**INFSCI 2725 Data Analytics**

**Assignment 2 - Statistics**

Date: 10/06/2015

Student name: Ting LI

Tutor name: Marek Druzdzel

# Executive Summary

Lower average retention rate may pose threat on the long-term development of a university. This report aims to explore factors the relationship between the average retention rate and another seven predictors including spend (average spending per student), top10(percentage of incoming freshmen who were among the top 10% students in their high schools), rejr(school's rejection rate), tstsc(average test scores of incoming freshmen), pacc(percent of admitted applicants who accept university's offer), strat(student-teacher ratio), and salar(average faculty salary). A dataset called retention with 170 data is analyzed using R, and the R code is placed in the appendix at the end. Initially, a statistical and graphic summary of this dataset are demonstrated, followed by a linear regression performed on different subset of predictors. The linear regression model suggests that tstsc, pacc and strat are the three most important variables in influencing average retention rate.

# 1. An overview of retention dataset

## A snapshot of retention data

Table 1 is a snapshot of the dataset we are going to analyze.

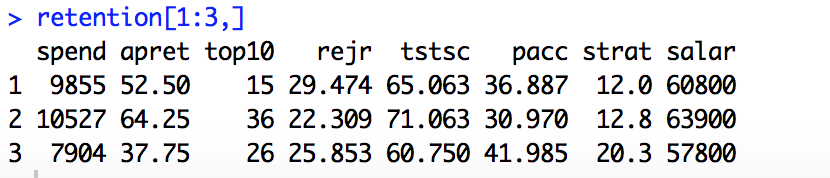


Table 1. Snapshot of retention data

## 1.2. Statistical Summary table

Table 2 summarizes some basic statistics for all predictors.

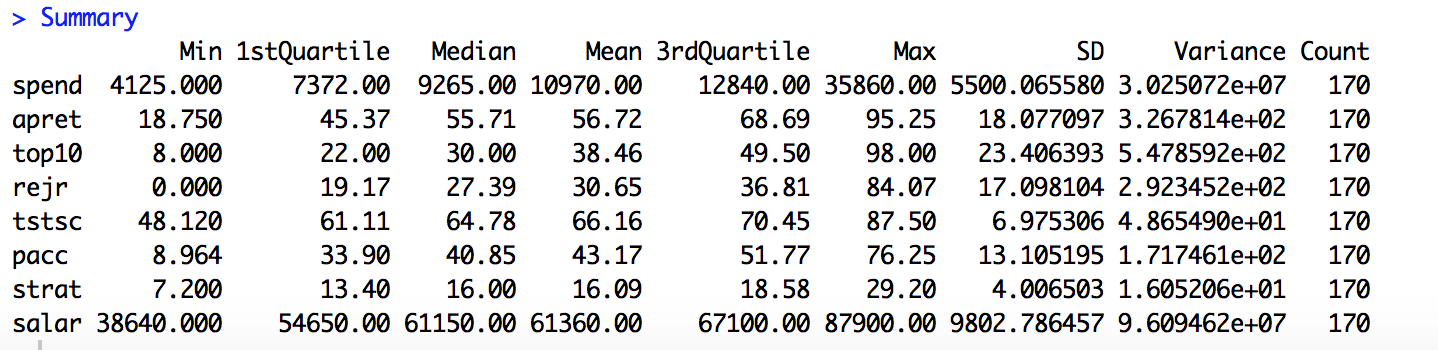


Table 2. Statistical Summary

## 1.3. Graphic Summary:

Figure 1, Figure 2 and Figure 3 are the histograms for variable apret, tstsc and salar respectively.

