Homework 6: Modeling effects of SAT Scores and GPA on getting accepted to NYU

Jaehyun (Alex Lee)

Prof Simonoff

Due Dec 21st, 2017

In the following regression analysis, I am aiming to model a relationship between probability of getting accepted to NYU given various academic standards. Specifically, I will be performing binary logistic regression where values of 1 represent observations that were accepted to NYU and 0 for those who were either denied or waitlisted. I have decided to model such binary response variable with potential predictors that are thought of being as academically most important in getting accepted to colleges. The predicting variables are as follows: Unweighted GPAs, SAT Math Scores, Writing Scores, and Critical Reading Scores. Note that the data are of applicants who applied as class of 2018 (the class I applied for), and the SAT format is of the old version that contains 3 separate sections respectively as opposed to the new SAT. I have collected my data from www.collegedata.com .

Note that the following regression analysis only uses predicting variables that are academic related to evaluate their effects on the probability of acceptance. The data represent a retrospective sample of 37 applicants who were accepted to NYU as class of 2018 and 37 applicants who were either rejected or waitlisted (since being waitlisted does not qualify as being adequate enough to be accepted at once). This is a retrospective study because the sampling was based on the response variable (acceptance), and has nothing to do with time-related aspect to the problem.

Before taking a look at the logistic regression output, I have made side by side boxplots of predicting variables to see if any of them has apparent predictive power.









There seems to be pretty clear separations between all 3 categories of SAT score – Math, Critical Reading, and Writing – between accepted students and students who were not accepted. Although the unweighted GPA seems to be separated, the separation does not seem as apparent as the other 3 variables which implies a bit of less predictive power for that variable. Note that the notion of non-constant variance is not relevant here. There also seems to be a couple of outliers that we may have to pay further attention to. But for now, let’s first take a look at the regression output to get a sense of the regression:

Binary Logistic Regression: Status versus UW, Math, CR, W

\* WARNING \* When the data are in the Response/Frequency format, the Residuals versus fits  
plot is unavailable.

Method

|  |  |
| --- | --- |
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 74 |

Response Information

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Count |  |
| Status | 1 | 37 | (Event) |
|  | 0 | 37 |  |
|  | Total | 74 |  |

Deviance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 4 | 20.353 | 5.0881 | 20.35 | 0.000 |
| UW | 1 | 1.314 | 1.3136 | 1.31 | 0.252 |
| Math | 1 | 1.213 | 1.2126 | 1.21 | 0.271 |
| CR | 1 | 2.341 | 2.3406 | 2.34 | 0.126 |
| W | 1 | 0.982 | 0.9818 | 0.98 | 0.322 |
| Error | 69 | 82.233 | 1.1918 |  |  |
| Total | 73 | 102.586 |  |  |  |

Model Summary

|  |  |  |
| --- | --- | --- |
| Deviance R-Sq | Deviance R-Sq(adj) | AIC |
| 19.84% | 15.94% | 92.23 |

Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Coef | SE Coef | VIF |
| Constant | -15.54 | 4.51 |  |
| UW | 1.127 | 0.997 | 1.11 |
| Math | 0.00456 | 0.00417 | 1.25 |
| CR | 0.00730 | 0.00484 | 1.43 |
| W | 0.00534 | 0.00545 | 1.79 |

Odds Ratios for Continuous Predictors

|  |  |  |
| --- | --- | --- |
|  | Odds Ratio | 95% CI |
| UW | 3.0851 | (0.4372, 21.7709) |
| Math | 1.0046 | (0.9964, 1.0128) |
| CR | 1.0073 | (0.9978, 1.0169) |
| W | 1.0054 | (0.9947, 1.0161) |

Regression Equation

|  |  |  |
| --- | --- | --- |
| P(1) | = | exp(Y')/(1 + exp(Y')) |

|  |  |  |
| --- | --- | --- |
| Y' | = | -15.54 + 1.127 UW + 0.00456 Math + 0.00730 CR + 0.00534 W |

Goodness-of-Fit Tests

|  |  |  |  |
| --- | --- | --- | --- |
| Test | DF | Chi-Square | P-Value |
| Deviance | 69 | 82.23 | 0.132 |
| Pearson | 69 | 72.85 | 0.353 |
| Hosmer-Lemeshow | 8 | 13.12 | 0.108 |

Measures of Association

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pairs | Number | Percent | Summary Measures | Value |
| Concordant | 1082 | 79.0 | Somers’ D | 0.58 |
| Discordant | 282 | 20.6 | Goodman-Kruskal Gamma | 0.59 |
| Ties | 5 | 0.4 | Kendall’s Tau-a | 0.30 |
| Total | 1369 | 100.0 |  |  |

