NON LINEAR REGRESSION MODEL TO PREDICT VAPOR PRESSURE OF MERCURY

**INTRODUCTION**

***Objective:*** Here our main objective is to load the inbuilt dataset ‘pressure’ in R and fit a suitable regression model and perform residual analysis. We also want to comment about the possibility of modelling nonlinear regression model for the dataset.

***Problem Description:*** Here, we are interested in fitting a suitable regression model for the given inbuilt data ‘pressure’ by taking temperature as the regressor and vapor pressure as response variable. We are further interested in performing the residual analysis and to comment about the possibility of modelling a nonlinear regression model.

***Data Description:*** The dataset under consideration consists of data on the relation between temperature in degrees Celsius and vapor pressure of mercury in millimeters (of mercury).

* The ***temperature in degrees Celsius(X)*** is the ***independent variable***.
* The ***vapor pressure of mercury in millimeters(Y)*** (of mercury) is the ***dependent variable***.

*#Loading the dataset pressure.*  
**data**(pressure)  
  
*#Obtaining the first few records of the dataset.*  
**head**(pressure)

## temperature pressure  
## 1 0 0.0002  
## 2 20 0.0012  
## 3 40 0.0060  
## 4 60 0.0300  
## 5 80 0.0900  
## 6 100 0.2700

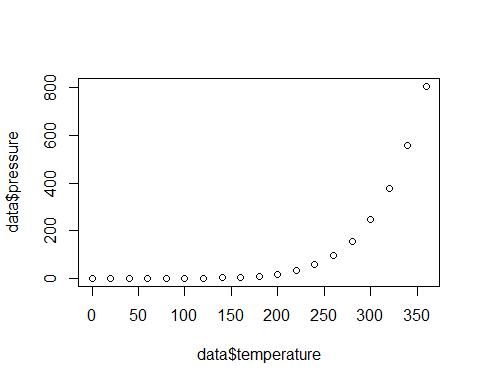
*#Renaming the 'pressure' dataset in the variable ‘data’.*  
data=pressure

**ANALYSIS**  
  
*#Attaching the 'pressure' dataset named as ‘data’.*  
**attach**(data)

## The following object is masked \_by\_ .GlobalEnv:  
##   
## pressure

## The following object is masked from package:datasets:  
##   
## pressure

*#Obtaining the scatter plot of the variable x i.e. temperature in degrees celcius and y i.e. vapor pressure of mercury in millimeters (of mercury).*  
**plot**(data**$**temperature,data**$**pressure)



* **From Figure 1, we observe that the plot is a curve, therefore we can either build,**

1. **A linear polynomial regression model**
2. **A nonlinear regression model.**

* **First we will try to fit a linear polynomial regression model to this dataset. Since in figure 1 the curve has only one bend we try to build a polynomial regression model of order 2(i.e quadratic polynomial regression model).**
* **Then we perform the residual analysis for the linear polynomial regression model. If the coefficient of determination value is greater than 0.5 and all the assumptions are validated to be true in the residual analysis then we can conclude that the linear polynomial regression model is a suitable and good fit model for this data.**
* **On the other hand if the assumptions about errors are violated we proceed for building up a nonlinear regression model for this dataset.**

*#Now, fitting a polynomial regression model of order 2.*  
reg=**lm**(data**$**pressure**~**data**$**temperature**+I**((data**$**temperature)**^**2))  
**summary**(reg)

##   
## Call:  
## lm(formula = data$pressure ~ data$temperature + I((data$temperature)^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -95.142 -54.391 -1.353 48.238 170.374   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 91.154379 46.262513 1.970 0.066354 .   
## data$temperature -2.706167 0.595775 -4.542 0.000333 \*\*\*  
## I((data$temperature)^2) 0.011718 0.001597 7.336 1.67e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 74.42 on 16 degrees of freedom  
## Multiple R-squared: 0.9024, Adjusted R-squared: 0.8902   
## F-statistic: 74 on 2 and 16 DF, p-value: 8.209e-09

**From the above summary we obtain the following polynomial regression model,**

**Y = B0+(B1\*x)+(B2\*(x^2)) = 91.154379+(-2.706167\*x)+( 0.011718\*(x^2))**

**i.e.**

**pressure = B0+(B1\*temperature)+(B2\*temperature^2)**

**pressure = 91.154379+(-2.706167\* temperature)+( 0.011718\*(temperature ^2))**

**From the above model we observe that,**

**The intercept is 91.154379 which means that when the temperature is 0 degree celcius the vapor pressure of mercury is 91.154379 millimeters.**

