Stats 101A Final Project: Predicting College Acceptance Rate

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

Research Question

- For our research model, we wanted to create a model that could accurately predict college acceptance rates based off of several different variables.
- Our dataset included variables such as SAT average, ACT average, high school gpa, the size of the college, and whether the university is private (denoted as 1) or public (denoted as 0).
- A model like this could be potentially useful as it could give students a potential structure of what kind of statistics they should aim to achieve if they want to get into a specific university.

Data

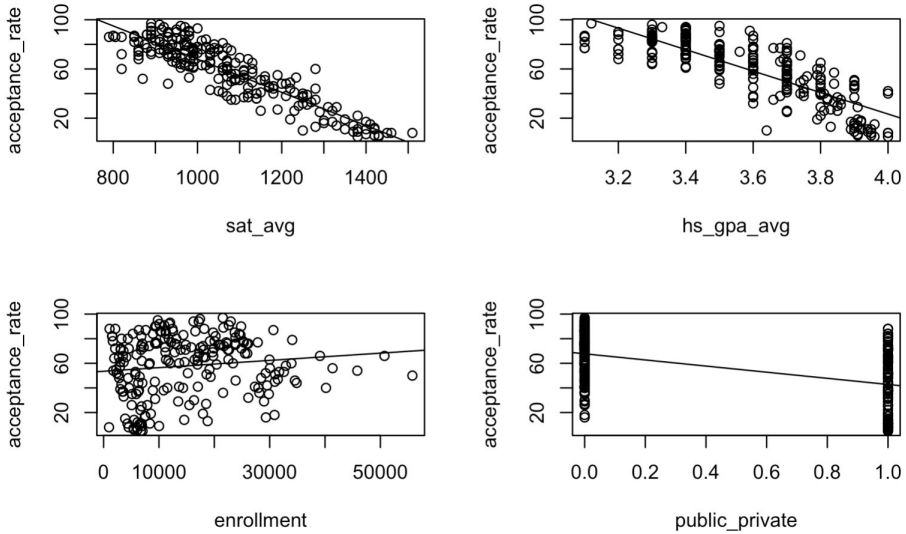
- Our original data was downloaded from Kaggle.
- The specific dataset that we used originally got its data by web scraping university data from US news, but this was not done properly and, there were several missing and incorrect values.
- For example, schools in the same system would often have an issue where a school's acceptance rate or average SAT score would be that of a different school in the same system.
- We replaced all missing values and as many incorrect values as we could find by manually looking up the data on collegesimply.com.

Single Variable Relationships

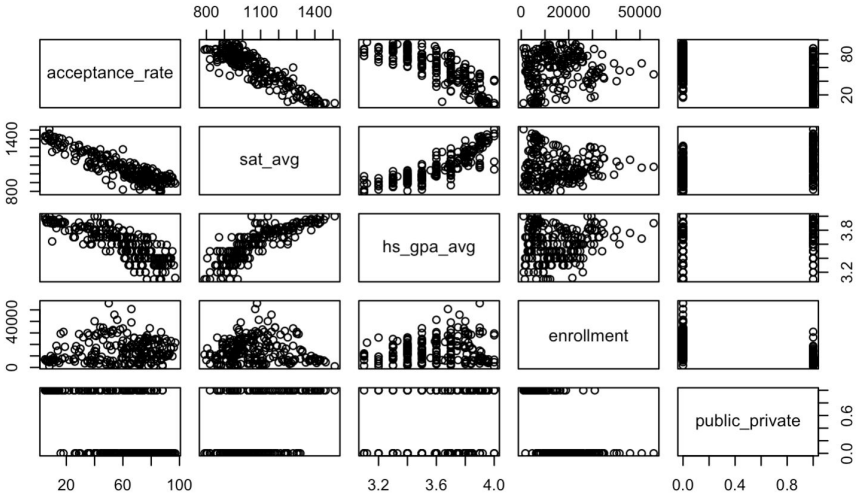
	acceptance_rate	sat_avg	hs_gpa_avg	enrollment	public_private
acceptance_rate	1.0000000	-0.89475093	-0.8125129	0.12438424	-0.4957500
sat_avg	-0.8947509	1.00000000	0.8434822	-0.06562565	0.4363641
hs_gpa_avg	-0.8125129	0.84348217	1.0000000	0.12047118	0.3092901
enrollment	0.1243842	-0.06562565	0.1204712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

