Pittsburgh Public Schools Retention/Mobility Research

Huiyi Guo
Master of Statistical Practice
Carnegie Mellon University

Jenny Luo
Master of Statistical Practice
Carnegie Mellon University

Yuhang Ying
Master of Statistical Practice
Carnegie Mellon University

Zach Branson
Project Advisor
Department of Statistics and Data Science
Carnegie Mellon University

Abstract

Pittsburgh Public Schools funds a Promise scholarship for post-secondary education of qualified students in their district. It hopes to evaluate factors that influence whether students received Promise scholarship and factors related to students' retention in college study. We conduct different regression models for both questions, considering factors including students' demographic information, academic performance in high school, Promise scholarships information, and enrollment records of post-secondary education. We find that students with high GPA, high attendance rate, and relatively low Keystone average score would be more likely to attend colleges in PA and receive the scholarship. In addition, there are discrepancies between race and gender in the status of qualifying and receiving Promise. Also, the effectiveness of Promise scholarship varies by race and becomes more apparent for more senior students, and other factors such as attendance rate, gender, and college enrollment semester also have significant effects on retention. Based on the findings, we suggest that the current criteria for Promise scholarship are appropriate, but the criteria shall be tailored for students with different racial identities. While the study has three major limitations, the next steps include adding more predictors or interaction terms in the model, performing two-stage least-squares analysis, diving deeper into surprising findings from this study.

1 Introduction

Pittsburgh Public Schools is a public school district for pre-K 12 students in Pittsburgh, Pennsylvania, United States. This organization funds a Promise scholarship for post-secondary education of students who are qualified for the scholarship requirements (ThePittsburghPromise, 2021). The major requirements of being qualified for Promise scholarship are graduating from a secondary school in Pittsburgh Public Schools, having a high-school cumulative GPA greater than or equal to 2.5, having a high-school attendance rate no less than 90%, and planning to enroll in a college or university in Pennsylvania (ThePittsburghPromise, 2021). Although the goal of the Promise scholarship is to help students from Pittsburgh Public Schools pursue and finish their college study, little is known about whether this award really helps or motivates students to pursue post-secondary education. Therefore, Pittsburgh Public Schools initiates a project to examine Promise scholarship use and post-secondary retention of students from Pittsburgh Public Schools. The client of this project is Pittsburgh Public Schools, and Steven Greene from Pittsburgh Public Schools is the Single Point of Contact from the client's side. The two primary research questions in this project are as follows:

- Investigate factors that influence whether students received Promise scholarships (abbreviated as scholarship analysis).
- Evaluate factors that influence students' retention in college and compare retention between different groups (abbreviated as retention analysis).

2 Data

For this project, we have 10 data sets in total. These 10 data sets cover information from 2014 to 2020. The table below shows the basic information of these 10 data sets.

Table 1: Basic Information of 10 Data Sets

Data	**Meaning of Data**	**Number of Observations**	**Number of Variables**
School Enrollment	All enrollment records to and from PPS schools	6833406	14
Attendance	Attendance data of students in high schools	109428	10
Demographics	Demographic information of students in each semester in high school	19039	11
NSC	Semester college enrollment records of students	5629	11
SAT	Highest SAT scores for students	3143	6
AP	AP exams and scores taken by students	5352	5
GPA	All end-of-year cumulative GPA during students' high school careers	19436	5
Keystone	Scores that students received on the Keystone Assessment based on different subject	37331	8
CTE	Career and Technical Education(CTE) certifications earned by students in high school	1179	6
Scholarship	Information about students eligibility for Promise scholarship and receipt of Promise scholarship	2265	8

For the first research question about investigating factors that influence whether students received Promise scholarships (abbreviated as scholarship analysis), we join all the data sets but the NSC data set. The total number of observations in the joined data set for scholarship analysis is 1708. Notice that the 1708 observations here are students who received education and graduated from secondary schools in the Pittsburgh Public Schools district.

For the second research question about investigating factors that influence students' retention in college (abbreviated as retention analysis), we join all of the data sets above but filter out students who did not go to college in Pennsylvania (abbreviated as PA). The reason why we only focus on students who went to PA colleges is that our second research question investigates how Promise scholarship helps students' retention in college, and only students who enroll in colleges in PA are able to receive and use Promise scholarship. Since we only consider students who went to PA colleges here, the data size for retention analysis is smaller than that for scholarship analysis. Specifically, the total number of observations in the joined data set for retention analysis is 1378. Among them, 13 observations began their college study in 2017, 574 observations began their college study in 2018, 698 students began their college study in 2019, and 93 students in 2020.

The definitions of variables used in the project are as follows.

Table 2: Variable Definition

Variable	**Definition**	**Data set**
RandomID	Unique student ID	All Data Sets
QualifiedforCorePromise	Eligibility for Promise(binary)	Scholarship
EverReceivedPromiseAward	Whether students received Promise(binary)	Scholarship
Gender	Gender of students	Demographics
Race	Race of students	Demographics
ELLStatus	English language level of students	Demographics
IEPGroup	Whether students need special education	Demographics
EconDisab	Economic status of students	Demographics
Num_AP(created)	Number of AP tests taken	AP
CumulativeGPA(created)	Cumulative GPA	GPA
AttendanceRate(created)	1-("absent unexcused"/ "total days")	Attendance
KeystoneMean(created)	Average keystone scores	Keystone
SAT_Total(created)	Highest SAT score	SAT
Num_CTE(created)	Number of Career and Technical Education(CTE) Certifications	CTE
MagnetInd	Whether students go to magnet schools(binary)	Enrollment
GradYear	Year in which students graduated from high school	Scholarship
Enrollment_Begin	When a student enrolled in a college semester	NSC
Enrollment_End	When the college semester ended	NSC
College_State	State where the college is located	NSC
Retention(created)	Enrollment_End-Enrollment_Begin	NSC
Start_College_Year(created)	Year in which a student first enrolled in college	NSC

Before we dive into formal statistical analyses, we first hope to compare and contrast students who are qualified for Promise scholarship and those who are not according to their racial groups and gender identity. We also want to compare and contrast students who received Promise scholarships versus those who did not according to a few different factors (not only race and gender but also include other associated variables this time). In addition, we hope to check how the current cut-offs on GPA and attendance relate to students' qualification and receipt status of Promise scholarship. To fulfill the three goals above, we conduct an initial exploratory data analysis in which we create proportional bar plots, box plots, and scatter plots. The results of the initial exploratory data analysis are shown as follows:

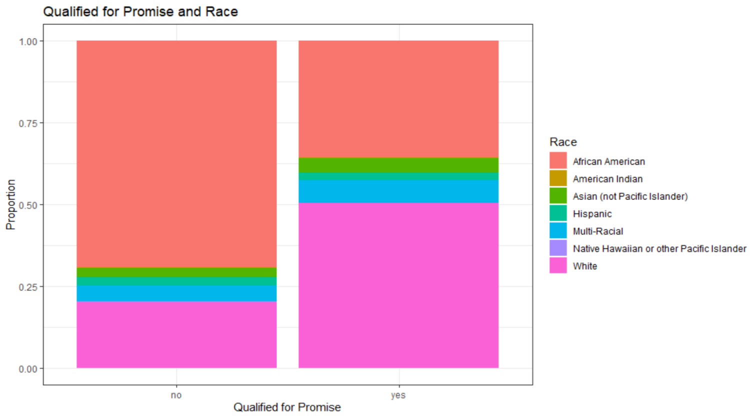


Figure 1: Students' Eligibility for Promise Scholarship in Different Racial Groups

Figure 1 displays the racial composition of students who are qualified for Promise scholarship and students who are not. The population of Figure 1 is all students in the joined data set of demographics and scholarship (2223 observations). From Figure 1, we observe that while the proportion of white students is highest among people who are eligible for Promise scholarship, the proportion of African American students is highest among people who are not eligible for Promise scholarship.

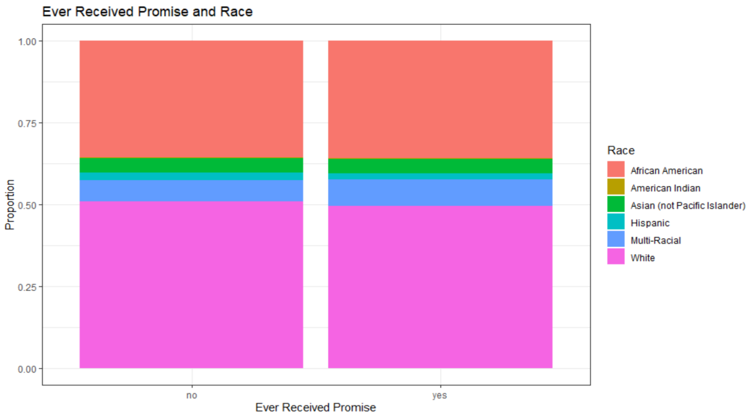


Figure 2: Students' Receipt of Promise Scholarship in Different Racial Groups

Figure 2 displays the racial composition of qualified students who received Promise scholarship and those who did not. The population of Figure 2 is all students who are qualified for Promise scholarship in the joined data set of demographics and scholarship (1560 observations). From Figure 2, we see that the distribution of race among students who received Promise scholarship is similar to that among students who did not receive it.

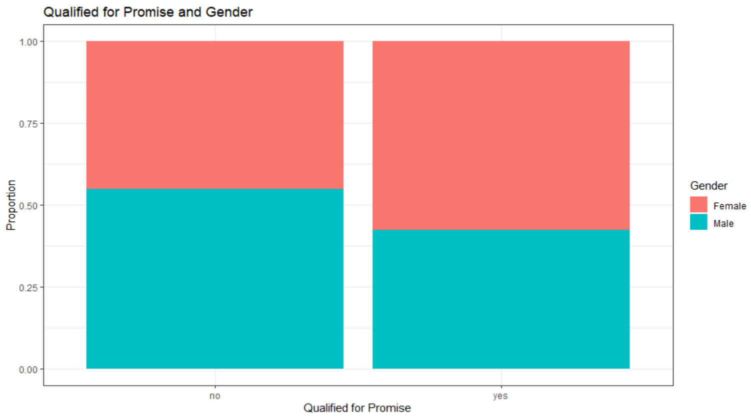


Figure 3: Students' Eligibility for Promise Scholarship in Different Gender Groups

Figure 3 displays the gender composition of students who are qualified for Promise scholarship and students who are not. The population of Figure 3 is all students in the joined data set of demographics and scholarship (2223 observations). From Figure 3, we find the proportion of females among students who are qualified for Promise scholarship is higher than that of males.

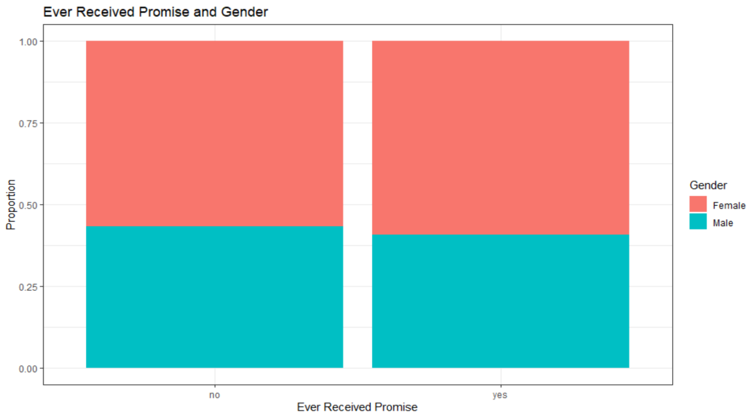


Figure 4: Students' Receipt of Promise Scholarship in Different Gender Groups

Figure 4 displays the gender composition of qualified students who received Promise scholarship and those who did not. The population of Figure 4 is students who are qualified for Promise scholarship in the joined data set of demographics and scholarship (1560 observations).

From Figure 4, we learn that the proportion of females among students who received Promise scholarship is relatively equal to that of males. This result is different from what we get from Figure 3, in which there is a slightly higher proportion of females that are qualified for Promise scholarship compared to males. The discrepancy in these two findings might imply that female students tend to have slightly better academic performance than male students in school and thus have a larger qualified proportion, but further study is needed to verify this guess.

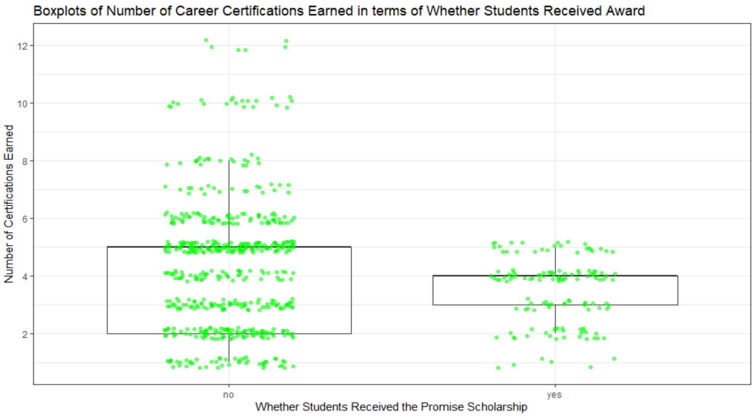


Figure 5: Box Plot of Number of Certifications Earned vs. Receipt of Promise Scholarship

Since the variable Number of Certifications Earned is a quantitative one, we use box plot, instead of proportional bar plots above, to evaluate whether the number of career certifications earned might differ by students' acceptance of Promise scholarships. The population of Figure 5 is all students in the joined data set of CTE and scholarship (698 observations). From Figure 5, we observe that students who received Promise scholarship on average earned less career certifications in high school than students who did not receive Promise scholarship, since the median line (the bold black horizontal line) of the box plot for students who did not receive scholarships is higher than that for students who did receive it. This finding indeed aligns with our expectation that students with more career certifications are less likely to enroll in college after high school and thus less likely to receive Promise scholarship.

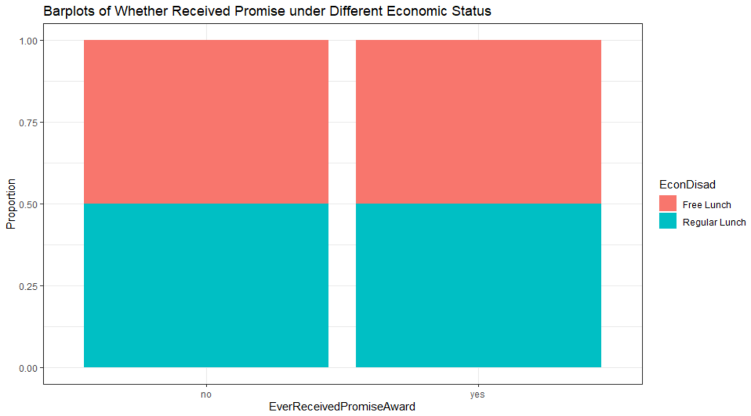


Figure 6: Students' Receipt of Promise Scholarship Under Different Economic Status

Figure 6 looks at the composition of economic status of qualified students who received Promise scholarship and who did not. The population of Figure 6 is all students in the joined data set of demographics and scholarship. From Figure 6, we learn that the distribution of students' economic status among students who received Promise scholarship is similar to that among students who did not receive. Thus, students' economic status might be unrelated to their likelihood of receiving a Promise scholarship.

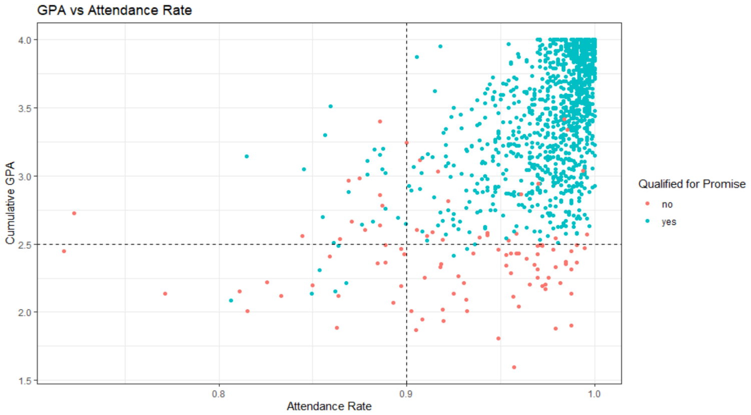


Figure 7: Scatter Plot of GPA and Attendance Based on Qualification for Promise Award

Figure 7 shows the scatter plot of GPA and attendance rate for all students. The points are colored based on whether students are qualified for Promise scholarship or not. The horizontal black dashed line is the GPA cut-off (i.e. GPA >= 2.5), and the vertical black dashed line is the attendance rate cut-off (i.e. attendance rate >= 90%). The two dashed lines divide the quadrant into four squares. The upper right square should only contain students who are qualified for Promise award, so ideally, all the points in this area should be green. Also, the dots in the other three squares should all be red in theory.

However, there exists red points falling in the upper-right square and green points falling out of this region. There might be two reasons. Firstly, we lack the data for students who transfer to schools outside the Pittsburgh Public Schools district. For these transferred students, their GPA or attendance rate could have changed, so their eligibility for Promise award also changed. Secondly, Promise scholarship makes exceptions for some students, so we have students who do not satisfy GPA and attendance criteria but are still qualified for Promise scholarship.

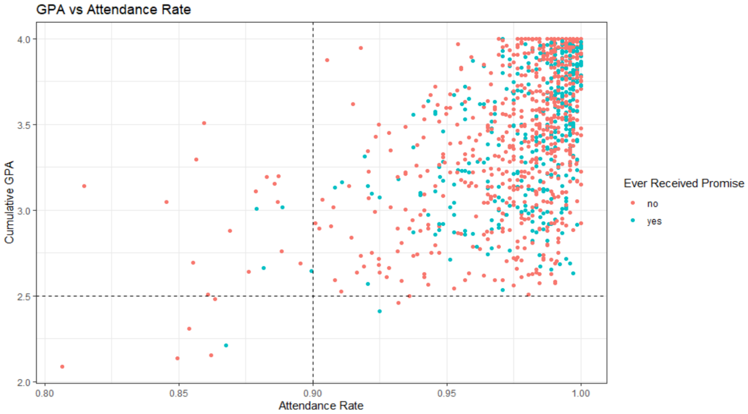


Figure 8: Scatter Plot of GPA and Attendance Based on Receipt of Promise Award

Figure 8 shows the scatter plot of GPA and attendance rate only for students who qualified for Promise, and each dot is colored based on whether the corresponding student received the Promise scholarship or not. From the plot, we know that most of the less-qualified students did not receive Promise scholarships. Also, some highly-qualified students did not receive Promise scholarships either. The reason might be that they attended college outside Pennsylvania and thus lost the right of using the Promise scholarship to pursue their post-secondary education.

3 Methods

All analyses in this project are carried out with R programming language and environment (RStudio Team, 2021). Also, the analyses of this project consist of two parts, one part for each research question. As a reminder, the first research question, which is scholarship analysis, is investigating factors that influence whether students received Promise scholarship; the second research question, which is retention analysis, evaluates factors that influence students' retention in college and make comparisons of retention between different groups.

3.1 Research Question 1: Scholarship Analysis

To answer the first research question, we conduct two logistic regression analyses. Logistic regression is analogous to linear regression, except that the outcome variable is binary instead of quantitative. The variables used for the two logistic regressions are the same, but the sample each logistic regression performs on is different. The first logistic regression includes all students in Scholarship data set; the second logistic regression only includes students who qualify for Promise scholarship. Thus, these analyses examine different aspects: the first assesses the general trend among all students, whereas the second assesses the trend specifically for students who are most likely to enroll in college in PA since they are already qualified for Promise given their GPA and attendance rate. The response binary variable is *EverReceivedPromiseAward* in scholarship data. We used the scholarship data as the base table and left joined predictor variables *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, CumulativeGPA, MagnetInd*. These chosen variables are the most related indicators that would affect students' enrollment in post-secondary institutions in PA based on the EDA (Exploratory Data Analysis). However, we modify the number of categories in the variable *Race*. The original variable *Race* includes seven different races: African American, American Indian, Asian (not Pacific Islander), Hispanic, Multi-racial, and White. Since the majority of students are African American and White, the other races have limited sample size. We categorize all races other than African American and White as Others, so the resulting variable *Race* has only three categories: African American, White, and Others.

For the first logistic regression, we use all 1708 records, which include all qualified/unqualified students for Promise. Then stepwise variable selection using AIC as a criterion is used to find the variables that have the most significant effects on student's enrollment in post-secondary institutions in PA. Stepwise variable selection is the process of adding or removing variables from a model sequentially, depending on certain criteria. In this case, the criteria we choose is AIC. For stepwise variable selection, the base model includes *Race* and *Gender*. Because based on EDA, we find there are discrepancies between different races and gender on whether they received Promise scholarship. The full model includes every variable we mentioned above. Then, we conduct forwards, backwards, and both-ways variable selection using AIC as a criterion; all three methods generate the same results: the variables that are selected as significant variables are GPA, attendance rate, Keystone average, ELL group, and Magnet school indicator. However, since *Race* and *Gender* are included in the base model, they are also included in the results table even though they are not significant. The results table and detailed analyses are given later in the Results section.

However, the above analyses only give a general understanding of how each variable is related to students' post-secondary enrollment. To understand what factors would affect students' enrollment in post-secondary institutions among those who already qualified for Promise, we run another logistic regression on the subset of students who are marked as "yes" for binary variable *QualifiedforCorePromise*. After filtering the qualified students, 1357 out of 1708 observations are preserved for further analysis. Then, the same procedure for the first logistic regression is performed. The base model had variables *Race* and *Gender*, and the full model had all variables we mentioned before. We conduct forwards, backwards, and both-ways stepwise variable selection using AIC, and all methods give the same results. The results show GPA, attendance rate, Keystone average, ELL group, and Magnet school indicator are significant variables. *Gender* and *Race* are not significant in affecting students' enrollment in college. The variables selected for the first regression and the second regression are identical, except that the coefficients and significant level are slightly different. We will give an in-depth discussion later in the Results section.

3.2 Research Question 2: Retention Analysis

To answer the second research question, we conduct both exploratory data analysis and regression analysis. Because the Promise scholarship is designed for students who chose colleges in Pennsylvania, we restrict our analysis only to students who went to colleges in Pennsylvania. A students' retention is calculated as the cumulative sum of the difference between *Enrollment_Begin* and *Enrollment_End* in the NSC data. In other words, retention refers to the total number of days a student has stayed in college. For fair comparisons of students' retention since the beginning of their colleges, we group students by the first year of their college enrollments.

3.2.1 Exploratory Data Analysis

In the exploratory data analysis (EDA), we create box plots of students' retention in different groups.

First, we compare retention between students who received Promise scholarships and those who did not. To conduct the first comparison, we join NSC and scholarship data by students' random ID, and use the variable *EverReceivedPromiseAward* to flag whether students received scholarships or not. Then, we create paired box plots to investigate whether college retention in days will differ between students who received promise scholarships and those who did not.

Second, we compare retention among students in different racial groups. To conduct the second comparison, we join NSC and demographics data by students' random ID. Similar to Research Question 1, we categorize the variable *Race* as White, Black, and Other. Then, we use paired box plots to observe racial differences in retention.

Third, we investigate the interaction between receipt of scholarship and race. We redo our analysis for racial difference separately for students who received the scholarship and students who did not.

To validate our insights, we apply Welch t-tests to examine statistical significance for both retention comparisons. We also conduct Bartlett tests first to check whether the variance of retention differs between groups, and choose either equal-variance t-test or unequal-variance t-test.

3.2.2 Multivariate Regression for Various Ranges

From the scholarship analysis, we learn that GPA and attendance, the requirement for Promise, are indeed very correlated with getting the Promise scholarship, so we want to ensure our comparison for scholarship on retention to be comparable at baseline. In the regression analysis, we focus on students with similar academic performance (in terms of GPA and Attendance), once they enrolled in college, to explore other possible factors that influence students' retention in college. The predictors we consider including in the multivariate regression are *AttendanceRate, Num_AP, Num_CTE, KeystoneMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, SAT_Total, CumulativeGPA, MagnetInd, EverReceivedPromiseAward, QualifiedforCorePromise, Semester*, and interaction between *Race* and *EverReceivedPromiseAward*. The outcome variable of the multivariate regression is *Retention*, which is the number of days a student has stayed in college. Since a student's retention in college is influenced by the first year of their college enrollments, we construct one model for each college starting year. We only focus on the year 2018 and 2019, because of the insufficient samples for 2017 and 2020.

• Whole Range of GPA and Attendance
For the whole data set, we conduct Poisson regressions for retention. Poisson regression is analogous to linear regression, except the outcome variable is a count. This nature of Poisson regression indeed works well for the analysis that uses *Retention* (number of days a student has stayed in college) as its outcome variable. In addition, the ANOVA test between linear regression and Poisson regression shows that the fitness of Poisson regression is slightly better than that of linear regression (please refer to technical Appendix C for more details). Thus, we focus on Poisson regression models rather than linear regression in multivariate regression analysis for the whole range of GPA and Attendance.

• Box Range of GPA and Attendance
To account for the effect that students with better academic performance might have better retention, we perform treatment analysis for 4 box ranges:

Attendance Lower	Attendance Upper	GPA Lower	GPA Upper
0.86	0.94	2.1	2.9
0.84	0.96	2.0	3.0
0.82	0.98	1.9	3.1
0.80	1.00	1.8	3.2

Figure 9 provides a visualization on students' high school performance, for all students who enrolled in college between 2017 and 2020. The smallest box consists of the most comparable students.

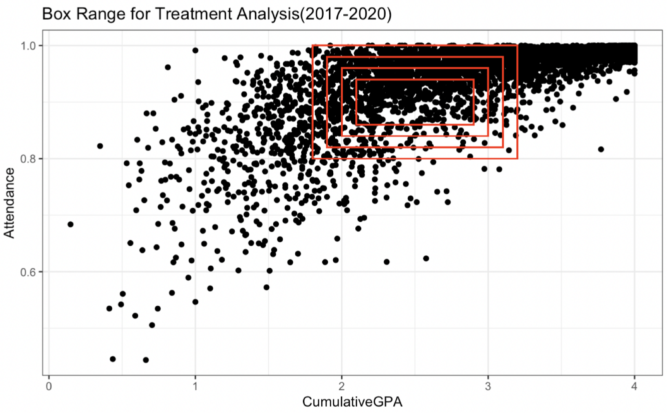


Figure 9: Visualization of Box Ranges

For each box, we first perform a treatment-control test between retention and *EverReceivedPromiseAward*. After we obtain the significance for the t-scores, we run linear regressions using the same predictors in the Poisson regression. We include other variables to test if the treatment effect of the scholarship still holds, even after adjusting for other variables that are not included in the t-test.

For both regressions, the predictors are chosen by stepwise selection method on AIC. After obtaining the final models from stepwise selection method, we perform model diagnosis with residual deviance test, diagnostics plots, and binned plots, and interpret the coefficients of the final model.

4 Results

4.1 Scholarship Analysis

4.1.1 Analysis on Qualified and Unqualified Students

In this part, we present the results of logistic regression which are based on all qualified and unqualified students. This logistic regression selects *Race* and *Gender* as the base variables. The model with its coefficients is presented and possible explanations to the coefficients is given. After conducting stepwise selection, the logistic model is of the form:

$$ln(odds(\textit{EverReceivedPromise})) = \beta_0 + \beta_1 RaceOthers + \beta_2 RaceWhite + \beta_3 GenderMale$$
$$+\beta_4 CumulativeGPA + \beta_5 AttendanceRate + \beta_6 KeystoneMean + \beta_7 ELLStatusNotinELL + \beta_8 MagnetInd$$

The values for all coefficients are displayed below. The coefficient describes the relationship between the predictor variable and the outcome variable. In logistic regression, a positive coefficient indicates this variable increases the likelihood of the outcome event. A negative coefficient indicates this variable decreases the likelihood of the outcome event. In this study, the outcome event is receiving Promise scholarship.

Table 3: Modeling Results of Analysis on Qualified and Unqualified Students

Variable	**Coefficient**	**Standard Error**	**P-Value**
(Intercept)	-3.725	2.868	0.194
RaceOthers	0.056	0.188	0.767
RaceWhite	-0.267	0.146	0.068 .
GenderMale	-0.091	0.117	0.435
CumulativeGPA	0.903	0.158	1.15e-08 ***
AttendanceRate	8.051	2.027	7.13e-05 ***
KeystoneMean	-0.006	0.002	4.58e-04***
ELLStatusNot in ELL	1.101	0.438	0.012 *
MagnetInd1 in ELL	0.249	0.116	0.032 *

The table shows the results after we perform stepwise variable selection based on AIC. By observing the coefficients, we find that students who have a high GPA and Attendance rate are more likely to receive Promise, which means they are more likely to successfully enroll in a PA college. This result is not surprising since Promise chooses eligible students based on their GPA and attendance rate. Also, students who are not in ELL group and students who ever attended a magnet school have a higher rate of receiving Promise scholarships. But one odd thing is the keystone mean score shows a negative relationship with students' enrollment in PA college. This is very different from our intuition that students who have higher keystone scores mean they have better academic performance at school; in this case, they are more likely to enroll in a college successfully. We think the reason for this situation might be that for those higher-achieving students in Keystone exams, they tend to go to college outside of PA; thus they will not be receiving the Promise scholarship even if they qualified. As for *Race* and *Gender* which are chosen as base variables, they are not significant in our model except *RaceWhite*. *RaceWhite* has P-value 0.068, with a significance level of 90%. *RaceWhite*'s coefficient is -0.267. This coefficient indicates that the log odds to receive Promise for white students are 0.267 lower than African American students. To be more specific, the odds to receive Promise for white students are $(1 - e^{-0.267}) * 100\% = 23.4\%$ lower than black students. We think the reasons are that even though a larger proportion of white students qualified for Promise, they either choose not to attend college or attend a college outside of PA.

4.1.2 Analysis on Qualified Students only

For this logistic regression analysis, we only focus on the subsets of qualified students. The same procedure and choices of variables are similar to the previous model. This model has the same form as above. The values for all coefficients are displayed below.

Table 4: Modeling Results of Analysis on Qualified and Unqualified Students

Variable	**Coefficient**	**Standard Error**	**P-Value**
(Intercept)	-4.917	3.335	0.140
RaceOthers	0.107	0.197	0.588
RaceWhite	-0.205	0.154	0.185
GenderMale	-0.023	0.123	0.852
CumulativeGPA	0.475	0.196	0.015 *
AttendanceRate	10.402	2.576	5.38e-05 ***
KeystoneMean	-0.005	0.002	0.002 **
ELLStatusNot in ELL	0.784	0.462	0.090 .
MagnetInd1 in ELL	0.225	0.122	0.066 .

The table shows the results after we perform stepwise variable selection based on AIC. By comparing the coefficients with the previous model, we find that the variable selection method chooses the same variables, except that RaceWhite no longer has a 90% significance level. And their relationship with the predictor variable *EverReceivedPromiseAward* remains the same. For example, *CumulativeGPA, AttendanceRate, ELLStatusNot in ELL, MagnetInd1* are positively related to the log odds of students' enrollment in postsecondary institutions in PA. Meanwhile, *KeystoneMean* is negatively related to the log odds of students' enrollment in postsecondary institutions in PA. *Race* and *Gender* are not significant; thus it is not necessary to give a detailed explanation for their variables.

4.2 Retention Analysis

4.2.1 EDA

We compare students' retention from two perspectives: whether a student received the scholarship and the student's race. We only conduct statistical tests for 2018 and 2019, due to limited observations for 2017(13 observations) and insufficient time lag for 2020. For racial comparison, we mainly focus on black and white students, as they constitute the majority of students. Our results suggest there exists an interaction between racial differences and whether a student received the scholarship.

- **Retention between students who received the scholarship and who did not**

Figure 10 shows that, in general, students who received the scholarship have higher retention on average except for the year 2019. After checking the equal variance assumption with the Bartlett test, the Welch t-test shows that the difference is significant for 2018(p-value = 5.98e-10) but not for 2019(p-value = 0.60). We conclude for 2018, students who received the scholarship have better retention. In particular, this indicates that students are dropping out during their sophomore year.

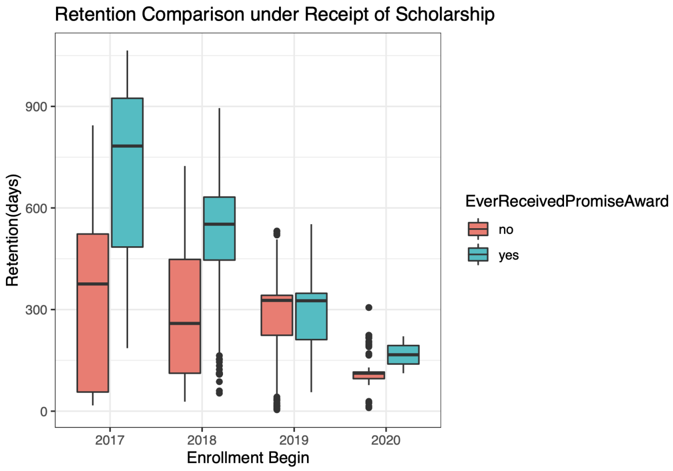


Figure 10: Students' Retention by the Receipt of Scholarship

- **Retention between different students' races**

Figure 11 shows a comparison of retention between different races. From the plot, we observe a big retention difference between races for 2018 and slight difference for 2019. Our one-way ANOVA test shows that the difference for both 2018 and 2019 are significant. We conclude that for both 2018 and 2019, the mean retention between races is not equal. Notice that the differences appear to be larger for 2018 than 2019, indicating that students drop out more during their sophomore year.

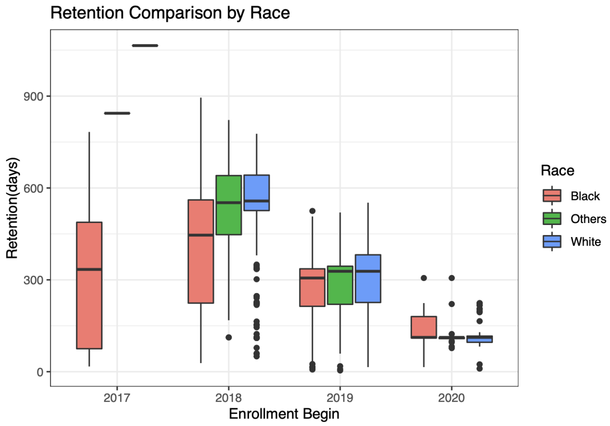


Figure 11: Students' Retention by Races

Next, we further investigate the differences between black and white students. After checking the equal variance assumption with the Bartlett test, the Welch t-test shows that the difference is significant for both 2018(p-value = 8.48e-06) and 2019(p-value = 8.78e-08).

- **Interaction between receipt of scholarship and race**

Given some significant differences for receipt of scholarship and race, we will investigate their interactions by accessing the racial difference separately for students who received the scholarship and students who did not. The left figure of Figure 12 shows the comparison of racial differences for students who received the scholarship. Our analysis shows that for 2018, the difference is significant, but not for 2019. For students who did not receive the scholarship on the right of Figure 12, our analysis shows that for 2018, the difference is not significant, but the difference is significant for 2019. The relevant p-values are shown in Table 5. Based on our results, we decide to include the interaction term between race and receiving of scholarship in the regression analysis, which is used to adjust for other variables beyond scholarship and race.

