568CCourseProject

Lakshmi

3/13/2020

Problem Statement

Education funding begins with Saving. The cost of a college education is skyrocketing. It's no surprise that student loans now make up the largest chunk of U.S. non-housing debt. While the cost of college and other education costs continue to rise, the good news is that there are smarter ways to save for education. If we know how much to save for college education today, we will have the flexibility of making financial adjustments without compromising on the quality of life and make more informed financial decisions. With this motto, I would like to answer the following questions through modelling:

- What would be the tuition costs for any 4-year public university program in USA in the future?
- Which factor affects Education costs the most?

Variables of Interest

Output variable:

Average Tuition and Fees and Room and Board (Enrollment-Weighted) in Current Dollars from 1986-2019 for any 4-year Public University degree.

Input variables:

- **Consumer price index (CPI):** The Consumer Price Index (CPI) is a measure that examines the average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. Changes in the CPI are used to assess price changes associated with the cost of living; the CPI is one of the most frequently used statistics for identifying periods of inflation or deflation.
- **Debt-to-GDP ratio:** A metric comparing a country's public debt to its gross domestic product (GDP). By comparing what a country owes with what it produces, the debt-to-GDP ratio reliably indicates that particular country's ability to pay back its debts. Often expressed as a percentage, this ratio can also be interpreted as the number of years needed to pay back debt, if GDP is dedicated entirely to debt repayment.
- **Year:** For this problem, we consider data for the years 1986-2019 to make inferences.

Data Source

Data is obtained from the below public sources for the years 1986-2019 and combined to make inferences:

Consumer Price Index, All Urban Consumers from Bureau of Labor Statistics

College price data from https://collegecost.ed.gov/affordability

Debt_GDP data from https://fred.stlouisfed.org/series/GFDEGDQ188S

Cleaning Data and Feature Engineering

The College price dataset is hierarchical in nature with years spread across columns, and Tuition and Fees for public 2-year, public 4-year and private non-profit 4-year as rows grouped by Dollars Value (current and 2019). The response variable for this analysis is ‘Average Tuition Fees for a 4-year public university program’. In order to use this data for further analysis, some cleansing needs to be done.

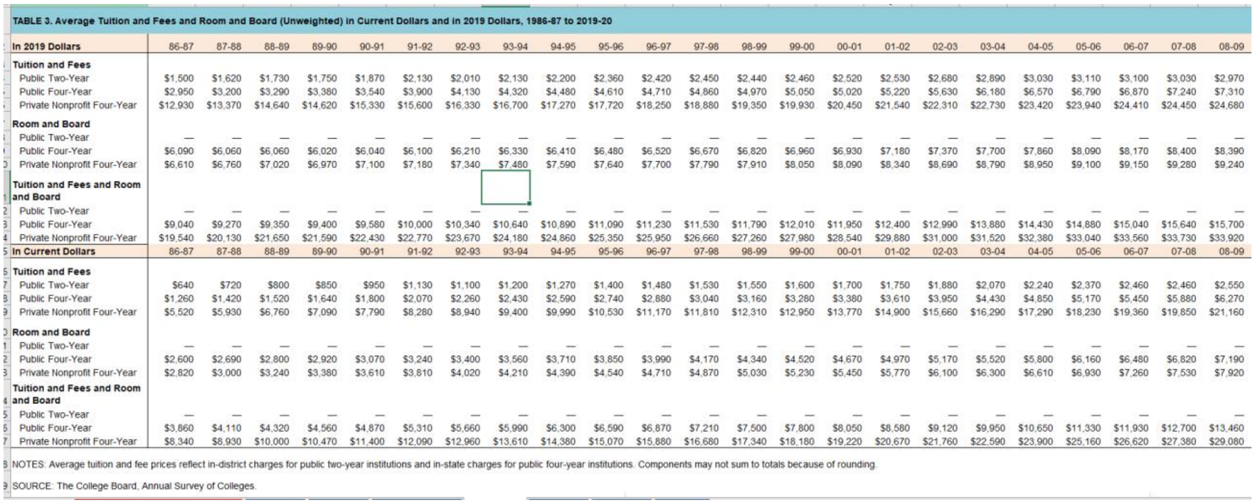


TABLE 3. Average Tuition and Fees and Room and Board (Unweighted) in Current Dollars and in 2019 Dollars, 1986-87 to 2019-20

In 2019 Dollars	86-87	87-88	88-89	89-90	90-91	91-92	92-93	93-94	94-95	95-96	96-97	97-98	98-99	99-00	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09
Tuition and Fees																							
Public Two-Year	$1,500	$1,620	$1,730	$1,750	$1,870	$2,130	$2,010	$2,130	$2,200	$2,360	$2,420	$2,450	$2,440	$2,460	$2,520	$2,530	$2,680	$2,890	$3,030	$3,110	$3,100	$3,030	$2,970
Public Four-Year	$2,950	$3,200	$3,290	$3,380	$3,540	$3,900	$4,130	$4,320	$4,480	$4,610	$4,710	$4,860	$4,970	$5,050	$5,020	$5,220	$5,630	$6,180	$6,570	$6,790	$6,870	$7,240	$7,310
Private Nonprofit Four-Year	$12,930	$13,370	$14,640	$14,620	$15,330	$15,600	$16,330	$16,700	$17,270	$17,720	$18,250	$18,880	$19,350	$19,930	$20,450	$21,540	$22,310	$22,730	$23,420	$23,940	$24,410	$24,450	$24,680
Room and Board																							
Public Two-Year	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Public Four-Year	$6,090	$6,060	$6,060	$6,020	$6,040	$6,100	$6,210	$6,330	$6,410	$6,480	$6,520	$6,670	$6,820	$6,960	$6,930	$7,180	$7,370	$7,700	$7,860	$8,090	$8,170	$8,400	$8,390
Private Nonprofit Four-Year	$6,610	$6,760	$7,020	$6,970	$7,100	$7,180	$7,340	$7,480	$7,590	$7,640	$7,700	$7,790	$7,910	$8,050	$8,090	$8,340	$8,690	$8,790	$8,950	$9,100	$9,150	$9,280	$9,240
Tuition and Fees and Room and Board																							
Public Two-Year	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Public Four-Year	$9,040	$9,270	$9,350	$9,400	$9,580	$10,000	$10,340	$10,640	$10,890	$11,090	$11,230	$11,530	$11,790	$12,010	$11,950	$12,400	$12,990	$13,880	$14,430	$14,880	$15,040	$15,640	$15,700
Private Nonprofit Four-Year	$19,540	$20,130	$21,650	$21,590	$22,430	$22,770	$23,670	$24,180	$24,860	$25,350	$25,950	$26,660	$27,260	$27,980	$28,540	$29,880	$31,000	$31,520	$32,380	$33,040	$33,560	$33,730	$33,920
In Current Dollars	86-87	87-88	88-89	89-90	90-91	91-92	92-93	93-94	94-95	95-96	96-97	97-98	98-99	99-00	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	08-09
Tuition and Fees																							
Public Two-Year	$640	$720	$800	$850	$950	$1,130	$1,100	$1,200	$1,270	$1,400	$1,480	$1,530	$1,550	$1,600	$1,700	$1,750	$1,880	$2,070	$2,240	$2,370	$2,460	$2,460	$2,550
Public Four-Year	$1,260	$1,420	$1,520	$1,640	$1,800	$2,070	$2,260	$2,430	$2,590	$2,740	$2,880	$3,040	$3,160	$3,280	$3,380	$3,610	$3,950	$4,430	$4,850	$5,170	$5,450	$5,880	$6,270
Private Nonprofit Four-Year	$5,520	$5,930	$6,760	$7,090	$7,790	$8,280	$8,940	$9,400	$9,990	$10,530	$11,170	$11,810	$12,310	$12,950	$13,770	$14,900	$15,660	$16,290	$17,290	$18,230	$19,360	$19,850	$21,160
Room and Board																							
Public Two-Year	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Public Four-Year	$2,600	$2,690	$2,800	$2,920	$3,070	$3,240	$3,400	$3,560	$3,710	$3,850	$3,990	$4,170	$4,340	$4,520	$4,670	$4,970	$5,170	$5,520	$5,800	$6,160	$6,480	$6,820	$7,190
Private Nonprofit Four-Year	$2,820	$3,000	$3,240	$3,380	$3,610	$3,810	$4,020	$4,210	$4,390	$4,540	$4,710	$4,870	$5,030	$5,230	$5,450	$5,770	$6,100	$6,300	$6,610	$6,930	$7,260	$7,530	$7,920
Tuition and Fees and Room and Board																							
Public Two-Year	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Public Four-Year	$3,860	$4,110	$4,320	$4,560	$4,870	$5,310	$5,660	$5,990	$6,300	$6,590	$6,870	$7,210	$7,500	$7,800	$8,050	$8,580	$9,120	$9,950	$10,650	$11,330	$11,930	$12,700	$13,460
Private Nonprofit Four-Year	$8,340	$8,930	$10,000	$10,470	$11,400	$12,090	$12,960	$13,610	$14,380	$15,070	$15,880	$16,680	$17,340	$18,180	$19,220	$20,670	$21,760	$22,590	$23,900	$25,160	$26,620	$27,380	$29,080

