Abstract: Linear Regression Model in R

ACD_ANR_PROJECT2.1

DONE BY : ASHWINI K
MENTOR NAME : ARVIND

Problem Statement

Imagine that the CEO of a DVD player sales company approaches you in order to predict the sale of DVDs. He also provides you the data such as the advertising budget (in thousands), sales (in thousands), number of times the song is played on the radio channel, Radio Mirchi per week and the attractiveness of the brand (rated on a scale of 1 to 10 by an independent agency).

Dataset -

:

Approach

: Splitting the data into 70 and 30

Inference

:

1 : it's a MLR, one response variable being sales and explanatory variables being advertise, attractiveness and plays.
2: sales=advertiseX1+attractivenssX2+playsX3
3: Correlation between response and explanatory variables are having moderate

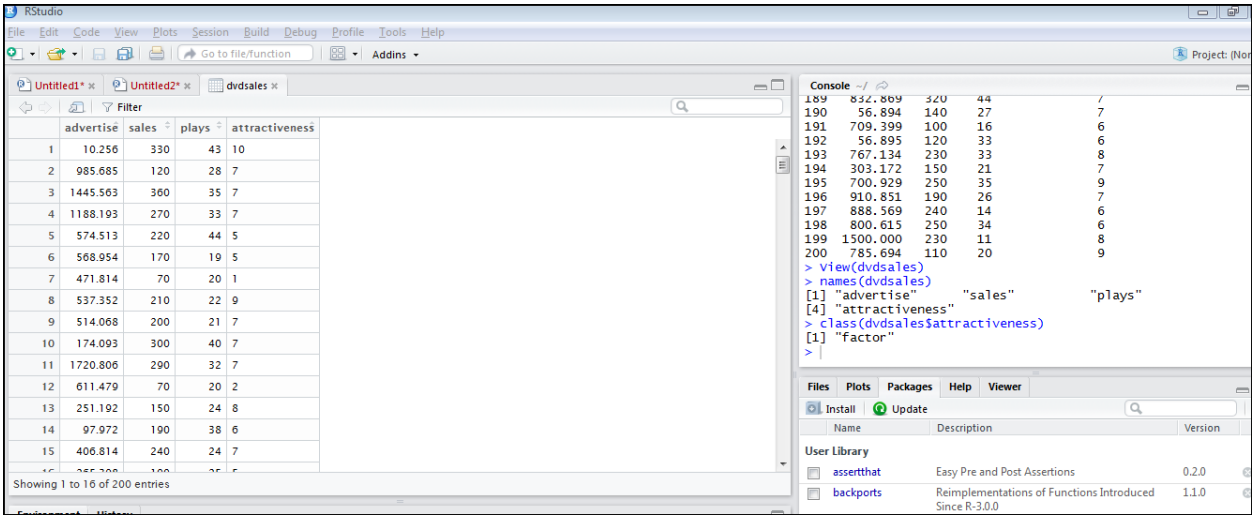
Source code with comments:

```
#Reading the data#
dvdsales<-read.csv("C:/All/R language/Final Project/fwdproject2/Sales_dataset.csv",header=TRUE)
dvdsales
View(dvdsales)
names(dvdsales)
class(dvdsales$attractiveness)

#Correlation betwwen predictors and response variable#
cor(dvdsales$sales,dvdsales$advertise)
cor(dvdsales$sales,dvdsales$plays)
cor(dvdsales$sales,dvdsales$attractiveness)

#Conversion from numeric to factor
dvdsales$attractiveness<-as.factor(dvdsales$attractiveness)
class(dvdsales$attractiveness)
summary(dvdsales)
nrow(dvdsales)
```

