

# Factors Affecting Graduate Admission Decisions for Top Computer Science Programs

*Superficial Intelligence (Nicholas Tanabe, Haodong Liu, Huiwen He, Xiaoyi Zhang)*

## Abstract

In this paper, we use data science to analyze graduate admissions data from Grad Cafe. We aim to better understand what factors are most important for graduate admissions and focus our analysis on the top 10 graduate computer science programs. We fit a logistic model with undergraduate GPA, GRE score, and student status as covariates. The model was not very accurate in prediction, with a McFadden  $R^2$  of .000. This likely means that there are other factors that were not represented in the data such as research experience or recommendations are more important to admission decisions than GPA and GRE scores alone.

## Introduction

Graduate school admission can be a very mysterious and esoteric process. While undergraduates and prospective graduate students constantly stress about whether their GPAs or GRE scores are “good enough” to get into a top graduate program, there are many potential factors that may be important to the final decision in a holistic admission process. In this paper, we aim to use data science to better understand what factors are most important to graduate admissions. Through our research we aim to answer several questions of interest:

- “What are the most important factors affecting graduate admissions decisions?”
- “Are international students held to a different standard in graduate admissions?”
- “Can we use data to model probabilities of acceptance?”

As students of data science and prospective graduate students, we will be focusing our analysis on top computer science programs. We start with exploratory data analysis to better understand the distributions and covariations of our variables of interest, and later aim to model admission probabilities for the top 10 graduate computer science programs using logistic regression. #The Dataset

For our analysis we will be using data on graduate admissions from GradCafe, a forum that allows students to submit their graduate school admissions decisions and details regarding their scores. A cleaned version of the dataset is provided by Debarghya Das on GitHub. The graduate school admission results database includes admission results and detailed student test scores, self-reported by prospective graduate students on <https://www.thegradcafe.com/>. The full dataset contains 345,303 observations and 19 variables with a mix of continuous and categorical data, but we will be limiting our analysis to the top 10 US graduate computer science programs, as ranked by US News. The dataset contains the following variables:

1. **rowid (integer)** - An integer id of the row.
2. **uni\_name (character)** - The name of the university.
3. **major(character)** - The subject of the program self-reported by students.
4. **degree (character)** - The type of degree program. The variable takes one of the following values: MS, MA, PhD, MFA, MBA, MEng, and Other.
5. **season (character)** - The season of application. The first letter indicates whether the program starts from the Fall semester or Spring semester, and then the letter is followed by the last 2 digits of the year the program starts.

6. **decision (character)** - The admission decision. Contains five categories - Accepted, Rejected, Wait-listed, Interview and Other.
7. **decision\_method (character)** - The method through which decision was communicated.
8. **decision\_date (character)** - The date that the decision was communicated.
9. **decision\_timestamp (integer)** - Timestamp of the decision.
10. **ugrad\_gpa (double)** - The respondent's undergraduate GPA. The scale of the GPA varies because some students use a 10-point scale while others use a 4-point scale.
11. **gre\_verbal (double)** - GRE verbal score, which varies from 130 to 170 for the new GRE and from 200 to 800 for the old GRE.
12. **gre\_quant (double)** - GRE quantitative score, which varies from 130 to 170 for the new GRE and from 200 to 800 for the old GRE.
13. **gre\_writing (double)** - GRE writing score that ranges from 0 to 6.
14. **is\_new\_gre (logical)** - Whether or not the applicant took the new GRE.
15. **gre\_subject (double)** - GRE subject test score on a 200 to 990 score scale.
16. **status (character)** - Status of the candidate. Can be "International", "International with US Degree", "American" or "Other".
17. **post\_data (character)** - The date in which the observation was posted on grad cafe.
18. **post\_timestamp (integer)** - Timestamp of the post.
19. **comments (character)** - Applicants' comments.

We decided to drop variables which either contain little information such as 'gre\_subject', which few candidates reported, and 'rowid' which is redundant, and variables which are not of interest to us, such as 'comments', 'decision\_method', 'post\_data', and 'post\_timestamp'. It is also important to note the limitations of the data. The dataset contains many missing values and may be biased data to self-reporting. In addition to this, there are several other factors that are widely acknowledged as being important in graduate admissions such as research experience, recommendations, and more that are not present in the data. While we tried to scrape the comments variable for relevant keywords such as "research experience," the majority of observations did not include relevant information in the comment field or did not include any comments at all.

## Exploratory Data Analysis

Table 1: Number of Reported Applications and Acceptance Rates  
for Top 10 Computer Science Programs

Rank	University Name	# Accepted	# Applied	Acceptance Rate
1	Carnegie Mellon University (CMU)	534	1427	37.4%
2	Massachusetts Institute Of Technology (MIT)	136	600	22.7%
3	Stanford University	246	927	26.5%
4	University Of California, Berkeley (UCB)	158	860	18.4%
5	University Of Illinois, Urbana-Champaign (UIUC)	367	957	38.3%
6	Cornell University	257	678	37.9%
7	University Of Washington, Seattle (UW)	169	714	23.7%
8	Georgia Institute Of Technology (GTech)	413	982	42.1%
9	Princeton University	94	388	24.2%
10	University Of Texas, Austin (UT Austin)	351	899	39.0%

First, we look at the top ten graduate programs in Computer Science as ranked by US News & World Report, which provides various rankings for US and international colleges. The full ranking can be observed at the following URL: <https://www.usnews.com/best-graduate-schools/top-science-schools/computer-science-rankings>. From Table 1, we can observe that there may be some bias in the reporting of the data, as the Acceptance Rates seem higher than those typically quoted for these programs. This could possibly mean that applicants that were accepted are more likely to report their results on GradCafe than those that were not. We also observe that acceptance rates for graduate programs seem to be higher than those of undergraduate programs at the same institution. For example, Stanford University has an undergraduate acceptance rate of 5%, yet the data for graduate programs in CS shows an acceptance rate of 26.5%.

Next, we look at the change in number of applications over time. The dataset has official data reported from 2006 to 2015. From Figure 1, we see that the number of applications for Master of Science in CS related programs have gradually trended up over the time period, while the number of PhD slightly drops in 2014. An increase in the number of applications with a limited number of open slots for admissions can lower acceptance rates and change standards for graduation. However, we believe that this increase over time can be attributed to an increase in responses on GradCafe.

