CLASSIFYING STUDENT BASED ON ADMISSION TO AN IVY-LEAGUE SCHOOL

There comes a point in every student's life when he or she needs to decide on their future steps. Academic inflation is already forcing more and more people into pursuing higher education, as now more than ever a university degree seems like the prerequisite for professional success. The ivy league is usually every high achieving student's first target. These are the 8 most highly competitive athletic colleges on the East Coast and include Harvard University (Massachusetts), Yale University (Connecticut), Princeton University (New Jersey), Columbia University (New York), Brown University (Rhode Island), Dartmouth College (New Hampshire), University of Pennsylvania (Pennsylvania) and Cornell University (New York). These schools have historically served as a breeding-ground for high achievers, with many Nobel Prize winners being former Ivy League graduates. The admission process for these elite schools is highly competitive.

Applying to an Ivy league school is a lengthy and complicated process with multiple components, some objective and some not. Making sense of the subjective biases that come into play in the application review of aspects such as a student's essay, recommendations, extra-curricular activities, and other academic achievements is frankly very difficult. Other aspects of an applicant's profile, such as ones' GPA and standardized test scores, are much more easily quantifiable and hence studied. In this study, I decided to use easily quantifiable aspects of a student's application, namely their high school GPA, their SAT and ACT test scores and their class rank to determine how these variables affect whether or not the student is offered a place at an Ivy league school.

Data Collection

This analysis is based on the data gathered from College Data's Admission tracker for the year 2019[1]. The CollegeData Admissions Tracker displays self-reported admissions outcomes for comparison purposes and includes only students who have created a College Data Admissions Profile, not all students who applied, or will apply, to the colleges. The information collected included a student's Grade Point Average on a scale of 0-4, standardized test scores such as the SAT on a 0-800 scale and the ACT on a scale of 1-36 and where the student's class rank, as decimal of the top (for example a data point with RANK at .1 corresponds to a student belong to the top 10% of his/hers class). These were collected along the admission outcome for each student, 'Denied' corresponding to 0 and 'Accepted' or 'Will Attend' corresponding to 1. I gathered a dataset containing all the data points available for the 8 Ivy League schools, and after removing any points with missing values, I ended up with a dataset of 80 points.

The study that follows is a cohort study, recording the initial test scores, grades and class ranking of students and "following up" with them by the time their admission decision is released. Since sampling is based on a predictor (not the response), it is a prospective design

The dataset follows:

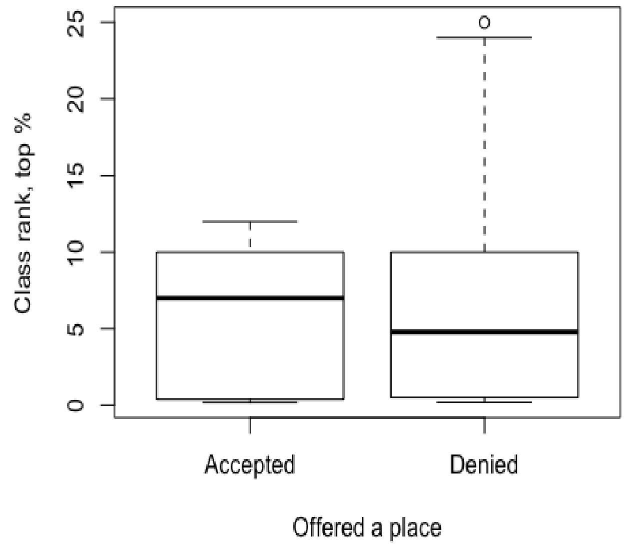
[1] The webpage I used can be found at: https://www.collegedata.com/en/prepare-and-apply/admissions-tracker/

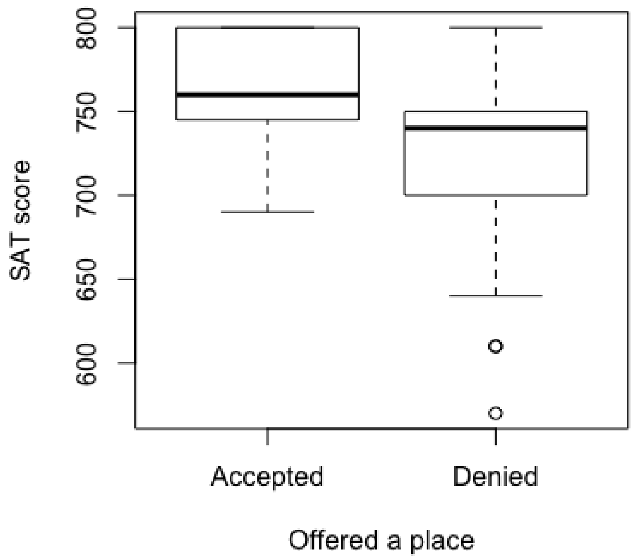
	A	B	C	D	E	F
1	GPA	SAT	ACT	RANK	DECISION	Binary
2	3.96	740	34	0.25	Denied	0
3	3.9	730	30	0.1	Denied	0
4	3.8	800	35	0.05	Accepted	1
5	4	780	34	0.1	Denied	0
6	4	750	31	0.1	Denied	0
7	3.9	700	31	0.25	Denied	0
8	3.67	750	32	0.2	Denied	0
9	3.83	800	34	0.2	Denied	0
10	3.74	730	34	0.1	Denied	0
11	4	800	36	0.1	Will Attend	1
12	4	740	35	0.1	Denied	0
13	3.8	610	34	0.3	Denied	0
14	3.91	710	32	0.1	Denied	0
15	4	740	32	0.1	Will Attend	1
16	3.7	800	21	0.3	Denied	0
17	4	780	34	0.1	Denied	0
18	3.96	640	26	0.25	Denied	0
19	4	750	31	0.1	Denied	0
20	3.9	700	31	0.1	Denied	0
21	3.9	760	33	0.25	Denied	0
22	4	800	36	0.1	Denied	0
23	3.97	660	29	0.1	Denied	0
24	4	800	36	0.15	Denied	0
25	3.91	710	32	0.1	Denied	0
26	4	710	30	0.1	Will Attend	1
27	4	740	31	0.1	Denied	0
28	4	750	31	0.1	Denied	0
29	3.67	750	32	0.1	Denied	0
30	4	800	36	0.1	Denied	0
31	4	760	32	0.05	Will Attend	1
32	3.16	690	29	0.1	Accepted	1
33	4	790	34	0.1	Will Attend	1
34	4	740	30	0.1	Denied	0
35	3.96	740	34	0.25	Denied	0
36	3.9	700	31	0.15	Denied	0
37	3.74	730	34	0.1	Denied	0
38	4	800	36	0.1	Accepted	1
39	3.97	660	29	0.1	Denied	0
40	4	760	32	0.1	Will Attend	1