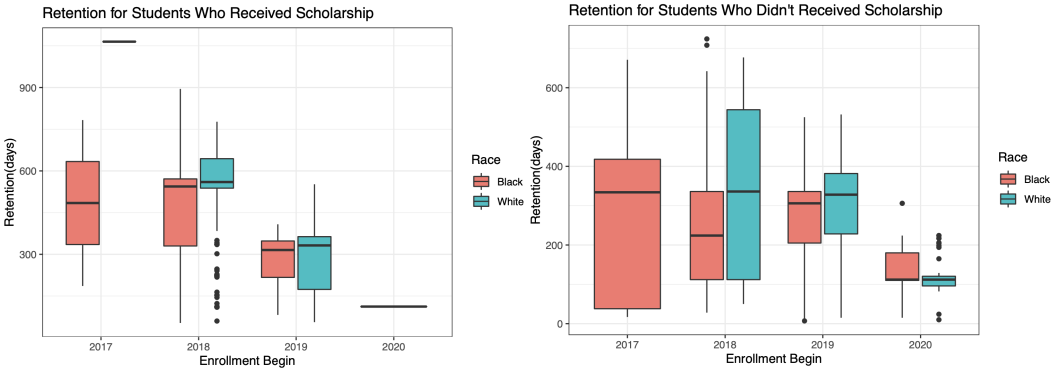


Figure 12: Racial difference in Retention and Receipt of Scholarship

Table 5: P-values for Racial Difference in Retention

Group	2018	2019
Received Scholarship	8.96e-05	0.51
NOT Received Scholarship	0.1	6.20e-09

4.2.2 Multivariate Regression for Various Ranges

- **Poisson regression on retention for students starting college education in 2018**

The null model for the Poisson regression is of the form:

$$\mathit{Retention} = \hat{\beta}_0 + \hat{\beta}_1 Race + \hat{\beta}_2 EverReceivedPromiseAward + \hat{\beta}_3 Race : EverReceivedPromiseAward$$

The selected Poisson regression model is of the form:

$$\mathit{Retention} = \hat{\beta}_0 + \hat{\beta}_1 AttendanceRate + \hat{\beta}_2 Num\_AP + \hat{\beta}_3 Num\_CTE + \hat{\beta}_4 KeystoneMean + \hat{\beta}_5 Race$$
$$+ \hat{\beta}_6 Gender + \hat{\beta}_7 ELLStatus + \hat{\beta}_8 IEPGroup + \hat{\beta}_9 EconDisab + \hat{\beta}_{10} SAT\_Total + \hat{\beta}_{11} CumulativeGPA$$
$$+ \hat{\beta}_{12} MagnetInd + \hat{\beta}_{13} EverReceivedPromiseAward + \hat{\beta}_{14} QualifiedforPromise$$
$$+ \hat{\beta}_{15} semester + \hat{\beta}_{16} Race : EverReceivedPromiseAward$$

where *semester* refers to the semester that a students starts college.

The results of the model selected from stepwise selection on AIC are as follows. A positive significant coefficient indicates a positive influence on retention or higher retention than its control group; A negative significant coefficient indicates a negative influence on retention or lower retention than its control group.

Notice that if a coefficient is positive and statistically significant, its corresponding quantitative or categorical variable will positively influence students' retention in college or have higher retention than its control group, respectively. On the contrary, if a coefficient is negative and statistically significant, its corresponding quantitative or categorical variable will negatively impact students' retention in college or have lower retention than its control group, respectively. In addition, when interpreting the results of any Poisson regression model, we shall first take the exponential form of each coefficient.

Table 6: Results of Poisson Regression Model on Retention (2018)

Variable	**Coefficient**	**Standard Error**	**P-Value**
(Intercept)	6.110	1.441e-01	<2e-16 ***
AttendanceRate	2.246	8.707e-02	<2e-16 ***
Num_AP	4.585e-03	1.041e-03	1.070e-05 ***
Num_ CTE	-2.825e-02	2.743e-03	<2e-16 ***
KeystoneMean	-2.166e-03	9.581e-05	<2e-16 ***
RaceOther	4.463e-02	2.040e-02	0.029 *
RaceWhite	6.300e-02	1.500e-02	2.680e-05 ***
GenderMale	-7.107e-02	4.611e-03	<2e-16 ***
ELLStatusNot in ELL	-5.165e-02	1.620e-02	0.001 **
IEPGroupIEP	2.181e-02	1.123e-02	0.052 .
IEPGroupNot IEP or Gifted	-5.117e-02	6.443e-03	1.99e-15 ***
EconDisadRegular Lunch	1.093e-02	4.630e-03	0.018*
SAT_Total	4.265e-04	2.367e-05	<2e-16 ***
CumulativeGPA	1.697e-01	7.631e-07	<2e-16 ***
MagnetInd	3.736e-02	4.682e-03	1.47e-15 ***
EverReceived-PromiseAward	3.162e-01	9.688e-03	<2e-16 ***
QualifiedforPromises	-4.058e-02	9.452e-03	1.760e-05 ***
semesterSpring	1.279e-01	4.022e-02	0.001***
RaceOther:Ever-Received-PromiseAward	-4.254e-03	2.116e-02	0.841
RaceWhite:Ever-Received-PromiseAward	-4.717e-02	1.540e-02	0.002 ***

From the Table 6, we know that students' attendance rate in high school, number of AP courses taken, SAT scores, cumulative GPA in high school are positively correlated with their retention in college. Among these four variables, the effects of attendance rate and cumulative GPA in high school on retention are more influential compared to the other two variables. Specifically, if a students' attendance rate in high school increases by 10%, his or her mean retention is expected to increase by $(e^{(2.246/10)} - 1) * 100\% = 25.2\%$; also, if a students' cumulative GPA in high school increases by 0.1 points, his or her mean retention is expected to increase by $(e^{(0.1697/10)} - 1) * 100\% = 1.71\%$.

In addition to the four quantitative variables above, some levels in students' racial identity, IEP group (indicator of whether students need special education in high school), economic status variable, magnet schools (indicator of whether students entered a magnet high school), ever received Promise scholarships (indicator of whether students received Promise scholarships), and semester (indicator of which semester students first enrolled in college) are also positively correlated with retention in college. The discrepancies in retention among students with different racial identity, between students who received scholarships and those who did not are relatively noticeable and worthy of our attention. Regarding racial identity, white students have $(e^{0.063} - 1) * 100\% = 6.50\%$ higher mean retention compared to black students. Also, black students who received Promise scholarship have $(e^{0.3162} - 1) * 100\% = 37.2\%$ higher mean retention than those who did not.

On the other hand, students' number of career certifications earned in high school and keystone scores are inversely related to their retention in college, but their effects on retention are not very influential. Regarding categorical variables, some levels in students' gender identity, ELL status (indicator of whether students joined a language learning program), IEP group (indicator of whether students need special education), and qualification for Promise scholarships are inversely correlated with their retention in college. In particular, the discrepancies in retention between males and females, and between qualified and unqualified students deserve our attention. Compared to female students, male students have $(e^{0.07107} - 1) * 100\% = 7.37\%$ lower mean retention; also, students who are qualified for Promise scholarships have $(e^{0.04058} - 1) * 100\% = 4.14\%$ lower retention than students who are not.

In addition, from the coefficient on the interaction term between white students and the binary indicator of whether students received Promise scholarship, we know that the positive effect of receiving Promise scholarships on retention for white students is about $(e^{0.3162} - e^{0.3162-0.04717}) * 100\% = 6.32\%$ less than that for black students.

• Poisson regression on retention for students starting college education in 2019

The null model for the Poisson regression is the same as 2018.

The selected Poisson regression model is of the form:

$$\begin{aligned}\textit{Retention} = \hat{\beta}_0 + \hat{\beta}_1 \textit{AttendanceRate} + \hat{\beta}_2 Num\_AP + \hat{\beta}_3 Num\_CTE + \hat{\beta}_4 KeystoneMean \\ + \hat{\beta}_5 Race + \hat{\beta}_6 Gender + \hat{\beta}_7 ELLStatus + \hat{\beta}_8 EconDisab + \hat{\beta}_9 SAT\_Total \\ + \hat{\beta}_{10} CumulativeGPA + \hat{\beta}_{11} MagnetInd + \hat{\beta}_{12} EverReceivedPromiseAward \\ + \hat{\beta}_{13} QualifiedforPromise + \hat{\beta}_{14} semester + \hat{\beta}_{15} Race : EverReceivedPromiseAward\end{aligned}$$

The results of the model selected from stepwise selection on AIC are as follows:

Table 7: Results of Poisson Regression Model on Retention (2019)

Variable	**Coefficient**	**Standard Error**	**P-Value**
(Intercept)	4.432	1.604e-01	<2e-16 ***
AttendanceRate	6.489e-01	8.490e-02	2.11e-14 ***
Num_AP	3.395e-03	1.218e-03	0.00531 **
Num_CTE	-1.305e-02	2.532e-03	2.07e-07 ***
KeystoneMean	-2.601e-04	1.077e-04	2.07e-07 ***
RaceOther	-3.699e-02	8.317e-03	8.71e-06 ***
RaceWhite	9.875e-04	6.405e-03	0.87748
GenderMale	-5.275e-02	5.012e-03	<2e-16 ***
ELLStatusNot in ELL	1.950e-01	1.834e-02	<2e-16 ***
EconDisadRegular Lunch	2.151e-02	5.154e-03	3.00e-05 ***
SAT_Total	2.291e-04	2.543e-05	<2e-16 ***
CumulativeGPA	1.701e-01	7.749e-03	<2e-16 ***
MagnetInd	2.889e-02	5.126e-03	1.74e-08 ***
Ever-Received-PromiseAward	2.282e-02	1.587e-02	0.15039
QualifiedforPromises	6.454e-02	9.687e-03	2.69e-11 ***
semesterSpring	-2.563e-01	1.147e-02	<2e-16 ***
RaceOther::Ever-Received-PromiseAward	3.113e-01	2.271e-02	<2e-16 ***
RaceWhite::Ever-Received-PromiseAward	-5.807e-02	2.075e-02	0.00514 **

Similarly, for students who started college in 2019 shown in Table 7, their attendance rate in high school, number of AP courses taken, SAT scores, and cumulative GPA are positively correlated with their retention in college. Also, the influences of attendance rate and cumulative GPA on college retention are remarkable compared to the other two variables. In particular, for a 10% increase in attendance rate in high school, students' retention in college is expected to increase by 6.7%; for a 0.1 point increase in cumulative GPA, students' retention in college is expected to increase by 1.7%. Regarding categorical variables that are positively related to students' retention in college, we know that some levels in ELL status (indicator of whether students joined language learning program), economic status, magnet schools (indicator of whether students went to magnet schools), and qualification for Promise scholarships are positively related to their retention in college. The discrepancy in retention between qualified and unqualified students is more noticeable and meaningful compared to the other variables. Specifically, students who are qualified for Promise scholarships have $(e^{0.06454} - 1) * 100\% = 6.67\%$ higher mean retention than those who are not.

Similar to students who started college in 2018, the number of career certifications earned in high school and keystone scores are inversely correlated with retention in college for students who started college in 2019. However, the effects of these two quantitative variables on retention are very limited. With respect to categorical variables, some levels in race and gender are negatively correlated with retention in college, and only the gender discrepancy in retention is worthy of our attention. Specifically, we find that male students have about 5.4% lower mean retention compared to female students.

Similar to the results of the 2018 Poisson regression model, the results of the 2019 one also show that attendance rate and cumulative GPA in high school are positively correlated with students' retention in college, and females tend to have higher retention than males. Nevertheless, the two models present very different results regarding the comparison of retention between qualified and unqualified students. While the 2018 model shows that qualified students have lower mean retention in college than unqualified students, the 2019 model shows the opposite. Also, while the variable *EverReceivedPromiseAward* indicates a strong and positive influence on students' retention in college in the 2018 model, the same variable is not statistically significant at all in the 2019 model.

- **Linear Regression on Different Box Ranges**

Our logistic regression for scholarship receipt analysis indicates that students with higher GPA and attendance rate are more likely to get Promise Award. Our treatment analysis focuses on students with similar GPA and attendance rate, so that students have similar likelihood of getting the scholarship. Table 8 and Table 9 below shows the treatment effect of receiving Promise Award on retention for 4 different box ranges, for the year 2018 and 2019:

Table 8: Treatment Analysis for 2018

	Mean Difference	**Lower bound**	**Upper Bound**	**P-value**	**Significance**
Box1	112.882	-233.952	8.187	0.067	No
Box2	151.920	-228.868	-74.972	0.000	Yes
Box3	143.685	-209.335	-78.035	0.000	Yes
Box4	178.061	-229.340	-126.782	0.000	Yes

Table 9: Treatment Analysis for 2019

	Mean Difference	**Lower bound**	**Upper Bound**	**P-value**	**Significance**
Box1	39.882	-142.535	62.771	0.344	No
Box2	25.170	-94.225	43.885	0.441	No
Box3	43.274	-96.661	10.112	0.105	No
Box4	24.121	-67.447	19.205	0.265	No

Table 10 and Table 11 below show the results of linear regressions for 2018 and 2019. We see that *EverReceivedPromiseAward* and its interaction between Race are selected in all box ranges. For 2018, the linear regression results align well with the t-test, so we conclude that there is indeed a significant effect from the receipt of scholarship on retention. For 2019, the linear regression confirms that there is no significant effect from the receipt of scholarship on retention. However, the regression for box 3 suggests that there is a significant effect from the interaction between *EverReceivedPromiseAward* and Race. Overall, the results align with the results of Poisson regression above. Both regressions find that receiving Promise scholarship has a positive significant effect on students' retention for 2018 but not for 2019, and that the racial interaction is more apparent in 2019 than in 2018. While both regressions select similar predictors, there are fewer significant coefficients in the linear regression of the box-range analysis.

Table 10: Linear Regression for 2018

	Number of Observations	**Variables Selected by Linear Regression(Year 2018)**	**Significant Treatment Effect**
Box1	34	*Race*EverReceivedPromiseAward, Gender, ELLStatus*	No
Box2	75	*AttendanceRate, CumulativeGPA, SAT_Total, Num_AP, Num_CTE, KeystoneMean Race*EverReceivedPromiseAward, Gender, ELLStatus,EconDisad, MagnetInd, QualifiedforCorePromise*	No, but close(p-value = 0.052)
Box3	150	Same as Box 2	Yes(p-value = 0.007)
Box4	235	Same as Box 2	Yes(p-value = 0.00)

Table 11: Linear Regression for 2019

	Number of Observations	**Variables Selected by Linear Regression(Year 2018)**	**Significant Treatment Effect**
Box1	57	*AttendanceRate, CumulativeGPA, SAT_Total, Num_AP, Num_CTE, KeystoneMean, Race*EverReceivedPromiseAward, Gender, EconDisad, MagnetInd, QualifiedforCorePromise*	No
Box2	110	Box 1 Variables + *ELLStatus*	No
Box3	203	Box 1 Variables + *ELLStatus*	No, but significant interaction between *Race* and *EverReceivedPromiseAward*(p-value = 0.037)
Box4	311	Box 1 Variables + *ELLStatus*	No

5 Discussion

5.1 Result Summary

5.1.1 Scholarship Analysis

Regarding the first research question about students' receipt of Promise scholarships, we conduct logistic regression and thus are able to understand what factors affect students' enrollment in post-secondary education in Pennsylvania (abbreviated as PA later). The first regression is conducted on all students from the Scholarship dataset. Based on the model results, we find that the variables that most significantly affected students' enrollment in PA college are *CumulativeGPA, AttendanceRate, KeystoneMean, ELLStatusNot in ELL, and MagnetInd1*. These variables all have at least 95% significance level. As for *Race* and *Gender*, only *RaceWhite* has 90% significance level. In this case, besides high GPA and attendance rate, a relatively lower Keystone average score, students who are not in ELL group, or students who are in a magnet school also add the likelihood of attending a college in PA and use the scholarship that Promise provided.

Moreover, the logistic regression on the subset of students who qualified for Promise provides us insight into what might affect their enrollment in PA college once they qualified for Promise. Comparing the results for both logistic regression models, we can find the significant variables that are selected and their relationship with outcome variables are very similar. The results for the second logistic regression indicate that GPA, attendance rate, and Keystone average will significantly affect students' enrollment in PA college. A student with high GPA, attendance rate, and relatively low Keystone mean score will be more likely to attend a college in PA. Students who are not in ELL group and who study in a magnet school add the likelihood of attending a college in PA. Variables *Race* and *Gender* are not significant in this model.

5.1.2 Retention Analysis

Regarding the second research question about students' retention in college, we conduct exploratory data analysis, t-tests, Poisson regression, and linear regression to explore what factors might influence students' retention in college.

From both EDA and t-test, we find that whether students received Promise scholarships is associated with their retention in college, and their retention in college differs between black students and white students.

From the Poisson regression for students who started college in 2018, we know that students with higher attendance rate, with higher cumulative GPA, and who received Promise scholarships are more likely to have higher retention in college. Additionally, male students have lower retention in college than female students, and students who are qualified for Promise scholarship have lower retention than those who are not. Also, while white students have higher mean retention than black students, the positive effect of Promise scholarship on retention is smaller for white students than for black students. In other words, while receiving Promise scholarship does help students who started college education in 2018 with their retention, the effectiveness of this help is better for black students compared to white students.

We find similar results from the Poisson regression for students who started college in 2019 as those from the 2018 regression. First, the output of the 2019 model also presents positive relationships between attendance rate, cumulative GPA, and retention in college. Second, it also shows that male students have lower retention than female students. Nevertheless, different from students who started college in 2018, students who are qualified for Promise Scholarship have higher retention than those who are not. Unlike the 2018 regression model, whether students received Promise scholarships is no longer a significant factor that influences their retention in college. The effect of Promise scholarship is significant for 2018 but not for 2019 might be because students are dropping out more frequently in their second year than their first year.

The box-range analysis shows that the effect of scholarship is significant for all boxes except for the first one for 2018. One possible explanation is limited observations. There are fewer observations in 2018, and box 1 only has 34 observations; such a small sample size also makes the model fit for box 1 worse than other boxes. Compared to 2018, there is no significant treatment effect for all box ranges in 2019. We suspect that the effect becomes more obvious for more senior students. Even though we found only box 3 has significant interaction effects, this is due to the small sample size for Other race students and its retention distribution is highly skewed. Besides the *EverReceivedPromiseAward*, we noticed that cumulative GPA also has a significant effect for box 3 and 4. This might be due to the wider range of GPA in these two boxes. We may replicate this box analysis when we obtain more observations in the future.

Therefore, based on our analysis for the year 2018 and 2019, we conclude that there are multiple factors that affect retention in college institutions. Regarding the Promise scholarship, there is no definite conclusion, but we may conclude that the Promise scholarship tends to help students with their retention in college, and the effectiveness varies by race and becomes more apparent for more senior students. There are also other variables that affect student's retention: attendance rate, cumulative GPA, and gender. Higher attendance rate and higher cumulative GPA in high school indicate higher retention in college institutions; retention of female students is higher than that of male students.

5.2 Take-home Policy

Based on the two logistic regression models in scholarship analysis, we find that the three most significant variables are cumulative GPA, attendance rate, and average Keystone score. Among the three variables, cumulative GPA and attendance are the two criteria for Promise scholarship eligibility. Despite the significant relationship between average Keystone score and the likelihood of receiving Promise scholarship, we still do not recommend adding this variable as an additional criterion for Promise scholarship eligibility. Firstly, the average Keystone score is strongly but negatively related to their likelihood of receiving Promise scholarship. Moreover, we lack a standardized measurement for students' performance in Keystone exams, since the number of Keystone exams as well as the subjects of Keystone exams taken by each student are different from one another. Thus, it is not appropriate to set a cut-off based on students' Keystone scores to evaluate who is more eligible for Promise scholarship. From the two Poisson regression models in retention analysis, we know that higher attendance rate and higher cumulative GPA in high school indicates higher retention in college institutions. Thus, using cumulative GPA and attendance rate in high school as the criteria for Promise scholarship does help Pittsburgh Public Schools give awards to students who will effectively use it to pursue post-secondary education. Therefore, based on the results of both scholarship analysis and retention analysis, we suggest that Pittsburgh Public Schools shall keep cumulative GPA and attendance rate as the criteria for its Promise scholarship.

From retention analysis, we also learn that the effectiveness of receiving Promise scholarships in improving students' retention in college differs by students' racial identity. Thus, we suggest that the criteria of Promise scholarship should be tailored for different racial groups. In other words, instead of having the same Promise award criteria for all students, Pittsburgh Public Schools might consider setting up different criteria for students with different racial identities.

5.3 Limitations

This project has three major limitations, and these limitations affect both the generalizability and credibility of our findings. First, for both research questions, we have a limited number of observations to study. Specifically, we only have 1708 observations for the first research question and 1378 observations for the second one. Second, when working on the analysis, we assume that students who are not included in the scholarship data are those who did not receive Promise scholarships for convenience, but this assumption might be invalid and thus ruins the credibility of our findings. Third, the fitness of the Poisson regression models in retention analysis is not very ideal, so the results returned by this model might not be very convincing.

5.4 Next Steps

For the logistic models that analyze the relationship between variables and students' receipt of Promise scholarship, we will try to add more predictors in the models and obtain more comprehensive and insightful findings compared to our current analysis. If no additional information is available, we will try to incorporate interaction terms to explore the relationship between existing predictors. For example, we could add the interaction term of average Keystone score and number of subjects, or the interaction between number of AP exams taken and average AP score. Although interpreting interaction terms is complicated, it might improve the fitness of the current models and provides more details about our results.

For the box-range retention analysis, we suggest performing a two-stage least squares (TSLS) analysis. There are students with low GPA or attendance rate who nonetheless got the scholarship, and there are also students that met the criteria but did not receive the scholarship. Hence, there are two stages for the retention regression: first, instead of using the raw data for the receipt of scholarship, we run a linear regression for *EverReceivedPromiseAward*, with a binary indicator for students that pass the criteria and other variables. Second, replace the *EverReceivedPromiseAward* in the retention regression with the predicted *EverReceivedPromiseAward* in the last stage. TSLS is useful for discontinuity design, where treatment is determined by the threshold of one or more covariates (Guido Thomas, 2007).

References

[1] Imbens, Guido, and Thomas Lemieux. “Regression Discontinuity Designs: A Guide to Practice.” Https://Www.nber.org/Papers/w13039, Working Paper 13039, Apr. 2007.

[2] RStudio Team (2020), RStudio: Integrated Development Environment for R, RStudio, PBC, Boston, MA. http://www.rstudio.com/

[3] Sheather, S.J. (2009), A Modern Approach to Regression with R. New York: Springer Science Business Media LLC.

[4] The Pittsburgh Promise. (n.d.). https://pittsburghpromise.org/.

Technical Appendix A: Initial EDA

Data Wrangling

```
# Library packages
library(plyr)
library(dplyr)
library(ggplot2)
library(readxl)
library(rvest)
library(robotstxt)
library(tidyverse)
```

Data Sets for Figure 1-4 & Figure 6

```
# Import scholarship data
scholarship <- read_excel("DR-826 Pittsburgh Promise scholarship use.xlsx", sheet = "Sheet1")

# Clean scholarship data
names(scholarship)[1]<-"RandomID"
scholarship$GradYear <- as.character(scholarship$GradYear)
```

```
# Import demographics data
demo1415 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1415")
demo1516 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1516")
demo1617 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1617")

# Clean demographics data
demo1416 <- join(demo1415,demo1516,type="full")
demo <- join(demo1416,demo1617,type="full")

# Filter observations for students' senior year
demo_senior <- demo %>%
  select(-c("Cohort")) %>%
  mutate(GradYear = substr(as.character(SchoolYear),5,8)) %>%
  filter(GradeCode == 12)
demo_senior$RandomID <- as.character(demo_senior$RandomID)
```

```
# Join demographic data in senior year and scholarship data together
# Only senior year since qualification is evaluated after graduating from high school
demo_scholarship <- inner_join(demo_senior, scholarship, by = c("RandomID", "GradYear"))

# Filter repetitive observations out
demo_scholarship_final <- distinct(demo_scholarship) #2223 obs in total
```

Data Sets for Figure 5

```
# Import CTE data
CTE <-
  read_excel("DR-837 Career and Tech. Education (CTE) outcomes (industrial-recognized certificates).xls
             sheet = "CTE data")

# Generate column that calculates the total number of certifications earned by each student
cte_id_check <- CTE %>%
  group_by(RandomID) %>%
  dplyr::summarise(num_cte = n())
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
# Generate new CTE data that includes a column showing the total number of
# certifications earned by each student.
CTE_new = left_join(CTE, cte_id_check, by = "RandomID")
```

```
# Manipulate CTE and Scholarship data to eliminate to prepare for later joint data set
CTE_new <- CTE_new %>%
  select(-c("Cohort"))
CTE_new$RandomID <- as.character(CTE_new$RandomID)

# Join CTE and Scholarship data
CTE_scholarship = inner_join(scholarship, CTE_new, by = "RandomID")
CTE_scholarship <- distinct(CTE_scholarship)
```

Data Sets for Figure 7 & Figure 8

```
# Import attendance data
attendance <- read_excel("DR-830 Attendance.xlsx",
    sheet = "Attendance data")
```

```
# Clean attendance data
attendance_new <- attendance %>%
  mutate_at(c(1, 3:9), as.factor)
names(attendance_new)[10] <- "CountofDays"
# Create attendance_scholarship table: attendance left_join scholarship
scholarship_new <- scholarship %>%
  mutate_at(c(2:8), as.factor)
attendance_scholarship <- attendance_new %>%
  left_join(scholarship_new, by = c("RandomID"="RandomID"))
# Create a long table where sum each student's recorded days based on attendance status
attendance_scholarship_long <- attendance_scholarship %>%
  dplyr::select(RandomID, AttendanceStatus, CountofDays) %>%
  group_by(RandomID, AttendanceStatus) %>%
  dplyr::summarise(TotalDays = sum(CountofDays))
# Create a wide table that each student takes 1 row
attendance_scholarship_wide <- attendance_scholarship_long %>%
  spread(AttendanceStatus, TotalDays) %>%
  dplyr::select(-’NULL’) %>%
  replace(is.na(.), 0)
# Add total count of days
attendance_scholarship_wide$totalDays <- rowSums(attendance_scholarship_wide[2:6])
```

```
# Create attendance_rate data
# attendance data is the joined data set between scholarship and attendance
wide_scholarship <- attendance_scholarship_wide %>%
  left_join(scholarship_new, by = c("RandomID"="RandomID")) %>%
  na.omit(.)
# Add excused percentage
wide_scholarship <- wide_scholarship %>%
  mutate(excused_pct = (‘Present‘+‘Present Excused‘+‘Present Unexcused‘+‘Absent Excused‘)/totalDays)
# Categorize excused absent rate
wide_scholarship <- wide_scholarship %>%
  mutate(excused_pct_cate = if_else(excused_pct >= 0.9, ">=90%", "<90%"))

attendance_rate <- wide_scholarship
```

```
# Import GPA data
gpa <- read_excel("DR-831 GPA.xlsx", sheet = "GPA")
```

```
# Clean GPA data
gpa_no_cohort <- gpa %>% dplyr::select(-Cohort)
gpa_no_cohort <- gpa_no_cohort[!duplicated(gpa_no_cohort),]

# Create senior_gpa data that includes cumulative GPA of each student
senior_students <- (gpa_no_cohort %>%
                      group_by(RandomID) %>%
                      dplyr::summarise(n = n()) %>%
                      filter(n==4))$RandomID
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
senior_gpa <- gpa_no_cohort %>%
  filter(RandomID %in% senior_students) %>%
  group_by(RandomID) %>%
  filter(SchoolYear == max(SchoolYear))
```

Visualizations

Figure 1: Qualification for Promise Scholarship vs. Race

```r
ggplot(demo_scholarship_final,
       aes(fill=Race, x=QualifiedforCorePromise)) +
  geom_bar(position="fill") +
  ggtitle("Qualified for Promise and Race") +
  labs(y="Proportion")
```

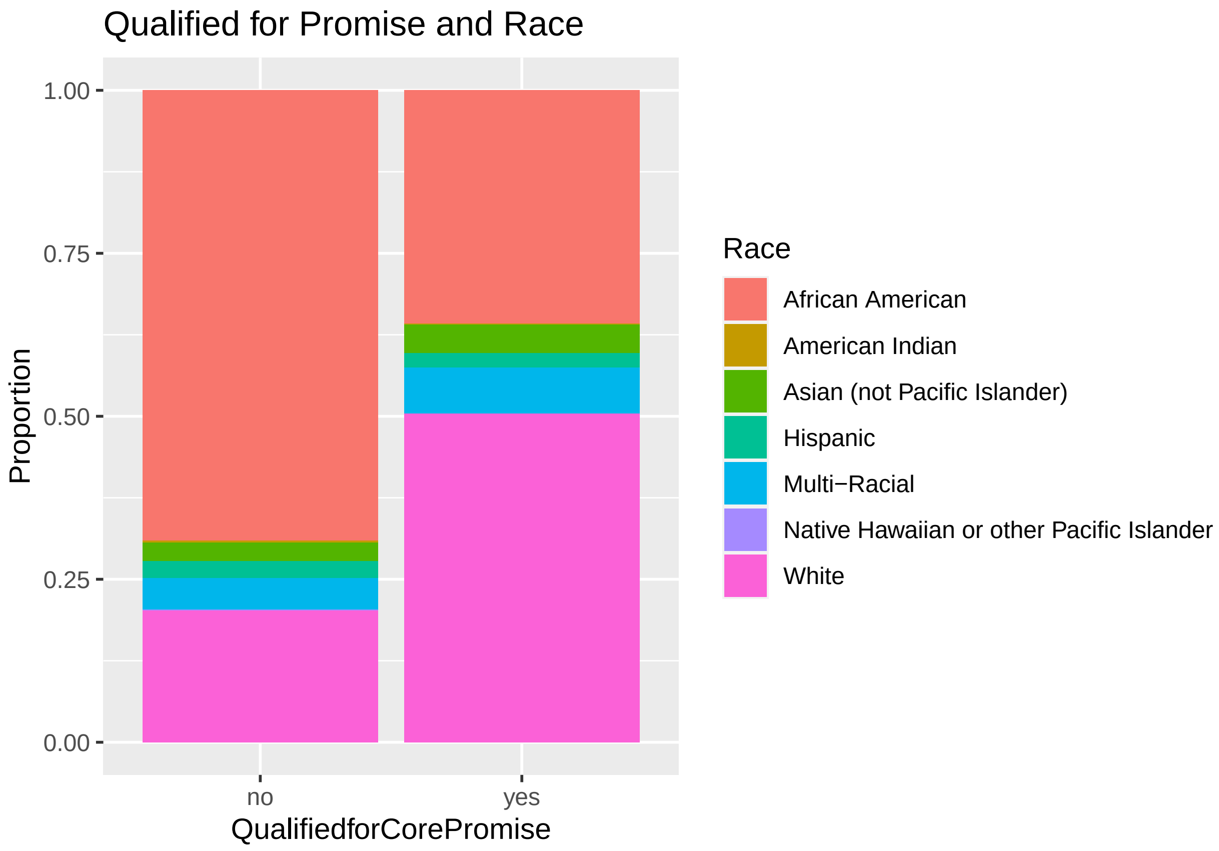


Figure 2: Receipt of Promise Scholarship vs. Race

```r
# Filter qualified students in the data set
qualified <- demo_scholarship_final %>%
  filter(QualifiedforCorePromise == "yes") #1560 observations

ggplot(qualified, aes(fill=Race, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Race") +
  labs(y="Proportion")
```

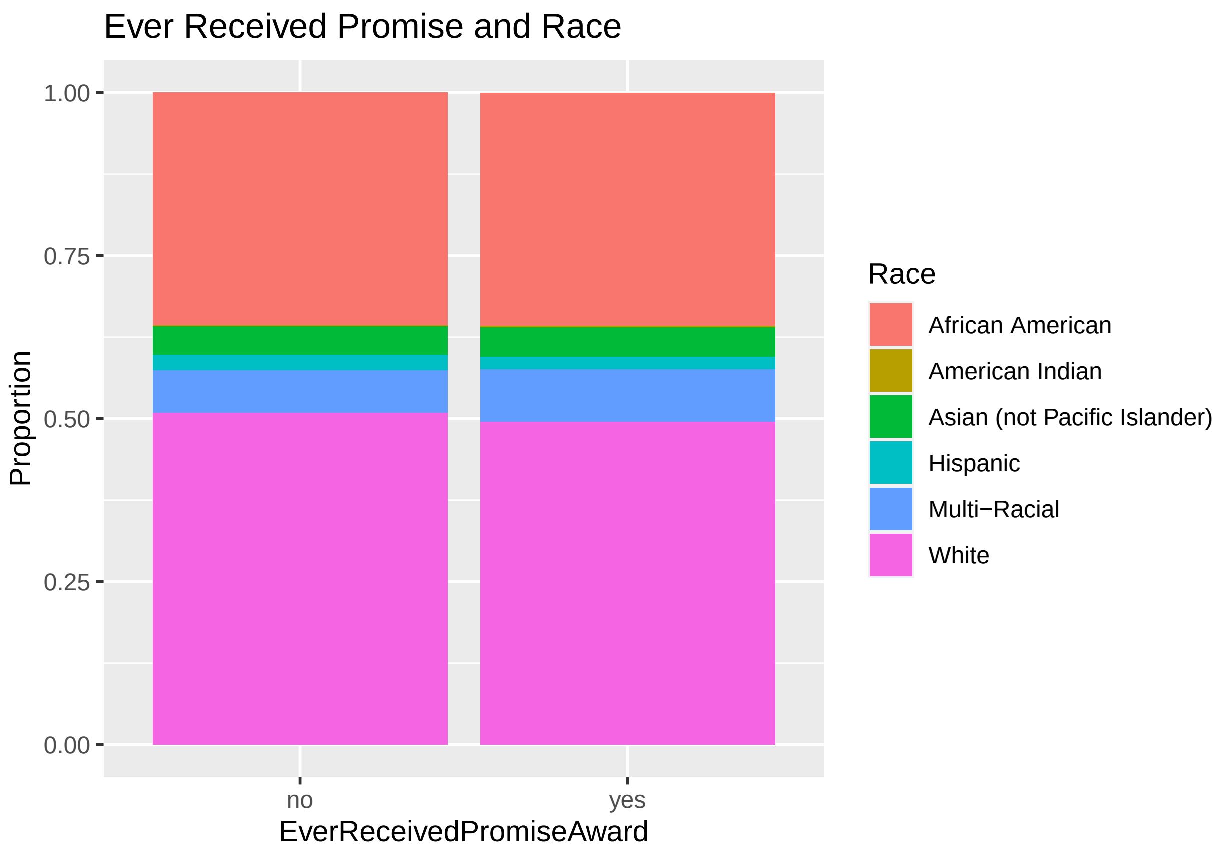


Figure 3: Qualification for Promise Scholarship vs. Gender

```
ggplot(demo_scholarship_final,
       aes(fill=Gender, x=QualifiedforCorePromise)) +
  geom_bar(position="fill") +
  ggtitle("Qualified for Promise and Gender") +
  labs(y="Proportion")
```