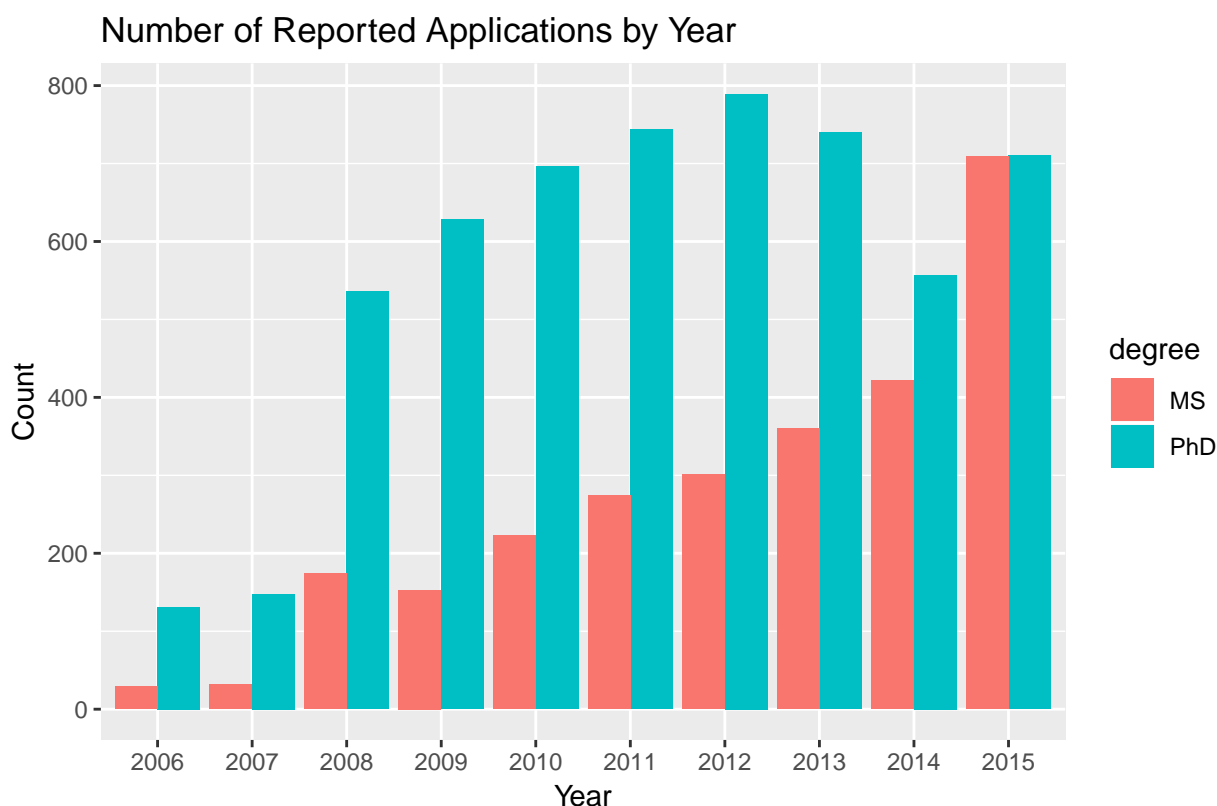


fig 1. Overall trend in the number of reported applications per year increasing

Next, we plot in Figure 2 the histograms of GRE test scores, broken down between GRE Verbal and GRE Quant, GRE Writing scores, and undergraduate GPA. In 2011, the GRE exam had a change in format, leading to a new grading scale. The dataset contains a variable “is\_new\_gre”, which distinguishes between old and new GRE scores, so we filter for only new GRE scores, as the majority of observations report new GRE scores. We see from Figure 3 that GRE verbal scores range from 130 to 170 with a bell shape, mostly concentrated between 155 - 160. GRE quant score are left skewed with a range of 130-170, with a large number of students getting a perfect score of 170. GRE writing scores range from 2 to 6 with a bell like shape, with most students getting a score of 4. We also see that the distribution of GPAs for the observations tend to be left skewed, with the majority of candidates having more than 3.6 GPA. This is accepted as grad programs tend to look at GPA as a major factor, and students who aim to attend a grad school would likely have higher GPAs. It can clearly be seen that the distributions of these variables are non-normal, which will

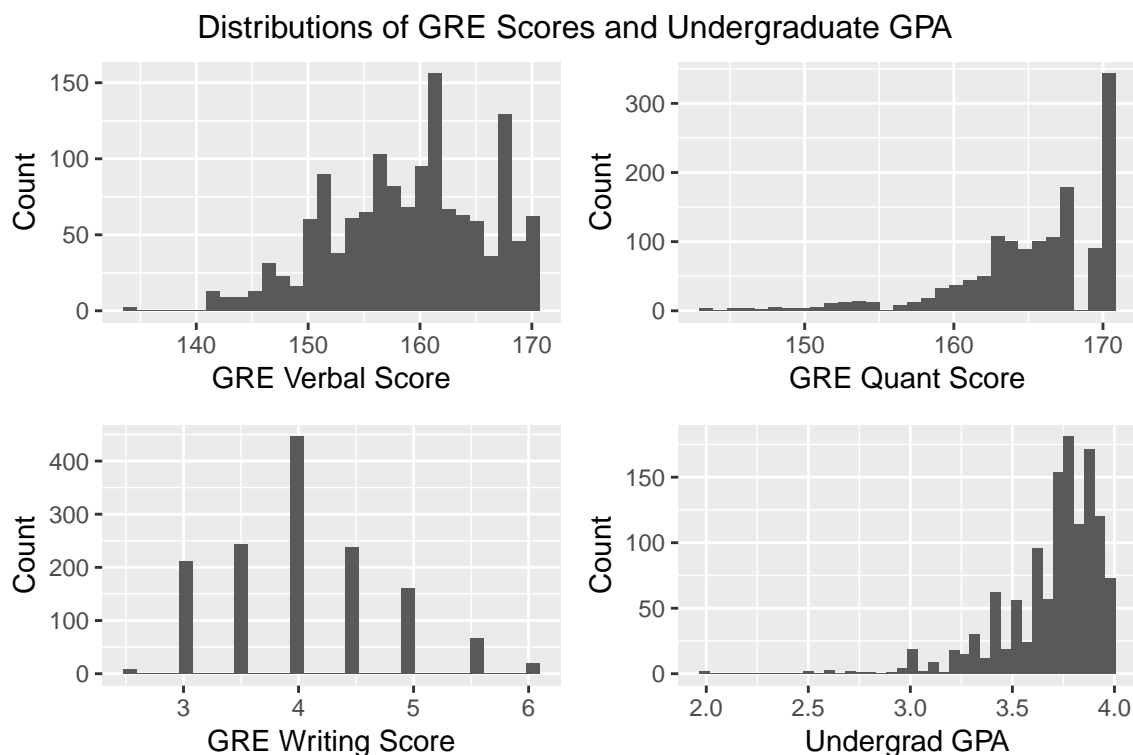


Figure 1: These figures show the distribution of student GRE quant scores, GRE verbal scores, GRE writing scores, and undergraduate GPA

likely present an issue in modeling.

Lastly, we look at the distribution of student status (international, US, international with US degree, etc)

From the chart above, we see that the majority of students applying are international students. In Immigration Status, around 70% of applicants are international students and the rest of them are American and students with unclear immigration status. We can tell that a big amount of graduate or Ph.D. students are coming from an international background.

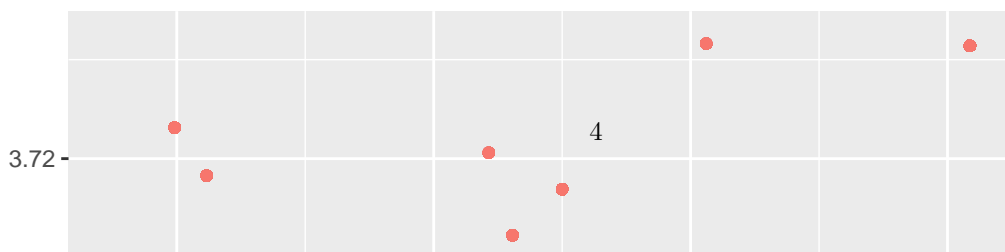
One covariation of interest is the influence of student status (international, US, etc.) on graduate admissions. Mainly, we are interested in understanding if International Students are held to a different standard for test scores in the admission process.

status	Accepted	Interview	Other	Rejected	Wait listed
American	178	13	4	215	4
International	308	24	10	490	6
International with US Degree	56	5	3	79	1

From Figure 4 above, it seems that US based students tend to have higher acceptance rates than international students, and international students with US degree.

Another covariation of interest is the relationship between GPA and GRE scores. For this we summed GRE verbal and GRE quant to get the full GRE score, and created a scatter plot against GPA. We filtered GPA to be less than 4, as GPA of different scales are not comparable.