NOTES: Average tuition and fee prices reflect in-district charges for public two-year institutions and in-state charges for public four-year institutions. Components may not sum to totals because of rounding.

SOURCE: The College Board, Annual Survey of Colleges.

Read the education costs data set, clean the data by converting missing values (literal NA’s and hyphens) to 0’s as costs are numeric in nature. Scrap off the irrelevant texts at the footer of the table. Convert all the costs into numeric values. Combine tuition fees and board room fees into total costs for public2year, public4year and privatenonprofit4year courses. To narrow the analysis and to produce more detailed insights about Tuition Fees, only the 4-year public university tuition fees in ‘In Current Dollars’ have been considered for analysis.

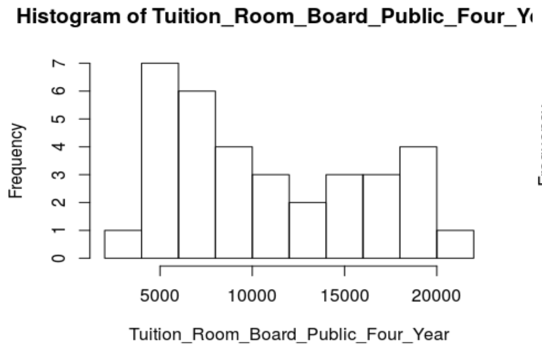
The aim is to assess the effects of variables namely Consumer Price Index(CPI), Debt_GDP ratio, Year on tuition costs, the inflation and Debt_to_GDP datasets are pulled from public sources listed under ‘Data Source’ for the same years as Education costs and merged with the Tuition Fees dataset. Column names are renamed appropriately after merging.

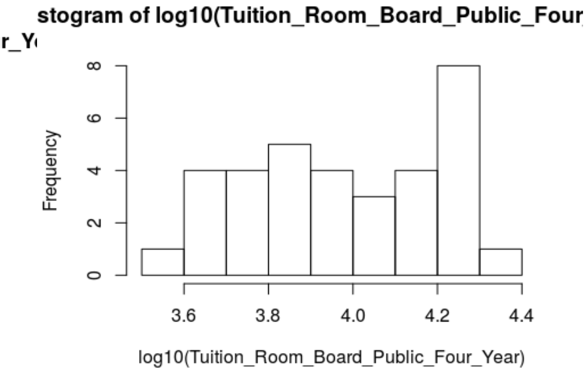
To get a preliminary understanding of the distributions of the chosen variables of interest, we plotted histograms of the numerical variables. As seen in the charts below, all the numerical variables of interest are not very normal.

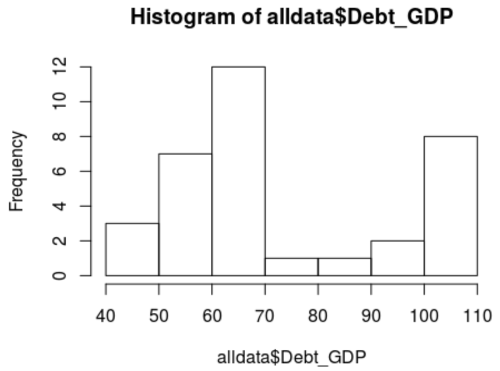
Descriptive Statistics

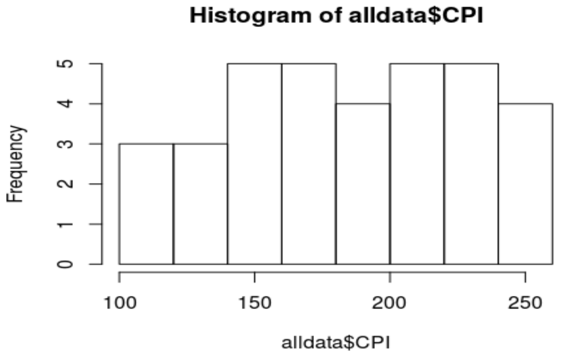
- The histogram shows that both the public4year Tuition fees, CPI and Debt GDP are not normal
- The log transformed public4year Tuition fees is also not very normal.
- Bar plot shows that there is an exponential increase in the Tuition fees with every increase of unit year.
- Over a period of 10 years, the Average tuition fees for a 4-year public university program has increased by 327%