**B1=-2.706167 and B2=0.011718 are the quadratic effect parameter.**

**Since B2 is positive i.e. B2=0.011718 which is positive which indicates that means the curvature is upward.**

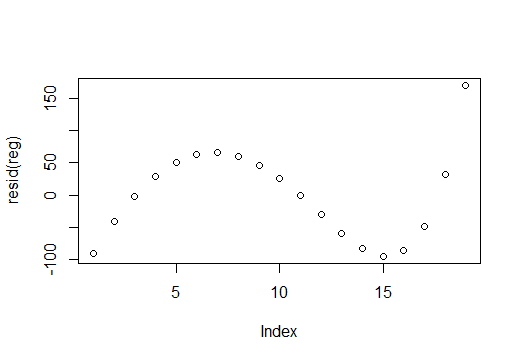
**Since, coefficient of determination is 0.8902 which is greater than 0.5, hence our fitted regression model is of good quality.**

**Also since the p value is negligible it can be concluded that the performance of the fitted regression model is good.**

*#Fitting the model in the plot.*

lines(smooth.spline(data$temperature,predict(reg)),col="blue",lwd=3)

plot(resid(reg))





**From figure 2 we observe that the obtained plot is not a straight line therefore it is evident that linear polynomial regression model is not a suitable regression model for this data. We also verify this by considering the assumption about the errors.**

*#Now, we try to validate the following assumtions regarding the fitted modeL,*

The following are the assumtions about the errors that we make,

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.

1. To check if the relationship beetween y and x is linear.

**From the figure 1 we observe that the relationship between y and x is not linear hence we observe that the assumption about the linearity is violated.**2. To check if the errors have zero mean.

#Obtaining the residuals of the fitted model.  
residuals=**resid**(reg)

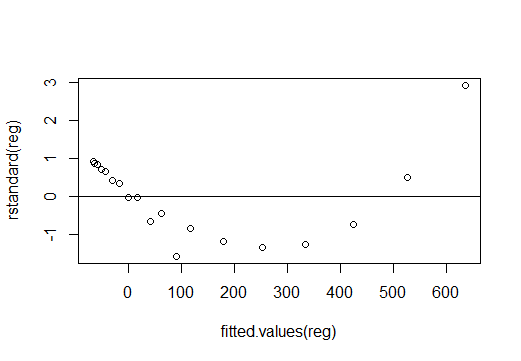
#Obtaining the mean of the residuals*.*  
**mean**(residuals)

## [1] 4.955831e-15

**Hence, from the above calculation we observe that the errors has mean 5.795529e-15 which is almost equal to zero hence the assumption of error mean to be zero is validated to be true.**

3. Assumption of homoscedasticity, i.e. the errors have constant variance.

*#Plot of fitted values against residuals.*  
fitted\_values<-**fitted.values**(reg)  
**plot**(**fitted.values**(reg),**rstandard**(reg))  
**abline**(0,0)





**From figure 3 i.e. the plot of fitted values against residuals we observe that all the points are not evenly spread around the horizontal line i.e. origin which means the errors have non-constant variance.***#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**(reg)

##   
## studentized Breusch-Pagan test  
##   
## data: reg  
## BP = 7.4533, df = 2, p-value = 0.02407

**From studentized Breusch-Pagan test weobserve that p value obtained is 0.02407 which is less than 0.05 hence we reject the null hypothesis and conclude that the errors have non-constant variance.**

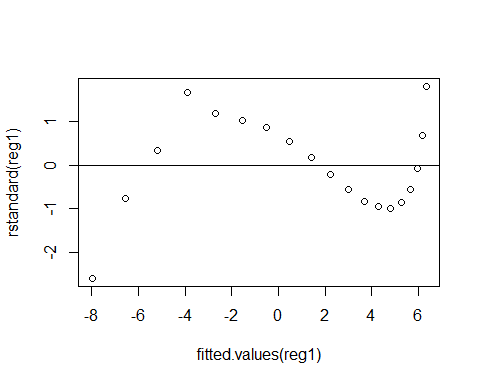
Since the assumption of homoscedasticity is violated we try to stabilize the variance in the following steps,

*#Step 1 – Transforming the response variable*  
y=**log**(data**$**pressure)  
  
*#Step 2 - Fit and validate a polynomial regression model of order 2 in the transformedvariable.*  
reg1=**lm**(y**~**data**$**temperature**+I**((data**$**temperature)**^**2))  
**summary**(reg1)

##   
## Call:  
## lm(formula = y ~ data$temperature + I((data$temperature)^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5219 -0.1840 -0.0177 0.1800 0.4008   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.995e+00 1.603e-01 -49.87 < 2e-16 \*\*\*  
## data$temperature 7.380e-02 2.065e-03 35.74 < 2e-16 \*\*\*  
## I((data$temperature)^2) -9.447e-05 5.536e-06 -17.07 1.09e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2579 on 16 degrees of freedom  
## Multiple R-squared: 0.9972, Adjusted R-squared: 0.9969   
## F-statistic: 2859 on 2 and 16 DF, p-value: < 2.2e-16

Now we check the assumption of homoscedasticity again.

