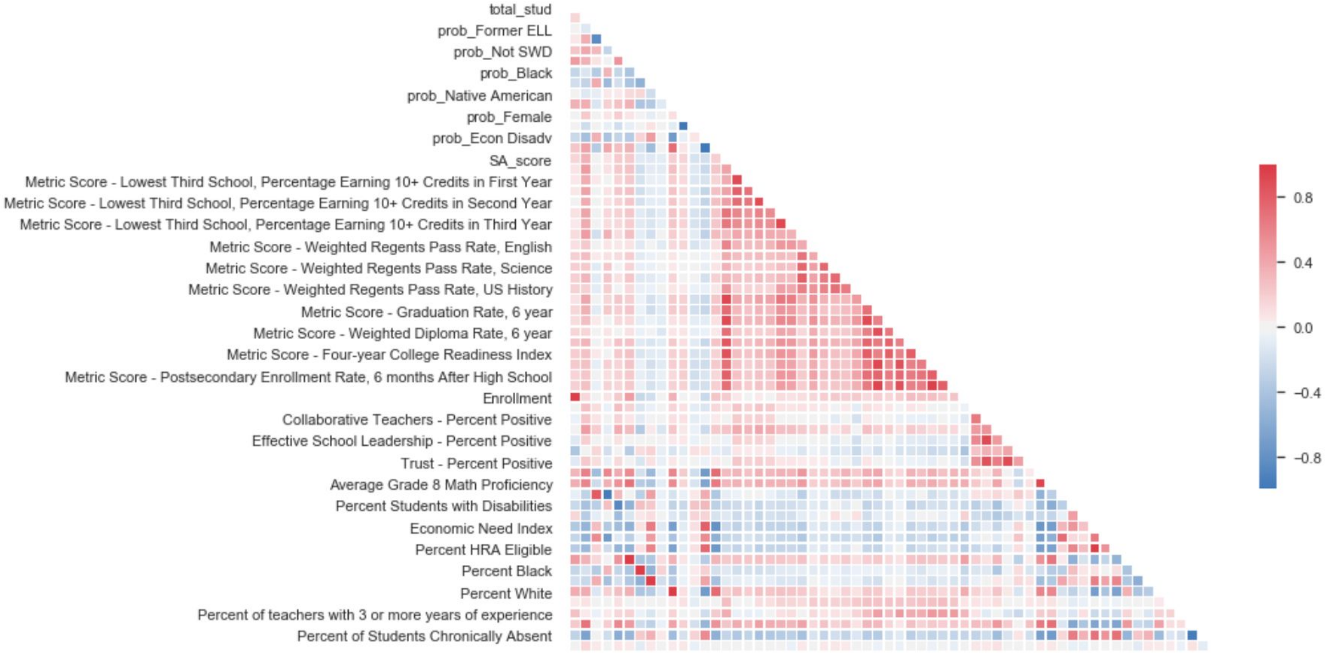
Relation between 2 main variables: Student Achievement Scores and Student Attendance Rate





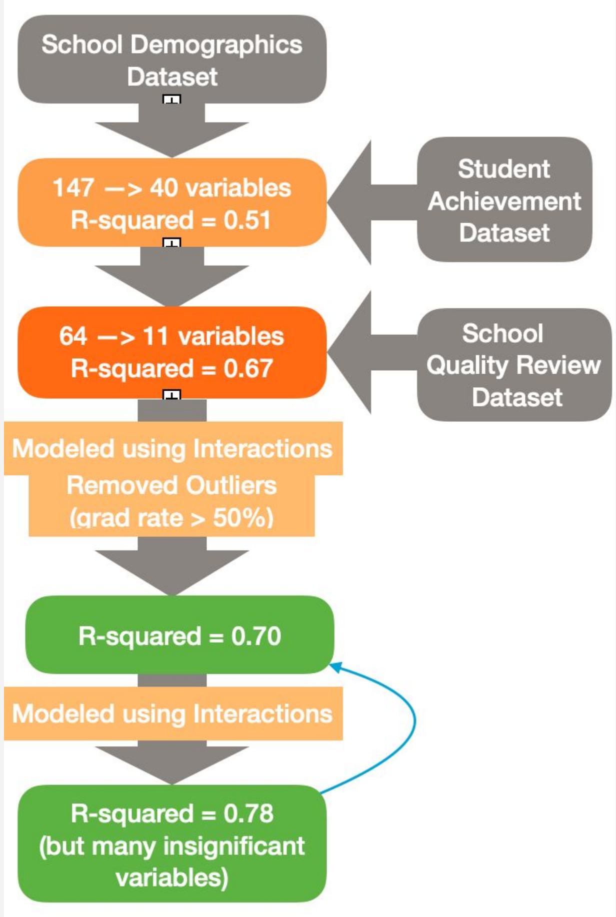
Empirical Approach (suite)