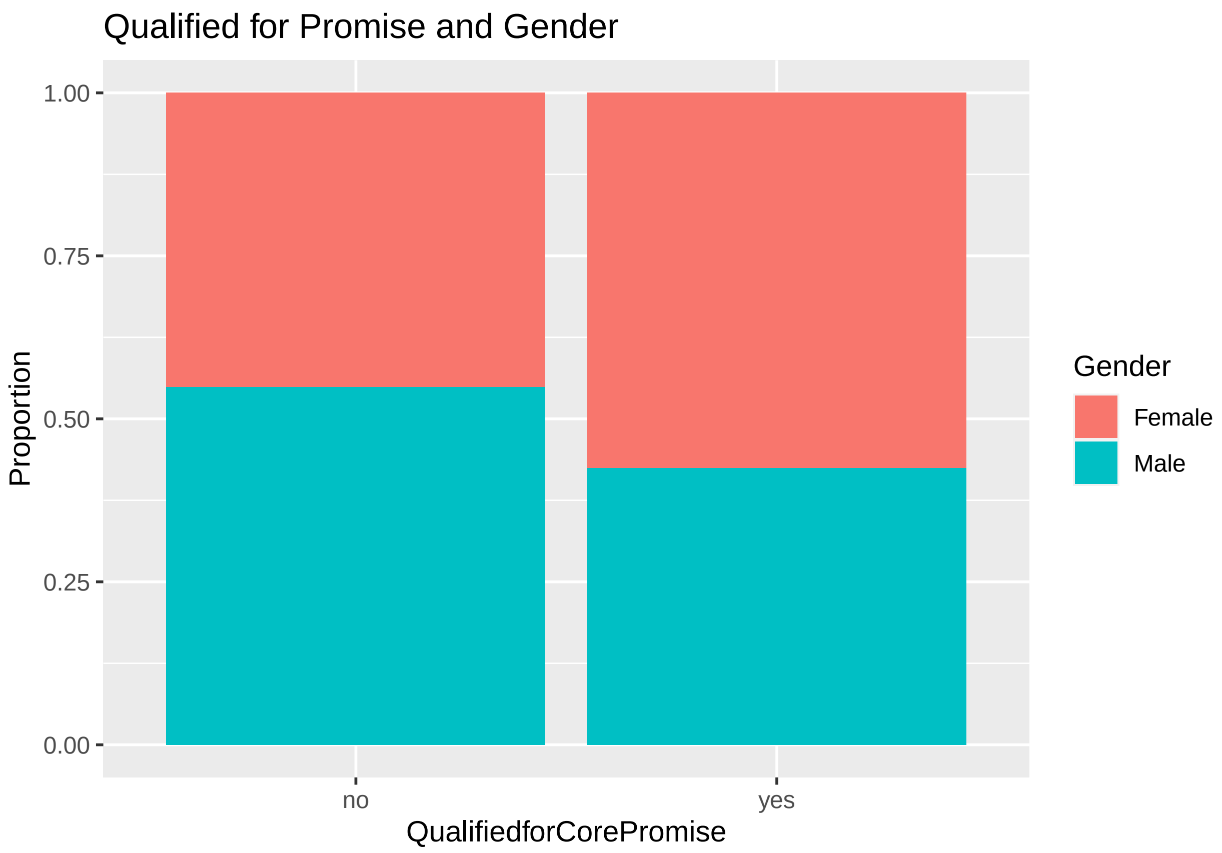


Figure 4: Receipt of Promise Scholarship vs. Gender

```
ggplot(qualified,
       aes(fill= Gender, x=EverReceivedPromiseAward)) +
  geom_bar(position="fill") +
  ggtitle("Ever Received Promise and Gender") +
  labs(y="Proportion")
```

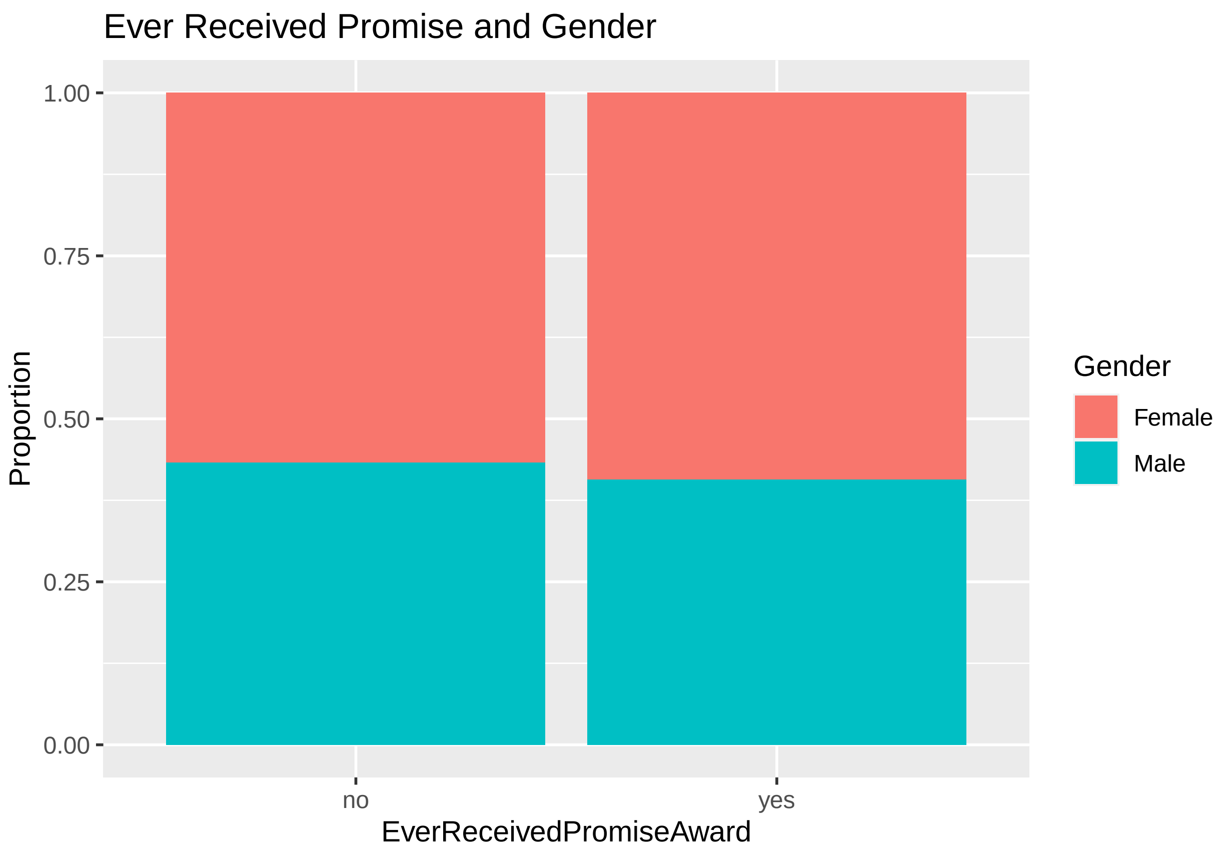


Figure 5: Receipt of Promise Scholarship vs. Career Certifications

```
ggplot(data = CTE_scholarship,
       aes(x = EverReceivedPromiseAward, y = num_cte)) +
  geom_boxplot(alpha = 0) +
  geom_jitter(alpha = 0.5, color = "green", width = 0.2, height = 0.2) +
  scale_y_continuous(breaks = c(0,2,4,6,8,10,12,14))+
  labs(title =
         "Boxplots of Number of Career Certifications Earned in terms of Whether Students Received Awar
       x = "Whether Students Received the Promise Scholarship",
       y = "Number of Certifications Earned")
```

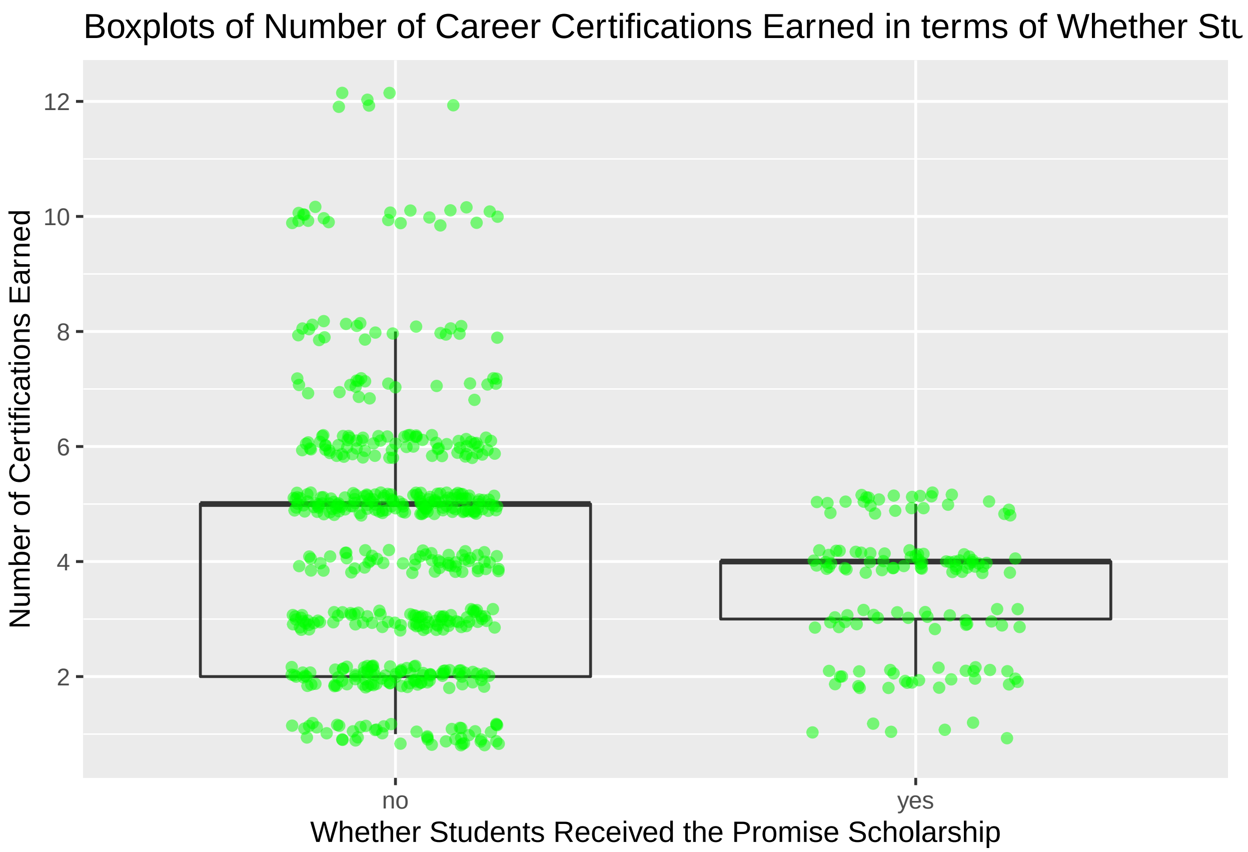


Figure 6: Receipt of Promise Scholarship vs. Economic Status

```
# Prepare data set for later visualization
join_data2c <-demo_scholarship_final %>%
  filter(QualifiedforCorePromise == "yes") %>%
  group_by(EconDisad,EverReceivedPromiseAward) %>%
  dplyr::summarise(counts = n())

ggplot(join_data2c, aes(x = EverReceivedPromiseAward, fill = EconDisad)) +
  geom_bar(position = "fill")+
  labs(title = "Barplots of Whether Received Promise under Different Economic Status",
       y = "Proportion")
```

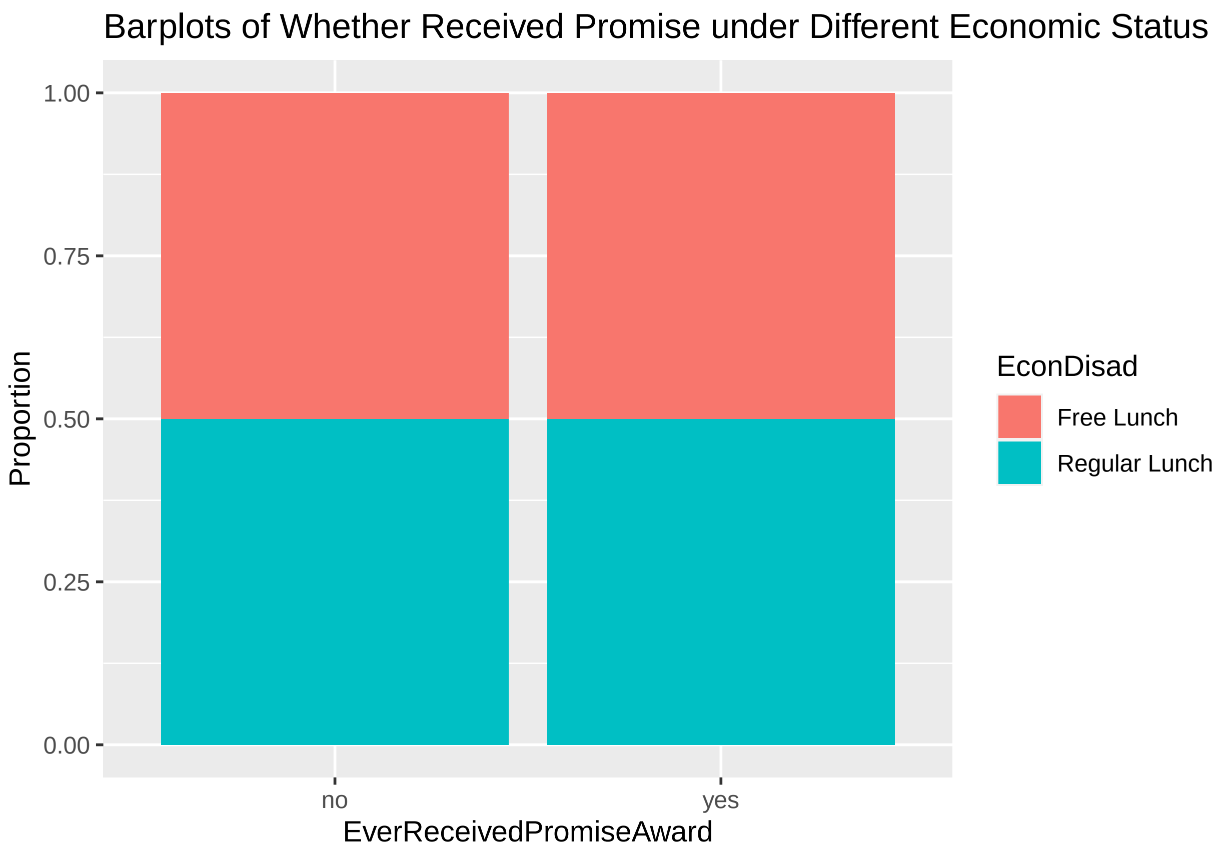


Figure 7: Cutoff vs. Qualification for Promise Scholarship

```r
# Prepare data set for later visualization
senior_gpa$RandomID <- as.character(senior_gpa$RandomID)
final_data <- attendance_rate %>%
  left_join(senior_gpa, by = c("RandomID"="RandomID"))
final_data_clean <- final_data %>%
  na.omit(final_data)

ggplot(data = final_data_clean,
       aes(x = excused_pct,
           y = CumulativeGPA,
           color = QualifiedforCorePromise)) +
  geom_point() +
  geom_hline(aes(yintercept=2.5), linetype="dashed") +
  geom_vline(aes(xintercept=0.9), linetype="dashed") +
  xlab("Attendance Rate") +
  ylab("Cumulative GPA") +
  ggtitle("GPA vs Attendance Rate") +
  guides(color = guide_legend("Qualified for Promise"))
```

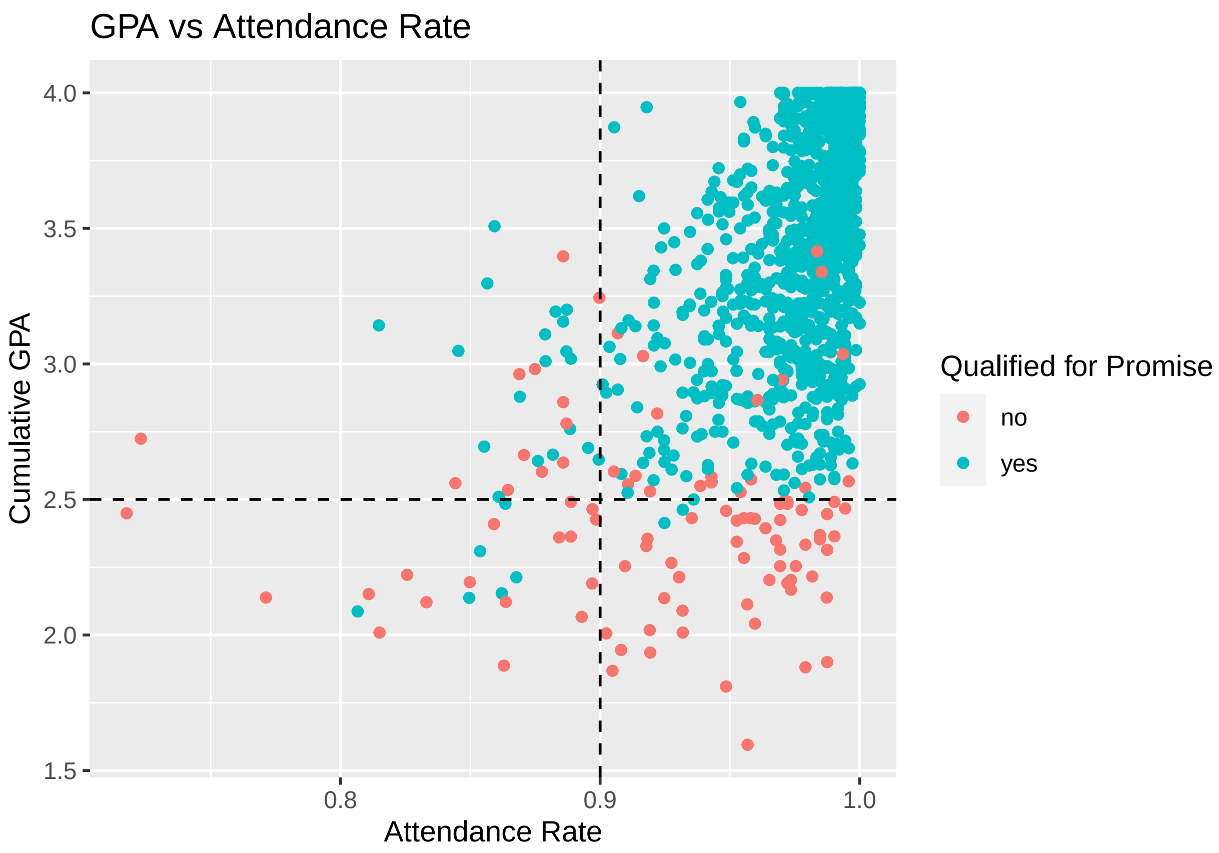
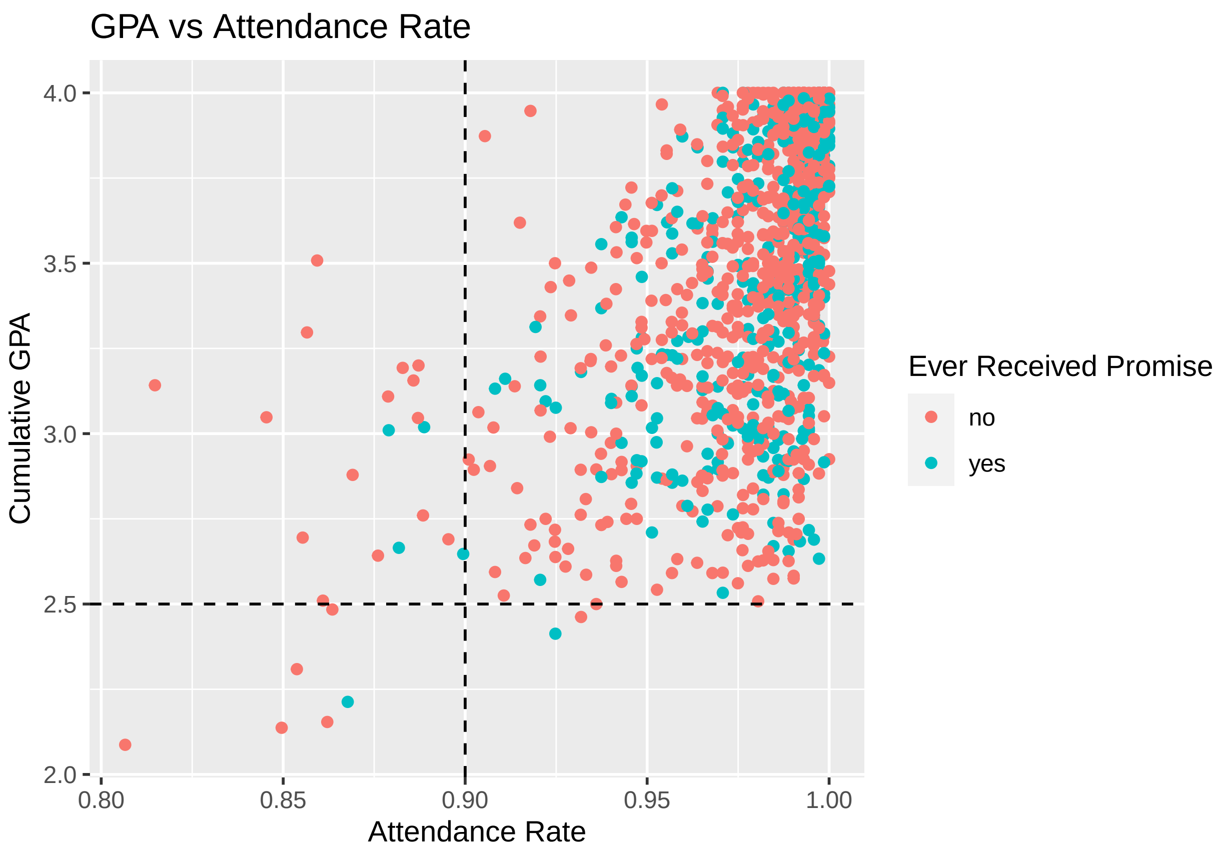


Figure 8: Cutoff vs. Receipt of Promise Scholarship

```
# Filter qualified student in the data set
final_data_clean_qualifiedYes <- final_data_clean %>%
  filter(QualifiedforCorePromise == "yes")

ggplot(data = final_data_clean_qualifiedYes,
       aes(x = excused_pct,
           y = CumulativeGPA,
           color = EverReceivedPromiseAward)) +
  geom_point() +
  geom_hline(aes(yintercept=2.5), linetype="dashed") +
  geom_vline(aes(xintercept=0.9), linetype="dashed") +
  xlab("Attendance Rate") +
  ylab("Cumulative GPA") +
  ggtitle("GPA vs Attendance Rate") +
  guides(color = guide_legend("Ever Received Promise"))
```


GPA vs Attendance Rate
Cumulative GPA
4.0
3.5
3.0
2.5
2.0
0.80
0.85
0.90
0.95
1.00
Attendance Rate
Ever Received Promise
● no
● yes

Technical Appendix B: Scholarship Analysis

Data Wrangling

```
# Library packages
library(tidyverse)
library(dplyr)
library(visdat)
library(stats)
library(ggplot2)
library(DHARMa)
library(arm)
library(readxl)
```

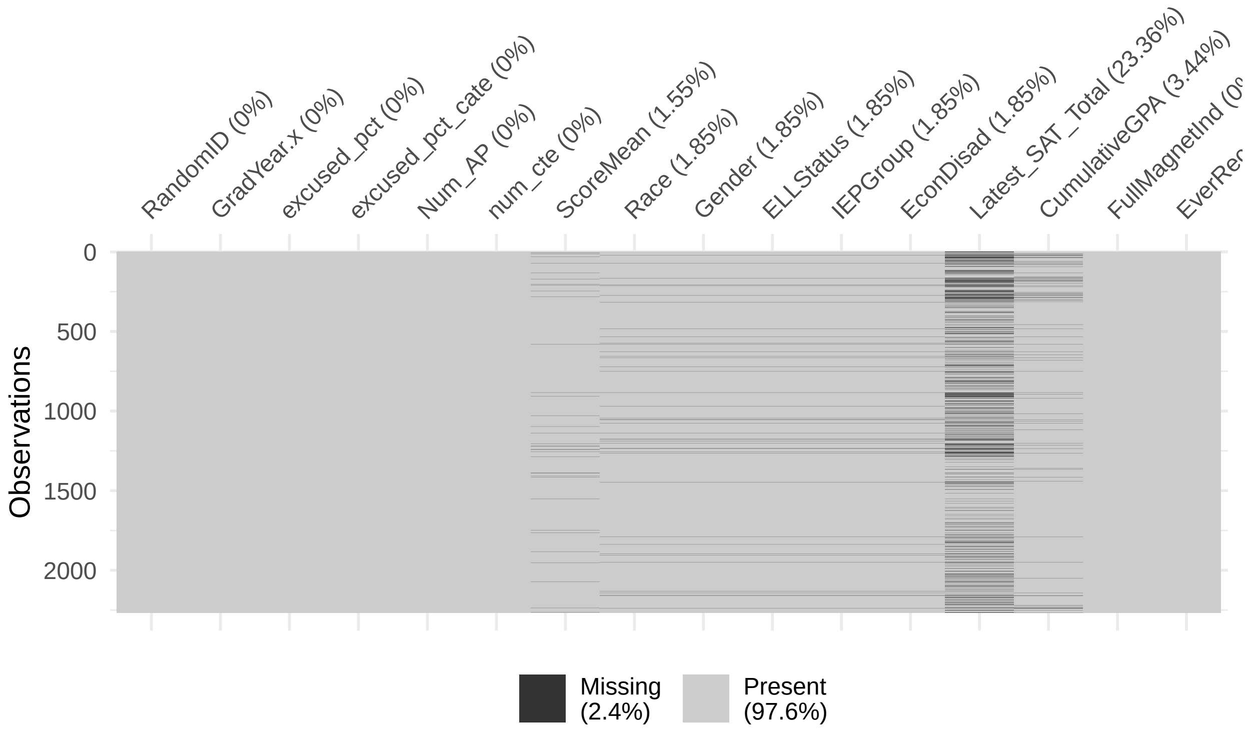
```
# Import data
scholarship <- read_excel("DR-826 Pittsburgh Promise scholarship use.xlsx", sheet = "Sheet1")
NSC <- read_excel("DR-844 NSC data_not including 1920.xlsx", sheet = "NSC data")
sat <- read_excel("DR-834 SAT scores.xlsx", sheet = "SAT Scores")
ap <- read_excel("DR-834 AP scores.xlsx", sheet = "AP Exam Scores")
keystone <- read_excel("DR-836 Keystone assessment results.xlsx", sheet = "Keystone data")
enrollment <- read_excel("DR-828 Enrollment.xlsx", sheet = "Enrollment records")
```

```
# Clean and reformat the data above
# Create new data sets for data compilation later

# Create magnet_model data
# Serve as indicator of whether students attend magnet school
enrollment_new <- enrollment %>%
  mutate_at(c(2:14), as.factor) %>%
  mutate(EntryDate=as.Date(as.character(EntryDate), tryFormats = c("%Y%m%d"))) %>%
  mutate(WithdrawDate=as.Date(as.character(WithdrawDate), tryFormats = c("%Y%m%d")))

magnet_model <- enrollment_new %>%
  filter(FullMagnetInd == "1") %>%
  distinct(RandomID, EnrolledSchoolID, SchoolReportName, FullMagnetInd)
magnet_model$RandomID <- as.character(magnet_model$RandomID)

# Create keystone_model data
# Provide information about students’ keystone scores
keystone_new <- keystone %>%
  mutate_at(c(1,3:7), as.factor) %>%
  mutate(AdminScaleScore=as.numeric(AdminScaleScore))

keystone_model <- keystone_new %>%
  na.omit() %>%
  group_by(RandomID)%>%
  dplyr::summarise(ScoreMean = mean(AdminScaleScore))
keystone_model$RandomID <- as.character(keystone_model$RandomID)

# Create sat_model data
# Provide information about each student's highest SAT score
sat_new <- sat %>%
  mutate_at(2, as.factor) %>%
  mutate(LATEST_ASSESSMENT_DATE=as.Date(LATEST_ASSESSMENT_DATE)) %>%
  mutate_at(c(4:6), as.numeric)

sat_rep <- sat_new$RandomID[duplicated(sat_new$RandomID)]

sat_model <- sat_new %>%
  distinct(RandomID, Latest_SAT_Total, .keep_all = TRUE)
sat_model$RandomID <- as.character(sat_model$RandomID)

# Create nsc_model data
# Joined data set between NSC and scholarship
nsc2 <- NSC %>%
  dplyr::select(-c("Cohort")) %>%
  mutate(GradYear = substr(as.character(HIGH_SCHOOL_GRAD_DATE),1,4)) %>%
  mutate(EnrollYear = substr(as.character(ENROLLMENT_BEGIN),1,4)) %>%
  filter(EnrollYear == GradYear)

nsc2$RandomID <- as.character(nsc2$RandomID)
scholarship$GradYear <- as.character(scholarship$GradYear)
names(scholarship)[1] <- "RandomID"

nsc_scholarship <- inner_join(nsc2, scholarship, by = c("RandomID", "GradYear"))

nsc_model <- nsc_scholarship %>%
  dplyr::select(-c("HIGH_SCHOOL_GRAD_DATE",
            "ENROLLMENT_BEGIN",
            "ENROLLMENT_END")) %>%
  distinct()
nsc_model$RandomID <- as.character(nsc_model$RandomID)
```

```
# Import data sets generated before
# Please refer to Appendix A for more details in code

# attendance_rate is the joined data set between attendance and scholarship
attendance_model <- read.csv("attendance_rate.csv")
attendance_model$RandomID <- as.character(attendance_model$RandomID)

# cte_scholarship is the joined data set between CTE and scholarship
cte_model <- read.csv("cte_scholarship.csv")
cte_model$RandomID <- as.character(cte_model$RandomID)

# senior_gpa is the cumulative GPA data generated from the original GPA data
senior_gpa <- read.csv("senior_gpa.csv")
senior_gpa$RandomID <- as.character(senior_gpa$RandomID)
```

```r
# demo_scholarship_final is the joined data set between demographics and scholarship
demographics_model <- read.csv("demo_scholarship_final.csv")
demographics_model$RandomID <- as.character(demographics_model$RandomID)

# ap_scholarship is the joined data set between ap and scholarship
ap_model <- read.csv("ap_scholarship.csv")
ap_model$RandomID <- as.character(ap_model$RandomID)
```

```r
# Data cleaning and reformatting
# Left join all data sets
data_join <- attendance_model %>%
  left_join(ap_model, by = c("RandomID"="RandomID")) %>%
  left_join(cte_model, by = c("RandomID"="RandomID")) %>%
  left_join(keystone_model, by = c("RandomID"="RandomID")) %>%
  left_join(demographics_model, by = c("RandomID"="RandomID")) %>%
  left_join(sat_model, by = c("RandomID"="RandomID")) %>%
  left_join(senior_gpa, by = c("RandomID"="RandomID")) %>%
  left_join(magnet_model, by = c("RandomID"="RandomID"))

# Select variables by column names
# Delete duplicated rows
data_variables <- data_join %>%
  dplyr::select(RandomID, GradYear.x, excused_pct, excused_pct_cate, Num_AP, num_cte,
         ScoreMean, Race, Gender, ELLStatus, IEPGroup, EconDisad, Latest_SAT_Total,
         CumulativeGPA, FullMagnetInd, EverReceivedPromiseAward.x) %>%
  distinct()

# Deal with NAs
# Replace missing num_AP, num_cte, FullMagnetInd with 0
data_na <- data_variables %>%
  replace_na(list(Num_AP=0, num_cte=0, FullMagnetInd=0))

# Visualize NAs
vis_miss(data_na)
```



```
# Omit all NAs
data_omit <- data_na %>%
  na.omit()

# Change response variable to 0,1
data_omit$EverReceivedPromiseAward.x <- as.numeric(as.factor(data_omit$EverReceivedPromiseAward.x))-1

# Change variables into appropriate format
data_everReceived <- data_omit %>%
  mutate_at(c(2, 4,8:12,15:16), as.factor)

# Rename columns
colnames(data_everReceived)<- c("RandomID","GradYear","AttendanceRate", "AttendaceRateCate",
                                "Num_AP", "Num_CTE", "KeystoneMean", "Race", "Gender",
                                "ELLStatus", "IEPGroup", "EconDisad", "SAT_Total",
                                "CumulativeGPA","MagnetInd", "EverReceivedPromiseAward")
data_everReceived$Race_new <- ifelse(data_everReceived$Race != "White" &
                                       data_everReceived$Race != "African American", "Others",
                                     as.character(data_everReceived$Race))
data_everReceived$Race_new <- as.factor(data_everReceived$Race_new)
```

Part I: Logistic Regression Analysis for All Students

```r
# Construct null model and full model
model_null <- glm(EverReceivedPromiseAward ~ Race_new+Gender, data = data_everReceived,
                  family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race_new
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_everReceived, family = "binomial")
```

```r
# Stepwise selection on AIC
# Backwards selection
backwards <- step(model_full, scope=list(lower=formula(model_null),upper=formula(model_full)),
                  direction="backward", trace = 0)

# Forwards selection
forwards <- step(model_null, scope=list(lower=formula(model_null), upper=formula(model_full)),
                 direction="forward", trace = 0)

# Selection on both Directions
bothways <- step(model_null, list(lower=formula(model_null), upper=formula(model_full)),
                 direction="both", trace=0)

# Summary of the selected models
# Choose the selected model on both directions as our final model
summary(bothways)
```

```
##
## Call:
## glm(formula = EverReceivedPromiseAward ~ Race_new + Gender +
##     CumulativeGPA + AttendanceRate + KeystoneMean + ELLStatus +
##     MagnetInd, family = "binomial", data = data_everReceived)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2897  -0.8882  -0.6592   1.2784   2.3905
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.725100   2.868124  -1.299 0.194015
## Race_newOthers       0.055956   0.188480   0.297 0.766559
## Race_newWhite       -0.266577   0.146155  -1.824 0.068163 .
## GenderMale          -0.091442   0.117153  -0.781 0.435075
## CumulativeGPA        0.903230   0.158258   5.707 1.15e-08 ***
## AttendanceRate       8.051132   2.027035   3.972 7.13e-05 ***
## KeystoneMean        -0.005891   0.001681  -3.504 0.000458 ***
## ELLStatusNot in ELL  1.100948   0.437938   2.514 0.011939 *
## MagnetInd1           0.248580   0.115826   2.146 0.031862 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2040.0  on 1707  degrees of freedom
## Residual deviance: 1919.4  on 1699  degrees of freedom
## AIC: 1937.4
##
## Number of Fisher Scoring iterations: 4
```

```
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + Race_new +
##     Gender + ELLStatus + CumulativeGPA + MagnetInd
```

```
formula(forwards)
```

```
## EverReceivedPromiseAward ~ Race_new + Gender + CumulativeGPA +
##     AttendanceRate + KeystoneMean + ELLStatus + MagnetInd
```

