Report: Generalized Linear Models

Group 6

Group members

1. Adhithya Unni Narayanan, r0776057
2. Cheung Wai Chun, r0817438
3. Lia Julien Azevedo Csuraji, r0824020
4. Sanne Kerckhoffs, r0667062
5. Sanya Anand, r0823086
6. Vancesca Dinh, r0830510
7. Yuanyuan Li, r0774789

Table of Content

Assignment 1

Background Information

In this study, we analyzed a data set from a survey of 1308 people, and they were asked how many homicide victims they know. The purpose is to study whether the race helps explain how many homicide victims a person knows. We defined the following variables for our study:

- resp: The number of victims the respondent knows which is a count response variable.
- race: The race of the respondent which is binary covariate with levels denoted by black and white.

From the histogram of "resp" in Figure 1, it ranges from 0 to 6 which are small integral values. Based on such observation, "resp" can be modelled as a Poisson distribution. Hence, Poisson regression model is considered as the first candidate model in our study.

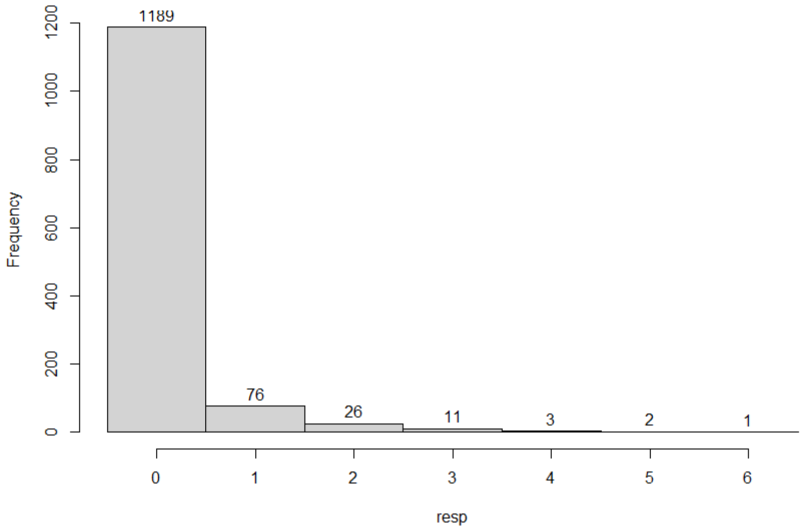

Figure 1: Histogram of resp

Poisson Regression Model

From the discussion in the previous section, the response variable "resp" is modelled with a Poisson distribution, and the systematic part of the Poisson regression model is specified as follows:

$$ln(\mathbb{E}[resp]) = \beta_0 + \beta_1 race$$

where race takes value 0 if it is categorized as white and takes value 1 if it is categorized as black.

As shown in Table 1, the covariate "race" is found to be statistically significant with an extremely small p-value. It suggests that the race of respondents is positively associated with the number of homicide victims that he/she knows.

	Estimate (Coefficients)	**Standard Error**	**P-value**
Intercept	-2.3832	0.0971	< 2e-16
Race: Black	1.7331	0.1466	< 2e-16

Table 1: Results of Poisson regression model

Risk Ratio Analysis

In this section, we presented the risk ratio of "race" with white as the reference group. The corresponding formula is given by

$$\text{Risk Ratio} = e^{\beta_1}$$

This risk ratio is also the ratio of the means of response variable for each race. Based on the results, the risk ratio is reported to be 5.6584.

The average number of homicide victims that a person with black race knows, is about 5.66 times larger than that of a person with white race. A person with black race is associated with a higher exposure to the number of homicide victims. The corresponding confidence interval for the risk ratio is calculated with both the Wald method and the profile-likelihood method. The 95% confidence interval constructed using the Wald method is a range between 4.25 and 7.54 while the 95% confidence interval constructed using the profile-likelihood method is a range between 4.24 and 7.53. Both methods yield similar results (Table 2). To conclude, we are 95% confident that the risk ratio is between 4.2 and 7.5 approximately.

Confidence Intervals

	Wald		**Profile-likelihood**	
	2.50%	**97.50%**	**2.50%**	**97.50%**
Race: Black	4.24557	7.54143	4.23633	7.53253

Table 2: Wald and profile-likelihood confidence intervals

Prediction

The prediction from the Poisson regression model for each race is summarized in Table 3. On average, the model predicts there are about 5.2 homicide victims known for every 10 people with a black race. There are about 1 homicide victim known for every 10 people with a white race.

Mean Predicted Response

Race: Black	**Race: White**	**Predicted Ratio**
0.52201	0.09225	5.65842

Table 3: Mean predicted response for black and white race

Model Diagnostics

To analyze the goodness-of-fit of the Poisson regression model, Pearson Chi-Squared test and Deviance test are firstly performed, with results shown in Table 4. Pearson Chi-Squared test suggests that the model needs to be rejected while Deviance test indicates that the fitted model is acceptable. As the results from both tests are not consistent, we resort to simulation-based method and rootogram.

Goodness-of-Fit Test

	Chi-Squared	Deviance	Df	P-value
Pearson test	2279.873	-	1306	7.074333e-56
Deviance test	-	844.707	1306	1

Table 4: Goodness of fit table for Poisson model

The rootogram is adopted to analyze whether the fitted Poisson model is suitable for the data or not. The blue line shown in Figure 2 represents the expected frequency while the bar plot represents the observed frequency. Obviously, the Poisson model overpredicts for value 1 and underpredicts for values 0, 2, 3 and 4. Hence, the model does not fit the data set well, and there is a possibility for zero inflation. Furthermore, by direct calculation, the actual proportion of zeros in the data is 90.9% but the prediction from the model is 87.3% which is 3.6% lower.

