~~~
---
title: "Final Project"
author: "Narih Lee & Chetna Mahajan"
date: "May 7, 2018"
output: html_notebook
---


```{r SETUP}
colleges <- read.csv("/Users/narihlee/Downloads/colleges_regions_types.csv")


names(colleges)
dim(colleges)

colleges$Total.cost.per.yr. = as.numeric(colleges$Total.cost.per.yr.)
na_count <-sapply(colleges, function(y) sum(length(which(is.na(y)))))
data.frame(na_count)
colleges <- na.omit(colleges)
colleges.new <- colleges[,-c(1,2,5)]
attach(colleges.new)
write.csv(colleges.new, "collegestrim2.csv")

library("ISLR")
library("usdm")
library("MASS")
library("pls")
library("leaps")
library("CombMSC")

all.x <- model.matrix(Salary~.,data=colleges.new)[,-1]
```

```{r VIF}
vifstep(all.x, th=10)
#VIF shows no severe multicollinearity problem when we set our VIF threshold as 10, however, because we see some
strong linear correlations
vifstep(all.x, th=8)
x.exclude.avgaid <- all.x[, -c(10)]
#cor.test(Avg.need.based.aid, Total.cost.per.yr., alternative = "two.sided")
cor.test(X4.yr.grad.rate, Total.cost.per.yr., alternative = "two.sided")
x.exclude.avgaid.x4 <-x.exclude.avgaid[, -c(8)]
vifstep(x.exclude.avgaid.x4, th=8)
```

```{r ALLSUBSET}
result1 = leaps(x=x.exclude.avgaid.x4,y=Salary,method="Cp")
result2 = leaps(x=x.exclude.avgaid.x4,y=Salary,method="adjr2")
which.min(result1$Cp)
which.max(result2$adjr2)
result1$which[51,]
#Mallow's Cp model is Salary~TypePrivate+TypePublic+RegionSE/MS+ +RegionW/SW+Admit.rate+Total.cost.per.yr.
model.cp <- lm(Salary~Type+Region+Admit.rate+Total.cost.per.yr.)
result2$which[91,]
#Adjusted R2 model is
Salary~TypePrivate+TypePublic+RegionNE+RegionSE/MS+REgionW/SW+Admit.rate+Total.cost.per.yr.+Avg.non.need.based.aid+X..o

model.adjr2 <-
lm(Salary~Type+Region+Admit.rate+Total.cost.per.yr.+Avg.non.need.based.aid+X..of.non.need.based.aid+Avg.debt.at.graduat


```

```{r STEPWISE}
full.model <- lm(Salary~Type + Region + Admit.rate+ Total.cost.per.yr. + Avg.non.need.based.aid +
X..of.non.need.based.aid + Avg.debt.at.graduation, data = colleges.new)
step(full.model, direction="both", k=2)
#AIC model is Salary ~ Type+Admit.rate+Total.cost.per.yr.+X..of.non.need.based.aid+Avg.debt.at.graduation
model.aic <- lm(Salary~Type+Admit.rate+Total.cost.per.yr.+X..of.non.need.based.aid+Avg.debt.at.graduation)
step(full.model, direction="both", k=log(length(Salary)))
#BIC model is Salary ~ Type + Admit.rate + Total.cost.per.yr.
model.bic <- lm(Salary ~ Type + Admit.rate + Total.cost.per.yr.)

#looking at model with interaction terms using AIC
step(full.model,.~.^2, direction="both", k=2)
model.aic.2 <-lm(Salary ~ Type + Region + Admit.rate + Total.cost.per.yr. + X..of.non.need.based.aid +
Avg.debt.at.graduation + Type:Avg.debt.at.graduation + Region:Admit.rate + Type:Total.cost.per.yr. +
Total.cost.per.yr.:Avg.debt.at.graduation +Admit.rate:Avg.debt.at.graduation)
#looking at model with interaction terms using BIC
step(full.model,.~.^2, direction="both", k=log(length(Salary)))
model.bic.2 <- lm(Salary ~ Type + Admit.rate + Total.cost.per.yr. + X..of.non.need.based.aid + Avg.debt.at.graduation
~~~

~~~
+ Type:Avg.debt.at.graduation + Admit.rate:X..of.non.need.based.aid)
```

```{r PCA}
pcr.fit<-pcr(Salary~.,data=colleges.new,scale=TRUE,validation="CV")
summary(pcr.fit)
Z = scale(all.x)%*%(pcr.fit$loadings)
pcr.real.model <- lm(Salary~Z[, 1:8])
```

```{r PRESS}
PRESS(model.cp)
PRESS(model.adjr2)
PRESS(model.aic)
PRESS(model.bic)
PRESS(pcr.real.model)
PRESS(model.aic.2)
PRESS(model.bic.2)
#Our AIC model with interaction terms seems to work the best based on the PRESS criterion, but still need to figure
out how to find interaction terms with Mallow's and Adjusted R2
```

```{r RESIDUAL ANALYSIS}
summary(model.aic.2)
#based on the t-test, we can get rid of admit.rate:avg.debt
plot(model.aic.2)
#use studentized residual
scatter.smooth(studres(model.aic.2)~predict(model.aic.2))
#Does not appear that we need to do box-cox transformation because variation is pretty constant and when we apply log
to the response, the model fit does not really change.

scatter.smooth(predict(model.aic.2), rstudent(model.aic.2))

model.aic.2.reduced.1<-lm(Salary ~ Type + Region + Admit.rate + X4.yr.grad.rate + Total.cost.per.yr. +
Avg.debt.at.graduation)

model.aic.2.reduced.type.debt<-lm(Salary ~ Type + Region + Admit.rate + Total.cost.per.yr. + X..of.non.need.based.aid
+ Avg.debt.at.graduation + Type:Total.cost.per.yr.+ Region:Admit.rate + Total.cost.per.yr.:Avg.debt.at.graduation
+Admit.rate:Avg.debt.at.graduation)

model.aic.2.reduced.type.cost<-lm(Salary ~ Type + Region + Admit.rate + Total.cost.per.yr. + X..of.non.need.based.aid
+ Avg.debt.at.graduation + Type:Avg.debt.at.graduation + Region:Admit.rate +
Total.cost.per.yr.:Avg.debt.at.graduation +Admit.rate:Avg.debt.at.graduation)

model.aic.2.reduced.region<-lm(Salary ~ Type + Region + Admit.rate + Total.cost.per.yr. + X..of.non.need.based.aid +
Avg.debt.at.graduation + Type:Avg.debt.at.graduation +  Type:Total.cost.per.yr. +
Total.cost.per.yr.:Avg.debt.at.graduation +Admit.rate:Avg.debt.at.graduation)

anova(model.aic.2.reduced, model.aic.2)
anova(model.aic.2.reduced.type.debt, model.aic.2)
anova(model.aic.2.reduced.type.cost,model.aic.2)
anova(model.aic.2.reduced.region, model.aic.2)
#Interaction terms contribute significantly to the model based on the partial f-test conducted by anova analysis
```
```{r INFOBS}
p=11
n=length(Salary)
cutoff=qf(0.5, df1=p, df2=n-p)
which(cooks.distance(model.aic.2)>cutoff)
#no influential observation
```
~~~