### Relationship between GPA and GRE Score



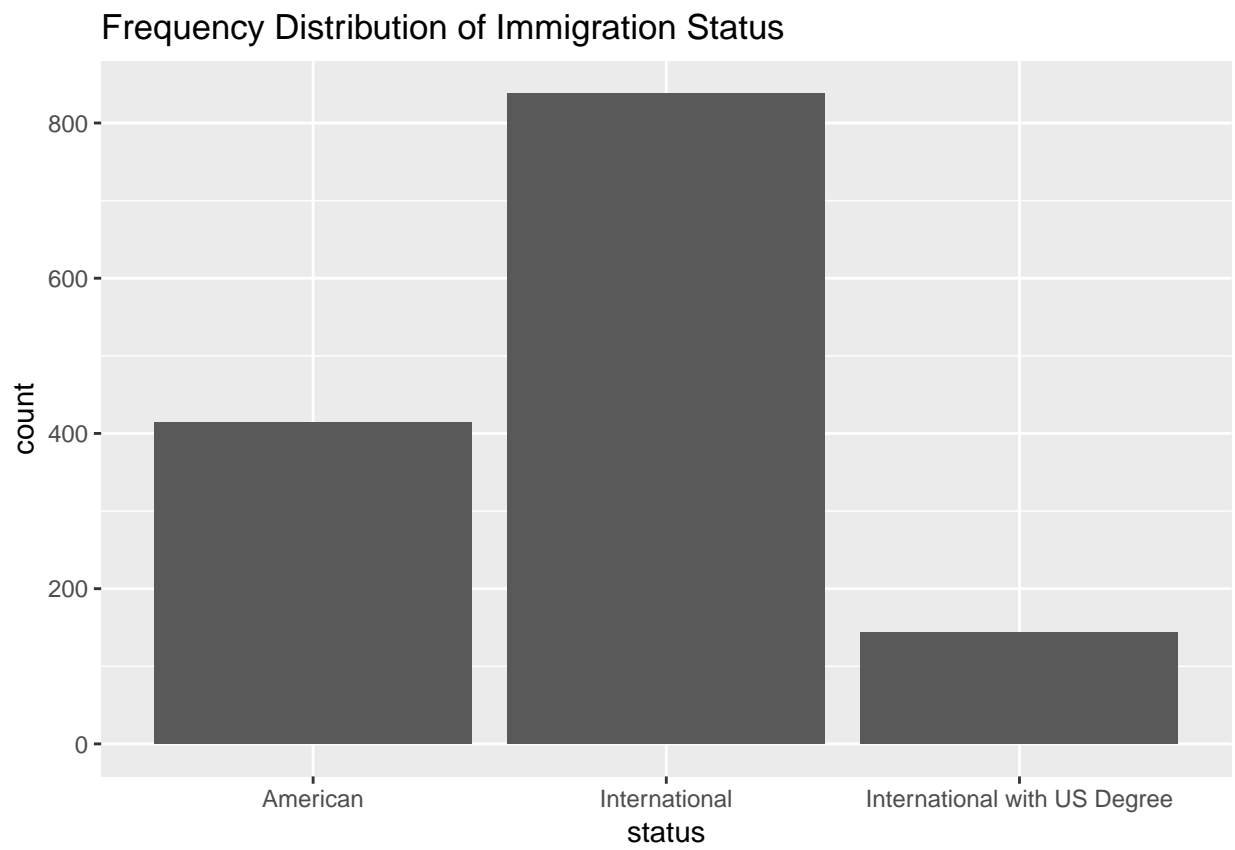


Figure 2: This figure shows the number of students for each immigration status category

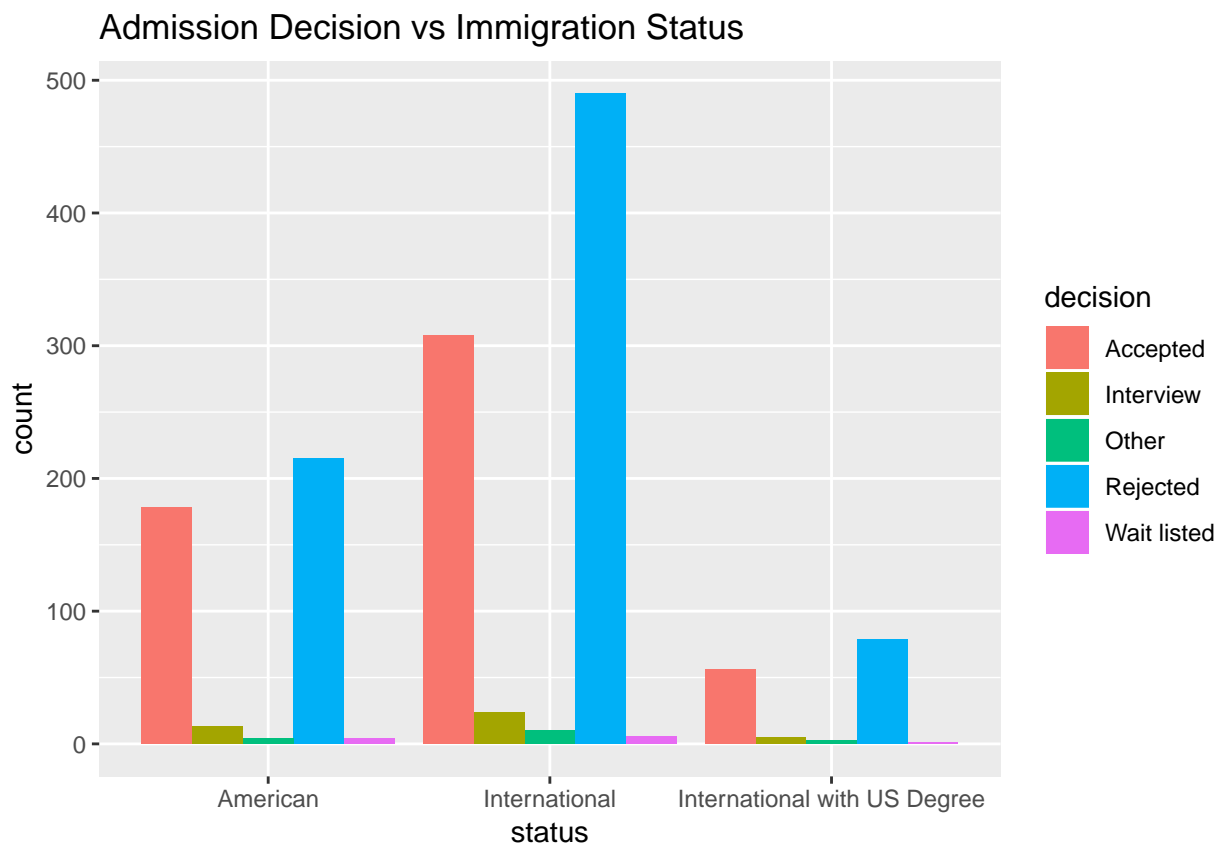


Figure 3: This figure shows the distribution of admission results with respect to student status

## Modeling

Next, we aim to fit a model to the data in order to better understand which factors are most important to admission decisions. Through our initial exploratory data analysis, we realised that the full dataset may be too large to get a clear picture of how different factors affect admissions decision. Because the full dataset contains a wide variety of schools and graduate programs (which would all have different standards for admission and a variety of interactions), we decided to narrow down the dataset to just the top 10 most popular Computer Science programs. The reason for this is that factors related to admission are likely not comparable across different schools (e.g. highly selective schools vs high acceptance rate schools) or different programs (e.g. factors that may be important to Computer Science programs would likely differ from a Fine arts program). We chose to focus on top computer science programs as they are of interest to us as students of data science, and prospective graduate students.

In order to model the probability of acceptance, we decide to use a logistic regression, as the result we want to model is binary (Accepted vs Not Accepted). For our covariates, we hypothesize that GPA, GRE Scores, and student status (American vs International Student) play a significant role in determining the admission decision. We also suspect that there may be interactions between student status and scores. In order to select variables, we start with a full model containing all of these variables and use the backwards elimination method of stepwise regression. This method minimizes a model selection criterion called “AIC” to fit a “best” model to the data.

We decided to fit a model both with and without interaction to better understand how significant the interaction terms are for prediction. Predictor variables included in the models are GRE total score (GRE verbal + GRE quant), undergraduate GPA, GRE writing score, and student status.

### Model With Interaction

```
##
## Call:
## glm(formula = decision1 ~ ugrad_gpa + gre_total + gre_writing +
##      status + ugrad_gpa:status + gre_writing:status - 1, family = binomial,
##      data = grad)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4044  -1.0133  -0.7955   1.2331   2.2644
##
## Coefficients:
##                                Estimate Std. Error
## ugrad_gpa                      0.158544    0.410870
## gre_total                      0.036466    0.008659
## gre_writing                    0.193395    0.162309
## statusAmerican                -13.717033    2.884330
## statusInternational            -17.458204    2.893370
## statusInternational with US Degree -13.063434    4.259133
## ugrad_gpa:statusInternational    1.621023    0.534849
## ugrad_gpa:statusInternational with US Degree -0.387502    1.054098
## gre_writing:statusInternational -0.591549    0.198851
## gre_writing:statusInternational with US Degree  0.221625    0.311482
##                                z value Pr(>|z|)
## ugrad_gpa                      0.386  0.69959
## gre_total                     4.211 2.54e-05 ***
## gre_writing                    1.192  0.23345
## statusAmerican                -4.756 1.98e-06 ***
```

```
## statusInternational -6.034 1.60e-09 ***
## statusInternational with US Degree -3.067 0.00216 **
## ugrad_gpa:statusInternational 3.031 0.00244 **
## ugrad_gpa:statusInternational with US Degree -0.368 0.71316
## gre_writing:statusInternational -2.975 0.00293 **
## gre_writing:statusInternational with US Degree 0.712 0.47676
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1935.3 on 1396 degrees of freedom
## Residual deviance: 1793.3 on 1386 degrees of freedom
## AIC: 1813.3
##
## Number of Fisher Scoring iterations: 4
```