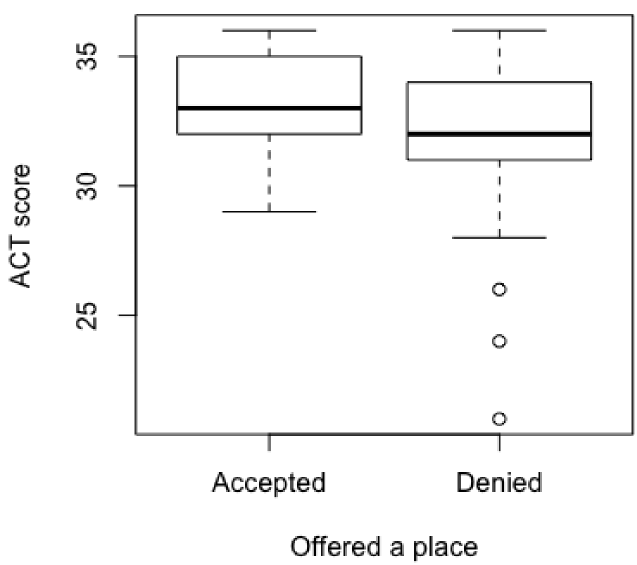
41	4	800	36	0.1	Denied	0
42	3.96	740	34	0.25	Denied	0
43	3.9	700	31	0.1	Denied	0
44	4	800	36	0.15	Denied	0
45	3.8	750	33	0.2	Denied	0
46	4	800	35	0.1	Will Attend	1
47	4	740	35	0.25	Denied	0
48	3.99	780	36	0.1	Will Attend	1
49	3.94	570	24	0.25	Denied	0
50	3.8	670	32	0.1	Denied	0
51	3.5	670	28	0.1	Denied	0
52	4	740	31	0.3	Denied	0
53	4	780	34	0.1	Denied	0
54	4	750	31	0.1	Denied	0
55	3.9	700	31	0.1	Denied	0
56	3.67	750	32	0.1	Accepted	1
57	3.83	800	34	0.05	Accepted	1
58	3.99	770	33	0.1	Denied	0
59	4	800	36	0.1	Accepted	1
60	3.8	750	33	0.1	Denied	0
61	3.97	650	30	0.1	Denied	0
62	3.6	720	30	0.25	Denied	0
63	3.09	610	28	0.25	Denied	0
64	3.7	710	32	0.1	Accepted	1
65	3.9	700	31	0.1	Denied	0
66	3.67	750	32	0.1	Will Attend	1
67	3.92	780	33	0.1	Will Attend	1
68	4	760	34	0.05	Accepted	1
69	3.85	740	32	0.1	Denied	0
70	3.8	720	34	0.1	Will Attend	1
71	3.87	780	33	0.05	Will Attend	1
72	3.9	730	30	0.1	Denied	0
73	4	780	34	0.25	Denied	0
74	3.7	710	32	0.1	Denied	0
75	3.9	700	31	0.05	Will Attend	1
76	3.83	800	34	0.1	Denied	0
77	4	800	36	0.1	Accepted	1
78	3.97	660	29	0.1	Denied	0
79	4	760	32	0.1	Will Attend	1
80	3.97	650	30	0.1	Denied	0
81	3.91	710	32	0.1	Denied	0
82						

