Determinants of Graduation Rates in Massachusetts

Gabriel Kagan,* Andrew Sharp,* William Delaney,* Abdoul Rachid Ayouba Mahamane,* and Zoe Liu*

Department of Mathematics and Statistics, University of Massachusetts, Amherst

E-mail: gkagan@umass.edu; awsharp@umass.edu; wdelaney@umass.edu; aayouba@umass.edu; ziyingliu@umass.edu

Abstract

The study our team conducted utilized regression models to analyze the relationship between various predictor variables and graduation rate. We incorporated both descriptive and regression statistics into the analysis. The results showed that the variables our team studied have a linear relationship with graduation rate, and the variables in our model are not significantly correlated with the others.

Introduction

Our group studied how various factors can predict the graduation rates of Massachusetts public high schools. In particular, we examined the average ($) amount spent per student, as well as the percentage of students whose first language is not English, who are disabled, who are economically disadvantaged, and who are high needs. Based on empirical research, we determined that these variables would be the most appropriate predictors of the graduation rate for each school. Our team expected for graduation rate to increase as expenditure increased;[1]

the more capital a school spends allows for better quality teachers due to higher salaries, and better learning equipment and facilities for students.[2] These symptoms of greater expenditure create better learning environment for students. Conversely, we expected graduation rates to decrease as the remaining variables increased. As these variables are measurements of disadvantages, such as language barrier, poverty, and disability; therefore a greater value for a measurement of disadvantage should result in difficulty for students to graduate.[3] The design of an accurate regression model may help governments determine how to allocate funding to schools in order to ensure all schools are able to maximize graduation rates.

Data set

The data used in the study was collected by the Massachusetts Department of Education.[4] The data includes: average expenditure per student, percentage of students whose first language is not English, percentage of disabled students, high needs students, economically disadvantaged students, and graduation rates for each school in Massachusetts. We analyzed 340 high schools from this data, since elementary and middle schools did not have graduation rates. This data set was available for download on Kaggle.[5]

Variable Definitions*

Average expenditure per student: Per student expenditures are calculated by dividing total expenditures by total average membership (full time equivalent students).

Disabled students: Percentage of students who participate in an Individualized Education Program (IEP).

Students whose first language is not English: Percentage of students whose first language is a language other than English. Students within this category are defined as those who have indicated a language other than English on the Home Language Survey. Note that this category does not reflect the level of proficiency in the English language.

High needs students: Percentage of students who have high needs students. Student are placed into this category if they are designated as either low income, economically disadvantaged, or ELL, or a student with disabilities.

Economically disadvantaged students: Percentage students who participate in one or more of the following state-administered programs: the Supplemental Nutrition Assistance Program (SNAP); the Transitional Assistance for Families with Dependent Children (TAFDC); the Department of Children and Families' (DCF) foster care program; and MassHealth (Medicaid).

Graduation rate: Percentage of students who graduate with a regular high school diploma within 4 years.

**Criteria for these variables was determined by the Massachusetts Department of Education and was provided in the public data set.*[5]