### Model Without Interaction

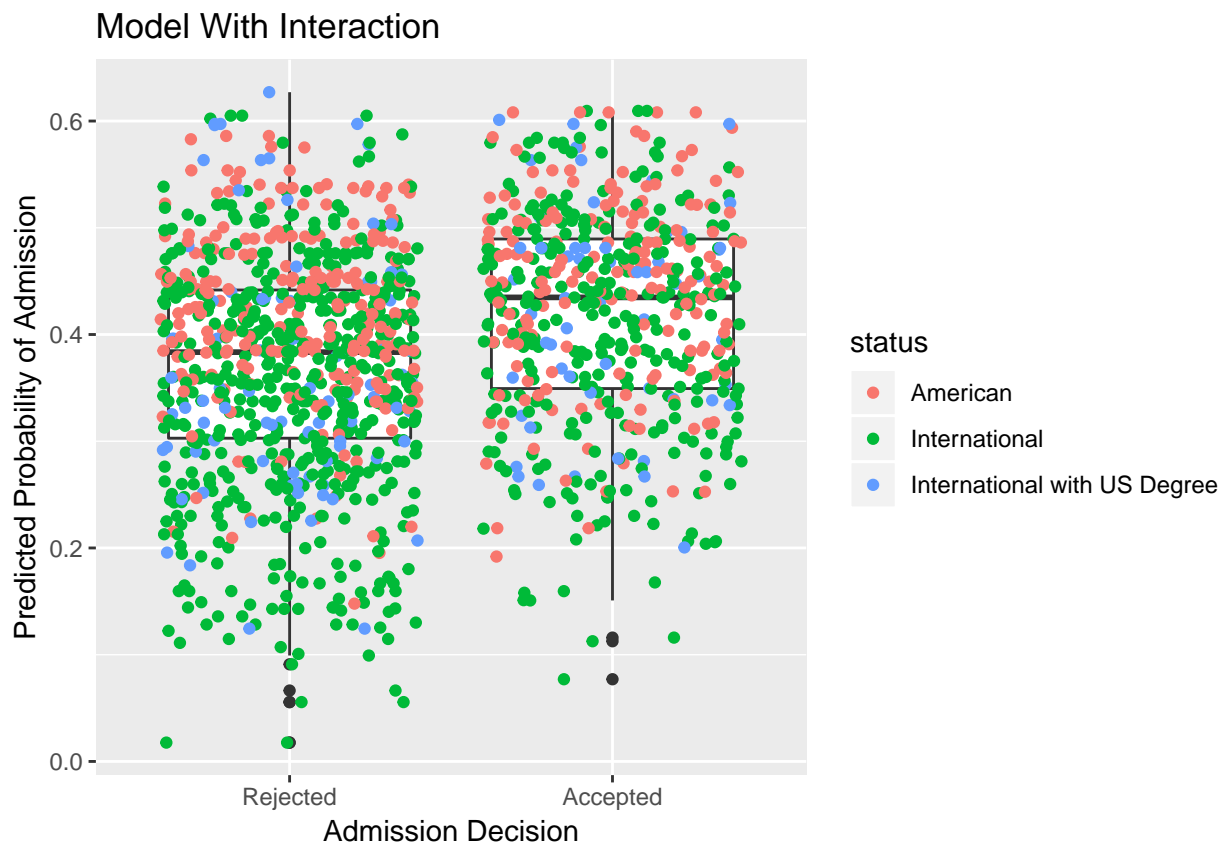
```
##
## Call:
## glm(formula = decision1 ~ ugrad_gpa + gre_total + status - 1,
##      family = binomial, data = grad)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3099 -1.0125 -0.8409  1.2668  2.0953
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## ugrad_gpa      1.109258   0.252718   4.389 1.14e-05
## gre_total       0.027741   0.007694   3.605 0.000312
## statusAmerican -13.534867   2.513739  -5.384 7.27e-08
## statusInternational -13.643998   2.478422  -5.505 3.69e-08
## statusInternational with US Degree -13.651008   2.494039  -5.473 4.41e-08
##
## ugrad_gpa ***
## gre_total ***
## statusAmerican ***
## statusInternational ***
## statusInternational with US Degree ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1935.3 on 1396 degrees of freedom
## Residual deviance: 1816.3 on 1391 degrees of freedom
## AIC: 1826.3
##
## Number of Fisher Scoring iterations: 4
```

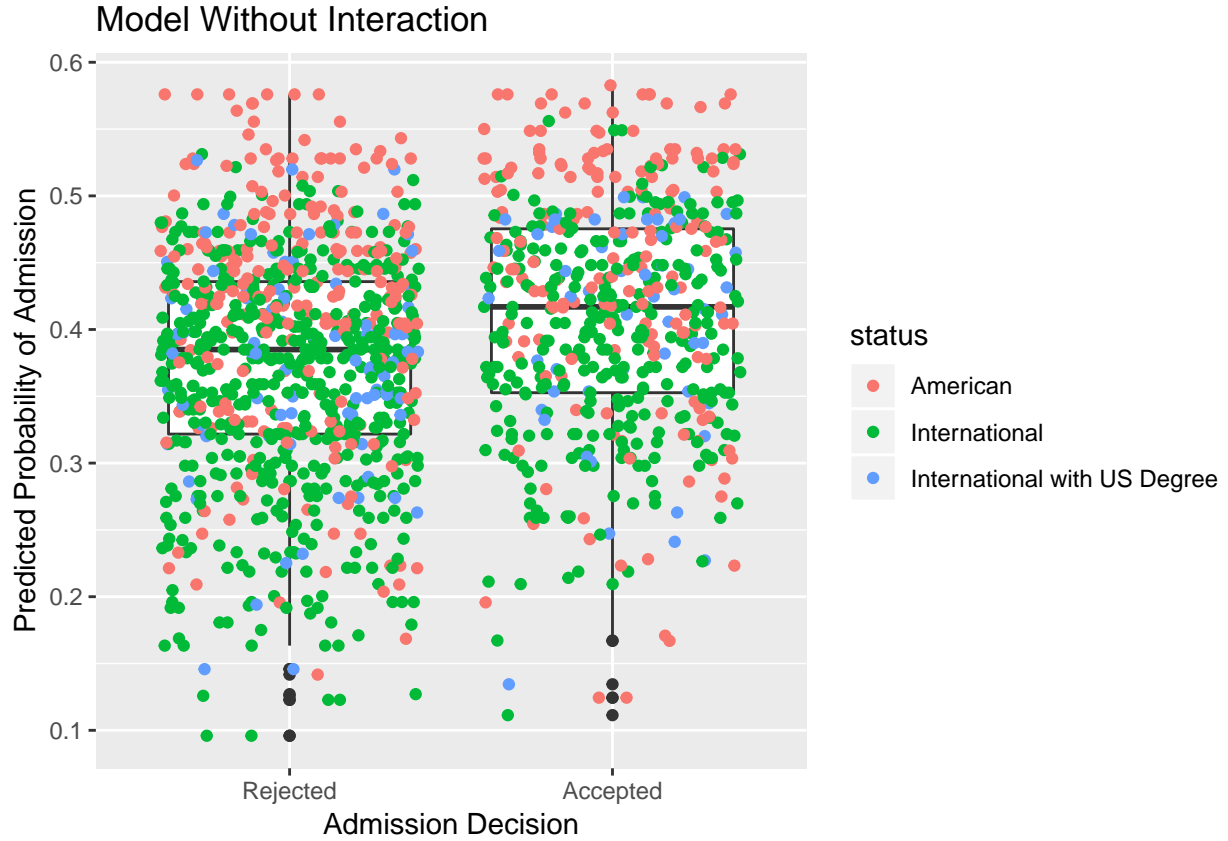
One interesting observation we see is that there is an interaction between GRE Writing and Student Status. This could be due to varying standards for writing ability based on Student Status. American students may



be held to a higher standard for writing quality than international students, which is to be expected.

Next, we plot box plots of admission decisions vs predicted probabilities to assess the predictive power of our models.





In model without interaction, the mean of predicted probabilities of rejected students is around 0.41, while the mean of predicted probabilities of accepted students is around 0.43, which is slightly higher than the mean of predicted probabilities of rejected students. Since their interval are overlapped, it means that the prediction may not be significant enough to explain the success of a student being accepted. In addition, the plots are fairly scattered, meaning that there does not exist a certain pattern to explain the trend.

In model with interaction, the mean of predicted probabilities of rejected students is around 0.4, while the mean of predicted probabilities of accepted students is around 0.45, which is slightly higher than the mean of predicted probabilities of rejected students. Since their interval are overlapped, it means that the prediction may not be significant enough to explain the success of a student being accepted. However, the plots are more densely concentrated than the one without interaction.