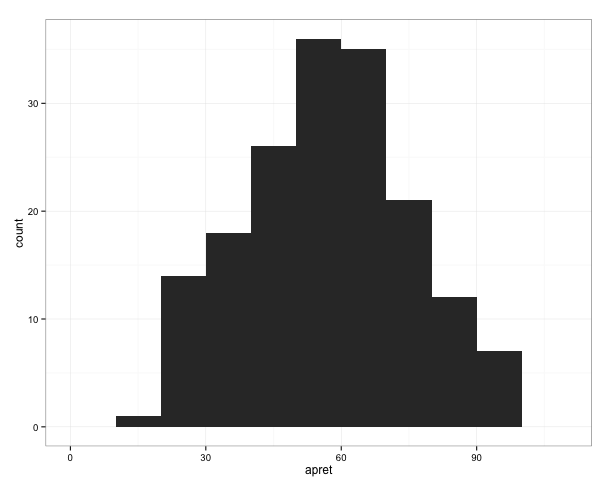


Figure 1. Histogram for apret

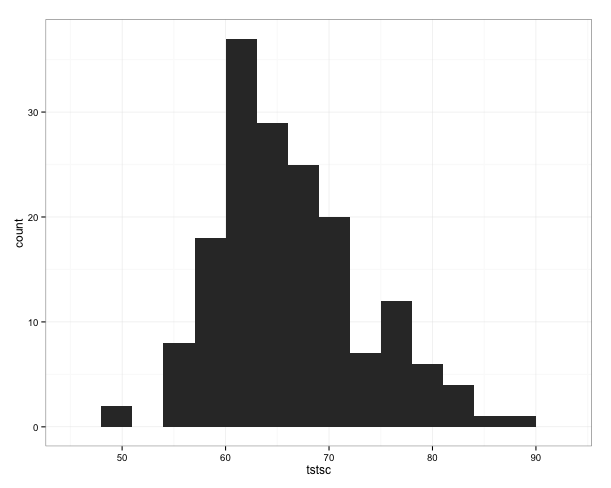


Figure 2. Histogram for tstsc

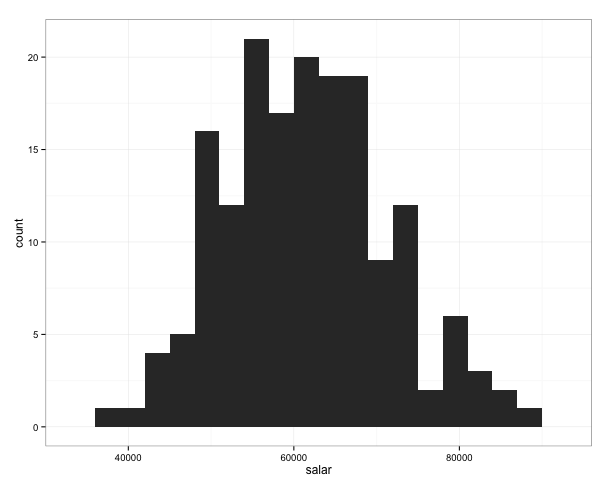
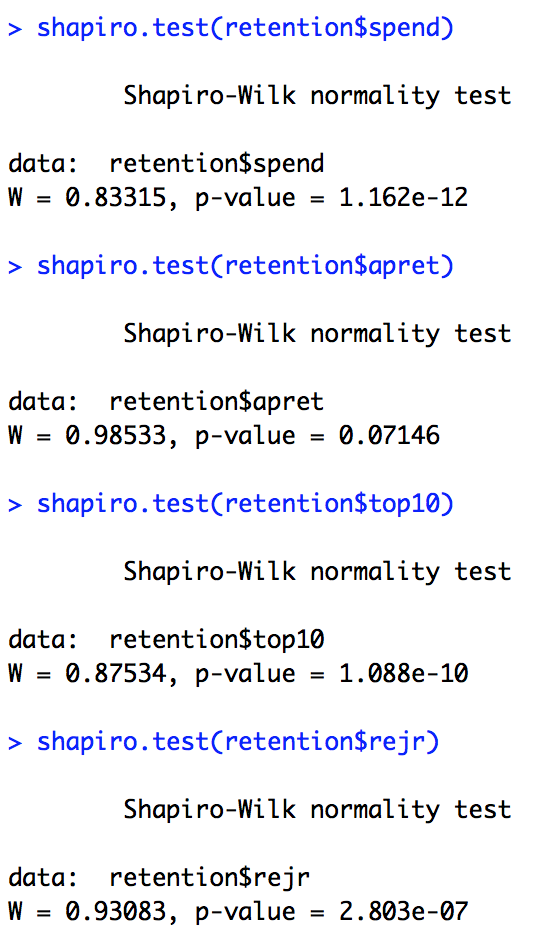
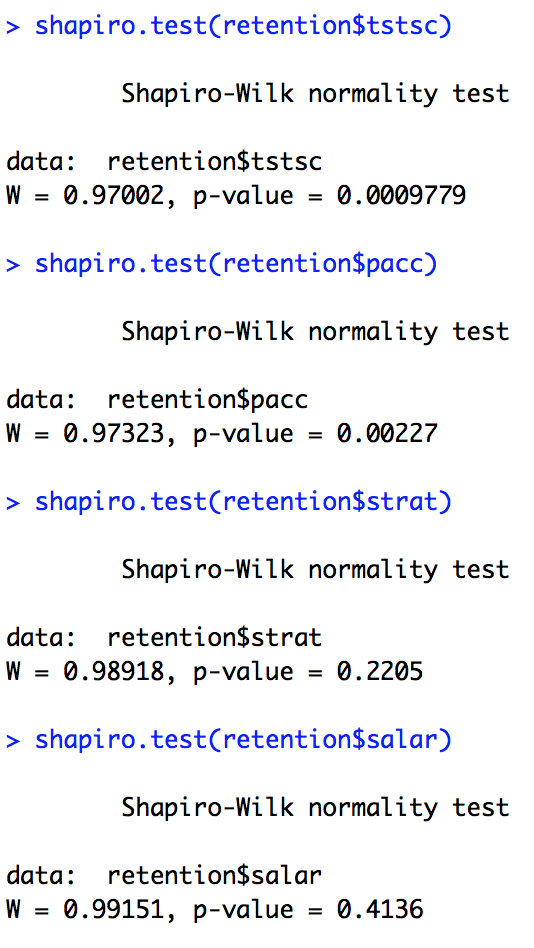


