Multiple Regression

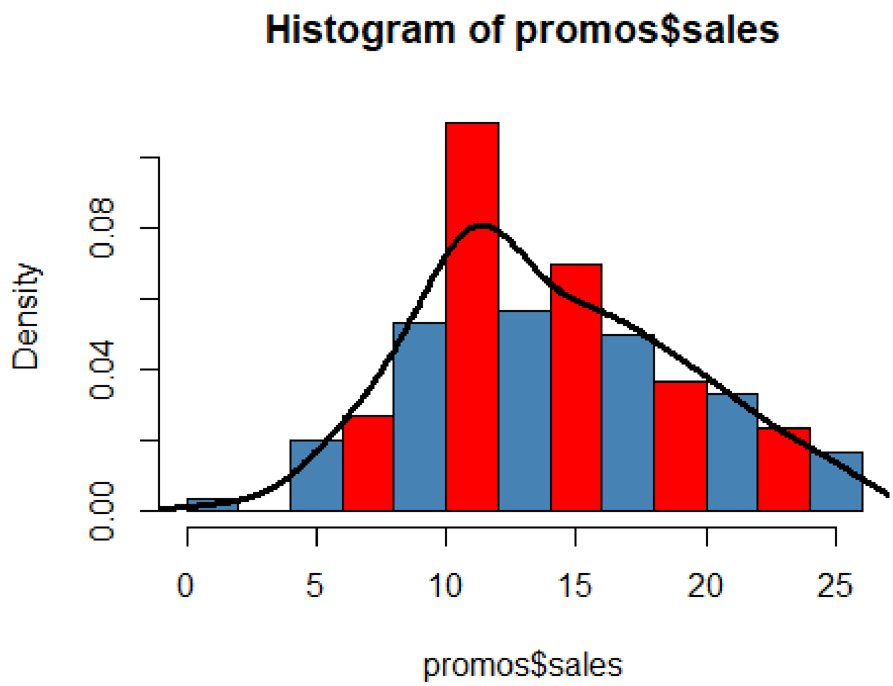
Suhail Shaikh

12/22/2019

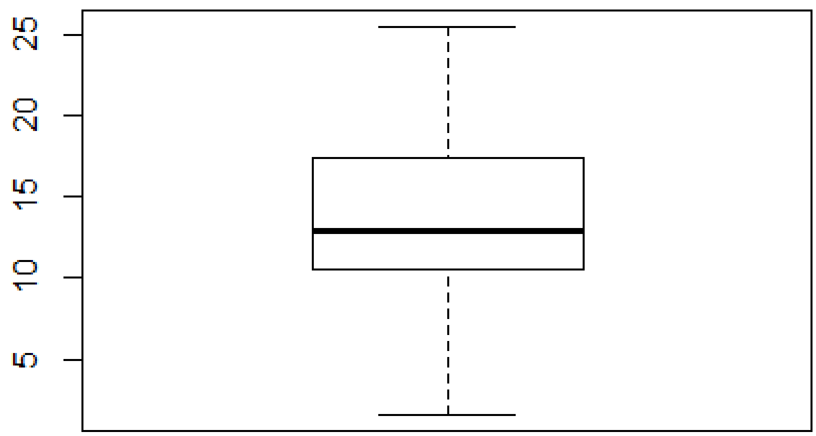
Q.1)Build an exhaustive multiple regression model to predict sales Provide a thoughtful and thorough explanation of the findings of this model.

```
setwd("C:\\Users\\shaik\\Desktop\\projects data mining\\General\\Multiple
Regression")
load("Multiple Regression.RData")

str(promos)

## Classes 'tbl_df', 'tbl' and 'data.frame':    150 obs. of  5 variables:
##  $ sales   : num  23.8 12.7 5.3 14.2 7.2 22.2 9.5 15.9 15.5 10.3 ...
##  $ region  : Factor w/ 4 levels "East","Midwest",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ online  : num  210.8 265.2 5.4 120.5 8.7 ...
##  $ paper   : num  37.7 43 9.4 14.2 75 72.3 7.4 22.9 26.2 9 ...
##  $ in_store: num  49.6 2.9 29.9 28.5 48.9 36.5 1.4 16.7 16.9 1.9 ...

head(promos)

##   sales region online paper in_store
## 1  23.8   East  210.8  37.7     49.6
## 2  12.7   East  265.2  43.0      2.9
## 3   5.3   East    5.4   9.4     29.9
## 4  14.2   East  120.5  14.2     28.5
## 5   7.2   East    8.7  75.0     48.9
## 6  22.2   East  250.9  72.3     36.5

#Step1 -Checking Normality of dependent Variable
library(psych)
summary(promos$sales)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60   10.53   12.90   14.01   17.38   25.50

describe(promos$sales)

##    vars   n  mean   sd median trimmed  mad min  max range skew kurtosis
## X1    1 150 14.01 5.04   12.9   13.83 4.97 1.6 25.5  23.9  0.3    -0.53
##      se
## X1 0.41

#Here in sales mean is approximately equal to median and sd is very small
compared to the mean Hence sales seems to have normally distributed. Lets
```