## Coefficient Interpretation

For the model that includes interaction:

The regression coefficient for  $ugrad\_gpa$  is  $\beta_{ugradgpa} = 1.146643$  meaning that for a one-unit increase in undergraduate GPA the logit-transformed probability of getting accepted to the program will increase by 1.15. Predictor  $GRE\_Total$  has a coefficient  $\beta_{GREtotal} = 0.031106$ , showing that for a one-unit increase in GRE total scores the log odds will increase by 0.03. We also include categorical variable  $status$  representing the applicant's status. The corresponding coefficient  $\beta_{American} = -13.403241$  shows that if the applicant is an American student, the log odds will decrease by 13.4, holding all other independent variables constant,  $\beta_{International} = -12.782405$  shows the change in log odds given the student is an international student, and  $\beta_{USdegree} = -15.544697$  shows the change in log odds given the student is an international student with a US degree.

$\beta(\hat{GRE}_{writing}) = -0.267686$  is the regression coefficients for GRE writing score, and  $\beta(\hat{GRE}_{writing} : \hat{International}) = -0.252731$  and for the interaction term  $\beta(\hat{GRE}_{writing} : US_{degree}) = 0.540781$  are the coefficients of GRE writing scores with respect to students status. However, the hypothesis tests for coefficient indicates that those terms would not significantly impact the prediction of our model.

```
## ugrad_gpa
## 0.3747516

## ugrad_gpa
## 0.06710505

## ugrad_gpa
## 0.4388602
```

We next check the prediction for the probability of a student getting accepted at mean level GPA, GRE total score, and writing score. According to our model that includes interaction, there's a 47.9% chance that the student will be admitted to the program if the student is an American student, and 57% and 15.6% respectively if the student is an international student or an international student with a US degree.

For the model that does not include interaction terms:

The regression coefficient for ugrad\_gpa is  $\beta(\hat{ugradgpa}) = 1.168482$ , which indicates that for a one-unit increase in undergraduate GPA the logit-transformed probability of getting accepted to the program will increase by 1.15.  $\beta(\hat{GRE}_{total}) = 0.030744$  is the coefficient for predictor GRE\_Total showing that for a one-unit increase in GRE total scores the log odds will increase by 0.03.  $\beta(\hat{GRE}_{writing}) = -0.359779$  shows that GRE writing score is negatively related with the probability of acceptance, and for every one unit increase in writing score leads to a 0.36 drop in log odds. If the applicant is an American students, our model predicts a drop equals to  $\beta(\hat{American}) = -12.892745$  in the log odds, holding all other independent variables constant. If the applicant is a international student, log odds decreases by  $\beta(\hat{International}) = -13.302409$ , and if the student has earned a US degree, log odds drops by  $\beta(\hat{US}_{degree}) = -12.981663$ .

```
## ugrad_gpa
## 9.065024e-25

## ugrad_gpa
## 9.001702e-25

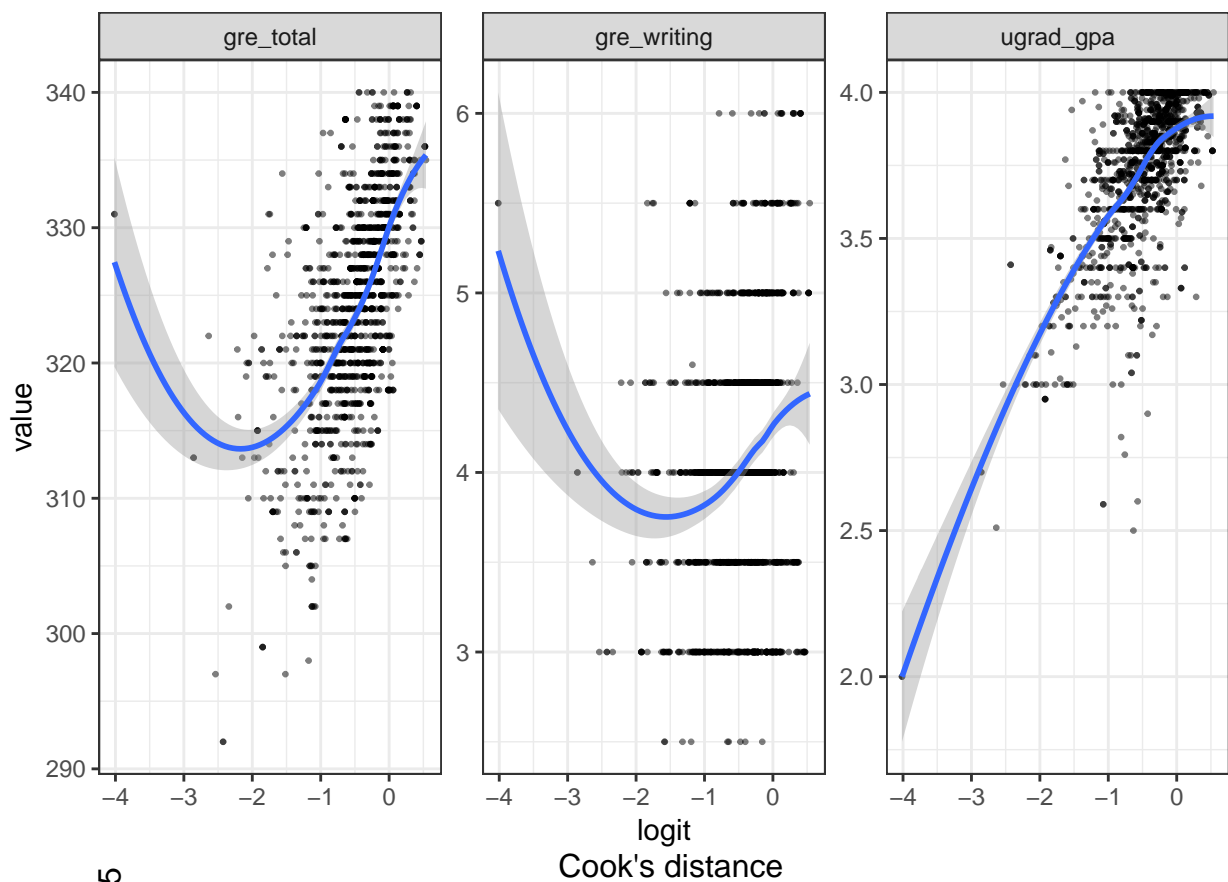
## ugrad_gpa
## NA
```

Using same mean level GPA, GRE total score and writing score, our simple logistic model predicts that the probability of an American student getting accepted to the program is 49.1% and the probability for an international student without a US degree and one with a US degree is 39% and 46.9% respectively.

## Assumption

Next, to ensure that our models are valid, we check the assumptions of logistic regression:

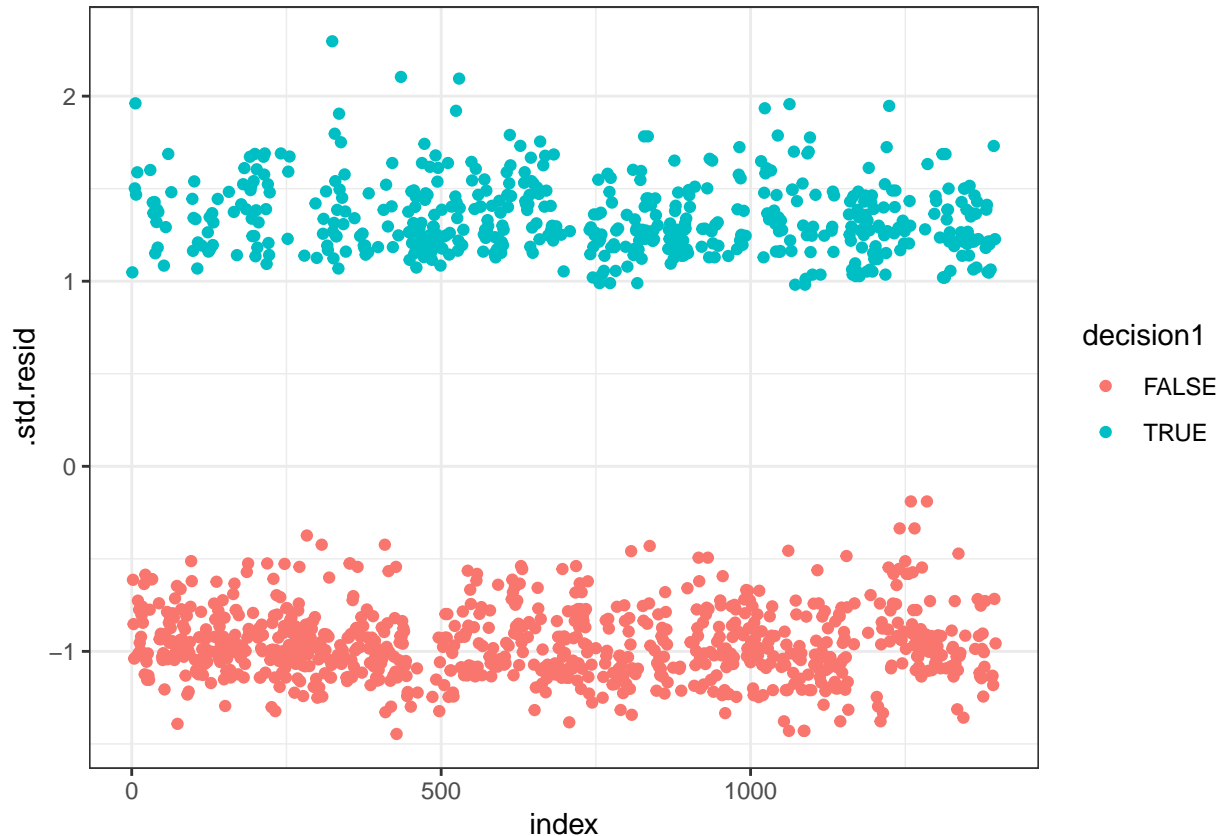
1. Outcome is binary
2. Linear relationship between the logit of the outcome and each predictor variables
3. No influential values
4. No high intercorrelations



