Stats 101A Final Project: Predicting College Acceptance Rates

Casey Tattersall, Stefan Abgarian, Kailyn Nguyen, Hanchen Huang, Gina Pak, Ahyoung Ju

3/11/2023

Contents

1 Introduction

1.1 Research Question

Can we reliably predict a college's acceptance rate given general information about the college and the test scores and high school GPAs of its students?

1.2 Gathering Data

We began our analysis with a Kaggle dataset that contained a variety of statistics, including acceptance rate, from well over 200 colleges. Many of those variables were not relevant to our research question, so we narrowed them down to the following four: Average SAT score, Average High School GPA, Students Enrolled, and whether the school was public (stored as 0) or private (stored as 1).

Unfortunately, we quickly encountered several issues. The dataset was generated by webscraping the US-News website, but this was often done incorrectly, resulting in a significant number of missing and incorrect values. While these errors were widespread, one specific issue we repeatedly encountered was that colleges from the same system (such as UC or Cal State) were often incorrectly given data from other colleges within the same system. To fix this, when we encountered missing and/or clearly incorrect information, we looked up that data point on collegesimply.com and manually replaced the value. Unfortunately, we were not able to find reliable data for some of the smaller and more obscure colleges in our original dataset, so we had to remove them.

After cleaning our data, we were left with a dataset that contained the acceptance rate and our chosen predictor variables for 216 colleges.

1.3 Modeling Methods

In order to answer our research question, we chose to create a multiple linear regression model. In the following sections, we will:

- Create an untransformed linear model to determine if there is a relationship between our variables.
- Explore transforming our variables through a Box-Cox transformation.
- Consider removing predictor variables from our model using forward selection and backward elimination.
- Choose a final model and discuss its effectiveness and limitations.

2 Data Description

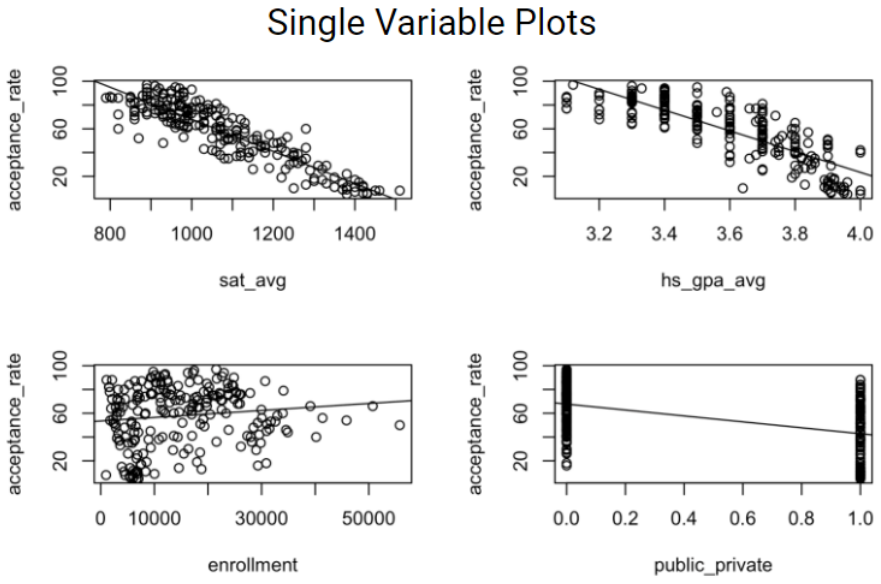
2.1 Variables

- acceptance_rate: The percentage of the university's applicants who are accepted.
  - Our dataset contains colleges that range from 5% to 97%, with a mean of 58%.
- sat_avg: The average SAT score of the university's students.
  - Our dataset contains colleges that range from 790 to 1510, with a mean of 1074.
- hs_gpa_avg: The average unweighted high school GPA (4.0 scale) of the university's students.
  - Our dataset contains colleges that range from 3.1 to 4.0, with a mean of 3.6.
- enrollment: The number of undergraduate students enrolled at the university.
  - Our dataset contains colleges that range from 979 to 55,766, with a mean of 14,994.
- public_private: Dummy variable that is 0 if the college is public, and 1 if the college is private.
  - Our dataset contains 85 private schools and 131 public schools.