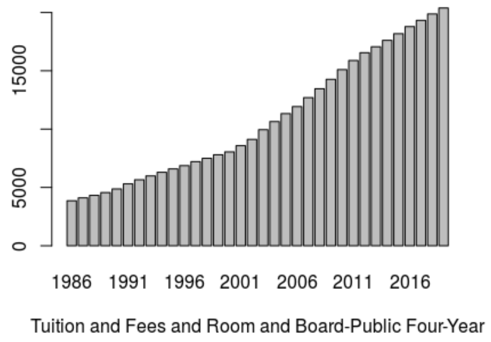
- Mean is 10874 and median is 9535 in current dollars.











```
#1) Read the worksheet number 5, skip 1st 11 rows
data <- readxl::read_xlsx('college_pricing.xlsx',sheet = 5, skip=11)

#2) Remove the last 3 lines from the dataset which are simplyextra texts
data <- head(data,-3)

#3)look for NA's
for(i in 1:dim(data)[2])
{
  print(paste(names(data)[i], length(which(is.na(data[,i])))))
}
```

```
## [1] "In Current Dollars 0"
## [1] "1986 0"
## [1] "1987 0"
## [1] "1988 0"
## [1] "1989 0"
## [1] "1990 0"
## [1] "1991 0"
## [1] "1992 0"
## [1] "1993 0"
## [1] "1994 0"
## [1] "1995 0"
## [1] "1996 0"
## [1] "1997 0"
## [1] "1998 0"
## [1] "1999 0"
## [1] "2000 0"
## [1] "2001 0"
## [1] "2002 0"
## [1] "2003 0"
## [1] "2004 0"
## [1] "2005 0"
## [1] "2006 0"
## [1] "2007 0"
## [1] "2008 0"
## [1] "2009 0"
## [1] "2010 0"
## [1] "2011 0"
## [1] "2012 0"
## [1] "2013 0"
## [1] "2014 0"
## [1] "2015 0"
## [1] "2016 0"
## [1] "2017 0"
## [1] "2018 0"
## [1] "2019 0"
## [1] "10-Year $ Change 0"
## [1] "10-Year % Change 0"

#4) Get each of row's data and remove 1st column as these are row headers
data_row1 <- data[1,-1]
data_row2 <- data[2,-1]
data_row3 <- data[3,-1]
data_row8 <- data[8,-1]
data_row9 <- data[9,-1]

#5) Get the length of the dataset from above step
length(data_row1)

## [1] 36

#6) Remove the last 2 values from the obtained row vectors in step 4 and convert all the value
s into #numeric values
Tuition_Public_Two_Year <- as.numeric(data_row1[c(-35,-36)])
Tuition_Public_Four_Year <- as.numeric(data_row2[c(-35,-36)])
Tuition_Private_Nonprofit_Four_Year <- as.numeric(data_row3[c(-35,-36)])
Tuition_Room_Board_Public_Four_Year <- as.numeric(data_row8[c(-35,-36)])
Tuition_Room_Board_Private_Nonprofit_Four_Year <- as.numeric(data_row9[c(-35,-36)])

#7) All entries from the row vectors in step 4 are converted to numeric
Tuition_Public_Two_Year_all <- as.numeric(data_row1)
```

```
#8) transpose the data df in order to get the row names or Years using rownames
data_transpose <- t(data)

#9) Remove the 1st and last 2 values from rownames(data_transpose) to get only the Years
Year <- rownames(data_transpose)[c(-1,-36)][-35]

#10) Look for values '–' and convert them into 0. Format row 4 and row 7 as numeric 0s
data[data == '–'] <- 0
data_row7 <- data[7,-1]
data_row4 <- data[4,-1]
Room_Board_Public_Two_Year <- as.numeric(data_row4)
Tuition_Room_Board_Public_Two_Year <- as.numeric(data_row7)

#11)Update the 7th row count as sum of  row 1 and row 4
Tuition_Room_Board_Public_Two_Year <- Tuition_Public_Two_Year_all + Room_Board_Public_Two_Year
Tuition_Room_Board_Public_Two_Year <- as.numeric(Tuition_Room_Board_Public_Two_Year[c(-35,-36)
])
data[7,-1] <- Tuition_Public_Two_Year_all + Room_Board_Public_Two_Year

#12a) Plot the barplots for the datasets in step 6  where x axis= Years
barplot(Tuition_Room_Board_Public_Four_Year, names.arg=Year, xlab='Tuition and Fees and Room a
nd Board-Public Four-Year')

# Get 10 Year change of 4-year cost
Tuition_Room_Board_Public_Four_Year[34] - Tuition_Room_Board_Public_Four_Year[1]

## [1] 16520

#Read the inflation dataset
inf <- read.csv("CPI.csv",stringsAsFactors = FALSE)

# Read the Debt to GDP data
GDP <- read.csv("Debt_to_GDP.csv",stringsAsFactors = FALSE)
colnames(GDP)[colnames(GDP) == "Government.Debt.as...of.GDP"] <- "Debt/GDP"

#  Combine data for regression analysis
comdata <- data.frame(Year,Tuition_Room_Board_Public_Two_Year,Tuition_Room_Board_Public_Four_Y
ear,Tuition_Room_Board_Private_Nonprofit_Four_Year)
alldata <- data.frame(comdata,inf$CPI,GDP$"Debt/GDP")
colnames(alldata) <- c("Year", "Public2Cost","Public4Cost","Private4Cost","CPI","Debt_GDP")

# Changing the Year to numeric value
alldata$Year <- as.numeric(as.character(alldata$Year))

# Get the statistics of 4-year cost, CPI and GDP
summary(alldata$Public4Cost)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3860    6372    9535   10874   15685   20380

summary(alldata$CPI)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   109.5   149.4   182.0   185.0   224.4   256.6

summary(alldata$Debt_GDP)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   47.58   58.28   63.59   72.89   96.40  105.46

hist(alldata$Public4Cost)
hist(alldata$CPI)

hist(alldata$Debt_GDP)
```

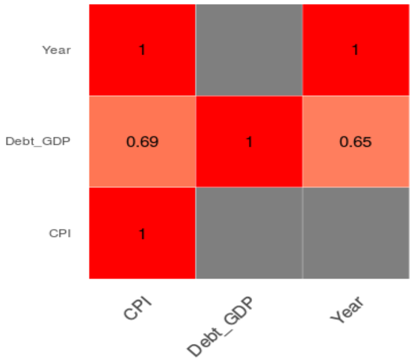
Evaluation of Models based on Training dataset

Split the public4Year tuition fees dataset into training and test datasets based on 70:30 ratio respectively.

```
#19)Training and test datasets
train_data <- alldata[1:24,]
test_data <- alldata[-c(1:24),]

predictor_names <- c("CPI","Debt_GDP","Year")

for(pname in predictor_names)
{
  print(typeof(train_data[,pname]))
}

## [1] "double"
## [1] "double"
## [1] "double"
```

Correlation between predictors

The correlation coefficients between predictor variables – Year, Debt_GDP and CPI tell us that there is a strong correlation between Year and CPI. This might affect the model results. Therefore, removed Year variable from model analysis.



Correlation between public4cost and predictor variables

When checking the correlation between Public4Cost and the predictors, CPI looks highly correlated with the Public4Cost output variable.

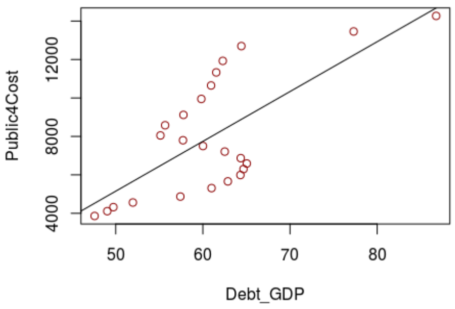
```
cor.test(train_data$Public4Cost,train_data$Debt_GDP, method = "pearson")

##
##  Pearson's product-moment correlation
##
## data:  train_data$Public4Cost and train_data$Debt_GDP
## t = 4.6462, df = 22, p-value = 0.0001245
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4194203 0.8623363
## sample estimates:
##      cor
## 0.703748

cor.test(train_data$Public4Cost,train_data$CPI, method = "pearson")

##
##  Pearson's product-moment correlation
##
## data:  train_data$Public4Cost and train_data$CPI
## t = 26.04, df = 22, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9631379 0.9932361
## sample estimates:
##       cor
## 0.9841616
```

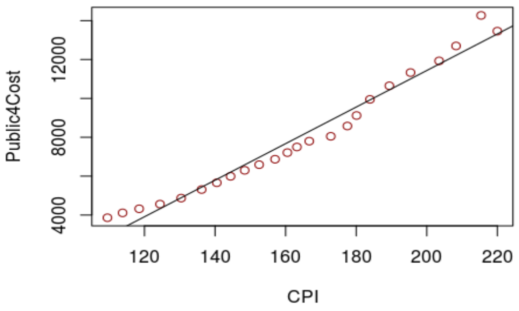
Linear Regression

After removing Year from the model analysis, the model equation becomes.