glm(decision1 ~ (ugrad\_gpa + gre\_total + gre\_writing) \* status - 1)

```
## # A tibble: 3 x 13
##   decision1 ugrad_gpa gre_total gre_writing status .fitted .se.fit .resid
##   <lgl>      <dbl>    <dbl>    <dbl> <chr>    <dbl>  <dbl>  <dbl>
```

```
## 1 FALSE      3.3      325      5.5 Inter~ 0.284 0.720 -1.30
## 2 TRUE       3.2      324      3   Inter~ -0.530 0.682 1.41
## 3 TRUE       3       305      3   Inter~ -1.51 0.811 1.85
## # ... with 5 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, index <int>
```



```
## # A tibble: 0 x 13
## #   ... with 13 variables: decision1 <lgl>, ugrad_gpa <dbl>,
## #     gre_total <dbl>, gre_writing <dbl>, status <chr>, .fitted <dbl>,
## #     .se_fit <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #     .std.resid <dbl>, index <int>
```

```
## # A tibble: 1,396 x 15
##   uni_name major degree season decision decision_date decision_timest~
##   <chr>      <chr> <chr> <chr> <chr>      <chr>              <dbl>
## 1 Univers~ ECE    MS    S16   Accepted 28-09-2015      1443412800
## 2 Univers~ CS     MS    F15   Rejected 27-01-2015      1422334800
## 3 Univers~ CS     MS    F15   Rejected 24-05-2015      1432440000
## 4 Carnegi~ ECE    MS    F15   Other    27-06-2015      1435377600
## 5 Georgia~ CS     MS    F15   Accepted 13-06-2015      1434168000
## 6 Carnegi~ ECE    MS    F15   Accepted 6-Feb-15        1433217600
## 7 Georgia~ ECE    MS    F15   Accepted 21-05-2015      1432180800
## 8 Georgia~ CS     PhD    F15   Rejected 15-05-2015      1431662400
## 9 Univers~ ECE    MS    F15   Accepted 14-04-2015      1428984000
## 10 Cornell~ CS     MS    F15   Rejected 4-Jul-15        1428379200
## # ... with 1,386 more rows, and 8 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## #   is_new_gre <lgl>, status <chr>, decision1 <lgl>, gre_total <dbl>
```

```
##          ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa          1.00      0.14      0.17      0.10
## gre_verbal          0.14      1.00     -0.10      0.50
## gre_quant           0.17     -0.10      1.00      0.11
## gre_writing         0.10      0.50      0.11      1.00
##
## n= 1396
##
##
## P
##          ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa          0e+00      0e+00      4e-04
## gre_verbal 0e+00          1e-04      0e+00
## gre_quant  0e+00      1e-04          0e+00
## gre_writing 4e-04      0e+00      0e+00
```

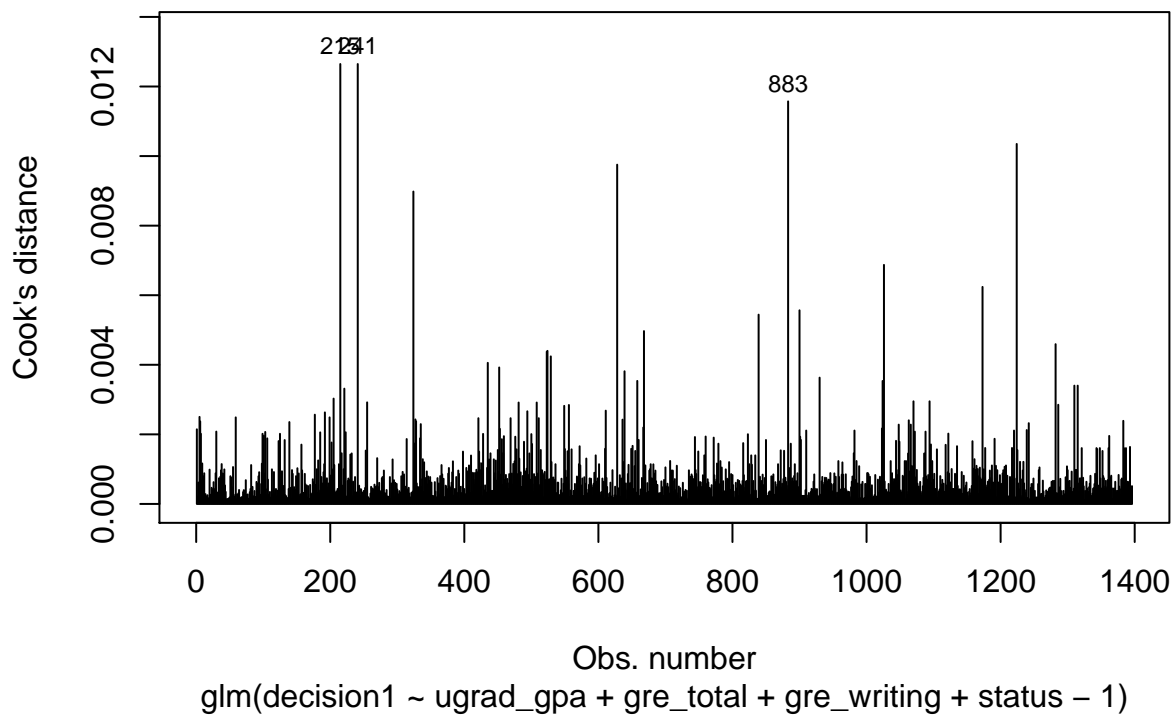
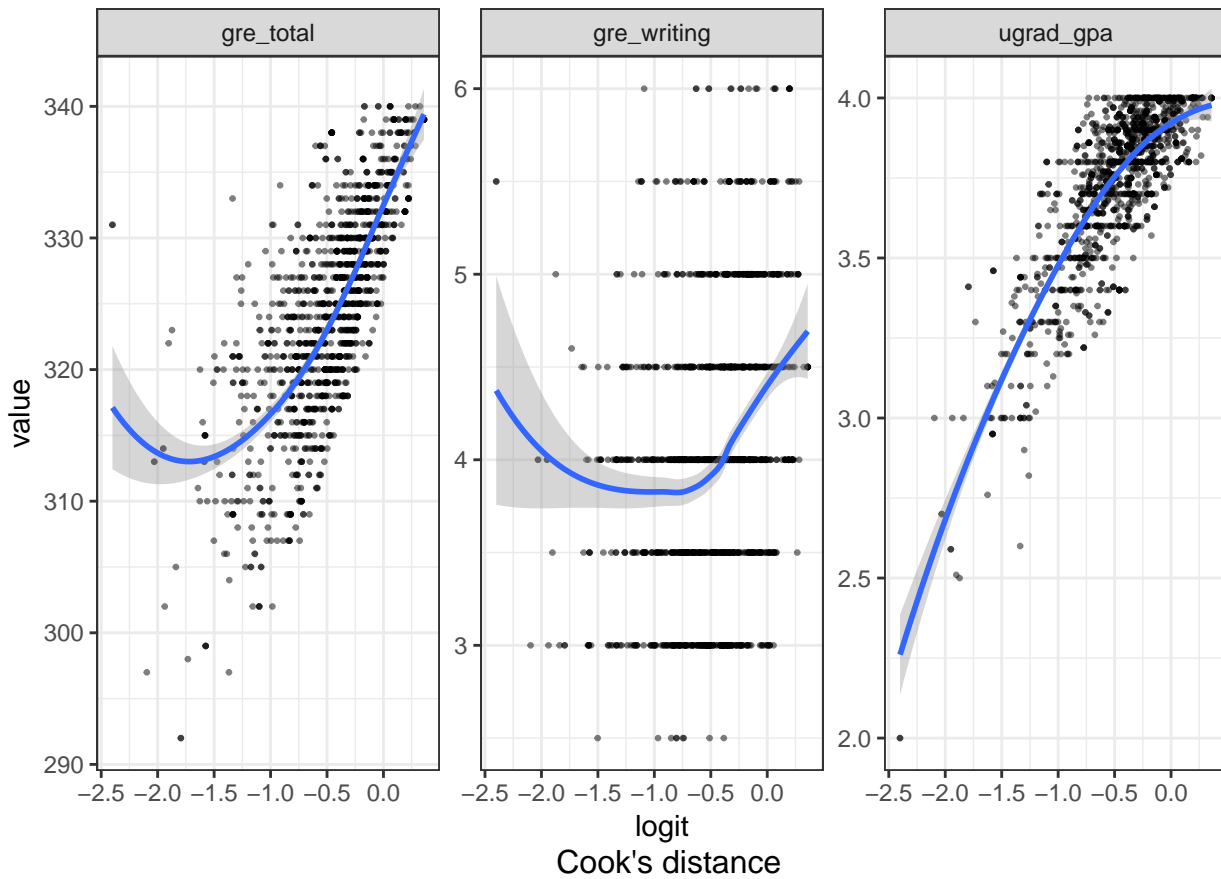
First, since we set the accepted decision as dependent variables and the decision is binary, either 1, accepted or 0, rejected. Therefore, the predicted probability is bind within the interval between 0 and 1. It meets the first assumption of dependent variable to be binary.

