IDA PROJECT REPORT

Topic 3 : Regression analysis for establishing a relation between response and regressor variables

Date of Submission: 30-11-2021

Group 3 :

Kongamudi Rokesh Reddy	-S20190010098
Y Rahul Rao	-S20190020264
Kamasani Bharath	-S20190010080
Uppatala Venkatesh	-S20190020257
Basheedu Sai Krishna	-S20190010016

UNDERSTANDING THE THEORY:

REGRESSION ANALYSIS:

Regression analysis is a statistical method to deal with the formulation of a mathematical model depicting relationships amongst variables, which can be used
for prediction of the values of dependent variable, given the values of independent variables.

Here we have 2 variables: -

1. *Base Pay (Independent Variable/ Control Variable/ Regressor Variable)*
2. *Sum of Overtime Pay, Other Pay and Benefits (Dependent Variable/ Response Variable)*

LINEAR REGRESSION ANALYSIS:

Here, we assume that the true relationship between Response Y and Regressor X is $Y = aX + b$, *but as the data cannot be explained only with X, i.e there will be some variability left, this variability is because there are some variables that we couldn't find out. So, we change the relation between Y and X a little bit, the new relationship will be* $Y = aX + b + \varepsilon$, *be where* ε *is the sum of all variability that we can't explain or couldn't find out.* ε *is the random error component.*

Some Properties of ε

1. ε *have zero mean and* σ^2 *as variance.*
2. *If we take any* ε_i *and* ε_j *are uncorrelated when* $i \neq j$.
3. ε_i *is normally distributed with mean 0 and variance* σ^2.

$E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, *where* σ^2 *is called the error of variance.*

When we try to fit the Y and X, we get a relation where $\hat{Y} = aX + b$, *where Y is the true value and* $\hat{Y}$ *is the predicted value or fitted value according to the model.*

LEAST SQUARE ERROR METHOD:

This method uses the concept of residual. A residual is essentially an error in the fit of the model

$y = ax + b$. Thus residual is $x\ e_i = y_i -$

$\hat{Y}_i$ where $i = 1,2,3,4 \ldots n$

Residual Sum of Squares also called as sum of squares of the errors(SSE) about the fitted line

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

(Source:- Slides)

We minimise SSE and find the parameters a,b

For minimum value of SSE $\frac{\partial (SSE)}{\partial a} = 0$ and $\frac{\partial (SSE)}{\partial b} = 0$

Thus, we can calculate the values of a,b

NON-LINEAR REGRESSION ANALYSIS:

Here the relationship between Response Y and Regressor X is $Y=aX^n+bX^{n-1}\ldots.+k_0$

MEASURING THE QUALITY OF FIT - R^2

The variability of the fitted model is determinedcalled as Coefficient of determination. by the quantity R^2, also $\sum_{=1}^{n} (y_i - \hat{y}_i)^2$ output, while y is the actual output/label.$_i$y Sum of squared errors, SSE =, where $\hat{}$ is the predicted

Total corrected sum of squares, $\sum_{i=1}^{n} (y_i - \bar{y})^2$ SST =, where y is the mean.

$R^2 = 1 - SSE/SST$

If the R^2 ~ 0, then the model is said to be poorly fit
If R_2~ 1, then it is said to be a good fit.

IMPLEMENTATION

PREPROCESSING AND CLEANING OF THE DATA:

R programming implementation of regression :

1. *We have first loaded the dataset, 'Salaries.csv' file into variable data using read.csv and we selected the columns "Base Pay", "overtime Pay", "Other Pay", "benefits", "year"*

```
> head(data)
  OvertimePay  OtherPay Benefits   BasePay Year
1         0.0 400184.25          167411.18 2011
2   245131.88 137811.38          155966.02 2011
3   106088.18   16452.6          212739.13 2011
4    56120.71  198306.9            77916.0 2011
5      9737.0 182234.59           134401.6 2011
6      8601.0 189082.74           118602.0 2011
> tail(data)
        OvertimePay     OtherPay     Benefits      BasePay Year
148649         0.00         0.00         0.00         0.00 2014
148650         0.00         0.00         0.00         0.00 2014
148651 Not Provided Not Provided Not Provided Not Provided 2014
148652 Not Provided Not Provided Not Provided Not Provided 2014
148653 Not Provided Not Provided Not Provided Not Provided 2014
148654         0.00      -618.13         0.00         0.00 2014
>
```