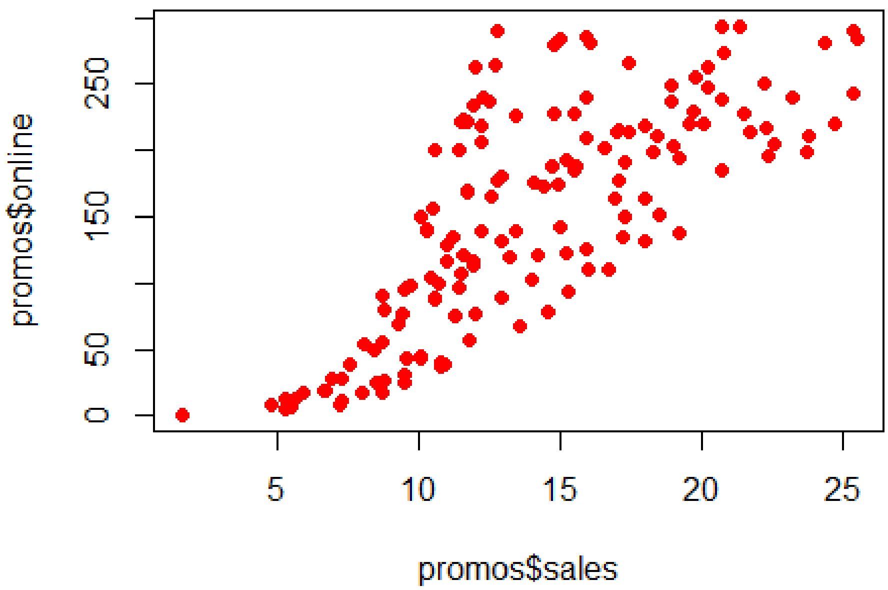
```r
check with the plots
hist(promos$sales,probability=T,col=c("steelblue", "red"))
lines(density(promos$sales),col="black",lwd=3)
```



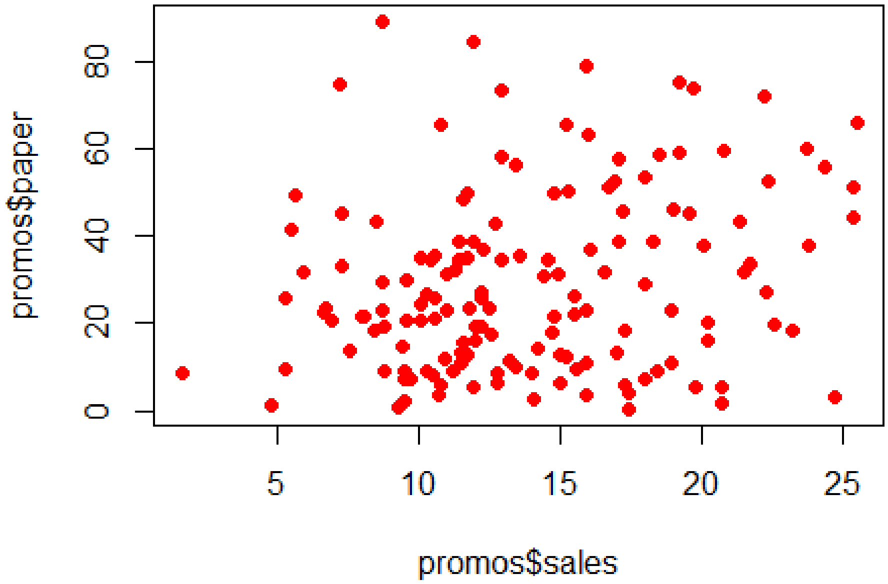

```r
#The histogram shows that the sales is normally distributed

boxplot(promos$sales)
```