Second, logistic regression also assumes the linearity of independent variables. As shown in “The linearity of independent variables”, the logit of GRE is quite linear to the accepted probability in logit scale. Even though there exists an U-shaped trend at the end of the parabala, the majority of gpa points associated linearly to the logit outcome of undergraduate gpa. However, the scatter plots of gre\_writing shows non\_linearity, similar to a cubic term.

Third, some outliers may be influential enough to alter the quality of the logistic regression model. Therefore, we calculated the Cook’s distance for each points; the higher the leverage and residuals of that point, the higher its Cook’s distance. As demonstrated in Cook’s distance graph, there exist couple of spikes in the graph. To further investigate this issue, the deviance residuals plots has ben constructed. Since it does not have any observations whose cook’s value is large than 3, we conclude that the dataset does not have any influential outliers.

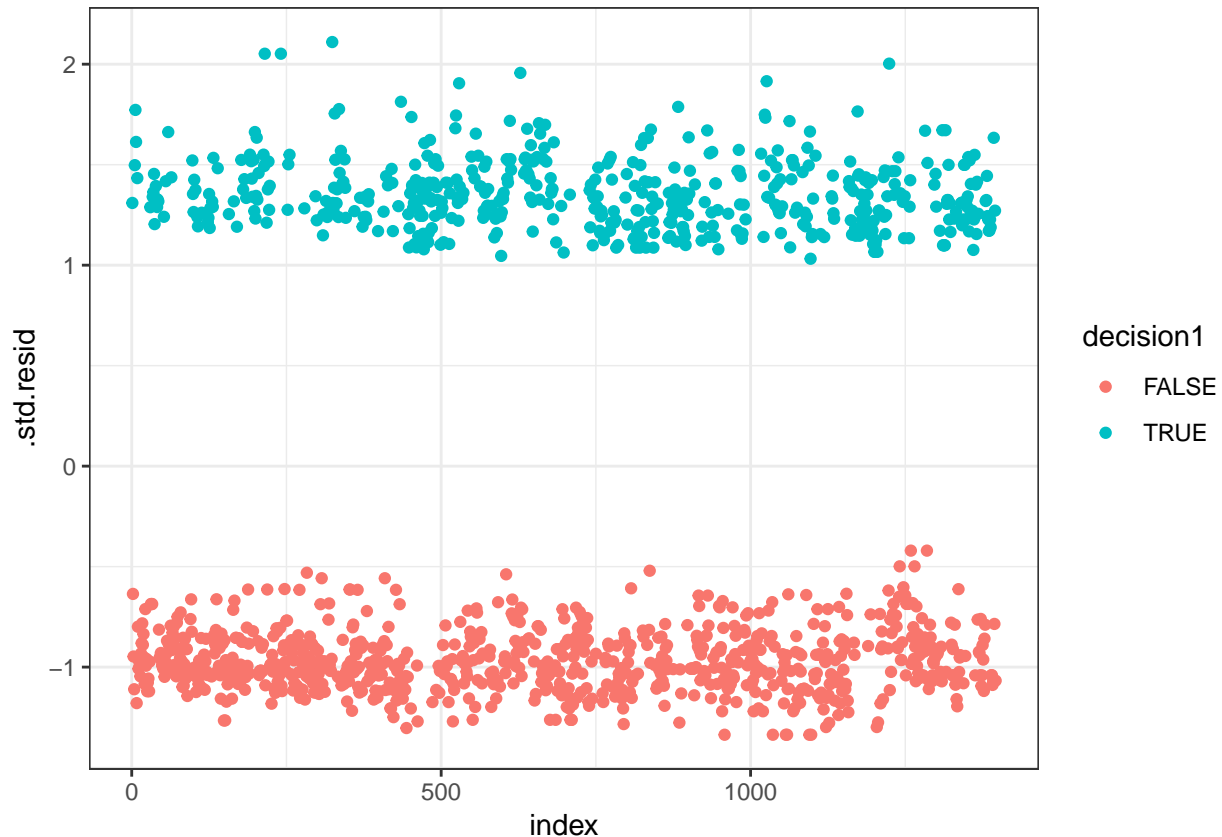
Last but not least, since the variables are intercorrelated, we take this into consideration and use interaction terms to overcome this issue.

## Assumption\_w/o interaction



## # A tibble: 3 x 13

```
## decision1 ugrad_gpa gre_total gre_writing status .fitted .se.fit .resid
## <lgl> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl>
## 1 TRUE 2.59 314 4 Ameri~ -1.95 0.311 2.04
## 2 TRUE 2.59 314 4 Ameri~ -1.95 0.311 2.04
## 3 TRUE 2.6 333 4 Ameri~ -1.34 0.327 1.77
## # ... with 5 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## # .std.resid <dbl>, index <int>
```



```
## # A tibble: 0 x 13
## # ... with 13 variables: decision1 <lgl>, ugrad_gpa <dbl>,
## # gre_total <dbl>, gre_writing <dbl>, status <chr>, .fitted <dbl>,
## # .se.fit <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## # .std.resid <dbl>, index <int>
```

```
## # A tibble: 1,396 x 15
## uni_name major degree season decision decision_date decision_timest~
## <chr> <chr> <chr> <chr> <chr> <chr> <dbl>
## 1 Univers~ ECE MS S16 Accepted 28-09-2015 1443412800
## 2 Univers~ CS MS F15 Rejected 27-01-2015 1422334800
## 3 Univers~ CS MS F15 Rejected 24-05-2015 1432440000
## 4 Carnegi~ ECE MS F15 Other 27-06-2015 1435377600
## 5 Georgia~ CS MS F15 Accepted 13-06-2015 1434168000
## 6 Carnegi~ ECE MS F15 Accepted 6-Feb-15 1433217600
## 7 Georgia~ ECE MS F15 Accepted 21-05-2015 1432180800
## 8 Georgia~ CS PhD F15 Rejected 15-05-2015 1431662400
## 9 Univers~ ECE MS F15 Accepted 14-04-2015 1428984000
## 10 Cornell~ CS MS F15 Rejected 4-Jul-15 1428379200
## # ... with 1,386 more rows, and 8 more variables: ugrad_gpa <dbl>,
```



```
## # gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## # is_new_gre <lgl>, status <chr>, decision1 <lgl>, gre_total <dbl>
```

First, since we set the accepted decision as dependent variables and the decision is binary, either 1, accepted or 0, rejected. Therefore, the predicted probability is bind within the interval between 0 and 1. It meets the first assumption of dependent variable to be binary.

Second, logistic regression also assumes the linearity of independent variables. As shown in “The linearity of independent variables”, the logit of GRE and undergraduate gpa are fairly linear to the accepted probability in logit scale. However, the scatter plots of gre\_writing fits a parabola, instead of a linear line.

