Done by : Ashwini k

Mentor Name : Arvind

Abstract: Multi Variate Regression Model in R

ACD\_ANR\_Project2.2

Problem Statement: **(Part I )**

Develop a Multivariate Regression Model for the data collected for several hundred used General Motors (GM) car values based on a variety of characteristics such as:

• Price: suggested retail price for the used GM car

• Mileage: number of miles the car has been driven

• Make: manufacturer of the car such as Cadillac, Pontiac, and Chevrolet

• Cylinder: number of cylinders in the engine

• Liter: a more specific measure of engine size

• Cruise: indicator variable representing whether the car has cruise control (1 = cruise)

• Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)

• Leather: indicator variable representing whether the car has leather seats (1 = leather)

Dataset :

Solution : Find a best fit model which verifies

Assumptions of Regression of model

APPROACH 1 :

STEPS TAKEN

* Whole data was split into 30 and 70 as train and test
* Usage of dummy Variables
* Introduction of boxcox transformation
* Since lamda was in between 0 to -0.5,log(target) and 1/sqrt(target) was applied

Source code with **comments**  :

gmds <- read.csv("C:/ASHWINI/Personal/ACAD/materials/project 2/gm\_dataset.csv",header=TRUE)

View(gmds)

summary(gmds)

gmds$Cadillac <- ifelse(gmds$Make=="Cadillac",1,0) #for cadillac dummy variable

gmds$Chevy <- ifelse(gmds$Make=="Chevrolet",1,0) #for Chevrolet dummy variable

names(gmds)

View(gmds)

nrow(gmds)

set.seed(1)

gmds1 <-sample(nrow(gmds), nrow(gmds)\*.7)

gmds1

gmtrain <- gmds[gmds1,]

gmtest<-gmds[-gmds1,]

View(gmtrain)

View(gmtest)

names(gmtrain)

plot(gmtrain$Cylinder,gmtrain$Price)

gmmod1 <- lm(Price~.,data=gmtrain)

summary(gmmod1)

plot(gmmod1)

gmmod6 <- lm(Price~.,data=gmds)

summary(gmmod6)

plot(gmmod6)

boxcox(lm(Price~.,data=gmds),lambda=seq(-2,2,by=.1))

class(gmmod6)

gmod7<-lm((1/sqrt(Price))~.,data=gmds)

summary(gmod7)

par(mfrow=c(2,2))

plot(gmod7)

gm3res<-data.frame(gmds,fittedval=fitted(gmod7),resi=resid(gmod7))

gm3res

View(gm3res)

gmod8<-lm(log(Price)~.,data=gmds)

summary(gmod8)

par(mfrow=c(2,2))

plot(gmod8)

View(gmds)

mean(gmod7$residuals)

mean(gmod8$residuals)

par(mfrow=c(2,2))

plot(gmmod1)

gm1res<-data.frame(gmdat,fittedval=fitted(gmmod1),resi=resid(gmmod1))

gm1res

head(gm1res)

library(MASS)

boxcox(lm(Price~.,data=gmds),lambda=seq(-2,2,by=.1))

plot(gm1res$Mileage,gm1res$resi)

plot(gm1res$Cylinder,gm1res$resi)

plot(gm1res$Liter,gm1res$resi)

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

#1/sqrt(Price) lambda when -0.5

gmmod2<-lm((1/sqrt(Price))~.,data=gmds)

gmmod2

summary(gmmod2)

gmmod2$residuals

par(mfrow=c(2,2))

plot(gmmod2)

gm2res<-data.frame(gmds,fittedval=fitted(gmmod2),resi=resid(gmmod2))

gm2res

head(gm2res)

plot(gm2res$Mileage,gm2res$resi)

plot(gm2res$Cylinder,gm2res$resi)

plot(gm2res$Liter,gm2res$resi)

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

#log(Price)

gmmod3<-lm(log(Price)~.,data=gmds)

gmmod3

summary(gmmod3)

gmmod3$residuals

par(mfrow=c(2,2))

plot(gmmod3)

gm3res<-data.frame(gmds,fittedval=fitted(gmmod2),resi=resid(gmmod2))

gm3res

head(gm3res)

plot(gm3res$Mileage,gm3res$resi)

plot(gm3res$Cylinder,gm3res$resi)

plot(gm3res$Liter,gm3res$resi)

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

gmdat[62,]

boxplot(gmds$Price)

gmdat31<-gmds[-62,]

gmmod4<-lm(Price~.,data=gmdat31)

summary(gmmod4)

par(mfrow=c(2,2))

plot(gmmod4)

boxcox(lm(Price~.,data=gmdat31),lambda=seq(-2,2,by=.1))

gmdatsqr<-lm((1/sqrt(Price))~.,data=gmdat31)

summary(gmdatsqr)

par(mfrow=c(2,2))

plot(gmdatsqr)

gmodln<-lm(log(Price)~.,data=gmdat31)

gmodln

summary(gmodln)

par(mfrow=c(2,2))

plot(gmodln)

acf(gmodln$residuals)

mean(gmodln$residuals)

mean(gmdatsqr$residuals)

CONCLUSION for APPROACH 1 :

Assumptions of regression model could not get in the best fit model so APPROACH 2 was taken.