Figure 2. Histogram for salar

## 1.4. Normality Test

shapiro.test in R is used to test the normality of variables. Based on the significance level 0.05, we compare the p-value derived from shapiro.test with 0.05 for each variable and conclude that only apret, strat and salar follow normal distribution.

# 2.Linear Regression

## 2.1. Linear Regression of apret on tstsc

Table 3 is the summary for linear regression between apret and tstsc. There's an expected increase of 2.0271 average retention rate for every 1 score increase in average test scores of incoming freshmen, meaning enrolling students with higher score in the very beginning could increase retention rate. The p-value is a probability that the coefficient is not significant. Big is bad because it indicates a high likelihood of insignificance. The coefficient for apret is significantly different from zero suggesting that tstsc and apret are linearly related. Moreover, as can be seen from Figure 4, the relationship between tstsc and apret is basically linear. The R-squared (0.6118) indicates that the model accounts for 61.18 percent of the variance in average retention rate.

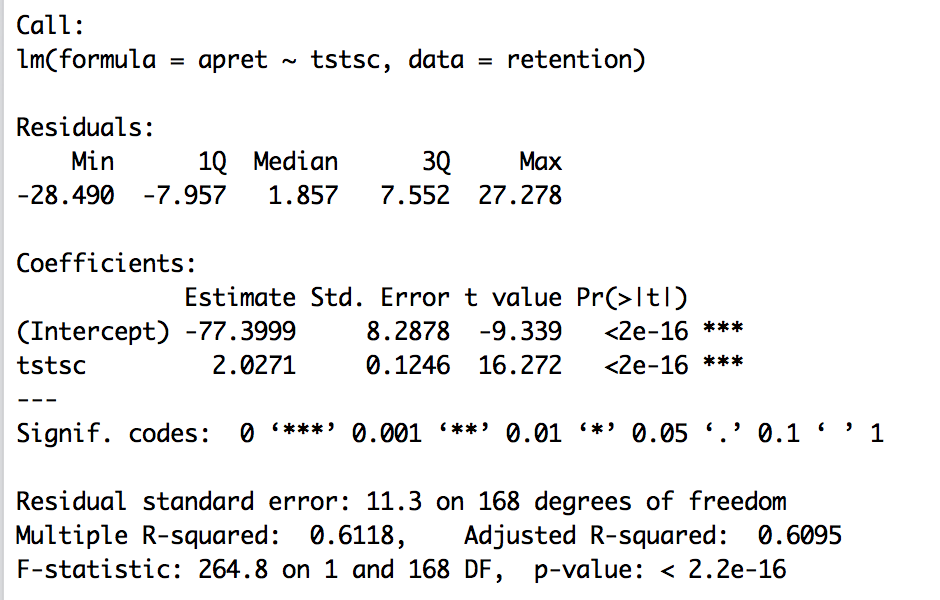


Table 3. Linear Regression of apret on tstsc

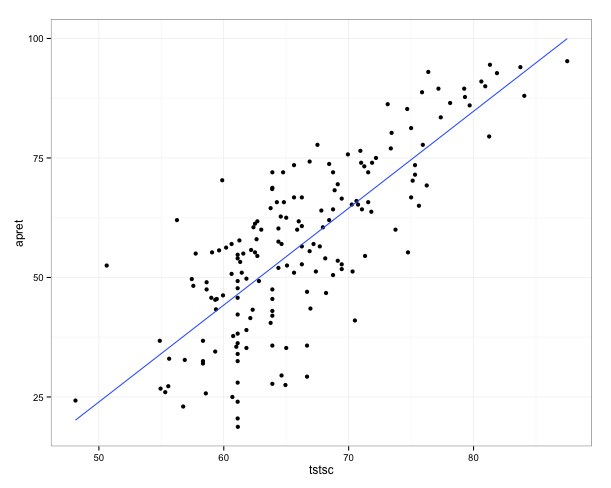


Figure 4. Relationship between tstsc and apret

## 2.2. Linear Regression of apret on salar

Table 4 is the summary for linear regression between apret and salar. There's an expected increase of 0.058 average retention rate for every 1-dollar increase in average faculty salary, meaning offering faculty higher salary could possibility enhance their teaching quality which can in turn increase retention rate. The coefficient for salar is significantly different from zero suggesting that salar and apret are linearly related. As can be seen from Figure 5, the relationship between salar and apret is basically linear. The R-squared (0.4043) indicates that the model accounts for 40.43 percent of the variance in average retention rate.

