R\_P4\_2048046

R Akhilandeshwari

2/5/2021

setwd("C:/Desktop/R/P4")  
getwd()

## [1] "C:/Desktop/R/P4"

**1.Import the dataset data\_marketing\_budget\_mo12 and do the exploratory data analysis.**

mark=read.csv('data-marketing-budget.csv')  
mark

## Month Spend Sales  
## 1 1 1000 9914  
## 2 2 4000 40487  
## 3 3 5000 54324  
## 4 4 4500 50044  
## 5 5 3000 34719  
## 6 6 4000 42551  
## 7 7 9000 94871  
## 8 8 11000 118914  
## 9 9 15000 158484  
## 10 10 12000 131348  
## 11 11 7000 78504  
## 12 12 3000 36284

**Finding the null values in the data**

complete.cases(mark)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

**This implies we donnot have null values in the given data**

**Knowing abt the summary of the data**

summary(mark)

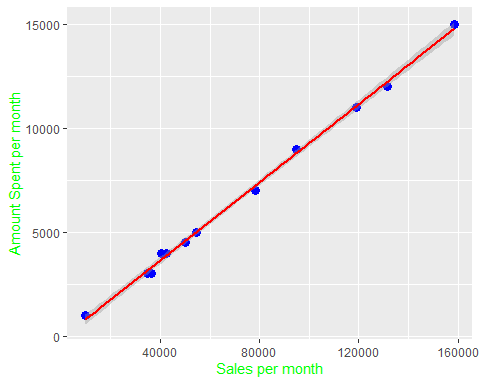
## Month Spend Sales   
## Min. : 1.00 Min. : 1000 Min. : 9914   
## 1st Qu.: 3.75 1st Qu.: 3750 1st Qu.: 39436   
## Median : 6.50 Median : 4750 Median : 52184   
## Mean : 6.50 Mean : 6542 Mean : 70870   
## 3rd Qu.: 9.25 3rd Qu.: 9500 3rd Qu.:100882   
## Max. :12.00 Max. :15000 Max. :158484

**In marketing We spend minimum Rs.1000 with the minimum sales of Rs.9,914** **In marketing We spend maximum Rs.15000 with the minimum sales of Rs.1,58,484**  **The average amount spent is Rs.6542 with an average sales of Rs.70,870**

1. **Use Scatter Plot To Visualize The Relationship**

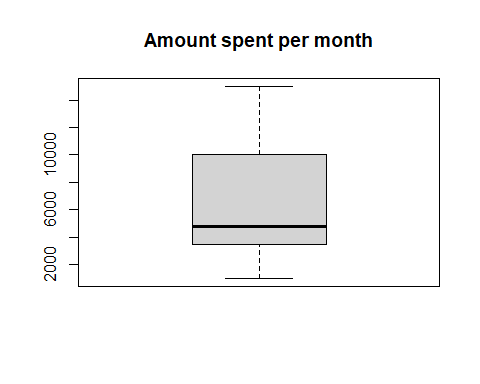
library(ggplot2)  
r<-ggplot(mark,aes(Sales, y=Spend)) + geom\_point(color="blue",size=3)  
r+geom\_smooth(method=lm,size=.8,color="red")+labs(y="Amount Spent per month",x="Sales per month")+theme(axis.title.x = element\_text(colour = "green"),axis.title.y = element\_text(colour = "green"))

## `geom\_smooth()` using formula 'y ~ x'

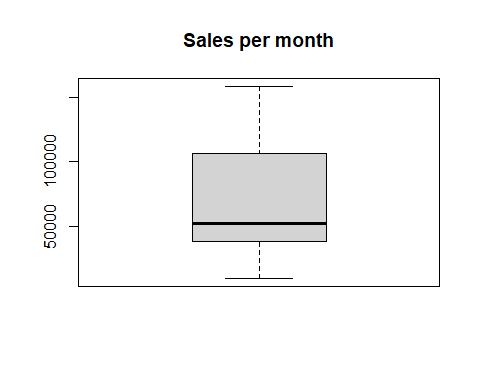


**3.Using BoxPlot To Check For Outliers**

boxplot(mark$Spend,main="Amount spent per month")