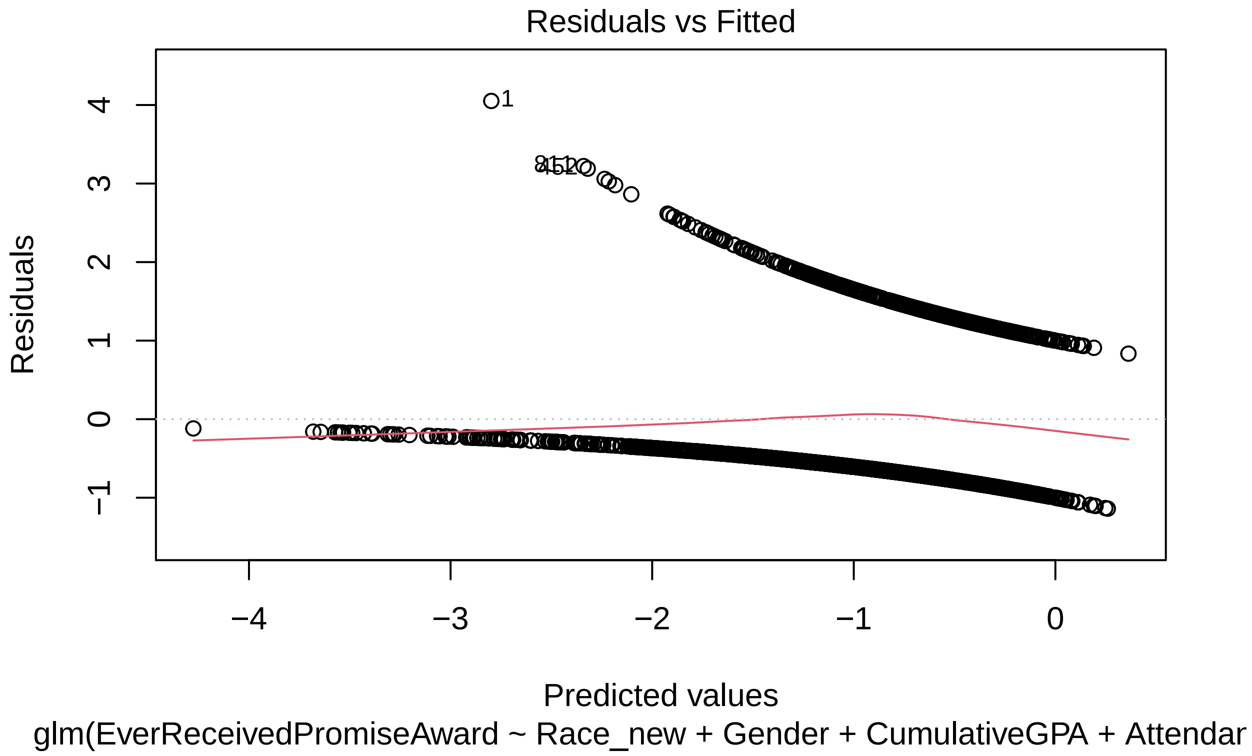
```
formula(bothways)
```

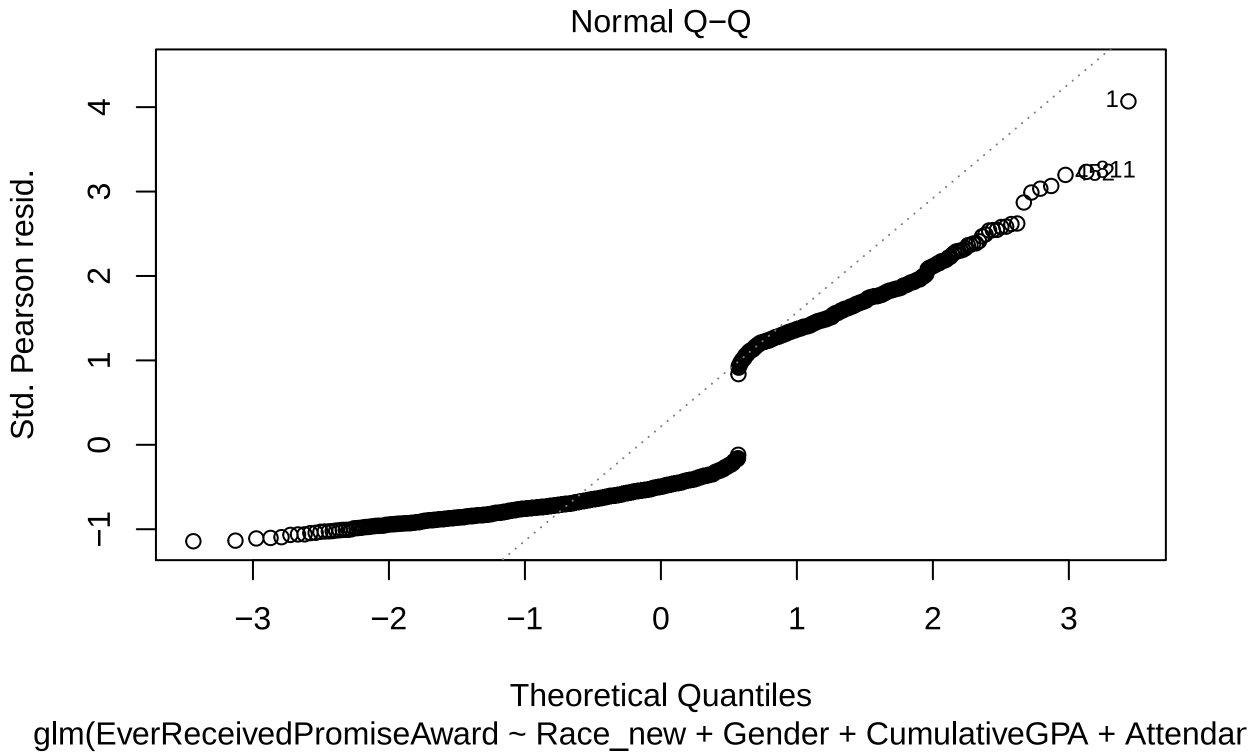
```
## EverReceivedPromiseAward ~ Race_new + Gender + CumulativeGPA +
##     AttendanceRate + KeystoneMean + ELLStatus + MagnetInd
```

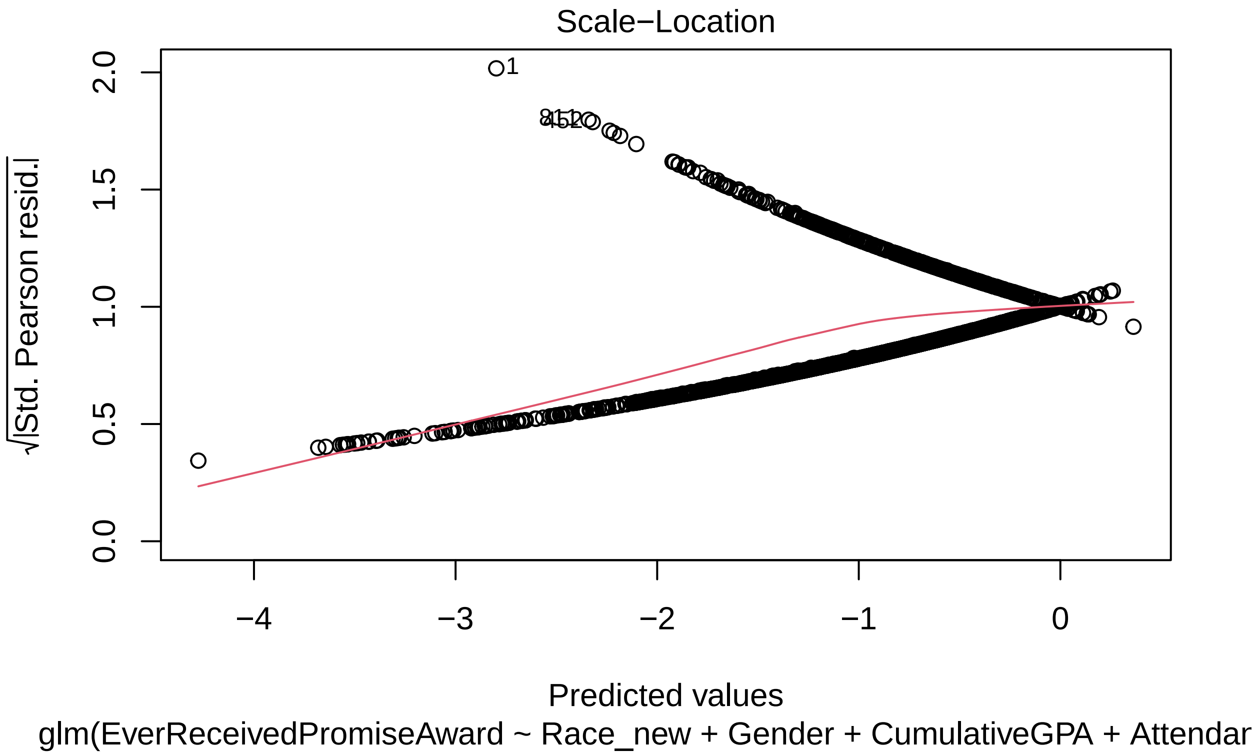
Interpretations for all significant variables:

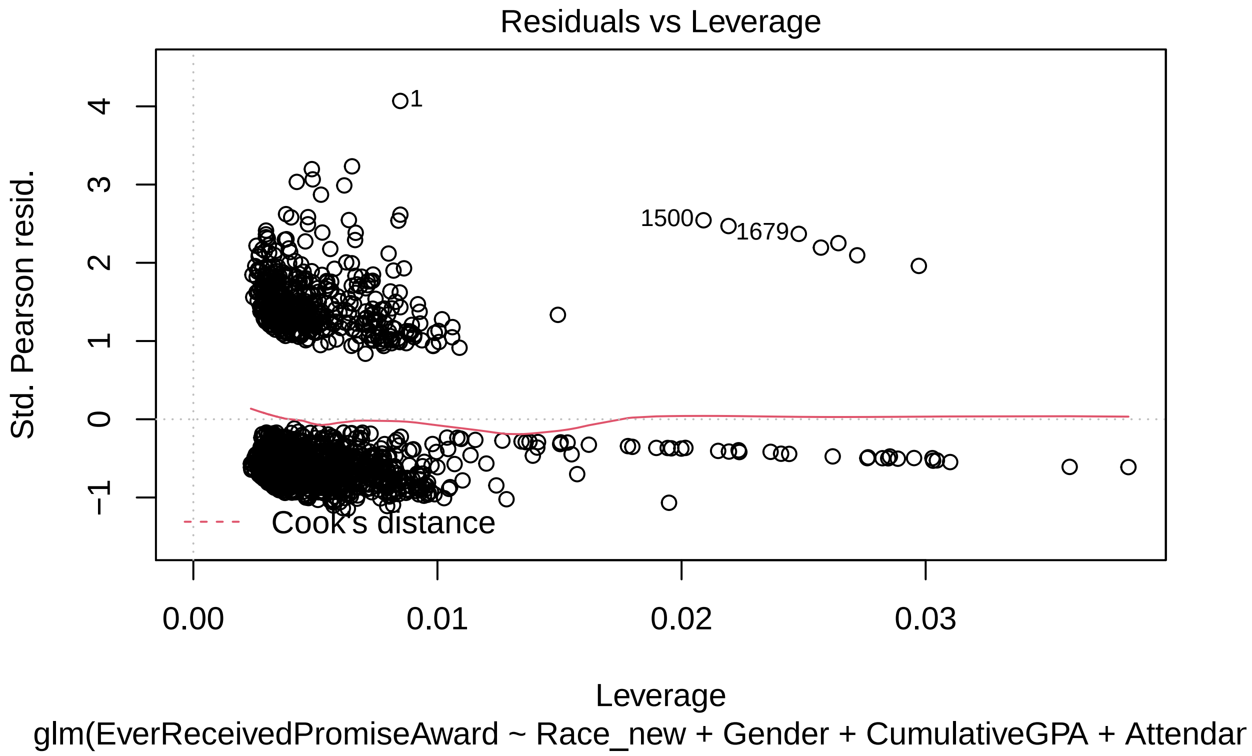
- When holding everything else fixed:
  - the odds to receive Promise for white students are 30.6% lower than black students;
  - the odds to receive Promise for student who are in magnet school are 38.3% higher than those in non-magnet school;
  - the odds for student who are not in ELL group are 200.7% higher than those in ELL group;
  - for 0.1 increase in GPA, we expect 9.4% increase in the odds of receiving Promise;
  - for 0.01 increase in attendance rate, we expect 8.4% increase in the odds of receiving Promise;
  - for 100 increase in Keystone mean, we expect 55.5% decrease in the odds of receiving Promise.

```
# Model diagnostics
plot(bothways)
```

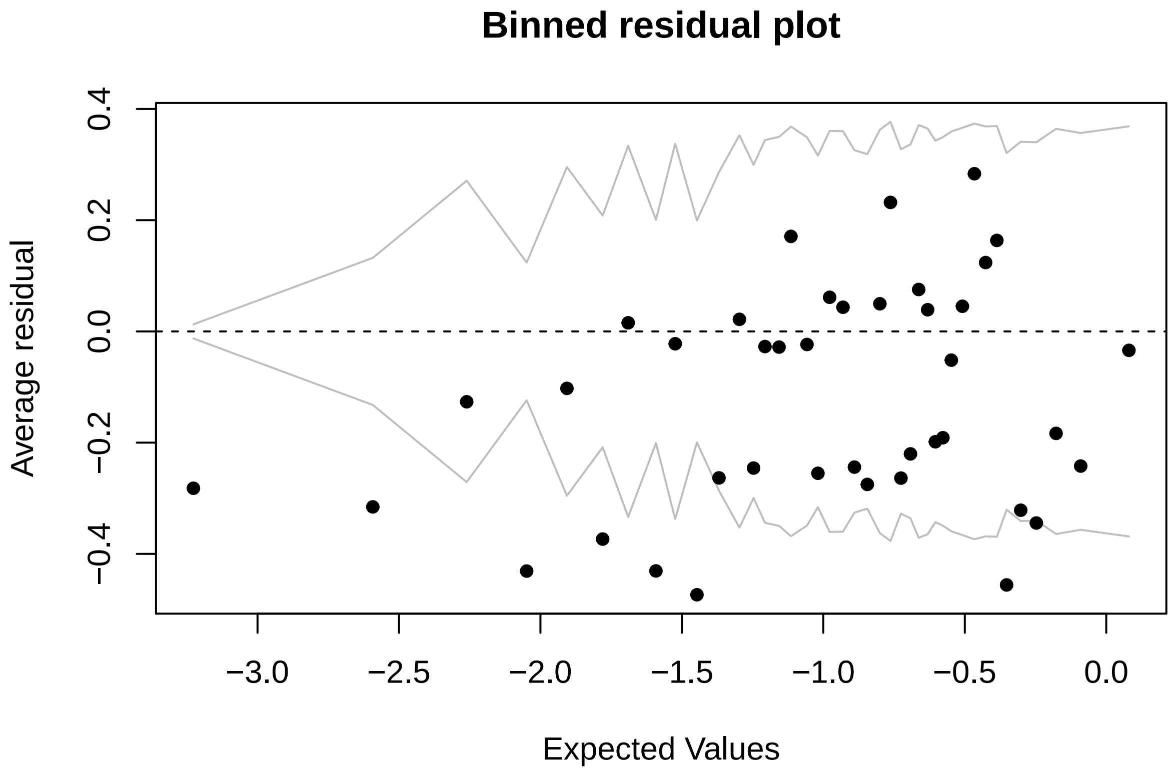

Residuals vs Fitted
Residuals
4
3
2
1
0
−1
1
811
452
−4
−3
−2
−1
0
Predicted values
glm(EverReceivedPromiseAward ~ Race_new + Gender + CumulativeGPA + Attendar


Normal Q−Q
Std. Pearson resid.
4
3
2
1
0
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles
1
452
811
glm(EverReceivedPromiseAward ~ Race_new + Gender + CumulativeGPA + Attendar

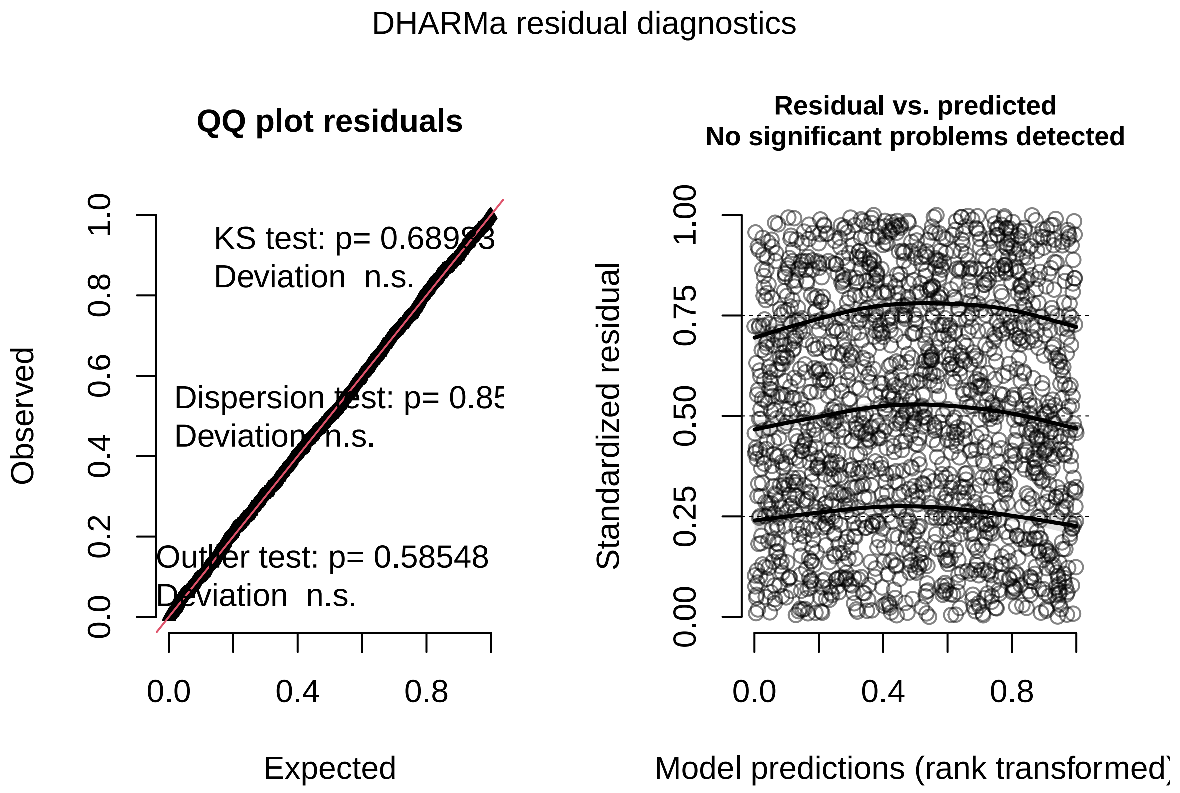

Scale−Location
2.0
1.5
1.0
0.5
0.0
√|Std. Pearson resid.|
1
811
452
−4
−3
−2
−1
0
Predicted values
glm(EverReceivedPromiseAward ~ Race_new + Gender + CumulativeGPA + Attendar



```r
y <- resid(bothways)
x <- predict(bothways)
binnedplot(x,y)
```



```
sim <- simulateResiduals(bothways)
plot(sim)
```



Conclusion of model diagnostics:

- From the binned plot, we know that the selected model is a good fit.
  - gray lines refer to the 95% confidence interval;
  - most points fall within the confidence interval;
  - residuals are randomly scattered with no obvious pattern.
- From QQ residuals plot, the residuals follow normal distribution.
- From residuals vs predicted plot, variance assumption of residuals is not violated by the selected model.

Part II: Logistic Regression Analysis for Qualified Students

```
# Prepare data sets for the analysis on only qualified students
data_qualified <- merge(data_everReceived,
                        attendance_model[, c("RandomID", "QualifiedforCorePromise")], by="RandomID")

data_qualifiedYes <- data_qualified %>%
  filter(QualifiedforCorePromise == "yes")

data_qualifiedYes$Race_new <- ifelse(data_qualifiedYes$Race != "White" & data_qualifiedYes$Race
                                     != "African American", "Other",
                                     as.character(data_qualifiedYes$Race))
data_qualifiedYes$Race_new <- as.factor(data_qualifiedYes$Race_new)
data_qualifiedYes$QualifiedforCorePromise <- as.factor(data_qualifiedYes$QualifiedforCorePromise)
```

```
# Construct null model and full model
model_null <- glm(EverReceivedPromiseAward ~ Race_new+Gender,
                  data = data_qualifiedYes, family = "binomial")
model_full <- glm(EverReceivedPromiseAward ~ AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race_new
              +Gender+ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd,
              data = data_qualifiedYes, family = "binomial")
```

```
# Stepwise selection on AIC
# Backwards selection
backwards <- step(model_full, scope=list(lower=formula(model_null), upper=formula(model_full)),
                  direction="backward", trace = 0)

# Forward selection
forwards <- step(model_null, scope=list(lower=formula(model_null),upper=formula(model_full)),
                 direction="forward", trace = 0)

# Selection on both directions
bothways <- step(model_null, list(lower=formula(model_null),upper=formula(model_full)),
                 direction="both", trace=0)

# Summary of selected model
# Choose the selected model on both directions as our final model
summary(bothways)
```

```
##
## Call:
## glm(formula = EverReceivedPromiseAward ~ Race_new + Gender +
##     AttendanceRate + MagnetInd + KeystoneMean + CumulativeGPA +
##     ELLStatus, family = "binomial", data = data_qualifiedYes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2264  -0.9362  -0.7954   1.3551   2.0967
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.916612   3.335304  -1.474  0.14045
## Race_newOther        0.106687   0.197056   0.541  0.58823
## Race_newWhite       -0.204815   0.154436  -1.326  0.18477
## GenderMale          -0.022979   0.123329  -0.186  0.85219
## AttendanceRate      10.401616   2.575524   4.039 5.38e-05 ***
## MagnetInd1           0.225356   0.122491   1.840  0.06580 .
## KeystoneMean        -0.005469   0.001796  -3.044  0.00233 **
## CumulativeGPA        0.475464   0.195891   2.427  0.01522 *
## ELLStatusNot in ELL  0.783758   0.462010   1.696  0.08981 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1722.8  on 1356  degrees of freedom
## Residual deviance: 1675.7  on 1348  degrees of freedom
## AIC: 1693.7
##
## Number of Fisher Scoring iterations: 4
```

```
formula(backwards)
```

```
## EverReceivedPromiseAward ~ AttendanceRate + KeystoneMean + Race_new +
##     Gender + ELLStatus + CumulativeGPA + MagnetInd
```

```
formula(forwards)
```

```
## EverReceivedPromiseAward ~ Race_new + Gender + AttendanceRate +
##     MagnetInd + KeystoneMean + CumulativeGPA + ELLStatus
```

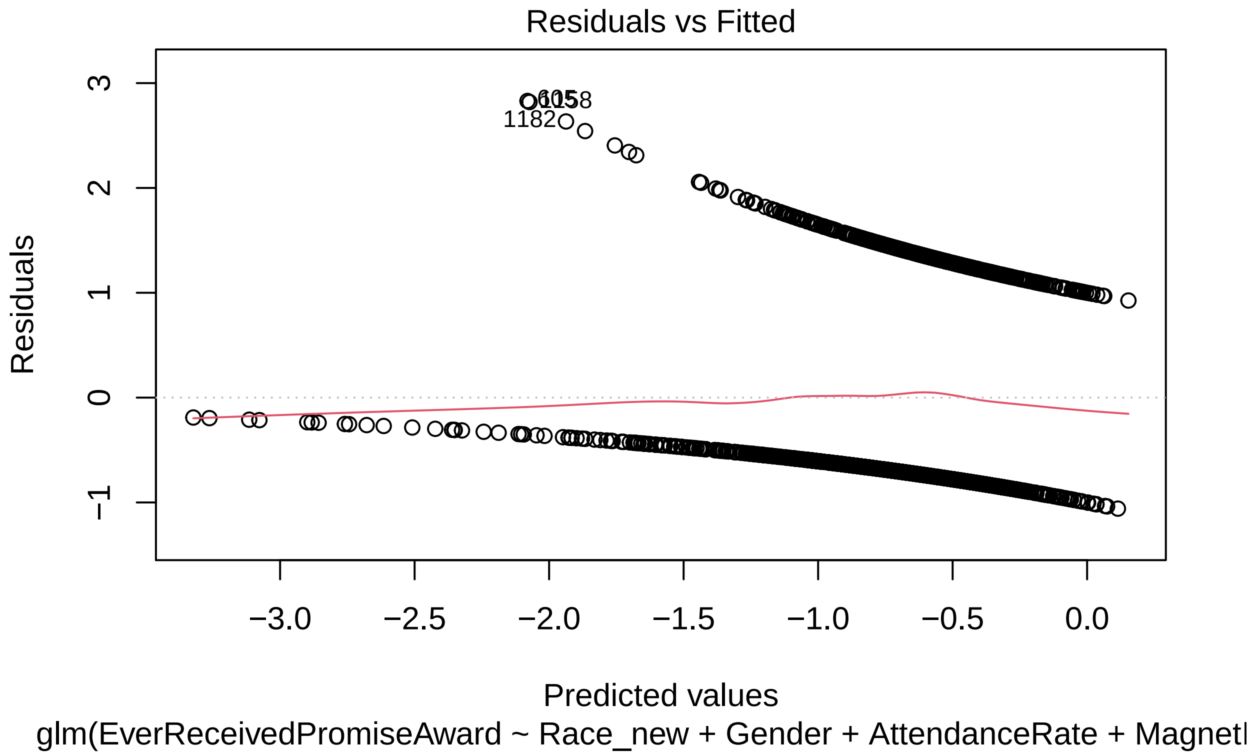
```
formula(bothways)
```

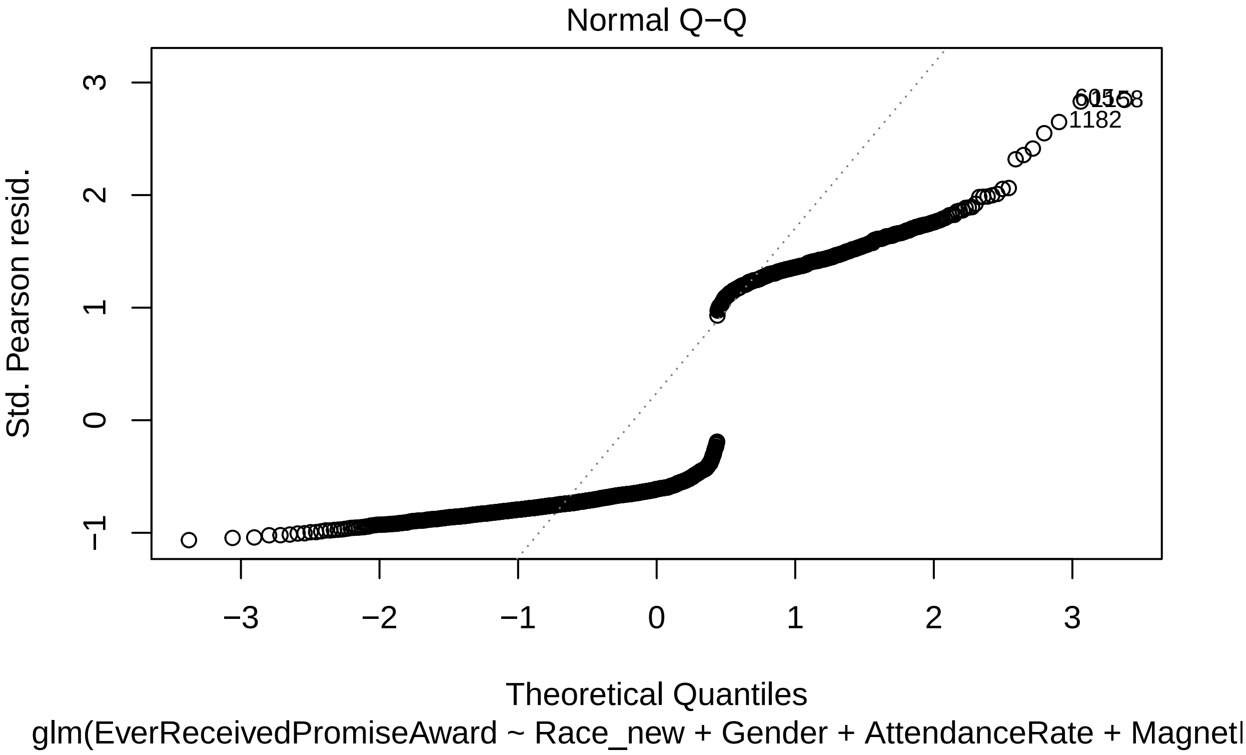
```
## EverReceivedPromiseAward ~ Race_new + Gender + AttendanceRate +
##     MagnetInd + KeystoneMean + CumulativeGPA + ELLStatus
```

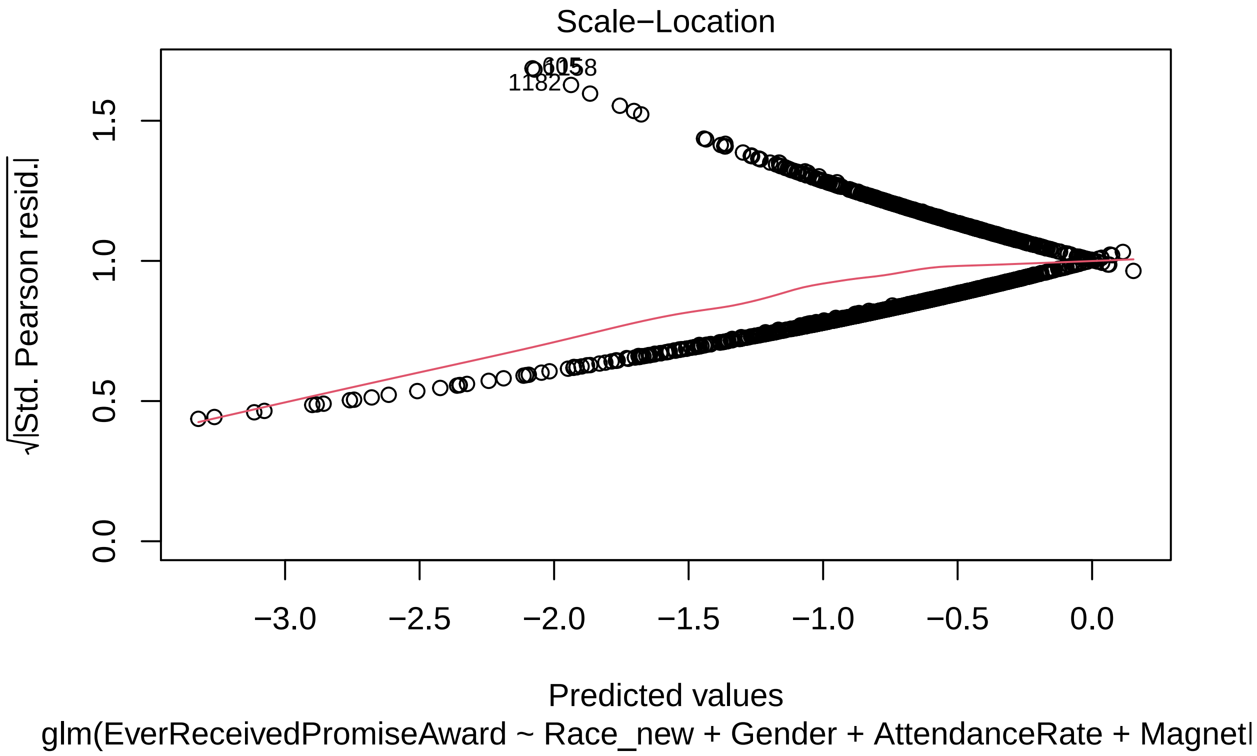
Interpretation for all significant variables:

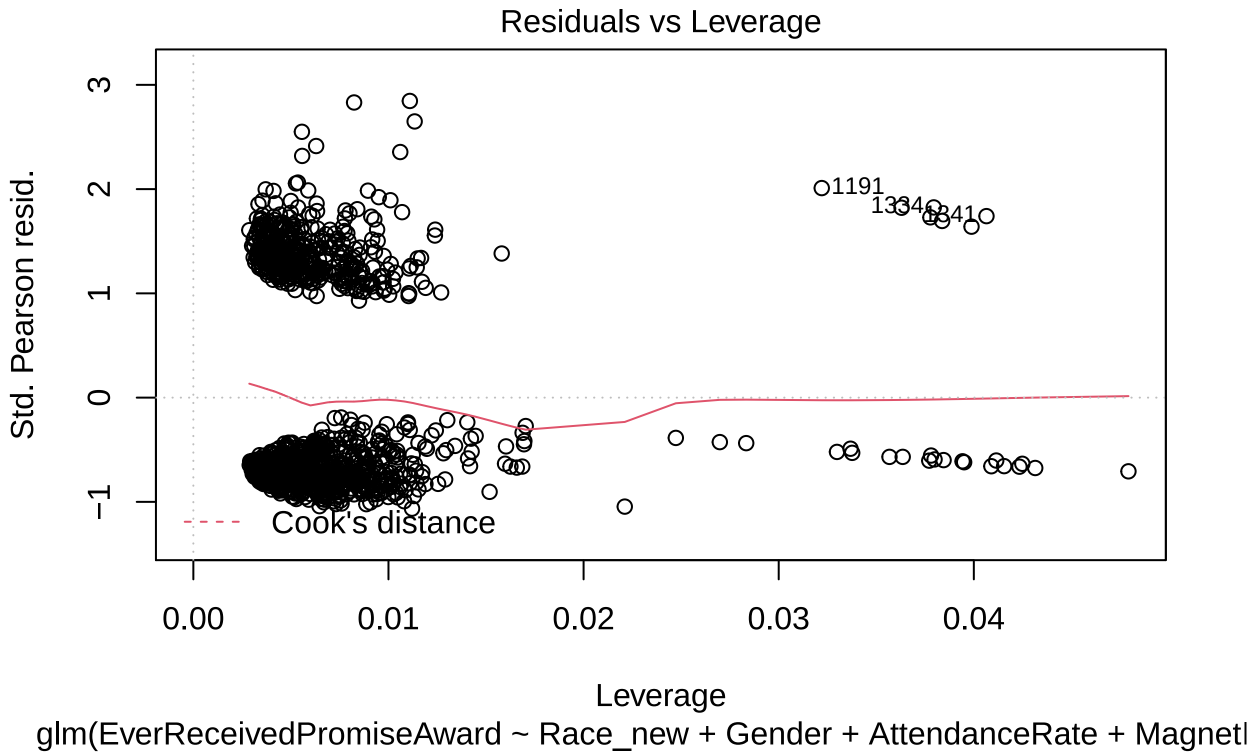
- When holding everything else fixed:
  - the odds to receive Promise for student who are in magnet school are 25.3% higher than those in non-magnet school;
  - the odds for student who are not in ELL group are 119% higher than those in ELL group;
  - for 0.1 increase in GPA, we expect 4.9% increase in the odds of receiving Promise;
  - for 0.01 increase in attendance rate, we expect 11% increase in the odds of receiving Promise;
  - for 100 increase in Keystone mean, we expect 57.9% decrease in the odds of receiving Promise.

```
# Model diagnostics
plot(bothways)
```

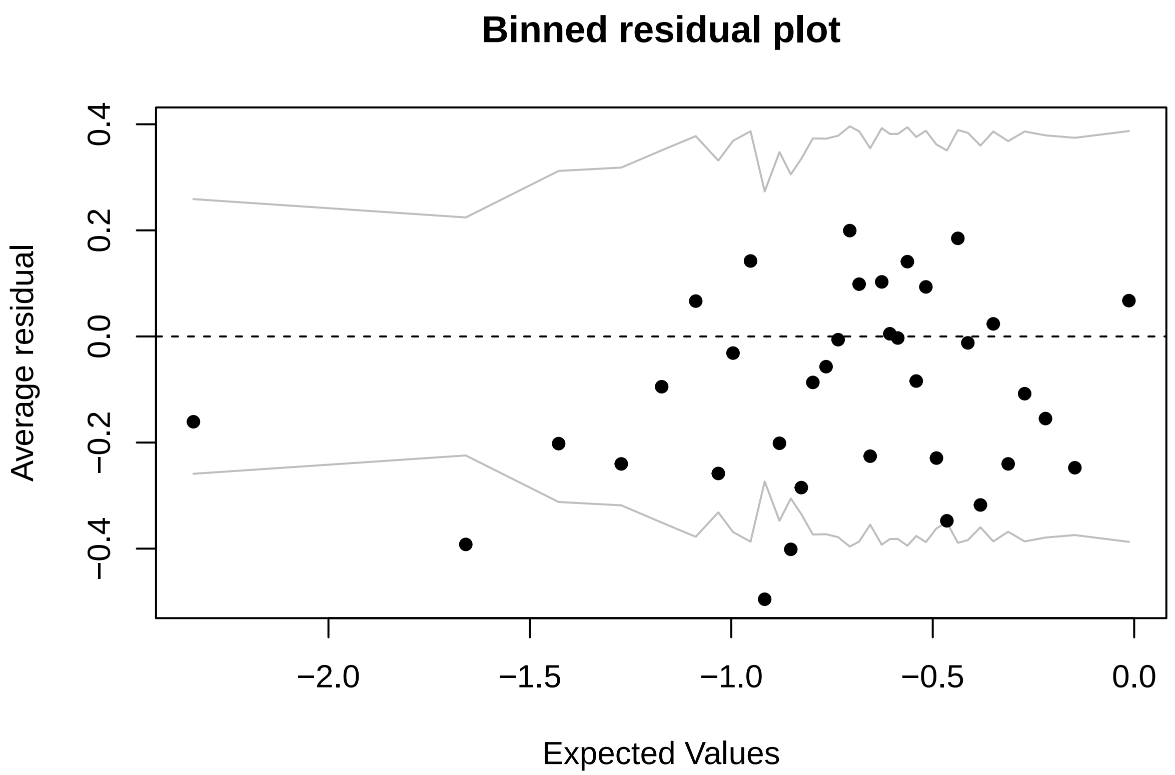

Residuals vs Fitted
605
1158
1182
3
2
1
0
−1
Residuals
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
Predicted values
glm(EverReceivedPromiseAward ~ Race_new + Gender + AttendanceRate + MagnetI


Normal Q−Q
Std. Pearson resid.
3
2
1
0
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(EverReceivedPromiseAward ~ Race_new + Gender + AttendanceRate + MagnetI
6055
6158
1182

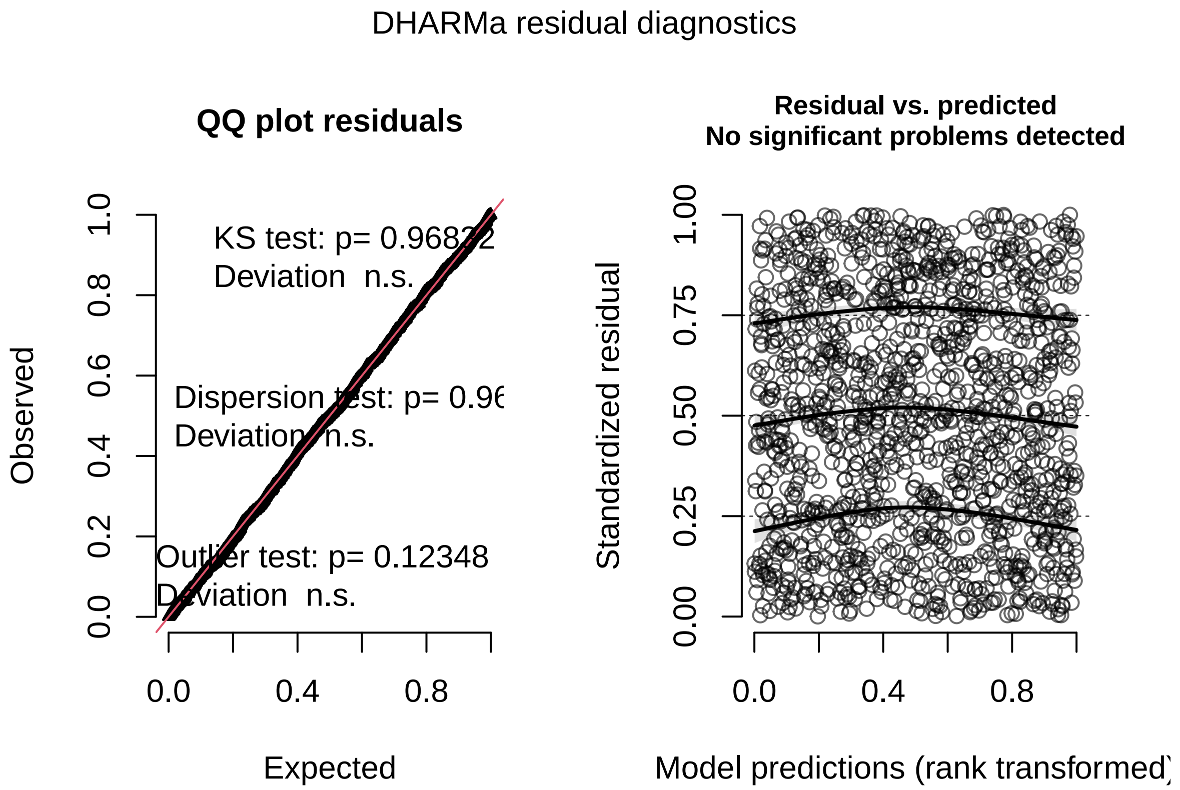

Scale−Location
605
1058
1182
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
Predicted values
glm(EverReceivedPromiseAward ~ Race_new + Gender + AttendanceRate + MagnetI



```r
y <- resid(bothways)
x <- predict(bothways)
binnedplot(x,y)
```



```
sim <- simulateResiduals(bothways)
plot(sim)
```



Conclusion of model diagnostics:

- From the binned plot, we know that the selected model is a good fit.
  - gray lines refer to the 95% confidence interval;
  - most points fall within the confidence interval;
  - residuals are randomly scattered with no obvious pattern.
- From QQ residuals plot, the residuals follow normal distribution.
- From residuals vs predicted plot, variance assumption of residuals is not violated by the selected model.

