GroupEAthleticsProject

Kiersten Hamby,Breanna Ranglall,Krutika Tekwani

What Factors Influence Recruitment Spending at Post-Secondary Education Institutions?

```
library(tidyverse)
library(skimr)
library(Stat2Data)
library(leaps)
library(dplyr)
library(car)
```

Abstract

Our goal was to build a multiple linear regression model using data collected by the U.S. Department of Education Office of Post secondary Education to predict the total recruitment expenses of US Post Secondary Education Institutions using data the universities provided. Using forward, backward, and best subsets for variable selection, our model used Total Student Aid, Head Coach Salary Men, Total Student Count, and Number of Head Coach Women as predictors. The model variables Total Student Aid and Head Coach Salary Men were log transformed for the purpose of this model. While the model proved to be statistically significant, it cannot be soundly used for inference purposes as the independence condition was violated due to relationships between universities in the data set. Ultimately, the model explains 79.72% of the variability in total recruiting expenses.

Introduction

The U.S. Department of Education Office of Postsecondary Education collects this data from all US institutions that have inter college athletic programs annually. This data is important because it helps the public, prospective students, and athletes being recruited aware of the school's commitment to equitable athletic opportunity.

Data Overview

The general topic we want to study is athletic recruitment costs of different U.S. post-secondary educational institutions.The data set is separated by male, female, and coed if the school has coed teams.

This data set includes information regarding athletic participation, athletic student aid, staffing, revenues, and expenses, by men's, women's, and coed varsity teams. It was collected by the U.S. Department of Education Office of Postsecondary Education.

While this data set included more than 4,000 variables, we hand selected 15 variables to be considered for inclusion in the model. These variables were institution name (character), state CD (character), EF Total Count (numeric) representing the number of students at the university, sector name (character), Student Aid Total (numeric) representing the sum of the financial student aid distributed, Head Coach Salary Men (numeric) representing the average amount of men's sport head coaching salary, Head Coach Salary Women (numeric) representing the average amount of women's sport head coaching salaries, Recruiting expense total (numeric) representing the sum of expenditures related to recruitment for both men's and women's sports, number of head coaches women's sport coaches (numeric), and number of head coaches women's sport coaches (numeric).

Our goal is to find the best model the best model to predict recruitment expenses of post secondary institutions.

Loading Data In

```
Athletics=read.csv("~/OneDrive - University of St. Thomas/STAT320/GroupE2AthleticsData/Gro

Athletics <- Athletics %>%
  select(institution_name, state_cd,classification_name,EFTotalCount,sector_name,
         STUDENTAID_TOTAL,HDCOACH_SALARY_MEN,HDCOACH_SALARY_WOMEN,
         RECRUITEXP_TOTAL,NUM_HDCOACH_MEN,NUM_HDCOACH_WOMEN)
```

We selected these variables because we wanted to create a multiple linear regression model and therefore wanted to include primarily quantitative variables. We kept Institution Name, State CD, and Classification Name for identification purposes throughout the data wrangling process.

First Model Attempt

```
oldbest <- regsubsets(RECRUITEXP_TOTAL ~ EFTotalCount+STUDENTAID_TOTAL+
                        HDCOACH_SALARY_MEN+HDCOACH_SALARY_WOMEN+
                        NUM_HDCOACH_MEN+NUM_HDCOACH_WOMEN,
                      data = Athletics, nbest = 1, method = "exhaustive", really.big = T)
with(summary(oldbest), data.frame(rsq, adjr2, cp, rss, outmat))
```

```
              rsq     adjr2        cp          rss EFTotalCount
1  ( 1 ) 0.7871808 0.7870735 623.11633 9.993651e+13
2  ( 1 ) 0.8119311 0.8117413 322.26488 8.831418e+13
3  ( 1 ) 0.8286181 0.8283586 120.07732 8.047820e+13
4  ( 1 ) 0.8332902 0.8329534  64.90867 7.828428e+13
5  ( 1 ) 0.8368515 0.8364393  23.33137 7.661194e+13            *
6  ( 1 ) 0.8383496 0.8378593   7.00000 7.590845e+13            *
         STUDENTAID_TOTAL HDCOACH_SALARY_MEN HDCOACH_SALARY_WOMEN
1  ( 1 )                                  *
2  ( 1 )                                  *
3  ( 1 )                                  *                    *
4  ( 1 )               *                  *                    *
5  ( 1 )               *                  *                    *
6  ( 1 )               *                  *                    *
         NUM_HDCOACH_MEN NUM_HDCOACH_WOMEN
1  ( 1 )
2  ( 1 )               *
3  ( 1 )               *
4  ( 1 )               *
5  ( 1 )               *
6  ( 1 )               *                *
```

