**College Outstate Tuition Fee Analysis**
**Data Ninjas**
**21st Dec 2015**

## Introduction:

The topic of college tuition, rising student debt and diminishing job prospects have dominated the news in the past three to five years. There are many factors that have contributed to the pressure that Colleges and Universities have felt and the data set we are using reveals some of them. Of the variables listed in the college data set the most interesting and relevant response to our audience and the analysis team was, Out of State College Tuition. Through process of elimination we trimmed the potential response variables based on project criteria and also interest.

The biggest question we continuously asked was, why do we care? The relevancy of out of state tuition leads to a much larger discussion about the direction of Colleges and Universities. Colleges and Universities were originally places for only higher education and extra-curricular activities but through time these institutions have adapted to provide a more lavish and luxurious experience. This adaptability was a result of increasing competition, large alumni re-investment and a means to market the school. The drawback of a more lavish and luxurious experience is the forwarding of cost to those pursuing higher education. As a result some have been shut out and others have drowned in debt. This project merely scrapes the surface of a very interesting and relevant issue of higher education expense and equal opportunity across social levels.

## Data Description:

Data is from the 1995 U.S. News report on American colleges and universities. This includes 18 variables which is demographic information on tuition, room & board costs, application/acceptance rates, student/faculty ratio, graduation rate, and more. The dataset is used for the 1995 Data Analysis Exposition, sponsored by the Statistical Graphics Section of the American Statistical Association. Total 770+ college data is available.

## Research Questions:

In this project we are attempting to answer the question how well tuition fees can be modeled using predictor variables. But using the same data we can answer different questions like, how we can cluster colleges into similar comparison groups? How can we find a reasonable way to rank the schools? But currently this analysis is out of scope.

## Objective:

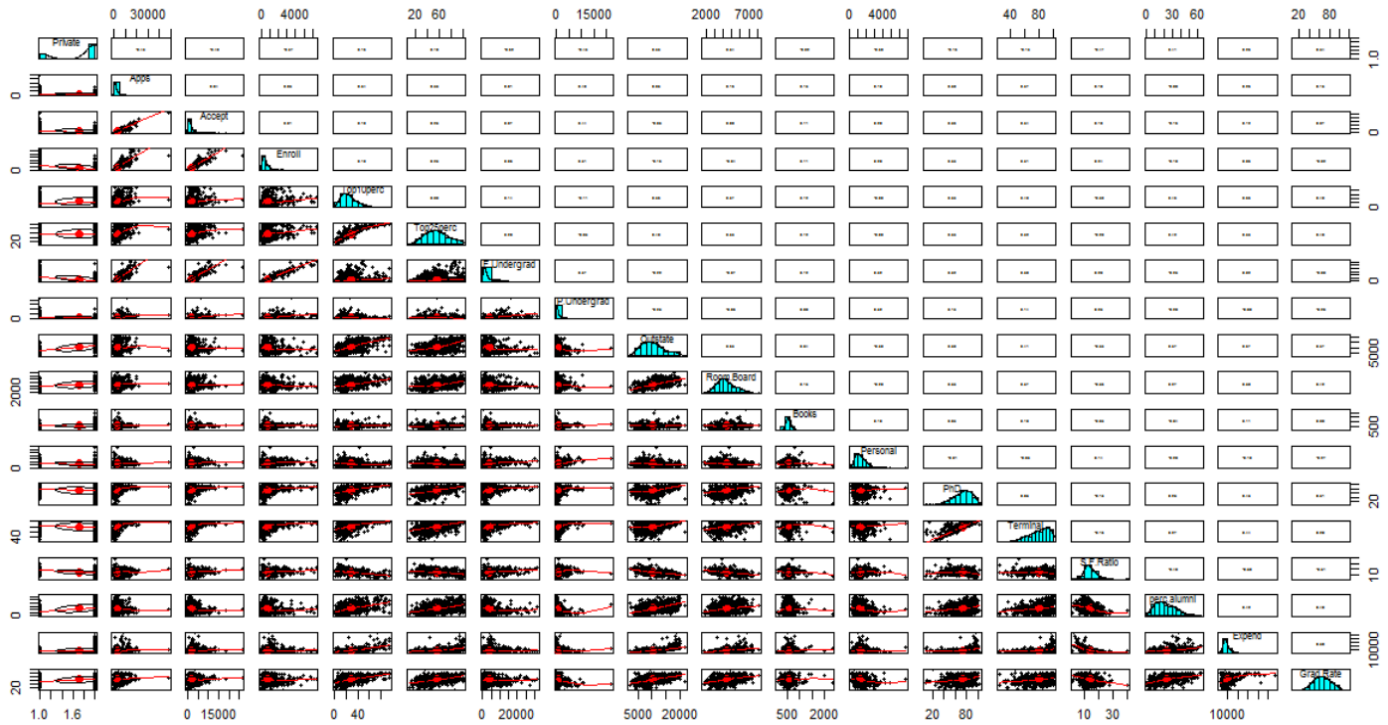
Analyze college data to understand relationship between Outstate tuition fee and other predictor variables.