Technical Appendix C: Retention Analysis

Part I: Exploratory Data Analysis & T-tests

Data Wrangling

```r
# Import packages
library(plyr)
library(dplyr)
library(lubridate)
library(tidyverse)
library(readxl)
library(ggplot2)
library(arm)
```

```r
# Import scholarship data
scholarship <- read_excel("DR-826 Pittsburgh Promise scholarship use.xlsx", sheet = "Sheet1")

# Clean scholarship data
scholarship <- scholarship %>%
  mutate(QualifiedforCorePromise = ifelse(QualifiedforCorePromise == "yes",1,0),
         QualifiedforExtensionPromise = ifelse(QualifiedforExtensionPromise == "yes",1,0),
         EverReceivedPromiseAward = ifelse(EverReceivedPromiseAward == "yes",1,0),
         StillReceivingAward = ifelse(StillReceivingAward == "yes",1,0),
         StillEligible = ifelse(StillEligible == "yes",1,0),
         HighSchool = as.factor(HighSchool)) %>%
  rename(RandomID = "Random ID")
head(scholarship)
```

```
## # A tibble: 6 x 8
##   RandomID GradYear QualifiedforCor~ QualifiedforExt~ EverReceivedPro~
##   <chr>       <dbl>            <dbl>            <dbl>            <dbl>
## 1 5829765      2018                0                0                0
## 2 5832055      2018                0                0                0
## 3 5833420      2018                1                0                0
## 4 5840516      2018                0                0                0
## 5 5841218      2018                0                0                1
## 6 5847024      2018                1                0                0
## # ... with 3 more variables: StillReceivingAward <dbl>, StillEligible <dbl>,
## #   HighSchool <fct>
```

```r
# Import NSC data
nsc <- read_excel("DR-844 NSC data_not including 1920.xlsx", sheet = "NSC data")
nsc$RandomID <- as.character(nsc$RandomID)
head(nsc)
```

```
## # A tibble: 6 x 11
##   RandomID Cohort HIGH_SCHOOL_GRA~ COLLEGE_STATE ‘2-YEAR_4-YEAR‘ PUBLIC_PRIVATE
##   <chr>     <dbl>            <dbl> <chr>         <chr>           <chr>
## 1 5841218    1415         20180608 PA            4-year          Public
## 2 5841218    1415         20180608 PA            4-year          Public
## 3 5847024    1415         20180608 NY            4-year          Private
## 4 5847024    1415         20180608 NY            4-year          Private
## 5 5847024    1415         20180608 NY            4-year          Private
## 6 5847024    1415         20180608 NY            4-year          Private
## # ... with 5 more variables: ENROLLMENT_BEGIN <dbl>, ENROLLMENT_END <dbl>,
## #   ENROLLMENT_STATUS <chr>, GRADUATED <chr>, GRADUATION_DATE <dbl>
```

```
# Import demographics data
demo1415 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1415")
demo1516 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1516")
demo1617 <- read_excel("DR-827 Demographics.xlsx", sheet = "Demographics for cohort 1617")

# Clean demographics data
demographics <- rbind(demo1415,demo1516,demo1617)
demographics <- demographics %>%
  dplyr::select(RandomID, Race) %>%
  distinct()

# Delete students who have more than 1 races
demographics <- demographics %>%
  filter(! RandomID %in% c(6008133,6040930,
                           6126484,6137723,
                           6207255,6213261,
                           6228923,6510341))
```

Comparison I: Retention vs. Scholarship Receipt

EDA

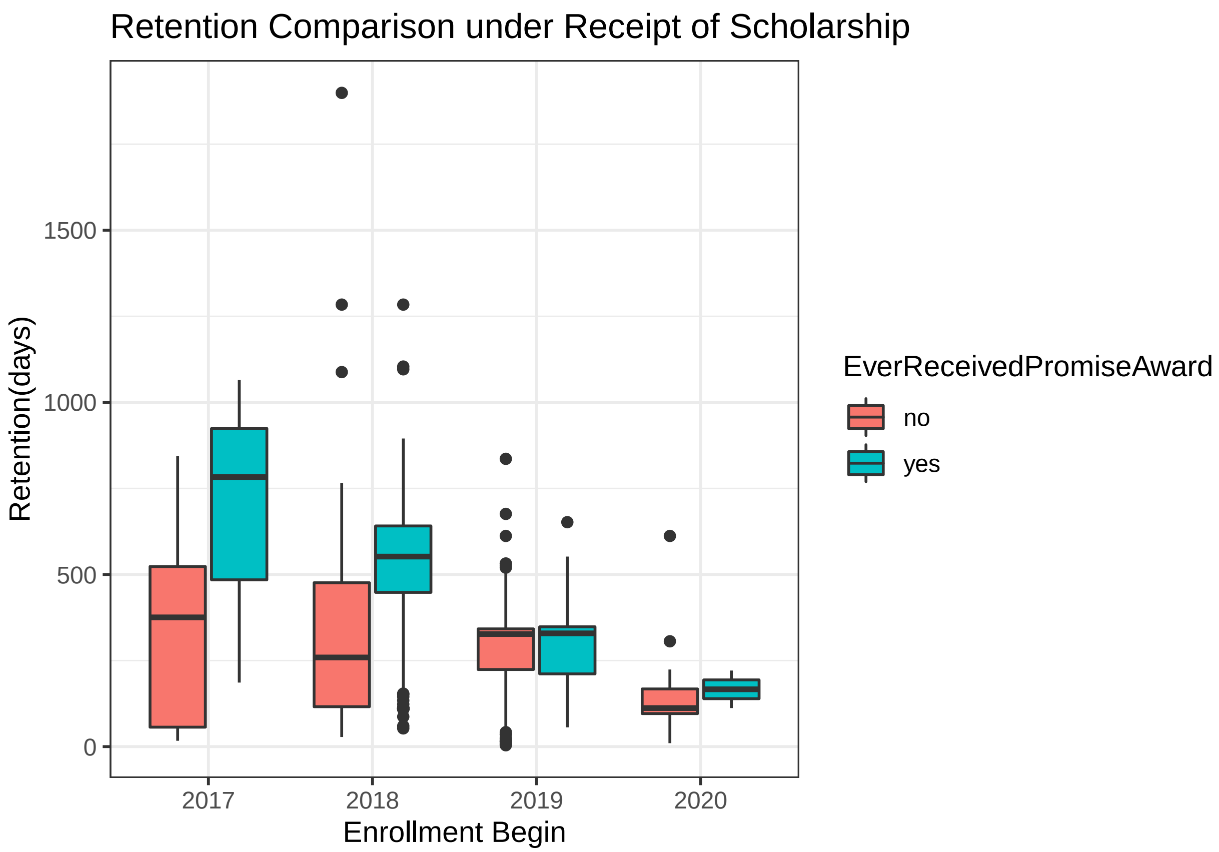
```
# Prepare data sets for the visualization of comparison plot
nsc_processed = nsc %>%
  filter(ENROLLMENT_BEGIN >0) %>%
  filter(ENROLLMENT_BEGIN %/% 10000 != 2016) %>%
  filter(COLLEGE_STATE == "PA") %>%
  mutate(ENROLLMENT_BEGIN = as.Date(paste(substr(ENROLLMENT_BEGIN, 1, 4),
                                          substr(ENROLLMENT_BEGIN, 5, 6),
                                          substr(ENROLLMENT_BEGIN, 7, 8),
                                          sep = "-")),
         ENROLLMENT_END = as.Date(paste(substr(ENROLLMENT_END, 1, 4),
                                        substr(ENROLLMENT_END, 5, 6),
                                        substr(ENROLLMENT_END, 7, 8),
                                        sep = "-")),
         elapsed = ENROLLMENT_END - ENROLLMENT_BEGIN)
```

```
# Issue: some students enroll in fall while others enroll in spring
# Create a new variable semester to indicate which semester students enroll in
```

```
nsc_retention = nsc_processed %>%
  group_by(RandomID) %>%
  summarise(retention = sum(elapsed),
            ENROLLMENT_BEGIN_year = year(min(ENROLLMENT_BEGIN)),
            semester = ifelse(month(min(ENROLLMENT_BEGIN))>6, "Fall","Spring"))
```

```
# Prepare data sets for the visualization of comparison plot
nsc_promise = nsc_retention %>%
  left_join(scholarship, by = "RandomID") %>%
  mutate(EverReceivedPromiseAward =
           ifelse(EverReceivedPromiseAward==1,"yes","no"),
         EverReceivedPromiseAward = replace_na(EverReceivedPromiseAward,"no"),
         EverReceivedPromiseAward = as.factor(EverReceivedPromiseAward),
         retention = as.numeric(retention)) %>%
  dplyr::select(-c(QualifiedforCorePromise,
            QualifiedforExtensionPromise,
            StillReceivingAward,
            StillEligible,HighSchool))
```

```
# Visualization of retention comparison under receipt of scholarship
ggplot(nsc_promise, aes(x=as.factor(ENROLLMENT_BEGIN_year),
                        y=retention,
                        fill=EverReceivedPromiseAward)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin",
       y = "Retention(days)",
       title = "Retention Comparison under Receipt of Scholarship") +
  theme_bw()
```



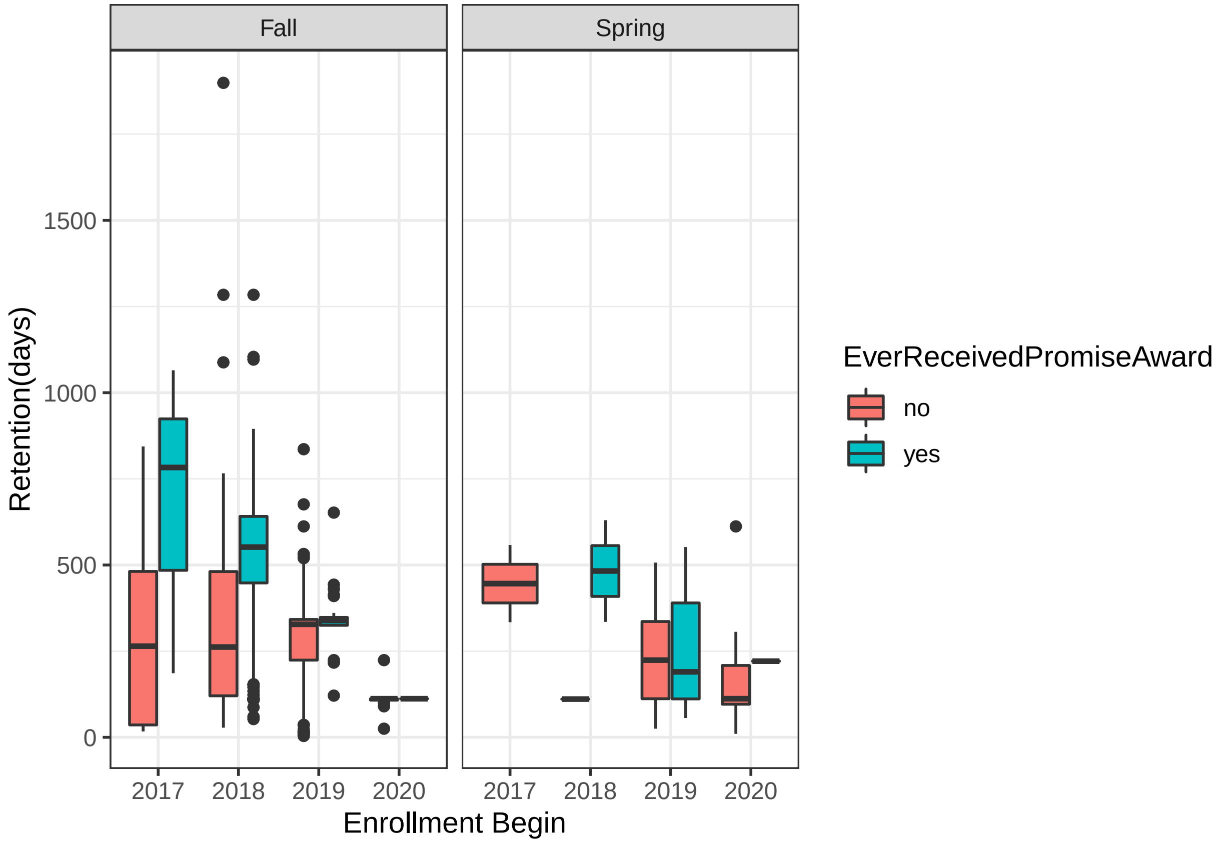
Findings:

- We see that students who received scholarship tend to have better retention except for year 2019.
- However, notice that we only have 13 observations for year 2017, and 93 observations for year 2020.

```
# Take a closer look by semester they enroll
nsc_promise %>%
  group_by(ENROLLMENT_BEGIN_year, semester) %>%
  count()
```

```
## # A tibble: 8 x 3
## # Groups:   ENROLLMENT_BEGIN_year, semester [8]
##   ENROLLMENT_BEGIN_year semester     n
##                   <dbl> <chr>    <int>
## 1                  2017 Fall        11
## 2                  2017 Spring       2
## 3                  2018 Fall       571
## 4                  2018 Spring       3
## 5                  2019 Fall       639
## 6                  2019 Spring      59
## 7                  2020 Fall        28
## 8                  2020 Spring      65
```

```
# Visualization of retention comparison under receipt of scholarship (by semester)
# Since we have limited observations for 2017 and all spring semesters,
# the resulting plot is just for reference
ggplot(nsc_promise, aes(x=as.factor(ENROLLMENT_BEGIN_year),
                        y=retention,
                        fill=EverReceivedPromiseAward)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin",
       y = "Retention(days)") +
  theme_bw() +
  facet_wrap(~semester)
```



Check for Significance for 2018 & 2019 (T-tests)

```
# Prepare data sets for 2018 & 2019
nsc_promise_2018 = nsc_promise %>% filter(ENROLLMENT_BEGIN_year == 2018)
nsc_promise_2019 = nsc_promise %>% filter(ENROLLMENT_BEGIN_year == 2019)

# Use Bartlett test to check for equal variance or not (2018)
bartlett.test(retention~EverReceivedPromiseAward, data = nsc_promise_2018) #unequal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by EverReceivedPromiseAward
## Bartlett's K-squared = 37.596, df = 1, p-value = 8.7e-10
```

```
# T-test to check for significance of difference in retention (2018)
oneway.test(retention~EverReceivedPromiseAward, data = nsc_promise_2018, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  retention and EverReceivedPromiseAward
## F = 44.755, num df = 1.00, denom df = 130.29, p-value = 5.98e-10
```

```
# Use Bartlett test to check for equal variance or not (2019)
bartlett.test(retention~EverReceivedPromiseAward, data = nsc_promise_2019) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by EverReceivedPromiseAward
## Bartlett's K-squared = 2.1978, df = 1, p-value = 0.1382
```

```
# T-test to check for significance of difference in retention (2019)
oneway.test(retention~EverReceivedPromiseAward, data = nsc_promise_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and EverReceivedPromiseAward
## F = 0.27045, num df = 1, denom df = 696, p-value = 0.6032
```

Findings:

- For 2018, the difference is significant; for 2019, the difference is not significant.
- Findings from significance tests align with what we observe in the box plot above.
- Thus, the difference in retention is quite obvious in 2018, whereas in 2019 is not very obvious. Maybe the retention difference would become more obvious for more senior students.

Comparison II: Retention vs.Race

EDA

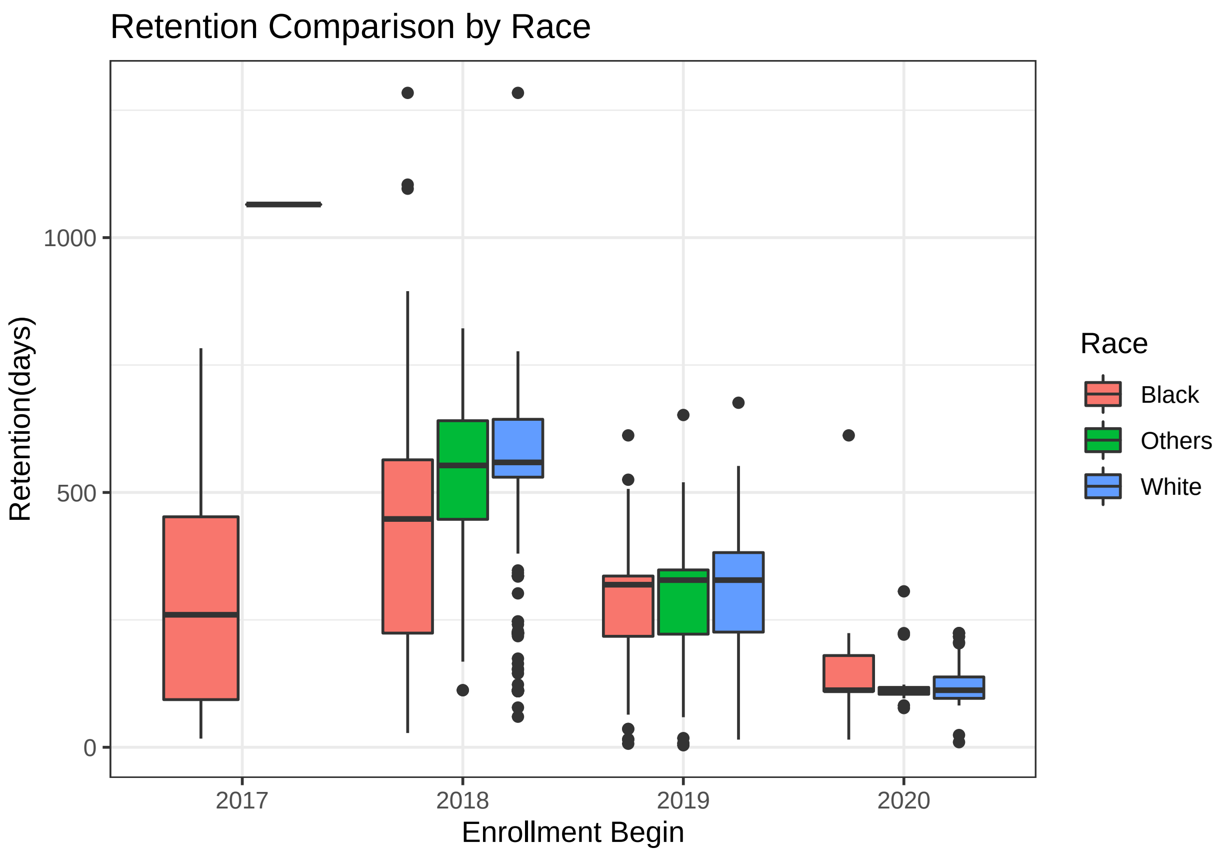
```
# Check racial composition
demographics %>% group_by(Race) %>% count() %>% arrange(-n)
```

```
## # A tibble: 7 x 2
## # Groups:   Race [7]
##   Race                                        n
##   <chr>                                   <int>
## 1 African American                         2766
## 2 White                                    1847
## 3 Multi-Racial                              327
## 4 Asian (not Pacific Islander)                  180
## 5 Hispanic                                       164
## 6 American Indian                                 10
## 7 Native Hawaiian or other Pacific Islander        4
```

- We see that the majority of students are blacks and whites, so we will explore the difference in student retention between these two races and categorize the rest racial groups as "Other".
- Also, we restrict on students who went to college in PA.

```
# Prepare data sets for the visualization of the comparison plot
demographics$RandomID <- as.character(demographics$RandomID)
nsc_demographics = nsc_promise %>%
  left_join(demographics, by = "RandomID") %>%
  mutate(Race = ifelse(Race != "White" & Race != "African American",
                       "Others",
                       ifelse(Race == "African American",
                              "Black",
                              "White"))) %>%
  na.omit(retention)
```

```
# Visualization of retention comparison under race
ggplot(nsc_demographics, aes(x=as.factor(ENROLLMENT_BEGIN_year),
                             y=retention,
                             fill= Race)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin",
       y = "Retention(days)",
       title = "Retention Comparison by Race") +
  theme_bw()
```



Check for Significance for 2018 & 2019 (T-tests)

```
# Prepare data set for the visualization of comparison plot
nsc_demographics2 = nsc_promise %>%
  left_join(demographics, by = "RandomID") %>%
  filter(Race %in% c("White","African American")) %>%
  mutate(Race = ifelse(Race == "White", "White","Black"))

# Prepare data sets for 2018 & 2019
nsc_demographics_2018 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2018)
nsc_demographics_2019 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2019)

# Use Bartlett test to check for equal variance or not (2018)
bartlett.test(retention~Race, data = nsc_demographics_2018) #unequal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 11.071, df = 1, p-value = 0.0008769
```

```
# T-test to check for significance of difference in retention (2018)
oneway.test(retention~Race, data = nsc_demographics_2018, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  retention and Race
## F = 20.256, num df = 1.00, denom df = 487.42, p-value = 8.48e-06
```

```
# Use Bartlett test to check for equal variance or not (2019)
bartlett.test(retention~Race, data = nsc_demographics_2019) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 0.077868, df = 1, p-value = 0.7802
```

```
# T-test to check for significance of difference in retention (2019)
oneway.test(retention~Race, data = nsc_demographics_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and Race
## F = 29.355, num df = 1, denom df = 591, p-value = 8.782e-08
```

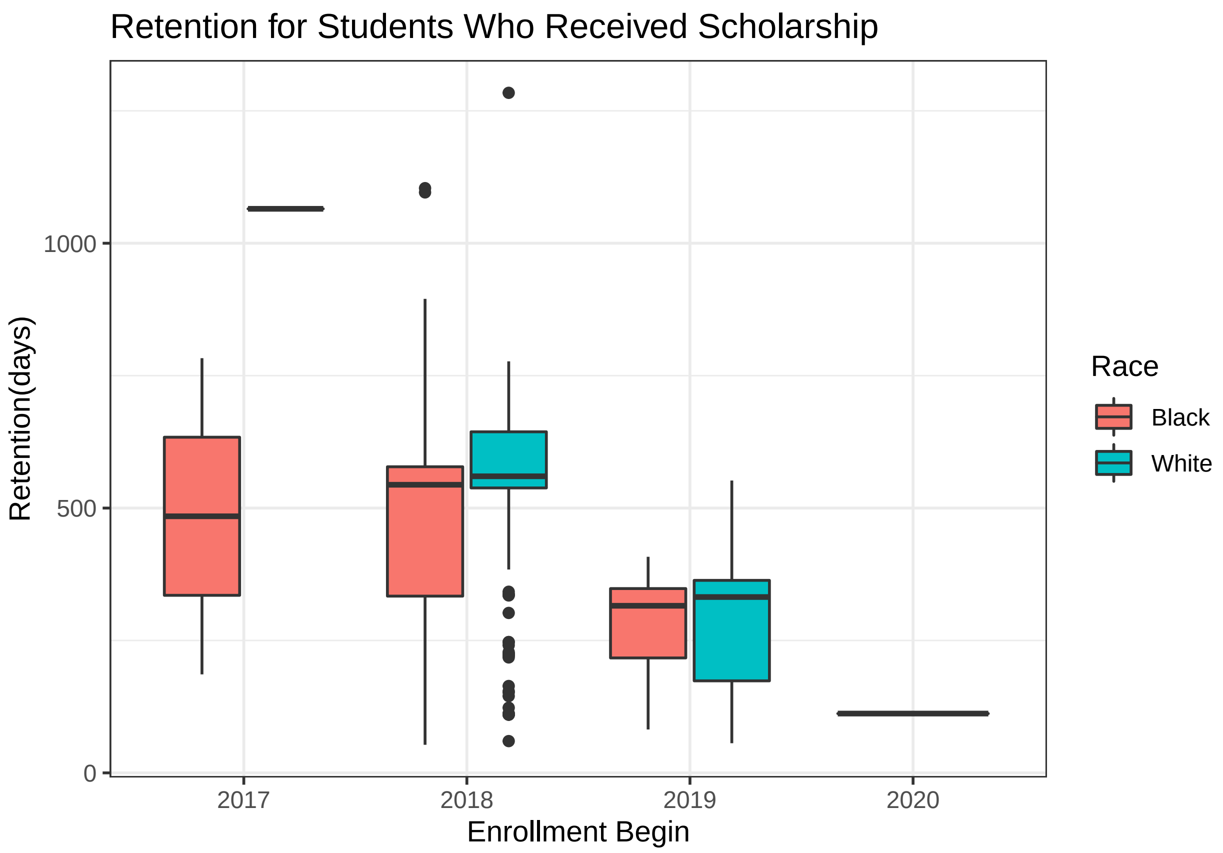
Findings:

- The differences are significant for both 2018 and 2019.
- Findings from significance tests align with our what we observe in the box plot above.
  - Well aligns with what we observe for 2018 enrollment year
  - Not perfectly aligns with what we see for 2019 enrollment year (i.e. no significant difference in retention between racial groups)

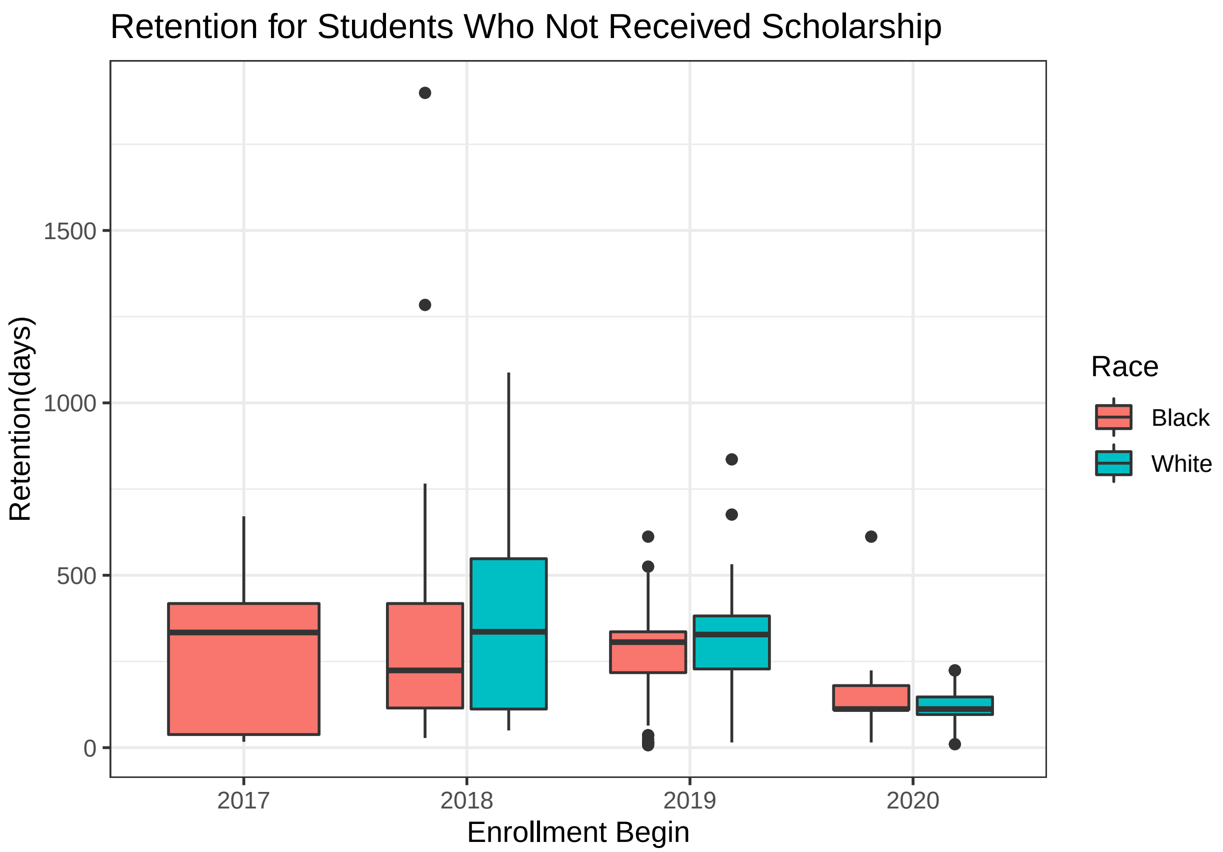
Comparison III: Retention vs. Interaction between Scholarship receipt and Race

EDA

```
# Students who received the scholarship
# Visualization of retention comparison under race
nsc_demographics2 %>%
  filter(EverReceivedPromiseAward=="yes") %>%
  ggplot(aes(x=as.factor(ENROLLMENT_BEGIN_year),
             y=retention, fill=Race)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin",
       y = "Retention(days)",
       title = "Retention for Students Who Received Scholarship") +
  theme_bw()
```



```
# Filter students who did not receive the scholarship
# Visualization of retention comparison under race
nsc_demographics2 %>%
  filter(EverReceivedPromiseAward=="no") %>%
  ggplot(aes(x=as.factor(ENROLLMENT_BEGIN_year),
             y=retention,
             fill=Race)) +
  geom_boxplot() +
  labs(x = "Enrollment Begin",
       y = "Retention(days)",
       title = "Retention for Students Who Not Received Scholarship") +
  theme_bw()
```



Check for Significance for 2018 & 2019

```
# Prepare data sets for 2018 & 2019
nsc_demographics2_2018 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2018,
         EverReceivedPromiseAward == "yes")
nsc_demographics2_2019 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2019,
         EverReceivedPromiseAward == "yes")

# Use Bartlett test to check for equal variance or not (2018)
bartlett.test(retention~Race, data = nsc_demographics2_2018) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 3.2004, df = 1, p-value = 0.07362
```

```
# T-test to check for significance of difference in retention (2018)
oneway.test(retention~Race, data = nsc_demographics2_2018, var.equal = TRUE)
```

```
##
```

```
##  One-way analysis of means
##
## data:  retention and Race
## F = 13.073, num df = 1, denom df = 407, p-value = 0.0003371
```

```
# Use Bartlett test to check for equal variance or not (2019)
bartlett.test(retention~Race, data = nsc_demographics2_2019) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 1.7403, df = 1, p-value = 0.1871
```

```
# T-test to check for significance of difference in retention (2019)
oneway.test(retention~Race, data = nsc_demographics2_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and Race
## F = 0.44839, num df = 1, denom df = 38, p-value = 0.5071
```

```
# Prepare data sets for 2018 & 2019
nsc_demographics3_2018 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2018,
         EverReceivedPromiseAward == "no")
nsc_demographics3_2019 = nsc_demographics2 %>%
  filter(ENROLLMENT_BEGIN_year == 2019,
         EverReceivedPromiseAward == "no")

# Use Bartlett test to check for equal variance or not (2018)
bartlett.test(retention~Race, data = nsc_demographics3_2018) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 0.79106, df = 1, p-value = 0.3738
```

```
# T-test to check for significance of difference in retention (2018)
oneway.test(retention~Race, data = nsc_demographics3_2018, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and Race
## F = 0.38341, num df = 1, denom df = 95, p-value = 0.5373
```

```
# Use Bartlett test to check for equal variance or not (2019)
bartlett.test(retention~Race, data = nsc_demographics3_2019) #equal variance
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  retention by Race
## Bartlett's K-squared = 0.018019, df = 1, p-value = 0.8932
```

```
# T-test to check for significance of difference in retention (2019)
oneway.test(retention~Race, data = nsc_demographics3_2019, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  retention and Race
## F = 29.444, num df = 1, denom df = 551, p-value = 8.633e-08
```

Findings:

- Among students who received the scholarship, we see that for 2018, there is a significant racial difference, but for 2019 the difference is not significant.
- Among students who did not receive the scholarship, we see reverse happens, the difference is not significant for 2018 but significant for 2019.