2.2 Single Variable Relationships

In the following single variable plots and correlation matrix, we can see that there is a relatively strong negative correlation between acceptance rate and both sat_avg and hs_gpa_avg. There appears to be a relationship between acceptance rate and public/private, but we can't tell for sure at this point. Enrollment does not appear to affect acceptance rate much at all.

	acceptance_rate	sat_avg	hs_gpa_avg	enrollment	public_private
acceptance_rate	1.0000000	-0.89475093	-0.8125129	0.12438424	-0.4957500
sat_avg	-0.8947509	1.00000000	0.8434822	-0.06562565	0.4363641
hs_gpa_avg	-0.8125129	0.84348217	1.0000000	0.12047118	0.3092901
enrollment	0.1243842	-0.06562565	0.1204712	1.00000000	-0.6327082
public_private	-0.4957500	0.43636409	0.3092901	-0.63270822	1.0000000

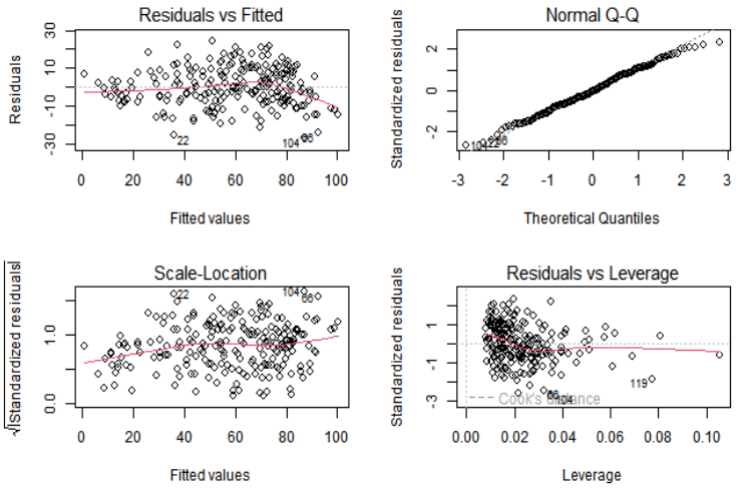


3 Modeling

3.1 Untransformed Model

When first fitting a multiple linear regression model to our data, we get the following output:

```
Call:
lm(formula = acceptance_rate ~ sat_avg + hs_gpa_avg + enrollment +
    public_private)

Residuals:
    Min      1Q  Median      3Q     Max
-27.067  -7.079  -0.342   7.830  23.988

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.570e+02  1.468e+01  17.506  < 2e-16 ***
sat_avg        -9.523e-02  8.537e-03 -11.155  < 2e-16 ***
hs_gpa_avg     -2.653e+01  6.070e+00  -4.371 1.94e-05 ***
enrollment      8.583e-05  9.698e-05   0.885  0.37716
public_private -6.093e+00  2.168e+00  -2.810  0.00542 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.26 on 211 degrees of freedom
Multiple R-squared:   0.83,     Adjusted R-squared:  0.8268
F-statistic: 257.5 on 4 and 211 DF,  p-value: < 2.2e-16
```



The R-squared value is 0.83 and the p-value is about 0, which shows that there is a clear relationship between the predictor variables and the response variable. The p-value for the enrollment variable is higher, which continues to suggest that the number of students enrolled may not have a statistically significant effect on the acceptance rate variable.

While not perfect, our diagnostic plots look reasonable. The data appears to be normally distributed with mostly constant variance, but there is a little bit of non-linearity and non-constant variance for colleges with very high and very low acceptance rates, and there are some leverage points. We will address these issues in more detail later.