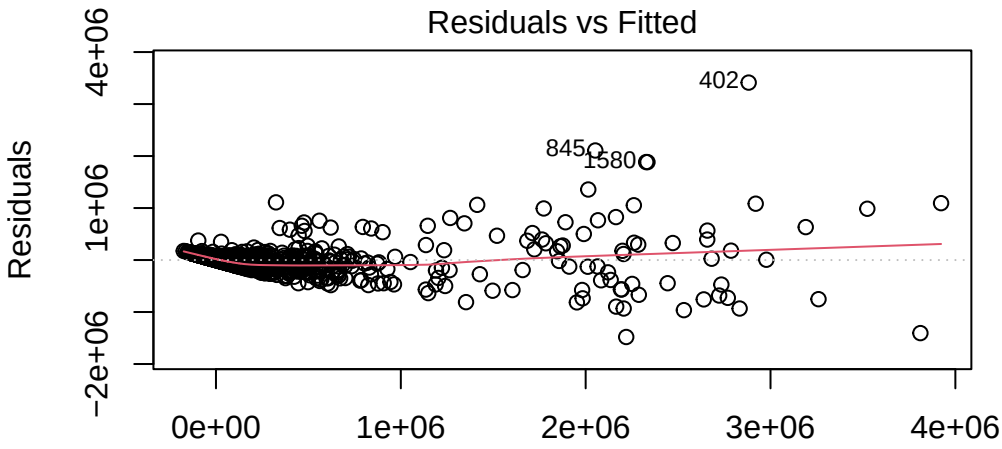
```
OldAthleticsRecruiting<-lm(RECRUITEXP_TOTAL ~ EFTotalCount+STUDENTAID_TOTAL+
                             HDCOACH_SALARY_MEN+HDCOACH_SALARY_WOMEN+
                             NUM_HDCOACH_MEN+NUM_HDCOACH_WOMEN, data = Athletics)
```

```
summary(OldAthleticsRecruiting)
```

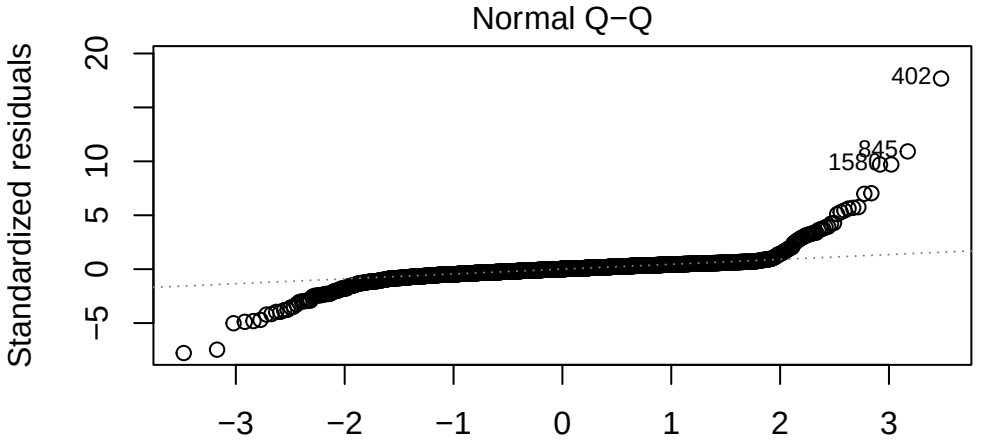
```
Call:
lm(formula = RECRUITEXP_TOTAL ~ EFTotalCount + STUDENTAID_TOTAL +
    HDCOACH_SALARY_MEN + HDCOACH_SALARY_WOMEN + NUM_HDCOACH_MEN +
```

```
    NUM_HDCOACH_WOMEN, data = Athletics)

Residuals:
     Min       1Q   Median       3Q      Max
-1480494   -58642     3037    60273  3412855

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.656e+05  1.008e+04 -16.437  < 2e-16 ***
EFTotalCount          7.739e+00  1.092e+00   7.085 1.92e-12 ***
STUDENTAID_TOTAL      1.370e-02  1.787e-03   7.669 2.71e-14 ***
HDCOACH_SALARY_MEN    1.089e+00  3.787e-02  28.761  < 2e-16 ***
HDCOACH_SALARY_WOMEN  1.763e+00  2.032e-01   8.676  < 2e-16 ***
NUM_HDCOACH_MEN       2.700e+04  3.330e+03   8.106 9.07e-16 ***
NUM_HDCOACH_WOMEN    -1.428e+04  3.336e+03  -4.282 1.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 195900 on 1978 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.8383,    Adjusted R-squared:  0.8379
F-statistic:  1710 on 6 and 1978 DF,  p-value: < 2.2e-16
```

```r
plot(OldAthleticsRecruiting,which = 1)
```




```
plot(OldAthleticsRecruiting, which=2)
```



As shown above, our original attempt to create a model violated the conditions of normality, equality of variance, and independence. Due to this, we had to adjust the model by cleaning the data and transforming certain variables that were resulting in these condition violations.