E(Public4Year) = beta0 + beta1 * CPI + beta2 * Debt_GDP

Null hypothesis test for each of the regression coefficients are statistically significant which implies the Public4Cost is linearly associated with CPI and Debt_GDP independently.





Confounding Effect

The coefficient of Debt_GDP decreases when CPI is added to the model because CPI and Debt_GDP are correlated. The estimate of coefficient of Debt_GDP when CPI predictor variable is not added only depends on the predictor variable Debt_GDP whereas the estimate of coefficient of Debt_GDP when CPI predictor variable is added not only depends on the predictor variable Debt_GDP, but also depends on CPI predictor variable. In effect we have “controlled” CPI (or adjusted for CPI). We say that CPI is “confounding” the association between Public4cost and Debt_GDP. By adjusting for CPI, we remove its confounding effect.

The Multiple R-squared, also called the coefficient of determination is the proportion of the variance in the response data that’s explained by the model. The more variance that is accounted for by the regression model, the data points will fall closer to the fitted regression line and higher the R^2 value and lower the residual SD. The more predictor variables we add, the larger the variability of data that is accounted for by the regression models and higher the R^2 value.

```
summary(lm(Public4Cost ~ CPI, data=train_data))

## 
## Call:
## lm(formula = Public4Cost ~ CPI, data = train_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -828.57 -452.81  -73.04  256.99 1384.31 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7375.685    599.942  -12.29 2.49e-11 ***
## CPI            94.064      3.612   26.04  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 562.4 on 22 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9671 
## F-statistic: 678.1 on 1 and 22 DF,  p-value: < 2.2e-16
```

```
summary(lm(Public4Cost ~ Debt_GDP, data=train_data))

## 
## Call:
## lm(formula = Public4Cost ~ Debt_GDP, data = train_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2870.4 -2051.1  -538.2  1955.5  3811.4 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -7811.7     3425.2  -2.281 0.032611 *  
## Debt_GDP       259.3       55.8   4.646 0.000125 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2254 on 22 degrees of freedom
## Multiple R-squared:  0.4953, Adjusted R-squared:  0.4723 
## F-statistic: 21.59 on 1 and 22 DF,  p-value: 0.0001245
```

```
summary(lm(Public4Cost ~ CPI + Debt_GDP, data=train_data))

## 
## Call:
## lm(formula = Public4Cost ~ CPI + Debt_GDP, data = train_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -689.5 -440.7 -141.9  327.0 1074.3 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7968.704    855.736  -9.312 6.65e-09 ***
## CPI            90.729      4.983  18.207 2.43e-14 ***
## Debt_GDP       18.688     19.207   0.973    0.342    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 563 on 21 degrees of freedom
## Multiple R-squared:  0.9699, Adjusted R-squared:  0.9671
## F-statistic: 338.7 on 2 and 21 DF,  p-value: < 2.2e-16
```

Interactions effect between predictors

The p-value for the interaction between Debt_GDP and CPI is statistically significant, therefore included the interaction effect in the model analysis for accuracy and better interpretation. As interaction term is significant, included the interaction term and reran the model.

```
summary(lm(Public4Cost ~ Debt_GDP * CPI, data=train_data))

##
## Call:
## lm(formula = Public4Cost ~ Debt_GDP * CPI, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.22 -229.41   57.76  162.09  522.71
3##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5622.6850  2315.1048   2.429   0.0247 *
## Debt_GDP     -218.9689    41.1441  -5.322 3.29e-05 ***
## CPI            17.4671    12.5321   1.394   0.1787
## Debt_GDP:CPI    1.2686     0.2105   6.026 6.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 343.8 on 20 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9877
## F-statistic: 617.6 on 3 and 20 DF,  p-value: < 2.2e-16
```

The multiple regression model equation now becomes

E(Public4Year) = beta0 + beta1 * CPI + beta2 * Debt_GDP + beta3 * CPI * Debt_GDP

```
# E(Cost) = alpha + beta1 * CPI + beta2 * Debt_GDP + beta3 * CPI * Debt_GDP
mr01 = lm(Public4Cost ~ CPI + Debt_GDP + Debt_GDP * CPI, data=train_data)
summary(mr01)

##
## Call:
## lm(formula = Public4Cost ~ CPI + Debt_GDP + Debt_GDP * CPI, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.22 -229.41   57.76  162.09  522.71
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5622.6850  2315.1048   2.429   0.0247 *
## CPI            17.4671    12.5321   1.394   0.1787
## Debt_GDP     -218.9689    41.1441  -5.322 3.29e-05 ***
## CPI:Debt_GDP    1.2686     0.2105   6.026 6.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 343.8 on 20 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9877
## F-statistic: 617.6 on 3 and 20 DF,  p-value: < 2.2e-16
```

The effect of one variable that forms the interaction depends on the level of the other variable in the interaction.

- The effect of CPI on Costs depends on the value of Debt_GDP = 17.4671 + 1.2686*Debt_GDP.

  The coefficient estimate of CPI is the average difference in the mean Public4cost per unit difference in CPI when Debt_GDP is 0.

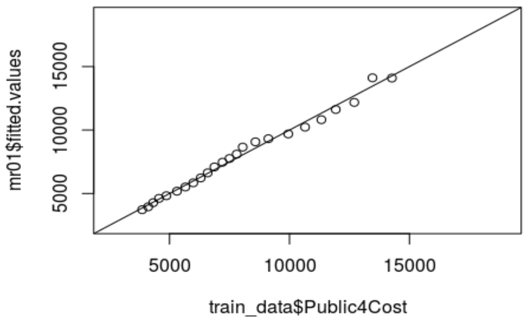
- The effect of Debt_GDP on Costs depends on the value of CPI and is -218.9689 + 1.2686*CPI

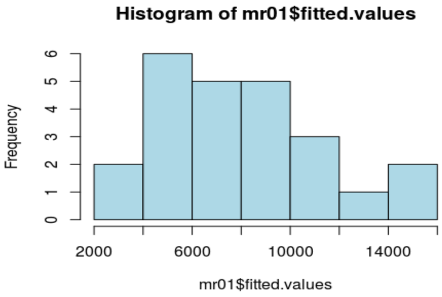
  The coefficient estimate of CPI is the average difference in the mean Public4cost per unit difference in CPI when Debt_GDP is 0.

