MA415 Lab: Deliverable 2

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

Abstract

In this project, we analyzed the dataset provided by Grad cafe website. Our main goal is to investigate the question of how different variables, such as GRE score and undergraduate GPA, relate to the admission decision. In this deliverable, we focus our analysis on graduate applications to US Top 10 computer science programs. We aim to better understand the factors affecting admission decisions by fitting a logistic model to the data. The covariates we use for the model include undergrad GPA, GRE Scores, Student Status, and interaction terms. We used the model to predict the probability of a student getting accepted. The model shows that, for an American student with average grade, the probability of getting is 49%. The probability for an average international student is 39%. For an average international student with a US degree, the probability is 46%. While our model was able to predict to some degree, which students were more likely to be accepted, the predicted probabilities were too variable to be useful for prediction. This is likely due to many variables being missing from the data such as research experience, reccomendations, reputation of undergraduate institution and so on.

Research Questions and Modeling Methods

Our general research question remains the same as in deliverable one: "How do different variables relate to admission decision?" However, through our 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 accross 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).

```
##                                        uni_name accepted    n      rate
## 1                 Carnegie Mellon University (CMU)      523 1414 0.3698727
## 2          Georgia Institute Of Technology (GTech)      413  985 0.4192893
## 3       University Of California, San Diego (UCSD)      349  954 0.3658281
## 4  University Of Illinois, Urbana-Champaign (UIUC)      367  954 0.3846960
## 5                               Stanford University      245  914 0.2680525
## 6         University Of California, Berkeley (UCB)      155  844 0.1836493
## 7                                Purdue University      335  745 0.4496644
## 8           University Of Washington, Seattle (UW)      167  713 0.2342216
## 9          University Of Texas, Austin (UT Austin)      282  706 0.3994334
## 10                              Cornell University      253  674 0.3753709
```

In order to model the probability of acceptance, we decide to use a logisitic 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 to fit a "best" model.

Fitting the Model

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 + gre_writing:status - 1, family = binomial, data = grad)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5368  -1.0575  -0.8203   1.2008   1.9261
##
## Coefficients:
##                                                Estimate Std. Error
## ugrad_gpa                                      1.146643   0.283057
## GRE_Total                                      0.031278   0.009388
## gre_writing                                   -0.268174   0.189752
## statusAmerican                               -13.403241   2.973278
## statusInternational                          -12.782405   2.852552
## statusInternational with US Degree           -15.544697   3.081698
## gre_writing:statusInternational               -0.252731   0.225536
## gre_writing:statusInternational with US Degree  0.540781   0.358197
##                                               z value Pr(>|z|)
## ugrad_gpa                                       4.051 5.10e-05 ***
## GRE_Total                                       3.332 0.000863 ***
## gre_writing                                    -1.413 0.157571
## statusAmerican                                 -4.508 6.55e-06 ***
## statusInternational                            -4.481 7.43e-06 ***
## statusInternational with US Degree             -5.044 4.55e-07 ***
## gre_writing:statusInternational                -1.121 0.262468
## gre_writing:statusInternational with US Degree  1.510 0.131112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1512.4  on 1091  degrees of freedom
## Residual deviance: 1438.4  on 1083  degrees of freedom
## AIC: 1454.4
##
## Number of Fisher Scoring iterations: 4
```

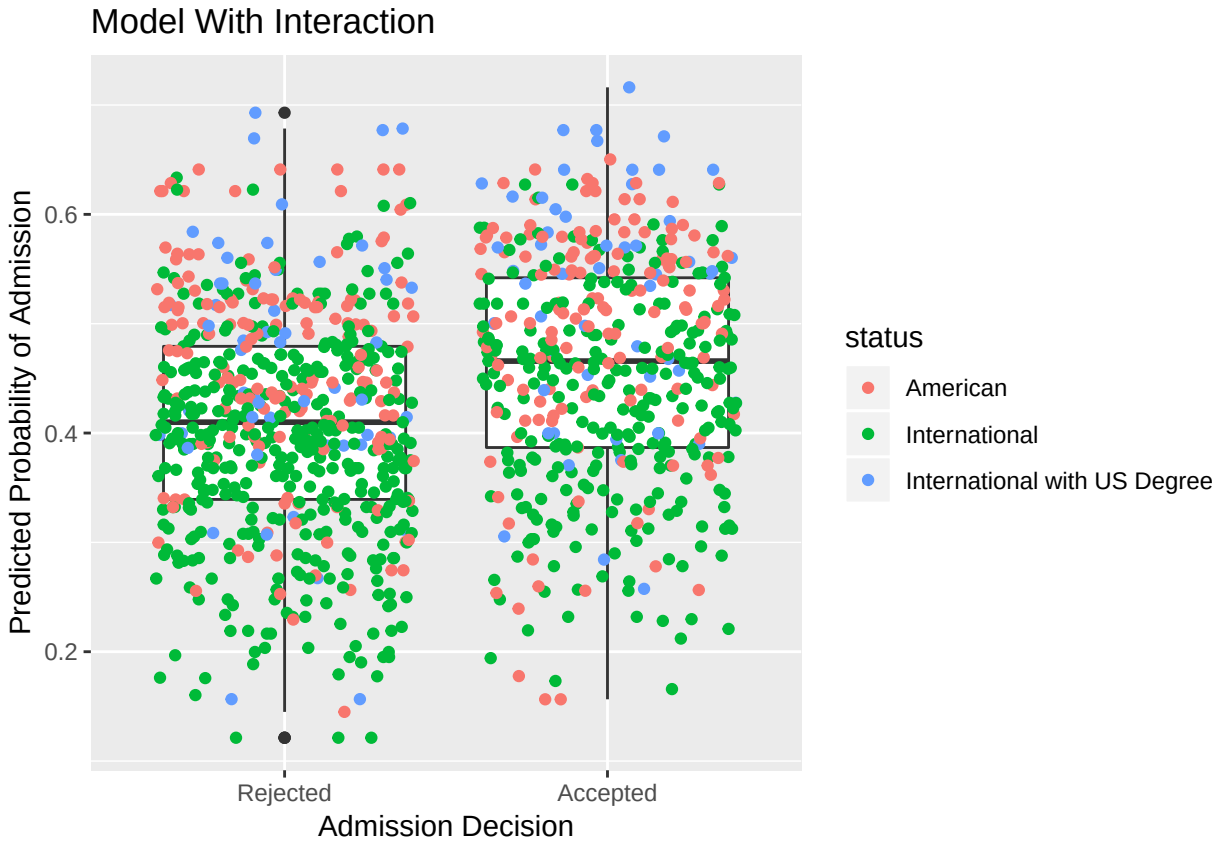
Model Without Interaction