Methods and results

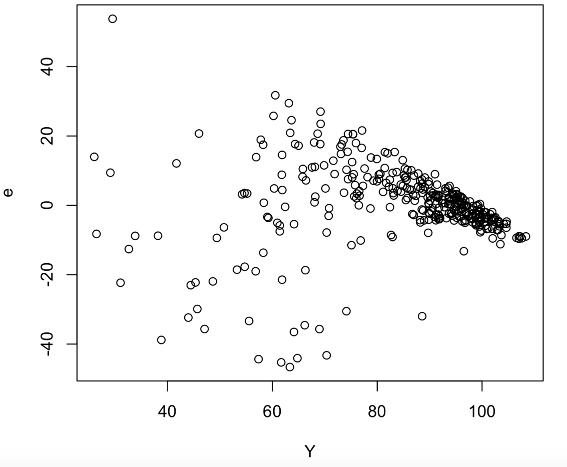


Figure [1] Fitted Y values vs. residuals for our first model

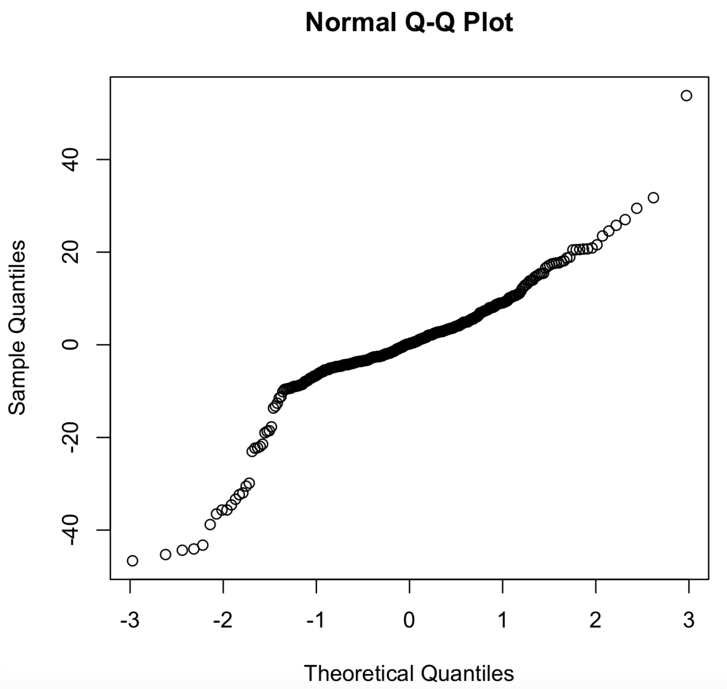


Figure [2] QQ plot of residuals for our first model

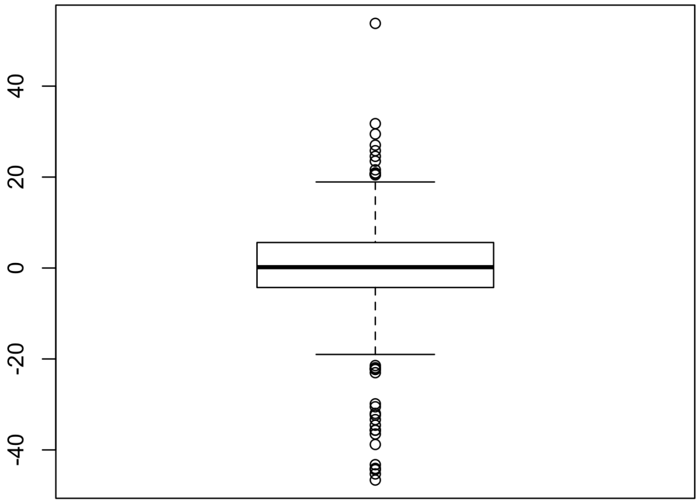


Figure [3] Box plot of residuals for our first model

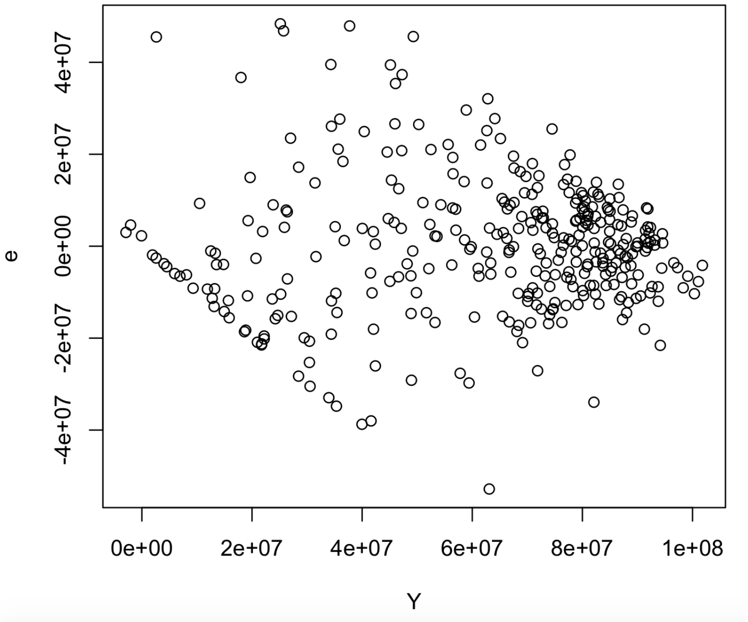


Figure [4] Fitted Y values vs. residuals for our second model

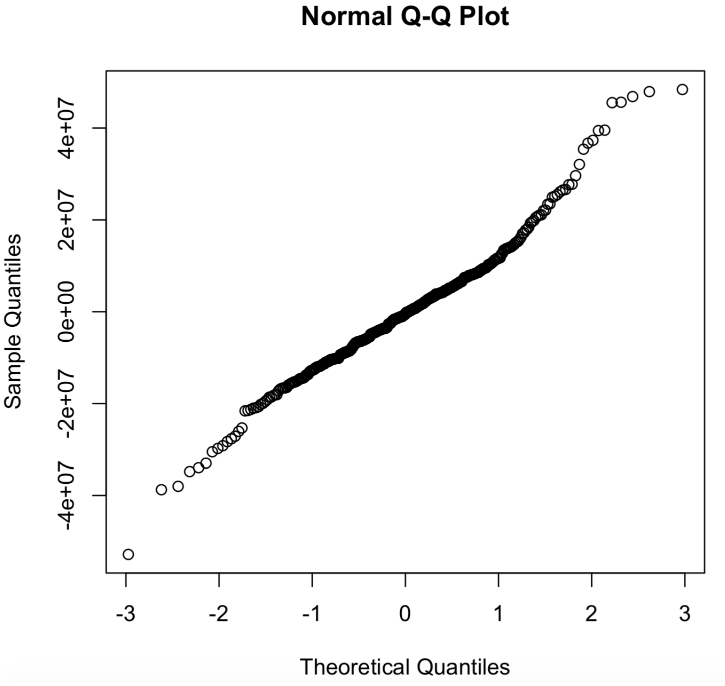


Figure [5] QQ plot of residuals for our second model

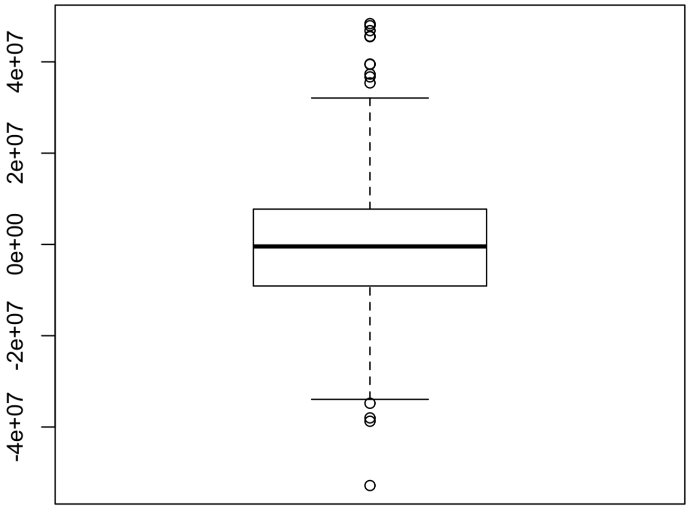


Figure [6] Box plot of residuals for our second model

We first attempted to use the following linear regression model:

$$
\begin{aligned}
&\textbf{Graduation rate}_i = \\
&\beta_0 + \beta_1(\text{Percent of students whose first language is not English})_i + \\
&\beta_2(\text{Percent of disabled students})_i + \beta_3(\text{Percent of high needs students})_i + \\
&\beta_4(\text{Percent of economically disadvantaged students}) + \beta_5(\text{Expenditure per student})_i + \epsilon_i