Checking the assumptions for Linear regression:

- Normality of error distribution or large sample size
- Constant variance of errors
- Linearity
- Independence

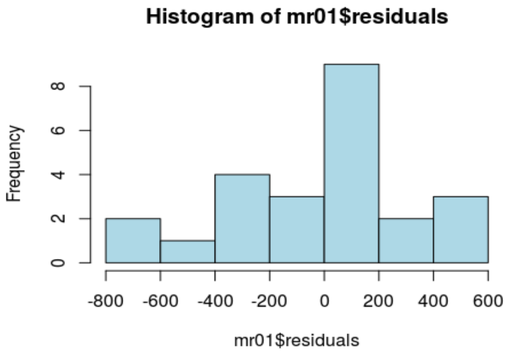
```r
## check out fitted values against the originals
#mr01
plot(train_data$Public4Cost, mr01$fitted.values, xlim = c(2500,19000), ylim = c(2500,19000))
abline(a=0, b=1)

hist(mr01$fitted.values,col="lightblue")
```

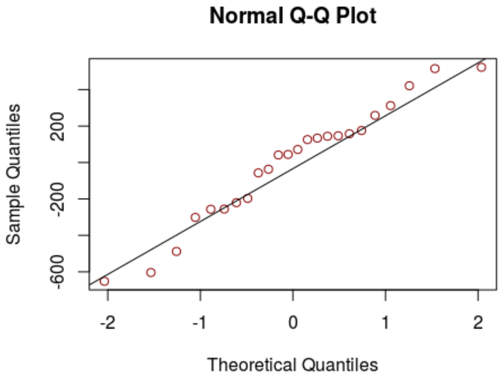


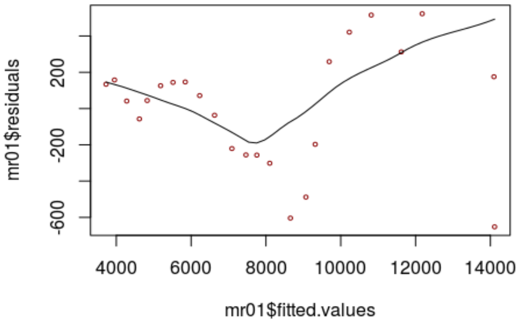


```r
#Check for normality using residuals - Data looks skewed
hist(mr01$residuals,col="lightblue")

qqnorm(mr01$residuals,col="darkred")
qqline(mr01$residuals)
```

```
# Check for constant-variance
scatter.smooth(mr01$fitted.values, mr01$residuals,cex=0.5,col="darkred")
```



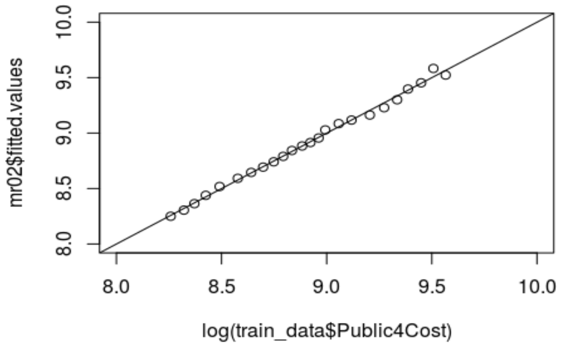


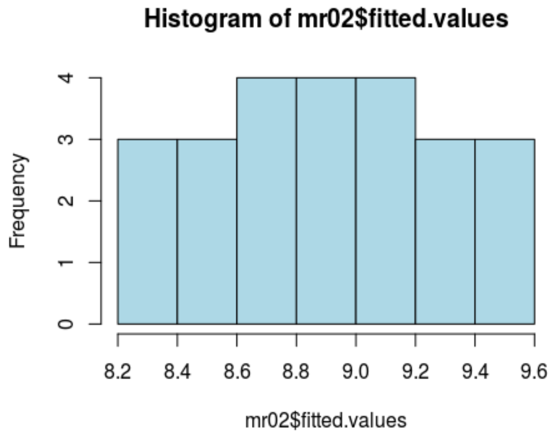


```
## MSE and RMSE
mean((mr01$fitted.values - train_data$Public4Cost)^2)

## [1] 98510.82

sqrt(mean((mr01$fitted.values - train_data$Public4Cost)^2))

## [1] 313.8643
```

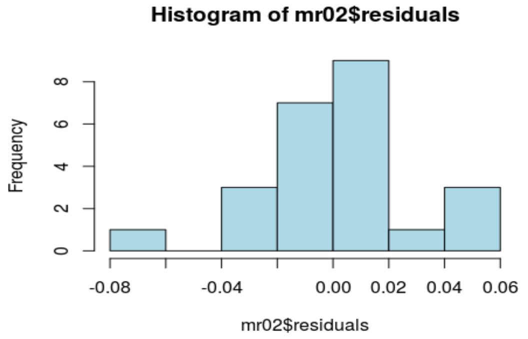
Linear regression did not meet the normality and constant variance assumptions. Also, the sample size is very small.

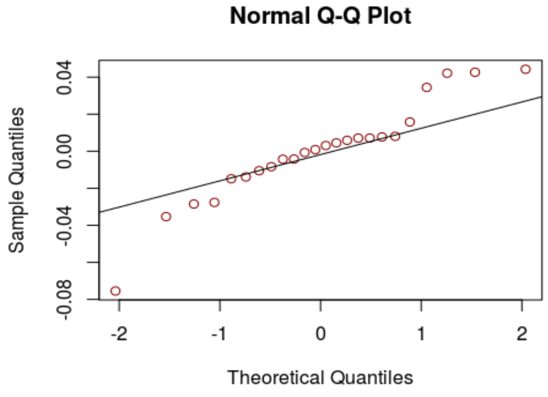
```
#Using log-tranformation of the response variable as residuals are not normal
mr02 <- (lm(log(Public4Cost) ~ CPI+Debt_GDP+CPI*Debt_GDP, data=train_data))

## check out fitted values against the originals
#mr02
plot(log(train_data$Public4Cost), mr02$fitted.values, xlim = c(8,10), ylim = c(8,10))
abline(a=0, b=1)

# normality of errors
hist(mr02$fitted.values,col="lightblue")
```

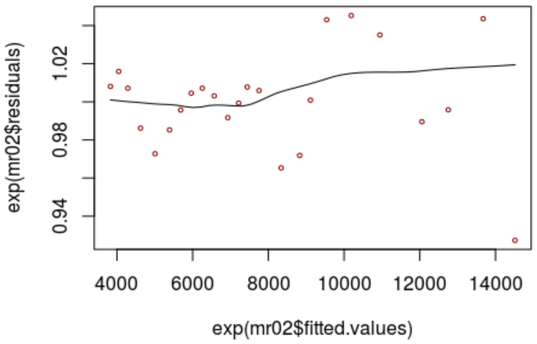




```
#Check for normality using residuals - Data looks slightly normal
hist(mr02$residuals,col="lightblue")

qqnorm(mr02$residuals,col="darkred")
qqline(mr02$residuals)

# Check for constant-variance
scatter.smooth(exp(mr02$fitted.values), exp(mr02$residuals),cex=0.5,col="darkred")
```







```
## MSE and RMSE
mean((exp(mr02$fitted.values) - train_data$Public4Cost)^2)

## [1] 92019.36

sqrt(mean((exp(mr02$fitted.values) - train_data$Public4Cost)^2))

## [1] 303.3469
```

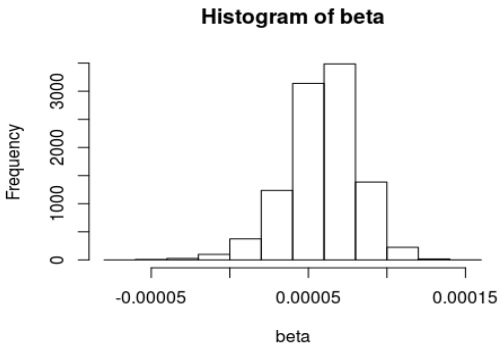
Log transformed Linear regression also did not meet the normality and constant variance assumptions. Also, the sample size is very small. So, let's try GLM , RF models.