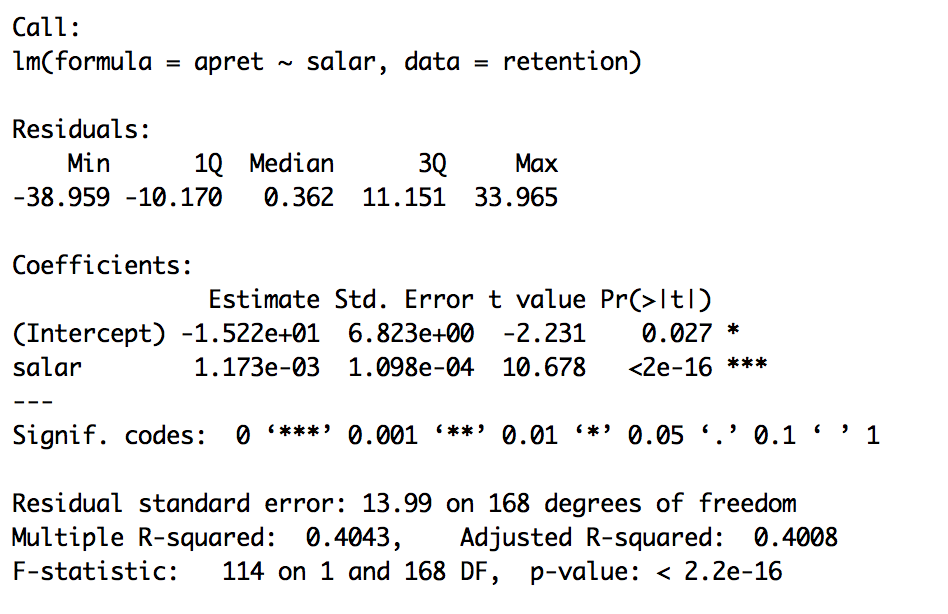


Table 4. Linear Regression of apret on salar

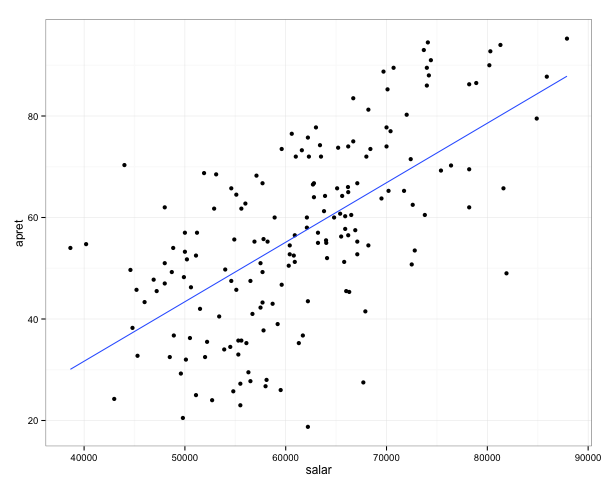


Figure 5. Relationship between salar and apret

2.3. Linear Regression of apret on tstsc and salar

Table 5 is the summary for linear regression of apret on tstsc and salar. While linear regression is performed separately on tstsc and salar, both variables have significant influences on response variable apret as elaborated in the above two sections. However, the coefficient for salar (p=0.0228) is not significantly different from zero suggesting that salar and apret are not linearly related. Compared with the R-squared (61.18) for linear regression performed only on tstsc, R-squared (0.6237) for this model did not make much difference. It is imprudent to say that salar is not linearly related with apret.