Third, some outliers may be influential enough to alter the quality of the logistic regression model. Therefore, we calculated the Cook’s distance for each points; the higher the leverage and residuals of that point, the higher its Cook’s distance. As demonstrated in Cook’s distance graph, there exist couple of spikes in the graph. To further investigate this issue, the deviance residuals plots has ben constructed. Since it does not have any observations whose cook’s value is large than 3, we conclude that the dataset does not have any influential outliers.

Last but not least, from the covariance matrix, we can tell that each term are correlated with each other since its p value is near 0. Therefore, we incorporate interaction terms in our further model to overcome this disadvantage.

## Tests for Significant Interaction

```
## # A tibble: 1,396 x 16
##   uni_name major degree season decision decision_date decision_timest~
##   <chr>      <chr> <chr> <chr> <chr>      <chr>              <dbl>
## 1 Univers~ ECE    MS    S16    Accepted 28-09-2015          1443412800
## 2 Univers~ CS     MS    F15    Rejected 27-01-2015          1422334800
## 3 Univers~ CS     MS    F15    Rejected 24-05-2015          1432440000
## 4 Carnegi~ ECE    MS    F15    Other    27-06-2015          1435377600
## 5 Georgia~ CS     MS    F15    Accepted 13-06-2015          1434168000
## 6 Carnegi~ ECE    MS    F15    Accepted 6-Feb-15            1433217600
## 7 Georgia~ ECE    MS    F15    Accepted 21-05-2015          1432180800
## 8 Georgia~ CS     PhD    F15    Rejected 15-05-2015          1431662400
## 9 Univers~ ECE    MS    F15    Accepted 14-04-2015          1428984000
## 10 Cornell~ CS     MS    F15    Rejected 4-Jul-15            1428379200
## # ... with 1,386 more rows, and 9 more variables: ugrad_gpa <dbl>,
## # gre_verbal <dbl>, gre_quant <dbl>, gre_writing <int>,
## # is_new_gre <lgl>, status <chr>, decision1 <lgl>, gre_total <dbl>,
## # pred <dbl>
```

## Interaction Plot for GRE Writing and Student Status

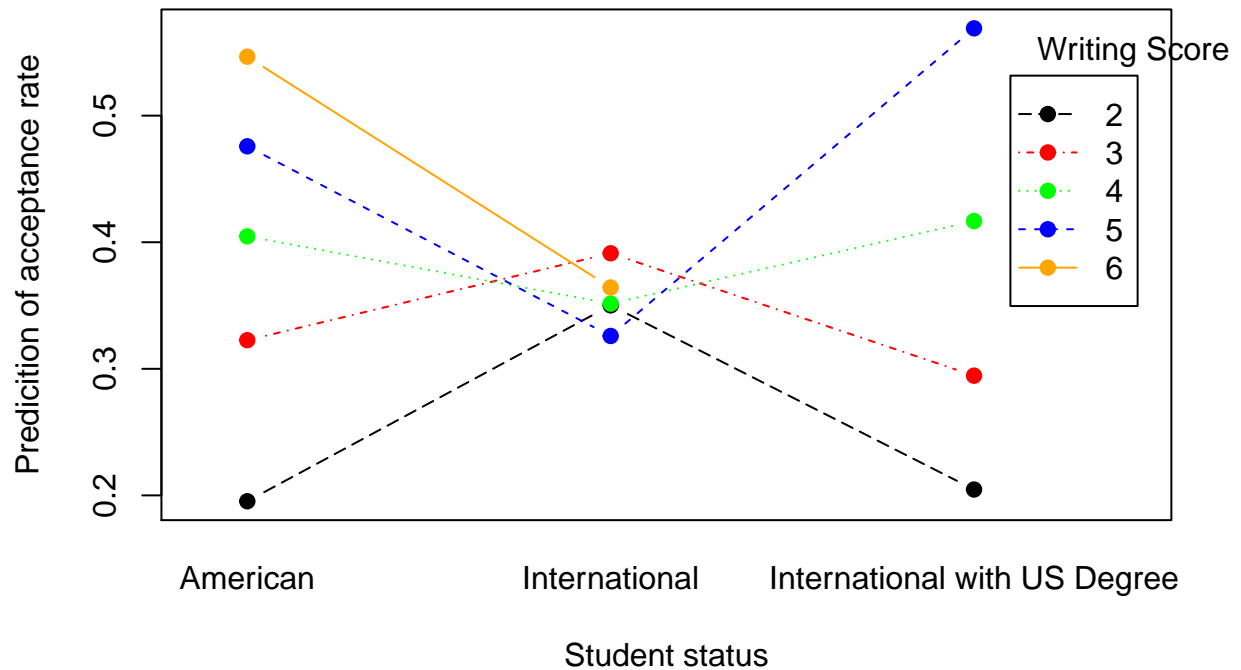


fig Interaction Plot Base on information of top 10 Computer Science programs

```
## Analysis of Deviance Table
##
## Model 1: decision1 ~ ugrad_gpa + gre_total + gre_writing + status - 1
## Model 2: decision1 ~ (ugrad_gpa + gre_total + gre_writing) * status -
##      1
##      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1390      1815.1
## 2      1384      1792.4  6    22.705 0.0009017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: decision1
##
## Terms added sequentially (first to last)
##
##              Df Deviance Resid. Df Resid. Dev
## NULL              1396      1935.3
## ugrad_gpa          1   64.089      1395      1871.2
## gre_total           1   17.451      1394      1853.7
## gre_writing         1    2.683      1393      1851.0
## status              3   35.983      1390      1815.1
## ugrad_gpa:status    2    8.073      1388      1807.0
## gre_total:status    2    3.198      1386      1803.8
## gre_writing:status  2   11.433      1384      1792.4
```

The plot suggests that the effect of GRE writing is not consistent across all three groups of students. For example, a writing score of 5 showed the greater mean probability of acceptance for American and international students with US Degree. for international student, a writing score of 3 gives the highest chance of acceptance. This suggests there may be a meaningful or significant interaction effect, but we will need to do a statistical test to confirm this hypothesis.

## Test for the inclusion of a Categorical Variable

$H_0$ : full\_mod = full\_mod

$H_a$ : full\_mod = full\_mod\_int

Significance Level: 0.05  $\Pr(>\text{Chi})$  for two models is 0.1581, which is bigger than significant level 0.05. Therefore, two models are not significantly different.  $\Pr(>\text{Chi})$  for ugrad\_gpa, GRE\_Total, gre\_writing and status are all smaller than significant level 0.05, while all the interaction effect is not significant. Therefore, the anova table indicates that the main effect are significant, and interaction effect is not significant. Our interpretation of this result is that, since our research focus on the top 10 Computer Science programs, most of the applicants have strong academic backgrounds, regardless of student status. For example, the acceptance rate for Carnegie Mellon University CS Program is ~6.5%. The distribution of applicants' grade are extremely left skewed. That means it is harder to differentiate international students and American students just by looking at their standard grades.

## Discussion

From this exploratory data analysis, we confirm many of the hypotheses that we had going into this project. The number of students applied for advanced degrees has increased over the years. We confirmed relationships between variables such as GPA and GRE scores. We learned several things as well. For example, we learned the distribution of GPA and GRE quant score are left skewed, and GRE verbal and writing scores have a bell like distribution, with several "spikes" among certain scores. We were surprised to see that American students tended to have higher rates of acceptance than international students.

While this is a very interesting and robust dataset to analyze, there are also several problems we encountered. First, the dataset is not very clean, as it is self-reported. For example, the names of Universities and Majors are not always consistent. For example, some students may write "Boston University (BU)" while others write the name of the specific college at BU such as "Boston University - Metropolitan College." We also noticed that scales of scores and GPA are not always consistent. For example, GPA is most often reported on a 4.0 scale, however, some responses included other scales such as 10 point scale. These will all be problems that we have to work around when going into modeling.

From the analysis above, we see that while GPA, GRE Scores, and Student Status have a significant affect on admissions decisions, they alone are not great predictors for admission results. We see from the box plots that while the model had a higher average predicted probability for students that were actually accepted, there is too much variance in the resulted predictions. This result is likely due to the fact that the dataset is missing many variables that may also be important for admission decisions, such as research experience, recommendations, reputation of undergraduate institution and so on. While it may be possible to extract this information from the 'comments', many observations did not include any comments and many more did not mention these factors in the comments. This leads us to believe that the admissions process is more than just a "numbers game," and likely includes many "intangibles" in order to determine the ultimate admission result of each student.