GLM MODEL

- Non-negative integer-type response variable (Price).
- As variance is proportional to the mean (hence not constant), so residuals are expected to display non-constant variance, which suits the data.
- There is no normality assumption for log-linear/Poisson regression

Considering the above factors, Poisson regression model works well for this dataset, but let's explore more.

```
##Permutation Test for small sample before GLM
glm01 <- glm(Public4Cost ~ CPI + Debt_GDP + Debt_GDP * CPI,family = poisson,data =train_data)

set.seed(0)
beta=rep(NA,10000)
for(i in 1:10000){
  x1=sample(train_data$CPI,size=length(train_data$Public4Cost),replace=FALSE)
  x = x1+train_data$Debt_GDP+x1 * train_data$Debt_GDP
  beta[i]=glm(Public4Cost ~ x, data=data.frame(Public4Cost=train_data$Public4Cost,x), family=p
oisson)$coef[2]
}

mean(abs(beta) > abs(glm01$coef[2]))

## [1] 0

hist(beta)
abline(v=glm01$coef[2],lty=2,col=2,lwd=2)
```



As sample size is very small (n =24 for training set), let’s do the permutation test. Permutation test p-value is statistically significant, so the sample size of 24 is good enough for doing GLM

Also, accounted for dispersion as calculation showed a value substantially greater than the default value of 1 when interpreting the regression coefficients. From the interpretation of coefficients for CPI and Debt_GDP, it is found that CPI is the most influential variable on costs.

```
D=summary(glm01)$deviance
df=summary(glm01)$df.residual
data.frame(D,df,dispersion=D/df)

##          D df dispersion
## 1 181.5282 20   9.076408

glmsum = summary(glm01,dispersion=D/df)
glmsum$coef

##                   Estimate   Std. Error   z value      Pr(>|z|)
## (Intercept)   6.612379e+00 2.798489e-01 23.628390 1.969005e-123
## CPI           1.375762e-02 1.401228e-03  9.818261  9.395011e-23
## Debt_GDP      6.018621e-03 4.859138e-03  1.238619  2.154866e-01
## CPI:Debt_GDP -3.085011e-05 2.347741e-05 -1.314034  1.888347e-01

##     Null deviance: 27181.71  on 23  degrees of freedom
## Residual deviance:   181.53  on 20  degrees of freedom
## AIC: 447.46
##
## Number of Fisher Scoring iterations: 3
```

- The effect of CPI on Costs depends on the value of Debt_GDP and is equal to exp(1.376e-02 + -3.085e-05*Debt_GDP)

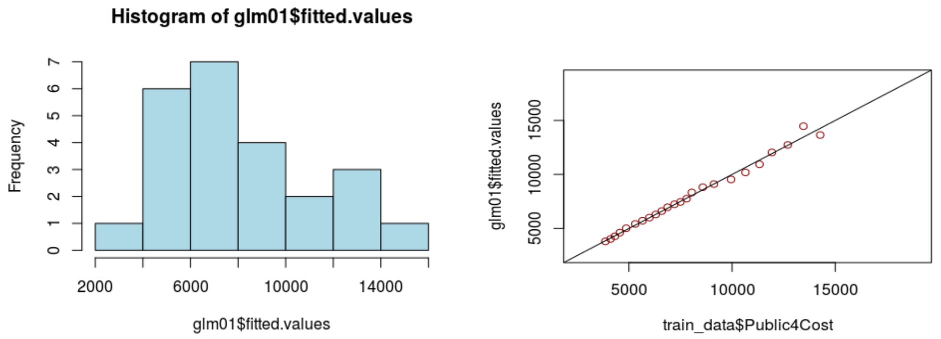
  The coefficient estimate of CPI is the average difference in the log of mean Public4cost per unit difference in CPI when Debt_GDP is 0 (or) mean costs increases by 14% for each increase of 10 unit CPI when Debt_GDP is 0

- The effect of Debt_GDP on Costs depends on the value of CPI and is equal to exp(6.018621e-03 + -3.085e-05*CPI)

  The coefficient estimate of Debt_GDP is the average difference in the mean Public4cost per unit difference in Debt_GDP when CPI is 0 (or) mean costs increases by 6.2% for each increase of 10 unit Debt_GDP when CPI is 0

```
hist(glm01$fitted.values ,col="lightblue")
```

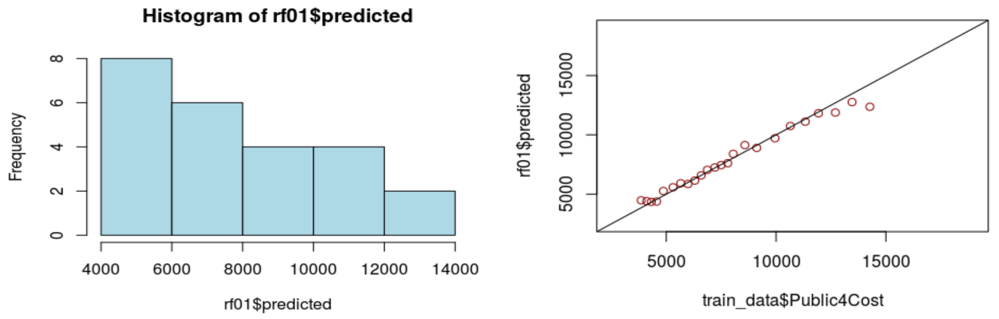
```
## check out fitted values against the originals
plot(train_data$Public4Cost, glm01$fitted.values, xlim = c(2500,19000), ylim =
c(2500,19000),col="darkred")
abline(a=0, b=1)
```



```
## MSE and RMSE
mean((glm01$fitted.values - train_data$Public4Cost)^2)

## [1] 89654.2

sqrt(mean((glm01$fitted.values - train_data$Public4Cost)^2))

## [1] 299.4231
```

Random Forest

```
# first, let's run a RF using a set of "standard" inputs.  Just to get a baseline
## set your seed so your model is reproducible
rf01 <- randomForest(Public4Cost ~ CPI + Debt_GDP + Debt_GDP * CPI,
                     data = train_data,
                     mtry = 2, ## roughly the standard pick for regression models
                     nodesize = 5, ## default for regression
                     maxnodes = NULL, ## again, default
                     ntree = 5000) ## 5 at first, to make sure it runs OK.  Then enlarge

## review the model summary
hist(rf01$predicted,col="lightblue")

## check out fitted values against the originals
plot(train_data$Public4Cost, rf01$predicted, xlim = c(2500,19000), ylim = c(2500,19000),col="d
arkred")
abline(a=0, b=1)
```



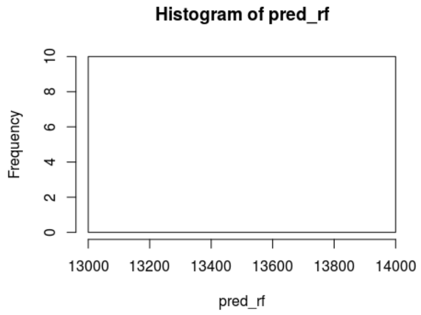
```
## MSE and RMSE
mean((rf01$predicted - train_data$Public4Cost)^2)

## [1] 242561.4
```