3.2 Transformed Model

We now ran a box-cox transformation on our numerical variables (not the public/private dummy variable) and got the following results:

```
bcPower Transformations to Multinormality
               Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
acceptance_rate    1.0017        1.00       0.8275       1.1758
sat_avg            0.2323        0.00      -0.3938       0.8585
hs_gpa_avg         5.1274        5.13       3.5475       6.7073
enrollment         0.3559        0.50       0.1988       0.5131

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                                LRT df       pval
LR test, lambda = (0 0 0 0) 249.9157  4 < 2.22e-16

Likelihood ratio test that no transformations are needed
                               LRT df       pval
LR test, lambda = (1 1 1 1) 96.6896  4 < 2.22e-16
```

Since transforming variables will make our interpretations more difficult, we decided to only transform a variable if it made the model clearly and significantly better in some way.

The only variable that we chose to transform was hs_gpa_avg, which we raised to the power of 5. While the transformation recommends using a log transformation on the sat_avg variable, this actually made our model worse, as it lowered our R^2 and worsened the outliers and leverage points, so we decided against transforming it. Additionally, transforming the enrollment variable didn't make much of a difference, and it will be removed in the next step either way.

3.3 Variable Reduction

In order to find the optimal model given our predictor variables, we conducted both forward selection and backward elimination with AIC and BIC. In all four tests, we finished with the same model: keeping all predictor variables except enrollment. The backward elimination process can be seen below, and the

forward selection output is in the appendix.

```
Start:  AIC=1002.59
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private +
    enrollment

                 Df Sum of Sq   RSS    AIC
- enrollment      1      99.8 21488 1001.6
<none>                       21388 1002.6
- public_private  1     845.3 22234 1009.0
- I(hs_gpa_avg^5) 1    2824.8 24213 1027.4
- sat_avg         1    9471.8 30860 1079.8

Step:  AIC=1001.6
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private

                 Df Sum of Sq   RSS    AIC
<none>                       21488 1001.6
- public_private  1    2338.0 23826 1021.9
- I(hs_gpa_avg^5) 1    2765.0 24253 1025.7
- sat_avg         1    9739.2 31228 1080.3

Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_privc

Coefficients:
    (Intercept)         sat_avg  I(hs_gpa_avg^5)   public_private
      178.02956        -0.08701         -0.03739         -7.54023
```

```
Start:  AIC=1019.47
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private +
    enrollment

                 Df Sum of Sq   RSS    AIC
- enrollment      1      99.8 21488 1015.1
<none>                       21388 1019.5
- public_private  1     845.3 22234 1022.5
- I(hs_gpa_avg^5) 1    2824.8 24213 1040.9
- sat_avg         1    9471.8 30860 1093.3

Step:  AIC=1015.1
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private

                 Df Sum of Sq   RSS    AIC
<none>                       21488 1015.1
- public_private  1    2338.0 23826 1032.0
- I(hs_gpa_avg^5) 1    2765.0 24253 1035.9
- sat_avg         1    9739.2 31228 1090.5

Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)

Coefficients:
    (Intercept)         sat_avg  I(hs_gpa_avg^5)   public_private
      178.02956        -0.08701         -0.03739         -7.54023
```

Additionally, our VIF for each variable is less than 5, so we don't have to worry about multicollinearity:

```
      sat_avg I(hs_gpa_avg^5)  public_private      enrollment
     4.498801        4.432735        2.288202        2.018236
```