Correlation Matrix

Single Variable Plots

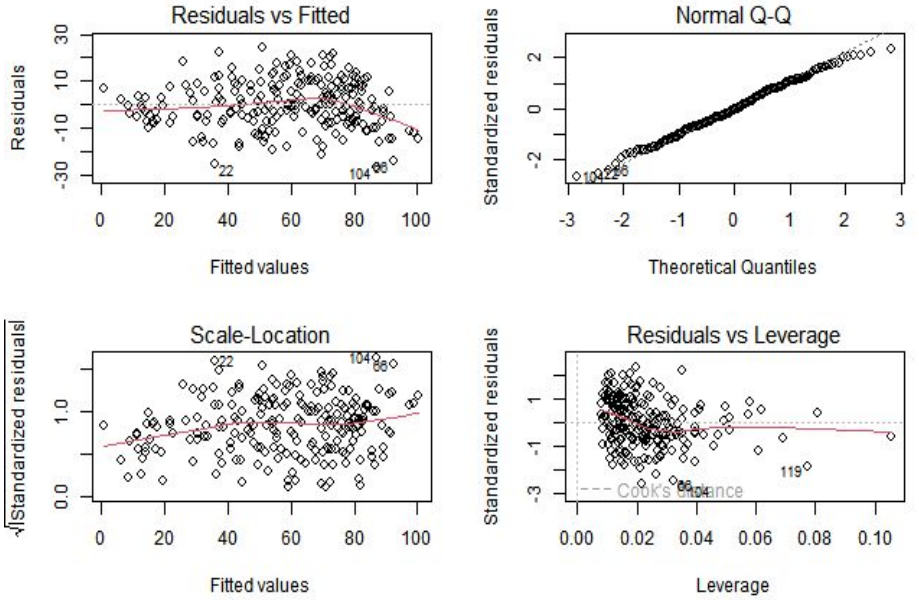


Pairwise plot



Untransformed Full Model

```
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
```



$R^2 = 0.83$

p-value = < 2.2e-16 ≈ 0

Transformed Full Model

- Transformed with Box-Cox Transformation (not the public/dummy variable)

```
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)

Likelihood ratio test that no transformations are needed
```

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (0 0 0 0)	249.9157	4	< 2.22e-16

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (1 1 1 1)	96.6896	4	< 2.22e-16

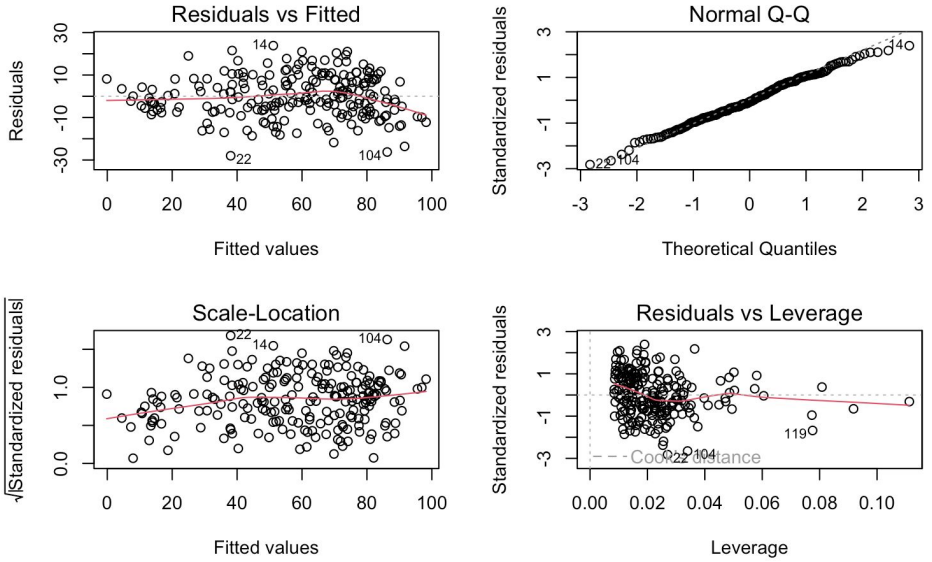
Transforming variables will make our interpretations more difficult.

Thus, we chose to transform only one variable **hs_gpa_avg**, not each variable.