Part II: Multivariate Regression Analysis for Whole Range

Data Wrangling

```
# Import data sets
# data_qualified is the final data set used in logistic regression modeling from scholarship analysis
# retention_demographics is the joined data set between NSC and demographics
data_qualified <- read_csv("data_qualified.csv")
retention_demographics <- read_csv("retention_demographics.csv") %>% dplyr::select(-c(1))
retention_demographics$EverReceivedPromiseAward <-
  ifelse(retention_demographics$EverReceivedPromiseAward == "yes", 1, 0)
```

```
# Inner join data_qualified and retention_demographics to obtain data set used for regression
retention_reg <- inner_join(data_qualified, retention_demographics, by = c("RandomID",
                                                                           "Race",
                                                                           "GradYear",
                                                                           "EverReceivedPromiseAward"))

# Categorize race as white, black, and other
retention_reg$Race <- ifelse(retention_reg$Race != "White" & retention_reg$Race != "African American",
                             "Other",
                             retention_reg$Race)
```

First Trial on Linear Regression

Linear Regression for Students Starting College in 2018

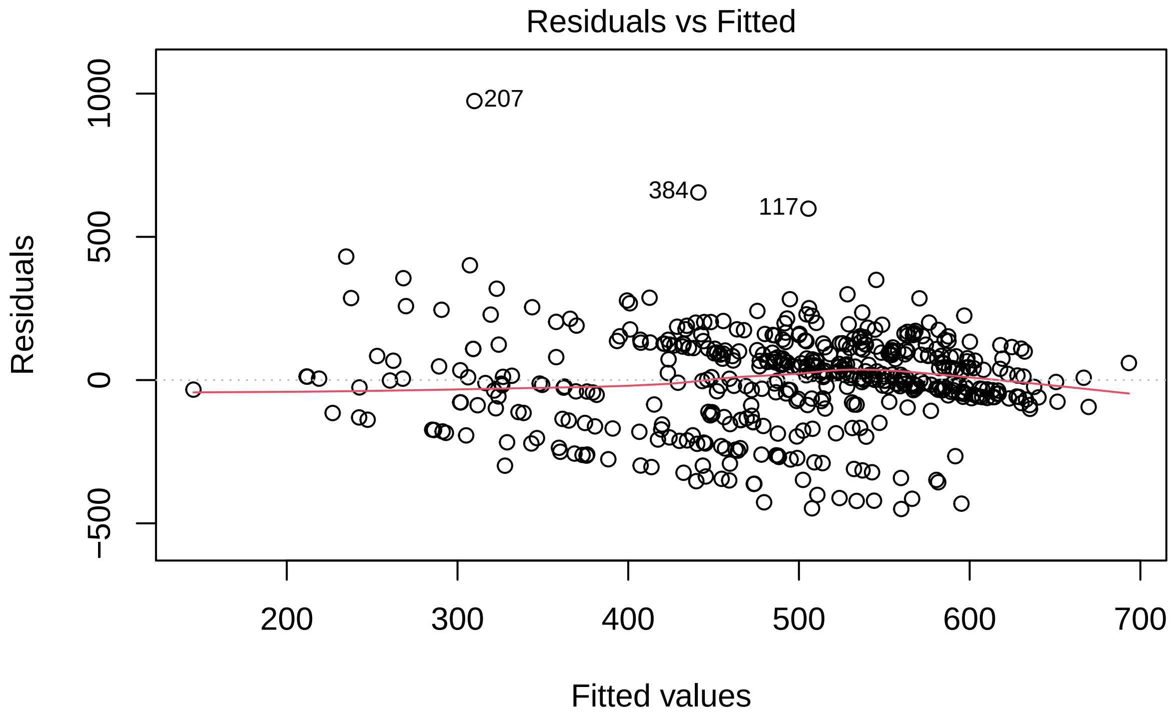
```
# Filter students who started college in 2018
retention2018 <- retention_reg %>%
  filter(ENROLLMENT_BEGIN_year == 2018) #485 obs
```

```
# Construct null and full model for linear regression 2018
md.null2018 = lm(retention ~ Race + EverReceivedPromiseAward + Race*EverReceivedPromiseAward,
                 data = retention2018)
md.full2018 = lm(retention~AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race+Gender+ELLStatus+
                   IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd+
                   EverReceivedPromiseAward+QualifiedforCorePromise+semester+
                   Race*EverReceivedPromiseAward, data = retention2018)
```

```
# Stepwise variable selection (both directions)
md.2018 = step(md.null2018,
               list(lower = formula(md.null2018),
                    upper = formula(md.full2018)),
               direction = "both", trace = 0)
summary(md.2018)
```

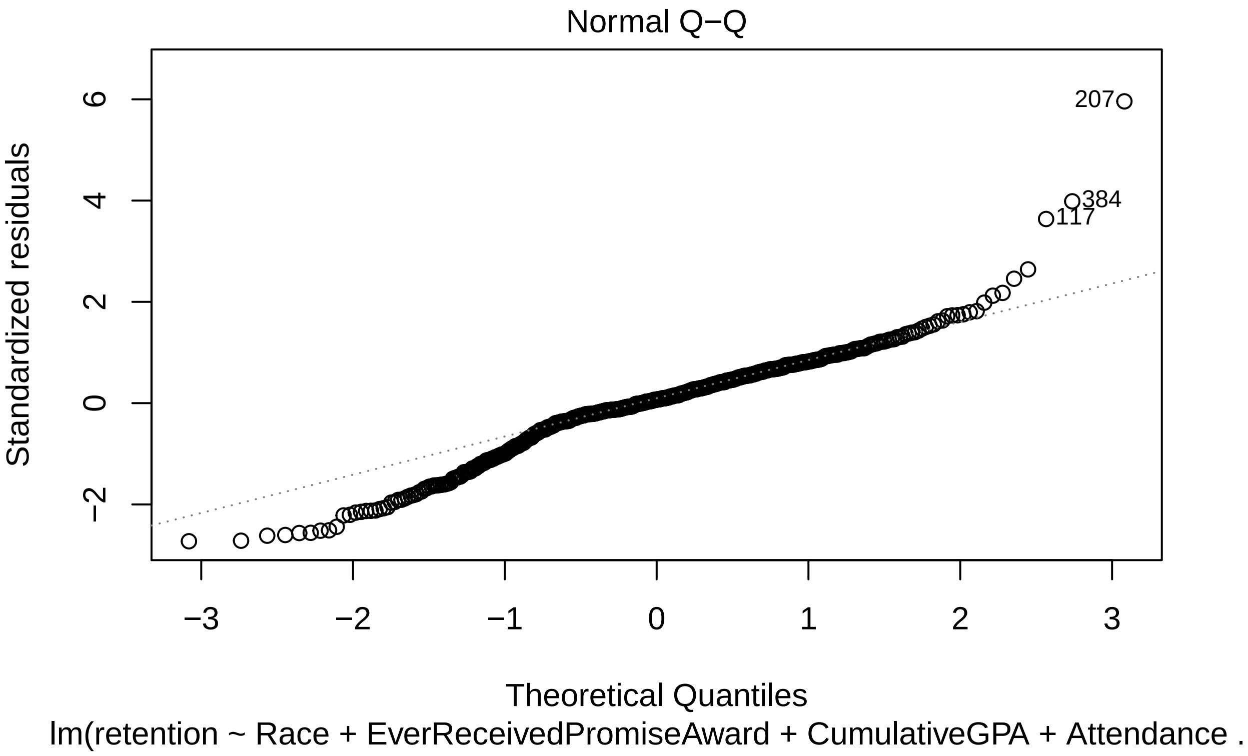
```
##
## Call:
## lm(formula = retention ~ Race + EverReceivedPromiseAward + CumulativeGPA +
##     AttendanceRate + Gender + SAT_Total + KeystoneMean + Num_CTE +
##     Race:EverReceivedPromiseAward, data = retention2018)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -449.94  -68.04   11.33   99.33  974.10
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        568.52533  491.96068   1.156 0.248415
## RaceOther                           22.48731   58.39909   0.385 0.700364
## RaceWhite                            6.72192   46.56124   0.144 0.885272
## EverReceivedPromiseAward           117.31801   28.94333   4.053  5.9e-05 ***
## CumulativeGPA                       71.52453   22.64271   3.159 0.001685 **
## AttendanceRate                     878.09592  277.32582   3.166 0.001644 **
## GenderMale                         -34.40616   16.54698  -2.079 0.038128 *
## SAT_Total                            0.27647    0.08116   3.407 0.000714 ***
## KeystoneMean                        -1.01836    0.33855  -3.008 0.002770 **
## Num_CTE                            -14.34949    9.39897  -1.527 0.127502
## RaceOther:EverReceivedPromiseAward  -2.34132   64.03626  -0.037 0.970849
## RaceWhite:EverReceivedPromiseAward   2.41024   48.48180   0.050 0.960371
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 165.7 on 473 degrees of freedom
## Multiple R-squared:  0.2545, Adjusted R-squared:  0.2371
## F-statistic: 14.68 on 11 and 473 DF,  p-value: < 2.2e-16
```

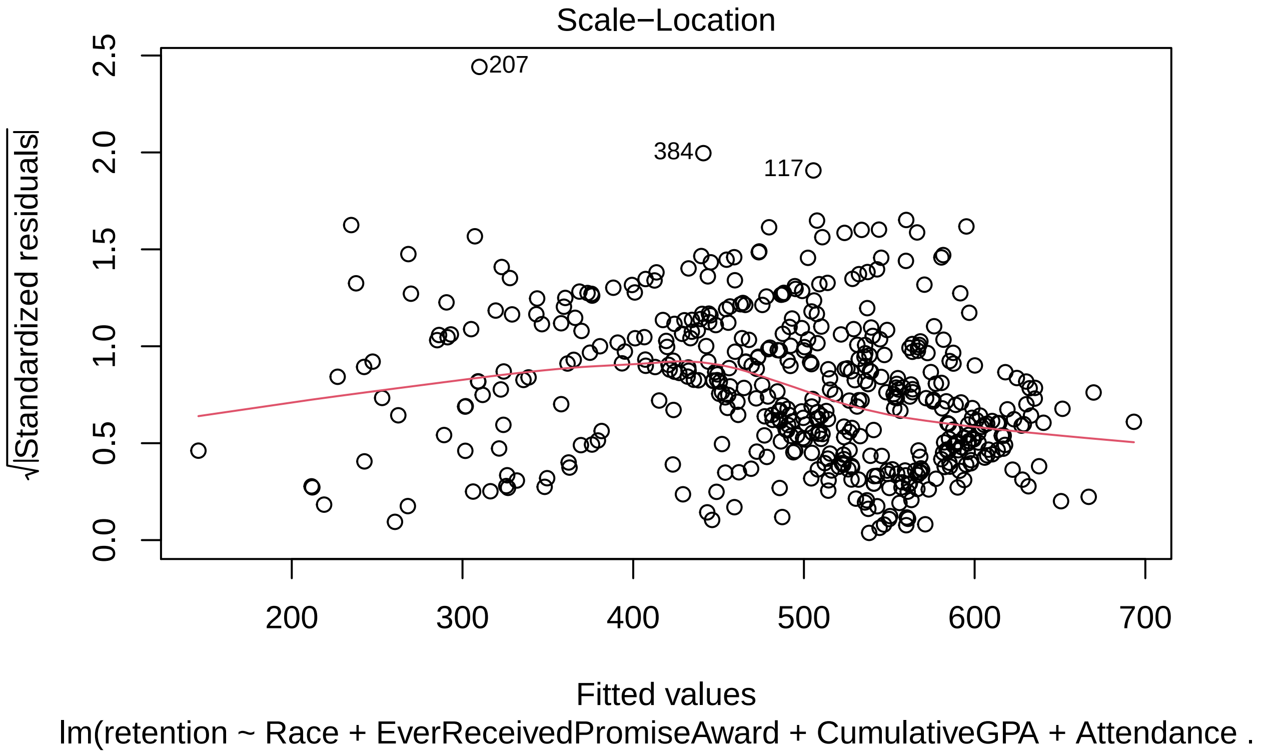
```r
# Model diagnostics
plot(md.2018)
```

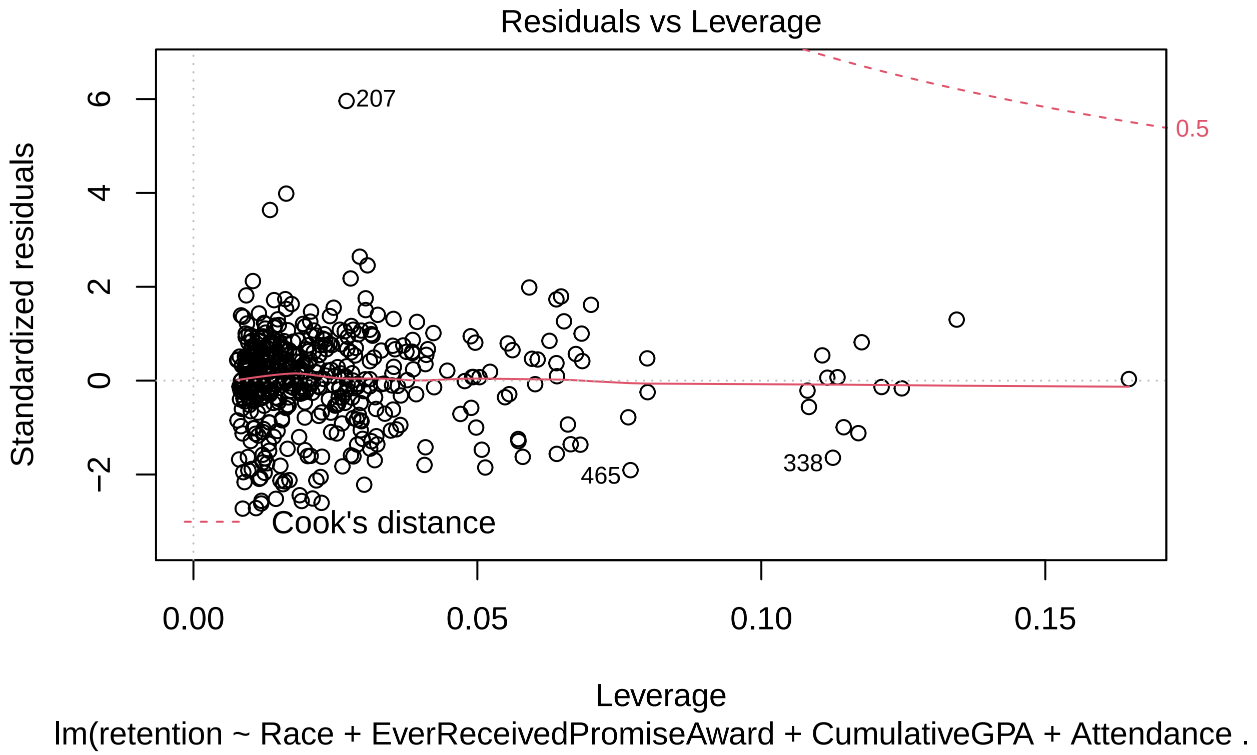


Residuals vs Fitted

lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + Attendance .


Normal Q−Q
Standardized residuals
6
4
2
0
−2
207
384
117
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + Attendance .


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
200
300
400
500
600
700
207
384
117
Fitted values
lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + Attendance .



Linear Regression for Students Starting College in 2019

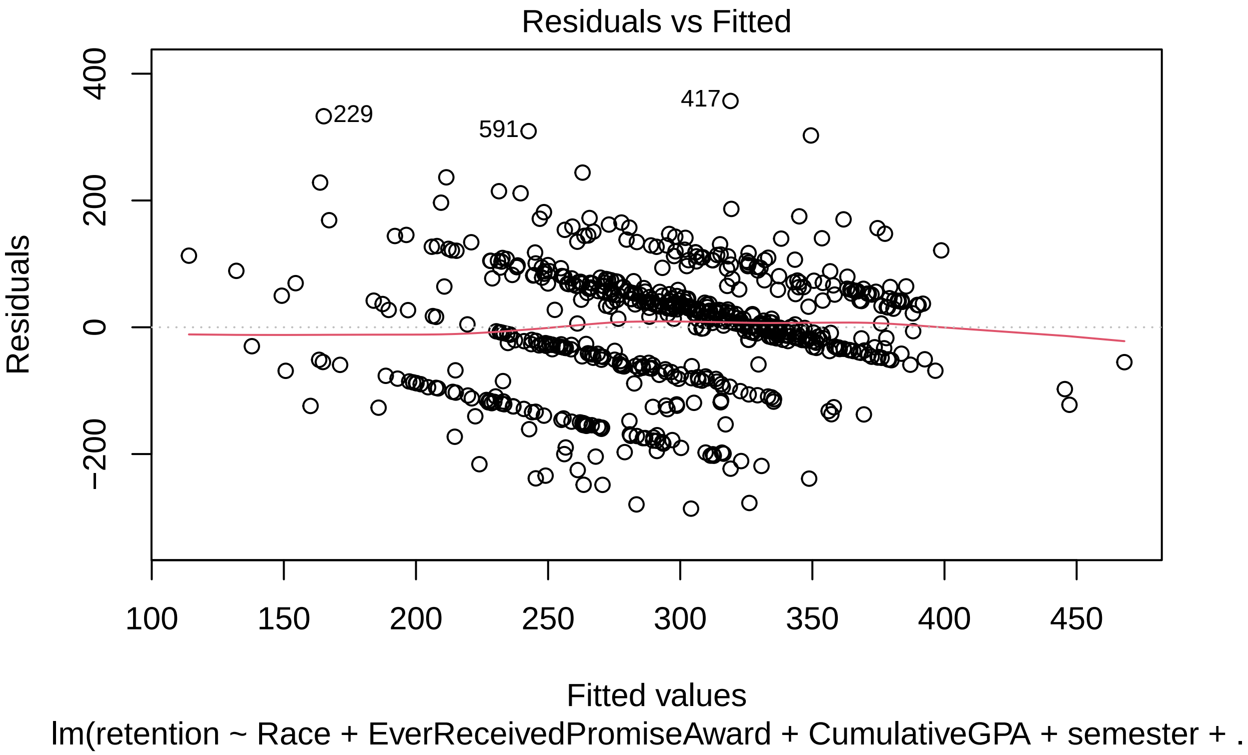
```
retention2019 <- retention_reg %>%
  filter(ENROLLMENT_BEGIN_year == 2019) #663 obs
```

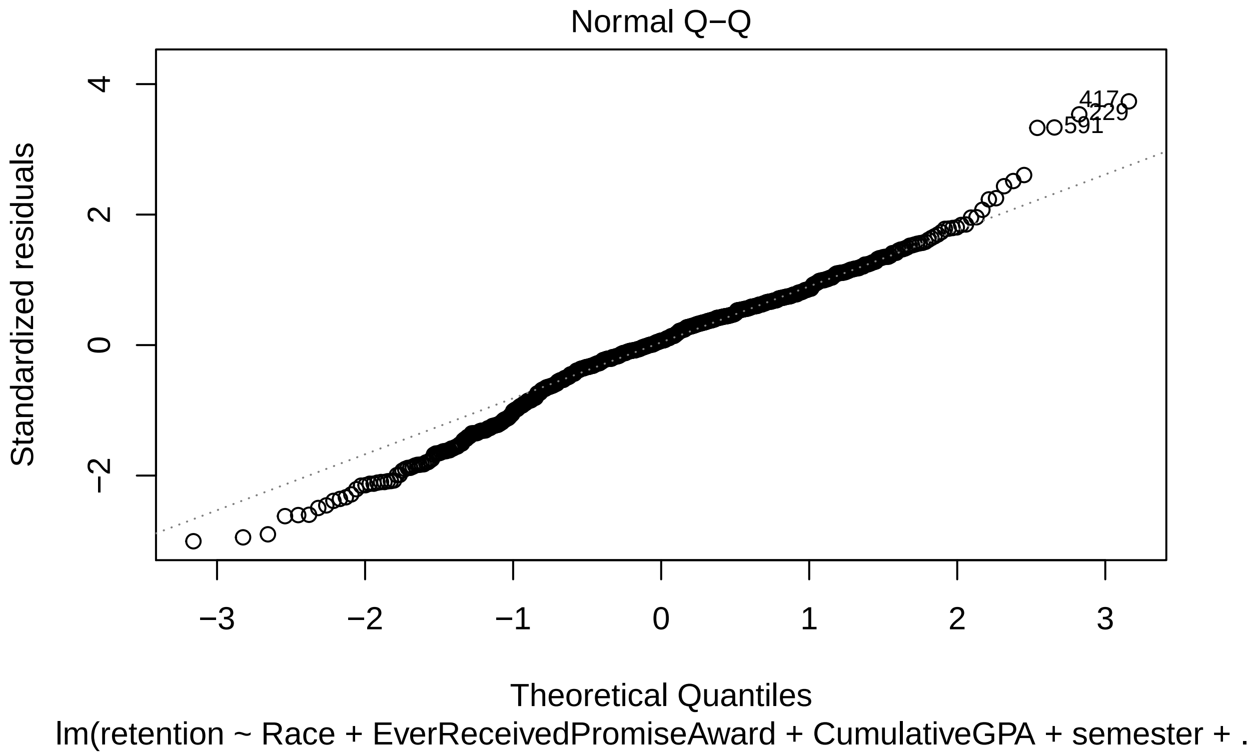
```
# Construct null and full model for linear regression 2019
md.null2019 = lm(retention ~ Race + EverReceivedPromiseAward + Race*EverReceivedPromiseAward,
                 data = retention2019)
md.full2019 = lm(retention~AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race+Gender+ELLStatus+
                   IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd+
                   EverReceivedPromiseAward+QualifiedforCorePromise+semester+
                   Race*EverReceivedPromiseAward, data = retention2019)
```

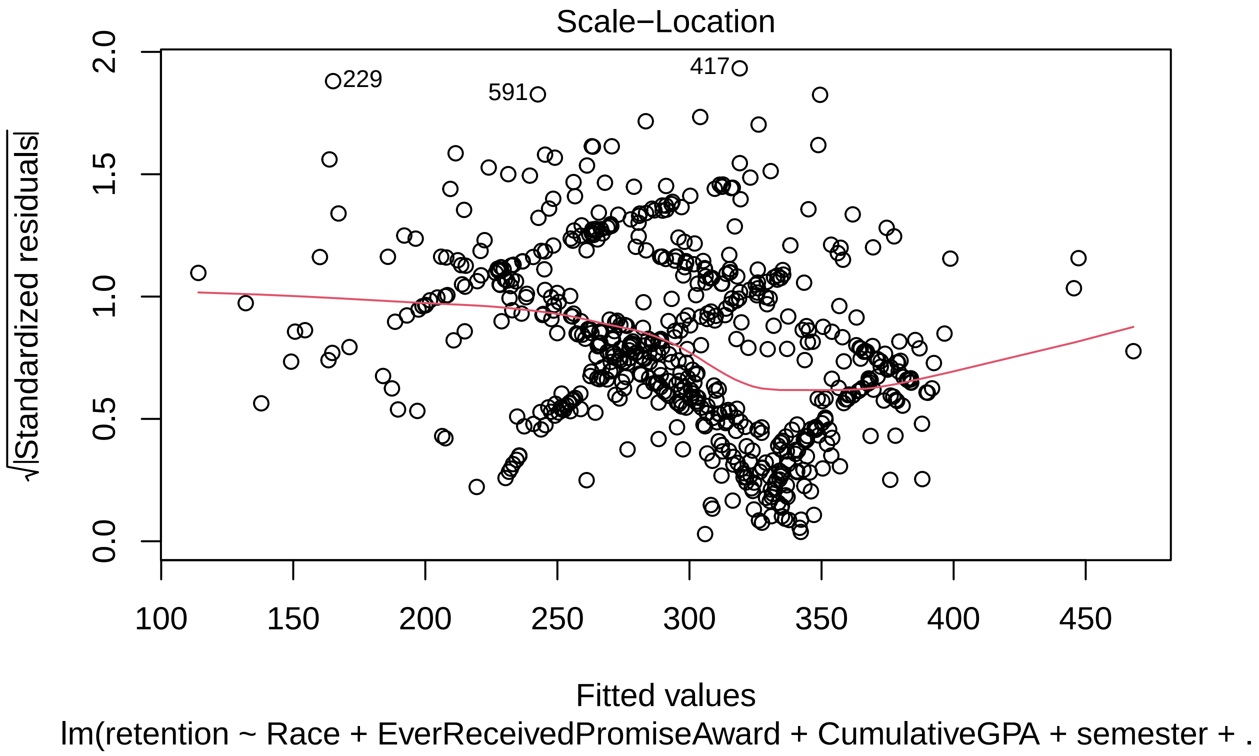
```
# Stepwise variable selection (both directions)
md.2019 = step(md.null2019,
               list(lower = formula(md.null2019),
                    upper = formula(md.full2019)),
               direction = "both", trace = 0)
summary(md.2019)
```

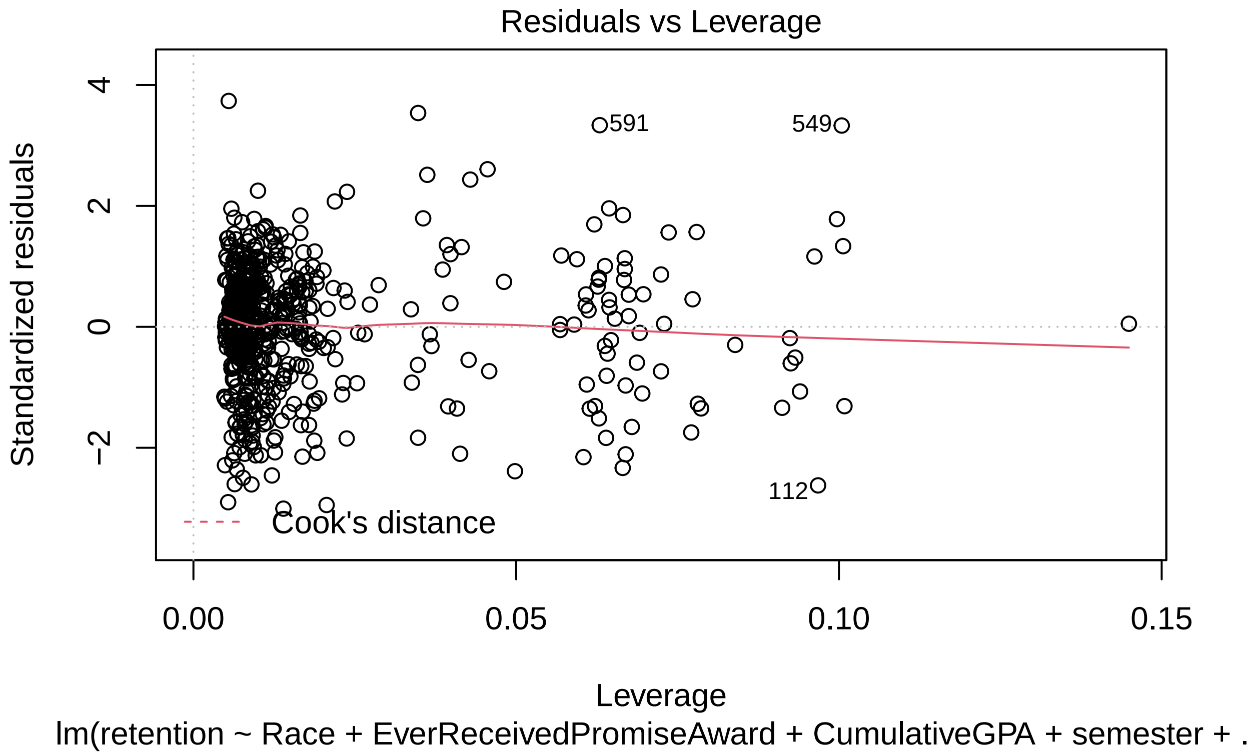
```
##
## Call:
## lm(formula = retention ~ Race + EverReceivedPromiseAward + CumulativeGPA +
##     semester + SAT_Total + Gender + ELLStatus + Race:EverReceivedPromiseAward,
##     data = retention2019)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -286.08  -51.18    6.52   59.19  356.98
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -7.03681   33.11796  -0.212  0.83180
## RaceOther                            -12.65312   13.11819  -0.965  0.33515
## RaceWhite                             -0.16154    9.94309  -0.016  0.98704
## EverReceivedPromiseAward              11.22200   24.74243   0.454  0.65031
## CumulativeGPA                         59.32630    9.13924   6.491 1.74e-10 ***
## semesterSpring                       -66.25741   16.80158  -3.944 8.94e-05 ***
## SAT_Total                              0.08155    0.02621   3.111  0.00195 **
## GenderMale                           -14.80311    8.04102  -1.841  0.06611 .
## ELLStatusNot in ELL                   44.55902   25.89806   1.721  0.08583 .
## RaceOther:EverReceivedPromiseAward  98.87681   38.34957   2.578  0.01016 *
## RaceWhite:EverReceivedPromiseAward -23.19223   33.06471  -0.701  0.48330
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 95.82 on 622 degrees of freedom
## Multiple R-squared:  0.2289, Adjusted R-squared:  0.2165
## F-statistic: 18.46 on 10 and 622 DF,  p-value: < 2.2e-16
```

```
# Model diagnostics
plot(md.2019)
```


Residuals vs Fitted
Residuals
400
200
0
−200
100
150
200
250
300
350
400
450
Fitted values
lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + semester + .
229
591
417


Normal Q−Q
Standardized residuals
4
2
0
−2
−3
−2
−1
0
1
2
3
417
229
591
Theoretical Quantiles
lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + semester + .


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
229
591
417
100
150
200
250
300
350
400
450
Fitted values
lm(retention ~ Race + EverReceivedPromiseAward + CumulativeGPA + semester + .



Second Trial on Poisson Regression

Poisson Regression for Students Starting College in 2018

```
# Construct null possion model and full poisson model
poisson2018.null = glm(retention~Race+EverReceivedPromiseAward+Race*EverReceivedPromiseAward,
                       family = "poisson", data = retention2018)
poisson2018.full = glm(retention~AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race+Gender+
                         ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd+
                         EverReceivedPromiseAward+QualifiedforCorePromise+semester+
                         Race*EverReceivedPromiseAward, family = "poisson", data = retention2018)
```

```
# Stepwise variable selection (both directions)
poisson2018 = step(poisson2018.full,
                   list(lower = formula(poisson2018.null),
                        upper = formula(poisson2018.full)),
                   direction = "both",
                   trace = 0)
summary(poisson2018)
```

```
##
## Call:
## glm(formula = retention ~ AttendanceRate + Num_AP + Num_CTE +
##     KeystoneMean + Race + Gender + ELLStatus + IEPGroup + EconDisad +
##     SAT_Total + CumulativeGPA + MagnetInd + EverReceivedPromiseAward +
##     QualifiedforCorePromise + semester + Race * EverReceivedPromiseAward,
##     family = "poisson", data = retention2018)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -23.933   -3.585    0.410    4.477   38.736
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         6.110e+00  1.441e-01  42.393  < 2e-16 ***
## AttendanceRate                      2.246e+00  8.707e-02  25.800  < 2e-16 ***
## Num_AP                              4.585e-03  1.041e-03   4.403 1.07e-05 ***
## Num_CTE                            -2.825e-02  2.743e-03 -10.300  < 2e-16 ***
## KeystoneMean                       -2.166e-03  9.581e-05 -22.603  < 2e-16 ***
## RaceOther                           4.463e-02  2.040e-02   2.187  0.02873 *
## RaceWhite                           6.300e-02  1.500e-02   4.199 2.68e-05 ***
## GenderMale                         -7.107e-02  4.611e-03 -15.415  < 2e-16 ***
## ELLStatusNot in ELL                -5.165e-02  1.620e-02  -3.189  0.00143 **
## IEPGroupIEP                         2.181e-02  1.123e-02   1.942  0.05217 .
## IEPGroupNot IEP or Gifted          -5.117e-02  6.443e-03  -7.942 1.99e-15 ***
## EconDisadRegular Lunch              1.093e-02  4.630e-03   2.361  0.01822 *
## SAT_Total                           4.265e-04  2.367e-05  18.015  < 2e-16 ***
## CumulativeGPA                       1.697e-01  7.631e-03  22.244  < 2e-16 ***
## MagnetInd                           3.736e-02  4.682e-03   7.979 1.47e-15 ***
## EverReceivedPromiseAward            3.162e-01  9.688e-03  32.636  < 2e-16 ***
## QualifiedforCorePromiseyes         -4.058e-02  9.452e-03  -4.293 1.76e-05 ***
## semesterSpring                      1.279e-01  4.022e-02   3.180  0.00147 **
## RaceOther:EverReceivedPromiseAward -4.254e-03  2.116e-02  -0.201  0.84068
## RaceWhite:EverReceivedPromiseAward -4.717e-02  1.540e-02  -3.064  0.00219 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 41127  on 484  degrees of freedom
## Residual deviance: 31222  on 465  degrees of freedom
## AIC: 35106
##
## Number of Fisher Scoring iterations: 5
```

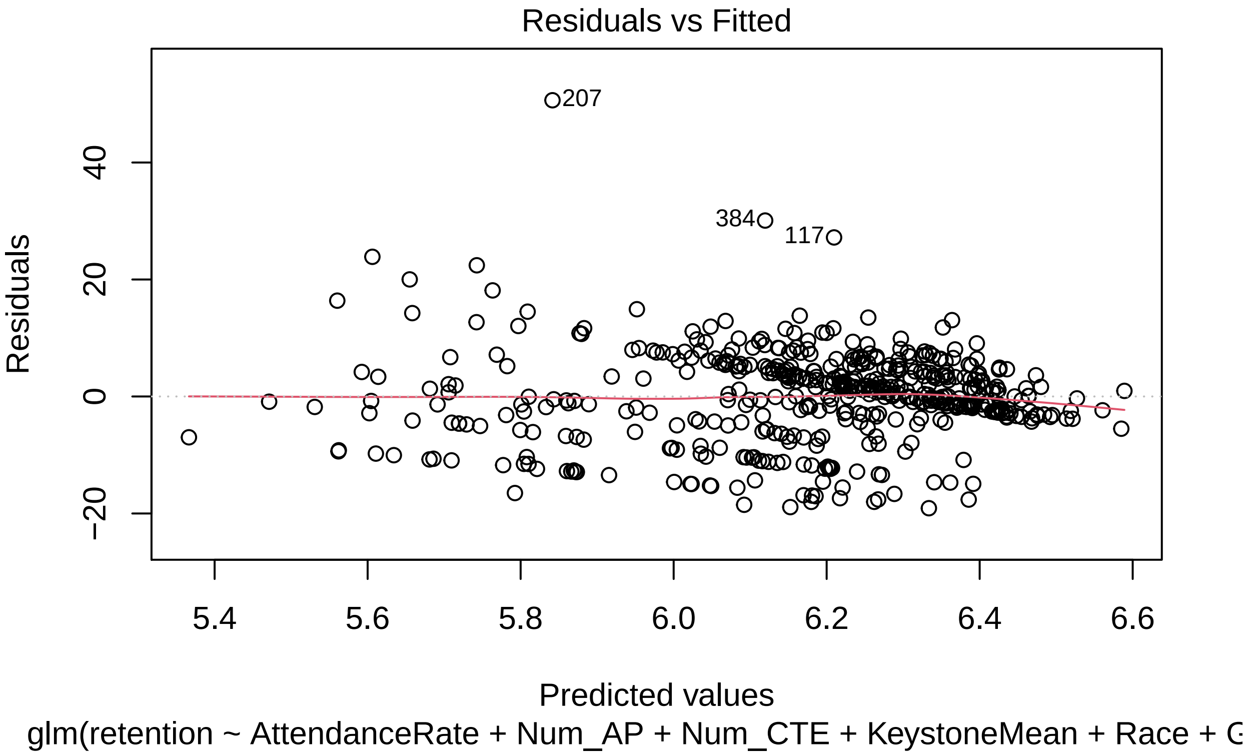
```
# Model diagnostics
# Residual deviance
pchisq(poisson2018$deviance,
       poisson2018$df.residual,
       lower.tail = FALSE)
```