\end{aligned}
$$

However, when checking the residual plots for this model, we found the model failing some of the necessary assumptions. The plot of the fitted values versus the residuals (Figure [1]) shows that although the relationship appears to be linear, the error terms do not have constant variance. The variance is smaller for small and large values of Y while values of Y in the middle of this range typically have larger residuals. The QQ plot of the residuals for this model (Figure [2]) also shows a potential problem, as the plot is not a straight line. This indicates that the data may not follow a normal distribution. Finally, the box plot of the residuals (Figure [3]) shows that there are a number of outliers on both ends. In order to fix some of these problems, we decided to transform our data by replacing Y with Y^4. This resulted in the following model instead:

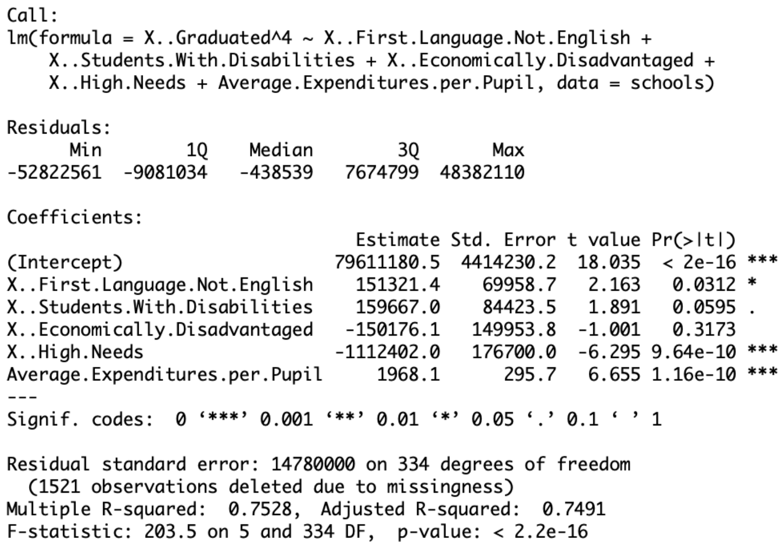
$$\textbf{Graduation rate}_i^4 =$$

$$\beta_0 + \beta_1(\text{Percent of students whose first language is not English})_i +$$