Correlation Matrix with all variables



- Since this is a Pearson Coefficient, the values near to 1 or -1 have high correlation.
- We drop “ethnicity probabilities” and all but one “metric score,” and then
- We start to execute our linear regression model with the non-correlated variables.

Modelisation & Validation



School Demographics Dataset

147 —> 40 variables
R-squared = 0.51

Student Achievement Dataset

64 —> 11 variables
R-squared = 0.67

School Quality Review Dataset

Modeled using Interactions
Removed Outliers
(grad rate > 50%)

R-squared = 0.70

Modeled using Interactions

R-squared = 0.78
(but many insignificant variables)

The different steps before selecting our final model:

- Adding variables
- Modeling with transformation
- Checking errors normality and heteroscedasticity

Final Model

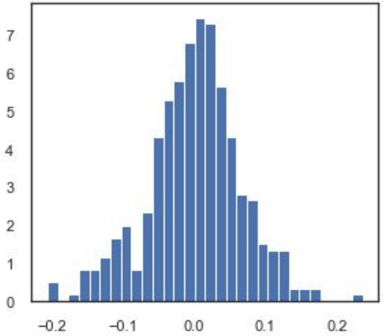
Final Model

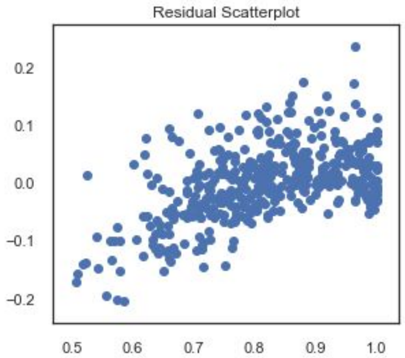
Grad_rate =

-0.8649 +
0.2975 (prob Former English Language Learner) +
0.5254 (prob_Not English Language Learner) +
0.0086 (Student Achievement Score) +
0.018 (% Earning 10 Credits in Year 1) +
0.2235 (Supportive Environment- % Positive) +
0.0469 (Avg Grade 8 English Proficiency) +
0.4109 (% English Language Learners) +
-0.3683 (% in Temp Housing) +
0.6355 (Student Attendance Rate) +
0.2678 (Teacher Attendance Rate)