*Association is between the response variable and predicted probabilities*

Fits and Diagnostics for Unusual Observations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Obs | Observed Probability | Fit | Resid | Std Resid |  |  |
| 3 | 0.000 | 0.854 | -2.419 | -2.47 | R |  |
| 7 | 0.000 | 0.806 | -2.040 | -2.08 | R |  |
| 40 | 1.000 | 0.626 | 0.772 | 0.89 |  | X |
| 73 | 0.000 | 0.885 | -2.775 | -2.85 | R |  |

*R  Large residual  
X  Unusual X*

Pearson Residual Plots for Status

First noticeable thing is that overall test on all four coefficients is very statistically significant with a p value of 0. However, contrary to my expectations, none of the four predictors were significant at a statistical level as they all hold p values that exceed 0.05. Of the four, Critical Reading is slightly marginally statistically significant with a p value of 0.126. Although the individual coefficients are not statistically significant, if we were to interpret them, they would be as follows: the Math coefficient says that an increase of one point in the Math section of the SAT is associated with an increase in the odds of getting into NYU as a class of 2018 by 0.4% holding all else fixed; an increase of one point on the Critical Reading section is associated with an increase in the odds of getting into NYU as a class of 2018 by 0.73% holding all else fixed; an increase of one point on the Writing section is associated with an increase in the odds of getting into NYU as a class of 2018 by 0.54%. One thing to note here is that a point increase in any of the 3 sections in SAT is associated with an increase of very small odds in getting into NYU as a class of 2018. This is as such because a point on each of the 3 sections in SAT has relatively small effect on the overall score as each section is out of 800 points. Moreover, more often than not, getting a single question wrong on any of these sections accounts for way more than a single point decrease. Thus, it seems reasonable that a single point increase in SAT would be associated with a very small increase in the odds of getting accepted to NYU. On the contrary, a point increase in Unweighted GPA is associated with an increase of odds of getting into NYU as a class of 2018 by more than 200%. This seems like an oddly huge number. However, taking into account that Unweighted GPA is out of 4.0 scale and that most students who apply for top notch school like NYU have GPA varying at mostly 3.0 ~ 4.0 level (rarely at 2.0 level), such output makes intuitive sense (since a single grade point difference in GPA implies a huge difference in high school academic results for students). Although the individual coefficients do not seem to be statistically significant, the model seems to fit reasonably. All three goodness of fit tests hold p values that exceed 5%; specifically, the Hosmer-Lemeshow test has a p value of 0.108 which implies a marginally adequate goodness-of-fit (Hosmer-Lemeshow is the one that is applicable in this analysis). Although the Hosmer-Lemeshow result implies a reasonable goodness of fit, it seems to be fitting well only at a statistically marginal level, and thus we should make further effort in improving the goodness of fit. Finally, there are 79% of concordant pairs and 20.6 percent of discordant pairs which imply good to excellent separation.

Before taking a look at unusual observations however, it seems more important that we first address statistical significance of the predictors. It was a little puzzling to find out that the overall regression was very statistically significant but the individual predicting variables were not. There seems to be correlation between predicting variables and it seems intuitively correct that we address this problem through simplifying the model. My original intention of using SAT scores by section was to see if I could extract insight on what section of the SAT could provide predictive power on being accepted to NYU. However, the output implies that 3 score variables could be correlated to each other, which I will fix through combining the 3 scores into one variable that is out of 2400 points total.

After doing so provides with the following result:

Binary Logistic Regression: Status versus SAT\_Tot, UW

\* WARNING \* When the data are in the Response/Frequency format, the Residuals versus fits  
plot is unavailable.

Method

|  |  |
| --- | --- |
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 74 |

Response Information

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Count |  |
| Status | 1 | 37 | (Event) |
|  | 0 | 37 |  |
|  | Total | 74 |  |

Deviance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 2 | 20.175 | 10.088 | 20.18 | 0.000 |
| SAT\_Tot | 1 | 14.712 | 14.712 | 14.71 | 0.000 |
| UW | 1 | 1.359 | 1.359 | 1.36 | 0.244 |
| Error | 71 | 82.410 | 1.161 |  |  |
| Total | 73 | 102.586 |  |  |  |

Model Summary

|  |  |  |
| --- | --- | --- |
| Deviance R-Sq | Deviance R-Sq(adj) | AIC |
| 19.67% | 17.72% | 88.41 |

Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Coef | SE Coef | VIF |
| Constant | -15.46 | 4.38 |  |
| SAT\_Tot | 0.00572 | 0.00171 | 1.04 |
| UW | 1.104 | 0.960 | 1.04 |

Odds Ratios for Continuous Predictors

|  |  |  |
| --- | --- | --- |
|  | Odds Ratio | 95% CI |
| SAT\_Tot | 1.0057 | (1.0024, 1.0091) |
| UW | 3.0172 | (0.4598, 19.7979) |