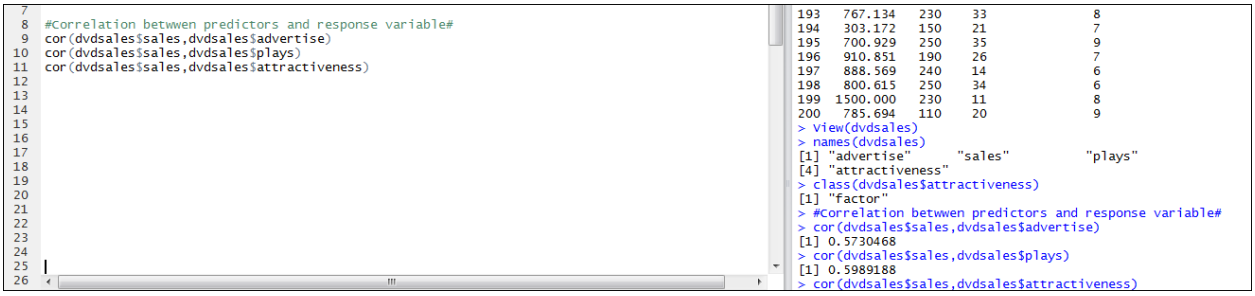
```
#Split the data into 70 and 30
set.seed(1)
dvdds<-sample(nrow(dvdsales),nrow(dvdsales)*0.7)

#Conversion of changing qualifying into levels
dvdsales$attractiveness1<-ifelse(dvdsales$attractiveness< 3,1,ifelse(dvdsales$attractiveness<
5,2,ifelse(dvdsales$attractiveness< 7,3,ifelse(dvdsales$attractiveness< 9,4,5))))
dvdsales$attractiveness1<-as.factor(dvdsales$attractiveness1)
dvdsales$attractiveness<-as.factor(dvdsales$attractiveness)
summary(dvdsales)

#attractiveness isn't required since it's been qualifies into levels
dvdtrain<-dvdsales[dvdds,-4]
summary(dvdtrain)
dvdtest<-dvdsales[-dvdds,-4]
summary(dvdtest)

#target variable is not required in test and train input
dvdt<-dvdtest[,-2]
dvdt

#Model Creation
smod1<-lm(sales~.,data=dvdtrain)
summary(smod1)

#Verifying Assumptions in model
plot(smod1)

#Prediction in test data
dvdt$testres<-predict(smod1,newdata = dvdt)
summary(dvdt)
dvdt$testres
```

Screenshot's for the solution after running code :

dvdsales



correlation between predictors



conversion from numeric to factor

```
#Reading the data#
dvdsales<-read.csv("C:/All/R language/Final Project/fwdproject2/Sales_dataset.csv",header=TRUE)
dvdsales
view(dvdsales)
names(dvdsales)
class(dvdsales$attractiveness)

#Correlation betwwen predictors and response variable#
cor(dvdsales$sales,dvdsales$advertise)
cor(dvdsales$sales,dvdsales$plays)
cor(dvdsales$sales,dvdsales$attractiveness)




#Conversion from numeric to factor
dvdsales$attractiveness<-as.factor(dvdsales$attractiveness)
class(dvdsales$attractiveness)
view(dvdsales)
summary(dvdsales)
nrow(dvdsales)
```

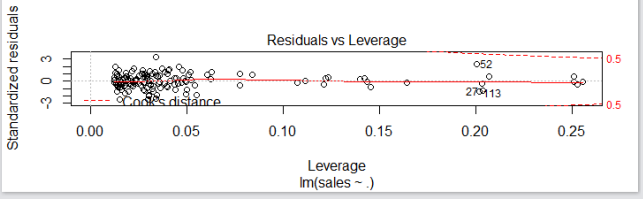
```
)
> class(dvdsales$attractiveness)
[1] "factor"
> view(dvdsales)
Error in view(dvdsales) : could not find function "view"
> summary(dvdsales)
   advertise          sales           plays
 Min.   :  9.104   Min.   : 10.0   Min.   : 0.00
 1st Qu.: 215.609   1st Qu.:137.5   1st Qu.:19.75
 Median : 526.311   Median :200.0   Median :28.00
 Mean   : 608.592   Mean   :193.2   Mean   :27.50
 3rd Qu.: 903.380   3rd Qu.:250.0   3rd Qu.:36.00
 Max.   :2271.860   Max.   :360.0   Max.   :63.00

 attractiveness
 7      :73
 6      :44
 8      :44
 5      :17
 9      :12
 4      : 4
 (other): 6
> nrow(dvdsales)
[1] 200
>
```

