MULTI LINEAR REGRESSION

1. COMPUTER DATA:

a) model.computer <- lm(price~speed+hd+ram+screen+cd+multi+premium+ads+trend)

This is the model prepared with all the input variables. R^2 obtained for this model is 0.7756. we also observe that the p values for all the input variables are less than alpha (0.05), which determines the model is a good one. But the R^2 value is less than 0.85, in order to increase that value we can apply transformations.

b) model.computer1<-lm(sqrt(price)~speed+hd+ram+screen+cd+multi+premium+ads+trend)

This is the model where we have applied transformation(sqrt). For this model R^2=0.7853

And all the p values are less than alpha (0.05)

c) model.computer2<-lm(log(price)~speed+hd+ram+screen+cd+multi+premium+ads+trend)

This is the model where we have applied transformation(log). For this model R^2=0.7832

And all the p values are less than alpha (0.05)

2.TOYOTA COROLLA:

a) model.toyoto <- lm(Price ~ Age_08_04 + KM + HP + cc + Doors + Gears + Quarterly_Tax + Weight)

This is the model prepared with all the input variables. R^2 obtained for this model is 0.8638.

P value of cc is 0.17909>0.05 and p value of doors is 0.96777>0.05.

b)We now make a model with only “cc” as input variable - model.toyoto_cc <- lm(Price ~ cc)

R^2 = 0.01597

c)We now make a model with only “doors” as input variable - model.toyoto_Doors <- lm(Price ~ Doors)

R^ = 0.03435

d)we make a model with both “cc” and “doors” as the input variables - model.toyoto_cD <- lm(Price ~ cc + Doors)

R^2 = 0.04688

Now, when we use the influence plots, we can observe that the data point “81” is influencing our model the most.so we tend to remove the entire 81st observation.

e) model.toyoto1 <- lm(Price ~ Age_08_04 + KM + HP + cc + Doors + Gears + Quarterly_Tax + Weight, data = ToyotaCorolla1[-81,])

R^2 = 0.8694 ; p value of gears = 0.4878>0.05

stepAIC recommends us to build us a model without using “cc” and “doors” for a better model.

f) model.final <-lm(Price ~ Age_08_04 + KM + HP + Gears + Quarterly_Tax + Weight)

R^2 = 0.8636; all the p values are good (<0.05)

g) model.final1 <- lm(Price ~ Age_08_04 + KM + HP + Gears + Quarterly_Tax + Weight, data = ToyotaCorolla1[-81,])

In this model we have removed the 81 observation along with both “cc” and “doors”

R^2 = 0.8632

3.X_50 start_ups:

a) model.X50<-lm(Profit~X50_Startups1$`R&D Spend`+Administration+`Marketing Spend`)

This is the model prepared with all the input variables. R^2 obtained for this model is 0.9507.

P value of Administration is 0.602>0.05 and p value of Marketing spend is 0.105>0.05.

b) We now make a model with only “Administration” as input variable - model.X50_AD<-lm(Profit~Administration)

R^2 = 0.04029

c) We now make a model with only “Marketing Spend” as input variable - model.X50_MS<-lm(Profit~`Marketing Spend`)

R^2 = 0.5592

d)we make a model with both “Administration” and “Marketing Spend” as the input variables - model.X50_AM<-lm(Profit~Administration+`Marketing Spend`)

R^2 = 0.6097

Now, when we use the influence plots, we can observe that the data point “50” is influencing our model the most.so we tend to remove the entire 50th observation.

e) model.X50_1 <- lm(Profit~`R&D Spend`+Administration+`Marketing Spend`,data = X50_Startups1[-50,])

R^2 = 0.9613; p value of administration is 0.6071 >0.05. And p value of Marketing Spend is 0.0746>0.05

stepAIC recommends us to build us a model without using “Administration” for a better model.

f) model.final<-lm(Profit~`R&D Spend`+`Marketing Spend`,data = X50_Startups1)

R^2 = 0.9505; p value of Marketing Spend = 0.06>0.05.

g) model.final1<-lm(Profit~`R&D Spend`+`Marketing Spend`,data = X50_Startups1[-50,])

In this model we have removed the 50th observation along with both “Administration”.

R^2 = 0.9611