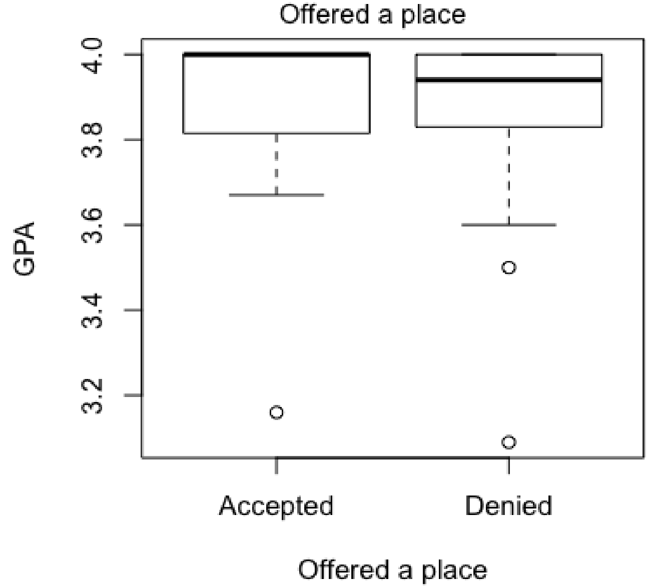
Data Analysis

A good way to get a feeling for the predictive power of the individual variables is to construct side-by-side boxplots, to see if there is separation between the two groups on the variables. This does not take into account the variables having joint effects and doesn't necessarily imply that a linear logistic model is appropriate but is a helpful visualization. The plots follow:









The predictors show clear separation between accepted and not accepted students, in the ways that would have been expected. Student accepted tend to have higher ACT and SAT scores, higher GPA, when compared to students not admitted. Interestingly, separation in terms of class rank isn't clear, however we do see more variability in the students who weren't admitted. Logistic regression can be used to analyze the relationship between the individual student's performance variables and the probability of admission more precisely. The indicator of admission is the target variable, 0 standing for rejection while 1 standing for acceptance.

Here is the output for a logistic regression model fit to these data.

```
glm(formula = Binary ~ GPA + SAT + ACT + RANK, family = binomial,
    data = df, maxit = 500)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3868  -0.7725  -0.5729   1.0946   2.1976

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.372756   7.084998  -0.899   0.3684
GPA         -3.094077   1.916105  -1.615   0.1064
SAT          0.018423   0.008517   2.163   0.0305 *
ACT          0.116999   0.138813   0.843   0.3993
RANK        -0.016224   0.050039  -0.324   0.7458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.984  on 79  degrees of freedom
Residual deviance: 83.837  on 75  degrees of freedom
AIC: 93.837

Number of Fisher Scoring iterations: 5
```

Linear Predictor

Y = -6.37 + -3.09 * GPA + 0.02 * SAT + 0.12 * ACT + -0.02 * RANK + e

A 95% confidence interval for the odds ratio

```
                   2.5 %      97.5 %
(Intercept) 1.590690e-09 1832.772837
GPA         1.059906e-03    1.937541
SAT         1.001731e+00    1.035740
ACT         8.563566e-01    1.475603
RANK        8.919917e-01    1.085293
```

Odds ratios

```
      GPA       SAT       ACT      RANK
0.0453168 1.0185935 1.1241186 0.9839066
```

The likelihood ratio test for whether all slopes equal 0

```
        gstat
[1,] 12.14677 0.01629257
```

Likelihood ratio tests for each slope, along with AIC values for the model that omits that variable

```
Single term deletions

Model:
Binary ~ GPA + SAT + ACT + RANK
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      83.837 93.837
GPA     1   86.306 94.306 2.4695  0.11607
SAT     1   88.534 96.534 4.6970  0.03022 *
ACT     1   84.596 92.596 0.7588  0.38371
RANK    1   83.943 91.943 0.1062  0.74452
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

VIF Values

```
     GPA      SAT      ACT     RANK
1.554194 1.639199 1.679801 1.253206
```

```
> AIC(R1)
[1] 93.83691
```

The AIC value given here is not the same as that given by Minitab, but comparisons between AIC values for different models will be the same (which is all that matters).

```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  df$Binary, fitted(R1)
X-squared = 10.598, df = 8, p-value = 0.2255
```

Summary measures of association in the stats component, with Somers D being represented by Dxy

```
Sum of squared errors     Expected value|H0                SD                 Z                 P
           14.1006999            14.0589613         0.2381351         0.1752728         0.8608653