Data Cleaning

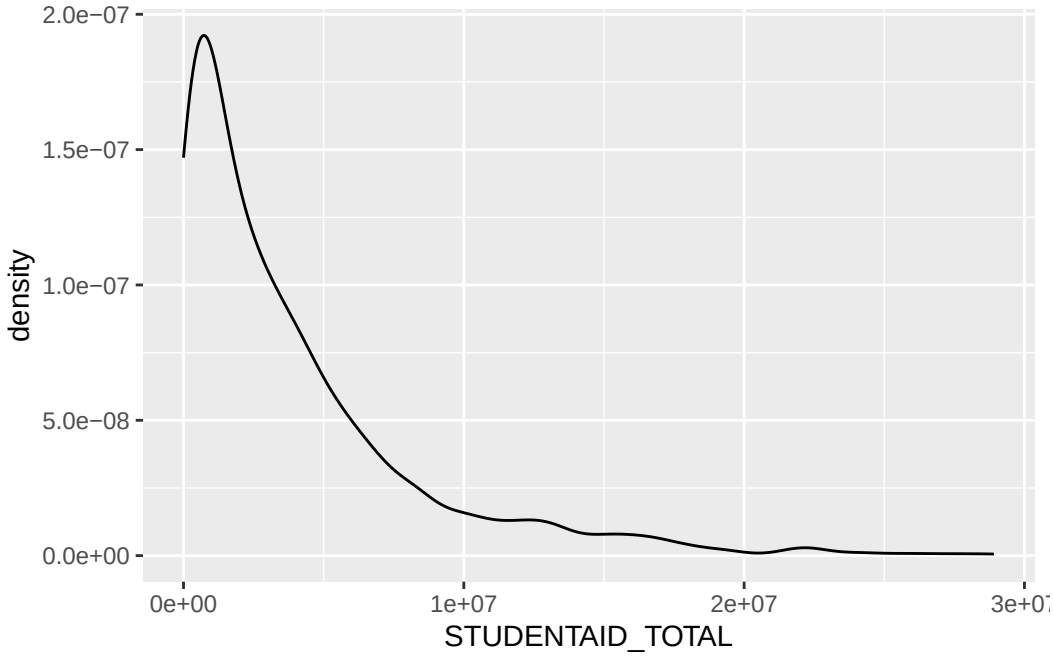
```
Athletics[Athletics==0] <- NA

Athletics <- Athletics %>%
  na.omit(Athletics)
```

We omitted rows that had missing variables and values of 0 so we could perform transformations on the variables to be able to create and appropriate model. This reduced our data set from 2027 observations to 1162, a decrease of 865 observations. This will be important to keep in mind when determining the usefulness of the model.

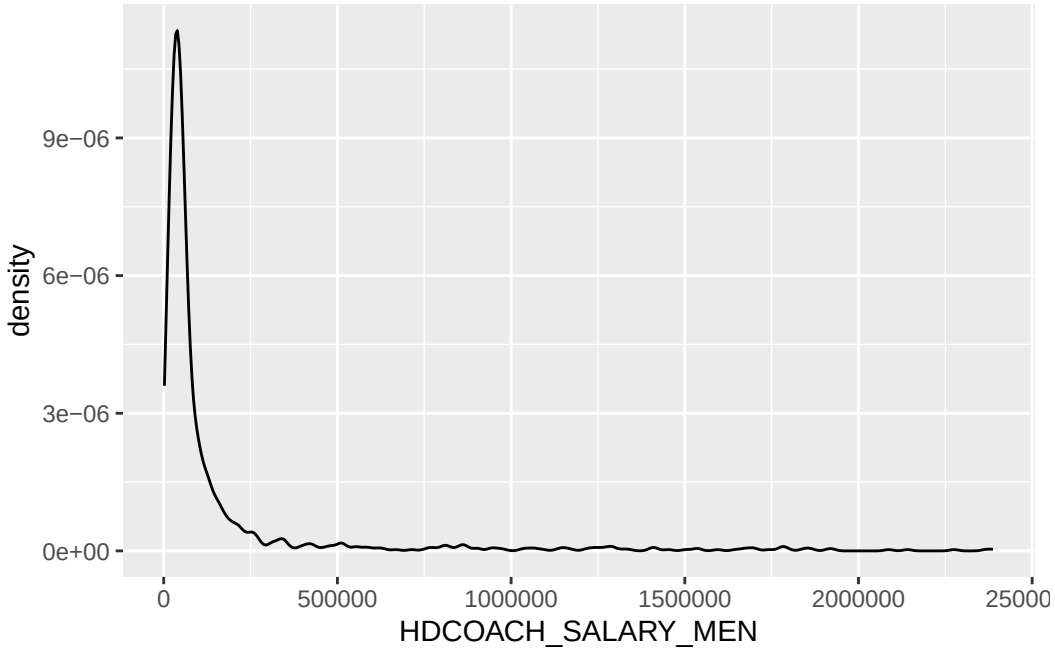
Data Transformations

```
ggplot(Athletics)+geom_density(aes(x=RECRUITEXP_TOTAL))
```