APPROACH 2 :

STEPS TAKEN

* Whole data set was considered,R2 of 0.88 was achieved.SO a transformation was applied to get a better fit model
* Introduction of boxcox transformation, Since lamda was in between 0 to -0.5,log(target) and 1/sqrt(target) was applied
* 1/sqrt(target) provided a R2 of 0.91 .

Log(target)

* Log(target) was applied and Residual was fitted assumption was verified.and outliers was observed
* Oultiers was excluded and transformation applied on new dataset and achieved a R2 of 0.94(A BEST FIT MODEL)
* Through gvlma package observed that all assumptions are acceptable

Source code with **comments**  :

gmds <- read.csv("C:/ASHWINI/Personal/ACAD/materials/project 2/gm\_dataset.csv",header=TRUE)

View(gmds)

summary(gmds)

md1 <- lm(Price~.,data=gmds)

summary(md1)

plot(md1)

boxcox(lm(Price~.,data=gmds),lambda=seq(-2,2,by=.1))

md2<-lm((1/sqrt(Price))~.,data=gmds)

summary(md2)

par(mfrow=c(2,2))

plot(md2)

md2df<-data.frame(gmds,fittedval=fitted(md2),resi=resid(md2))

md2df

View(md2df)

plot(md2df)

md3<-lm((log(Price))~.,data=gmds)

summary(md3)

par(mfrow=c(2,2))

plot(md3)

gmdsnew<-gmds[-c(62:65),]

gmdsnew

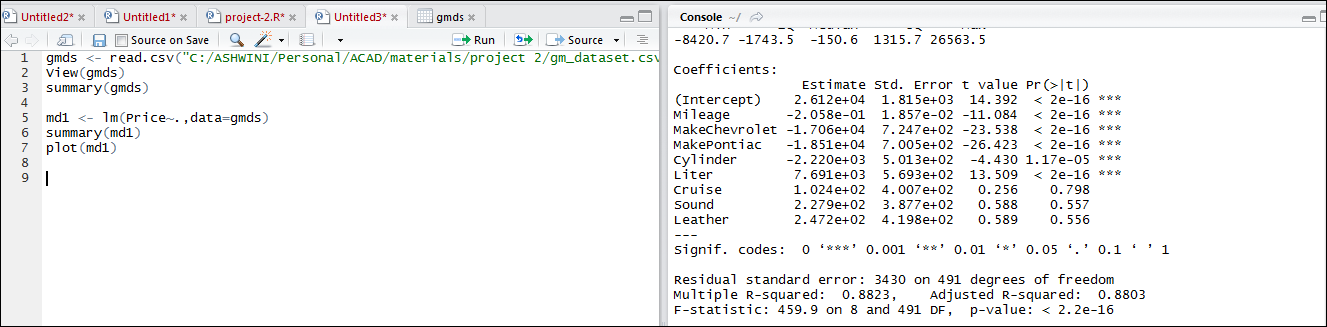
md4<-lm((log(Price))~.,data=gmdsnew)

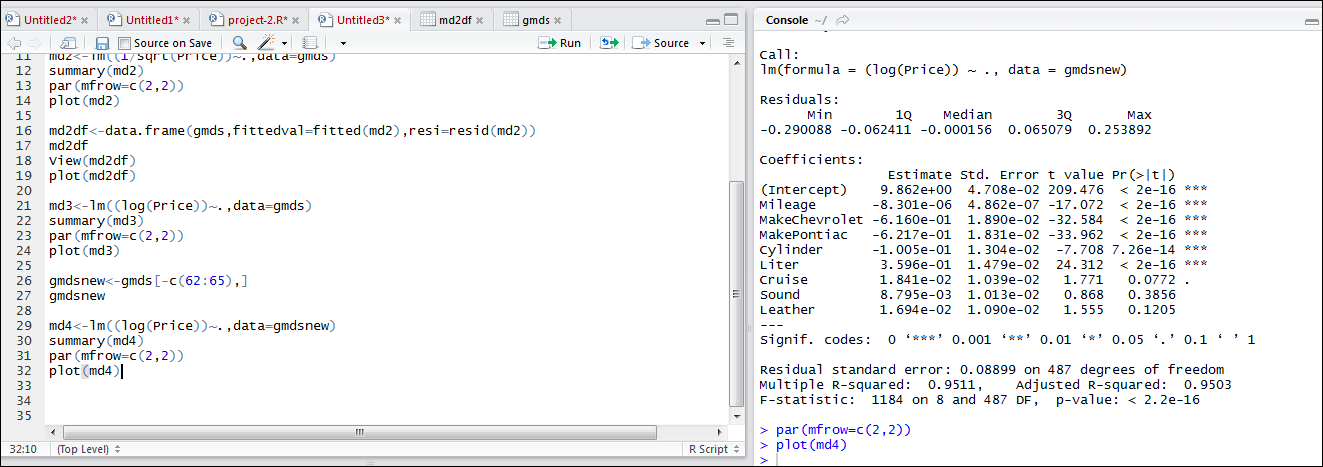
summary(md4)

par(mfrow=c(2,2))

plot(md4)

Screenshots





THANK YOU