```
##
## Call:
## glm(formula = decision1 ~ ugrad_gpa + GRE_Total + gre_writing +
```

```
##     status - 1, family = binomial, data = grad)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4334  -1.0585  -0.8327   1.2071   1.9526
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## ugrad_gpa                          1.168482   0.283037   4.128 3.65e-05
## GRE_Total                          0.030744   0.009342   3.291 0.000998
## gre_writing                       -0.359779   0.110128  -3.267 0.001087
## statusAmerican                   -12.892745   2.846590  -4.529 5.92e-06
## statusInternational              -13.302409   2.834117  -4.694 2.68e-06
## statusInternational with US Degree -12.981663   2.846232  -4.561 5.09e-06
##
## ugrad_gpa                         ***
## GRE_Total                         ***
## gre_writing                       **
## 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: 1512.4  on 1091  degrees of freedom
## Residual deviance: 1444.7  on 1085  degrees of freedom
## AIC: 1456.7
##
## 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.


Model With Interaction
Predicted Probability of Admission
0.6
0.4
0.2
Rejected
Accepted
Admission Decision
status
American
International
International with US Degree


Model Without Interaction
Predicted Probability of Admission
0.6
0.5
0.4
0.3
0.2
0.1
Rejected
Accepted
Admission Decision
status
American
International
International with US Degree

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 interation.

Coefficient Interpretation

For the model that includes interaction:

The regression coefficient for ugrad_gpa is $\hat{\beta_(ugradgpa)} = 1.146643$ 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 $\hat{\beta_(GREtotal)} = 0.031106$, 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 $\hat{\beta_(American)} = -13.403241$ shows that if the applicant is an American student, the log odds will decrease by 13.4, holding all other independent variables constant, $\hat{\beta_(International)} = -12.782405$ shows the change in log odds given the student is an international student, and $\hat{\beta_(USdegree)} = -15.544697$ shows the change in log odds given the student is an international student with a US degree.

