# Author - Romil Gupta

# Course - Statistical Learning, BDA, Hyderabad

# Data set used here is Concrete\_Data.xls and we have converted it to csv format for easier usage

# Import and understand the data. Look at the range of the various attributes

setwd ("E:/Data Science/Assignments/Statistics/25-Dec-17")

Concrete\_Data <- read.csv('Concrete\_Data.csv',header = TRUE)

View(Concrete\_Data)

str(Concrete\_Data)

# Making the name of heading shorter

names(Concrete\_Data) <- c('cement\_component','blast\_furnace\_slag','fly\_ash','water','superplasticizer', 'coarse\_aggregate','fine\_aggregate','age','compressive\_strength')

# Print out the summary statistics of all variables and arrange them in a neat tabular format.

View(summary(Concrete\_Data))

# Creating a test data

set.seed(1)

test <- sample(nrow(Concrete\_Data),200)

#Create Scatter Plot of Compressive Strength versus the other variables, taking each predictive variable at a time. Clearly label the graphs.

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$cement\_component)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$blast\_furnace\_slag)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$fly\_ash)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$water)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$superplasticizer)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$coarse\_aggregate)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$fine\_aggregate)

plot(Concrete\_Data$compressive\_strength , Concrete\_Data$age)

# Produce pairwise correlation coefficient table. Comment on the values of the correlations between each predictor and response. Do you think there is any pairwise correlation between predictors which may cause worry?

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$cement\_component)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$blast\_furnace\_slag)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$fly\_ash)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$water)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$superplasticizer)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$coarse\_aggregate)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$fine\_aggregate)

cor(Concrete\_Data$compressive\_strength , Concrete\_Data$age)

# As per our analysis, pairwise correlation between predictors is not a cause worry

cor(Concrete\_Data)

#Build Multiple Linear Regression model of Compressive Strength on ALL the predictors. Report multiple R2 of the model.

lm.model <- lm(compressive\_strength~.,data=Concrete\_Data)

summary(lm.model)

#Sum of square

lm.modelsum<- sum((lm.model$residuals)^2)

lm.modelsum

# As we can see the p-values of the t-statistic of all the coefficients are close to zero except for that of coarse\_aggregate and fine\_aggregate

# If we look at summary data, after p-value there is dot for coarse\_aggregate and fine\_aggregate

# Usually it means probablity between 0.5 and 0.10 and we do not include these in mode

# This indicates that other than the coefficients of coarse\_aggregate and fine\_aggregate, all the other coefficients of the model are significant for the accuracy of the fit

# Removing the coarse\_aggregate and fine\_aggregate attributes from the model

lm.model2 <- lm(compressive\_strength~.-coarse\_aggregate-fine\_aggregate,data=Concrete\_Data)

summary(lm.model2)

#Sum of square

lm.modelsum2<- sum((lm.model2$residuals)^2)

lm.modelsum

# Removal of "coarse\_aggregate" and "fine\_aggregate" attributes doesn't improve the R-squared value of the model and the mean squared error of the model data doesn't decrease

# Therefore we will go ahead with the first model, lm.model

# Residual Standard Error of lm.model

sqrt(deviance(lm.model)/lm.model$df.residual)

# Mean response of the dataset

mean(Concrete\_Data$compressive\_strength)

# The residual standard error of lm.model is around 10.340and the mean response is about 36.00.

# The error rate is as following.

# Error rate

10.40/36\*100

28.88

# R Squared value

summary(lm.model)$r.squared

# The R-squared value of lm.model is 0.61. This indicates that 61% of the variability in the response has been explained by the model.

# In our model multicollinearity is not there, as there is no independent variable which are highly co related

# There s no definate cut off but we are considering any correlation greater than or less than 0.7 is cause for concern.

# So we will sticking to model 1

#################################################################

# Predicting The Model

predict\_strength = predict(lm.model,Concrete\_Data[test,])

mean((Concrete\_Data[test,]$compressive\_strength-predict\_strength)^2)

str(Concrete\_Data)

# Mean Squared Error of the lm.model2 on test data

predict\_strength2 = predict(lm.model2,Concrete\_Data[test,])

mean((Concrete\_Data[test,]$compressive\_strength-predict\_strength2)^2)

#Correlation between response in the test data and the response of lm.model

cor(Concrete\_Data[test,]$compressive\_strength,predict(lm.model,Concrete\_Data[test,]))

# Comparison of mean squared error between training data and test data

# Mean squared error of lm.model on training data

mean((Concrete\_Data[-test,]$compressive\_strength-predict(lm.model,Concrete\_Data[-test,]))^2)

# Mean squared error of lm.model on test data

mean((Concrete\_Data[test,]$compressive\_strength-predict(lm.model,Concrete\_Data[test,]))^2)

# Testing a model on a single new data

new\_data <- data.frame(cement\_component=310,

blast\_furnace\_slag=0,

fly\_ash=0,

water=150,

superplasticizer=5,

coarse\_aggregate=1047,

fine\_aggregate=676,

age=28)

predict(lm.model,new\_data,interval="confidence")

#############################################################################################

####################End of Assignment##################