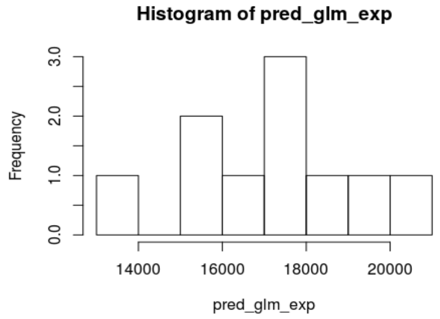
```r
sqrt(mean((rf01$predicted - train_data$Public4Cost)^2))

## [1] 492.5052
```

Evaluation of models - Predict Test Data

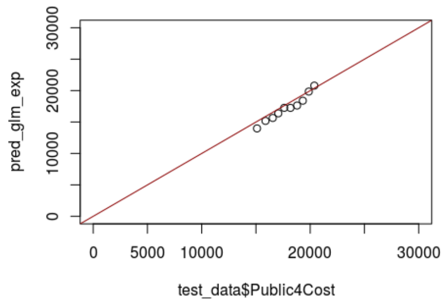
At this point, we are excluding Linear regression and Log transformed Linear regression from further analysis.

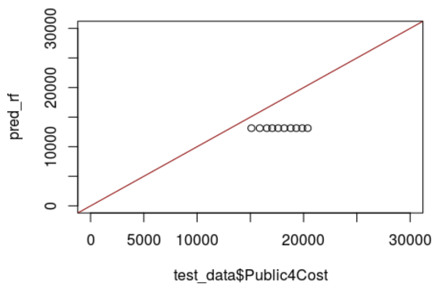
```r
## rf01 and glm01
# Infer predictors - Infer how Debt_GDP and CPI predicts college 4 Year degree fees
pred_glm <- predict(glm01, newdata=test_data)
pred_rf <- predict(rf01, newdata=test_data)
pred_glm_exp <- exp(pred_glm)

hist(pred_rf)

hist(pred_glm_exp)
```





```r
## RMSE
sqrt(mean(pred_glm_exp - test_data$Public4Cost)^2)

## [1] 636.6545

sqrt(mean(pred_rf - test_data$Public4Cost)^2)

## [1] 4740.12

plot(test_data$Public4Cost, pred_glm_exp,xlim = c(0, 30000), ylim = c(0, 30000))
abline(a=0,b=1,col="darkred")

plot(test_data$Public4Cost, pred_rf,xlim = c(0, 30000), ylim = c(0, 30000))
abline(a=0, b=1,col="darkred")
```



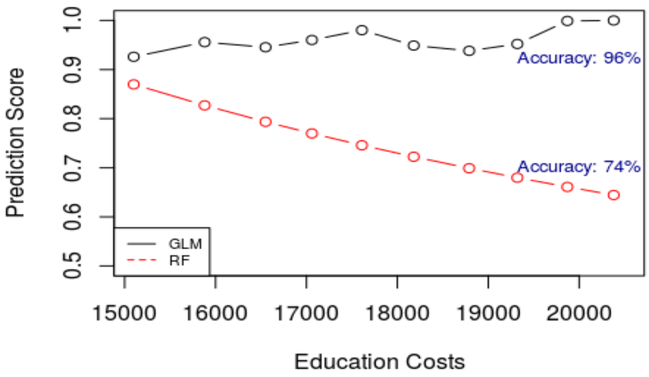


Accuracy of Models

RMSE is higher in Random Forest for training and prediction of test data, therefore GLM wins.

RMSE\Models	GLM (Poisson)	Random Forest
Training Dataset	299.4231	492.5052
Predict Test Dataset	636.6545	4740.12

```r
# Model Accuracy
probs_glm <- pmin((pred_glm_exp / test_data$Public4Cost), 1)
glm_accuracy = mean(probs_glm) * 100
probs_rf <- pmin((pred_rf / test_data$Public4Cost), 1)
rf_accuracy = mean(probs_rf) * 100

par(xpd=TRUE)
plot(test_data$Public4Cost, probs_glm , type = "b", xlab = "Education Costs", ylab =
"Prediction Score",xlim = c(15100, 20500),ylim=c(0.5,1))
lines(test_data$Public4Cost, probs_rf , type = "b", col = "red",xlim = c(15100, 20500))
legend("bottomleft",c("GLM","RF"),lty=1:2,col=1:2,cex=0.65)
text(20000,0.92,"Accuracy: 96%",col="darkblue",cex=0.8)
text(20000,0.7,"Accuracy: 74%",col="darkblue",cex=0.8)
```



Utility Function

Let consider this scenario where based on the actual tuition fees, bounds are calculated for utility calculation. Average Tuition Costs for 2019 = 22000 based on original data. Let the threshold be +-5% for model prediction. Then lower bound = 20900 and upper bound = 23100. Base on these bounds, consider the following scenarios.

- If the education saving is greater than 20900 and less than 23100, we don’t lose anything, and we are safe.

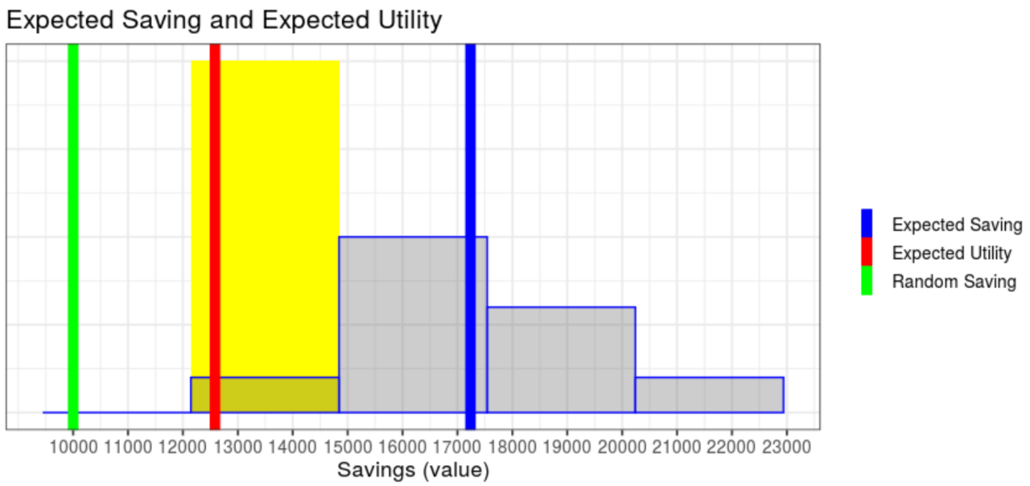
- If we save less than 20900, we end up borrowing more in the future as actual value is higher than this amount. Let the interest rate levied be 10% on the borrowed amount (lose 10%).

- If we are saving more than the actual amount, and we want to withdraw the extra savings for family trips, let’s assume a 10% penalty is imposed as withdrawal penalty and we lose 10% value (example, 529 plan).

- In 2nd and 3rd scenario where savings is not within the threshold range, we lose 10% of the value.