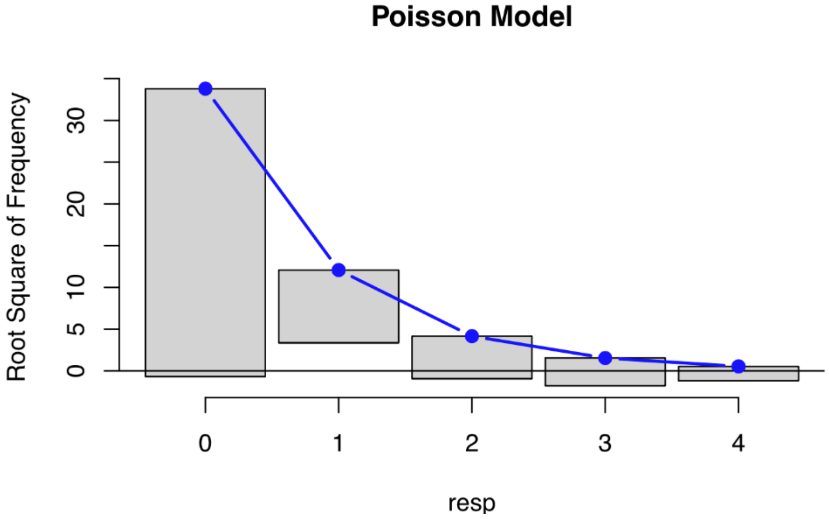

Figure 2: Rootogram of Poisson model

As mentioned, the randomized quantile residuals method is also included in the goodness-of-fit analysis. The uniformity test is performed on the simulated residuals. The QQ-plot below is used to detect overall deviations from the expected distribution, by default with added tests for the uniform distribution (KS test), dispersion and outliers. The results shown in Figure 3 shows that there is no significant deviation from uniformity while the effects of dispersion and outliers are significant.

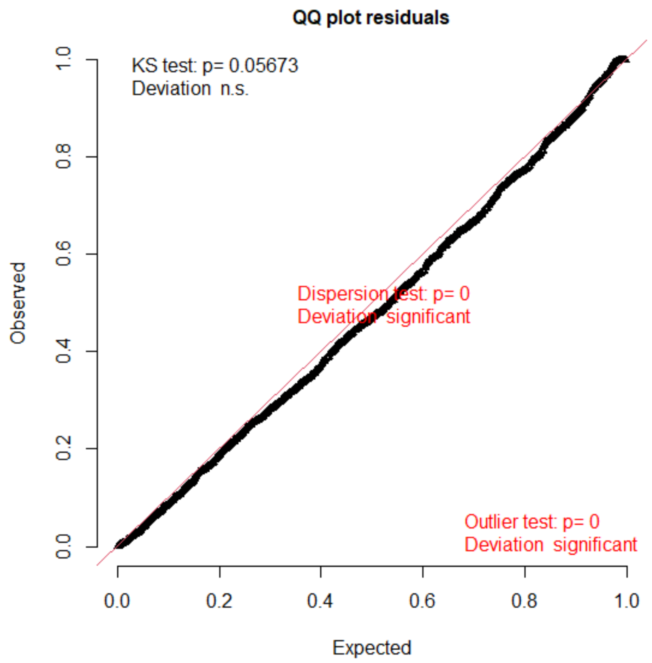

Figure 3: Randomized quantile residual plot

In Figure 4, the simulated values of standard deviation are much less than the actual standard deviation. Together with the dispersion test, we conclude that there is a problem of overdispersion.

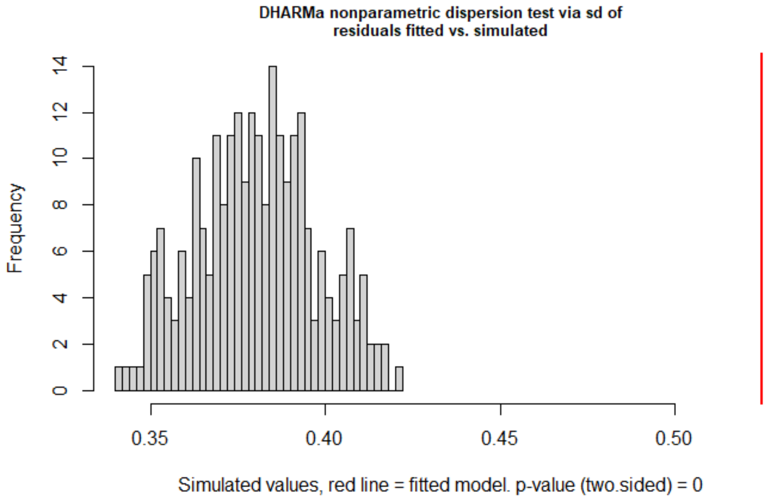

Figure 4: Overdispersion: Fitted mean deviance residuals vs simulated

In order to resolve the problem of overdispersion, three possible solutions can be taken: sandwich estimator for the covariate, negative binomial model and quasi-likelihood estimation. The results of the sandwich estimator (robust variance-covariance estimator) are shown in Table 5. It yields the same significant estimated coefficients with larger standard errors. The other two possible solutions will be discussed in the next section.