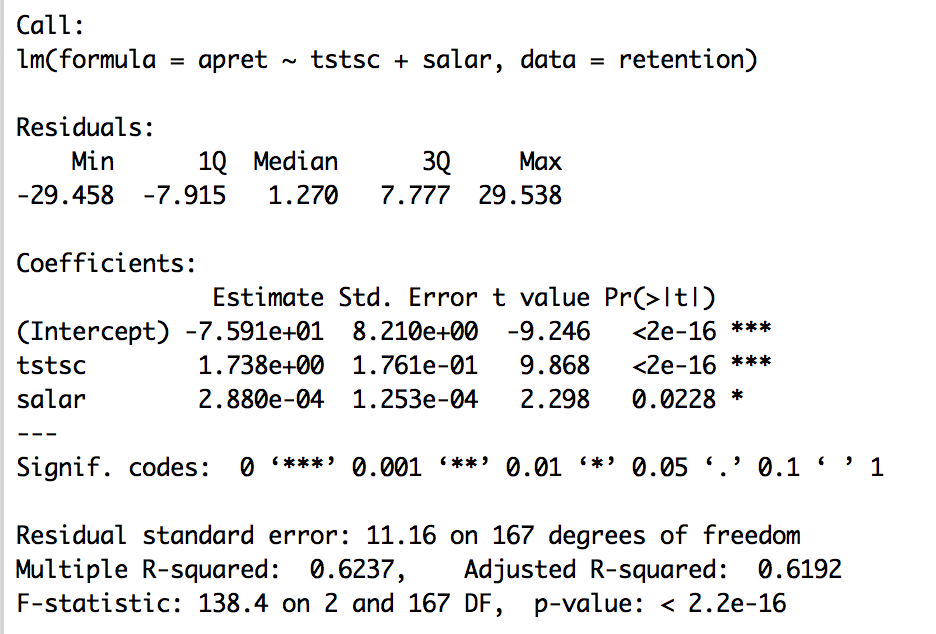


Table 5. Linear Regression of apret on salar and tstsc

2.4. Linear Regression of apret on all predictors

Table 6 is the summary for linear regression of apret on all other predictors. R-squared is a measure of the model's quality – the fraction of the variance of y that is explained by the regression model and bigger is better. However, taken together, the predictor variables account for only 65.71 percent of the variance in average retention rate, signifying that this linear model is not great. In fact, referring to the second column of Figure 6 which pictures the relationship between apret with all the other predictors, we can notice that the relationship between apret with most other predictors are not linear. With regard to the summary table, it seems that tstsc (average test scores of incoming freshmen) is the one who is mostly linearly related with average retention rate, followed by pacc (percent of admitted applicants who accept university’s offer) and strat (student-teacher ratio). Additionally, we can observe that a backward stepwise variable selection also left us with the linear regression model with only three predictors aforementioned, tstsc”, ”pacc” and “strat”. Only tstsc has a positive impact on apret, both pacc and strat have negative impact on apret. This phenomenon indicates that enroll better students initially, and reduce the admission rate and student-teacher ratio can contribute to increase the average retention rate.