$\hat{\beta_(GREwriting)} = -0.267686$ is the regression coefficients for GRE writing score, and $\hat{\beta_(GREwriting : International)} = -0.252731$ and for the interaction term $\hat{\beta_(GREwriting : USdegree)} = 0.540781$ 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
## 0.4785392
```

```
## ugrad_gpa
## 0.5700877
```

```
## ugrad_gpa
## 0.1562274
```

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 $\hat{\beta_(ugradgpa)} = 1.168482$, 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. $\hat{\beta_(GREtotal)} = 0.030744$ 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. $\hat{\beta_(GREwriting)} = -0.359779$ 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 $\hat{\beta_(American)} = -12.892745$ in the log odds, holding all other independent variables

constant. If the aaplicant is a international student, log odds decreases by $\hat{\beta_(International)} = -13.302409$, and if the student has earned a US degree, log odds drops by $\hat{\beta_(USdegree)} = -12.981663$.

```
## ugrad_gpa
## 0.4909546
```

```
## ugrad_gpa
##  0.390348
```

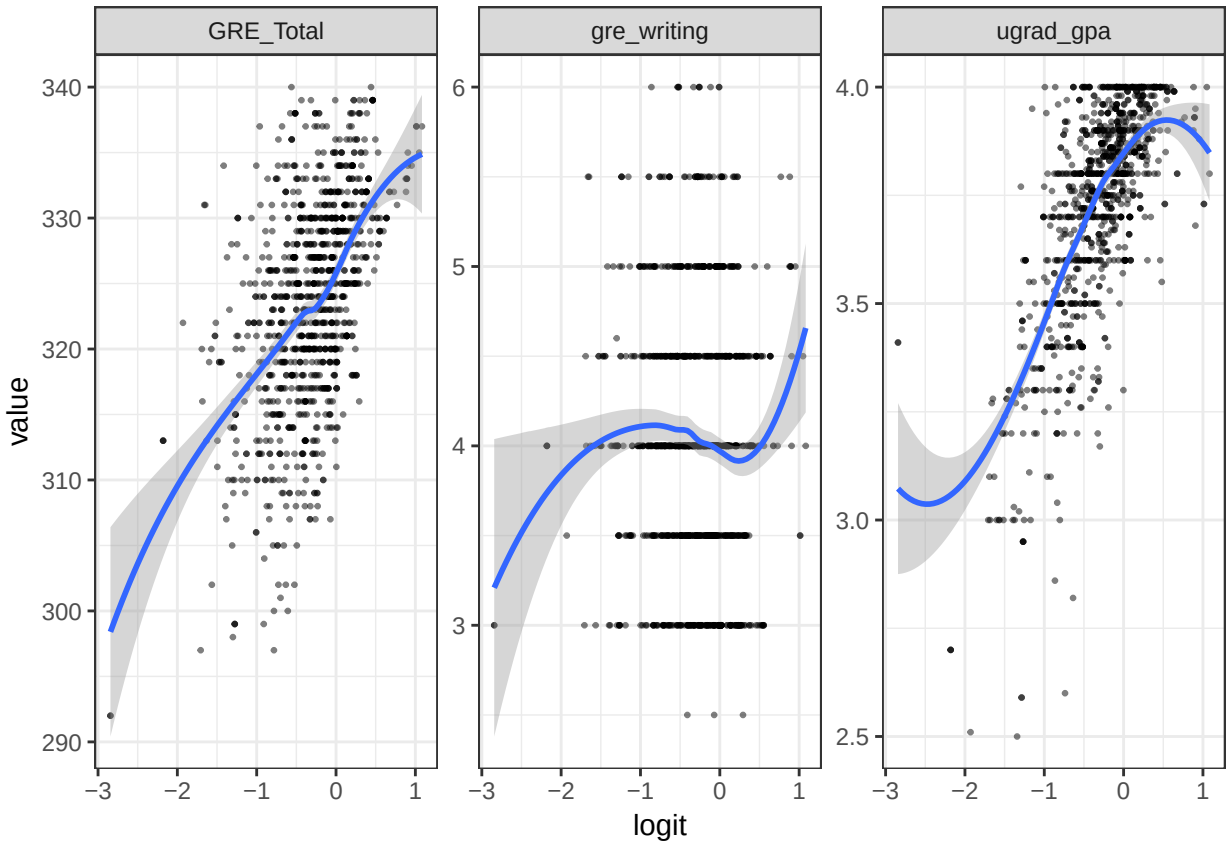
```
## ugrad_gpa
##  0.468765
```

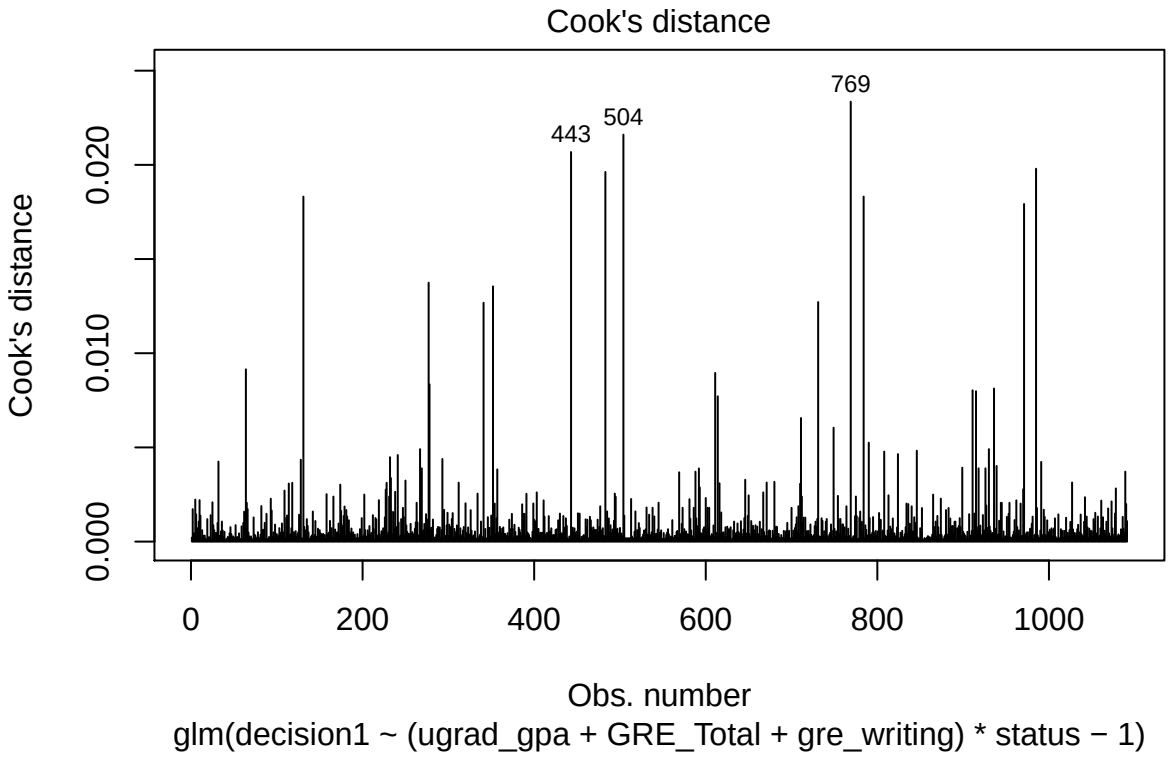
Using same mean level GPA, GRE total score and writing score, our simple logistic model predicts that the probability of an American student getting accepted to the program is 49.1% and the probability for an international student without a US degree and one with a US degree is 39% and 46.9% respectively.

Assumption

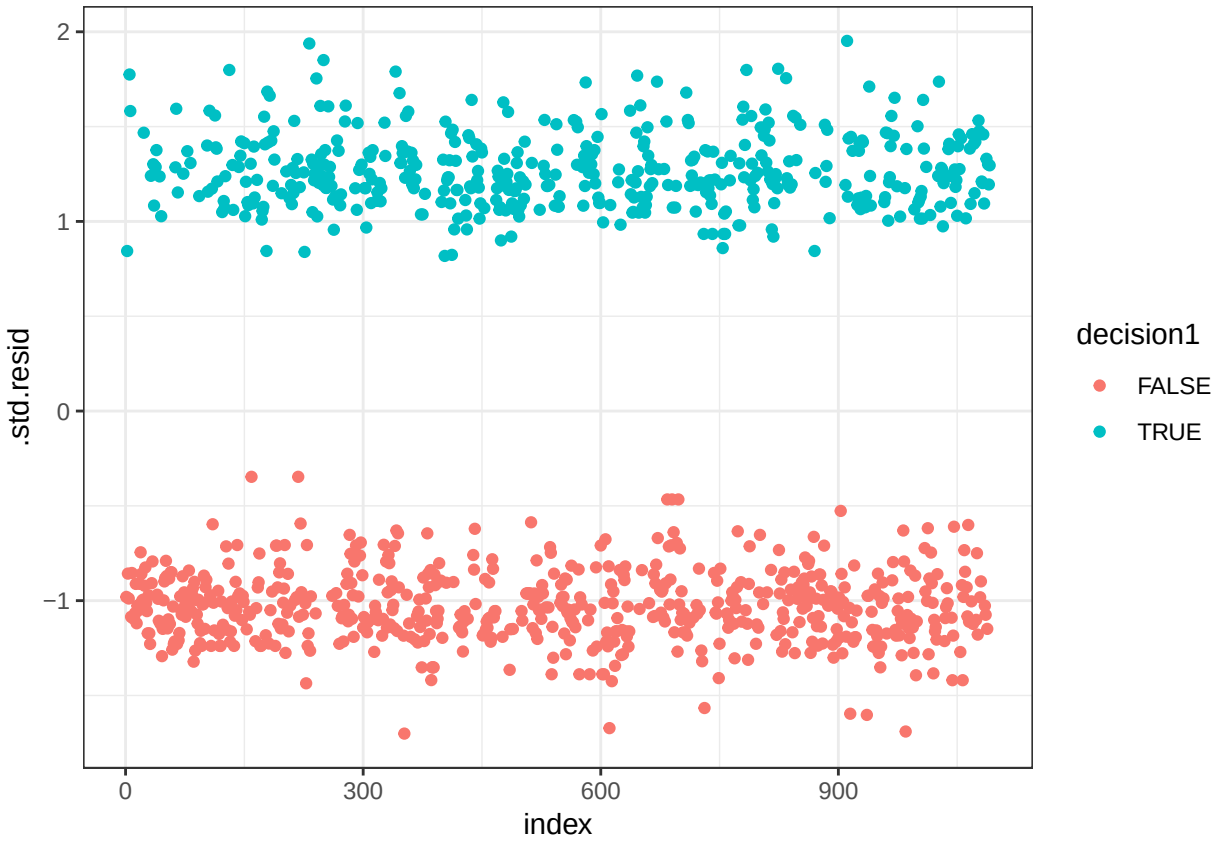
Next, to ensure that our models are valid, we check the assumptions of logistic regression:

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