Linear Regression Assumptions

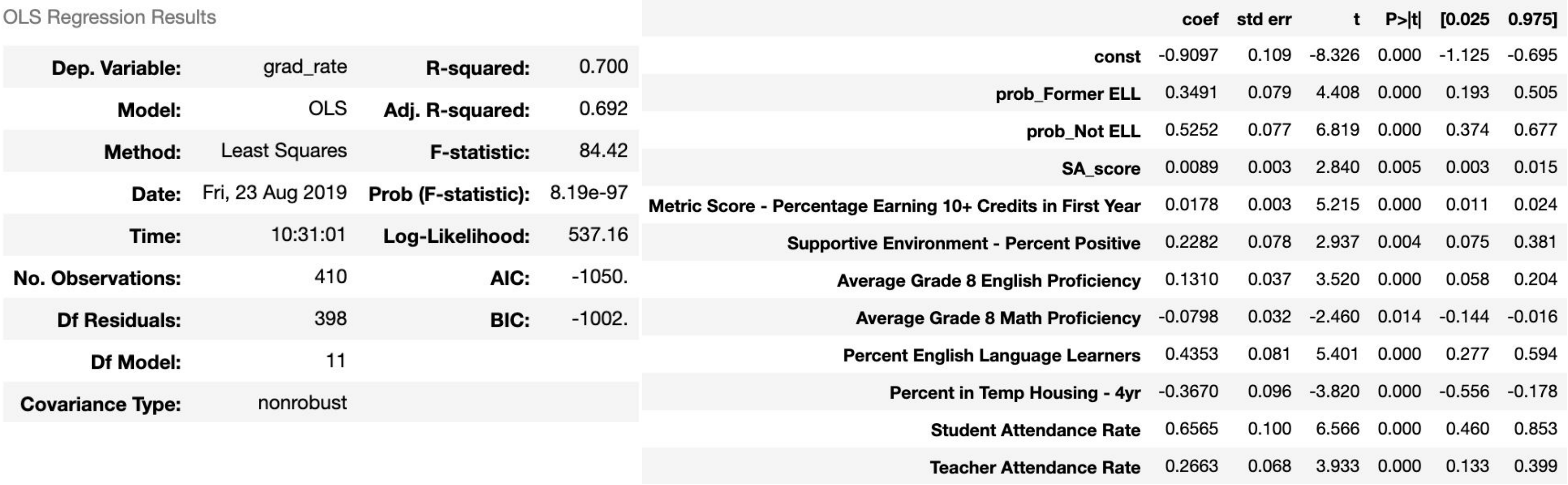
- Our model explains 70% of the variances of the variables
- Error histogram → approximates a normal distribution
- The residual scatter plot doesn't show a real pattern for the fitted value - we can assume the error is homoscedastic.





Results and Interpretation

Model Summary Table



OLS Regression Results

Dep. Variable:	grad_rate	**R-squared:**	0.700
Model:	OLS	**Adj. R-squared:**	0.692
Method:	Least Squares	**F-statistic:**	84.42
Date:	Fri, 23 Aug 2019	**Prob (F-statistic):**	8.19e-97
Time:	10:31:01	**Log-Likelihood:**	537.16
No. Observations:	410	**AIC:**	-1050.
Df Residuals:	398	**BIC:**	-1002.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-0.9097	0.109	-8.326	0.000	-1.125	-0.695
prob_Former ELL	0.3491	0.079	4.408	0.000	0.193	0.505
prob_Not ELL	0.5252	0.077	6.819	0.000	0.374	0.677
SA_score	0.0089	0.003	2.840	0.005	0.003	0.015
Metric Score - Percentage Earning 10+ Credits in First Year	0.0178	0.003	5.215	0.000	0.011	0.024
Supportive Environment - Percent Positive	0.2282	0.078	2.937	0.004	0.075	0.381
Average Grade 8 English Proficiency	0.1310	0.037	3.520	0.000	0.058	0.204
Average Grade 8 Math Proficiency	-0.0798	0.032	-2.460	0.014	-0.144	-0.016
Percent English Language Learners	0.4353	0.081	5.401	0.000	0.277	0.594
Percent in Temp Housing - 4yr	-0.3670	0.096	-3.820	0.000	-0.556	-0.178
Student Attendance Rate	0.6565	0.100	6.566	0.000	0.460	0.853
Teacher Attendance Rate	0.2663	0.068	3.933	0.000	0.133	0.399

At 95% confidence level, all our coefficients expect ‘Average Grade 8 Math Proficiency’ are statistically significant

1. NYC Public High School graduation rates can be explained mainly by factors relevant to the school environment and attendance

2. Gender and ethnicity don’t impact graduation rates in any of our models

3. More analysis are required to find out why male/female population and other demographic attributes are not statistically significant