	Estimate (Coefficients)	Standard Error	P-value
Intercept	-2.3832	0.12594	< 2.2e-16
Race: Black	1.7331	0.20551	< 2.2e-16

Table 5: Results of Poisson regression model using robust covariance

Alternative Models

Negative Binomial Regression Model

A negative binomial regression model is fitted as an attempt to resolve the problem of overdispersion. It assumes a quadratic relationship between the mean and the variance. The coefficient for "race" is significant and the same as the coefficient from the Poisson regression model. However, the corresponding standard error is much larger for the negative binomial model.

	Estimate (Coefficients)	Standard Error	P-value
Intercept	-2.3832	0.1172	< 2e-16
Race: Black	1.7331	0.2385	3.66e-13

Table 6: Results of negative binomial model

The AIC of the negative binomial model is 1001.8, whereas the AIC of the Poisson model is 1122. The AIC of the negative binomial model is smaller and thus indicates a better fit of the data. This could have been expected due to overdispersion of the Poisson model. Compared to Figure 2, the rootogram for the negative binomial model (Figure 5) also shows a better fit as there are less discrepancies between the observed and the expected frequency. In particular, the rootogram does not suggest a zero-inflation problem.

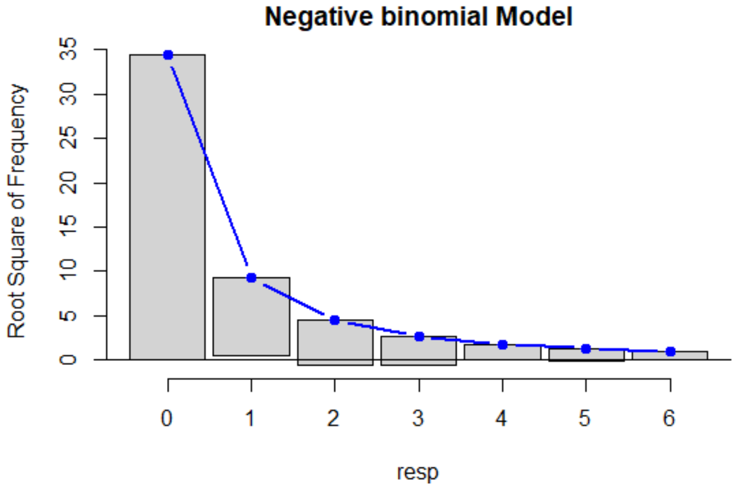

Figure 5: Rootogram of negative binomial model

The randomized quantile residuals method is also adopted for goodness-of-fit analysis. The results in Figure 6 show that there is no significant deviation from uniformity, and no significant dispersion and outliers' effects.

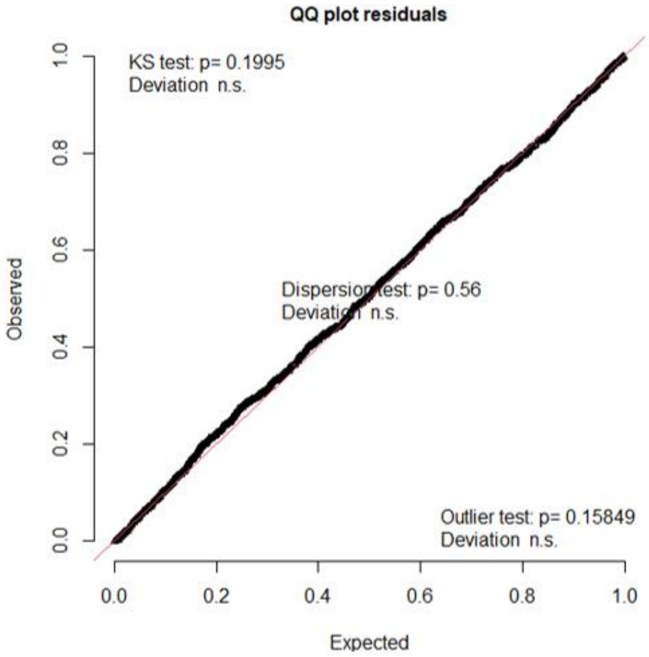

Figure 6: Randomized quantile residual plot

Figure 7 shows the comparison between fitted and simulated standard deviations, and there the fitted standard deviation does not exceed those simulated values. Together with the dispersion test, we conclude that the negative binomial model is better compared to the Poisson model because it resolved the problem of overdispersion.

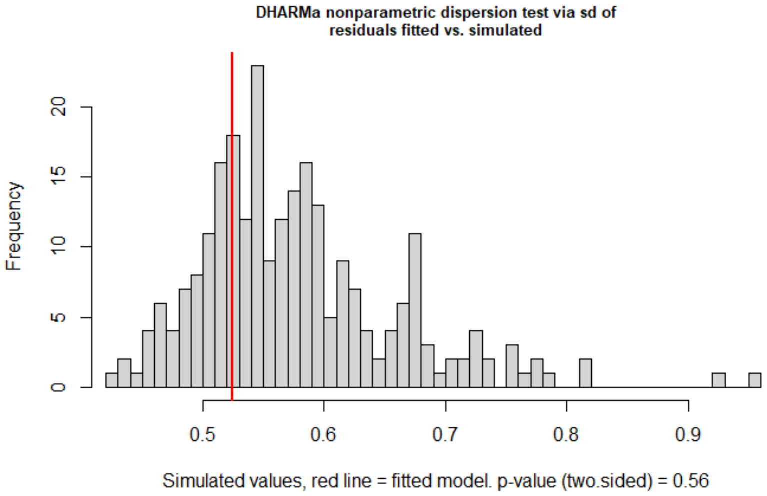

Figure 7: Overdispersion: Fitted mean deviance residuals vs simulated

At last, we compare the observed variance from the data and the variance from the negative binomial model. The observed variance for respondents with black race is lower than the estimated variance. The observed variance for respondents with a white race is slightly higher than the estimated variance. Still, there are some discrepancies between observed and estimated variances using the model.