Regression Equation

|  |  |  |
| --- | --- | --- |
| P(1) | = | exp(Y')/(1 + exp(Y')) |

|  |  |  |
| --- | --- | --- |
| Y' | = | -15.46 + 0.00572 SAT\_Tot + 1.104 UW |

Goodness-of-Fit Tests

|  |  |  |  |
| --- | --- | --- | --- |
| Test | DF | Chi-Square | P-Value |
| Deviance | 71 | 82.41 | 0.167 |
| Pearson | 71 | 72.97 | 0.413 |
| Hosmer-Lemeshow | 8 | 12.25 | 0.140 |

Measures of Association

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pairs | Number | Percent | Summary Measures | Value |
| Concordant | 1081 | 79.0 | Somers’ D | 0.58 |
| Discordant | 283 | 20.7 | Goodman-Kruskal Gamma | 0.59 |
| Ties | 5 | 0.4 | Kendall’s Tau-a | 0.30 |
| Total | 1369 | 100.0 |  |  |

*Association is between the response variable and predicted probabilities*

Pearson Residual Plots for Status

As expected, the p value of the new predictor variable, SAT\_Tot, which is a combined version of the previous 3 test scores is very statistically significant with a p value of 0. The goodness of fit has not changed much as the p value of Hosmer-Lemeshow as gone up from 0.108 to 0.140 and thus we should still investigate the effects of unusual observations if any. The Somer’s D is 0.58 which is satisfactory but seems as if it could be improved. The regression overall is very statistically significant as expected and with a new strongly statistically significant variable, SAT\_Tot, it seems that this variable may be mostly accounting for the significance of the overall regression.

|  |  |  |
| --- | --- | --- |
| SPEARRES | HI | COOK |
| 0.61547 | 0.042226 | 0.0055669 |
| -0.5627 | 0.029602 | 0.0032196 |
| -2.50414 | 0.036372 | 0.0788949 |
| -0.6652 | 0.025138 | 0.0038034 |
| 0.75968 | 0.027334 | 0.0054061 |
| -1.89628 | 0.055592 | 0.0705569 |
| -2.03937 | 0.038275 | 0.0551735 |
| -0.56708 | 0.059589 | 0.0067922 |
| 1.42125 | 0.044791 | 0.0315728 |
| 0.68195 | 0.036014 | 0.0057914 |
| -0.42982 | 0.079848 | 0.0053439 |
| 0.52426 | 0.03298 | 0.0031245 |
| 0.51051 | 0.033042 | 0.0029685 |
| 0.59095 | 0.035594 | 0.0042963 |
| -0.63085 | 0.040918 | 0.0056596 |
| -0.69124 | 0.079075 | 0.0136758 |
| 1.38629 | 0.032785 | 0.0217139 |
| 0.74344 | 0.037021 | 0.0070829 |
| 0.64393 | 0.035687 | 0.0051149 |
| -0.91091 | 0.036166 | 0.0103781 |
| -0.38044 | 0.059495 | 0.0030519 |
| -0.62801 | 0.029433 | 0.0039867 |
| 2.16462 | 0.04131 | 0.0673012 |
| -0.4438 | 0.034414 | 0.0023399 |
| -0.81753 | 0.035665 | 0.0082394 |
| 0.87061 | 0.027787 | 0.0072211 |
| 1.33651 | 0.032763 | 0.0201684 |
| 0.35335 | 0.03672 | 0.0015865 |
| 0.58198 | 0.031221 | 0.0036385 |
| 1.00352 | 0.027892 | 0.0096316 |
| 1.00247 | 0.018085 | 0.0061697 |
| -1.29514 | 0.037599 | 0.0218442 |
| -1.56554 | 0.026689 | 0.0224017 |
| -0.23767 | 0.032134 | 0.0006251 |
| 0.85924 | 0.040372 | 0.0103536 |
| -1.59873 | 0.028392 | 0.0248957 |
| 0.63847 | 0.026905 | 0.003757 |
| -0.40248 | 0.037622 | 0.0021109 |
| -1.1839 | 0.103033 | 0.0536675 |
| 0.66042 | 0.032079 | 0.0048185 |
| -1.54432 | 0.027425 | 0.0224173 |
| -0.59543 | 0.028297 | 0.0034415 |
| -0.5305 | 0.082391 | 0.008423 |
| 0.37346 | 0.036967 | 0.0017846 |
| 0.27591 | 0.034768 | 0.000914 |
| 1.74068 | 0.053361 | 0.0569322 |
| -0.27621 | 0.043955 | 0.0011692 |
| -0.75548 | 0.06488 | 0.0131998 |
| -0.88087 | 0.051661 | 0.0140898 |
| -0.65598 | 0.041922 | 0.0062763 |
| -0.59543 | 0.028297 | 0.0034415 |
| 1.37503 | 0.0324 | 0.0211033 |
| 0.45386 | 0.035597 | 0.0025344 |
| -0.39456 | 0.066316 | 0.0036857 |
| 1.25805 | 0.058299 | 0.0326605 |
| -0.94315 | 0.051777 | 0.0161906 |
| -0.61834 | 0.030425 | 0.0039994 |
| 0.59095 | 0.035594 | 0.0042963 |
| 0.7604 | 0.02227 | 0.00439 |
| 0.88461 | 0.041314 | 0.0112408 |
| 1.47894 | 0.026748 | 0.0200374 |
| -0.83753 | 0.024272 | 0.0058165 |
| -0.63085 | 0.040918 | 0.0056596 |
| 0.40772 | 0.0369 | 0.002123 |
| 0.4608 | 0.034843 | 0.0025552 |
| 0.97349 | 0.020549 | 0.0066276 |
| 1.14508 | 0.020117 | 0.0089731 |
| 0.29744 | 0.035497 | 0.0010853 |
| -0.5027 | 0.08678 | 0.0080046 |
| 0.42595 | 0.091293 | 0.0060758 |
| 1.03595 | 0.022505 | 0.0082359 |
| -0.63591 | 0.02929 | 0.0040673 |
| -2.62059 | 0.036921 | 0.0877586 |
| -0.80386 | 0.047792 | 0.0108111 |