```
## # 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           3.17       324         3.5 Inter~ -0.299     0.749   1.31
## 2 TRUE           3.2        324         3   Inter~ -0.325     0.755   1.32
## 3 FALSE          3.3        325         5.5 Inter~ -0.00218   0.863  -1.18
## # ... 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: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected (2, 11, 2015)       1446440400
##  2 Univers~ Comp~ MS     S16    Accepted (28, 9, 2015)       1443412800
##  3 Univers~ (Com~ MS     F15    Rejected (24, 5, 2015)       1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    (27, 6, 2015)       1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted (2, 6, 2015)        1433217600
##  6 Univers~ Elec~ MS     F15    Accepted (14, 4, 2015)       1428984000
##  7 Univers~ Comp~ PhD    F15    Other    (20, 4, 2015)       1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected (7, 4, 2015)        1428379200
##  9 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## 10 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## #   is_new_gre <lgl>, status <chr>, comments <chr>, decision1 <lgl>,
## #   GRE_Total <dbl>, gre_total <dbl>
```

```
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa        1.00       0.19      0.18        0.13
## gre_verbal       0.19       1.00     -0.08        0.54
## gre_quant        0.18      -0.08      1.00        0.03
```

```
## gre_writing      0.13       0.54      0.03        1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa             0.0000     0.0000    0.0000
## gre_verbal  0.0000               0.0069    0.0000
## gre_quant   0.0000    0.0069               0.3114
## gre_writing 0.0000    0.0000     0.3114
```

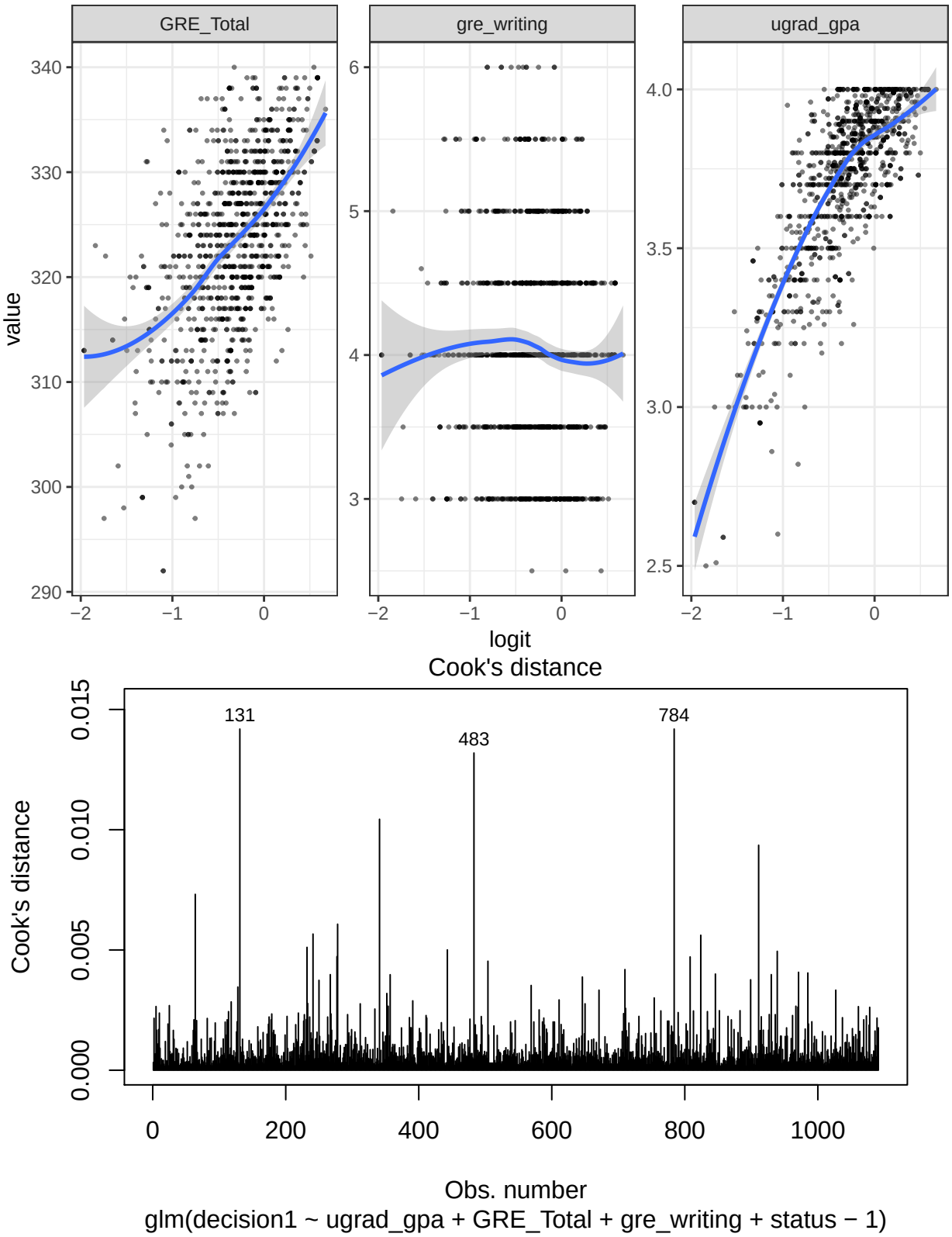
First, since we set the accepted decision as dependent variables and the decision is binary, either 1, accepted or 0, rejected. Therefore, the predicted probability is bind within the interval between 0 and 1. It meets the first assumption of dependent variable to be binary.

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.

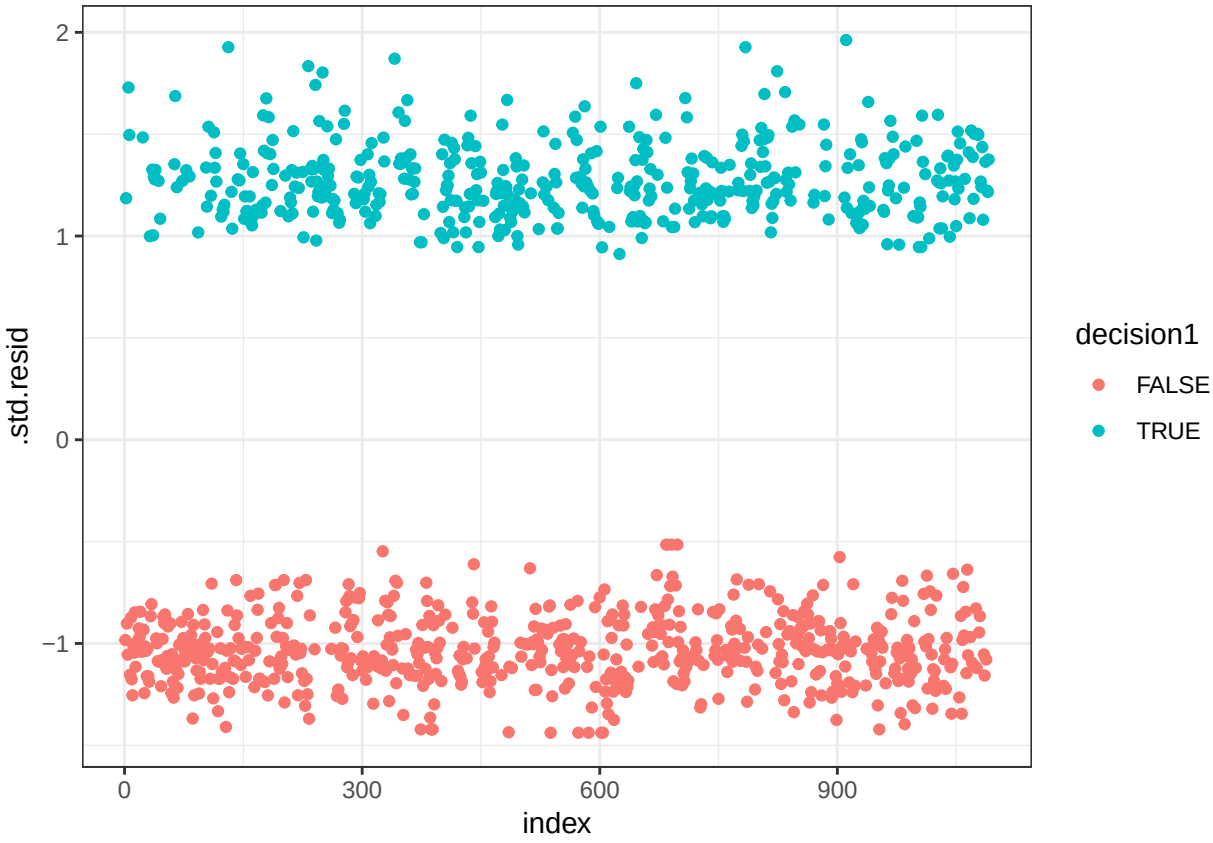
Last but not least, since the variables are intercorrlated, we take this into consideration and use interaction terms to overcome this issue.

Assumption_w/o interation