4 Results and Discussion

4.1 Final Model

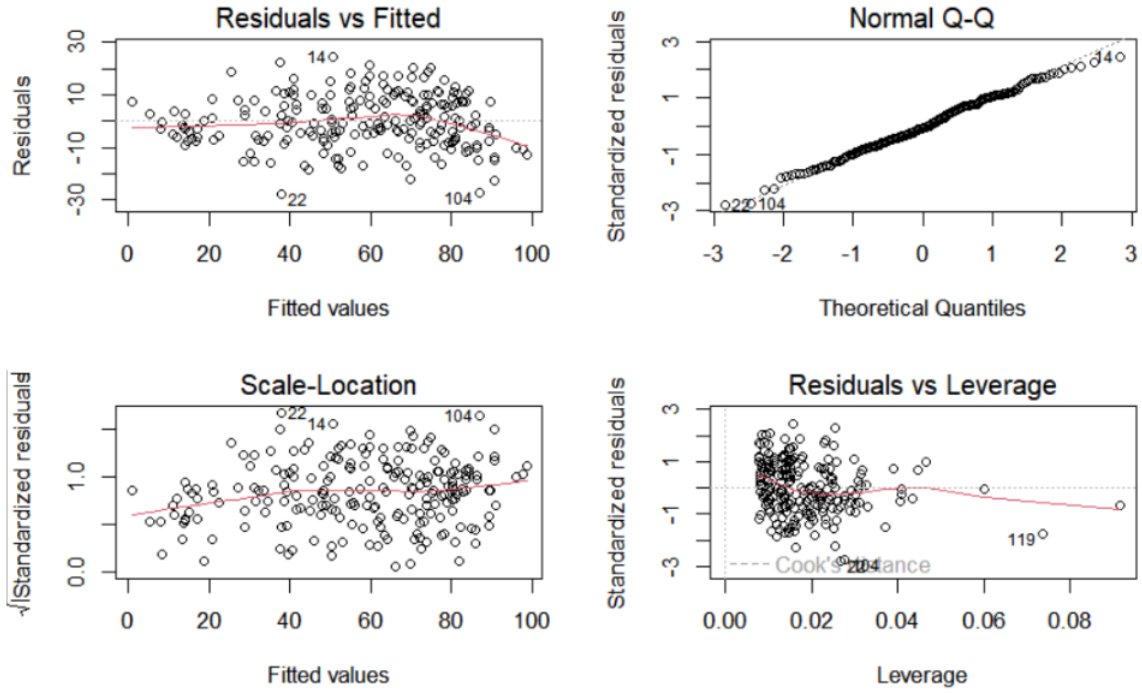
After transforming hs_gpa_avg and removing enrollment, we are left with the following model:

```
Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)

Residuals:
     Min       1Q   Median       3Q      Max
-27.8329  -6.9711  -0.6892   7.4151  24.4105

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     178.029556   5.972222  29.810  < 2e-16 ***
sat_avg          -0.087011   0.008877  -9.802  < 2e-16 ***
I(hs_gpa_avg^5)  -0.037391   0.007159  -5.223 4.20e-07 ***
public_private   -7.540226   1.569971  -4.803 2.96e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.07 on 212 degrees of freedom
Multiple R-squared:  0.8355,    Adjusted R-squared:  0.8331
F-statistic: 358.8 on 3 and 212 DF,  p-value: < 2.2e-16
```

This gives us the following relationship:

$$\textbf{acceptance\_rate} = \mathbf{178.029556} - \mathbf{0.087011}(\textbf{sat\_avg}) - \mathbf{0.037391}(\textbf{hs\_gpa\_avg}^{\mathbf{5}}) - \mathbf{7.540226}(\textbf{private})$$

And these are our final diagnostic plots:



Our R^2 and p-values are all excellent, and the diagnostic plots suggest that our model is mostly valid, with a few minor issues that we will discuss below.

4.2 Coefficient Interpretations

- Each one unit increase in SAT score is associated with a 0.087% decrease in acceptance rate. We were expecting a small negative coefficient, as we can expect colleges with a higher average SAT score to have a lower acceptance rate, and since the SAT is scored out of 1600 points, a one point increase shouldn't affect the magnitude of acceptance rate by very much.
- The Average GPA predictor variable is a bit more difficult to interpret, since we transformed it. However, the negative coefficient suggests that an increase in average GPA leads to a decrease in acceptance rate, which is what we would expect. Interestingly, by raising this variable to the power of 5, its values (originally between 3.1 and 4.0) now range from 286 to 1024, which is similar to the average SAT score variable.
- With all else held equal, a private school is expected to have a 7.54% lower acceptance rate than a public school. This number may be inflated by the fact that our dataset had a lot of public state schools with very high acceptance rates; for more prestigious schools, it is possible that the difference may be a bit smaller.

4.3 Limitations

While our final model is quite effective at predicting acceptance rate, especially given our issues with the dataset, there are a few issues and inconsistencies with it.