> R1.lrm$stats
         Obs    Max Deriv   Model L.R.         d.f.            P            C          Dxy        Gamma        Tau-a
8.000000e+01 9.601720e-09 1.214677e+01 4.000000e+00 1.629257e-02 7.498093e-01 4.996186e-01 5.073586e-01 2.072785e-01
          R2        Brier            g           gr           gp
2.016040e-01 1.762587e-01 1.217874e+00 3.379993e+00 1.960981e-01
```

The likelihood ratio test for whether all slopes equal 0 labeled by R as gstat provides a test of the overall regression. In this case the test statistic equals 12.1467, with a p-value less than .01629, signaling that at least one slope equals 0, so we strongly reject the null hypothesis of no relationship. Similarly, individual slopes are tested. It can be seen here that only SAT scores seem to be statistically significant with a p-value of .03022, while GPA, ACT scores and class rank are not statistically significant at a 95% significance level.

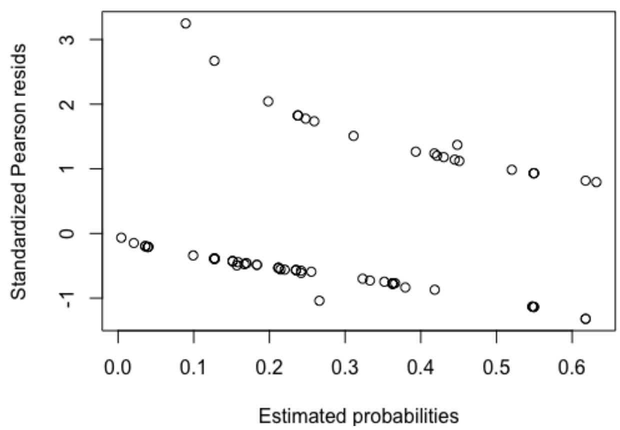
The analysis shows that a percentage point higher GPA is associated with multiplying the odds of a student getting into an IV league school by .0453, a percentage point increase in SAT is associated with 1.859% higher odds of getting accepted, while a percentage point increase in ACT is associate with 12.4% higher odds of getting accepted. A percentage point increase in Rank would actually lower the odds of getting accepted by 1.61%.

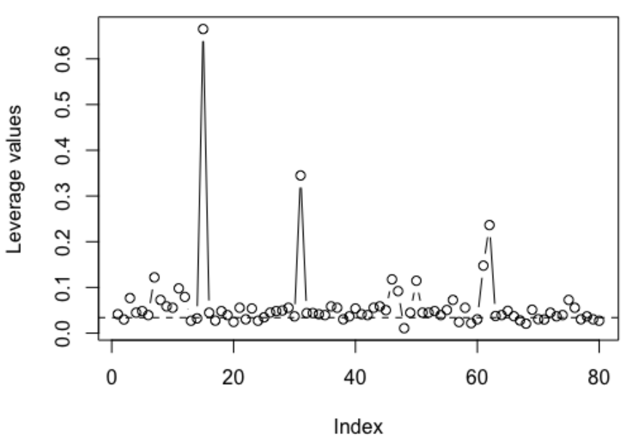
The goodness-of-fit tests are designed to test whether the logistic model fits the data adequately. The Hosmer-Lemeshow test is a goodness of fit test with a relatively high p-value at .2255 so the linear logistic model seems to fit these data relatively well. Lack of fit does not seem to be an issue. The VIF values, even though only approximate in this type of model, they can still be used as an approximate value. They are relatively low therefore multicollinearity does not seem to be an issue here either. Somers' D, the difference between the concordant and discordant proportions, is a measure of how well the successes are separated from the failures. In this model we have Somer's D of .499 which signals Fair separation in our data.

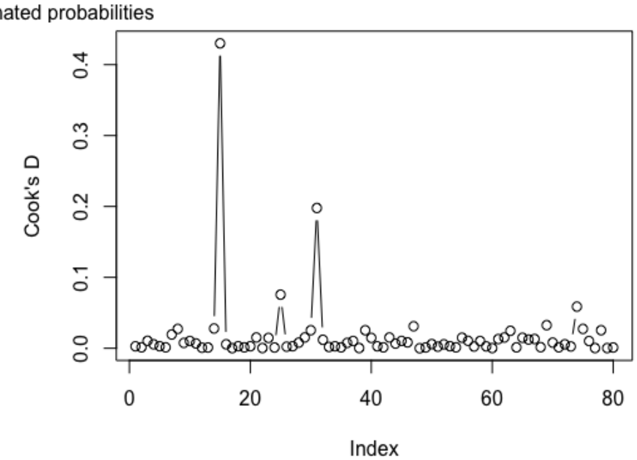
We should consider simplifying our model:
I used a best subsets analysis for generalized linear models (including logistic regression). I chose the measure of comparison here to be AIC which is technically not valid in the logistic regression context, is be a useful way of trading off fit versus complexity. The model with the smallest aic is preferred, which here seems to be the one predictor model, using SAT scores.

```
> logitbest$Subsets
   Intercept    X1    X2    X3    X4 logLikelihood      AIC
0       TRUE FALSE FALSE FALSE FALSE     -47.99184 95.98368
1*      TRUE FALSE  TRUE FALSE FALSE     -43.33785 88.67570
2       TRUE  TRUE  TRUE FALSE FALSE     -42.46735 88.93470
3       TRUE  TRUE  TRUE  TRUE FALSE     -41.97155 89.94311
4       TRUE  TRUE  TRUE  TRUE  TRUE     -41.91846 91.83691
```

Before rerunning the regression however, I decided to investigate the data for unusual observations that could be having a strong effect on our fitted logistic regression model. Diagnostics corresponding to standardized residuals, leverage values, Cook’s distances, the latter two of which are only approximate, are given below. I created a plot of residuals versus fitted probabilities:







```
> cbind(spearson1,R1diag$cook,R1diag$h)
     spearson1
1  -0.56618792 2.771518e-03 0.04143689
2  -0.45859160 1.299190e-03 0.02996257
3   0.79420580 1.048749e-02 0.07675265
4  -0.77191358 5.587540e-03 0.04478715
5  -0.53156068 2.863112e-03 0.04822128
6  -0.38973560 1.249042e-03 0.03949189
7  -0.83472333 1.937064e-02 0.12204037
8  -1.32011320 2.732317e-02 0.07269449
9  -0.77788642 7.520644e-03 0.05850724
10  0.93205866 1.021808e-02 0.05554367
11 -0.55947295 6.798611e-03 0.09796184
12 -0.20972696 7.568778e-04 0.07922139
13 -0.42812801 1.022717e-03 0.02714114
14  2.04339735 2.784608e-02 0.03226883
15 -1.04034845 4.300381e-01 0.66517565
16 -0.77191358 5.587540e-03 0.04478715
17 -0.14763969 1.229052e-04 0.02741950
18 -0.53156068 2.863112e-03 0.04822128
19 -0.38973560 1.249042e-03 0.03949189
20 -0.74598406 2.802989e-03 0.02456577
21 -1.13599095 1.517863e-02 0.05554367
22 -0.20653403 2.667576e-04 0.03032019
23 -1.13071150 1.458039e-02 0.05394510
24 -0.42812801 1.022717e-03 0.02714114
25  3.24850797 7.564846e-02 0.03460257
26 -0.48491415 2.209067e-03 0.04486561
27 -0.53156068 2.863112e-03 0.04822128
28 -0.87002754 7.883094e-03 0.04949433
29 -1.13599095 1.517863e-02 0.05554367
30  1.82529197 2.542055e-02 0.03674771
31  1.37067750 1.978488e-01 0.34492455
32  1.14252212 1.196667e-02 0.04382786
33 -0.44412443 1.812740e-03 0.04393238
34 -0.56618792 2.771518e-03 0.04143689
35 -0.38973560 1.249042e-03 0.03949189
36 -0.77788642 7.520644e-03 0.05850724
37  0.93205866 1.021808e-02 0.05554367
38 -0.20653403 2.667576e-04 0.03032019
39  1.82529197 2.542055e-02 0.03674771
40 -1.13071150 1.458039e-02 0.05394510
41 -0.56618792 2.771518e-03 0.04143689
42 -0.38973560 1.249042e-03 0.03949189
43 -1.13599095 1.517863e-02 0.05554367
44 -0.72790299 6.597645e-03 0.05861125
45  0.98542256 1.029632e-02 0.05034680
46 -0.55656972 8.259012e-03 0.11762799
47  1.23833317 3.104968e-02 0.09193274
48 -0.06539927 9.372302e-06 0.01083772
49 -0.33967717 1.080331e-03 0.04472225
50 -0.47584413 5.866265e-03 0.11468347
51 -0.48491415 2.209067e-03 0.04486561
52 -0.77191358 5.587540e-03 0.04478715
53 -0.53156068 2.863112e-03 0.04822128
54 -0.38973560 1.249042e-03 0.03949189
55  1.18088039 1.487784e-02 0.05064390
56  0.81689453 1.046262e-02 0.07269449
57 -0.69954670 2.441506e-03 0.02433848
58  0.93205866 1.021808e-02 0.05554367
59 -0.76818310 2.637474e-03 0.02185899
60 -0.19599315 2.397115e-04 0.03025753
61 -0.61174295 1.296860e-02 0.14768180
62 -0.49383473 1.509914e-02 0.23639058
63  1.77578635 2.456145e-02 0.03748438
64 -0.38973560 1.249042e-03 0.03949189
65  1.20143779 1.479669e-02 0.04875557
66  1.26544597 1.233454e-02 0.03708463
67  1.50911946 1.305964e-02 0.02787259
68 -0.59175495 1.490440e-03 0.02083797
69  1.73566444 3.256833e-02 0.05128269
70  1.12092406 7.930916e-03 0.03059471
71 -0.45859160 1.299190e-03 0.02996257
72 -0.77191358 5.587540e-03 0.04478715
73 -0.57545440 2.533489e-03 0.03684375
74  2.67133827 5.868062e-02 0.03949189
75 -1.32011320 2.732317e-02 0.07269449
76  0.93205866 1.021808e-02 0.05554367
77 -0.20653403 2.667576e-04 0.03032019
78  1.82529197 2.542055e-02 0.03674771
79 -0.19599315 2.397115e-04 0.03025753
80 -0.42812801 1.022717e-03 0.02714114
```

It becomes clear that Student 15 is an outlier, with both a signifactly higher leverage value and Cook's distance. Upon reexaming the data I saw that the specific student had very different statistics, achieving the maximum possible grade in their SAT at 800 and a surprisingly low ACT score at 21. The specific student was ranked very low relative to the rest of the data, at the 25th percentile. It is also important to note that admission certainly does not only depend on grades, therefore with many factors at play it is possible that other aspects of the student's profile influenced the final decision to reject him or her. Therefore I rerun the regression having removed the outlier.

```
glm(formula = Binary ~ GPA + SAT + ACT + RANK, family = binomial,
    data = df2, maxit = 500)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4537  -0.8011  -0.5270   1.1576   2.0964

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.375012   7.295442  -1.011   0.3121
GPA         -2.986765   1.930232  -1.547   0.1218
SAT          0.026683   0.012245   2.179   0.0293 *
ACT         -0.055937   0.218658  -0.256   0.7981
RANK        -0.001222   0.051889  -0.024   0.9812
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.301  on 78  degrees of freedom
Residual deviance: 82.495  on 74  degrees of freedom
AIC: 92.495

Number of Fisher Scoring iterations: 5
```

Linear Predictor

Y = -7.38 + -2.99 * GPA + 0.03 * SAT + -0.06 * ACT + 0 * RANK + e

The likelihood ratio test for whether all slopes equal 0

```
      gstat2
[1,] 12.80592 0.0122641
```

Odds ratios

```
       GPA        SAT        ACT       RANK 
0.05045037 1.02704237 0.94559902 0.99877839 
```

A 95% confidence interval for the odds ratio

```
                   2.5 %      97.5 %
(Intercept) 3.865290e-10 1016.165193
GPA         1.147752e-03    2.217587
SAT         1.002686e+00    1.051990
ACT         6.160057e-01    1.451541
RANK        9.021956e-01    1.105701
```

Likelihood ratio tests for each slope, along with AIC values for the model that omits that variable

