Residual Analysis of fitted regression model of Income vs expenditure for 30 households in India

**INTRODUCTION:**

***Objective:*** Here in this practical assignment our main objective is to plot a appropriate residual plot for our fitted regression model of our dataset IncomeData to check the assumption of homoscedasticity and comment on it, we also want to test for hemoscedasticity using a studentized Breusch-Pagan test. We are also interested in explaining the remedial measure that we would take if our assumption of homoscedasticity is violated with the help of an analysis. We also want to check if the residuals are uncorrelated and validate it using Durbin-Watson test.

***Data Description:*** Here we are considering the dataset consisting of annual consumption and disposable income for 30 households in India.

Here in this data we have,

Annual consumption i.e. **expenditure** of 30 households as **dependent variable.**

Disposable **income** of 30 households as **independent variable.**

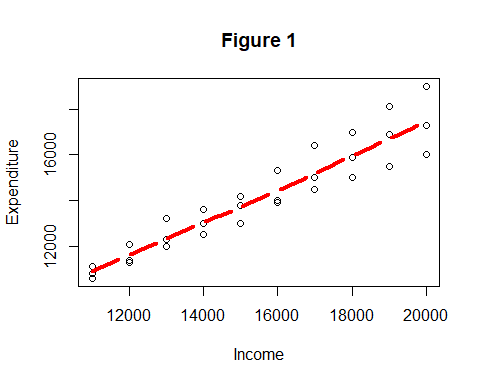
*#Loading the package 'readxl' required to load the dataset 'IncomData'.*  
**library**(readxl)  
  
*#Loading the dataset 'IncomeData'.*  
IncomeData <- **read\_excel**("IncomeData.xlsx")  
  
*#Obtaining the first few records of our dataset 'IncomeData'.*  
**head**(IncomeData)

## # A tibble: 6 x 3  
## S.No. Expenditure Income  
## <dbl> <dbl> <dbl>  
## 1 1 10600 11000  
## 2 2 11400 12000  
## 3 3 12300 13000  
## 4 4 13000 14000  
## 5 5 13800 15000  
## 6 6 13900 16000

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

**ANALYSIS**

*#Obtaining the scatter plot for the two variables Income and Expenditure to check whether there exist a linear relationship between the variables.*  
**scatter.smooth**(Income,Expenditure,main="Figure 1", lpars = **list**(col="red", lwd = 4, lty = 5))



*#Fitting a simple linear regression model to our dataset 'IncomeData'.*  
reg=**lm**(Expenditure**~**Income)  
reg

##   
## Call:  
## lm(formula = Expenditure ~ Income)  
##   
## Coefficients:  
## (Intercept) Income   
## 2799.3939 0.7327

*#Obtaining the fitted values of the model.*  
fit\_values=**fitted.values**(reg)  
fit\_values

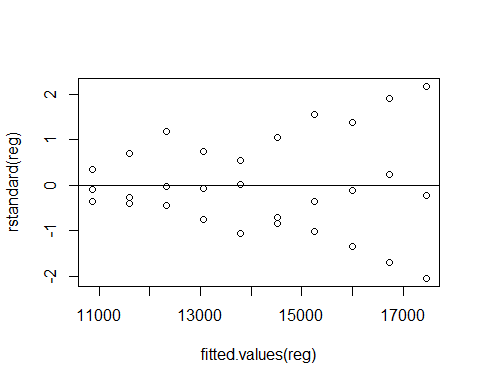
## 1 2 3 4 5 6 7 8   
## 10859.39 11592.12 12324.85 13057.58 13790.30 14523.03 15255.76 15988.48   
## 9 10 11 12 13 14 15 16   
## 16721.21 17453.94 10859.39 11592.12 12324.85 13057.58 13790.30 14523.03   
## 17 18 19 20 21 22 23 24   
## 15255.76 15988.48 16721.21 17453.94 10859.39 11592.12 12324.85 13057.58   
## 25 26 27 28 29 30   
## 13790.30 14523.03 15255.76 15988.48 16721.21 17453.94

*#Obtaining the residuals that is residuals=observed value - fitted values.*  
resid=**resid**(reg)  
resid

## 1 2 3 4 5 6   
## -259.39394 -192.12121 -24.84848 -57.57576 9.69697 -623.03030   
## 7 8 9 10 11 12   
## -255.75758 -88.48485 178.78788 -153.93939 -59.39394 -292.12121   
## 13 14 15 16 17 18   
## -324.84848 -557.57576 -790.30303 -523.03030 -755.75758 -988.48485   
## 19 20 21 22 23 24   
## -1221.21212 -1453.93939 240.60606 507.87879 875.15152 542.42424   
## 25 26 27 28 29 30   
## 409.69697 776.96970 1144.24242 1011.51515 1378.78788 1546.06061

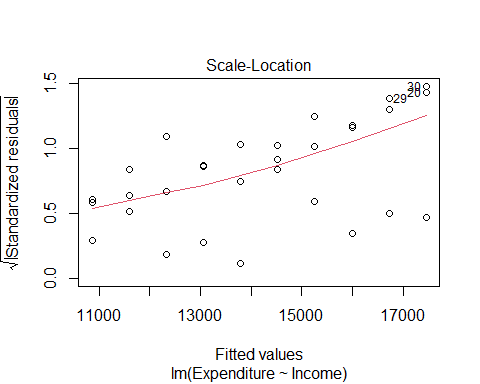
To plot appropriate residual plot to check the assumption of homoscedasticity and comment on it and also test for hemoscedasticity using a suitable statistical test.