There are 2 apparent outliers which are observation 3 and 73. Observation 3 has Unweighted GPA of 3.97 and cumulative SAT score of 2250 (Math 760, Writing 770, and Critical Reading of 720), and observation 73 has Unweighted GPA of 4.0 and cumulative SAT score of 2260 (Math 720, Writing 740, and Reading 800). Although the rationale for these students not having been accepted is unknown (maybe their outstanding academic record made their application to NYU seem like a “safety school”), they both are clear outliers as they both hold SPEARRES values greater than + or – 2.5. There aren’t any points whose Cook’s Distance value exceeds 1 and I consequently did not flag any points on that.

Here are the residual plots. We should use these residual plots as a rough guideline to get a sense of what is going on. Although not too apparent from the normal probability plot, there are 2 points towards the bottom left that exceed the -2.5 boundary which are the points that I have just previously mentioned. Also, there seems to be a marginally unusual point towards the far top right who was a student accepted with Unweighted GPA of 3.26 and cumulative Sat score of 1810. A potential explanation could be that this student might have gotten accepted to a program that does not value academic standing as much as other programs such as any majors related to arts (in which case portfolio could have more weight on acceptance decision as opposed to academic standing).



If we omit these to observations and try to fit a model using the 2 predictor model, the regression output is as follows:

Binary Logistic Regression: Status versus SAT\_Tot, UW

\* WARNING \* When the data are in the Response/Frequency format, the Residuals versus fits  
plot is unavailable.

Method

|  |  |
| --- | --- |
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 72 |

Response Information

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Count |  |
| Status | 1 | 37 | (Event) |
|  | 0 | 35 |  |
|  | Total | 72 |  |

Deviance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 2 | 26.412 | 13.206 | 26.41 | 0.000 |
| SAT\_Tot | 1 | 19.476 | 19.476 | 19.48 | 0.000 |
| UW | 1 | 2.262 | 2.262 | 2.26 | 0.133 |
| Error | 69 | 73.346 | 1.063 |  |  |
| Total | 71 | 99.758 |  |  |  |

Model Summary

|  |  |  |
| --- | --- | --- |
| Deviance R-Sq | Deviance R-Sq(adj) | AIC |
| 26.48% | 24.47% | 79.35 |

Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Coef | SE Coef | VIF |
| Constant | -19.75 | 5.10 |  |
| SAT\_Tot | 0.00721 | 0.00198 | 1.01 |
| UW | 1.51 | 1.03 | 1.01 |

Odds Ratios for Continuous Predictors

|  |  |  |
| --- | --- | --- |
|  | Odds Ratio | 95% CI |
| SAT\_Tot | 1.0072 | (1.0033, 1.0112) |
| UW | 4.5091 | (0.6034, 33.6935) |

Regression Equation

|  |  |  |
| --- | --- | --- |
| P(1) | = | exp(Y')/(1 + exp(Y')) |

|  |  |  |
| --- | --- | --- |
| Y' | = | -19.75 + 0.00721 SAT\_Tot + 1.51 UW |

Goodness-of-Fit Tests

|  |  |  |  |
| --- | --- | --- | --- |
| Test | DF | Chi-Square | P-Value |
| Deviance | 69 | 73.35 | 0.338 |
| Pearson | 69 | 67.51 | 0.528 |
| Hosmer-Lemeshow | 8 | 13.49 | 0.096 |