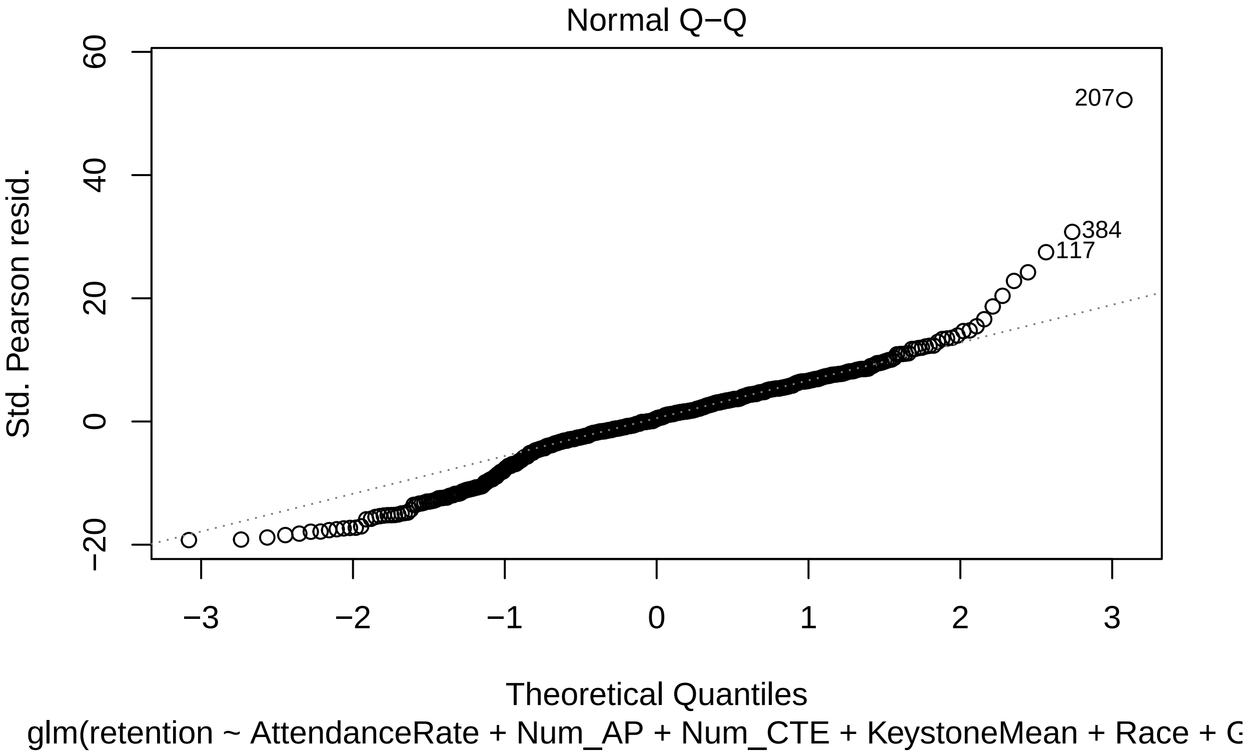
```
## [1] 0
```

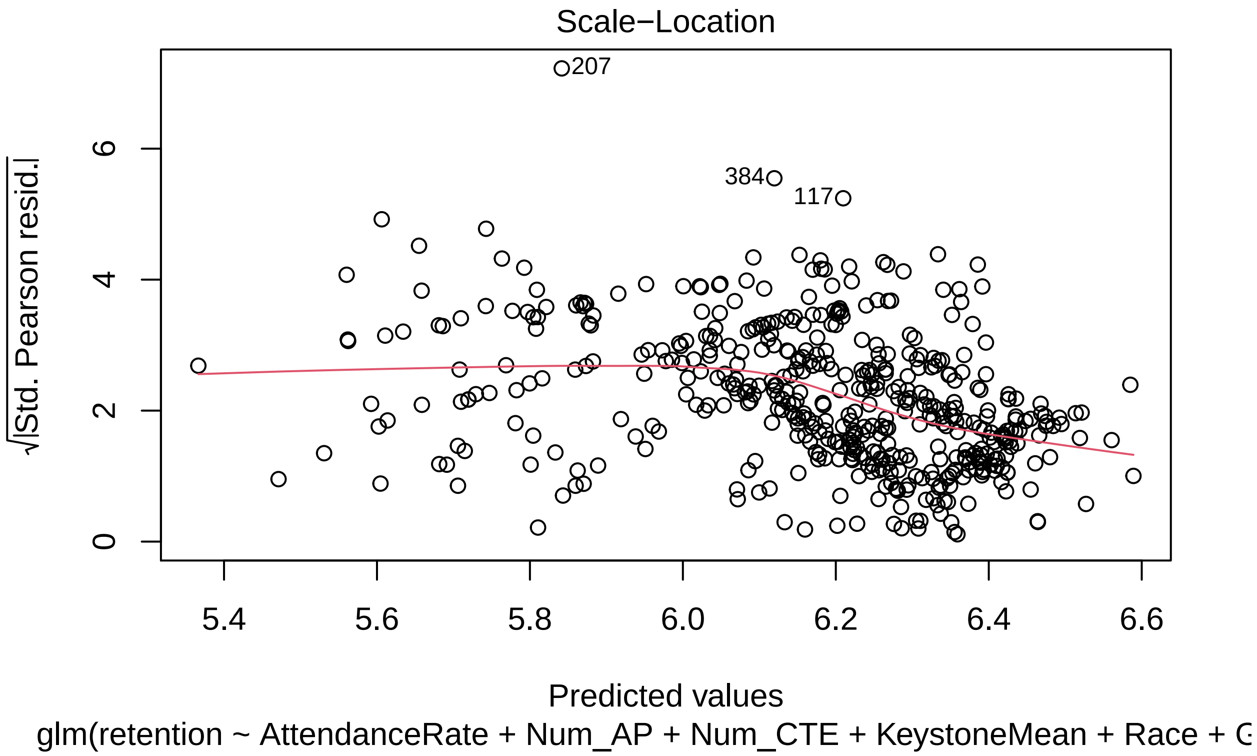
```
# Diagnostics plot
plot(poisson2018)
```

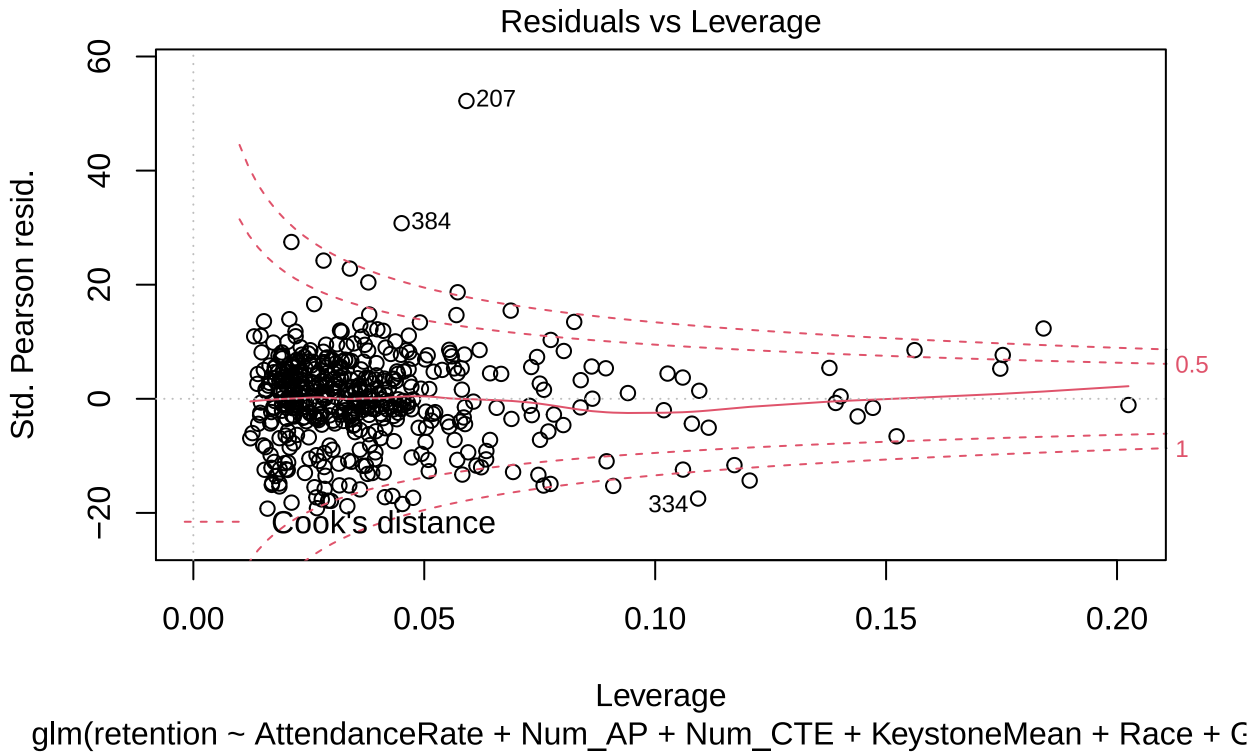
```
## Warning: not plotting observations with leverage one:
```

```
## 171
```

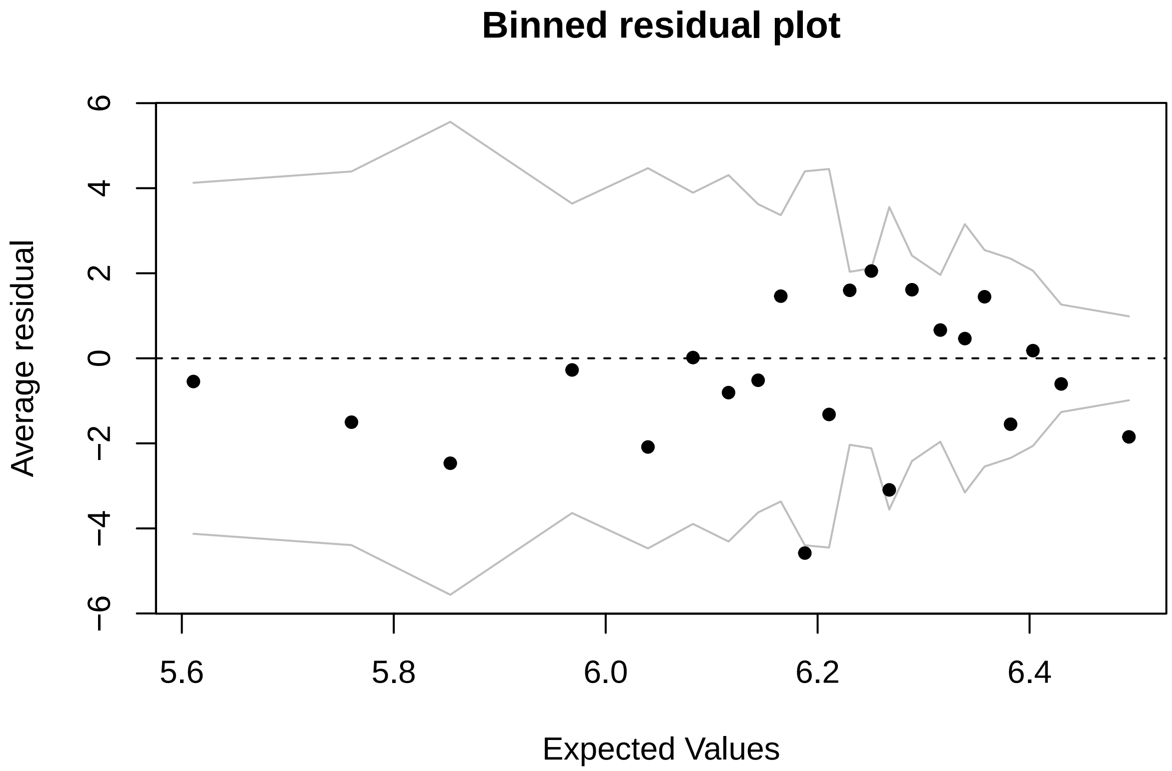

Residuals vs Fitted
Residuals
Predicted values
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G
207
384
117
40
20
0
−20
5.4
5.6
5.8
6.0
6.2
6.4
6.6


Normal Q−Q
Std. Pearson resid.
60
40
20
0
−20
207
384
117
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G


Scale−Location
207
384
117
√|Std. Pearson resid.|
0
2
4
6
5.4
5.6
5.8
6.0
6.2
6.4
6.6
Predicted values
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G



```r
# Binned plot
x.2018 = predict(poisson2018)
y.2018 = residuals(poisson2018)
binnedplot(x.2018,y.2018)
```



Poisson Regression for Students Starting College in 2019

```
# Construct null possion model and full poisson model
poisson2019.null = glm(retention~Race+EverReceivedPromiseAward+Race*EverReceivedPromiseAward,
                       family = "poisson", data = retention2019)
poisson2019.full = glm(retention~AttendanceRate+Num_AP+Num_CTE+KeystoneMean+Race+Gender+
                         ELLStatus+IEPGroup+EconDisad+SAT_Total+CumulativeGPA+MagnetInd+
                         EverReceivedPromiseAward+QualifiedforCorePromise+semester+
                         Race*EverReceivedPromiseAward, family = "poisson", data = retention2019)
```

```
# Stepwise variable selection (both directions)
poisson2019 = step(poisson2019.full, list(lower = formula(poisson2019.null),
                                          upper = formula(poisson2019.full)),
                   direction = "both", trace = 0)
summary(poisson2019)
```

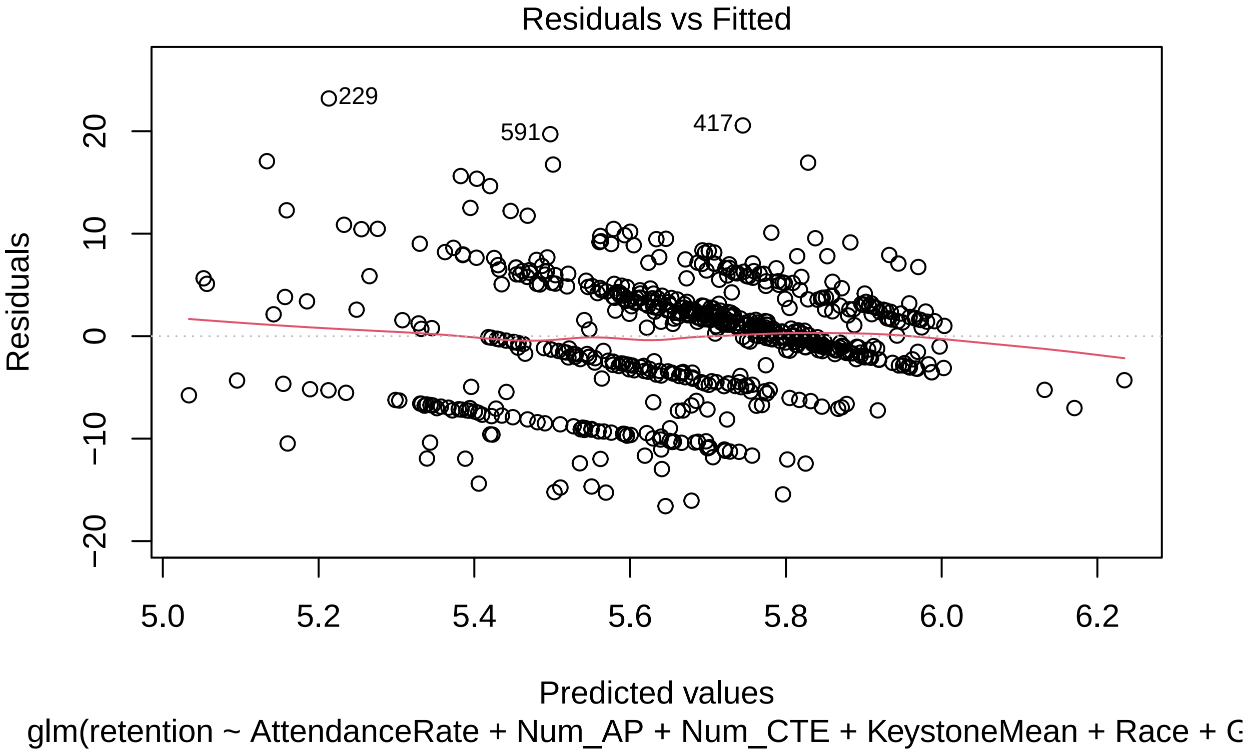
```
##
## Call:
## glm(formula = retention ~ AttendanceRate + Num_AP + Num_CTE +
##     KeystoneMean + Race + Gender + ELLStatus + EconDisad + SAT_Total +
##     CumulativeGPA + MagnetInd + EverReceivedPromiseAward + QualifiedforCorePromise +
##     semester + Race:EverReceivedPromiseAward, family = "poisson",
##     data = retention2019)
```

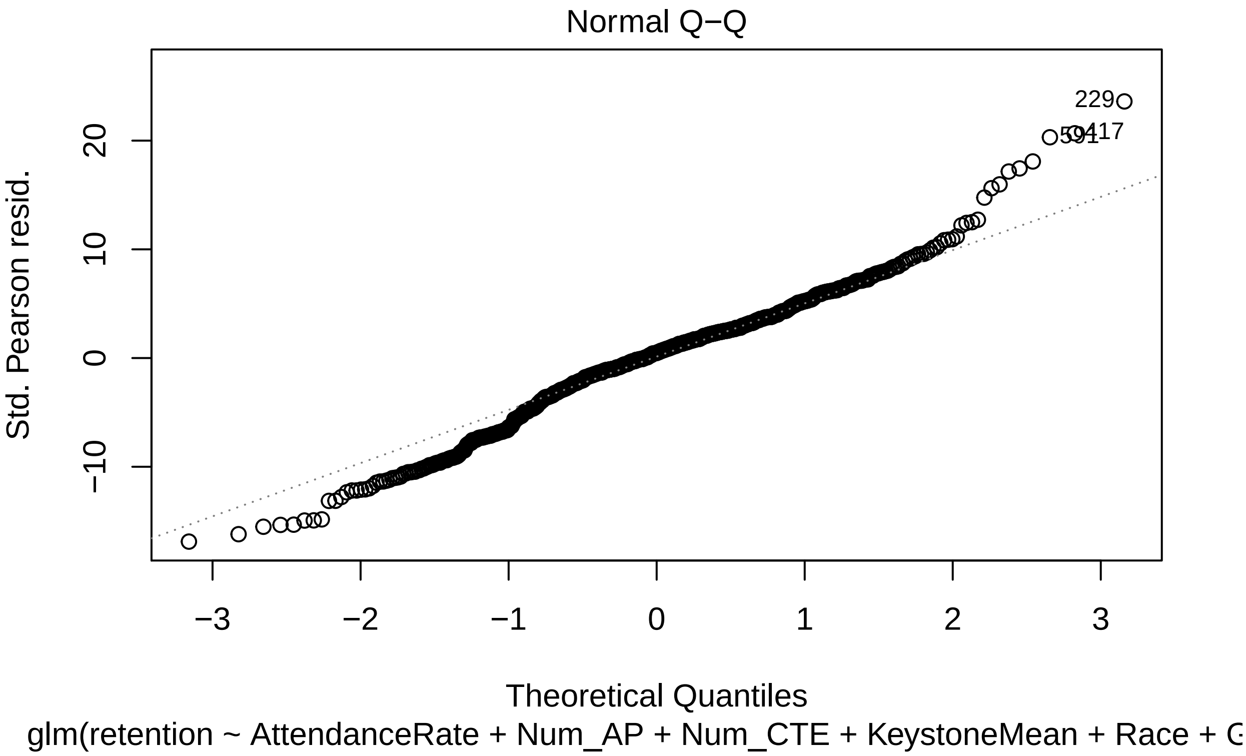
```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -22.8894   -3.1941    0.4624    3.2503   19.0998
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        4.432e+00  1.604e-01  27.634  < 2e-16 ***
## AttendanceRate                     6.489e-01  8.490e-02   7.644 2.11e-14 ***
## Num_AP                             3.395e-03  1.218e-03   2.787  0.00531 **
## Num_CTE                           -1.305e-02  2.532e-03  -5.152 2.57e-07 ***
## KeystoneMean                      -2.601e-04  1.077e-04  -2.415  0.01575 *
## RaceOther                         -3.699e-02  8.317e-03  -4.447 8.71e-06 ***
## RaceWhite                          9.875e-04  6.405e-03   0.154  0.87748
## GenderMale                        -5.275e-02  5.012e-03 -10.524  < 2e-16 ***
## ELLStatusNot in ELL                1.950e-01  1.834e-02  10.636  < 2e-16 ***
## EconDisadRegular Lunch             2.151e-02  5.154e-03   4.174 3.00e-05 ***
## SAT_Total                          2.291e-04  2.543e-05   9.011  < 2e-16 ***
## CumulativeGPA                      1.701e-01  7.749e-03  21.951  < 2e-16 ***
## MagnetInd                          2.889e-02  5.126e-03   5.636 1.74e-08 ***
## EverReceivedPromiseAward           2.282e-02  1.587e-02   1.438  0.15039
## QualifiedforCorePromiseyes         6.454e-02  9.687e-03   6.663 2.69e-11 ***
## semesterSpring                    -2.563e-01  1.147e-02 -22.343  < 2e-16 ***
## RaceOther:EverReceivedPromiseAward  3.113e-01  2.271e-02  13.703  < 2e-16 ***
## RaceWhite:EverReceivedPromiseAward -5.807e-02  2.075e-02  -2.798  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 29317  on 632  degrees of freedom
## Residual deviance: 23250  on 615  degrees of freedom
## AIC: 27983
##
## Number of Fisher Scoring iterations: 5
```

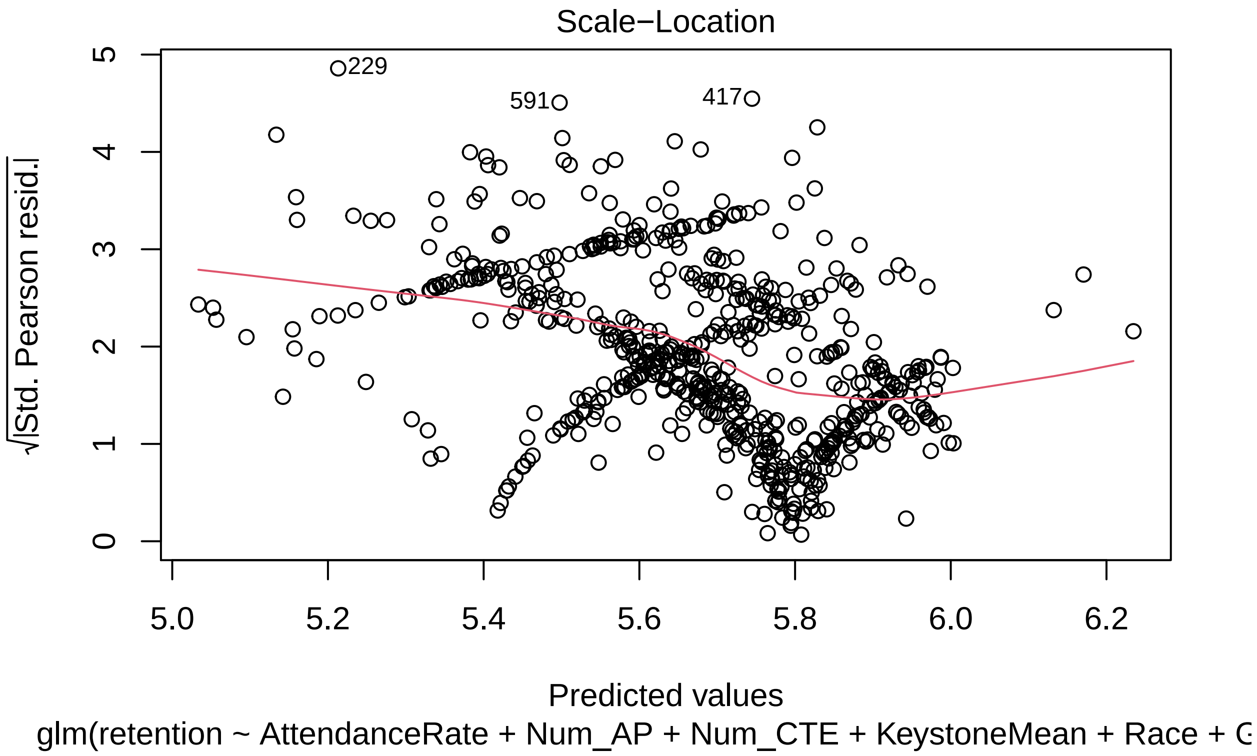
```
# Model diagnostics
# Residual deviance
pchisq(poisson2019$deviance, poisson2019$df.residual, lower.tail = FALSE)
```

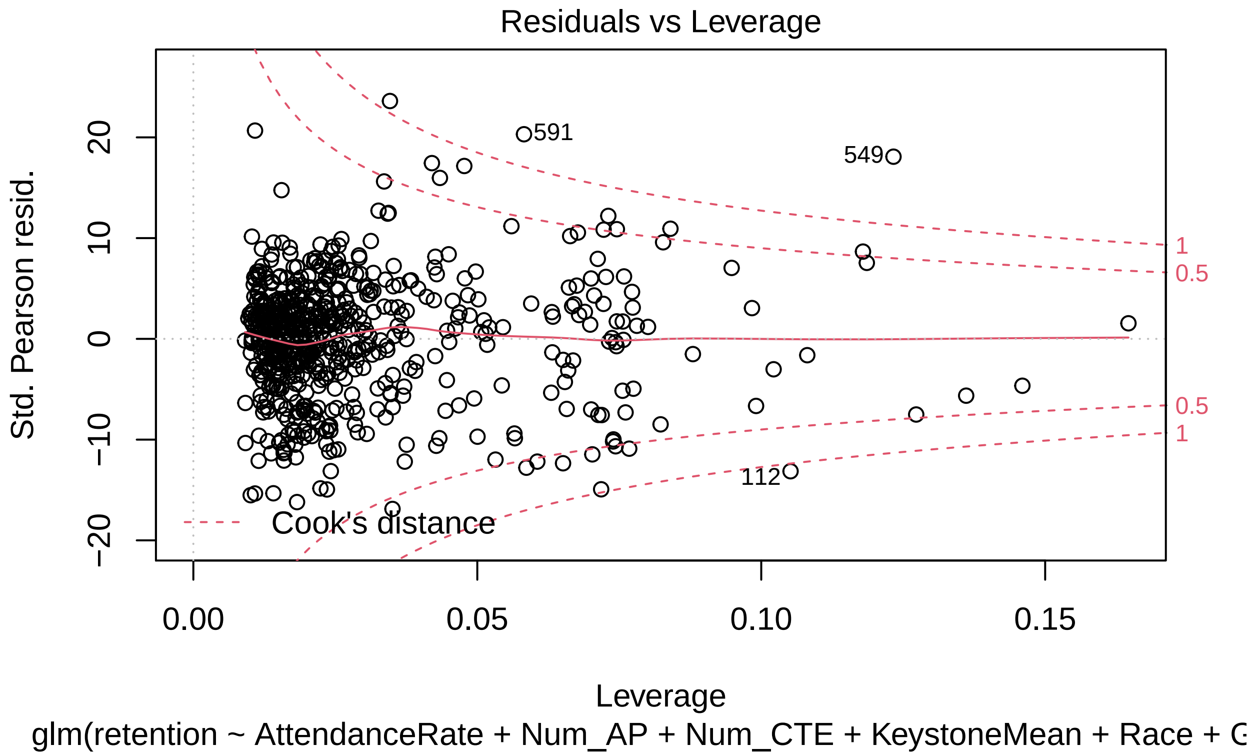
```
## [1] 0
```

```
# Diagnostics plot
plot(poisson2019)
```

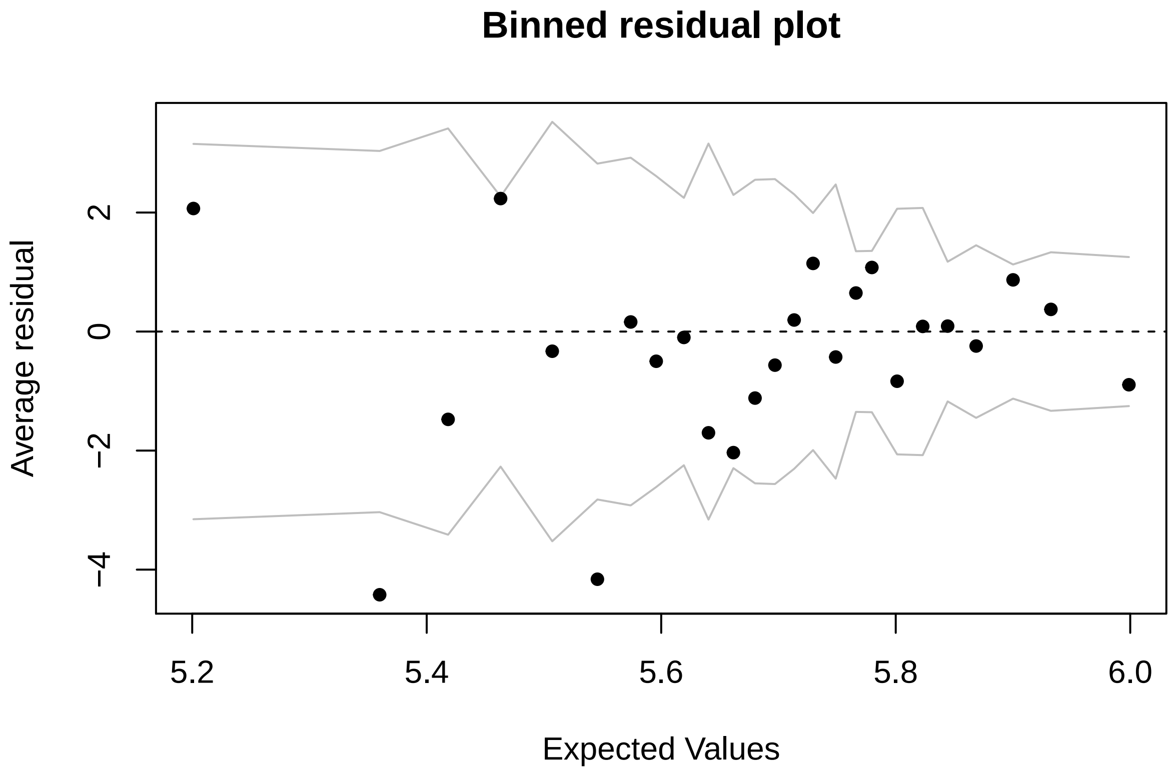

Residuals vs Fitted
Residuals
Predicted values
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G
229
591
417
20
10
0
−10
−20
5.0
5.2
5.4
5.6
5.8
6.0
6.2


Normal Q−Q
Std. Pearson resid.
−10
0
10
20
−3
−2
−1
0
1
2
3
Theoretical Quantiles
229
591
417
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G


Scale−Location
√|Std. Pearson resid.|
0
1
2
3
4
5
5.0
5.2
5.4
5.6
5.8
6.0
6.2
229
591
417
Predicted values
glm(retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean + Race + G



```
# Binned plot
x.2019 = predict(poisson2019)
y.2019 = residuals(poisson2019)
binnedplot(x.2019,y.2019)
```



Binned residual plot

Comparison between Poisson Regression and Linear Regression

Retention for Students Starting College in 2018

```
# Compare fitness of linear regression and poisson regression
# md.2018 for linear regression of 2018 retention
# poisson2018 for poisson regression of 2018 retention
anova(md.2018, poisson2018)
```

```
## Analysis of Variance Table
##
## Model 1: retention ~ Race + EverReceivedPromiseAward + CumulativeGPA +
##     AttendanceRate + Gender + SAT_Total + KeystoneMean + Num_CTE +
##     Race:EverReceivedPromiseAward
## Model 2: retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean +
##     Race + Gender + ELLStatus + IEPGroup + EconDisad + SAT_Total +
##     CumulativeGPA + MagnetInd + EverReceivedPromiseAward + QualifiedforCorePromise +
##     semester + Race * EverReceivedPromiseAward
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1    473 12986100
## 2    465    31222  8  12954877 24117 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Retention for Students Starting College in 2019

```
# Compare fitness of linear regression and poisson regression
# md.2019 for linear regression of 2019 retention
# poisson2019 for poisson regression of 2019 retention
anova(md.2019, poisson2019)
```

```
## Analysis of Variance Table
##
## Model 1: retention ~ Race + EverReceivedPromiseAward + CumulativeGPA +
##     semester + SAT_Total + Gender + ELLStatus + Race:EverReceivedPromiseAward
## Model 2: retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean +
##     Race + Gender + ELLStatus + EconDisad + SAT_Total + CumulativeGPA +
##     MagnetInd + EverReceivedPromiseAward + QualifiedforCorePromise +
##     semester + Race:EverReceivedPromiseAward
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1    622 5711445
## 2    615   23250  7   5688195 21495 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions:

- Since the p-value is very small ($< 2.2 \times 10^{-16}$), we say that poisson model fits better than linear regression model for 2018 retention.
- Since the p-value is also very small ($< 2.2 \times 10^{-16}$), we say that poisson model fits better than linear regression model for 2019 retention.
- Thus, we will employ poisson regression model for students' retention in college.