## Reading and Cleaning Data:

After reading input file which is in csv format, we realized that first column is University/College Names. Then we fixed that column as row names and removed that column as variable from dataset. Upon reviewing structure of data we realized that there is one categorical variable and remaining are quantitative variables.

## Exploratory Data Analysis:

Just to have good idea about how variables are correlated to each other, we plotted pairwise scatterplot




**Modeling Steps:**

1. **Removing Multicollinearity:**

   After running vfstep function we identified that two variables 'Accept' and 'Enroll' have strong multicollinearity problem since they have VIF value greater than 10. Hence we removed those variables. Now we are left with only with 15 predictor variables.

```
---------- VIFs of the remained variables --------
      Variables      VIF
1   PrivateYes 2.399406
2         Apps 4.105399
3    Top10perc 6.809897
4    Top25perc 5.573362
5  F.Undergrad 4.998398
6  P.Undergrad 1.691743
7   Room.Board 1.762908
8        Books 1.107194
9     Personal 1.292643
10         PhD 4.080675
11    Terminal 3.987629
12   S.F.Ratio 1.886845
13 perc.alumni 1.763968
14       Expend 2.812110
15   Grad.Rate 1.809994
```

2. **Variable Selection**

   After removing highly correlated variables, we ran different variables selection methods to get optimal model. First we used all subset method using Mallow's Cp criteria and Adjusted R2. Then also applied Stepwise regression using AIC and BIC criteria.

   Let's look into details:

   a. Mallow's Cp Criteria:

      result1=leaps(Y, Outstate, int=TRUE, method=c("Cp"), nbest=16)

      The model with the smallest Cp value is one with predictors Private, Top10perc, Room.Board, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend and Grad.Rate.

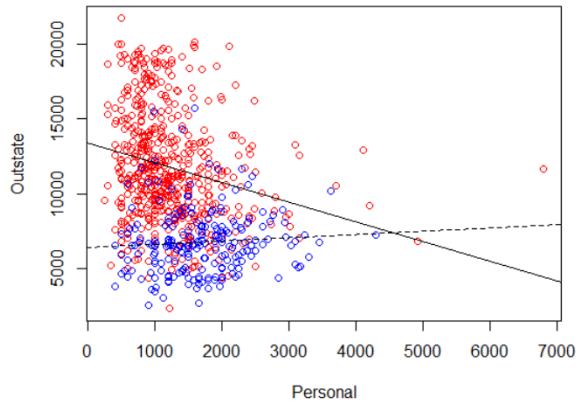
      And Cp value is 10.35987.

b. Adjusted R2:
result2=leaps(Y, Outstate, int=TRUE, method=c("adjr2"), nbest=16)
The model with the largest Adjusted-R2 value is one with predictors Private, Apps, Top10perc, F.Undergrad, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend and Grad.Rate. and Adjusted R2 value is 0.7515843.

c. AIC Criteria:
fit.AIC = step(model1,direction="both",k=2)
The model with smallest AIC value is one with predictors PrivateYes, Apps, Top10perc, F.Undergrad, Room.Board, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend and Grad.Rate. And AIC value is 11828.69

d. BIC Criteria:
fit.BIC = step(model1,direction="both",k=log(777))
The model with smallest BIC value is one with predictors PrivateYes, Room.Board, Personal, Terminal, perc.alumni, Expend and Grad.Rate. And BIC value is 11871.47

3. **Finding Interaction Terms:**
After selecting variables using all subset method and stepwise regression method, we plotted a scatter plot between Outstate and every quantitative variable for each subgroup of Private.
These different scatterplots showed that Private variable is interacting with 5 other variables named Personal, Terminal, S.F.Ratio, Grad.Rate and perc.alumni. The diagram below shows one such interaction:




4. **Partial F-test:**
After including the interaction term in models we observed that Private:S.F.Ratio, Private:Grad.Rate, and Private:perc.alumni interaction terms are not significant at 0.05 significance level. Hence we conducted partial-F test to check whether we can remove all these terms fromthe model. For all four models using Mallow's Cp, Adjusted R2, AIC, BIC, F-Statistic is small number hence p-value is greater than 0.05 and we concluded that we failed to reject null hypothesis and we have enough evidence to remove these interaction terms from the models
We removed these interaction terms from model and refitted all four models.