Observed Variance Black	Observed Variance White	Estimated Variance Black	Estimated Variance White
1.1498	0.1552	1.8689	0.1343

Table 7: Observed and estimated variances for black and white race

Quasi-likelihood Poisson Model

For quasi-likelihood Poisson model, it assumes there is a linear relationship between the mean and the variance. The corresponding dispersion parameter is 1.745 which implies that the variance is 74.5% larger than the mean, and the standard error is about 45% larger.

	Estimate (Coefficients)	**Standard Error**	**P-value**
Intercept	-2.3832	0.1283	< 2e-16
Race: Black	1.7331	0.1937	< 2e-16

Table 8: Results of quasi-likelihood Poisson model

Summary

In total, four different models have been discussed. We started with a Poisson regression model, but it has a problem of overdispersion and a possibility of zero inflation. In order to resolve these problems, three possible solutions have been suggested: sandwich estimator of covariate, negative binomial model, and quasi-likelihood Poisson model.

All discussed models have the same estimated coefficients which are significant. The standard errors (Table 9) are the smallest for the Poisson model. The standard errors for the sandwich estimator of covariate and quasi-likelihood Poisson model are similar, whereas negative binomial model yields the largest standard errors.

	Poisson	**Sandwich Estimator**	**Negative Binomial**	**Quasi-Likelihood**
Intercept	0.0971	0.1259	0.1172	0.1283
Race: Black	0.1466	0.2055	0.2385	0.1937

Table 9: Comparison of the standard errors

By using AIC as a criterion, we found that the negative binomial model indicates a better fit of the data compared to the Poisson model. The negative binomial model is better compared to the Poisson model because it resolved the problem of overdispersion and zero inflation. If we consider the mean-variance relationship for the quasi-likelihood Poisson and negative binomial models, as shown in Figure 8, the quasi-likelihood Poisson model performs better in addressing the problem of overdispersion because its mean-variance relationship is closer to the observed data. For the negative binomial model, the estimated variances are much larger than the observed values. The sandwich estimator of covariance has similar standard error estimates compared to the quasi-likelihood Poisson model. Therefore, we prefer the quasi-likelihood Poisson model over the negative binomial model, and the sandwich estimator of covariance can be used as a reference.

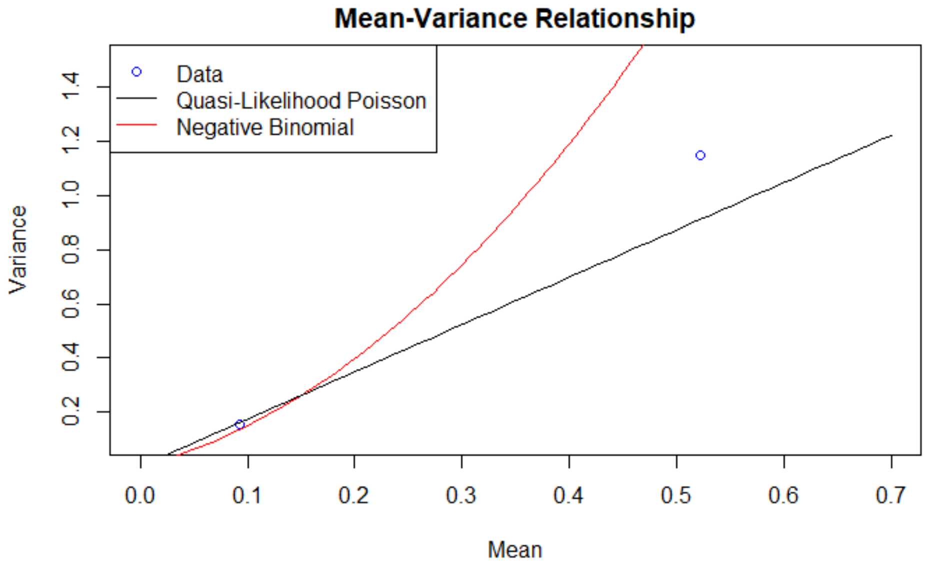

Figure 8: Mean-variance relationship for quasi-likelihood Poisson and negative binomial models

Assignment 2

Background Information

In this study, we analyzed the effects of Graduate Record Exam scores, grade point average and prestige of the undergraduate institution on the admission into the graduate schools. We defined the following variables for our study:

- GRE: Graduate Record Exam scores of the respondents which is a continuous covariate.
- GPA: Grade point average of the respondents, which is a continuous covariate.
- rank: Prestige of the undergraduate institution, which is a categorical covariate with 4 levels, denoted by rank1, rank2, rank3 and rank4.
- admit: Admission into the graduate schools which is a binary dependent variable with 2 levels: 1 for admitted and 0 for not being admitted.

The descriptive statistics are summarized in the following tables. In our study, "GRE" ranges from 220 to 800 and "GPA" ranges from 2.26 to 4.00. For "rank", all 4 categories are non-empty.