4.3.1 Nonlinearity at High Acceptance Rates

While the majority of our dataset appears to have a linear relationship, this isn't quite the case for colleges that accept nearly 100% of their students. This issue can be seen on the far right of the Residuals vs Fitted diagnostic plot. A likely explanation for this is that it's not possible for colleges to exceed a 100% acceptance rate, so as the average GPA and SAT scores continue to decrease, the acceptance rate can only asymptotically approach 100%. While our model appears to be fine for the vast majority of colleges,

we should take its projections for colleges with higher acceptance rates with a grain of salt. This could be fixed by adding a nonlinear element to the model in some way.

4.3.2 Outliers and Leverage Points

Our diagnostic plots show a surprisingly small number of outliers, likely because we had to remove many of the more obscure colleges due to a lack of reliable data, but there are still a few, and point 119 is right on the borderline of being a bad leverage point. Interestingly, a bunch of our outliers (including point 119) are small rural private schools that have relatively high test scores and GPAs, but higher than expected acceptance rates, suggesting that less qualified candidates simply aren't applying there. If we could find an additional predictor variable that somehow accounted for colleges like these, then that could fix the issue.

4.3.3 Nonconstant Variance at Low Acceptance Rates

Colleges that have a moderate acceptance rate (30-80%) tend to have a significant variance in their statistics, but students attending extremely prestigious colleges all have very strong profiles. Near-perfect GPAs and test scores are a necessity to be accepted to a college with an acceptance rate under 15%, so there is much less variance for these schools, as can be seen in the Residuals vs Fitted chart. Exploring the use of a weighted linear model could help account for this.