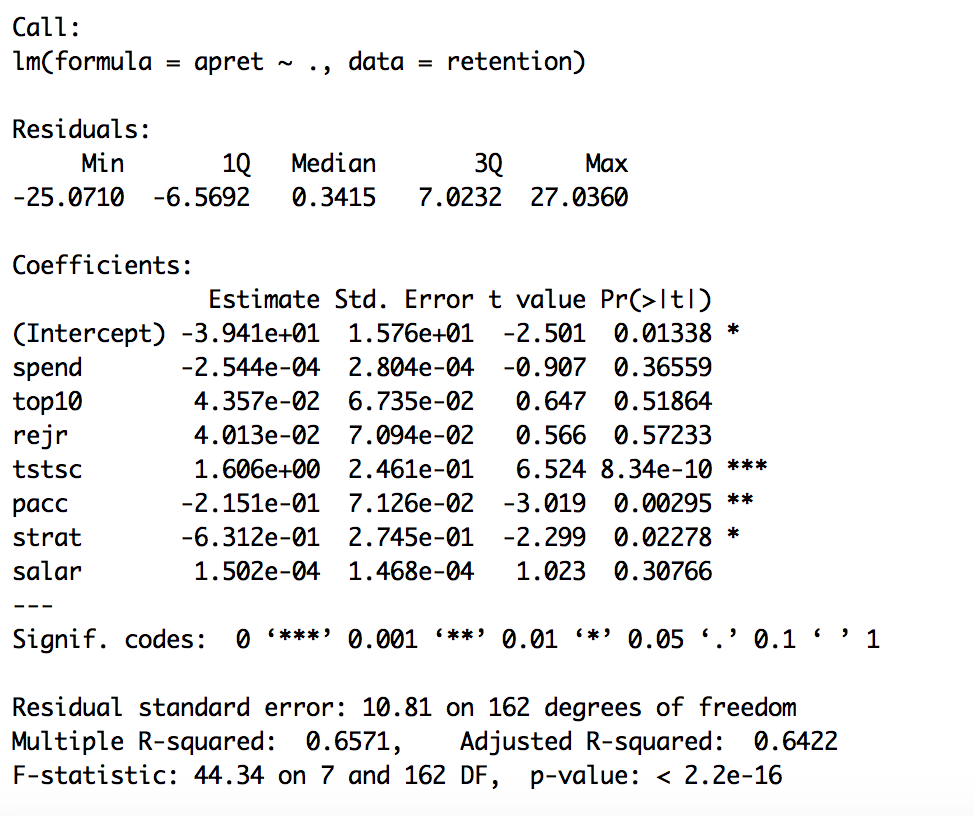


Table 6. Linear Regression of apret on all predictors