```
Single term deletions

Model:
Binary ~ GPA + SAT + ACT + RANK
       Df Deviance    AIC    LRT Pr(>Chi)  
<none>      82.495 92.495                  
GPA     1   84.802 92.802 2.3072  0.12878  
SAT     1   88.516 96.516 6.0212  0.01414 *
ACT     1   82.560 90.560 0.0658  0.79762  
RANK    1   82.495 90.495 0.0006  0.98120  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

VIF Values

```
     GPA      SAT      ACT     RANK 
1.554194 1.639199 1.679801 1.253206 
```

```
> AIC(R2)
[1] 92.49472
```

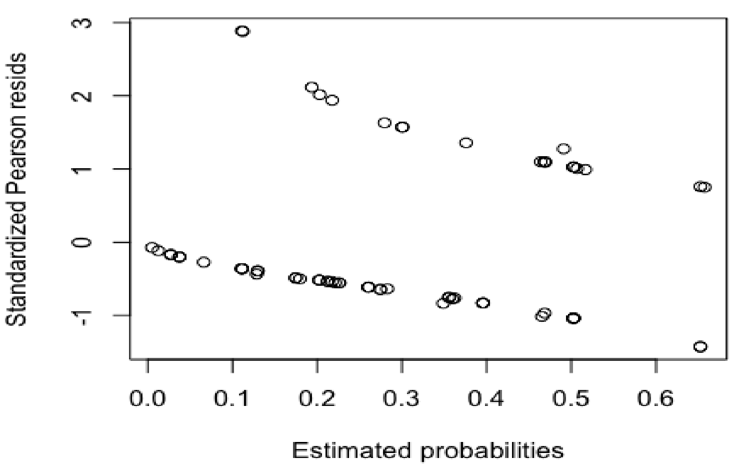
```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  df2$Binary, fitted(R2)
X-squared = 7.9746, df = 8, p-value = 0.436
```

Summary measures of association in the stats component, with Somers D being represented by Dxy

```
         Obs    Max Deriv   Model L.R.         d.f.            P            C          Dxy        Gamma        Tau-a           R2
7.900000e+01 8.233471e-08 1.280592e+01 4.000000e+00 1.226410e-02 7.507764e-01 5.015528e-01 5.094637e-01 2.096722e-01 2.135610e-01
       Brier            g           gr           gp
1.762095e-01 1.294938e+00 3.650771e+00 2.022292e-01
```

The test statistic for the likelihood ratio test rises to 12.80592, with a p-value falling to .0122641, leading us to strongly reject the null hypothesis of no relationship. It can once again be seen here that only SAT scores seem to be even more statistically significant with their p-value falling to .01414 while GPA, ACT scores and class rank are not statistically significant at a 95% significance level. The analysis shows that a percentage point higher GPA is associated with multiplying the odds of a student getting into an IV league school by .05045, a percentage point increase in SAT is associated with 2.7% higher odds of getting accepted, while a percentage point increase in ACT is associated with 5.441% lower odds of getting accepted. A percentage point increase in Rank would now only lower the odds of getting accepted by .13%. The Hosmer-Lemeshow test is a goodness of fit test with a shows a much higher p-value at .436 so the linear logistic model seems to fit these data quite well. Lack of fit does not seem to be an issue. The VIF values are still relatively low therefore multicollinearity does not seem to be an issue here either. Somers' D, the difference increases to .5016 which still signals Fair separation in our data. The outlier seems to have been removed in the Standardized Pearson residuals too.



```
> cbind(spearson2,Rdiag2$cook,Rdiag2$h)
     spearson2
1  -0.51615037 2.655984e-03 0.04748067
2  -0.55576530 3.566564e-03 0.05458337
3   0.75130887 9.395843e-03 0.07683320
4  -0.83030886 7.264166e-03 0.05004705
5  -0.61363176 5.268719e-03 0.06538697
6  -0.36054290 1.048163e-03 0.03875424
7  -1.01364013 3.700154e-02 0.15258695
8  -1.42684245 3.459160e-02 0.07830273
9  -0.64341925 7.804534e-03 0.08614067
10  1.02831491 1.450912e-02 0.06420093
11 -0.48558796 5.370203e-03 0.10223246
12 -0.11512261 1.311967e-04 0.04716184
13 -0.39190002 9.523559e-04 0.03007172
14  2.01455150 2.679922e-02 0.03196156
15 -0.83030886 7.264166e-03 0.05004705
16 -0.16973699 2.189988e-04 0.03661496
17 -0.61363176 5.268719e-03 0.06538697
18 -0.36054290 1.048163e-03 0.03875424
19 -0.75284358 2.848611e-03 0.02451403
20 -1.03918115 1.481737e-02 0.06420093
21 -0.19994463 2.469006e-04 0.02995468
22 -1.03767011 1.423742e-02 0.06201258
23 -0.39190002 9.523559e-04 0.03007172
24  2.88003171 8.203726e-02 0.04712202
25 -0.53343934 3.160697e-03 0.05261496
26 -0.61363176 5.268719e-03 0.06538697
27 -0.96817488 1.190393e-02 0.05970581
28 -1.03918115 1.481737e-02 0.06420093
29  1.57342367 3.138523e-02 0.05960908
30  1.27668257 1.864641e-01 0.36386946
31  1.10029533 1.155347e-02 0.04554287
32 -0.55511041 5.273821e-03 0.07882748
33 -0.51615037 2.655984e-03 0.04748067
34 -0.36054290 1.048163e-03 0.03875424
35 -0.64341925 7.804534e-03 0.08614067
36  1.02831491 1.450912e-02 0.06420093
37 -0.19994463 2.469006e-04 0.02995468
38  1.57342367 3.138523e-02 0.05960908
39 -1.03767011 1.423742e-02 0.06201258
40 -0.51615037 2.655984e-03 0.04748067
41 -0.36054290 1.048163e-03 0.03875424
42 -1.03918115 1.481737e-02 0.06420093
43 -0.77451957 8.128021e-03 0.06344855
44  0.99290161 1.056840e-02 0.05087340
45 -0.48950424 6.454444e-03 0.11869735
46  1.35825279 4.072190e-02 0.09939636
47 -0.07142051 1.384601e-05 0.01339042
48 -0.27262233 7.597211e-04 0.04862430
49 -0.50050201 7.225782e-03 0.12604669
50 -0.53343934 3.160697e-03 0.05261496
51 -0.83030886 7.264166e-03 0.05004705
52 -0.61363176 5.268719e-03 0.06538697
53 -0.36054290 1.048163e-03 0.03875424
54  1.09542991 1.476731e-02 0.05796542
55  0.76038874 9.824037e-03 0.07830273
56 -0.75266497 3.320877e-03 0.02847562
57  1.02831491 1.450912e-02 0.06420093
58 -0.76231729 2.679130e-03 0.02253175
59 -0.16970950 1.641593e-04 0.02770891
60 -0.83474228 4.189880e-02 0.23115594
61 -0.43439560 1.051021e-02 0.21782767
62  1.93800468 3.363798e-02 0.04286129
63 -0.36054290 1.048163e-03 0.03875424
64  1.09703862 1.485790e-02 0.05813937
65  1.09881533 1.495409e-02 0.05831576
66  1.63080241 1.705671e-02 0.03107095
67 -0.63601214 2.099655e-03 0.02529646
68  2.11635357 6.735045e-02 0.06992793
69  1.00837750 9.095838e-03 0.04281183
70 -0.55576530 3.566564e-03 0.05458337
71 -0.83030886 7.264166e-03 0.05004705
72 -0.53802377 2.498457e-03 0.04137041
73  2.88541716 6.713237e-02 0.03875424
74 -1.42684245 3.459160e-02 0.07830273
75  1.02831491 1.450912e-02 0.06420093
76 -0.19994463 2.469006e-04 0.02995468
77  1.57342367 3.138523e-02 0.05960908
78 -0.16970950 1.641593e-04 0.02770891
79 -0.39190002 9.523559e-04 0.03007172
```