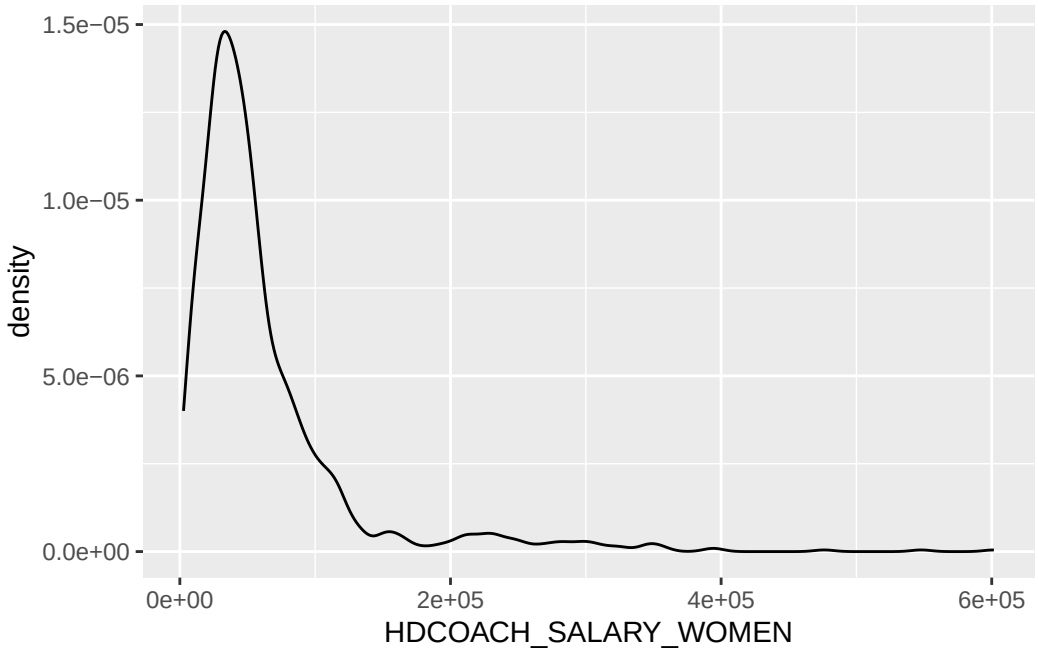
```
ggplot(Athletics)+geom_density(aes(x=STUDENTAID_TOTAL))
```




```
ggplot(Athletics)+geom_density(aes(x=HDCOACH_SALARY_MEN))
```

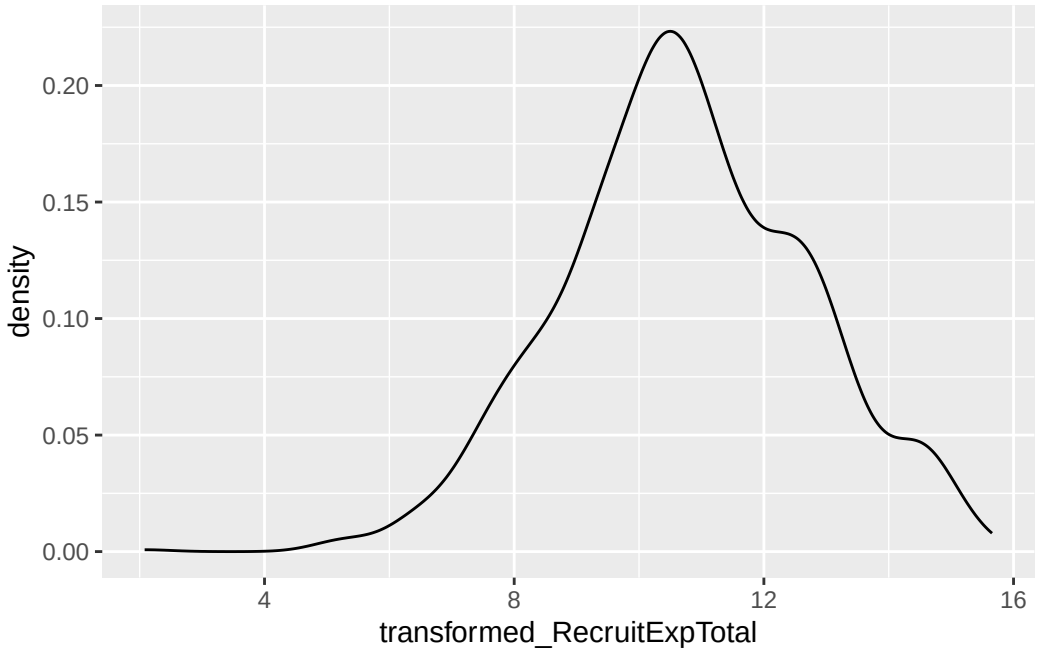



```
ggplot(Athletics)+geom_density(aes(x=HDCOACH_SALARY_WOMEN))
```