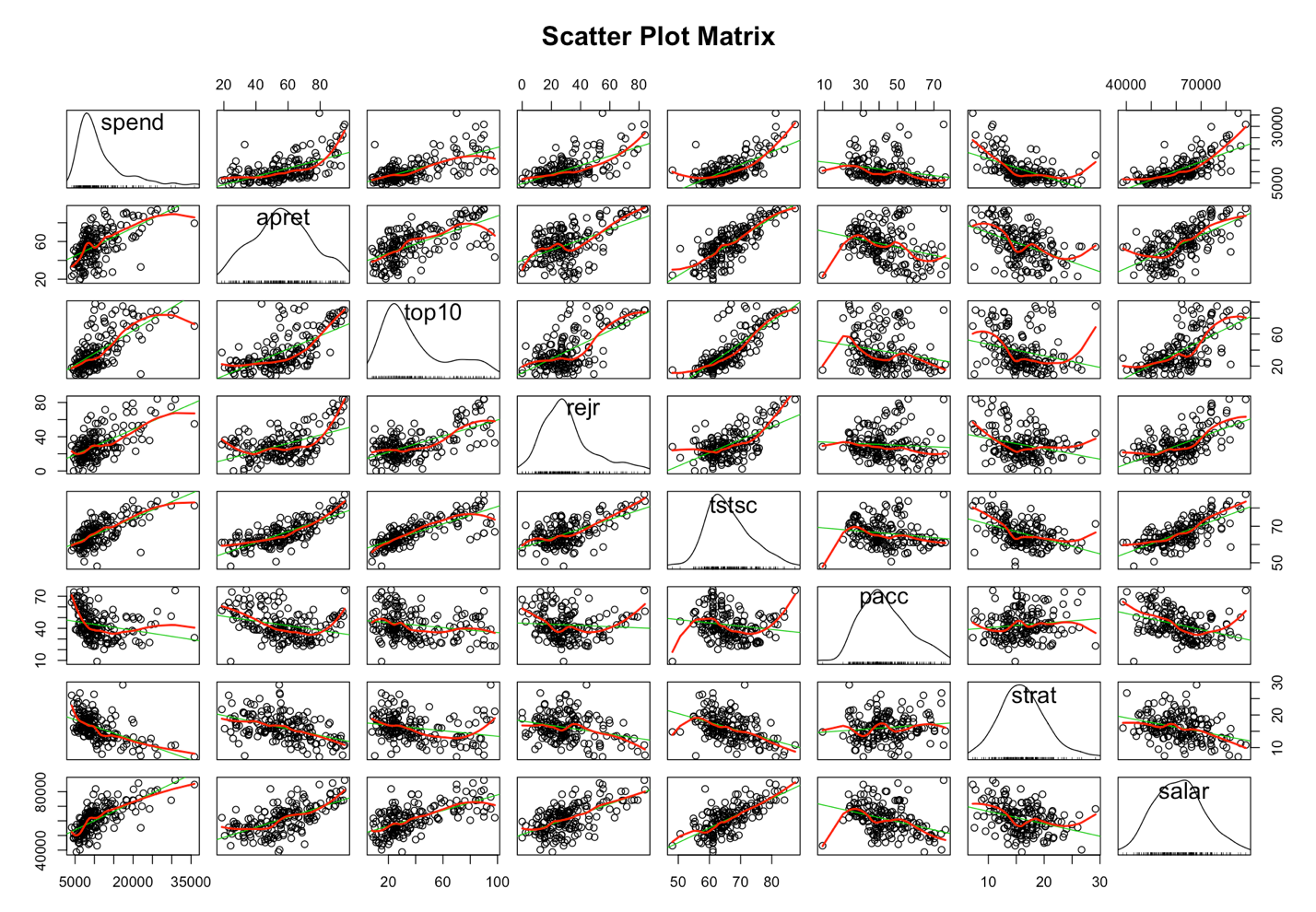
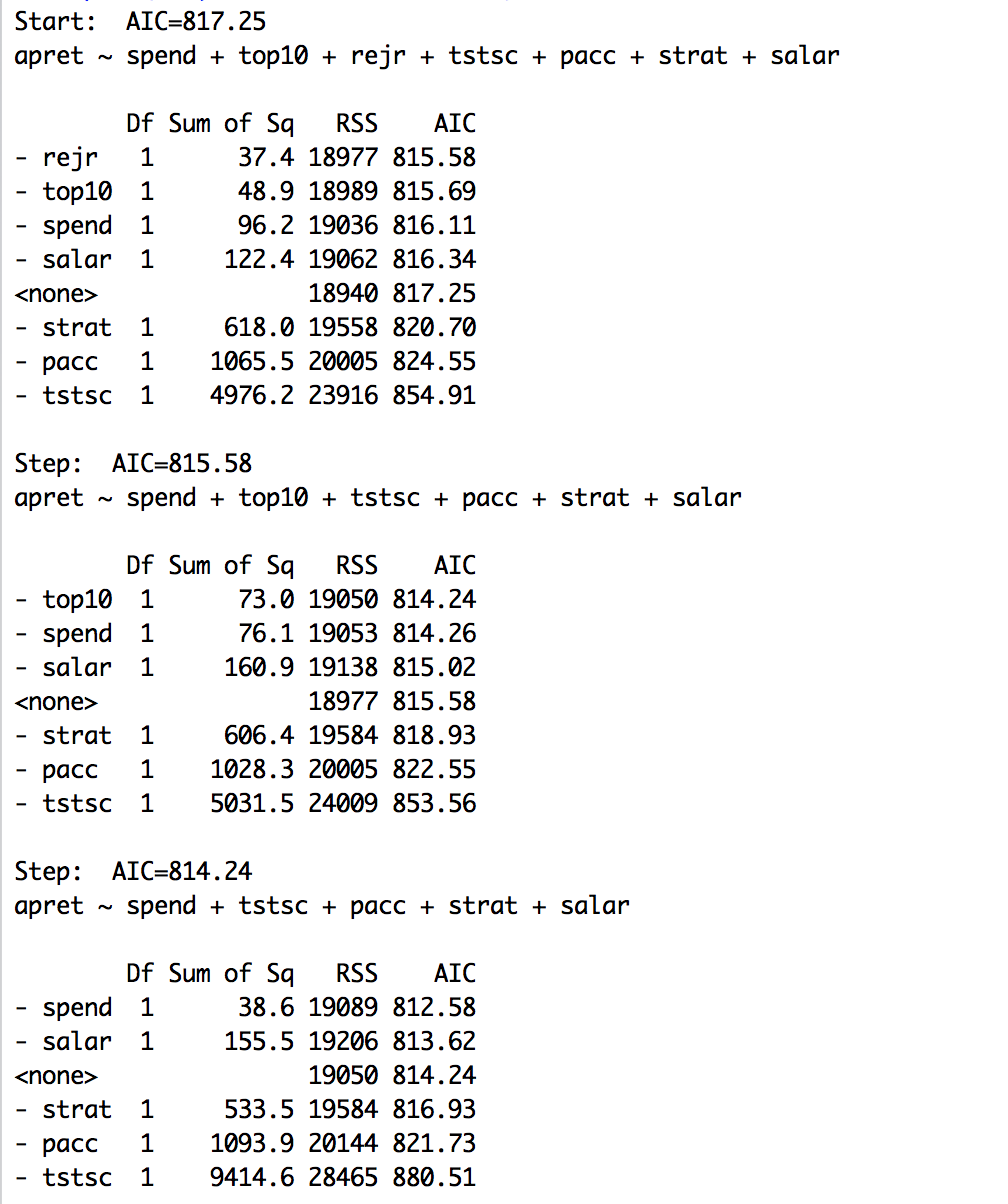
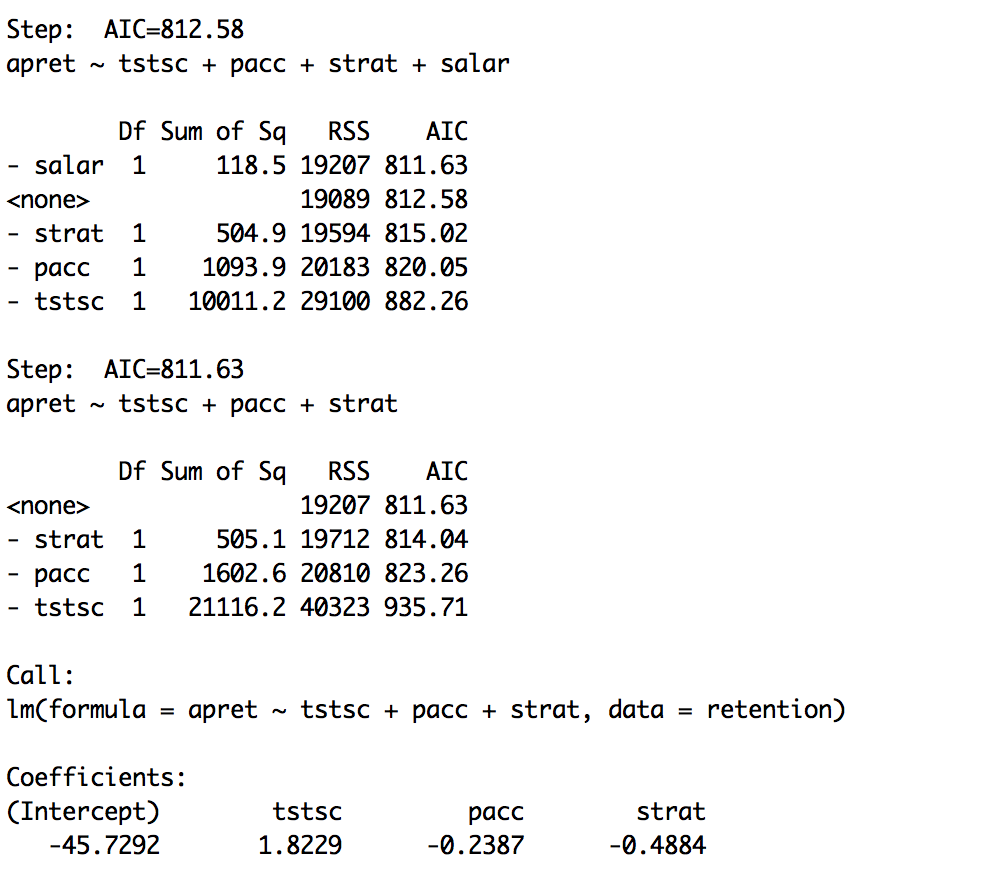