Since not all of the predictors are significant, I rerun best subsets:

```
> logitbest2$Subsets
   Intercept    X1    X2    X3    X4 logLikelihood      AIC
0       TRUE FALSE FALSE FALSE FALSE     -47.65032 95.30064
1       TRUE FALSE  TRUE FALSE FALSE     -42.59995 87.19990
2*      TRUE  TRUE  TRUE FALSE FALSE     -41.28090 86.56180
3       TRUE  TRUE  TRUE  TRUE FALSE     -41.24764 88.49527
4       TRUE  TRUE  TRUE  TRUE  TRUE     -41.24736 90.49472
```

The model that minimizes AIC is a 2-predictor model with GPA and SAT

```
Call:
glm(formula = Binary ~ GPA + SAT, family = binomial, data = df2,
    maxit = 500)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4312  -0.8251  -0.5372   1.1506   2.1202

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.346358   6.488813  -1.132  0.25757
GPA         -3.034853   1.828600  -1.660  0.09698 .
SAT          0.024436   0.008143   3.001  0.00269 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.301  on 78  degrees of freedom
Residual deviance: 82.562  on 76  degrees of freedom
AIC: 88.562

Number of Fisher Scoring iterations: 5
```

Linear Predictor

$\boxed{Y = -7.35 + -3.03 * GPA + 0.02 * SAT + e}$

The likelihood ratio test for all slopes equal 0.

```
       gstat3
[1,] 12.73884 0.002303362
```

Odds Ratio

```
       GPA        SAT
0.04808173 1.02473741
```

Confidence Interval for Odds Ratio

```
                   2.5 %      97.5 %
(Intercept) 1.932988e-09 215.181590
GPA         1.334975e-03   1.731757
SAT         1.008513e+00   1.041223
```

Likelihood ratio tests for each slope, along with AIC values for the model that omits that variable

```
Model:
Binary ~ GPA + SAT
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      82.562 88.562
GPA     1   85.200 89.200  2.6381 0.1043281
SAT     1   95.269 99.269 12.7068 0.0003643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

VIF Values

```
     GPA      SAT
1.388964 1.388964
```

```
> AIC(R3)
[1] 88.5618
```

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  df2$Binary, fitted(R3)
X-squared = 12.87, df = 8, p-value = 0.1164
```

```
         Obs    Max Deriv   Model L.R.         d.f.            P            C          Dxy        Gamma        Tau-a           R2
7.900000e+01 3.753485e-08 1.273884e+01 2.000000e+00 1.713156e-03 7.364130e-01 4.728261e-01 4.891566e-01 1.976631e-01 2.125301e-01
       Brier            g           gr           gp
1.763303e-01 1.284579e+00 3.613148e+00 2.021066e-01
```