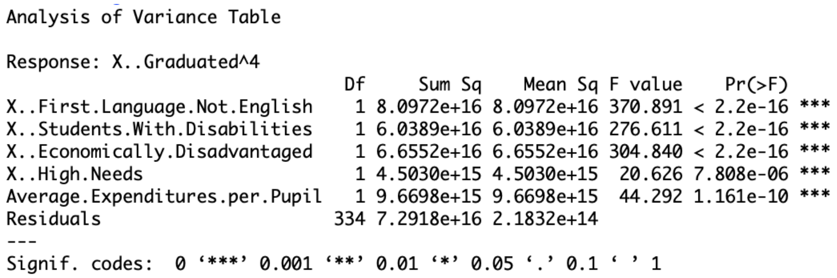
$$\beta_2(\text{Percent of disabled students})_i + \beta_3(\text{Percent of high needs students})_i +$$

$$\beta_4(\text{Percent of economically disadvantaged students}) + \beta_5(\text{Expenditure per student})_i + \epsilon_i$$

The residual plots for this new model show that some of the assumptions are more satisfied. The QQ plot (Figure [5]) now appears to indicate that the residuals follow a normal distribution. The box plot of the residuals (Figure [6]) also shows that there are fewer outliers in this model, although some still exist. However, the variance of the errors still does not appear to be completely constant for all values of Y^4 as seen in Figure [4]. This means that there are still potential problems with our model, and in particular that it cannot accurately estimate the variance σ^2. The estimates for the coefficients in the model may also be less accurate because of this. However, we still chose to proceed with this model. The results of this linear regression from R can be seen in the following tables.