Measures of Association

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pairs | Number | Percent | Summary Measures | Value |
| Concordant | 1069 | 82.5 | Somers’ D | 0.65 |
| Discordant | 222 | 17.1 | Goodman-Kruskal Gamma | 0.66 |
| Ties | 4 | 0.3 | Kendall’s Tau-a | 0.33 |
| Total | 1295 | 100.0 |  |  |

*Association is between the response variable and predicted probabilities*

The output given above is of 2 predictor model with outliers removed. The first thing we can notice is that the p value of Unweighted GPA has gone down to 0.133 which I could say is marginally statistically significant. The percentage of concordant pairs have gone up to 82.5% and discordant pairs have gone down to 17.1%, indicating the model’s better separation. However, one puzzling thing is that the p value of Hosmer-Lemeshow goodness of fit test has gone down slightly to 0.096 which still implies that the data fits but not at a perfect level. This could be potentially attributed to small expected counts (the observation is small enough but getting rid of two points may have had an effect). The Somer’s D has increased some amount to 0.65 which implies adequate to decent separation. This is somewhat expected given that we have removed two outliers.

Having run a 2-predictor regression model after removing outliers, I naturally came to question if removing outliers from the original model could possibly improve goodness of fit while giving each of the SAT score predictors predictive power. In other words, I want to check out if removal of outliers from the original model could reduce correlation of predictors (if any) while also improving goodness of fit. If the regression output implies such result, then I would be justified to used such model.

The following output is of the original 4 predictor model indicating clues of unusual observations:

|  |  |  |
| --- | --- | --- |
| SPEARRES | HI | COOK |
| 0.63399 | 0.055028 | 0.004681 |
| -0.53814 | 0.038711 | 0.002332 |
| -2.46655 | 0.038168 | 0.048284 |
| -0.65162 | 0.028885 | 0.002526 |
| 0.8276 | 0.077675 | 0.011536 |
| -1.79361 | 0.082761 | 0.058053 |
| -2.08362 | 0.041152 | 0.037265 |
| -0.54768 | 0.068308 | 0.004398 |
| 1.46365 | 0.077868 | 0.03618 |
| 0.67504 | 0.036755 | 0.003477 |
| -0.46661 | 0.104678 | 0.005091 |
| 0.48959 | 0.079695 | 0.004151 |
| 0.49641 | 0.040292 | 0.002069 |
| 0.60042 | 0.03814 | 0.002859 |
| -0.60088 | 0.062201 | 0.00479 |
| -0.6955 | 0.094019 | 0.01004 |
| 1.4911 | 0.058284 | 0.027522 |
| 0.72685 | 0.03993 | 0.004395 |
| 0.67539 | 0.046345 | 0.004434 |
| -0.89853 | 0.072132 | 0.012552 |
| -0.39343 | 0.100714 | 0.003467 |
| -0.62694 | 0.042692 | 0.003506 |
| 1.99267 | 0.092679 | 0.081118 |
| -0.43927 | 0.03962 | 0.001592 |
| -0.79431 | 0.06255 | 0.00842 |
| 0.87764 | 0.035247 | 0.005628 |
| 1.3786 | 0.071317 | 0.02919 |
| 0.36032 | 0.040812 | 0.001105 |
| 0.5584 | 0.03888 | 0.002523 |
| 1.03845 | 0.042767 | 0.009636 |
| 1.06736 | 0.086393 | 0.021546 |
| -1.28538 | 0.039367 | 0.013542 |
| -1.46257 | 0.0861 | 0.040306 |
| -0.20824 | 0.055825 | 0.000513 |
| 0.83107 | 0.059525 | 0.008743 |
| -1.85097 | 0.132236 | 0.104419 |
| 0.63315 | 0.027581 | 0.002274 |
| -0.40032 | 0.044502 | 0.001493 |
| -1.30936 | 0.1395 | 0.055587 |
| 0.88671 | 0.241303 | 0.050014 |
| -1.44346 | 0.070888 | 0.031794 |
| -0.62311 | 0.063128 | 0.005232 |
| -0.52653 | 0.097307 | 0.005977 |
| 0.37723 | 0.050577 | 0.001516 |
| 0.26953 | 0.035958 | 0.000542 |
| 1.69276 | 0.064487 | 0.039504 |
| -0.2677 | 0.044005 | 0.00066 |
| -0.77178 | 0.080281 | 0.010399 |
| -0.86795 | 0.076512 | 0.012483 |
| -0.6758 | 0.04677 | 0.004482 |
| -0.58083 | 0.041538 | 0.002924 |
| 1.35658 | 0.057174 | 0.022319 |
| 0.45773 | 0.043164 | 0.00189 |
| -0.38908 | 0.078812 | 0.00259 |
| 1.45868 | 0.173852 | 0.089551 |
| -1.00418 | 0.081738 | 0.017952 |
| -0.62786 | 0.057051 | 0.00477 |
| 0.64169 | 0.060366 | 0.005291 |
| 0.71122 | 0.070839 | 0.007713 |
| 0.84525 | 0.113227 | 0.018245 |
| 1.48316 | 0.050638 | 0.023467 |
| -0.84728 | 0.07546 | 0.011718 |
| -0.78235 | 0.15903 | 0.023149 |
| 0.39119 | 0.04366 | 0.001397 |
| 0.46351 | 0.039328 | 0.001759 |
| 0.95087 | 0.050702 | 0.009658 |
| 1.14569 | 0.034233 | 0.009305 |
| 0.28549 | 0.036814 | 0.000623 |
| -0.45603 | 0.11664 | 0.005492 |
| 0.41994 | 0.091618 | 0.003557 |
| 1.03096 | 0.027987 | 0.006121 |
| -0.6933 | 0.064153 | 0.00659 |
| -2.85463 | 0.054674 | 0.094261 |
| -0.88964 | 0.086757 | 0.015038 |