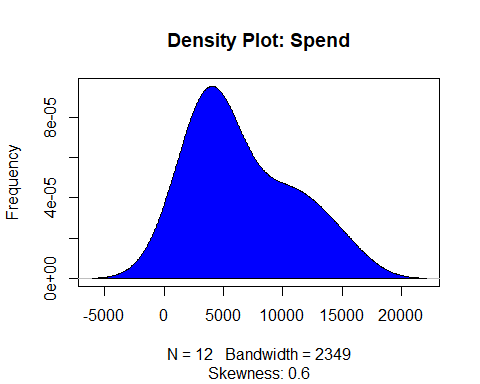
boxplot(mark$Sales,main="Sales per month")



**There are no outliers in our data**

1. **Using Density Plot To Check If Response Variable Is Close To Normal.**

library(e1071) # for skewness function  
#par(mfrow=c(1,2)) # divide graph area in 2 columns  
plot(density(mark$Spend), main="Density Plot: Spend", ylab="Frequency",sub=paste("Skewness:",round(e1071::skewness(mark$Spend),2)))   
polygon(density(mark$Spend), col="blue")



**The skewness of the graph is 0.6 which is greater than 0. So, the data is positively skewed and it follows Platykurtic distribution.**

1. **Check the Correlation Analysis**

cor(mark$Spend,mark$Sales)

## [1] 0.9988322

**The amount spent and sales are highly positively correlated**

1. **Build the Linear Regression Model**

model=lm(Spend~Sales,mark)  
model

##   
## Call:  
## lm(formula = Spend ~ Sales, data = mark)  
##   
## Coefficients:  
## (Intercept) Sales   
## -114.67027 0.09392

**The regression line with the data is of the form Spend = 0.093sales -114.67. The per unit spend is increased by 0.09392times in the sales.**

1. **Using p-value Check for Statistical Significance**

**P-value Significance**

**Null Hypothesis: The model is not significant**

**Alternative Hypothesis: The model is significant**

sum\_report=summary(model)  
sum\_report

##   
## Call:  
## lm(formula = Spend ~ Sales, data = mark)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -293.22 -165.15 -20.82 188.67 312.02   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.147e+02 1.196e+02 -0.959 0.36   
## Sales 9.392e-02 1.437e-03 65.378 1.71e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 217.5 on 10 degrees of freedom  
## Multiple R-squared: 0.9977, Adjusted R-squared: 0.9974   
## F-statistic: 4274 on 1 and 10 DF, p-value: 1.707e-14

**ModelCoefficients**

modelCoeffs <- sum\_report$coefficients   
print(modelCoeffs)

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -114.67027453 119.62169406 -0.9586077 3.603591e-01  
## Sales 0.09392275 0.00143662 65.3776136 1.706704e-14

**Estimated Amount to be spent**

estimate=modelCoeffs["Sales","Estimate"]  
estimate

## [1] 0.09392275

**Standard\_Error**

se=modelCoeffs["Sales","Std. Error"]  
se

## [1] 0.00143662

**t-statistic**

t=estimate/se  
t

## [1] 65.37761

p<-2\*pt(-abs(t),df=length(mark$Sales)-1)  
p

## [1] 1.329412e-15

**The p-value is less than 0.05. So, we reject null hypothesis. Therefore, the model is significant**

**8.Capture the summary of the linear model**

sum\_report

##   
## Call:  
## lm(formula = Spend ~ Sales, data = mark)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -293.22 -165.15 -20.82 188.67 312.02   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.147e+02 1.196e+02 -0.959 0.36   
## Sales 9.392e-02 1.437e-03 65.378 1.71e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 217.5 on 10 degrees of freedom  
## Multiple R-squared: 0.9977, Adjusted R-squared: 0.9974   
## F-statistic: 4274 on 1 and 10 DF, p-value: 1.707e-14

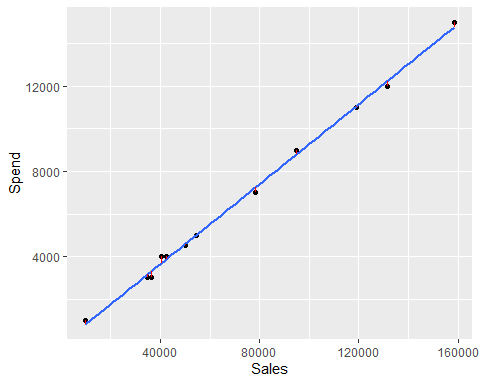
1. **Also perform the Linear Diagnostics for the given data set(Hint:plot(lmmodel))**