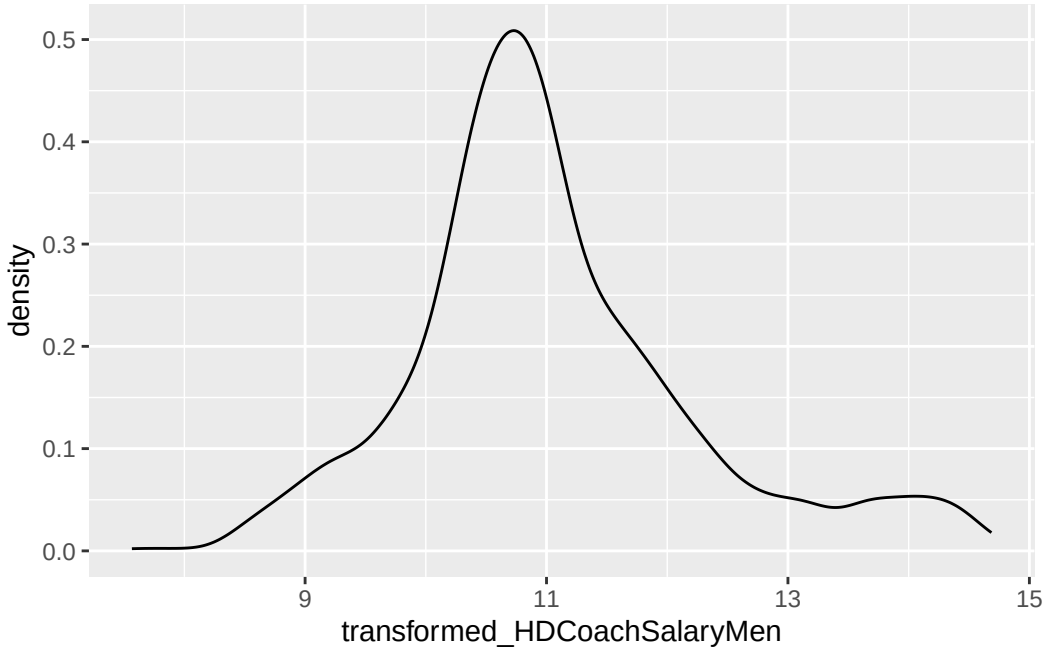


We used Tukey's Bulging Rule to determine that each of these variables should be log-transformed based on the skew of the graphs.

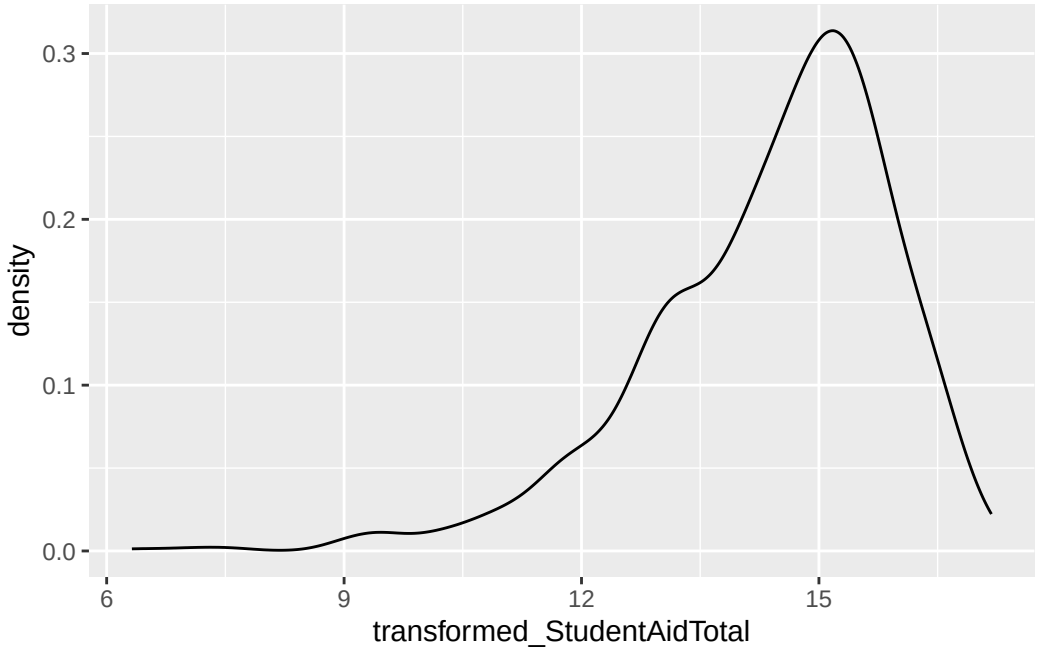
```
Athletics <- Athletics %>%
  mutate(transformed_RecruitExpTotal = log(RECRUITEXP_TOTAL))

Athletics <- Athletics %>%
  mutate(transformed_StudentAidTotal = log(STUDENTAID_TOTAL))

Athletics <- Athletics %>%
  mutate(transformed_HDCoachSalaryMen = log(HDCOACH_SALARY_MEN))

Athletics <- Athletics %>%
  mutate(transformed_HDCoachSalaryWomen = log(HDCOACH_SALARY_WOMEN))


ggplot(Athletics)+geom_density(aes(x=transformed_RecruitExpTotal))
```




```
ggplot(Athletics)+geom_density(aes(x=transformed_HDCoachSalaryMen))
```




```
ggplot(Athletics)+geom_density(aes(x=transformed_StudentAidTotal))
```



```
ggplot(Athletics)+geom_density(aes(x=transformed_HDCoachSalaryWomen))
```

After transforming the variables and rechecking the skew, the log transformation improved each of the variables. The student aid variable is now skewed to the left but it is still improved from the original variable.