There are 2 outliers that I can identify. Again, there is not a point whose Cook’s Distance exceeds the value of 1. The two outliers are the same ones as before, and I will go ahead and remove the 2 points to see if doing that makes any difference to the regression output of the original model.

Doing so yields the following results:

Binary Logistic Regression: Status versus UW, Math, CR, W

\* WARNING \* When the data are in the Response/Frequency format, the Residuals versus fits  
plot is unavailable.

Method

|  |  |
| --- | --- |
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 72 |

Response Information

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Count |  |
| Status | 1 | 37 | (Event) |
|  | 0 | 35 |  |
|  | Total | 72 |  |

Deviance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 4 | 27.0254 | 6.7563 | 27.03 | 0.000 |
| UW | 1 | 2.4703 | 2.4703 | 2.47 | 0.116 |
| Math | 1 | 1.8467 | 1.8467 | 1.85 | 0.174 |
| CR | 1 | 4.4294 | 4.4294 | 4.43 | 0.035 |
| W | 1 | 0.7914 | 0.7914 | 0.79 | 0.374 |
| Error | 67 | 72.7322 | 1.0856 |  |  |
| Total | 71 | 99.7576 |  |  |  |

Model Summary

|  |  |  |
| --- | --- | --- |
| Deviance R-Sq | Deviance R-Sq(adj) | AIC |
| 27.09% | 23.08% | 82.73 |

Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Coef | SE Coef | VIF |
| Constant | -20.49 | 5.41 |  |
| UW | 1.66 | 1.09 | 1.10 |
| Math | 0.00598 | 0.00446 | 1.19 |
| CR | 0.01100 | 0.00541 | 1.30 |
| W | 0.00503 | 0.00570 | 1.58 |

Odds Ratios for Continuous Predictors

|  |  |  |
| --- | --- | --- |
|  | Odds Ratio | 95% CI |
| UW | 5.2664 | (0.6232, 44.5053) |
| Math | 1.0060 | (0.9972, 1.0148) |
| CR | 1.0111 | (1.0004, 1.0218) |
| W | 1.0050 | (0.9939, 1.0163) |

Regression Equation

|  |  |  |
| --- | --- | --- |
| P(1) | = | exp(Y')/(1 + exp(Y')) |

|  |  |  |
| --- | --- | --- |
| Y' | = | -20.49 + 1.66 UW + 0.00598 Math + 0.01100 CR + 0.00503 W |

Goodness-of-Fit Tests

|  |  |  |  |
| --- | --- | --- | --- |
| Test | DF | Chi-Square | P-Value |
| Deviance | 67 | 72.73 | 0.295 |
| Pearson | 67 | 65.00 | 0.546 |
| Hosmer-Lemeshow | 8 | 9.47 | 0.304 |

Measures of Association

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pairs | Number | Percent | Summary Measures | Value |
| Concordant | 1070 | 82.6 | Somers’ D | 0.65 |
| Discordant | 222 | 17.1 | Goodman-Kruskal Gamma | 0.66 |
| Ties | 3 | 0.2 | Kendall’s Tau-a | 0.33 |
| Total | 1295 | 100.0 |  |  |

*Association is between the response variable and predicted probabilities*

The regression output seems pretty decent after removing two outliers from the original model. Although not all of the 3 SAT score predictor variables show statistical significance, Critical Reading has a p value of 0.035, indicating its predictive power. The overall regression is still very strong with good separation implied by 82.6% of concordant pairs and 17.1% of discordant pairs which is around the same as the improved 2 predictor model. Compared to the original model without removing outliers, the regression output seems to have improved overall. I in fact like this model compared to the improved 2 predictor model first because despite some potential existence of correlation among SAT score predictor variables, we could still infer some information out of this model. One thing that has become clearer is the importance of Critical Reading section’s score out of the three sections in increasing the odds of getting into NYU as a class of 2018. Moreover, given the very big p value (0.374) of Writing score, it seems that Writing section does not provide much predictive power in estimating the odds of our response variable. Finally, Math section seems to be marginally statistically significant, although our model still may have some degree of correlation (although the low VIF values imply that it does not exist..) between test score predictors. The second reason why I prefer this model over the simplified 2 predictor model is because, although the 2-predictor model is simpler, Hosmer-Lemeshow statistics reveal a stronger evidence against the null (same value for Somer’s D).