*#Plot of fitted values against residuals,to check the assumption of homoscedasticity.*  
**plot**(**fitted.values**(reg),**rstandard**(reg))  
**abline**(0,0)



***Interpretation:*** Since from the plot we observe the upward open funnel shape which indicates non constant variance which violates the assumption of homoscedasticity.

*#Obtaining the scale-location graph.*  
**plot**(reg)



***Interpretation:*** Since from the scale-location plot we observe that the points are not evenly spread above and below the horizontal line which indicates non constant variance which violates the assumption of homoscedasticity.

*#Examining the assumption of homoscedasticity with the help of bptest.*  
**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

**bptest**(reg)

##   
## studentized Breusch-Pagan test  
##   
## data: reg  
## BP = 11.373, df = 1, p-value = 0.000745

***Interpretation:*** Since p value 0.000745 which is less than 0.05 hence we reject the null hypothesis and conclude that errors are having non constant variance that means the assumption of homoscedasticity is violated.

To explain remedial measures will be taken if the constant variance assumption is violated and explain it with the complete analysis.

Since from the above test and plots we observe that variance of error series is not constant we stabilize the variable by,

1. Transforming the response variable. We do it using either of the following transformation:

* T=log(y)
* T=sqrt(y)
* T=1/y

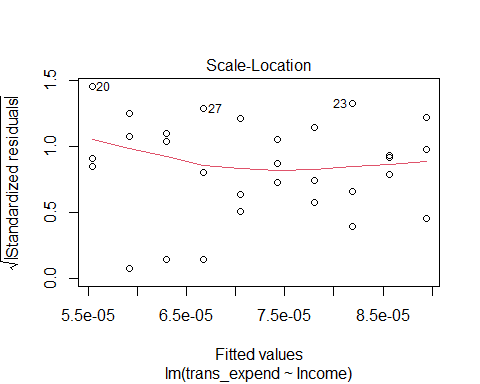
1. Fit and validate the model using the transformed variables.
2. Finally, transform the predicted values back into the original units.

*#Here we are tranforming our response variable that is expenditure which is required to stabilize the model*  
trans\_expend=1**/**Expenditure  
trans\_expend

## [1] 9.433962e-05 8.771930e-05 8.130081e-05 7.692308e-05 7.246377e-05  
## [6] 7.194245e-05 6.666667e-05 6.289308e-05 5.917160e-05 5.780347e-05  
## [11] 9.259259e-05 8.849558e-05 8.333333e-05 8.000000e-05 7.692308e-05  
## [16] 7.142857e-05 6.896552e-05 6.666667e-05 6.451613e-05 6.250000e-05  
## [21] 9.009009e-05 8.264463e-05 7.575758e-05 7.352941e-05 7.042254e-05  
## [26] 6.535948e-05 6.097561e-05 5.882353e-05 5.524862e-05 5.263158e-05

*#Fitting a new regression model after transforming the response variable.*  
reg\_new=**lm**(trans\_expend**~**Income)

*#Obtaining the scale location plot to check the assumption of homoscedasticity after transforming the response variable.*  
**plot**(reg\_new)



***Interpretation:*** Since from the scale-location plot we observe that the points are evenly spread above and below the horizontal line which indicates constant variance which proves the assumption of homoscedasticity to be true.

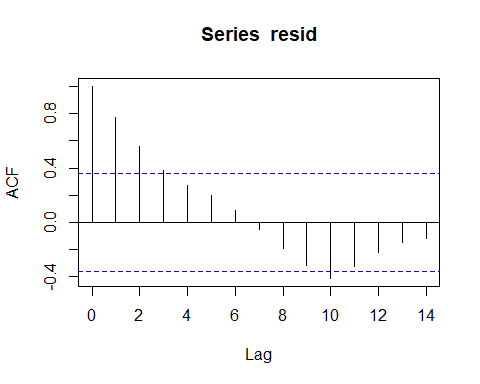
*#Performing bptest to check the assumption of homoscedasticity after transforming the response variable.*  
**bptest**(reg\_new)

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_new  
## BP = 1.0701, df = 1, p-value = 0.3009

***Interpretation:*** Since p value 0.3009 which is greater than 0.05 hence we fail to reject the null hypothesis and conclude that errors are having constant variance that means the assumption of homoscedasticity is proved to be true.

To check if the residuals uncorrelated and validate our answer with a suitable test procedure.

*#Obtaining the acf plot to check if the residuals uncorrelated i.e. to check if there is no autocorrelation in our residual series.*  
**acf**(resid)



***Interpretation:*** Since few lag values crosses the threshhold line (blue line) which indicates that there is an autocorrelation.

*#Performing Durbin-Watson test to check if the residuals are uncorrelated.*  
**dwtest**(reg)

##   
## Durbin-Watson test  
##   
## data: reg  
## DW = 0.30717, p-value = 1.553e-10  
## alternative hypothesis: true autocorrelation is greater than 0

***Interpretation:*** Since p value is 1.553e-10 less than 0.05 we reject the null hypothesis and conclude that there is an autocorrelation in our residuals.