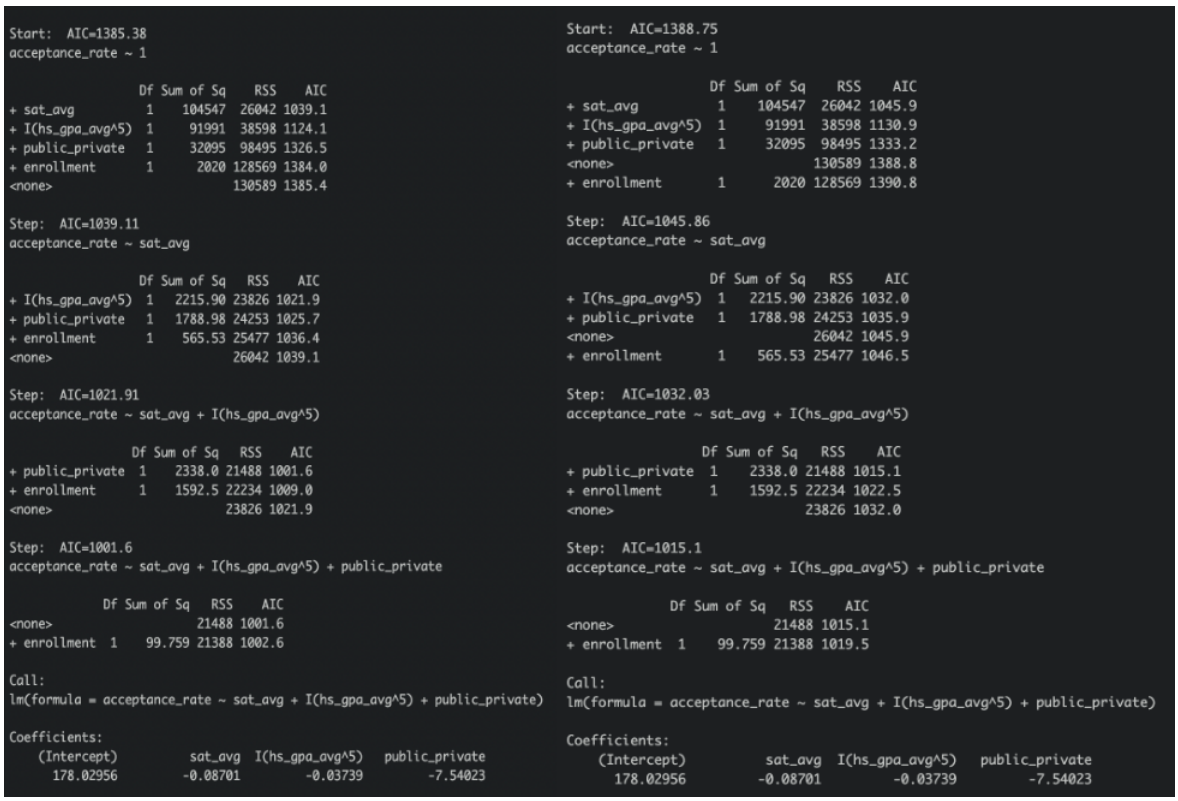
4.4 Conclusion

Despite these limitations, we are happy with our results, and believe our model could be a useful tool for students to estimate their chances of being accepted into different tiers of colleges. There are already many similar calculators that one can find online, and we believe that a model similar to ours is being used behind the scenes. While it would be interesting to see how much the model could be improved by using a larger and more consistent dataset and accounting for the nonlinear relationships and nonconstant variance at the two extreme ends of acceptance rate, we believe the model is still quite effective in its current state.

5 Appendix

5.1 Additional Figures

5.1.1 Variable Selection: Forward Selection



```
Start:  AIC=1385.38
acceptance_rate ~ 1

                  Df Sum of Sq    RSS    AIC
+ sat_avg          1    104547  26042 1039.1
+ I(hs_gpa_avg^5)  1     91991  38598 1124.1
+ public_private   1     32095  98495 1326.5
+ enrollment       1      2020 128569 1384.0
<none>                        130589 1385.4

Step:  AIC=1039.11
acceptance_rate ~ sat_avg

                  Df Sum of Sq   RSS    AIC
+ I(hs_gpa_avg^5)  1   2215.90 23826 1021.9
+ public_private   1   1788.98 24253 1025.7
+ enrollment       1    565.53 25477 1036.4
<none>                         26042 1039.1

Step:  AIC=1021.91
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5)

                 Df Sum of Sq   RSS    AIC
+ public_private  1    2338.0 21488 1001.6
+ enrollment      1    1592.5 22234 1009.0
<none>                        23826 1021.9

Step:  AIC=1001.6
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private

             Df Sum of Sq   RSS    AIC
<none>                    21488 1001.6
+ enrollment  1    99.759 21388 1002.6

Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)

Coefficients:
    (Intercept)         sat_avg  I(hs_gpa_avg^5)   public_private
      178.02956        -0.08701         -0.03739         -7.54023
```

```
Start:  AIC=1388.75
acceptance_rate ~ 1

                  Df Sum of Sq    RSS    AIC
+ sat_avg          1    104547  26042 1045.9
+ I(hs_gpa_avg^5)  1     91991  38598 1130.9
+ public_private   1     32095  98495 1333.2
<none>                        130589 1388.8
+ enrollment       1      2020 128569 1390.8

Step:  AIC=1045.86
acceptance_rate ~ sat_avg

                  Df Sum of Sq   RSS    AIC
+ I(hs_gpa_avg^5)  1   2215.90 23826 1032.0
+ public_private   1   1788.98 24253 1035.9
<none>                         26042 1045.9
+ enrollment       1    565.53 25477 1046.5

Step:  AIC=1032.03
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5)

                 Df Sum of Sq   RSS    AIC
+ public_private  1    2338.0 21488 1015.1
+ enrollment      1    1592.5 22234 1022.5
<none>                        23826 1032.0

Step:  AIC=1015.1
acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private

             Df Sum of Sq   RSS    AIC
<none>                    21488 1015.1
+ enrollment  1    99.759 21388 1019.5

Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)

Coefficients:
    (Intercept)         sat_avg  I(hs_gpa_avg^5)   public_private
      178.02956        -0.08701         -0.03739         -7.54023
```

5.2 Citations

- Kaggle Dataset:
  https://www.kaggle.com/datasets/theriley106/university-statistics
- CollegeSimply:
  https://www.collegesimply.com/
- Project Github (R files):
  https://github.com/Kc1227/Stats-101A-Group-Project