Model Selection

```
backward <- regsubsets(transformed_RecruitExpTotal ~ EFTotalCount+
                         transformed_StudentAidTotal+
                         transformed_HDCoachSalaryMen+transformed_HDCoachSalaryWomen+
                         NUM_HDCOACH_MEN+NUM_HDCOACH_WOMEN, data = Athletics,
                       nbest = 1, nvmax = 9,
                       method = "backward")
summary(backward)
```

```
Subset selection object
Call: regsubsets.formula(transformed_RecruitExpTotal ~ EFTotalCount +
    transformed_StudentAidTotal + transformed_HDCoachSalaryMen +
    transformed_HDCoachSalaryWomen + NUM_HDCOACH_MEN + NUM_HDCOACH_WOMEN,
```

```
    data = Athletics, nbest = 1, nvmax = 9, method = "backward")
6 Variables  (and intercept)
                              Forced in Forced out
EFTotalCount                      FALSE      FALSE
transformed_StudentAidTotal       FALSE      FALSE
transformed_HDCoachSalaryMen      FALSE      FALSE
transformed_HDCoachSalaryWomen    FALSE      FALSE
NUM_HDCOACH_MEN                   FALSE      FALSE
NUM_HDCOACH_WOMEN                 FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: backward
         EFTotalCount transformed_StudentAidTotal transformed_HDCoachSalaryMen
1  ( 1 ) " "          " "                         "*"
2  ( 1 ) " "          "*"                         "*"
3  ( 1 ) " "          "*"                         "*"
4  ( 1 ) "*"          "*"                         "*"
5  ( 1 ) "*"          "*"                         "*"
6  ( 1 ) "*"          "*"                         "*"
         transformed_HDCoachSalaryWomen NUM_HDCOACH_MEN NUM_HDCOACH_WOMEN
1  ( 1 ) " "                            " "             " "
2  ( 1 ) " "                            " "             " "
3  ( 1 ) " "                            " "             "*"
4  ( 1 ) " "                            " "             "*"
5  ( 1 ) "*"                            " "             "*"
6  ( 1 ) "*"                            "*"             "*"
```

```
with(summary(backward), data.frame(cp, outmat))
```

```
                cp EFTotalCount transformed_StudentAidTotal
1  ( 1 ) 505.554729
2  ( 1 )  51.797988                                      *
3  ( 1 )   8.108735                                      *
4  ( 1 )   5.942268           *                          *
5  ( 1 )   6.216886           *                          *
6  ( 1 )   7.000000           *                          *
         transformed_HDCoachSalaryMen transformed_HDCoachSalaryWomen
1  ( 1 )                            *
2  ( 1 )                            *
3  ( 1 )                            *
4  ( 1 )                            *
5  ( 1 )                            *                              *
```

```
6  ( 1 )                          *                                *
         NUM_HDCOACH_MEN NUM_HDCOACH_WOMEN
1  ( 1 )
2  ( 1 )
3  ( 1 )                                *
4  ( 1 )                                *
5  ( 1 )                                *
6  ( 1 )               *                *
```

```
best <- regsubsets(transformed_RecruitExpTotal ~ EFTotalCount+
                     transformed_StudentAidTotal+
                     transformed_HDCoachSalaryMen+
                     transformed_HDCoachSalaryWomen+
                     NUM_HDCOACH_MEN+NUM_HDCOACH_WOMEN, data = Athletics,
                   nbest = 1, method = "exhaustive", really.big = T)
```

```
forward <- regsubsets(transformed_RecruitExpTotal ~ EFTotalCount+
                        transformed_StudentAidTotal+
                        transformed_HDCoachSalaryMen+transformed_HDCoachSalaryWomen+
                        NUM_HDCOACH_MEN+NUM_HDCOACH_WOMEN, data = Athletics,
                      nbest = 1,
                      nvmax = 8, method = "forward")
```

All selection methods lead us to choose a model based on 4 variables, Transformed Student Aid Total, Transformed Head Coach Salary Men, Total Student Count, and Number of Head Coaches Women to predict Transformed Recruiting Expense Total because of the low CP value.

```
AthleticsRecruiting<-lm(transformed_RecruitExpTotal~transformed_StudentAidTotal+
                          transformed_HDCoachSalaryMen+EFTotalCount+NUM_HDCOACH_WOMEN,
                        data=Athletics)
```