Attractiveness response removed

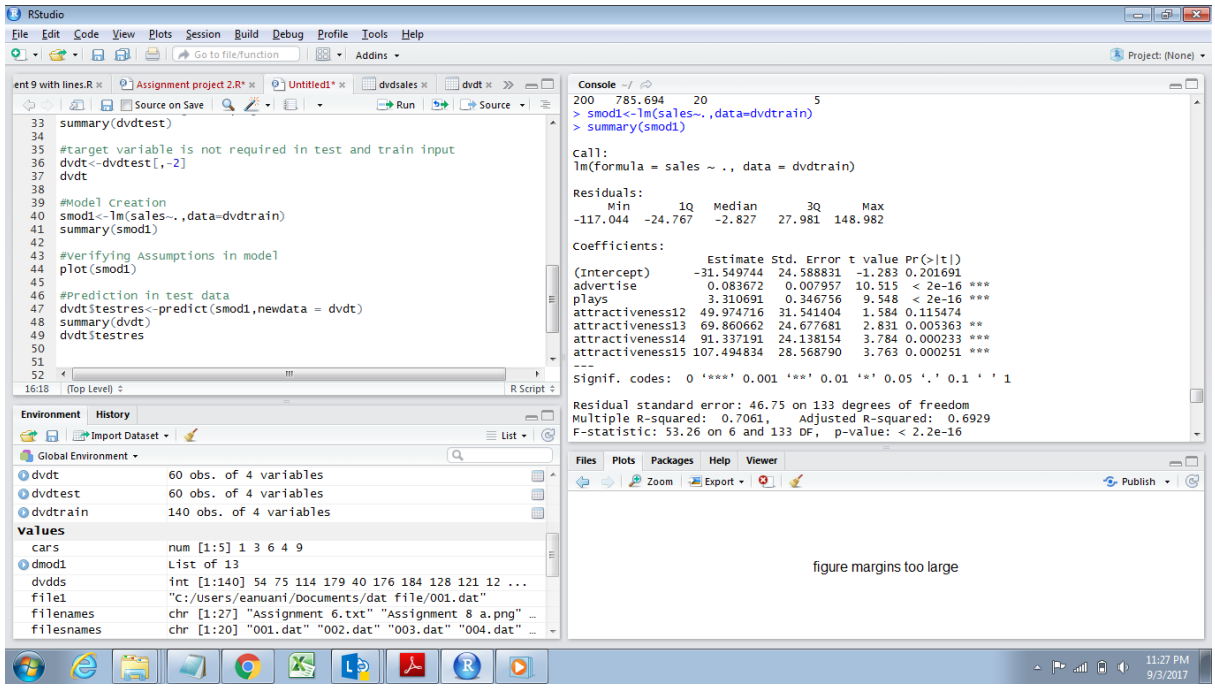
```
dvdds<-sample(nrow(dvdsales),nrow(dvdsales)*0.7)

#Conversion of changing qualifying into levels#
dvdsales$attractiveness1<-ifelse(dvdsales$attractiveness< 3,1,ifelse(
dvdsales$attractiveness1<-as.factor(dvdsales$attractiveness1)
dvdsales$attractiveness<-as.factor(dvdsales$attractiveness)
summary(dvdsales)

#attractiveness isnt required since its been qulaifies into levels
dvdtrain<-dvdsales[dvdds,-4]
summary(dvdtrain)
dvdtest<-dvdsales[-dvdds,-4]
summary(dvdtest)

#target variable is not required in test and train input
dvdt<-dvdtest[,-2]
dvdt

#Model Creation
```

```
> dvdtrain<-dvdsales[dvdds,-4]
> summary(dvdtrain)
   advertise          sales           plays       attractiveness1
 Min.   :  10.65   Min.   : 30.0   Min.   : 0.00   1: 4
 1st Qu.: 221.57   1st Qu.:120.0   1st Qu.:19.00   2: 5
 Median : 544.05   Median :195.0   Median :27.50   3:38
 Mean   : 631.18   Mean   :191.9   Mean   :26.62   4:84
 3rd Qu.: 918.27   3rd Qu.:242.5   3rd Qu.:35.00   5: 9
 Max.   :2271.86   Max.   :360.0   Max.   :55.00
> dvdtest<-dvdsales[-dvdds,-4]
> summary(dvdtest)
   advertise          sales           plays       attractiveness1
 Min.   :   9.104   Min.   : 10.0   Min.   : 4.00   1: 0
 1st Qu.: 209.918   1st Qu.:150.0   1st Qu.:21.00   2: 0
 Median : 513.701   Median :205.0   Median :28.00   3:23
 Mean   : 555.886   Mean   :196.3   Mean   :29.55   4:33
 3rd Qu.: 834.390   3rd Qu.:250.0   3rd Qu.:37.50   5: 4
 Max.   :1547.159   Max.   :340.0   Max.   :63.00
> dvdt<-dvdtest[,-2]
> dvdt
    advertise plays attractiveness1
```

