PREDICTION OF SALARY OF AN EMPLOYEE OF A COMPANY BASED ON HIS EXPERIENCE

**INTRODUCTION**

***Objective:***

Here our main objective is to choose a suitable data set for simple linear regression and analyse it with the following sub-objectives and prepare a report based on the analysis,

1. To add some outlying observations to the chosen data set and fit a simple linear regression in the presence of outliers with respect to both OLS and LAD regression.
2. To comment on the fit of the regression line using both the methods in part 1.
3. To perform the residual analysis based on the best fitted model from part 2.

***Data Description:***

Here, in this problem we have taken a dataset that consists of salary distribution of an employee in a company based on years of experience they have. The dataset consists of records of 30 employees.

The variables taken into consideration are described as follows,

* ***Years of experience*** of an employee is denoted by ***y, dependent variable.***
* ***Salary*** of an employee is denoted by ***x, independent variable.***

*#Importing the Salary dataset from current working directory.*  
**library**(readr)

## Warning: package 'readr' was built under R version 4.0.3

SalaryData <- **read\_csv**("SalaryData.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## Years\_Experience = col\_double(),  
## Salary = col\_double()  
## )

*#Since we observed that the dataset is not a dataframe we are converting it to a dataframe.*  
SalaryData=**data.frame**(SalaryData)  
  
*#Obtaining first few records of the dataset.*  
**head**(SalaryData)

## Years\_Experience Salary  
## 1 1.1 39343  
## 2 1.3 46205  
## 3 1.5 37731  
## 4 2.0 43525  
## 5 2.2 39891  
## 6 2.9 56642

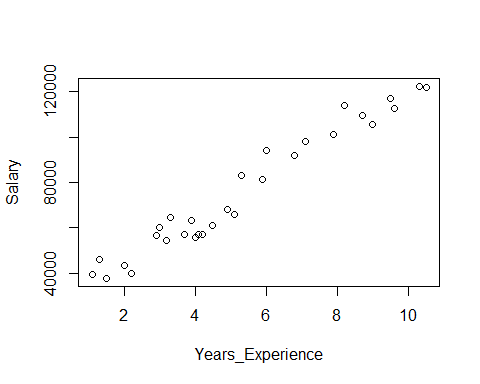
**ANALYSIS**

*#We use attach function to access the variables present in the dataframe without calling the dataframe.*  
**attach**(SalaryData)

*#Obtaining the length of the Salary dataset.*  
**length**(Salary)

## [1] 30

*#Obtaining the scatter plot of Years of Experience versus experience of employees.*  
**plot**(Years\_Experience,Salary)





From figure 1 we observe that the years of experience and salary of an employee are positively linearly related that is more the years of experience more salary the employee will get.

1. **To add some outlying observations to the chosen data set and fit a simple linear regression in the presence of outliers with respect to both OLS and LAD regression.**

*#Assigning our salary dataset to another variable 'SalaryData1'.*  
SalaryData1<-SalaryData[1**:**30,]

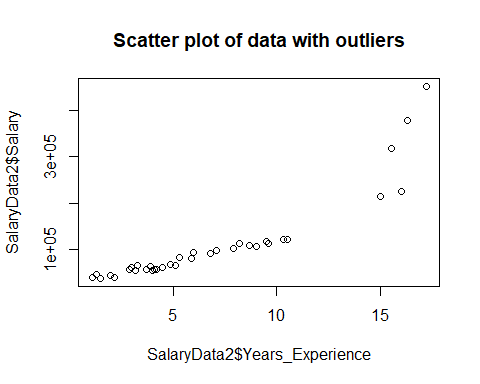
*#Creating a dataframe of outliers.*  
outliers<-**data.frame**(Years\_Experience=**c**(15,15.5,16,16.3,17.2),Salary=**c**(214693,317514,225679,378412,451121))  
outliers

## Years\_Experience Salary  
## 1 15.0 214693  
## 2 15.5 317514  
## 3 16.0 225679  
## 4 16.3 378412  
## 5 17.2 451121