Figure 6. Relationship between apret and all other predictors





# Appendix

retention = read.csv("/Users/Tillie/Desktop/retention.csv", header=TRUE)

retention[1:3,]

summary(retention)

rm=(list=ls())

#list all the statistical information for variables

Rowname = c("spend","apret","top10","rejr","tstsc","pacc","strat","salar");

Columname = c("Min","1stQuartile","Median","Mean","3rdQuartile","Max","SD","Variance","Count")

spend=c(as.vector(summary(retention$spend)),sd(retention$spend),var(retention$spend),length(retention$spend))

apret=c(as.vector(summary(retention$apret)),sd(retention$apret),var(retention$apret),length(retention$apret))

top10=c(as.vector(summary(retention$top10)),sd(retention$top10),var(retention$top10),length(retention$top10))

rejr = c(as.vector(summary(retention$rejr)), sd(retention$rejr),var(retention$rejr),length(retention$rejr))

tstsc= c(as.vector(summary(retention$tstsc)),sd(retention$tstsc),var(retention$tstsc),length(retention$tstsc))

pacc=

c(as.vector(summary(retention$pacc)),sd(retention$pacc),var(retention$pacc),length(retention$pacc))

strat=

c(as.vector(summary(retention$strat)),sd(retention$strat),var(retention$strat),length(retention$strat))

salar=

c(as.vector(summary(retention$salar)),sd(retention$salar),var(retention$salar),length(retention$salar))

Summary =matrix(c(spend,apret,top10,rejr,tstsc,pacc,strat,salar),nrow=8,ncol=9,byrow=TRUE,dimnames = list(Rowname,Columname))

Summary

#plot histogram for apret, tstsc, and salar

library('ggplot2')

theme\_set((theme\_bw()))

ggplot(retention, aes(x=apret))+geom\_histogram(binwidth=10)

ggplot(retention, aes(x=tstsc))+geom\_histogram(binwidth=3)

ggplot(retention, aes(x=salar))+geom\_histogram(binwidth=3000)

#Normality test

shapiro.test(retention$spend)

shapiro.test(retention$apret)

shapiro.test(retention$top10)

shapiro.test(retention$rejr)

shapiro.test(retention$tstsc)

shapiro.test(retention$pacc)

shapiro.test(retention$strat)

shapiro.test(retention$salar)

#perform linear regression of apret on tstsc

fit1=lm(apret~tstsc,data=retention)

summary(fit1)

ggplot(retention, aes(x=tstsc,y=apret))+geom\_point()+geom\_smooth(method=lm,se=FALSE)

#perform linear regression of apret on salar

fit2=lm(apret~salar,data=retention)

summary(fit2)

ggplot(retention, aes(x=salar,y=apret))+geom\_point()+geom\_smooth(method=lm,se=FALSE)

#perform linear regression of apret on both tstsc and salar.

fit3=lm(apret~tstsc+salar,data=retention)

summary(fit3)

#perform linear regression of apret on all predictors.

fit4=lm(apret~.,data=retention)

summary(fit4)

#Scatter plot matrix

library('car')

suppressWarnings( ## (avoid printing the warnings)

scatterplotMatrix(retention, spread=FALSE, lty.smooth=2,

main="Scatter Plot Matrix")

)

#backward stepwise selection

library(MASS)

fit = lm(apret~., data=retention)

stepAIC(fit, direction="backward")