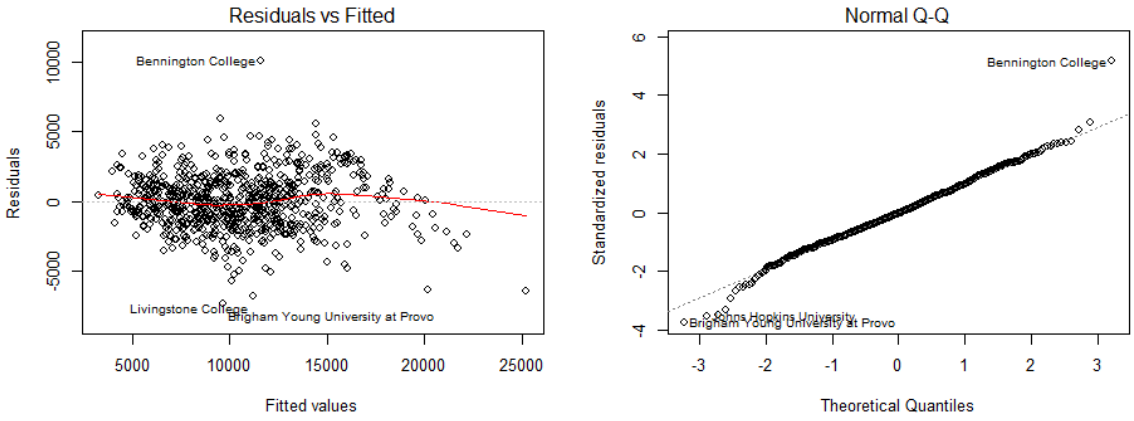
5. **Cross Validation:** The purpose of cross validation is to test how well your model able to get trained by some data and then predict independent data effectively. After refitting model we wanted to check which model gives minimum prediction error using PRESS and k-fold validation method.

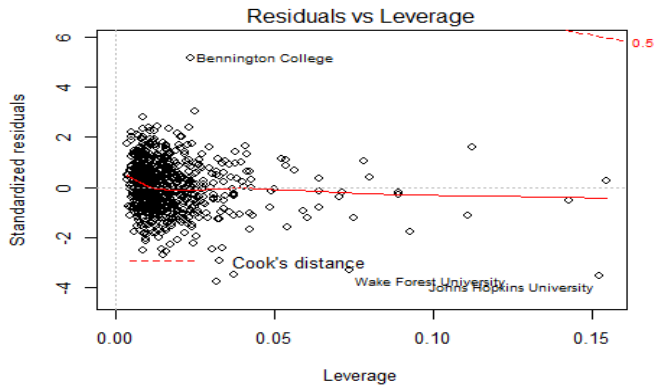
| | PRESS | K-fold |
|---|---|---|
| **All subset method(Mallow's Cp)** | 3124506861 | 4010605 |
| **All subset method(Adjusted R2)** | 3171141666 | 4088727 |
| **Stepwise Regression (AIC)** | 3165547476 | 4080693 |
| **Stepwise Regression (BIC)** | 3153020388 | 4061999 |

We selected model which uses Mallow's Cp criteria as optimal model since its giving less prediction error in both PRESS and k-fold methods.

6. **Residual Analysis:**

As we can observe in Residuals vs Fitted plot that there is no trend in a residual plot and vertical variation across the x-axis is hence satisfying linear and constant variance assumption. Also there is no cyclical or systematic pattern in the plot hence satisfying the independence assumption as well.
QQ Plt compares standardized residual and theoretical quantiles and all points are on a straight line hence it holds normality assumption. Some deviation is there but it is at the ends and it is small.

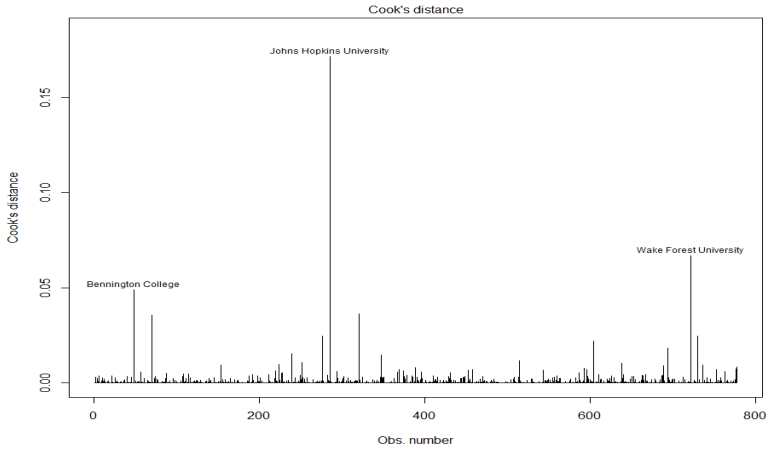



In a plot of standardized residuals against leverage, standardized residuals are centralized around zero. As we know leverage is measure of how each data point influences the regression. Since the regression line must pass through centroid, points that lie far from centroid have greater leverage and then leverage increases if

there are fewer points nearby. Here we can observe that Wake Forest University and John Hopkins University have large leverage and Bennington College has large standardized residual.