The test statistic for the likelihood ratio test fell marginally to 12.738, with a p-value falling drastically to .00230, leading us to even more strongly reject the null hypothesis of no relationship. SAT scores still however seem to be the only

statistically significant with their p-value falling further to .0003643 while GPA p-value fell to .1043281, still significantly higher than the 95% level.
The analysis shows that a percentage point higher GPA is associated with multiplying the odds of a student getting into an IV league school by .048, a percentage point increase in SAT is associated with 2.4737% higher odds of getting accepted. The Hosmer-Lemeshow test shows a much lower however p-value at .1164, still however high enough for us say that the linear logistic model seems to fit these data quite well and lack of fit does not seem to be an issue. The VIF values are still low therefore multicollinearity does not seem to be an issue here either. Somers' D, falls marginally to .4728 which still signals Fair separation in our data. AIC does fall significantly to 88.56

-- [side note]

Since GPA is not statistically significant and the single term deletion table shows the AIC to fall when the predictor is excluded, I decided to rerun the regression using only SAT as the predictor.

```
glm(formula = Binary ~ SAT, family = binomial, data = df2, maxit = 500)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2333  -0.8544  -0.5608   1.1225   2.0456

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.076672   5.178048  -2.912  0.00360 **
SAT           0.019009   0.006853   2.774  0.00554 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.301  on 78  degrees of freedom
Residual deviance: 85.200  on 77  degrees of freedom
AIC: 89.2

Number of Fisher Scoring iterations: 4
```

Linear Predictor

$Y = -15.08 + 0.02 * SAT + e$

The likelihood ratio test for all slopes equal 0.

```
        gstat4
[1,] 10.78378 0.001023934
```

Confidence Interval for Odds Ratio

```
                   2.5 %      97.5 %
(Intercept) 1.108466e-11 0.007241806
SAT         1.005592e+00 1.032973071
```

Odds Ratio

```
     SAT
1.019191
```

```
> AIC(R4)
[1] 89.1999
```

Likelihood ratio tests for each slope, along with AIC values for the model that omits that variable

```
Single term deletions

Model:
Binary ~ SAT
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      85.200 89.200
SAT     1   95.301 97.301 10.101 0.001482 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  df2$Binary, fitted(R4)
X-squared = 8.4989, df = 8, p-value = 0.3863
```

```
        Obs   Max Deriv  Model L.R.        d.f.           P           C         Dxy       Gamma       Tau-a
79.000000000 0.001135550 10.100734086 1.000000000 0.001482103 0.717779503 0.435559006 0.474218090 0.182083739
          R2       Brier           g          gr          gp
 0.171285411 0.182521569 1.076247554 2.933650506 0.180615772
```

The test statistic for the likelihood ratio test fell marginally to 10.7837, with a p-value falling drastically to .00102, leading us to even more strongly reject the null hypothesis of no relationship. SAT scores are very statistically significant with their p-value of .0010239. The analysis shows that a percentage point higher GPA is associated with multiplying the odds of a student getting into an IV league school by 1.019191. The Hosmer-Lemeshow test shows a much higher p-value at .3863 therefore the linear logistic model seems to fit these data quite well and lack of fit does not seem to be an issue. Somers' D, actually falls marginally to .435559 which still signals Fair separation in our data. AIC does increase marginally to 89.1999. It seems that the previous regression is an overall better fit.

I created a classification matrix using the 3rd regression, based on whether the estimated probability is above or below .5:

```
  R3.predict
     0  1
0 47  9
1  9 14
```

Rows: Actual			Columns Predicted
	0	1	all
0	47	9	56
1	9	14	23
All	56	23	79

About 77.2% of the students were correctly classified, which is significantly higher than the Cmax of 70.89%. In order to take into account that we are fitting the model onto the same data, I also compared to the Cpro of 73.4% [Cpro = 1.25 ×[56/79 * 56/79 + 23/79 * 23/79]]. Thus, it is reasonable to assume that the two admission statistics do a relatively good job of classifying students into admitted students and rejected groups. Taking into account the outlier, which was removed, the model classified correctly 76.25% of students, which is still higher than the Cmax of 71.25%. Unfortunately, there was no additional data available, so I wasn't able to test the predictive ability of the model on new data. I was planning on testing it using data from another year, but it seems that the website only offers last year's information, while 2020 information has yet to be updated.

Discussion and Conclusion

A logistic regression model was successfully generated, using students' GPAs and SAT scores as the predictors. While SAT scores seem to be strongly statistically significant with a p-value of .0003643, GPA is not statistically significant at a 95% level with a p-value of .1043281. Prediction seems to be possible, with 77.2% of the students correctly classified (76.25% taking into account the removed outlier). The data however doesn't show perfect separation with a Somer's D of .4728.

It is important to note that there are important flaws in the data that were beyond the control of this study. Out of the 400 datapoints initially collected, 320 of them had missing values and therefore were discarded. This deletion of rows with missing values, reduced the sample size and affected the informativeness of the regression. The final dataset consisted of 80 data points.

In order to create a possibly more accurate model, we would need to incorporate more subjective aspects of an applicant's portfolio such as application essay, extracurricular activities or leadership roles. The techniques necessary for such processing are beyond the scope of this class.

[NOTE: I was actually planning on using admission statistics for Imperial College London, specifically their master's in Financial Technology program which they just started offering last year. However, there is no college data website equivalent outside of the US and since the program is very new, very little information was publicly available. I requested the necessary information from their school, and they are able to provide the statistics, but they will do so by the beginning of June.]