Summary Output

```
summary(AthleticsRecruiting)
```

```
Call:
lm(formula = transformed_RecruitExpTotal ~ transformed_StudentAidTotal +
```

```
    transformed_HDCoachSalaryMen + EFTotalCount + NUM_HDCOACH_WOMEN,
    data = Athletics)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1140 -0.4513  0.1074  0.5448  2.5860

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -5.736e+00  3.323e-01 -17.261  < 2e-16 ***
transformed_StudentAidTotal  4.287e-01  2.794e-02  15.343  < 2e-16 ***
transformed_HDCoachSalaryMen 8.980e-01  4.042e-02  22.218  < 2e-16 ***
EFTotalCount                 1.063e-05  5.210e-06   2.040   0.0415 *
NUM_HDCOACH_WOMEN            4.670e-02  8.288e-03   5.635  2.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8863 on 1157 degrees of freedom
Multiple R-squared:  0.7972,    Adjusted R-squared:  0.7965
F-statistic:  1137 on 4 and 1157 DF,  p-value: < 2.2e-16
```

$$\hat{TotalRecruitingExpense} = -5.736 + 0.4287 * log(StudentAidTotal) + 0.898 * log(HDCoachSalaryMen)$$

$$+ 0.00001063 * EFTotalCount \; + \; 0.0467 * NumHDCoachWomen$$

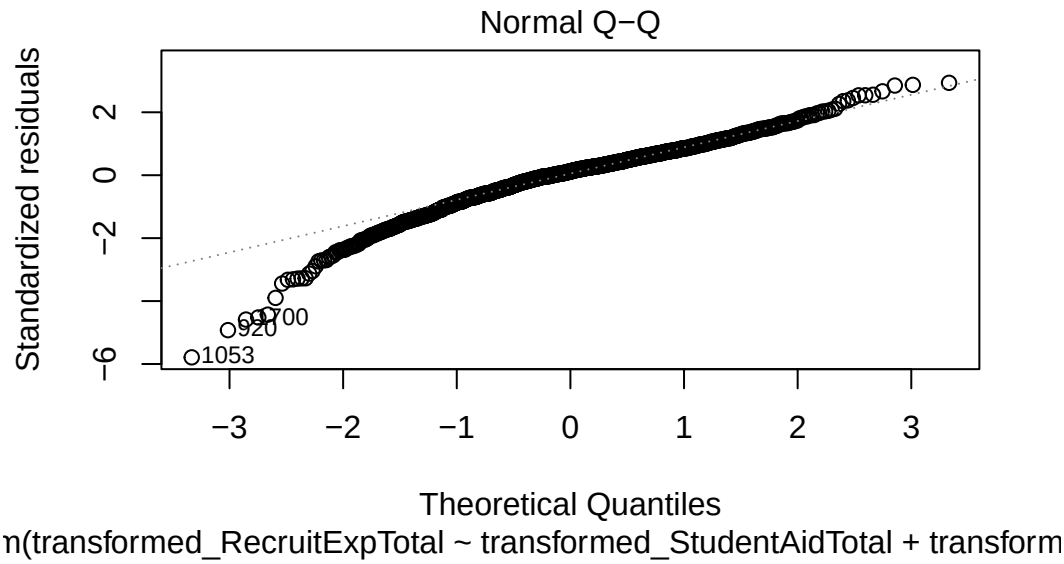
All the variables are significant with p-values of less than 0.05.

The model has a p-value of almost 0 and therefore we have evidence that it is statistically significant.

Condition Check

QQ Plot to Check Normality

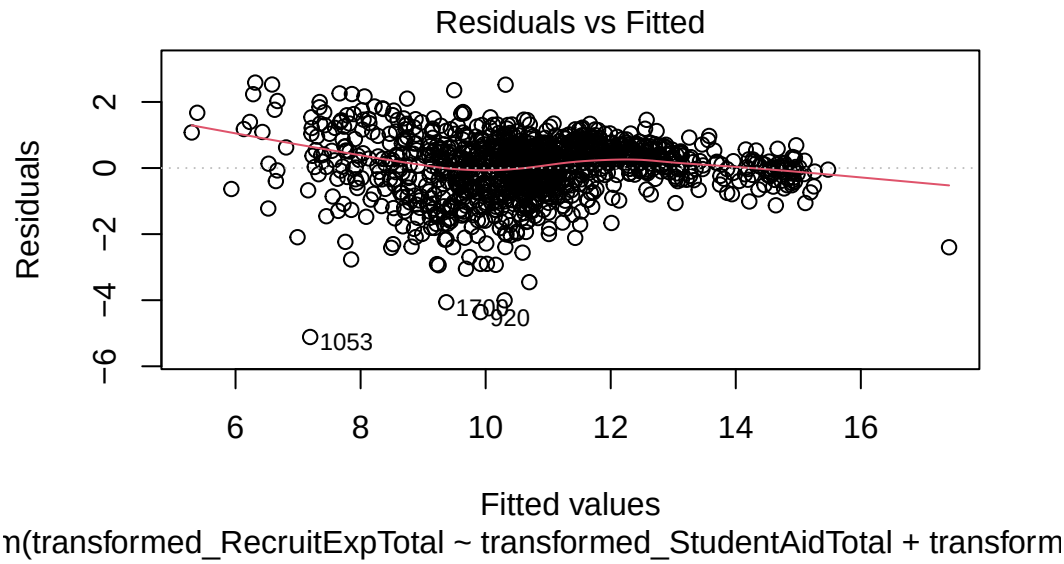
```r
plot(AthleticsRecruiting,which = 2)
```



The normality condition is not completely satisfied as the QQ plot shows points deviating from the line on both ends creating an s-shape. It is however a significant improvement from the model with untransformed variables and is not so bad that we would not continue forward with this model.