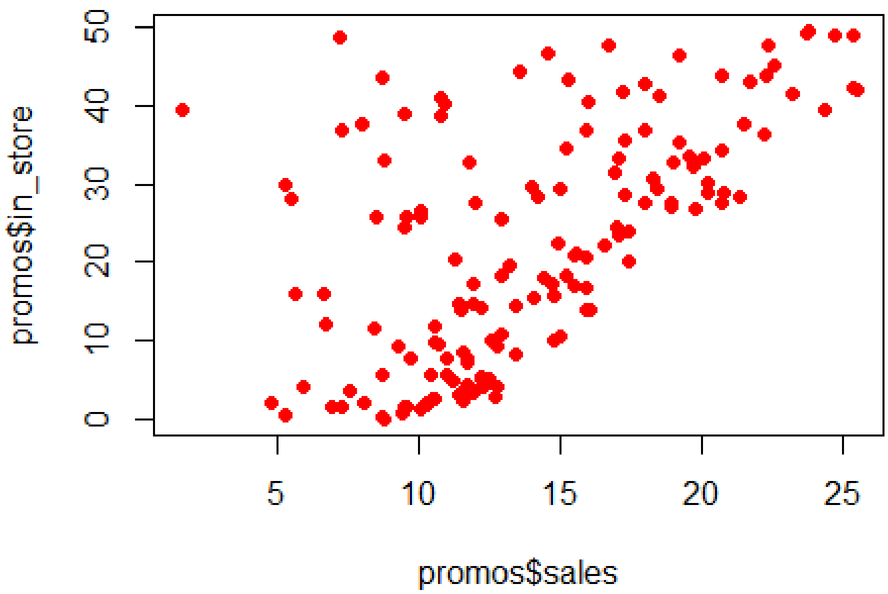

```r
#Boxplot shows that there are no outliers in the variable accuracy but the
sales is slightly left skewed

#Hence from all above the sales is roughly normaly distributed

#step2- Checking linear relationship between IVs and DVS
cor.test(promos$sales, promos$online)

##
##  Pearson's product-moment correlation
##
## data:  promos$sales and promos$online
## t = 14.344, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6862556 0.8223890
## sample estimates:
##       cor
## 0.7626416

# cor
# 0.7626416
#High  positive correlation
plot(promos$sales, promos$online, pch=16, col="red")
```




```
cor.test(promos$sales, promos$paper)

##
##  Pearson's product-moment correlation
##
## data:  promos$sales and promos$paper
## t = 2.5154, df = 148, p-value = 0.01296
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0436345 0.3513407
## sample estimates:
##       cor
## 0.2024801

# cor
# 0.20
#There is no correlation of paper with sales
plot(promos$sales, promos$paper, pch=16, col="red")
```




```
cor.test(promos$sales, promos$in_store)

##
##  Pearson's product-moment correlation
##
## data:  promos$sales and promos$in_store
## t = 8.8381, df = 148, p-value = 2.636e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4719505 0.6836244
## sample estimates:
##       cor
## 0.5877567

# cor
# 0.5877
#There is positive correlation of in_store with sales
plot(promos$sales, promos$in_store, pch=16, col="red")
```




```
#Step 3- multiple regression model

model1<-lm(sales ~ online+paper+in_store+region, data=promos)
model1

##
## Call:
## lm(formula = sales ~ online + paper + in_store + region, data = promos)
##
## Coefficients:
##   (Intercept)         online          paper       in_store  regionMidwest
##       3.01513        0.04441       -0.00112        0.18785       -0.28232
##   regionSouth     regionWest
##       0.46457        0.47778

summary(model1)

##
## Call:
## lm(formula = sales ~ online + paper + in_store + region, data = promos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5930 -0.6353  0.2350  1.0313  2.7123
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    3.015128   0.506528   5.953 1.95e-08 ***
## online         0.044408   0.001656  26.820  < 2e-16 ***
## paper         -0.001120   0.007042  -0.159    0.874
## in_store       0.187849   0.009686  19.393  < 2e-16 ***
## regionMidwest -0.282320   0.404105  -0.699    0.486
## regionSouth    0.464575   0.421249   1.103    0.272
## regionWest     0.477785   0.448051   1.066    0.288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.684 on 143 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8885
## F-statistic: 198.9 on 6 and 143 DF,  p-value: < 2.2e-16

#Findings from model1
#The adjusted R square is 0.8885 i.e 88.85% of variance in sales is explained
by the model.
#The variables online and in_store are significant with sales as have p-
values less than 0.05
#One unit increase in online increase the sales by 0.044408  units.
#One unit increase in in_store increases the sales by 0.187849 units

#Since this model has some insignificat variables like region and paper we
have to remove them
```

Q.2)Build a new model that is only based on the Independent Variables that are “good” predictors. Explain your findings

```
#Since this model developed in question 1 has some insignificat variables
(from summary of model checking p values)like region and paper we have to
remove them

#lets select variables online and in_store in model as they are significant
i.e have p-value less than 0.05 in summary(model1)
model2<-lm(sales ~ online+in_store, data=promos)
model2

##
## Call:
## lm(formula = sales ~ online + in_store, data = promos)
##
## Coefficients:
## (Intercept)       online     in_store
##     3.15162      0.04435      0.18645

summary(model2)

##
## Call:
## lm(formula = sales ~ online + in_store, data = promos)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.9663 -0.6180  0.2886  1.0737  2.7107 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.151619   0.346661   9.091 6.22e-16 ***
## online      0.044353   0.001659  26.731  < 2e-16 ***
## in_store    0.186455   0.009280  20.092  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.697 on 147 degrees of freedom
## Multiple R-squared:  0.8883, Adjusted R-squared:  0.8868 
## F-statistic: 584.6 on 2 and 147 DF,  p-value: < 2.2e-16
```