The below piece of code calculates the predicted tuition costs and the utilities based on the above calculations for GLM model. If the random saving is 10000 dollars, model's predicted saving is 17237 dollars. Based on the Utility function calculations, the Expected Utility is 12584 dollars which is better than randomly guessed saving of 10000.

```
# Plot it
ggplot(D_plot_glm) +
  # Histograms of posterior predictive
  geom_histogram(aes(x=utilities_glm), bins=5, fill='yellow') +
  geom_histogram(aes(x=values_glm), bins=5, col='blue', alpha=0.3) +

  # Lines for utility
  geom_vline(aes(xintercept=mean(utilities_glm), color='Expected Utility'), lwd=3, show.legend
=TRUE) +
  geom_vline(aes(xintercept=mean(values_glm), color='Expected Value'), lwd=3) +
  geom_vline(aes(xintercept=16000, color='Random Saving'), lwd=3) +

  # Styling
  labs(x = 'Savings (value)', y='', title='Expected value and Expected Utility') +
  scale_x_continuous(breaks=seq(10000, 25000, by=1000)) +
  theme_bw(13) +
  theme(axis.ticks.y = element_blank(), axis.title.y = element_blank(), axis.text.y = element_
blank()) +

 # Legend
  scale_color_manual(name = "", values = c('Expected Value'='blue', 'Expected Utility' = 'red'
, 'Random Saving'='green'))
```

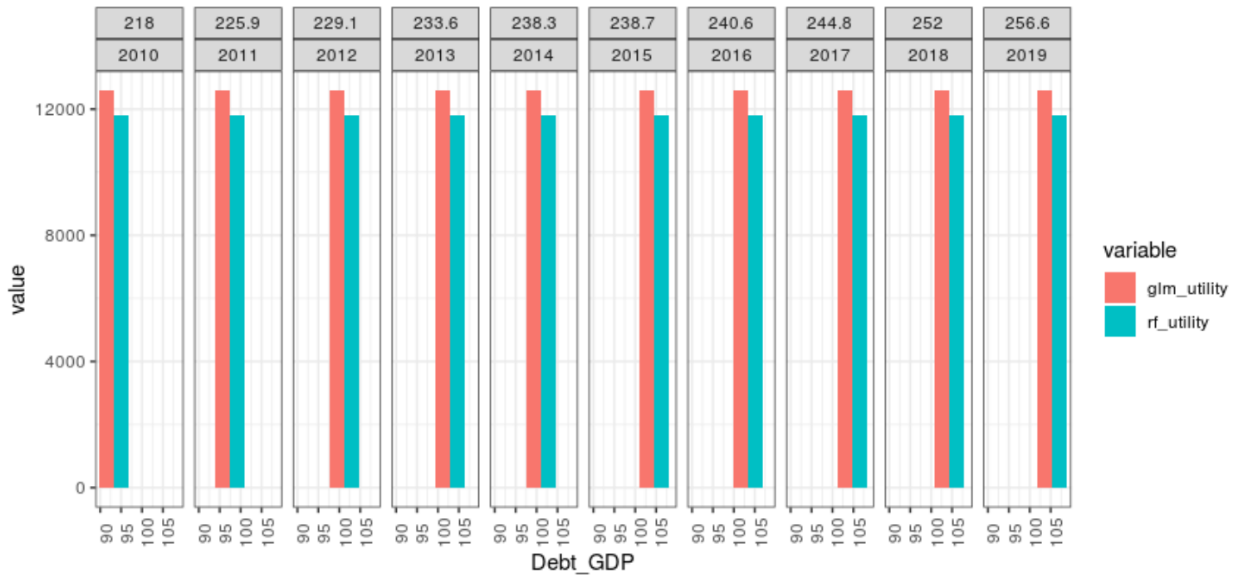


Utility Function – RF vs GLM

GLM outperforms Random Forest as seen from the graph.

```
## need to predict the tuition costs given 'Debt_GDP', 'Year' and Inflation index 'CPI' data.
Predict tuition costs using test data and apply a utility function to it.

  values_glm = exp(predict(glm01, newdata=test_data))
  utilities_glm = ifelse(values_glm > 20900 && values_glm < 22000, values_glm * 1.0, values_gl
m * 0.9)
```

```
  values_rf = predict(rf01, newdata=test_data)
  utilities_rf = ifelse(values_glm > 20900 && values_glm < 22000, values_rf * 1.0, values_rf *
0.9)

  # Put it in a data.frame
D_plot_glm = data.frame(values_glm, utilities_glm)
D_plot_rf = data.frame(values_rf, utilities_rf)

tdata = data.frame(CPI=test_data$CPI,Debt_GDP=test_data$Debt_GDP,Year=test_data$Year)
df1 <- tdata %>% rowwise %>% do(W = as_tibble(.)) %>% ungroup

predicted_utility_glm = function(CPI,Debt_GDP,Year) {
  # Posterior predictive samples (values)
  values_glm = exp(predict(glm01, newdata=data.frame(CPI,Debt_GDP,Year), summary=FALSE))

  # Associated utilities. Posterior predictive utilities
  utilities_glm = ifelse(values_glm > 20900 && values_glm < 22000, values_glm * 1.0, values_gl
m * 0.9)
  utilities_glm  # Return it
}

predicted_utility_rf = function(CPI,Debt_GDP,Year) {
  # Posterior predictive samples (values)
  values_rf = predict(rf01, newdata=data.frame(CPI,Debt_GDP,Year), summary=FALSE)

  # Associated utilities. Posterior predictive utilities?
  utilities_rf = ifelse(values_rf  > 20900 && values_rf < 22000, values_rf * 1.0, values_rf *
0.9)
  utilities_rf  # Return it
}

D_plot_rf = tdata %>%
  mutate(rf_utility = predicted_utility_rf(tdata$CPI,tdata$Debt_GDP,tdata$Year)) %>% unnest

## Warning: `cols` is now required.
## Please use `cols = c()`

D_plot_glm = tdata %>%
  mutate(glm_utility = predicted_utility_glm(tdata$CPI,tdata$Debt_GDP,tdata$Year)) %>% unnest

## Warning: `cols` is now required.
## Please use `cols = c()`

#D_plot_all_one <- merge(D_plot_glm, D_plot_rf, by=c("CPI","Debt_GDP","Year"))
D_plot_all <- merge(D_plot_glm, D_plot_rf, by=c("CPI","Debt_GDP","Year"))
D_plot_l <- melt(D_plot_all, id.vars = c("CPI","Debt_GDP","Year"))
p <- ggplot(data = D_plot_l, aes(x = Debt_GDP, y = value, group = variable, fill = variable))
p <- p + geom_bar(stat = "identity", width = 7, position = "dodge")
p <- p + facet_grid(. ~ CPI+Year)
p <- p + theme_bw()
p <- p + theme(axis.text.x = element_text(angle = 90))
p

save.image("FinalModelEval.RData")

load("FinalModelEval.RData")
```



Conclusion:

- We are able to predict the 4-year public University costs using GLM model with 96% accuracy
- Consumer Price Index is the most influential variable on costs.

Recommendations/Next Steps:

- Evaluate the models with more predictors such as difference between the mean earnings with and without college degree, interest rates etc. to get some interesting insights.
- Collect more data (monthly wise) rather than yearly and verify if the prediction results change.
- Time series analysis to understand the trends and seasonality of inflation/deflation and it’s subsequent impact on Education costs.