```
## # 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~   -1.65   0.342   1.91
## 2 TRUE           2.6        333           4 Ameri~   -1.06   0.369   1.65
## 3 TRUE           2.59       314           4 Ameri~   -1.65   0.342   1.91
## # ... 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: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected (2, 11, 2015)       1446440400
##  2 Univers~ Comp~ MS     S16    Accepted (28, 9, 2015)       1443412800
##  3 Univers~ (Com~ MS     F15    Rejected (24, 5, 2015)       1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    (27, 6, 2015)       1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted (2, 6, 2015)        1433217600
##  6 Univers~ Elec~ MS     F15    Accepted (14, 4, 2015)       1428984000
##  7 Univers~ Comp~ PhD    F15    Other    (20, 4, 2015)       1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected (7, 4, 2015)        1428379200
##  9 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## 10 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
```

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

```
## # A tibble: 1,091 x 3
##    ugrad_gpa gre_writing gre_total
##        <dbl>       <dbl>     <dbl>
##  1      3.5           3.5       325
##  2      3.68          4.5       335
##  3      3.96          5         318
##  4      3.93          5         332
##  5      3.3           4         314
##  6      3.76          5         325
##  7      4             5         337
##  8      3.25          3.5       325
##  9      3.7           3.5       322
## 10      3.7           3         322
## # ... with 1,081 more rows
```

```
##             ugrad_gpa gre_writing gre_total
## ugrad_gpa        1.00        0.13      0.27
## gre_writing      0.13        1.00      0.48
## gre_total        0.27        0.48      1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_writing gre_total
## ugrad_gpa               0           0
## gre_writing  0                      0
## gre_total    0         0
```