Transformed Full Model

- Transformed model with the transformation of **hs_gpa_avg** to **(hs_gpa_avg)^5**

```
Likelihood ratio test that no transformations are needed

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

Residuals:
     Min       1Q   Median       3Q      Max
-28.0167  -6.7305  -0.6803   7.6850  23.8480

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.766e+02  6.143e+00  28.750  < 2e-16 ***
sat_avg         -8.618e-02  8.916e-03  -9.666  < 2e-16 ***
I(hs_gpa_avg^5) -3.964e-02  7.508e-03  -5.279 3.22e-07 ***
public_private  -6.125e+00  2.121e+00  -2.888  0.00429 **
enrollment       9.329e-05  9.404e-05   0.992  0.32232
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.07 on 211 degrees of freedom
Multiple R-squared:  0.8362,    Adjusted R-squared:  0.8331
F-statistic: 269.3 on 4 and 211 DF,  p-value: < 2.2e-16
```



Variable Selection

No multicollinearity was observed among the predictors because all variance inflation factors were less than 5.

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

Variable 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
```

Forward selection with AIC and BIC

Variable Selection

```
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_priv

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
```

Backward elimination with AIC and BIC

Final Model

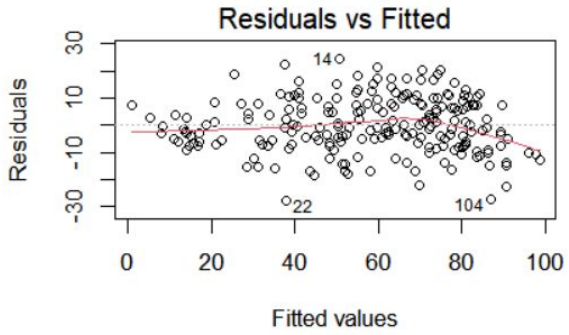
Acceptance Rate = 178.02 - 0.087 * sat_avg - 0.037 * hs_gpa_avg - 7.54 * is_private

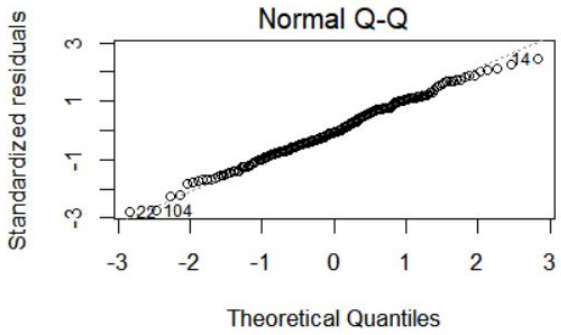
```
Call:
lm(formula = acceptance_rate ~ sat_avg + I(hs_gpa_avg^5) + public_private)

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
```

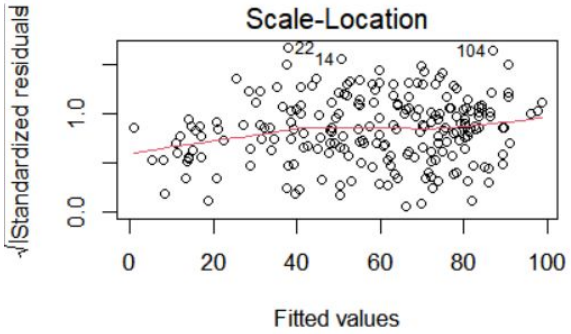
$R^2 = 0.836$

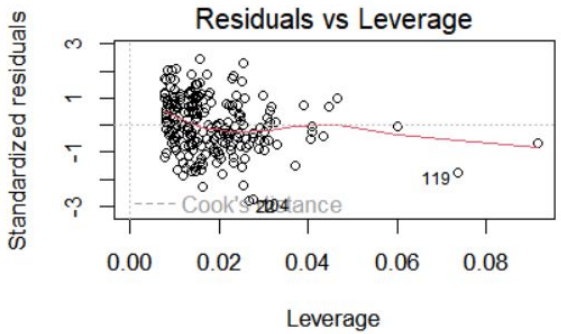
p-value ≈ 0

Final Model









Deficiencies

- Slight nonlinearity near 100% predicted acceptance rate (as AR can't exceed 100%)
- A few outliers and leverage points (such as point 119) often caused by small private colleges with high acceptance rates and high GPAs