Residual vs Fitted Plot for Linearity and Equality of Variance

```
plot(AthleticsRecruiting,which=1)
```



The linearity condition looks pretty good as there is no obvious curvature to the line shown in the residuals vs fitted plot.

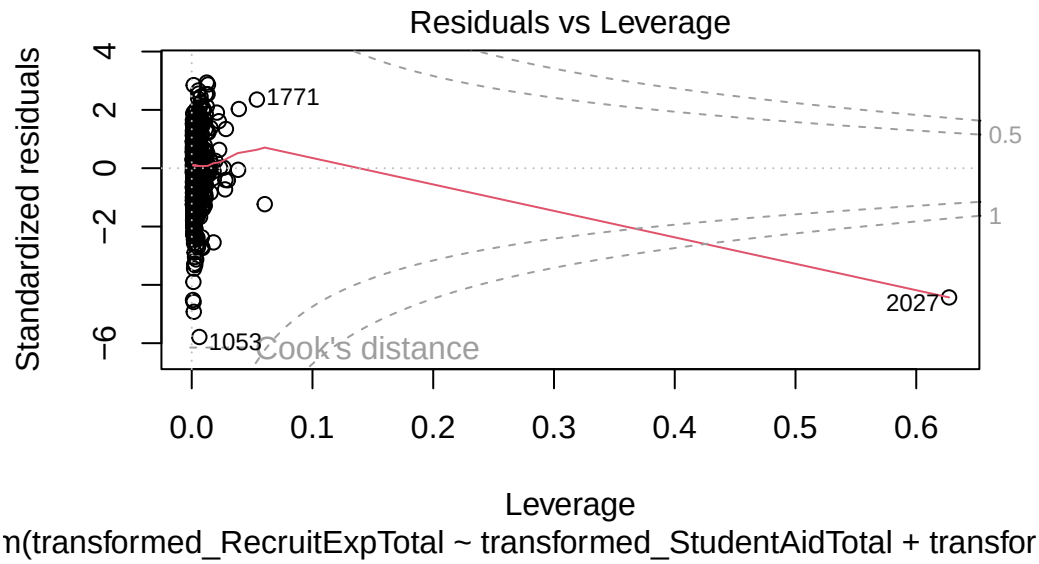
The normality condition is not completely satisfied as the bands of residuals are not uniform throughout the graph, however, this is also a significant improvement from the untransformed model.

Independence

Because we were unable to filter out campuses that are related we weren't able to maintain the independence condition, however, this is something we were aware of and should be kept in mind when trying to draw any conclusions from the model.

Outliers

```
plot(AthleticsRecruiting, which = 5)
```



There appears to be one outlier based on the Residuals vs Leverage plot. Due to the low number of outliers and the high number of observations, we chose to move forward with the model without removing the outlier.

When researching further, the outlier is The Pennsylvania State University and appears to have an abnormally large Student Aid Total and Total Number of Students.

Multicollinearity Test

```
vif(AthleticsRecruiting)
```

```
transformed_StudentAidTotal transformed_HDCoachSalaryMen
                   2.779621                     3.415905
                EFTotalCount           NUM_HDCOACH_WOMEN
                   2.222141                     1.669834
```

None of the VIF values are greater than 5 so multicollinearity is not concerning at this point.

Coefficient Interpretation

For a 1% increase in Total Student Aid, holding all other variables constant, we expect Total Recruiting Expenses to increase by .4287%.

For a 1% increase in Head Coach Salary Men, holding all other variables constant, we expect Total Recruiting Expenses to increase by .898%

For a 1-student increase in EF Total Count, holding all other variables constant, we expect Total Recruiting Expense to increase by 0.00001063%.

For a 1-person increase in Number of Head Coaches Women, holding all other variables constant, we expect Total Recruiting Expenses to increase by 0.0467%.

Conclusion

With this project we hoped to make a multiple linear regression model that could predict U.S. Post Secondary Education Institution's total recruitment expenses using various explanatory variables collected by the U.S. Department of Education Office of Post secondary Education.

We rejected the null hypothesis that there was no relationship between total recruiting expenses and the explanatory variables used in this model. That being said, our significant p-value suggests there is evidence to suggest that there is a relationship between Total Student Aid (transformed), Total Student Count, Head Coach Salary Men (transformed), and Number of Head Coaches Women and Total Recruiting Expenses (transformed).

The model is faced with limitations in that the data collected is from US schools during the 2021-2022 academic year and therefore any conclusions or predictions being made outside of those characteristics are not sound. Further with the dropping of 865 observations due to missing values, we have to be aware that there is a possibility of the model missing statistically significant data that would change the model if included.

The next steps we identified are to gather sufficient data from all schools so broader conclusions can be drawn from the model. Further, determining a way to filter out related schools would help satisfy the independence condition so that the model could be used to draw inferences. We were also interested into diving more into the data associated with gender to identify potential inequities.