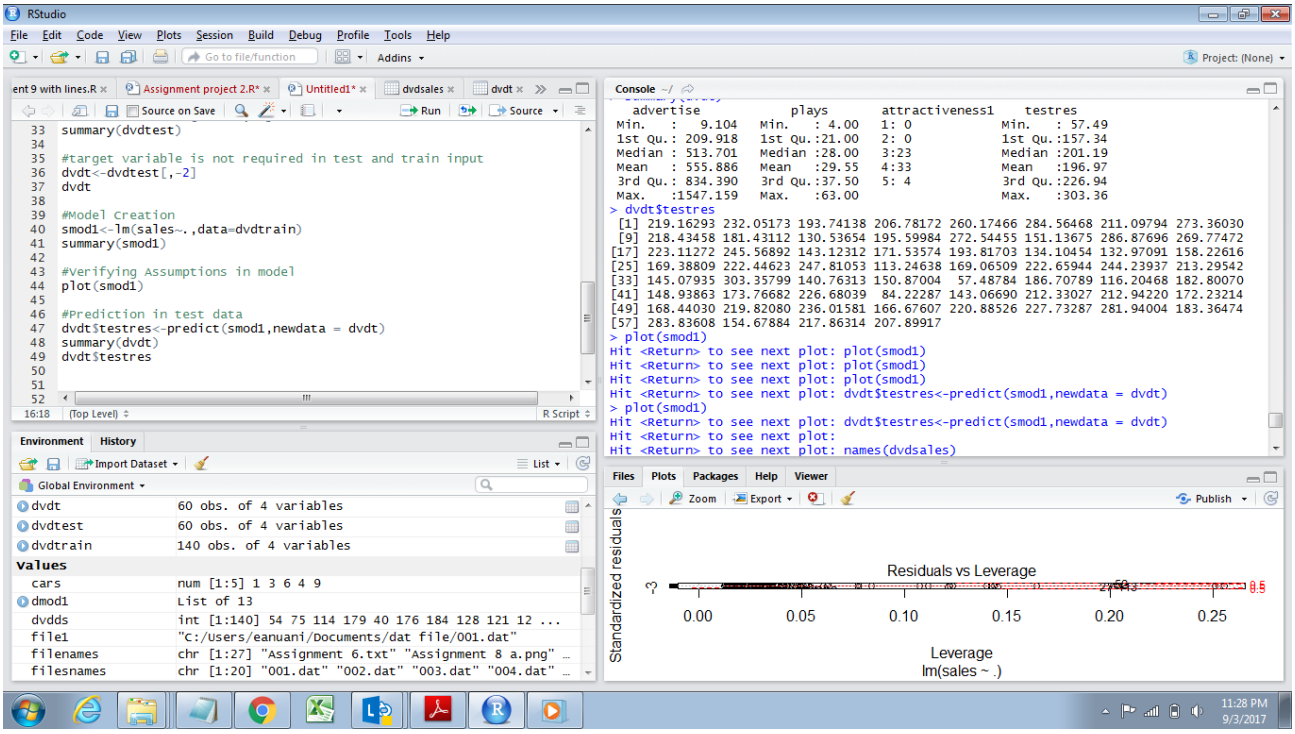
```
Global Environment
dvdt        60 obs. of 4 variables
dvdtest     60 obs. of 4 variables
dvdtrain    140 obs. of 4 variables
values
cars        num [1:5] 1 3 6 4 9
dmod1       List of 13
dvdds       int [1:140] 54 75 114 179 40 176 184 128 121 12 ...
file1       "C:/Users/eanuani/Documents/dat file/001.dat"
filenames   chr [1:27] "Assignment 6.txt" "Assignment 8 a.png" ...
filesnames  chr [1:20] "001.dat" "002.dat" "003.dat" "004.dat" ...
```



Model Creation



testres



CONCLUSION :

- Since P value is less than 0.5 the relationship is linear.
- Predictors, attractiveness 12 has no influence on target variable
- Advertise and plays have minimal error has their standard error is less
- Assumptions criteria are verified.

Predictors have direct relationship with target.

THANK YOU