## 7. Remove Influential Outliers:




After reviewing residual plots, we also explored cook's distance to identify influential outliers, which measures how much regression would change if an observation is deleted. Simply Cook's distance is used to compare predicted response with or without a particular observation in the model fitting process if the change in the predicted values is significant after removing one observation then that observation is influential.

## 8. Final Model:

The fitted model equation is

$\hat{}$

Outstate = -313.18 + 325.18 Private + 7.76 Top10perc + 0.869 Room.Board + 0.299 Personal +18.79 PhD - 14.38 Terminal - 46.5 S.F.Ratio + 30.95 Perc.Alumni + 0.2259 Expend + 28.63 Grad.Rate - 0.8521 Private:Personal + 44.66 Private:Terminal. After removing influential observation model adjusted-R2 increases from 0.75 to 0.77 that is model now explains total 77% variation in response Outstate tuition fees. Model has standard error of 1920.

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -313.1855  1097.7004   -0.29  0.77548
PrivateYes          325.1804  1056.1824    0.31  0.75826
Top10perc             7.7666     6.0194    1.29  0.19735
Room.Board            0.8690     0.0817   10.64  < 2e-16 ***
Personal              0.2999     0.1992    1.51  0.13264
PhD                  18.7904     8.4432    2.23  0.02634 *
Terminal            -14.3883    13.2091   -1.09  0.27638
S.F.Ratio           -46.5046    23.7904   -1.95  0.05098 .
perc.alumni          30.9533     7.3824    4.19  3.1e-05 ***
Expend                0.2259     0.0238    9.50  < 2e-16 ***
Grad.Rate            28.6332     5.2352    5.47  6.1e-08 ***
PrivateYes:Personal   -0.8521     0.2376   -3.59  0.00036 ***
PrivateYes:Terminal   44.6604    12.4511    3.59  0.00036 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1920 on 761 degrees of freedom
Multiple R-squared:  0.774,     Adjusted R-squared:  0.77
F-statistic:  217 on 12 and 761 DF,  p-value: <2e-16
```

## 9. Regression Coefficient Interpretation:

b1 = 325.18, Private colleges outstate tuition fee is $325.18 greater than public college, keeping all other variables constant.
b2 = 7.76, If Top 10 percent student increase by 1 then fees increases by $7.76, keeping all other variables constant.
b3 = 0.869, If room and board cost increases by $1 then fees increases by 0.869, keeping all other variables constant.
b4 = 0.299, If personal expenditure increases by $1 then fees increases by $0.299, keeping all other variables constant.
b5 = 18.79, If percent of Phd faculty increases by 1 then fees increases by $18.79, keeping all other variables constant.
b6 = -14.38, If percent of faculty with Terminal degree increases by 1 then fees decrease by $14.38, keeping all other variables constant.
b7 = -46.5, If student-faculty ratio was 1 then fees decreases by $46.5, keeping all other variables constant.
b8 = 30.95, If alumni donation percentage increase by 1 then fees increases by $30.95, keeping all other variables constant
b9 = 0.2259, If instructional expenditure increases by 1 then fees increase by $0.2259, keeping all other variables constant.
b10 = 28.63, If graduation rate increase by 1 then fees increases by $28.63, keeping all other variables constant.
b11 = -0.8521, For private college if personal expenditure increase by $1 then fees decreases by $0.8521, keeping all other variables constant.
b12 = 44.66, For private college if percent of terminal degree increases by 1 then fees increases by 44.66, keeping all other variables constant.

**Conclusion:**

After above analysis processes, we achieved a model with outstate tuition as response variables, including ten first-order variables along with two interaction terms. There are two practical uses of the model we built. First, the model can be used by universities to decide their reasonable outstate tuition so that they will be able to draw more students without compromising their profitability. The other use is to help students judge whether the tuition of a university is consistent with its other attributes, and thus assist them in choosing which university to apply.

**References:**


http://www.amstat.org/publications/jse/datasets/colleges.txt
http://www.amstat.org/publications/jse/datasets/usnews.txt
http://www.amstat.org/publications/jse/datasets/usnews3.dat.txt