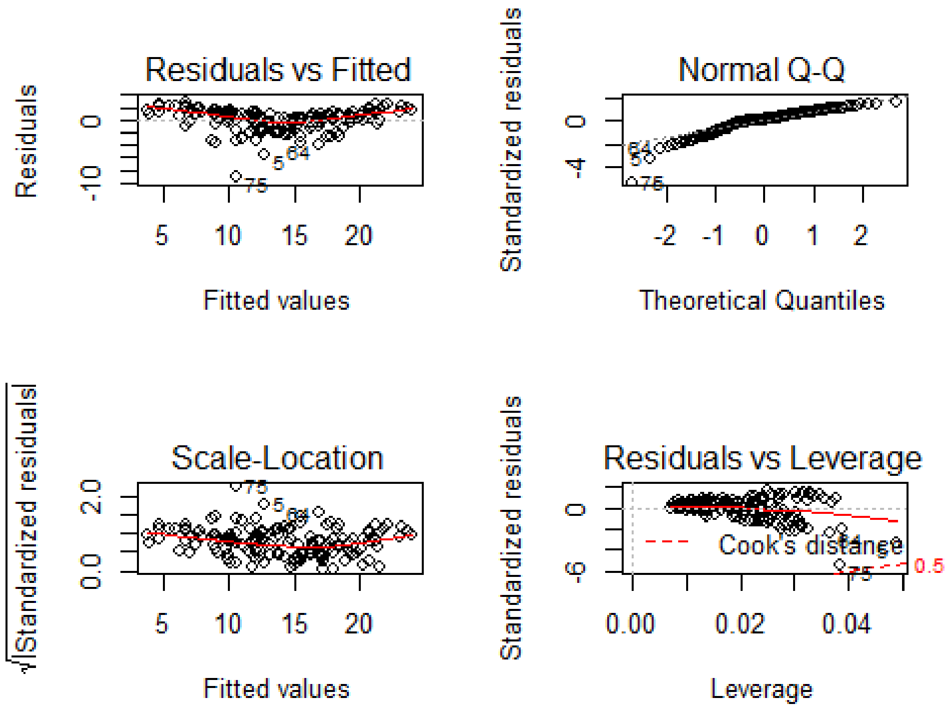
```
summary(model2)$adj.r.squared-summary(model1)$adj.r.squared
```

```
## [1] -0.001688289
```

```
#The adjusted R-square value is actually decreasing by 0.1% but we still have
to remove the variables region and paper as they are not significant with the
sales and are violating the assumptions of the regression model

#Findings from model2
#The adjusted R square is 0.8868
#The variables online and in_store are significant with sales as have p-
values less than 0.05
#One unit increase in online increase the sales by 0.044353  units.
#One unit increase in in_store increases the sales by 0.186455 units
```

Q.3) Perform a regression diagnosis. What concerns do you have with your model, and explain how would you address them?

```
#Run diagnostic tests
par(mfrow=c(2,2))
plot(model2)
```




```
outliers= c(5,64,75)
promos1<-promos[-outliers,]
model2.1<-lm(sales ~ online+in_store, data=promos1)

summary(model2.1)

##
## Call:
## lm(formula = sales ~ online + in_store, data = promos1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1096 -0.6367  0.1738  0.9271  2.6262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.362951   0.291380   11.54   <2e-16 ***
## online      0.042607   0.001437   29.66   <2e-16 ***
## in_store    0.194534   0.007953   24.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.419 on 144 degrees of freedom
## Multiple R-squared:  0.9191, Adjusted R-squared:  0.918
## F-statistic: 817.8 on 2 and 144 DF,  p-value: < 2.2e-16

summary(model2.1)$adj.r.squared-summary(model2)$adj.r.squared
```

```
## [1] 0.03115488
```

```r
#The adjusted R-square value is actually increased by 3% after removing the outliers the outliers are removed as in QQ plotit was not showing the approx diagonal line and hence was less normal due to the outliers
#After removing the outliers QQ plot can be  considered to be normal



car::vif(model2.1)
```

```
##   online in_store
## 1.009669 1.009669
```

```r
#there is no problem of multicollinearity here. The multicollinearity is checked because while developing regression model we assume that there is no collinearity among the independent variables.
#If the vif is greater than 10 there is multicollinearity and independent variables are not completely independent which affects the model. Hence multicollinearity is tested in diagnostics.



#Findings from model
#The adjusted R square is increased to 0.918
#The variables online and instore are significant woith sales
#One unit increase in online increase the sales by 0.042 units.
#One unit increase in in_store increases the sales by 0.1945 units
```