The new residual plots reveal an existence of one clear outlier. Although there is no end in identifying and removing unusual observations, doing so continuously would distort the model’s representability of the reality, but I am tempted to do so one more time to see if doing so could increase the predictive power of my individual variables. If I were to go ahead and remove an additional outlier, my rationale for doing so would be as follows: since this model aims to model the odds of being accepted to NYU given academic results of students, any observations that represent acceptance to the school despite seemingly inadequate academic standings could be justified of removal (for example, art majors). However, following the logic it does not seem justifiable to remove observations which have been declined or waitlisted despite seemingly adequate academic standings (because there is a lack of other academic aspects such as extracurricular activities which I have not included in the model which could account for getting denied despite high SAT Scores and Unweighted GPA). From the plot above, the noticeable outlier is of the latter and thus it should not be removed following my rationale.

Given the full model with outliers removed, it is of my interest to see if I can simplify the model even further while retaining similar level of fit. To do this, it is ideal to use the best subset model to provide me a guidance on which predicting variables I could use to simplify my model further. The quick and dirty way of getting the best subsets is to use the ordinary least squares best subsets regression, but best subsets regression is not offered for logistic model in minitab.

Although it is of my best interest to further simplify the full model with outliers removed, given this disadvantage, I am inclined to use the 4 predictor model. I cannot manually assess the best subsets, however the least I could do is to try to remove one or two predictors that I think are not adding predictive power to the model and see how Somer’s D and p values of other predictor variables change (to see trade of fit with predicting power). If the results are not satisfactory, then I will stick to the simplified 4 predictor model.

The regression output of the model with Writing Score (high p value) removed is as follows:

Binary Logistic Regression: Status versus UW, Math, CR

\* WARNING \* When the data are in the Response/Frequency format, the Residuals versus fits  
plot is unavailable.

Method

|  |  |
| --- | --- |
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 72 |

Response Information

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Count |  |
| Status | 1 | 37 | (Event) |
|  | 0 | 35 |  |
|  | Total | 72 |  |

Deviance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 3 | 26.234 | 8.745 | 26.23 | 0.000 |
| UW | 1 | 3.907 | 3.907 | 3.91 | 0.048 |
| Math | 1 | 3.531 | 3.531 | 3.53 | 0.060 |
| CR | 1 | 9.641 | 9.641 | 9.64 | 0.002 |
| Error | 68 | 73.524 | 1.081 |  |  |
| Total | 71 | 99.758 |  |  |  |

Model Summary

|  |  |  |
| --- | --- | --- |
| Deviance R-Sq | Deviance R-Sq(adj) | AIC |
| 26.30% | 23.29% | 81.52 |

Coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Coef | SE Coef | VIF |
| Constant | -20.72 | 5.44 |  |
| UW | 1.96 | 1.04 | 1.00 |
| Math | 0.00746 | 0.00408 | 1.02 |
| CR | 0.01332 | 0.00476 | 1.03 |

Odds Ratios for Continuous Predictors

|  |  |  |
| --- | --- | --- |
|  | Odds Ratio | 95% CI |
| UW | 7.1346 | (0.9347, 54.4579) |
| Math | 1.0075 | (0.9995, 1.0156) |
| CR | 1.0134 | (1.0040, 1.0229) |

Regression Equation

|  |  |  |
| --- | --- | --- |
| P(1) | = | exp(Y')/(1 + exp(Y')) |

|  |  |  |
| --- | --- | --- |
| Y' | = | -20.72 + 1.96 UW + 0.00746 Math + 0.01332 CR |

Goodness-of-Fit Tests

|  |  |  |  |
| --- | --- | --- | --- |
| Test | DF | Chi-Square | P-Value |
| Deviance | 68 | 73.52 | 0.302 |
| Pearson | 68 | 63.44 | 0.634 |
| Hosmer-Lemeshow | 8 | 9.59 | 0.295 |

Measures of Association

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pairs | Number | Percent | Summary Measures | Value |
| Concordant | 1068 | 82.5 | Somers’ D | 0.65 |
| Discordant | 225 | 17.4 | Goodman-Kruskal Gamma | 0.65 |
| Ties | 2 | 0.2 | Kendall’s Tau-a | 0.33 |
| Total | 1295 | 100.0 |  |  |

*Association is between the response variable and predicted probabilities*

Pearson Residual Plots for Status

The output seems quite favorable in fact. The Somer’s D has not changed with the value of 0.65 (which I can say at this point seems like a satisfactory fit). The regression overall is still very statistically significant and all of the 3 predictors are statistically significant except of Math which has a p value of 0.06 and is marginally significant (we could still consider it as adding predictive power to the overall model). As this new model improved individual predictors’ statistics while retaining the goodness of fit, I am justified to favor this model over the previous two models that have been discussed.