Part III: Multivariate Regression Analysis for Box Range

Data Wrangling

```
# Current GPA and attendance Cutoff: GPA 2.5, attendance 90%
# Create different cutoffs for GPA and attendance
gpa_upper = seq(2.9, 3.2, by = 0.1)
gpa_lower = seq(2.1, 1.8, by = -0.1)
attn_upper = seq(0.94, 1, by = 0.02)
attn_lower = seq(0.86,0.8, by = -0.02)
range_df = data.frame(attn_lower,attn_upper, gpa_lower,gpa_upper)
range_df
```

```
##   attn_lower attn_upper gpa_lower gpa_upper
## 1       0.86       0.94       2.1       2.9
## 2       0.84       0.96       2.0       3.0
## 3       0.82       0.98       1.9       3.1
## 4       0.80       1.00       1.8       3.2
```

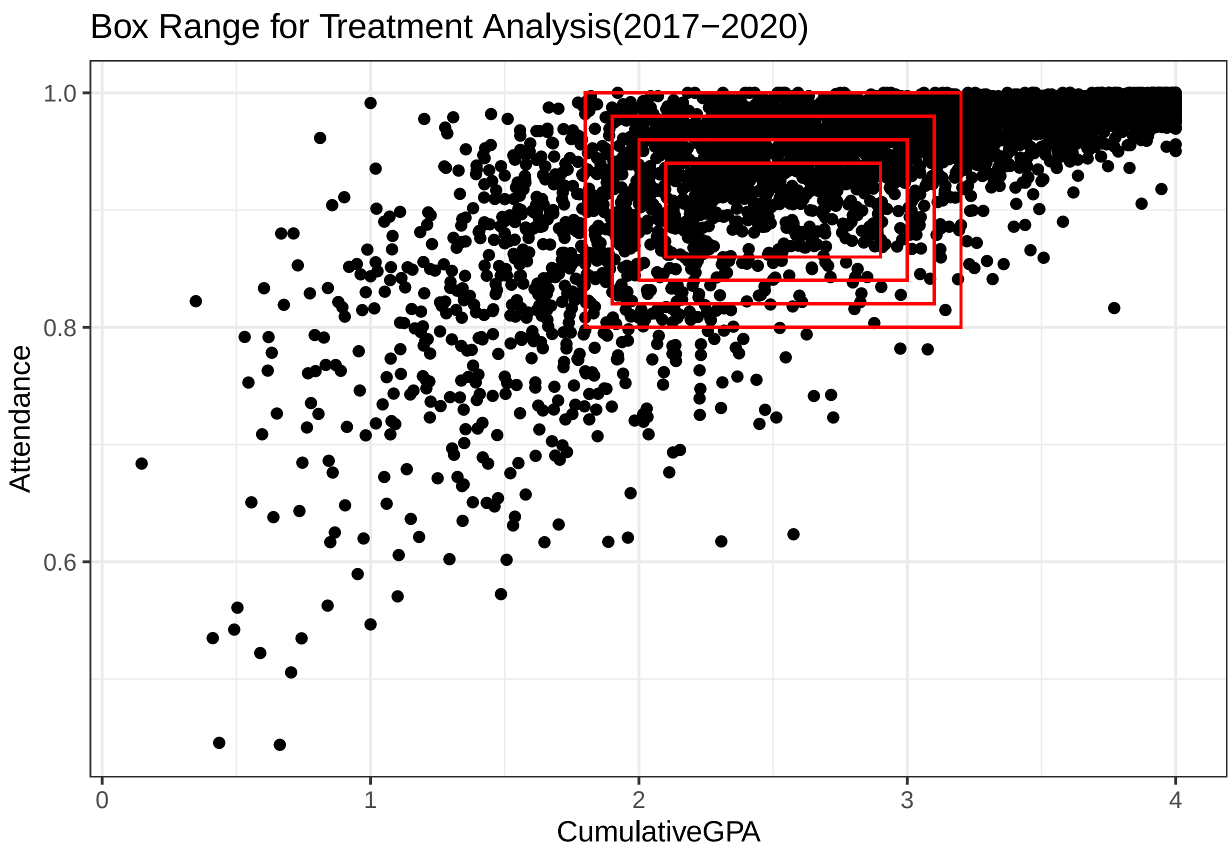
```
# Clean data
# retention_reg data is the final data set used for multivariate regression analysis
# for whole range of GPA and attendance
retention_joint = read.csv("retention_reg.csv", stringsAsFactors = T)
retention_joint$RandomID <- as.character(retention_joint$RandomID)
retention_joint <- retention_joint %>%
  mutate(Gender = ifelse(Gender == "Female",1,0),
         ELLStatus = ifelse(ELLStatus == "ELL",1,0),
         EconDisad = ifelse(EconDisad == "Free Lunch",1,0),
         QualifiedforCorePromise = ifelse(QualifiedforCorePromise == "yes",1,0)) %>%
  dplyr::select(-c(X, AttendaceRateCate, GradYear, retention)) %>%
  inner_join(nsc_retention, by = c("RandomID","ENROLLMENT_BEGIN_year", "semester")) %>%
  mutate(retention = as.numeric(retention))
head(retention_joint)
```

```
##   RandomID AttendanceRate Num_AP Num_CTE KeystoneMean             Race Gender
## 1  5841218      0.7552156      0       3     1435.357 African American      1
## 2  5849658      0.9944367      6       0     1516.250            White      0
## 3  5861017      0.9833102      0       0     1500.400 African American      0
## 4  5862641      0.9568846      0       0     1424.000 African American      0
## 5  5866142      0.9972184      3       0     1547.500            White      0
## 6  5869349      0.9860918      1       0     1464.167            White      1
##   ELLStatus          IEPGroup EconDisad SAT_Total CumulativeGPA MagnetInd
## 1         0               IEP         1       810         2.438         0
## 2         0 Not IEP or Gifted         0      1210         3.408         0
## 3         0 Not IEP or Gifted         1       970         2.117         1
## 4         0               IEP         1       740         2.856         0
## 5         0               IEP         1      1230         3.644         1
## 6         0               IEP         1       890         2.982         0
##   EverReceivedPromiseAward QualifiedforCorePromise ENROLLMENT_BEGIN_year
## 1                        1                       0                  2018
## 2                        1                       1                  2018
## 3                        0                        0                  2018
## 4                        1                        1                  2018
## 5                        1                        1                  2018
## 6                        1                        1                  2018
##   semester retention
## 1     Fall       217
## 2     Fall       548
## 3     Fall       337
## 4     Fall       544
## 5     Fall       563
## 6     Fall       724
```

```
# Prepare data sets for box range analysis
# gpa_scholarship_attendance data is the joined data set of gpa, scholarship, and attendance
joint_df = read.csv("gpa_sholarship_attendance.csv")
names(joint_df)[11] <- "present_pct"
names(joint_df)[12] <- "present_pct_cate"
joint_df$RandomID <- as.character(joint_df$RandomID)

box_df = joint_df %>%
  dplyr::select(RandomID, CumulativeGPA, excused_pct) %>%
  inner_join(nsc_promise,  by = "RandomID") %>%
  filter(CumulativeGPA >=2, CumulativeGPA<=3) %>%
  filter(excused_pct >=0.85, excused_pct<=0.95)
```

EDA

```
# Visualize the box ranges
ggplot(joint_df, aes(x =CumulativeGPA, y= excused_pct)) +
  geom_point() +
  labs(y = "Attendance") +
  geom_rect(aes(xmin = 2.1, xmax = 2.9, ymin = 0.86, ymax = 0.94), color = "red", alpha = 0)+
  geom_rect(aes(xmin = 2, xmax = 3, ymin = 0.84, ymax = 0.96), color = "red", alpha = 0)+
  geom_rect(aes(xmin = 1.9, xmax = 3.1, ymin = 0.82, ymax = 0.98), color = "red", alpha = 0)+
  geom_rect(aes(xmin = 1.8, xmax = 3.2, ymin = 0.8, ymax = 1), color = "red", alpha = 0) +
  theme_bw() +
  labs(title = "Box Range for Treatment Analysis(2017-2020)")
```



```
# Exclude semester and IEPGroup because of insufficient factor levels
get_box_df <- function(df, attn_lo, attn_high, gpa_lo, gpa_high){
  res = df %>%
    filter(AttendanceRate >=attn_lo, AttendanceRate<=attn_high) %>%
    filter(CumulativeGPA >=gpa_lo, CumulativeGPA<=gpa_high)
  return(res)
}
```

T-test for the treatment (EverReceivedPromiseAward)

```
# T-test for 2018
t.test.summary.2018 <- data.frame(estimate=numeric(),
                 lower=numeric(),
                 upper = numeric(),
                 pval=numeric(),
                 significant = logical())

# Conduct T-test for each box range
year = 2018
for(ind in 1:4){
  row = unlist(range_df[ind,])
  box_df = get_box_df(retention_joint,row[1], row[2],row[3],row[4])
  box_df_year = box_df %>% filter(ENROLLMENT_BEGIN_year == year)
  a = t.test(retention~EverReceivedPromiseAward, data = box_df_year)
  res = data.frame(meanDiff=round(a$estimate[2]-a$estimate[1],3),
                  lower=round(a$conf.int[1],3),
                  upper = round(a$conf.int[2],3),
                  pval=round(a$p.value,3),
                  significant = a$p.value<=0.05)
  t.test.summary.2018 = rbind(t.test.summary.2018, res)

}
rownames(t.test.summary.2018) <- paste0("Box", 1:4)
t.test.summary.2018
```

```
##      meanDiff    lower   upper  pval significant
## Box1  145.118 -291.280   1.045 0.052       FALSE
## Box2  137.080 -228.130 -46.030 0.004        TRUE
## Box3  118.876 -201.006 -36.747 0.005        TRUE
## Box4  159.944 -223.736 -96.152 0.000        TRUE
```

```
# T-test for 2019
t.test.summary.2019 <- data.frame(estimate=numeric(),
                  lower=numeric(),
                  upper = numeric(),
                  pval=numeric(),
                  significant = logical())

# Conduct T-test for each box range
year = 2019
for(ind in 1:4){
  row = unlist(range_df[ind,])
  box_df = get_box_df(retention_joint,row[1], row[2],row[3],row[4])
  box_df_year = box_df %>% filter(ENROLLMENT_BEGIN_year == year)
  a = t.test(retention~EverReceivedPromiseAward, data = box_df_year)
  res = data.frame(meanDiff=round(a$estimate[2]-a$estimate[1],3),
                  lower=round(a$conf.int[1],3),
                  upper = round(a$conf.int[2],3),
                  pval=round(a$p.value,3),
                  significant = a$p.value<=0.05)
  t.test.summary.2019 = rbind(t.test.summary.2019, res)

}
rownames(t.test.summary.2019) <- paste0("Box", 1:4)
t.test.summary.2019
```

```
##      meanDiff    lower   upper  pval significant
## Box1  110.344 -431.535 210.846 0.361       FALSE
## Box2   51.920 -164.177  60.337 0.327       FALSE
## Box3   61.896 -136.690  12.898 0.098       FALSE
## Box4   33.696  -84.802  17.409 0.188       FALSE
```

Findings:

- We see that for 2018, the treatment effect is significant for all boxes except for the first one. This is probably because of limited observations in the smallest box (24 observations).
- We see that for 2019, none of the boxes has significant treatment effect. This aligns with our box plot for the whole data set above.

Multivariate Linear Regression

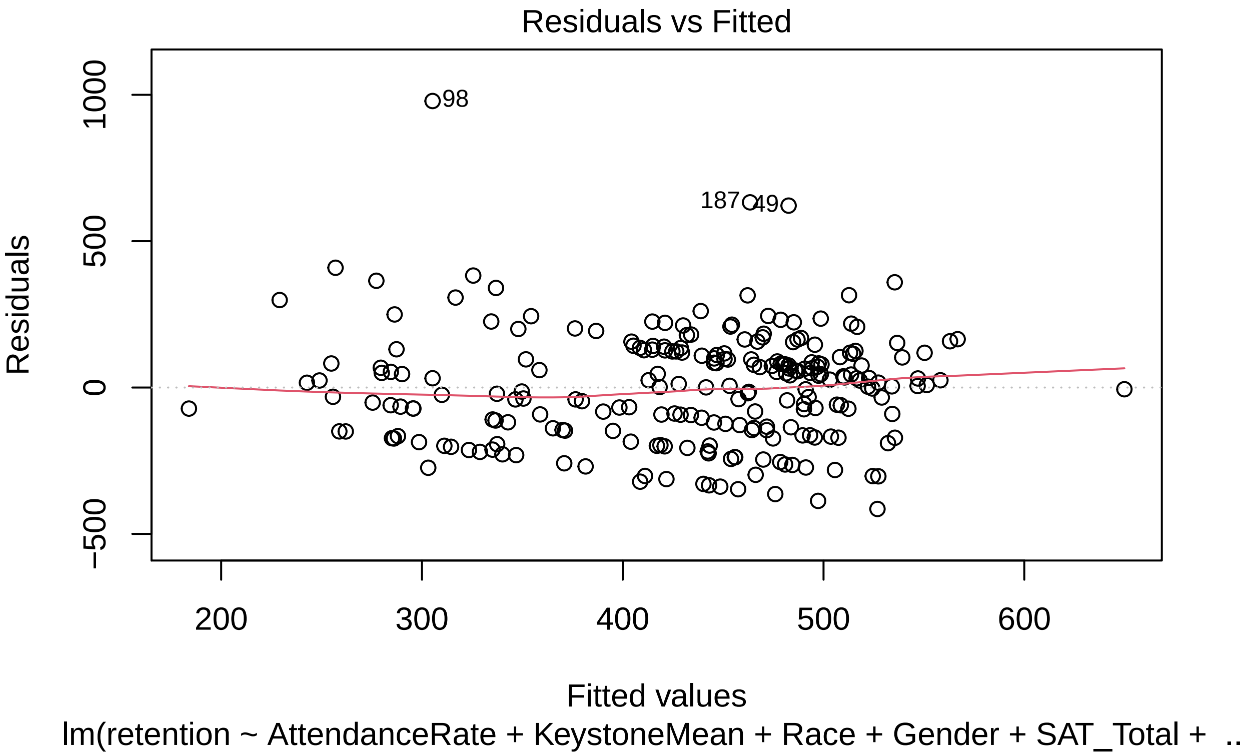
- We perform linear regression on retention with all other predicting variables.
- The linear regression serves as an extra step to verify the treatment effect of scholarship on retention.

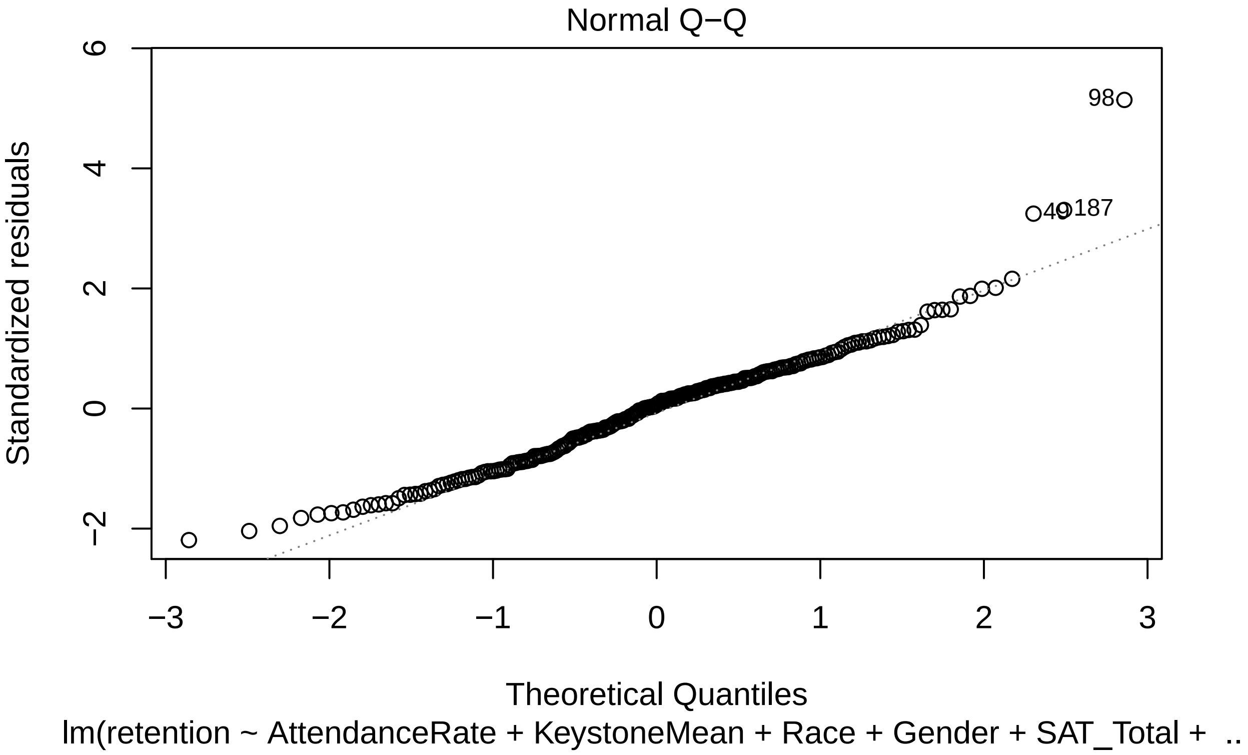
```
# Define null model and full model for multivariate linear regression
form.0 = as.formula("retention ~ Race + EverReceivedPromiseAward + Race*EverReceivedPromiseAward")
form.full = as.formula("retention ~ AttendanceRate + Num_AP + Num_CTE + KeystoneMean +
                       Race + Gender + ELLStatus + EconDisad + SAT_Total + CumulativeGPA +
                       MagnetInd + EverReceivedPromiseAward + QualifiedforCorePromise +
                       Race*EverReceivedPromiseAward")

# Create a function box_yearly to perform multivariate linear regression for each box range
box_yearly <- function(range_ind, year = 2018){
  row = unlist(range_df[range_ind,])
  box_df = get_box_df(retention_joint,row[1], row[2],row[3],row[4])
  box_df_year = box_df %>% filter(ENROLLMENT_BEGIN_year == year)
  print(paste("Number of students:", nrow(box_df_year)))
  lm.full.year = lm(form.full, data = box_df_year)
  res = step(lm.full.year, list(lower = form.0, upper = form.full), direction = "both", trace = 0)
  return(res)
}
```

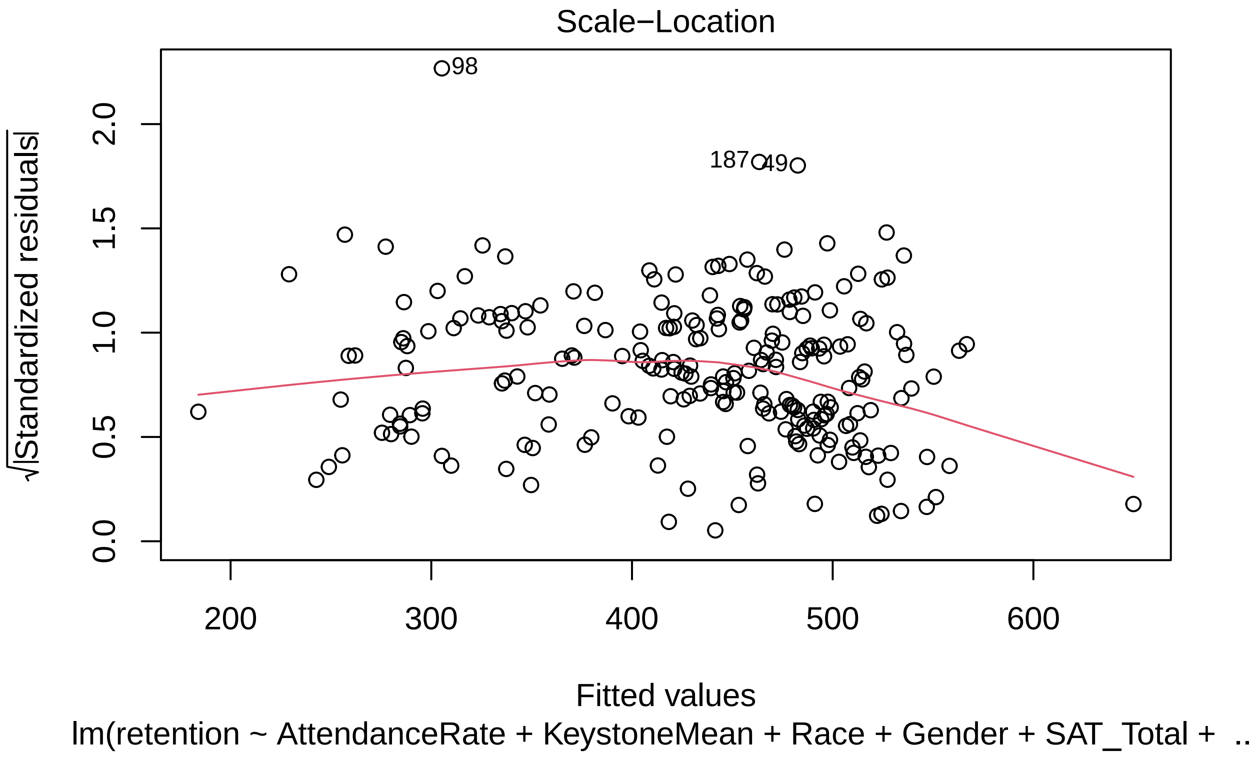
```
# Construct linear regression model for each box range of 2018 retention
res.2018 = lapply(1:4, box_yearly)
```

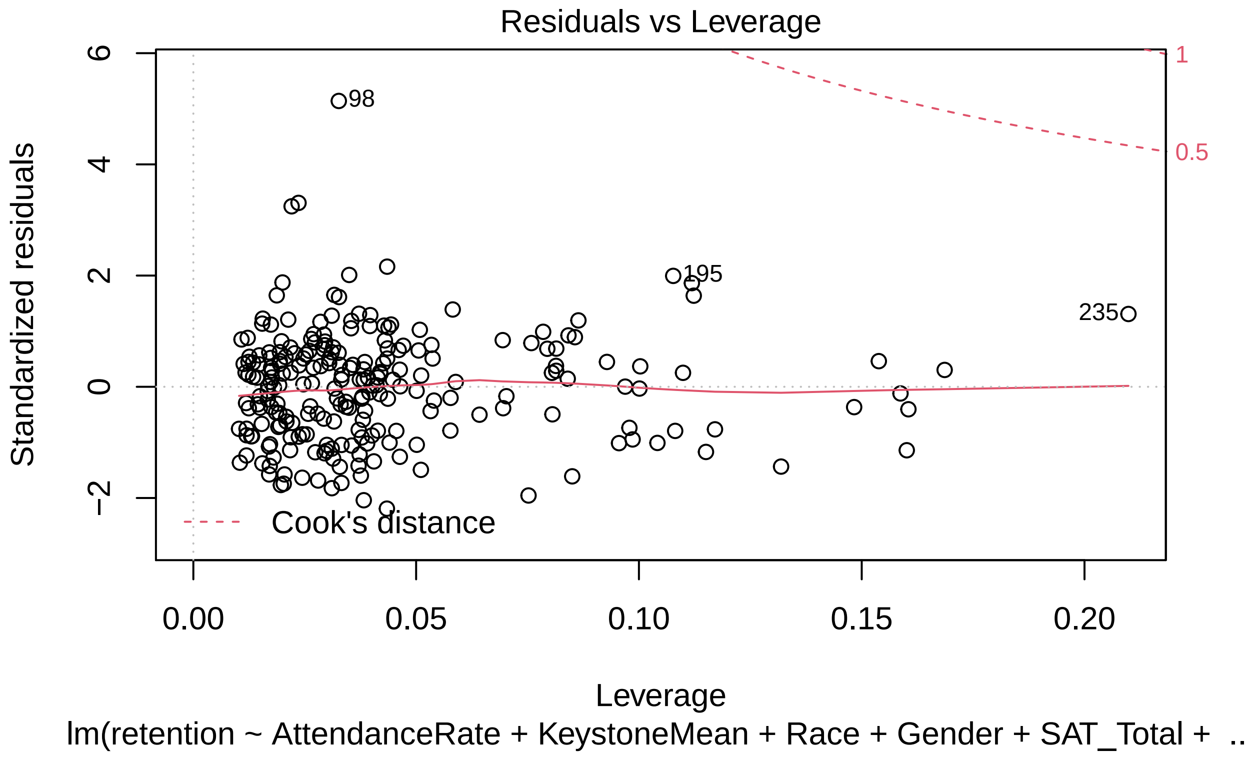
```
## [1] "Number of students: 34"
## [1] "Number of students: 75"
## [1] "Number of students: 150"
## [1] "Number of students: 235"
```

```
# Model diagnostics for the last box
plot(res.2018[[4]])
```


Residuals vs Fitted
Residuals
1000
500
0
−500
98
187
49
200
300
400
500
600
Fitted values
lm(retention ~ AttendanceRate + KeystoneMean + Race + Gender + SAT_Total + ..


Normal Q−Q
Standardized residuals
6
4
2
0
−2
98
49
187
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(retention ~ AttendanceRate + KeystoneMean + Race + Gender + SAT_Total + ..


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
98
187
49
200
300
400
500
600
Fitted values
lm(retention ~ AttendanceRate + KeystoneMean + Race + Gender + SAT_Total + ..



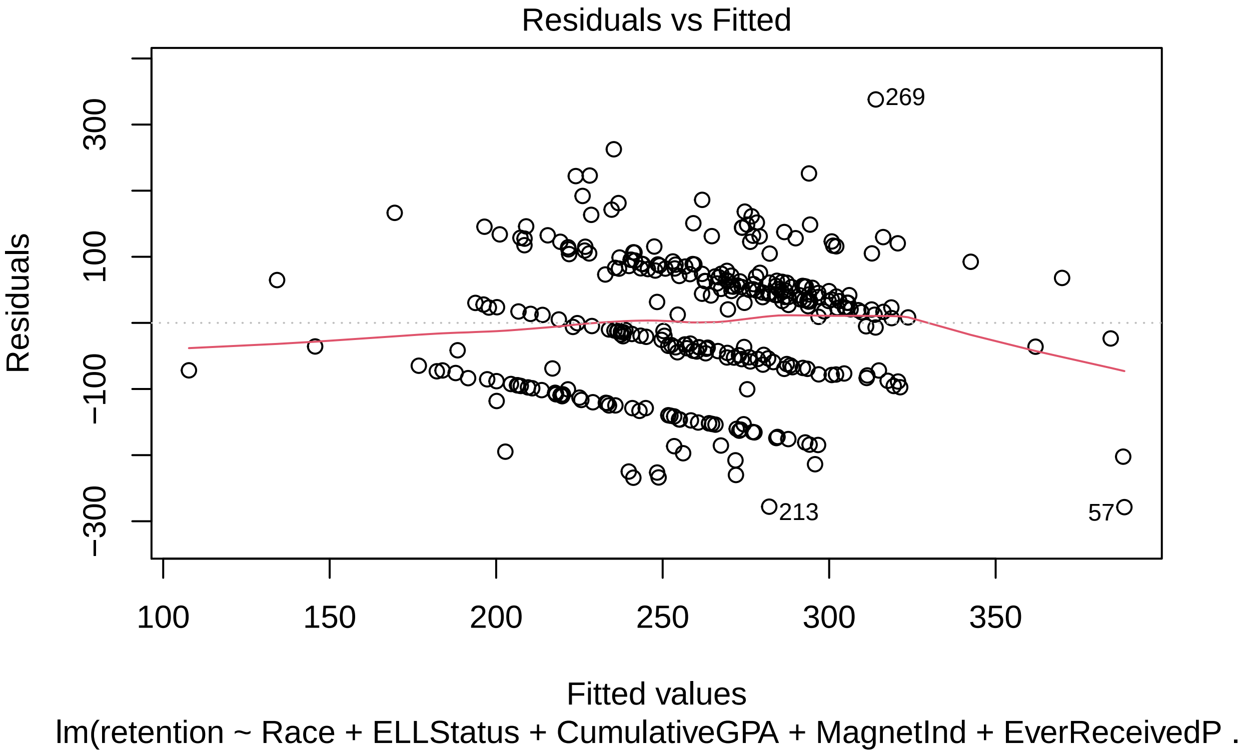
```
# Coefficient output for the last box
(round((summary(res.2018[[4]]))$coef[,"Pr(>|t|)"],3))
```

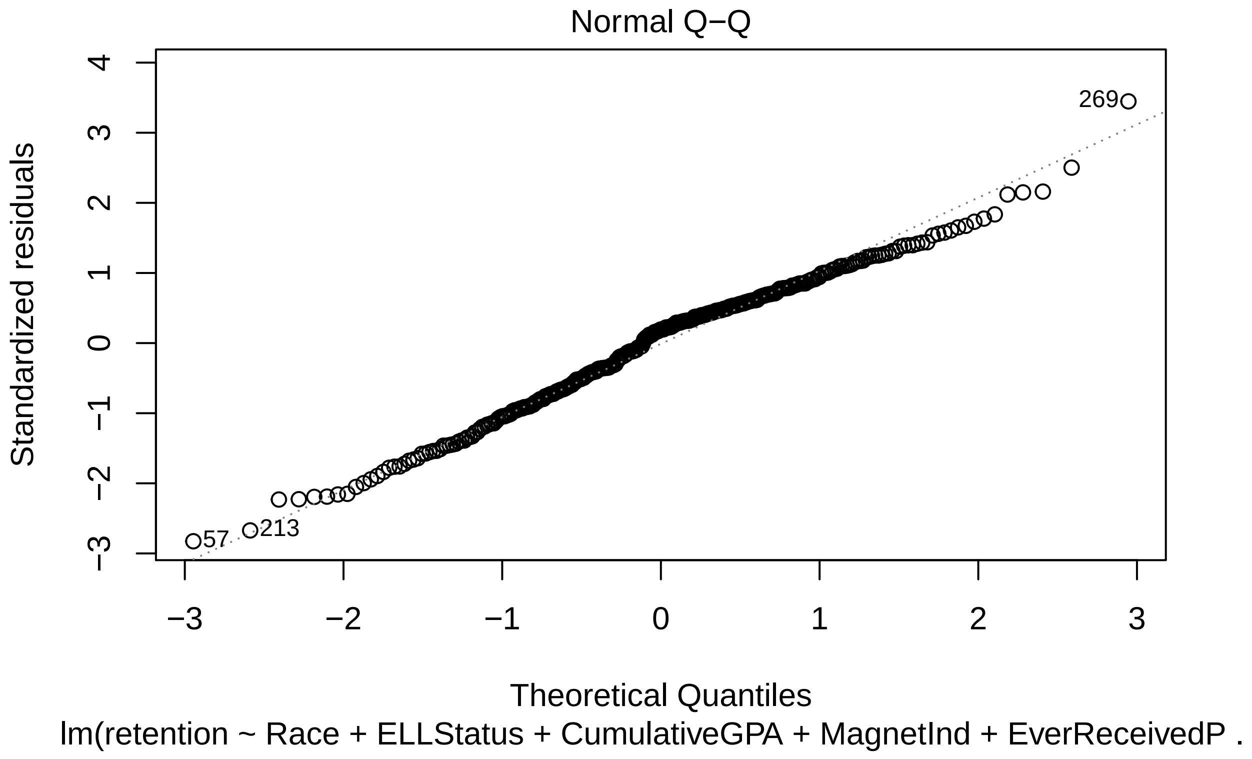
```
##                          (Intercept)                       AttendanceRate
##                                0.370                                0.063
##                         KeystoneMean                            RaceOther
##                                0.084                                0.849
##                            RaceWhite                               Gender
##                                0.613                                0.047
##                            SAT_Total            EverReceivedPromiseAward
##                                0.013                                0.000
## RaceOther:EverReceivedPromiseAward RaceWhite:EverReceivedPromiseAward
##                                0.801                                0.562
```

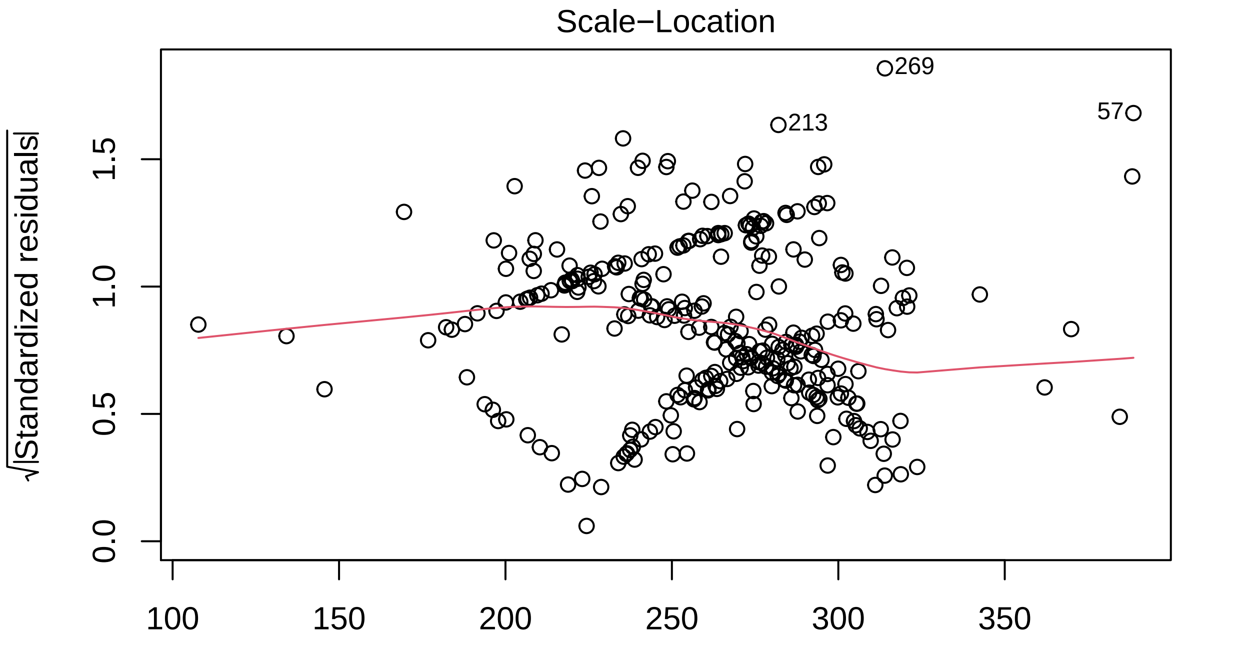
```
# Construct linear regression model for each box range of 2019 retention
res.2019 = lapply(1:4, box_yearly, year = 2019)
```

```
## [1] "Number of students: 57"
## [1] "Number of students: 110"
## [1] "Number of students: 203"
## [1] "Number of students: 311"
```

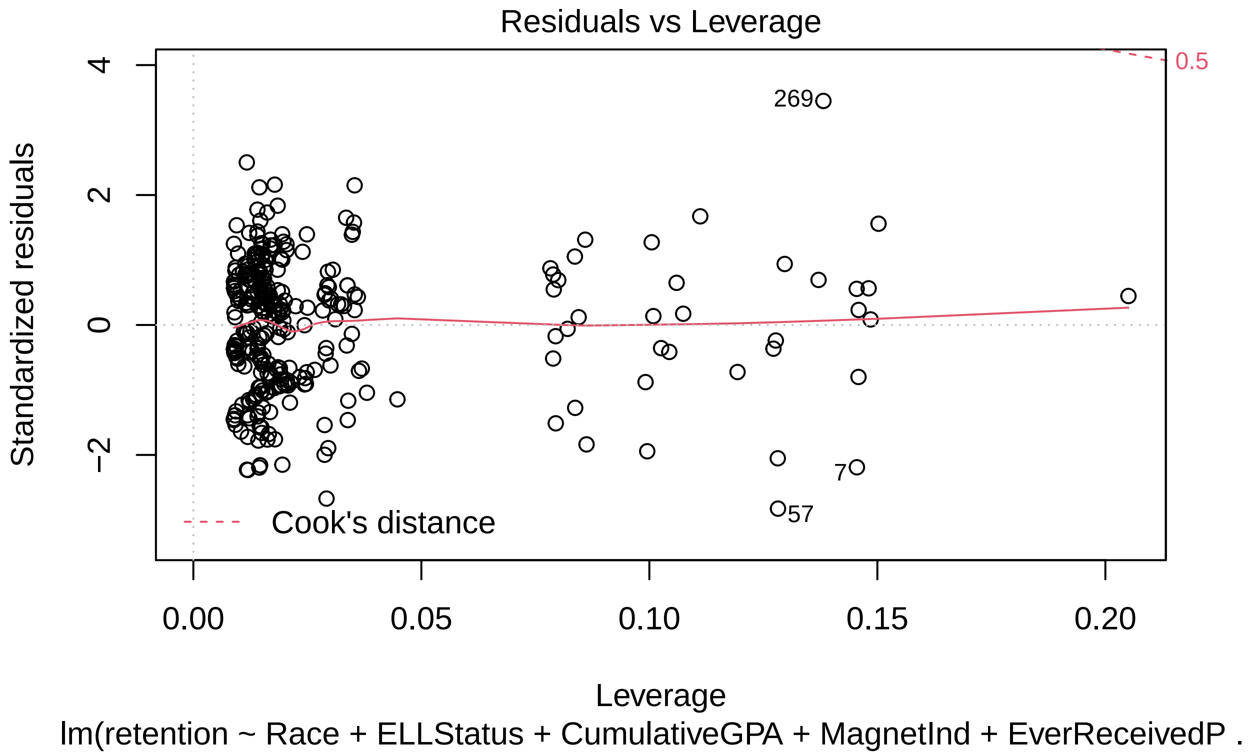
```
# Model diagnostics for the last box
plot(res.2019[[4]])
```


Residuals vs Fitted
Residuals
300
100
−100
−300
100
150
200
250
300
350
Fitted values
lm(retention ~ Race + ELLStatus + CumulativeGPA + MagnetInd + EverReceivedP .
269
213
57


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
269
57
213
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(retention ~ Race + ELLStatus + CumulativeGPA + MagnetInd + EverReceivedP .



lm(retention ~ Race + ELLStatus + CumulativeGPA + MagnetInd + EverReceivedP .



```
# Coefficient output for the last box
(round((summary(res.2019[[4]]))$coef[,"Pr(>|t|)"],3))
```

```
##                         (Intercept)                          RaceOther
##                               0.919                              0.785
##                           RaceWhite                          ELLStatus
##                               0.840                              0.054
##                       CumulativeGPA                          MagnetInd
##                               0.000                              0.054
##             EverReceivedPromiseAward RaceOther:EverReceivedPromiseAward
##                               0.381                              0.026
## RaceWhite:EverReceivedPromiseAward
##                               0.419
```

Notice:

- For simplicity, we only demonstrate the model diagnosis for the last box.
- Replacing the index "4" with 1 through 3 to see model performance and coefficients for other 3 smaller boxes.
- More results of the linear regression will be discussed in the paper.