Variables	Minimum	Median	Maximum
GRE	220	580	800
GPA	2.26	3.40	4.00

Table 10: Descriptive statistics for "GPA" and "GRE"

	Rank 1	Rank 2	Rank 3	Rank 4
Count	61	151	121	67

Table 11: Descriptive statistics for "rank"

Logistic Regression Model

As "admit" is a binary outcome variable, we modelled it with Bernoulli distribution; therefore, we adopted a logistic regression model. The systematic component of the logistic model is:

$$\text{logit}(\mathbb{E}[\text{admit} = 1]) = \beta_0 + \beta_1\text{GRE} + \beta_2\text{GPA} + \beta_3\text{rank}_2 + \beta_4\text{rank}_3 + \beta_5\text{rank}_4$$

The results for the logistic regression model are summarized in Table 12. The estimated coefficients for GRE and GPA are positive and significant which imply a positive association between the admission and higher scores, respectively and holding other covariates constants. For the estimated coefficients from rank 2 through 4, they are negative and significant. Compared to rank 1, there is a negative association between the admission and other ranks.

	Estimate (Coefficients)	Standard Error	P-value
Intercept	-3.989979	1.13995	0.000465
GRE	0.002264	0.00109	0.038465
GPA	0.804038	0.33182	0.015388
Rank 2	-0.675443	0.31649	0.032829
Rank 3	-1.340204	0.34531	0.000104
Rank 4	-1.551464	0.41783	0.000205

Table 12: Results of the logistic regression model

The profile-likelihood and Wald confidence intervals for the coefficients are reported in Table 13. Both methods yield similar confidence intervals for all covariates. In the section of odds ratio analysis, we will interpret these coefficients in terms of odds ratio.

Confidence Intervals

	Profile-likelihood		Wald	
	2.50%	**97.50%**	**2.50%**	**97.50%**
Intercept	-6.27162	-1.79255	-6.22424	-1.75572
GRE	0.00014	0.00444	0.00012	0.00441
GPA	0.16030	1.46414	0.15368	1.45439
Rank 2	-1.30089	-0.05675	-1.29575	-0.05513
Rank 3	-2.02767	-0.67037	-2.01699	-0.66342
Rank 4	-2.40003	-0.75354	-2.37040	-0.73253

Table 13: Profile-likelihood and Wald confidence intervals

Effects of Rank

As “rank” is a categorical variable, we conducted the likelihood ratio test for the overall effect of “rank”. It is equivalent to compare the reduced model which does not include “rank” against the full model. As shown in Table 14, the effect of “rank” is significant with a small p-value. Hence, we conclude that "rank” has an overall effect in our study.

	Degree of Freedom	Deviance	P-value
Reduced model	397	480.34	-
Full model	394	458.52	0.00007

Table 14: Likelihood ratio test for “rank”

Apart from the overall effect of "rank", we also want to know if the coefficient for rank 2 is equal to the coefficient for rank 3. In other words, we constructed the following hypotheses:

$$H_0 : \beta_3 = \beta_4$$
$$H_A : \beta_3 \neq \beta_4$$

for which H_0 represents the null hypothesis and H_A represents the alternative hypothesis. Under H_0, $\beta_3 - \beta_4 = 0$, and confidence intervals are constructed based on the estimator for $\beta_3 - \beta_4$. The hypothesis testing is performed using Wald's test and bootstrapping, and the corresponding results are summarized in Table 15. Although the numerical values are not all the same, these confidence intervals show consistent results. As the confidence intervals do not cover 0, the null hypothesis is rejected at 5% level of significance. We conclude that, relative to rank 1, the effect of rank 2 is greater than that of rank 3.

Confidence Intervals

	2.50%	**97.50%**
Wald's test	0.1095	1.2201
Bootstrap - Normal	0.0843	1.2178
Bootstrap - Basic	0.0588	1.1703
Bootstrap - Percentile	0.1593	1.2707
Bootstrap - BCa	0.1572	1.2572

Table 15: Confidence intervals using Wald's test and bootstrapping

In our study, the bootstrap framework has been repeated for 1000 times. The corresponding bootstrapped statistics are summarized in Figure 9. Based on the histogram, the distribution for bootstrapped statistics seems to be symmetric. From the QQ-plot, there may be an issue with heavy tails, but it does not deviate much from the normal distribution in general.

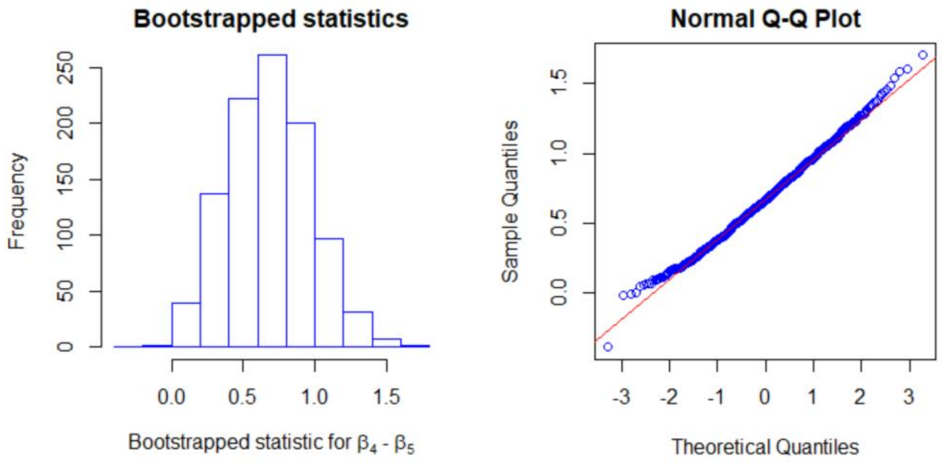

Figure 9: Distribution of bootstrapped statistics and its normal QQ-plot

Odds Ratios Analysis

In this section, we discuss and interpret the odds ratios associated with each covariate. Odds ratio is a ratio of two odds in which odds is a ratio of the probability that the event occurred to the probability that the event does not occur. Odds ratio describes how the odds change given a unit change in a covariate, holding other covariates fixed. Odds ratio is given by the following formula:

$$\text{OR for i-th covariate} = e^{\beta_i}$$

A unit increase in GRE score is associated with about 0.23% increase in the odd of being admitted. In other words, 10 units increase in GRE score is associated with 2.29% increase in odds of being admitted with 95% confidence interval ranges from 0.14% to 4.54%.

A unit increase in GPA score is associated with about 123% increase in the odd of being admitted. In other words, 0.1 unit increase in GPA score is associated with 8.37% increase in the odd of being admitted with 95% confidence interval ranges from 1.62% to 15.8%.

Relative to the highest prestige undergraduate institution (i.e. rank 1), rank 2 is associated with 49.1% decrease in the odd of being admitted, and rank 3 is associated with 73.8% decrease in the odd of being admitted, and rank 4 is associated with 78.8% decrease in the odd of being admitted.

	Odds ratio	**2.50%**	**97.50%**
GRE	1.002267	1.000138	1.004446
GPA	2.234545	1.173858	4.323835
Rank 2	0.508931	0.272290	0.944834
Rank 3	0.261792	0.131642	0.511518
Rank 4	0.211938	0.090716	0.470696

Table 17: Odds ratio for covariates

Prediction

In this section, we reported the predicted probabilities for each rank using the logistic regression model with GRE and GPA scores held at their means.

The predicted probability of getting an admission in a graduate program for a student from the highest prestige undergraduate institution i.e., rank 1, is approximately 52% on average. For the students from the lowest ranked institutions (rank 4) have the lowest predicted probability of 18.5% of getting accepted into the graduate program. The predicted probability of getting admission into the graduate school decreases for the students that belong to lower ranking institutions. The predicted probability decreases drastically from rank 1 to rank 2 (about 16.4%) and from rank 2 to rank 3 (13.4%) but mildly from rank 3 to rank 4 (about 3.4%). It suggests that whether a student comes from an undergraduate institution at rank 3 or rank 4 makes less difference on the probability of admission into the graduate school, holding other covariates as constants.

	Predicted Probabilities
Rank 1	0.5166016
Rank 2	0.3522846
Rank 3	0.2186120
Rank 4	0.1846684

Table 18: Predicted probabilities for "rank"

Model Diagnostics

To assess the goodness-of-fit for the logistic regression model, we conducted Hosmer-Lemeshow test which is designed for ungrouped data. As shown in Table 19, the p-value is insignificant which fails to reject the null hypothesis. Therefore, we conclude that the logistic regression model fits the data.

Chi-Square	P-value
11.08547	0.196903

Table 19: Hosmer-Lemeshow test result

To assess the possibility of the existence of influential data points, we examined three commonly used measures, namely: delta chi-square, delta deviance and Cook's distance. The delta chi-square values are plotted against the case numbers. In Figure 10, we found that delta chi-square values for observations 40, 156, 198, and 342 are far-away from the rest of delta chi-square values.

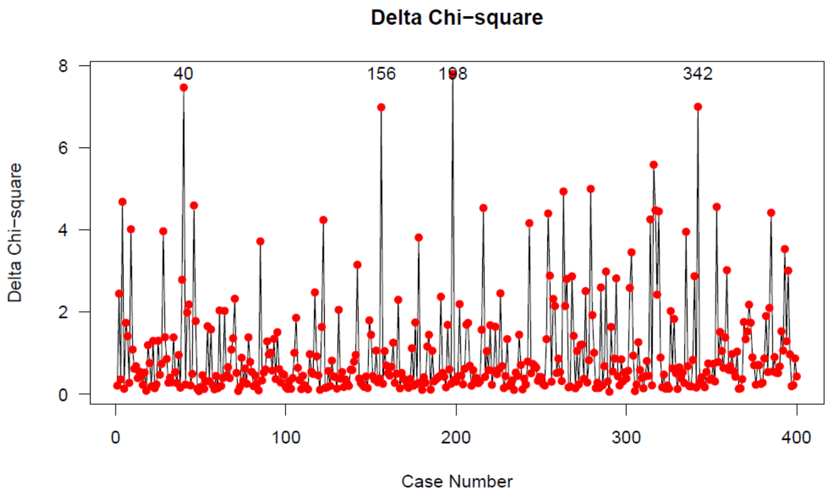

Figure 10: Delta chi-square values against case numbers

Similarly, the delta deviance values are plotted against the case numbers as shown in Figure 11. We found that observations 40, 156, 198, and 342 delta values are still the 4 highest spikes, but the difference among them and the rest of delta deviance values is less obvious. Similar findings for Cook's distance plot in Figure 12.