Here is a plot of residuals, from which we can identify an outlier towards the far bottom left. Below shows diagnostics for the three-predictor model; although there is still an indication for some unusual observations, following my logic previously mentioned on removing outliers, I will not pursue further in removing this one.

SPEARRES\_3 HI\_3 COOK\_3

0.48428 0.051819 0.003204

-0.45786 0.039157 0.002136

-0.59161 0.030658 0.002767

0.91565 0.045492 0.009990

-1.81544 0.074162 0.066001

-2.49024 0.038811 0.062599

-0.46712 0.064627 0.003769

1.22954 0.062464 0.025181

0.53898 0.039991 0.003025

-0.52473 0.103122 0.007915

0.27385 0.040323 0.000788

0.41005 0.037354 0.001631

0.46415 0.039641 0.002223

-0.62627 0.058979 0.006146

-0.53595 0.072635 0.005624

1.37559 0.055111 0.027591

0.58164 0.042912 0.003792

0.54851 0.048320 0.003819

-1.14417 0.045492 0.015599

-0.21367 0.036583 0.000433

-0.50235 0.033938 0.002216

2.01350 0.097797 0.109867

-0.42806 0.034803 0.001652

-0.67576 0.051222 0.006163

0.89091 0.028121 0.005741

1.10531 0.046833 0.015007

0.23862 0.030962 0.000455

0.41778 0.039283 0.001784

1.14392 0.030185 0.010182

0.73222 0.037086 0.005162

-1.47076 0.044301 0.025068

-1.73719 0.085132 0.070205

-0.18144 0.041609 0.000357

0.75627 0.060615 0.009226

-1.71136 0.090001 0.072415

0.53051 0.031281 0.002272

-0.28890 0.033934 0.000733

-1.38240 0.147912 0.082933

1.16782 0.162551 0.066180

-1.64073 0.073564 0.053440

-0.47443 0.041437 0.002432

-0.52466 0.098088 0.007484

0.22184 0.029633 0.000376

0.16248 0.022383 0.000151

1.65284 0.067944 0.049786

-0.22487 0.036180 0.000475

-0.71897 0.079627 0.011180

-1.08126 0.057046 0.017682

-0.63967 0.051008 0.005498

-0.60141 0.035864 0.003364

1.12061 0.042317 0.013872

0.39449 0.035560 0.001435

-0.42539 0.065587 0.003175

2.00902 0.089163 0.098777

-1.36277 0.054786 0.026911

-0.74958 0.025554 0.003684

0.52753 0.060389 0.004471

0.48653 0.055517 0.003479

0.81713 0.104560 0.019492

1.81823 0.034354 0.029404

-1.08249 0.038885 0.011852

-0.99090 0.160579 0.046958

0.24448 0.031129 0.000480

0.32155 0.034458 0.000922

0.99712 0.036627 0.009450

1.23012 0.026803 0.010419

0.17840 0.024558 0.000200

-0.30300 0.075199 0.001866

0.31349 0.072736 0.001927

0.85629 0.024515 0.004607

-0.64277 0.063256 0.006975

-0.90674 0.095472 0.021695

Since this is a retrospective study, these cannot be interpreted as genuine probabilities that a student with given academic standing will be accepted to NYU (at 2018 standards). However, they can be still be used to classify observations as accepted or not accepted. The following is a classification matrix, based on whether the estimated probability is above or below 0.5.

Tabulated Statistics: Status, Predict

Rows: Status   Columns: Predict

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | All |
|  |  |  |  |
| 0 | 24 | 11 | 35 |
|  | 33.33 | 15.28 | 48.61 |
|  |  |  |  |
| 1 | 11 | 26 | 37 |
|  | 15.28 | 36.11 | 51.39 |
|  |  |  |  |
| All | 35 | 37 | 72 |
|  | 48.61 | 51.39 | 100.00 |

*Cell Contents  
      Count  
      % of Total*

69.44% of the students were correctly classified, which compares to Cpro value of

(1.25)[0.4861 x 0.4861 + 0.5139 x 0.5139] = 0.63 which is 63%. This result is somewhat dismal but not quite unexpected given the mediocre strength of the logistic regression which we attempted to continuously improve. Note that we have taken 2 outliers out which could have been most likely misclassified, which would have lowered our classification rate.