```
#Final Project R Code - Student Performance
#Load data
library("openxlsx")
setwd("C:/Users/ISABE/OneDrive/Desktop/Stats/")
Final<-read.xlsx("Student_Performance.xlsx", sheet="Student_Performance",
check.names=TRUE)

#Normalize data so that variables can be compared to each other
Final$HStudy<-scale(Final$Hours.Studied)
Final$PrevScore<-scale(Final$Previous.Scores)
Final$HSleep<-scale(Final$Sleep.Hours)
Final$NumPaperPrac<-scale(Final$Sample.Question.Papers.Practiced)
Final$PI<-scale(Final$Performance.Index)

#Convert categorical variable (Extracurricular activities) to a dummy variable
Final$exa_contrast <- ifelse(Final$Extracurricular.Activities=='Yes', 1, 0)

#Save normalized data as new file
write.csv(Final,"/Users/ISABE/OneDrive/Desktop/Stats/NormalizedData.csv", row.names
= FALSE)

#Create full model with all parameters and no interactions
full <-lm(data=Final, PI~HStudy+PrevScore+exa_contrast+HSleep+NumPaperPrac)
summary(full)

#Assess assumptions for full model
#Added variable plots to assess linearity
library(car)
avPlots(full)
#All plots look linear so no problems with non linear relationships

#Influence plot and cooks distance to check for influential points
influencePlot(full)
cooks.distance(full)
#Compare cooks distance to F(6, 9994) for 50%
qf(0.50, 6, 9994)
#No cooks distance above threshold of F(6,9994), so no major influential points

#Check for multicollinearity using VIF
vif(full)
#No VIF greater than 10 so no multicollinearity issues

#Qq plot to check for normality
qqnorm(resid(full))
#QQ plot appears normal

#K-fold cross validation
library(MASS)
library(leaps)
library(caret)
```

```
set.seed(123)
train.control <-trainControl(method="cv",number=10)
full.k<-train(PI~HStudy+PrevScore+exa_contrast+HSleep+NumPaperPrac, data=Final,
method="leapBackward",
              tuneGrid=data.frame(nvmax=5),
              trControl=train.control)
full.k$results

#Research question 1:
#Ho: B Hours slept=B Extracurriculars=0
#Ha: One or both B do not equal zero
reduced_model1 <- lm(PI ~ HStudy + PrevScore + NumPaperPrac, data = Final)

#Assess assumptions for reduced model for question 1:
#Added variable plots to assess linearity
library(car)
avPlots(reduced_model1)
#All plots look linear so no problems with non linear relationships

#Influence plot and cooks distance to check for influential points
influencePlot(reduced_model1)
cooks.distance(reduced_model1)
#Compare cooks distance to F(4, 9996) for 50%
qf(0.50, 4, 9996)
#No cooks distance above threshold of F(4,9996), so no major influential points

#Check for multicollinearity using VIF
vif(reduced_model1)
#No VIF greater than 10 so no multicollinearity issues

#Qq plot to check for normality
qqnorm(resid(reduced_model1))
#QQ plot appears normal

#K-fold cross validation
library(MASS)
library(leaps)
library(caret)
set.seed(123)
train.control <-trainControl(method="cv",number=10)
reduced1.k<-train(PI~HStudy+PrevScore+NumPaperPrac, data=Final,
method="leapBackward",
                  tuneGrid=data.frame(nvmax=3),
                  trControl=train.control)
reduced1.k$results

#Compare full and reduced models
anova(reduced_model1, full)
#P is <0.05 so reject the null. Either hours slept, extracurricular activities, or
both have a B /= 0.