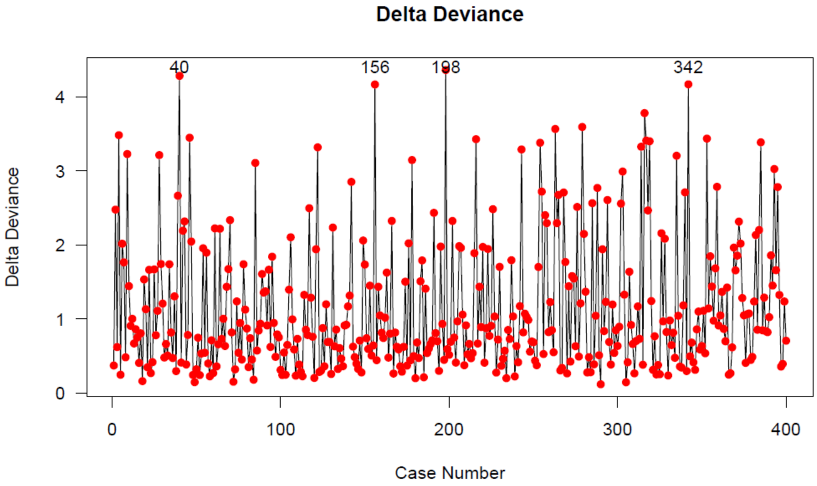

Figure 11: Delta deviance values against case numbers

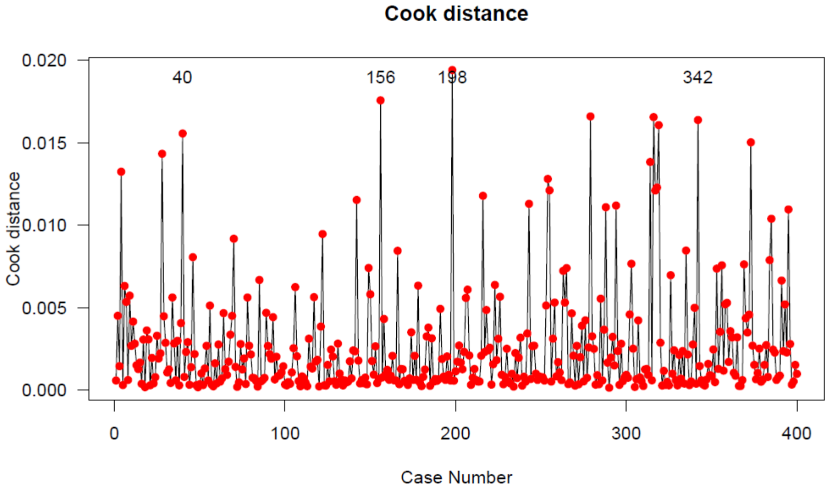

Figure 12: Cook's distance values against case numbers

To summarize, there is no obvious evidence for influential observations. Although there are few observable spikes in the delta chi-square plot, they are not extreme in other plots.

To assess the possibility of higher order terms, we investigated the residuals plot against the continuous covariates and applied lowess curve to explore the structure of the residuals. The blue curve denotes lowess curve of residuals and the red curves are the corresponding 95% confidence intervals. In Figure 13, we found that the lowess curve and its confidence intervals deviate from the horizontal line at 0 for "GPA". Still, the lowess curve is fluctuating around the horizontal line in general. The corresponding lowess curve for "GPA" is shown in Figure 14, we found that there are some patterns for “GPA”. However, it remains unclear to us whether there is a quadratic effect from “GPA”.

In Figure 15, the lowess curve for “GRE” does not deviate from the horizontal line, and in Figure 16, there are no special patterns found in the lowess curve. For “GRE”, we did not find evidence for including higher order terms.