*#Merging the outliers and original data to obtain a dataset of years of experience and salary with outliers.*  
SalaryData2<-**rbind**(SalaryData1,outliers)  
**head**(SalaryData2)

## Years\_Experience Salary  
## 1 1.1 39343  
## 2 1.3 46205  
## 3 1.5 37731  
## 4 2.0 43525  
## 5 2.2 39891  
## 6 2.9 56642

*#Obtaining the plot of the dataset with outliers.*  
**plot**(SalaryData2**$**Years\_Experience,SalaryData2**$**Salary,main="Scatter plot of data with outliers")





*#The par( ) function helps us include the option mfrow=c(nrows, ncols) to create a matrix of nrows x ncols plots that are filled in by row.*  
**par**(mfrow=**c**(1,2))  
  
*#Fitting a simple linear regression with respect to OLS method to the data with outliers .*  
ols\_reg<-**lm**(SalaryData2**$**Salary**~**SalaryData2**$**Years\_Experience,data=SalaryData2)  
**plot**(SalaryData2**$**Years\_Experience,SalaryData2**$**Salary,main="OLS method") **+**  
**abline**(**lm**(SalaryData2**$**Salary**~**SalaryData2**$**Years\_Experience,data=SalaryData2),col="yellow",lwd=3,lty=3)

## integer(0)

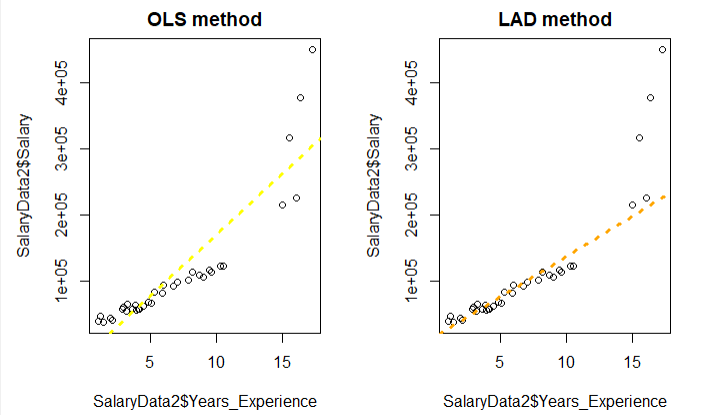
*#Loading the package 'L1pack'.*  
**library**(L1pack)

## Warning: package 'L1pack' was built under R version 4.0.3

## Loading required package: fastmatrix

## Warning: package 'fastmatrix' was built under R version 4.0.3

*#Fitting a simple linear regression with respect to LAD method to the data with outliers.*  
lad\_reg<-**lad**(SalaryData2**$**Salary**~**SalaryData2**$**Years\_Experience,data=SalaryData2,method="BR")  
**plot**(SalaryData2**$**Years\_Experience,SalaryData2**$**Salary,main="LAD method") **+**  
**abline**(**lad**(SalaryData2**$**Salary**~**SalaryData2**$**Years\_Experience,data=SalaryData2),col="orange",lwd=3,lty=3)





From the Figure 3 we observe that when our data has outliers then on fitting the regression line to the datasset using least absolute deviation (LAD) method gives the best fit as compared to ordinary least square method.

1. **To comment on the fit of the regression line using both the methods in part 1.**

From part 1 we observe that LAD method gives better fit than OLS method when we have outliers in our dataset. Hence we prefer method of least absolute deviation over the method of OLS to fit the simple linear regression since it is suitable even if we have outliers in our dataset.

1. **To perform the residual analysis based on the best fitted model obtained from part 2 i.e. based on the regression model obtained by the method of LAD.**

The following are the assumtions regarding the fitted model,

1. The relationship beetween y and x is linear.
2. Errors have zero mean.
3. Assumption of homoscedasticity, i.e. the errors have constant variance.
4. Errors are uncorrelated.
5. Errors are normally distributed random variables.

*#To check if x i.e. salary and y i.e. years of experience are linearly related.*

From Figure 2 we observe the years of experience and salary of an employee positively linearly related. Hence this assumption is satisfied.