```

```
#Research question 2:
#Ho: B Hours studied=B Previous Scores = B new
#Ha: B Hours studied /= B Previous Scores
reduced_model2 <- lm(PI ~ I(HStudy+PrevScore) + exa_contrast + HSleep +
NumPaperPrac, data = Final)

#Assess assumptions for reduced model for question 2:
#Added variable plots to assess linearity
library(car)
avPlots(reduced_model2)
#All plots look linear so no problems with non linear relationships

#Influence plot and cooks distance to check for influential points
influencePlot(reduced_model2)
cooks.distance(reduced_model2)
#Compare cooks distance to F(5, 9995) for 50%
qf(0.50, 5, 9995)
#No cooks distance above threshold of F(5,9995), so no major influential points

#Check for multicollinearity using VIF
vif(reduced_model2)
#No VIF greater than 10 so no multicollinearity issues

#Qq plot to check for normality
qqnorm(resid(reduced_model2))
#QQ plot appears close to normal

#K-fold cross validation
library(MASS)
library(leaps)
library(caret)
set.seed(123)
train.control <-trainControl(method="cv",number=10)
reduced2.k<-train(PI~I(HStudy+PrevScore)+exa_contrast+HSleep+NumPaperPrac,
data=Final, method="leapBackward",
                  tuneGrid=data.frame(nvmax=4),
                  trControl=train.control)
reduced2.k$results

#Compare full and reduced models
anova(reduced_model2, full)
#P is <0.05 so reject the null. Hours studied and previous score do not have
equivalent B

#Research question 3:
#Ho: B hours study = B hours slept = B new
#Ha: B hours study /= B hours slept
reduced_model3 <- lm(PI ~ I(HStudy+HSleep) + PrevScore + exa_contrast +
NumPaperPrac, data = Final)
```

```
#Assess assumptions for reduced model for question 3:
#Added variable plots to assess linearity
library(car)
avPlots(reduced_model3)
#All plots look linear so no problems with non linear relationships

#Influence plot and cooks distance to check for influential points
influencePlot(reduced_model3)
cooks.distance(reduced_model3)
#Compare cooks distance to F(5, 9995) for 50%
qf(0.50, 5, 9995)
#No cooks distance above threshold of F(5,9995), so no major influential points

#Check for multicollinearity using VIF
vif(reduced_model3)
#No VIF greater than 10 so no multicollinearity issues

#Qq plot to check for normality
qqnorm(resid(reduced_model3))
#QQ plot appears close to normal

#K-fold cross validation
library(MASS)
library(leaps)
library(caret)
set.seed(123)
train.control <-trainControl(method="cv",number=10)
reduced3.k<-train(PI~I(HStudy+HSleep)+PrevScore+exa_contrast+NumPaperPrac,
data=Final, method="leapBackward",
                  tuneGrid=data.frame(nvmax=4),
                  trControl=train.control)
reduced3.k$results

#Compare full and reduced models
anova(reduced_model3, full)
#P is <0.05 so reject the null. Hours slept and hours studied do not have equivalent
B.

#Research question 4:
#Ho:B Hours studied*extracurricular activities = 0
#Ha:B Hours studied*extracurricular activities /= 0
#Reduced model for this question is the original full model
full_model4 <- lm(PI ~ HStudy*exa_contrast + HSleep + PrevScore + NumPaperPrac, data
= Final)

#Assess assumptions for full model for question 4:
#Added variable plots to assess linearity
library(car)
avPlots(full_model4)
```

```
#All plots look linear so no problems with non linear relationships

#Influence plot and cooks distance to check for influential points
influencePlot(full_model4)
cooks.distance(full_model4)
#Compare cooks distance to F(7, 9993) for 50%
qf(0.50, 7, 9993)
#No cooks distance above threshold of F(7,9993), so no major influential points

#Check for multicollinearity using VIF
vif(full_model4)
#Must use predictor type since interaction present
vif(full_model4, type='predictor')
#No VIF greater than 10 so no multicollinearity issues

#Qq plot to check for normality
qqnorm(resid(full_model4))
#QQ plot appears close to normal

#K-fold cross validation
library(MASS)
library(leaps)
library(caret)
set.seed(123)
train.control <-trainControl(method="cv",number=10)
full4.k<-train(PI~HStudy*exa_contrast+PrevScore+HSleep+NumPaperPrac, data=Final,
method="leapBackward",
               tuneGrid=data.frame(nvmax=6),
               trControl=train.control)
full4.k$results

#Compare full and reduced models
anova(full, full_model4)
#P is >0.05 so fail to reject the null. The interaction between hours studied and
extracurricular activities does not improve the model

#RMSE of full_model4 and full model are same (=0.106), all other models have larger
RMSE
#Research question 4 null hypothesis is not rejected
#Final model should not include the interaction effect between hours studied and
extracurricular activities

#Run Best Subset Algorithm to see if any other parameters can be excluded
#Find best subset of parameters
library(ALSM)
#Reorder columns so that parameters are last
Final <- Final[,c(1,2,3,4,5,6,11,7,8,9,10,12)]
bs<-BestSub(Final[,8:12],Final$PI,num=1)
bs
#Best subset is model with all 5 parameters, so keep all 5 in final model
```

```
#Model summary of final model selected (model with all 5 parameters)
summary(full)
#Rescaling intercept and coefficients to get model using original scales for each
parameter
#Scaled Y intercept = -0.0158
Intercept=sd(Final$Performance.Index)*(-0.0158)+mean(Final$Performance.Index)
Intercept

#Scaled B1=0.385
B1=0.385*(sd(Final$Performance.Index)/sd(Final$Hours.Studied))
B1

#Scaled B2=0.919
B2=0.919*(sd(Final$Performance.Index)/sd(Final$Previous.Scores))
B2

#B3 is categorical so no need to rescale

#Scaled B4=0.042
B4=0.042*(sd(Final$Performance.Index)/sd(Final$Sleep.Hours))
B4

#Scaled B5=0.029
B5=0.029*(sd(Final$Performance.Index)/sd(Final$Sleep.Hours))
B5

#Scaled Interaction B=0.003
B6=0.003*(sd(Final$Performance.Index)/sd(Final$HStudy))
B6
```