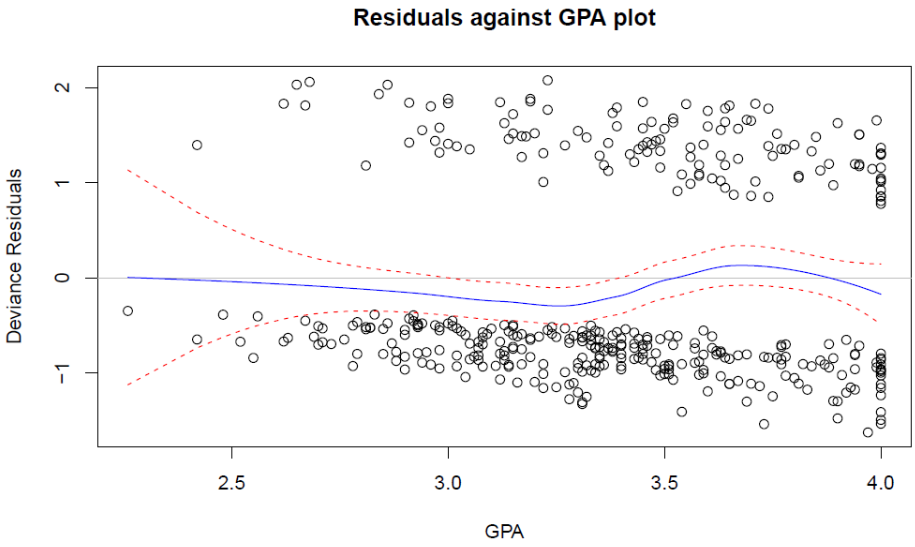

Figure 13: Deviance residuals against GPA

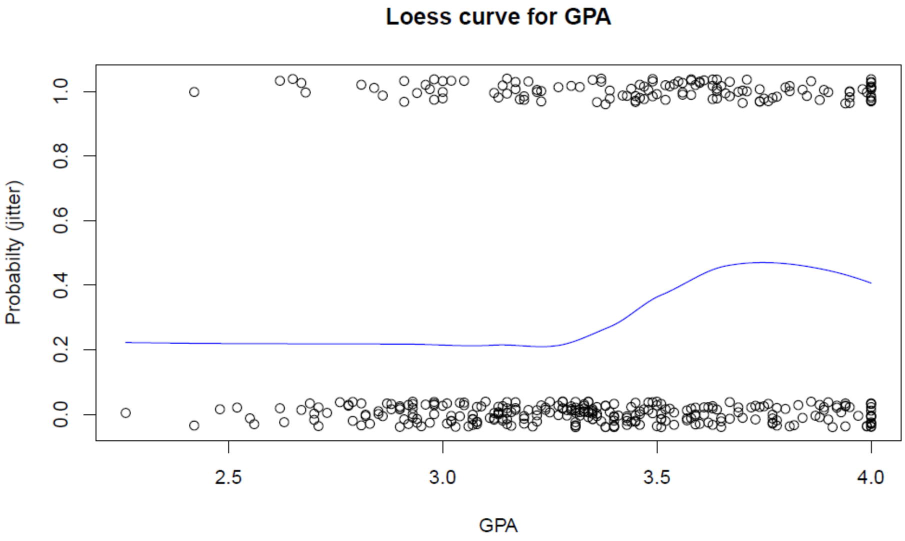

Figure 14: Lowess curve for GPA

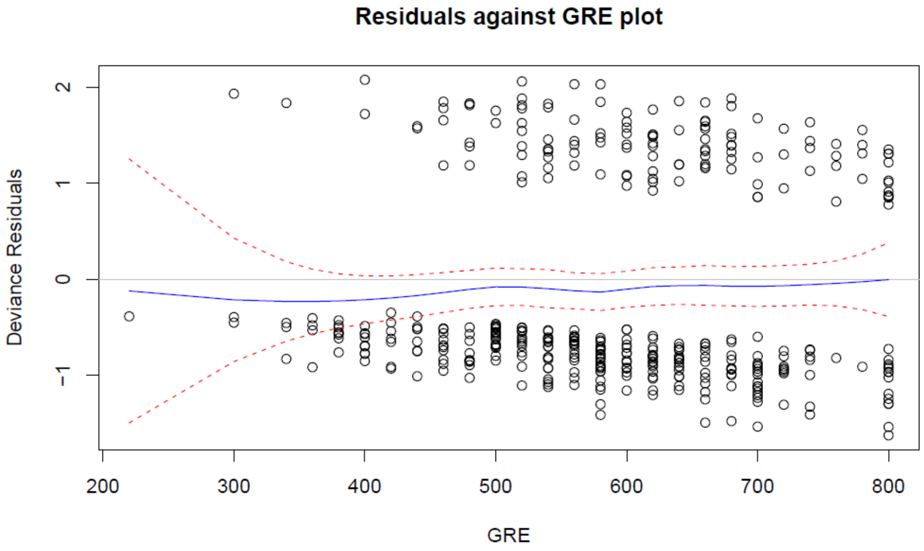

Figure 15: Deviance residuals against GRE

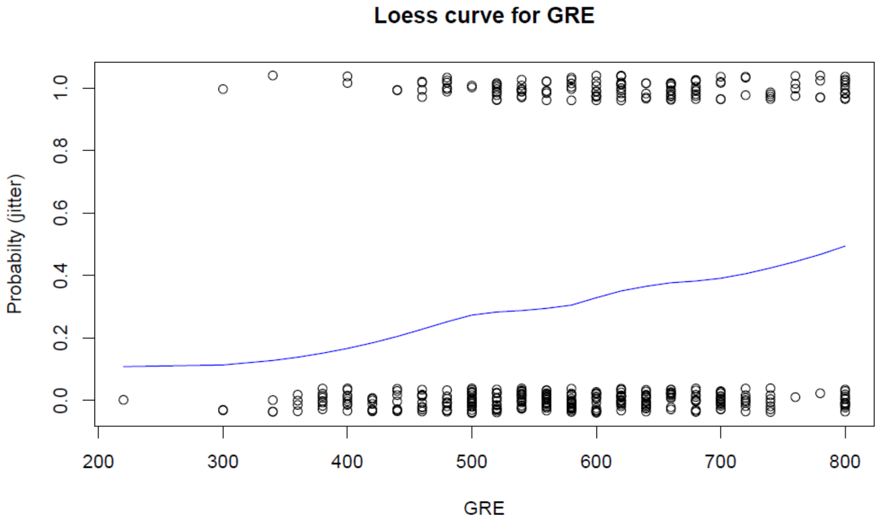

Figure 16: Lowess curve for GRE

Summary

The logistic regression model is used in our study to model binary response outcome. We found that the continuous covariates, namely “GPA” and " GRE”, are significant and positively associated with the outcome, and the categorical covariate, “rank”, is significant and negatively associated with the outcome with reference to highest ranked institution. Moreover, we found the overall effects of “rank” are significant and further discussed the prediction at different ranks. Furthermore, we discussed the interpretations of model coefficients in terms of odds ratio. The report is ended with model diagnostics section from which we did not find evidence for lack-of-fit, outliers and higher order terms.