#broom: creates a tidy data frame from statistical test results  
library(broom)   
model.diag.metrics <- augment(model)  
head(model.diag.metrics)

## # A tibble: 6 x 8  
## Spend Sales .fitted .resid .std.resid .hat .sigma .cooksd  
## <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1000 9914 816. 184. 0.971 0.245 218. 0.153   
## 2 4000 40487 3688. 312. 1.53 0.124 201. 0.166   
## 3 5000 54324 4988. 12.4 0.0600 0.0953 229. 0.000189  
## 4 4500 50044 4586. -85.6 -0.415 0.102 227. 0.00982   
## 5 3000 34719 3146. -146. -0.725 0.140 223. 0.0429   
## 6 4000 42551 3882. 118. 0.578 0.118 225. 0.0225

ggplot(model.diag.metrics, aes(Sales,Spend)) +  
 geom\_point() +  
 stat\_smooth(method = lm, se = FALSE) +  
 geom\_segment(aes(xend = Sales, yend = .fitted), color = "red", size = 0.3)

## `geom\_smooth()` using formula 'y ~ x'

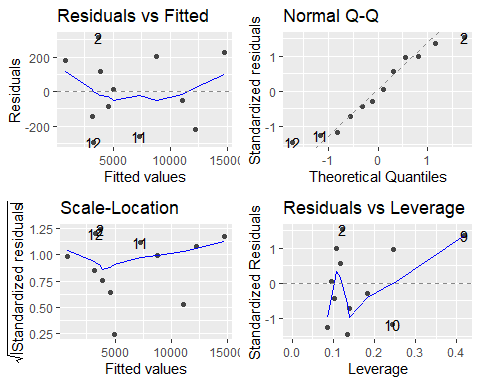


**We can notice the residual error is very small**

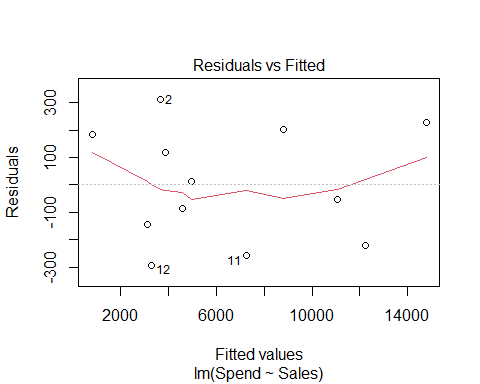
##Create the diagnostic plots using ggfortify

library(ggfortify)  
autoplot(model)

## Warning: `arrange\_()` is deprecated as of dplyr 0.7.0.  
## Please use `arrange()` instead.  
## See vignette('programming') for more help  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

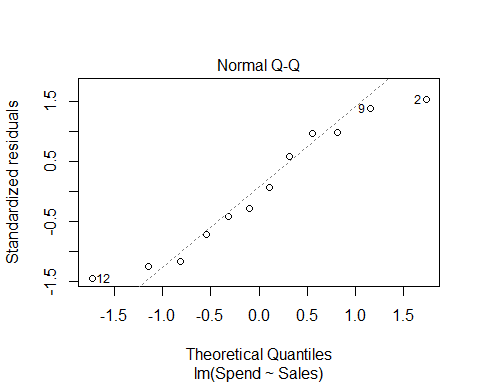


#Residual Vs Fitted Value  
plot(model, 1)



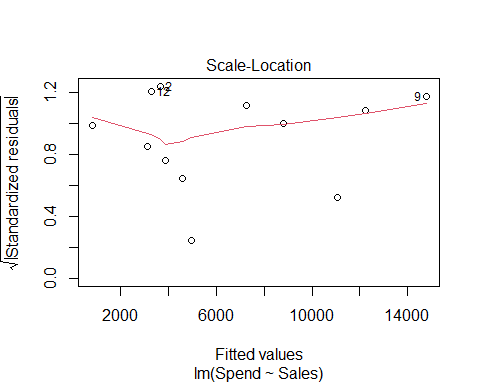
As the data is spread across without following any pattern, we can say that the regression model is best fit.