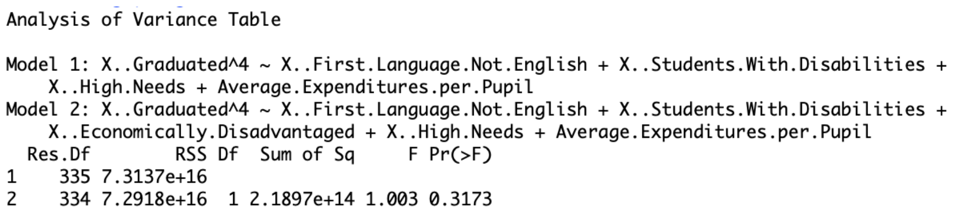
```
Call:
lm(formula = X..Graduated^4 ~ X..First.Language.Not.English +
    X..Students.With.Disabilities + X..Economically.Disadvantaged +
    X..High.Needs + Average.Expenditures.per.Pupil, data = schools)

Residuals:
      Min        1Q    Median        3Q       Max
-52822561  -9081034   -438539   7674799  48382110

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    79611180.5  4414230.2  18.035  < 2e-16 ***
X..First.Language.Not.English    151321.4    69958.7   2.163   0.0312 *
X..Students.With.Disabilities    159667.0    84423.5   1.891   0.0595 .
X..Economically.Disadvantaged   -150176.1   149953.8  -1.001   0.3173
X..High.Needs                  -1112402.0   176700.0  -6.295 9.64e-10 ***
Average.Expenditures.per.Pupil     1968.1      295.7   6.655 1.16e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14780000 on 334 degrees of freedom
  (1521 observations deleted due to missingness)
Multiple R-squared:  0.7528,  Adjusted R-squared:  0.7491
F-statistic: 203.5 on 5 and 334 DF,  p-value: < 2.2e-16
```



```
Analysis of Variance Table

Response: X..Graduated^4
                                 Df     Sum Sq    Mean Sq F value    Pr(>F)    
X..First.Language.Not.English     1 8.0972e+16 8.0972e+16 370.891 < 2.2e-16 ***
X..Students.With.Disabilities     1 6.0389e+16 6.0389e+16 276.611 < 2.2e-16 ***
X..Economically.Disadvantaged     1 6.6552e+16 6.6552e+16 304.840 < 2.2e-16 ***
X..High.Needs                     1 4.5030e+15 4.5030e+15  20.626 7.808e-06 ***
Average.Expenditures.per.Pupil    1 9.6698e+15 9.6698e+15  44.292 1.161e-10 ***
Residuals                       334 7.2918e+16 2.1832e+14                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We performed a hypothesis test on this data with the null hypothesis (H_0) that $\beta_1 = \ldots = \beta_5 = 0$ and the alternative hypothesis (H_a) that at least one of the β_i does not equal 0. The p-value for the resulting F-test was very small (< 2.2 x 10^{-16}), so we rejected the null hypothesis and concluded that there was a linear relationship between graduation rate and at least one of our independent variables. The R^2 value for our model was approximately 0.75, so the variables in our model account for most of the variance in graduation rate. However, this value is not incredibly close to 1, so there are most likely other variables influencing this as well.

Because the p-value for β_3 was not significant, we chose to test a reduced model with the third variable, the percent of economically disadvantaged students, removed. We then performed another hypothesis test comparing the reduced model to the full model with the null hypothesis that $\beta_3 = 0$ and the alternative hypothesis that β_3 is not 0. The resulting ANOVA table is shown below.



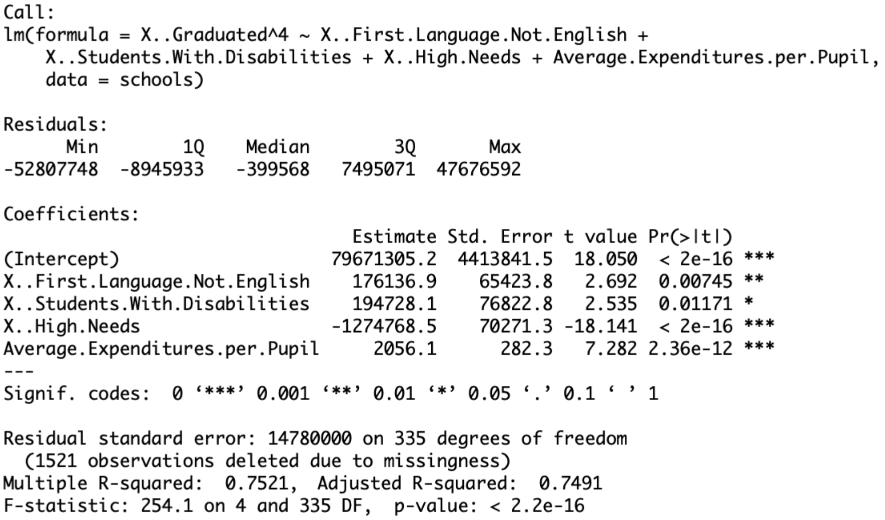
```
Analysis of Variance Table

Model 1: X..Graduated^4 ~ X..First.Language.Not.English + X..Students.With.Disabilities + 
    X..High.Needs + Average.Expenditures.per.Pupil
Model 2: X..Graduated^4 ~ X..First.Language.Not.English + X..Students.With.Disabilities + 
    X..Economically.Disadvantaged + X..High.Needs + Average.Expenditures.per.Pupil
  Res.Df        RSS Df  Sum of Sq     F Pr(>F)
1    335 7.3137e+16                           
2    334 7.2918e+16  1 2.1897e+14 1.003 0.3173
```

The p-value for the resulting F-value is high (0.32), so we fail to reject the null hypothesis. This indicates that it is possible that $\beta_3 = 0$, so we removed this variable from the model.

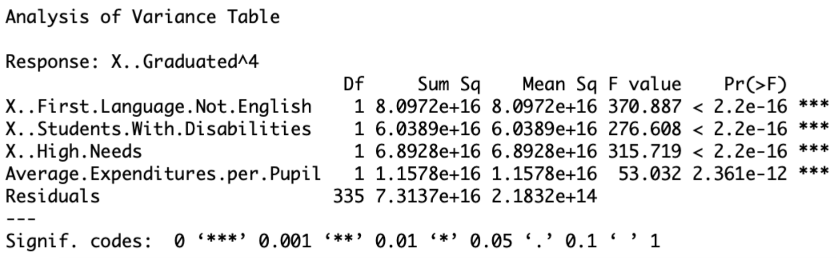
Our final model is the following:

$$\textbf{Graduation rate}_i^4 = \beta_0 + \beta_1(\text{Percent of students whose first language is not English})_i + \beta_2(\text{Percent of disabled students})_i + \beta_3(\text{Percent of high needs students})_i + \beta_4(\text{Expenditure per student})_i + \epsilon_i$$

The R outputs for this new model are shown below. The p-value for this model is still less than 2.2 x 10^{-16}, which indicates that these variables have a linear relationship with the graduation rate. The R^2 value is also still approximately 0.75, so these four variables account for most of the variance in Y. As shown in the ANOVA table, each of the independent variables explains a significant amount of the variation when added to this model.

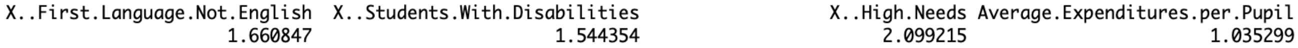