*#Plot of fitted values against residuals.*  
fitted\_values<-**fitted.values**(reg1)  
**plot**(**fitted.values**(reg1),**rstandard**(reg1))  
**abline**(0,0)



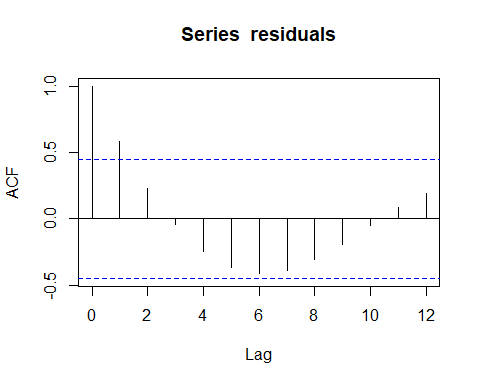
**From figure 4 i.e. the plot of fitted values against residuals we observe that all the points are not evenly spread around the horizontal line i.e. origin which means the errors have non-constant variance.**

*#Performing studentized Breusch-Pagan test to check the assumption of homoscedasticy.*  
**bptest**(reg1)

## studentized Breusch-Pagan test  
##   
## data: reg1  
## BP = 7.0995, df = 2, p-value = 0.02873

**From the studentized Breusch-Pagan test we again observe that the p-value= 0.006249<0.05 hence we reject the null hypothesis and conclude that the errors have non constant variance.Hence the assumption of homoscedasticity is violated.**4. To check if the errors are uncorrelated.

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





**From the figure 4 we observe that lag 1 crosses the threshold line therefore since all the lags are not inside the threshhold line we conclude that the residuals are auto correlated i.e. they arenot uncorrelated.**  
  
*#Performing Durbin-Watson test to check if the residuals are uncorrelated.*  
**dwtest**(reg)

##   
## Durbin-Watson test  
##   
## data: reg  
## DW = 0.41271, p-value = 1.226e-08  
## alternative hypothesis: true autocorrelation is greater than 0

**From the above test it is observed that the p value is negligible which is less than 0.05 hence we reject the null hypothesis and conclude that there exists some correlation between residuals and they are not uncorrelated. Hence this assumption is also violated.**

5. Errors are normally distributed random variables.

*#Now, performing Shapiro-Wilk normality test to check if the residuals are normaally distributed or not.*  
**shapiro.test**(residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: residuals  
## W = 0.93405, p-value = 0.2056

**OnperformingShapiro-Wilk normality test we observe that p value is 0.2056 which is greater than 0.05 hence we fail to reject the null hypothesis and conclude that the errors residuals are normally distributed.**

**All the assumptions about errors are violated except for normality assumption and assumption about mean of the errors.**

**CONCLUSION**

1. From the summary of the model we obtain the following polynomial regression model,

Y = B0+ (B1\*x) + (B2\*(x^2)) = 91.154379+ (-2.706167\*x) + (0.011718\*(x^2))

i.e.

Pressure = B0+ (B1\*temperature) + (B2\*temperature^2)

Pressure = 91.154379+ (-2.706167\* temperature) + (0.011718\*(temperature ^2))

From the above model we observe that,

The intercept is 91.154379 which means that when the temperature is 0 degree Celsius the vapor pressure of mercury is 91.154379 millimeters.

B1=-2.706167 and B2=0.011718 are the quadratic effect parameters.

Since B2 is positive i.e. B2=0.011718 which is positive which indicates that means the curvature is upward.

Since, coefficient of determination is 0.8902 which is greater than 0.5, hence our fitted regression model is of good quality.

Also since the p value is negligible it can be concluded that the performance of the fitted regression model is good.

1. On performing residual analysis we observe that all the assumptions except normality assumption and assumption about mean of the errors is violated therefore we cannot model a polynomial linear regression model for this data.
2. Since all the assumptions about the errors are violated for polynomial regression model, we go for modelling nonlinear regression model for this data.