#Normal Q-Q plot  
plot(model, 2)



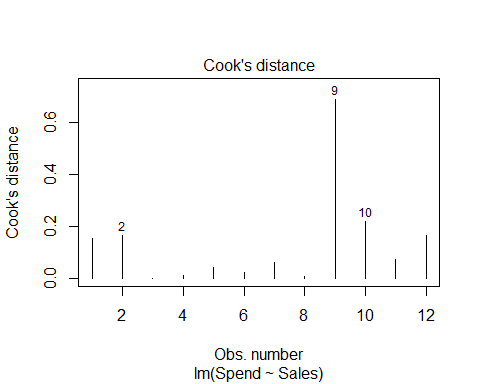
**The sales on the spend is linearly related.**

#Scale location @Heterosadasticity or homosedasticity  
plot(model, 3)

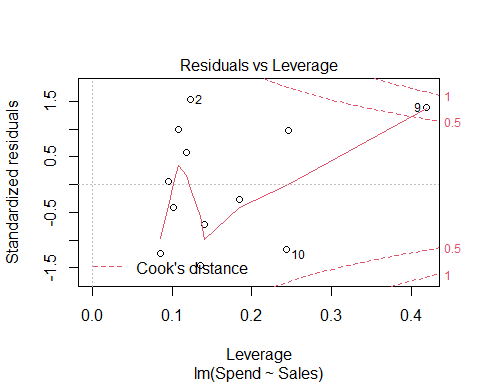


**As the residuals are equally distributed and the red line is horizontal. Hence, the variance is equal for the data i.e., data is homosedasticity.**

#Finding the Cook's distance  
plot(model, 4)



#Residual Vs Leverage  
plot(model, 5)



**By removing the 9th observation we may have great impact on the residual error.**

1. **Create the training and test data (70:30)**

set.seed(100)   
split\_dummy <- sample(c(rep(0, 0.7 \* nrow(mark)), rep(1, 0.3 \* nrow(mark))))  
data\_train <- mark[split\_dummy==0,]   
data\_test<-mark[split\_dummy==1, ]

1. **Fit the model on training data and predict sales on test data**

lin\_mod=lm(Spend~Sales,data=data\_train)  
pred=predict(lin\_mod,data\_test)  
pred

##   
## 1 9 11 12

935.2047 14642.4143 7263.3834 3368.1260

library(caret)

cat("MAE",MAE(pred, actual),"\n")

cat("RMSE",RMSE(pred, actual),"\n")

# MAE 263.4726

# RMSE 290.2385

**The linear model build is of the form Spend=0.9226Sales+20.52966. The per unit spend is increased by 0.922times in the sales.**  With the Root mean square error as 290.2385 and the Mean Absolute Error as 263.4726.

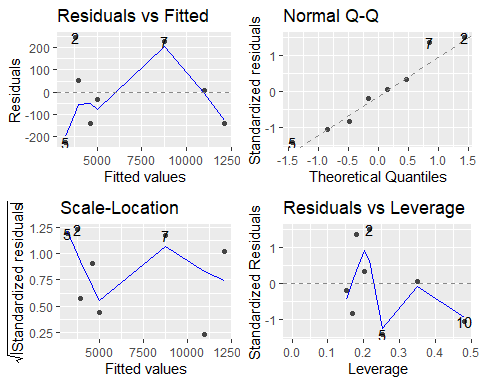
**11.Review the diagnostic measures**

#sum\_report  
summary(lin\_mod)

##   
## Call:  
## lm(formula = Spend ~ Sales, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -223.74 -137.93 -12.08 96.90 244.10   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.053e+01 1.442e+02 0.142 0.891   
## Sales 9.226e-02 1.815e-03 50.833 3.89e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 184 on 6 degrees of freedom  
## Multiple R-squared: 0.9977, Adjusted R-squared: 0.9973   
## F-statistic: 2584 on 1 and 6 DF, p-value: 3.888e-09

**The R2 is 0.997 which is good indicator of the fitted model. The linear regression model is the best fit for the given data.Moreover, the p-value high significance relationship between the spend and sale. And even the F-statistic value is 2584 for 1 which implies, a perfect good relationship between the spend and sales.**

autoplot(lin\_mod)



**It is similar to the previous model built. Implies the Spend is positively and highly related with the Sales, and they maintain a good positive relation. And there is only one point which have significant effect on the variables with respect to the residuals Vs Leverage plot we can indicate this. As the point in the plane are scattered throughout in horizontal plan(Scale-Location), the variance is less.**