*#To check if errors have mean zero.*  
**mean**(**resid**(lad\_reg))

## [1] 13338.08

From the above calculation we observe that the residuals have mean 13338.08 which is not equal to zero, hence the assumption of zero mean is also violated.

*#To check the assumption of homoscedasticity.*  
  
*#Loading the package 'lmtest' required to perform studentized Breusch-Pagan test.*  
**library**(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

*#Performing studentized Breusch-Pagan test to check the assumption of homoscedasticy.*  
**bptest**(lad\_reg)

##   
## studentized Breusch-Pagan test  
##   
## data: lad\_reg  
## BP = 13.978, df = 1, p-value = 0.000185

From the studentized Breusch-Pagan test we observe that p value=0.000185 which is less than 0.05, hence we reject the null hypothesis and conclude that errors are having non constant variance.

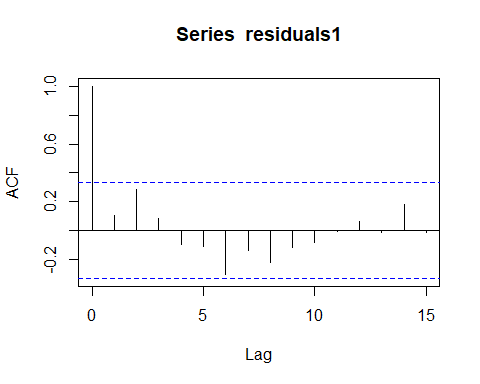
Since the errors are having non constant variance we try to stabilize the variance and check for the assumption of homoscedasticity again.

*#Stabilizing the variance consists of the following steps,*  
  
*#Step 1 - Transforming the response variable.*  
y\_trans=**log**(SalaryData2**$**Salary)  
  
*#Step 2 - Fit and validate the model in the transformed variable.*  
lad\_reg1<-**lad**(y\_trans**~**SalaryData2**$**Years\_Experience,data=SalaryData2,method="BR")  
  
*#Performing studentized Breusch-Pagan test to check the assumption of homoscedasticy.*  
**bptest**(lad\_reg1)

##   
## studentized Breusch-Pagan test  
##   
## data: lad\_reg1  
## BP = 14.308, df = 1, p-value = 0.0001552

From the above studentized Breusch-Pagan test we observe that the pvalue=0.0001552 which is less than 0.05 hence we reject the null hypothesis and conclude that the errors have non constant variance. Hence the assumption of homoscedasticity is violated.

*#To check if the errors are uncorrelated.*  
  
*#Obtaining the residuals of regression model with transformed response variable.*  
residuals1=**resid**(lad\_reg1)  
  
*#Obtaining the acf plot to check if the residuals uncorrelated i.e. to check if there is no autocorrelation in our residual series.*  
**acf**(residuals1)





From the Figure 4 we observe that all the lags are inside the threshold line therefore we can conclude that the errors are uncorrelated, hence the assumption is validated to be true.

*#To check if the errors are normally distributed.*  
  
*#Now, performing Shapiro-Wilk normality test to check if the residuals are normaally distributed or not.*  
**shapiro.test**(**resid**(lad\_reg1))

##   
## Shapiro-Wilk normality test  
##   
## data: resid(lad\_reg1)  
## W = 0.98481, p-value = 0.8995

From the shapiro-Wilk normality test we observe that pvalue = 0.8995 which is greater than 0.05 hence we fail to reject the null hypothesis and conclude that the errors are normally distributed. Hence the assumption of normality is also satisfied.

Hence from the above residual analysis we observe that all the assumptions except the assumption of zero mean and constant variance are satisfied.

**CONCLUSION**

From the above analysis we conclude that when we have outliers in our dataset then method of least absolute deviation give the best fit as compared to method of OLS.

But further on performing the residual analysis we observe that the all the assumptions about errors are not satisfied and therefore we cannot claim that it is a good model but we know that the method of LAD gives the best fit therefore we may assume that the dataset that is taken into consideration is not good.