```
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa        1.00       0.19      0.18        0.13
## gre_verbal       0.19       1.00     -0.08        0.54
## gre_quant        0.18      -0.08      1.00        0.03
## gre_writing      0.13       0.54      0.03        1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa               0.0000     0.0000    0.0000
## gre_verbal  0.0000                 0.0069    0.0000
## gre_quant   0.0000    0.0069                 0.3114
## gre_writing 0.0000    0.0000     0.3114
```

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.

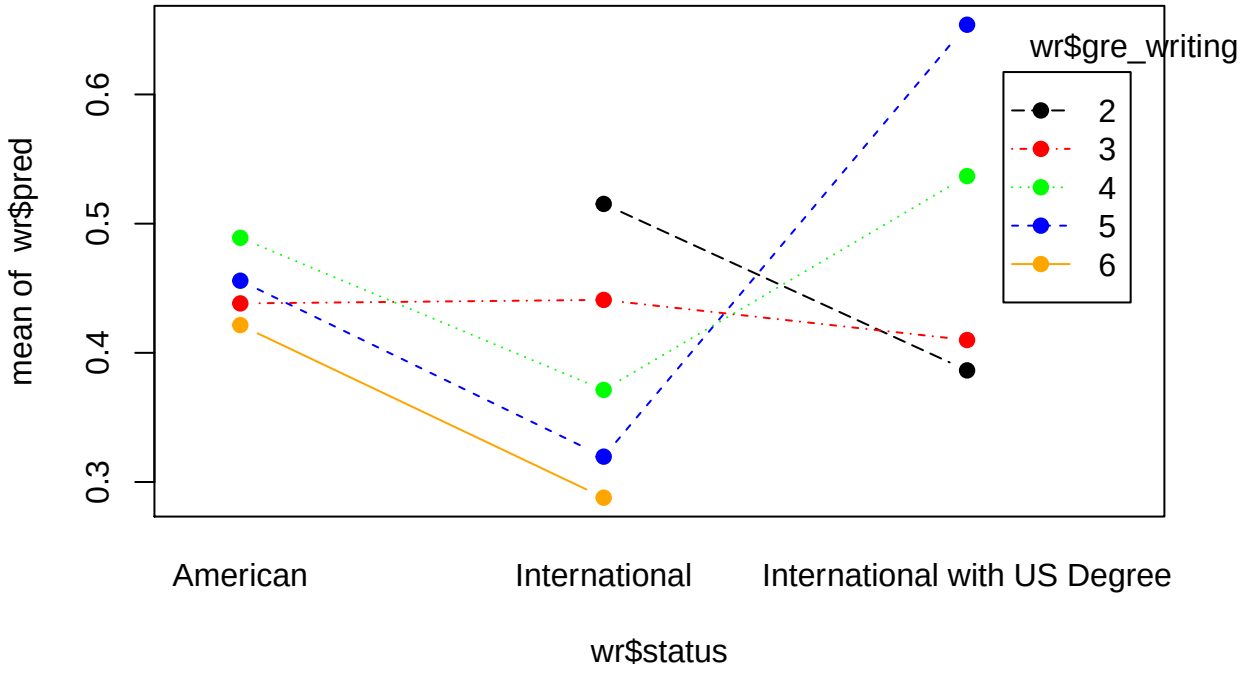
Second, logistic regression also assumes the linearity of independent variables.As shown in "The linearity of independent variables", the logit of GRE and undergraduate gpa are fairly linear to the accepted probability in logit scale. However, the scatter plots of gre_writing fits a parabola, instead 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 corrlated 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: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected (2, 11, 2015)       1446440400
##  2 Univers~ Comp~ MS     S16    Accepted (28, 9, 2015)       1443412800
##  3 Univers~ (Com~ MS     F15    Rejected (24, 5, 2015)       1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    (27, 6, 2015)       1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted (2, 6, 2015)        1433217600
##  6 Univers~ Elec~ MS     F15    Accepted (14, 4, 2015)       1428984000
##  7 Univers~ Comp~ PhD    F15    Other    (20, 4, 2015)       1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected (7, 4, 2015)        1428379200
##  9 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## 10 Univers~ Comp~ PhD    F15    Other    (16, 4, 2015)       1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <int>,
## #   is_new_gre <lgl>, status <chr>, comments <chr>, decision1 <lgl>,
## #   GRE_Total <dbl>, pred <dbl>
```




```
## 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      1085     1444.7
## 2      1079     1435.4  6   9.2779   0.1585
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: decision1
##                     LR Chisq Df Pr(>Chisq)
## ugrad_gpa             17.480  1  2.904e-05 ***
## GRE_Total             11.910  1  0.0005584 ***
## gre_writing           11.308  1  0.0007716 ***
## status                32.334  3  4.449e-07 ***
## ugrad_gpa:status       1.154  2  0.5616588
## GRE_Total:status       2.097  2  0.3504976
## gre_writing:status     3.014  2  0.2215837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 2 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(>Chi) for two models is 0.1581, which is bigger than siginificant level 0.05. Therefore, two models are not significantly different. Pr(>Chi) for ugrad_gpa, GRE_Total, gre_writing and status are all smaller than siginificant level 0.05, while all the interaction effect is not signficant. Therefore, the anova table indicates that the main effect are significant, and interaction effect is not significant.

Discussion

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, reccomendations, 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.