```
Call:
lm(formula = X..Graduated^4 ~ X..First.Language.Not.English +
    X..Students.With.Disabilities + X..High.Needs + Average.Expenditures.per.Pupil,
    data = schools)

Residuals:
      Min        1Q    Median        3Q       Max
-52807748  -8945933   -399568   7495071  47676592

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   79671305.2  4413841.5  18.050  < 2e-16 ***
X..First.Language.Not.English   176136.9    65423.8   2.692  0.00745 **
X..Students.With.Disabilities   194728.1    76822.8   2.535  0.01171 *
X..High.Needs                 -1274768.5    70271.3 -18.141  < 2e-16 ***
Average.Expenditures.per.Pupil    2056.1      282.3   7.282 2.36e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14780000 on 335 degrees of freedom
  (1521 observations deleted due to missingness)
Multiple R-squared:  0.7521,  Adjusted R-squared:  0.7491
F-statistic: 254.1 on 4 and 335 DF,  p-value: < 2.2e-16
```



```
Analysis of Variance Table

Response: X..Graduated^4
                                Df     Sum Sq    Mean Sq F value    Pr(>F)
X..First.Language.Not.English    1 8.0972e+16 8.0972e+16 370.887 < 2.2e-16 ***
X..Students.With.Disabilities    1 6.0389e+16 6.0389e+16 276.608 < 2.2e-16 ***
X..High.Needs                    1 6.8928e+16 6.8928e+16 315.719 < 2.2e-16 ***
Average.Expenditures.per.Pupil   1 1.1578e+16 1.1578e+16  53.032 2.361e-12 ***
Residuals                      335 7.3137e+16 2.1832e+14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Finally, we performed a multicollinearity test on the four variables in this model to check the correlation between them. The square roots of the VIF values for each of the variables is shown below. These values are all close to 1, which indicates that they are not significantly correlated with the other variables.



```
X..First.Language.Not.English  X..Students.With.Disabilities                  X..High.Needs Average.Expenditures.per.Pupil
                     1.660847                       1.544354                       2.099215                     1.035299
```

Conclusion

We can conclude that there is a linear relationship between the graduation rate at a Massachusetts high school (to the fourth power) and the expenditure per student, percent of students whose first languages are not English, percent of students with disabilities, and percent of high needs students at a high school. As we expected, as the amount of money spent per student increases, the graduation rate also tends to increase and as the percent of high needs students at a school increases, the graduation rate tends to decrease. However, graduation rate also tends to increase as the percent of students whose first language is not English or the percent of disabled students increase. We expected some students in these groups to struggle in school and to cause the graduation rate to decrease as the sizes of these groups increased. However, it is possible that students in these groups are more motivated to work hard or receive more assistance in school and are therefore more likely to graduate. Our model indicates that it is important for governments to ensure that high schools are well funded because spending more money on each student tends to significantly increase the percentage of students who are able to graduate.

Acknowledgement

We thank our honorable Professor Zheni Utic for her critical role in the inspiration of the research paper. With her guidance we were able to pursue an in depth study of graduation rates via regression analysis. We would also like to thank Professor Utic for her comedic abilities during lecture, which guided our research during the most difficult times.

References

(1) Ruggiero, J. *Education Economics* **2007**, *15*, 1–13.

(2) Sander, W. *Journal of Public Economics* **1993**, *52*, 403–416.

(3) Schifter, L. *Exceptional Children* **2011**, *77*, 409–422.

(4) Massachusetts Department of Education. `http://www.doe.mass.edu`, Accessed: 2018-11-07.

(5) Massachusetts Public Schools Data. `https://www.kaggle.com/ndalziel/massachusetts-public-schools-data/home`, Accessed: 2018-11-07.