2. *We have then selected only the rows which have entries from the year 2014 and stored in df.*

```
> head(data)
  OvertimePay  OtherPay Benefits   BasePay Year
1        0.00 342802.63 38780.04 129150.01 2014
2    10712.95  60563.54 89540.23 318835.49 2014
3        0.00  82313.70 96570.66 257340.00 2014
4        0.00  19266.72 91302.46 307450.04 2014
5        0.00  24165.44 91201.66 302068.00 2014
6     6009.22  67956.20 71580.48 270222.04 2014
> tail(data)
       OvertimePay     OtherPay     Benefits      BasePay Year
38118         0.00         0.00         0.00         0.00 2014
38119         0.00         0.00         0.00         0.00 2014
38120 Not Provided Not Provided Not Provided Not Provided 2014
38121 Not Provided Not Provided Not Provided Not Provided 2014
38122 Not Provided Not Provided Not Provided Not Provided 2014
38123         0.00      -618.13         0.00         0.00 2014
>
```

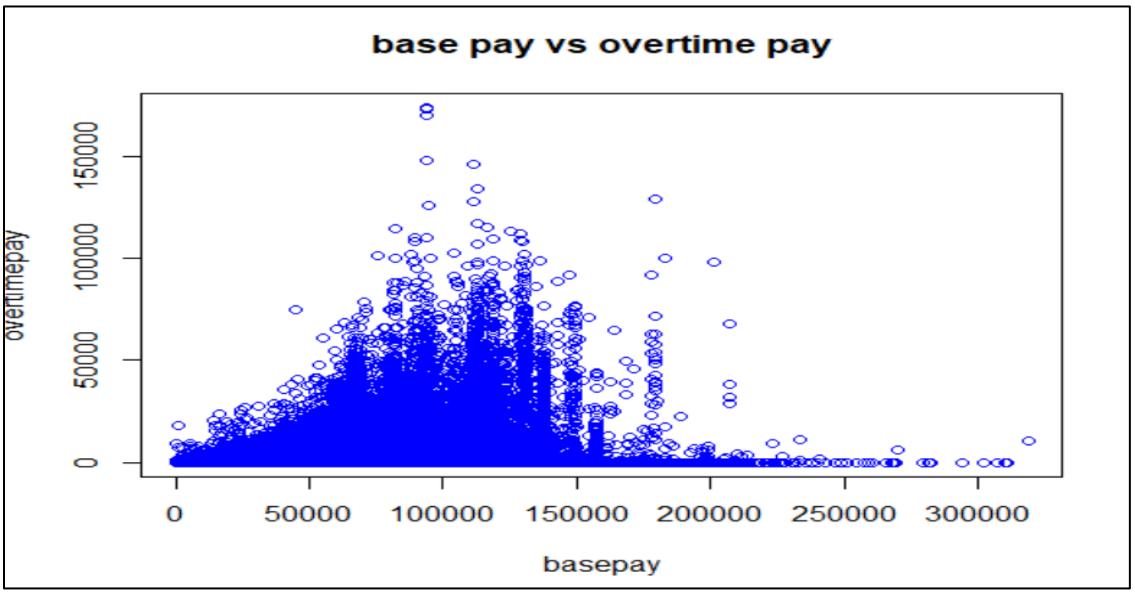
3. *The values in the columns Base Pay, Overtime Pay, Other Pay, Benefits are stored in 4 different arrays and NaN values are replaced by 0, negative and 'Not Possible' entries are removed.*

```
> head(data)
  OvertimePay  OtherPay Benefits  BasePay Year
1        0.00 342802.63 38780.04 129150.0 2014
2    10712.95  60563.54 89540.23 318835.5 2014
3        0.00  82313.70 96570.66 257340.0 2014
4        0.00  19266.72 91302.46 307450.0 2014
5        0.00  24165.44 91201.66 302068.0 2014
6     6009.22  67956.20 71580.48 270222.0 2014
> tail(data)
       OvertimePay OtherPay Benefits BasePay Year
38112         0.00        0        0       0 2014
38113         0.00        0        0       0 2014
38114         0.00        0        0       0 2014
38115         0.00        0        0       0 2014
38116         0.00        0        0       0 2014
38117         0.00        0        0       0 2014
>
```

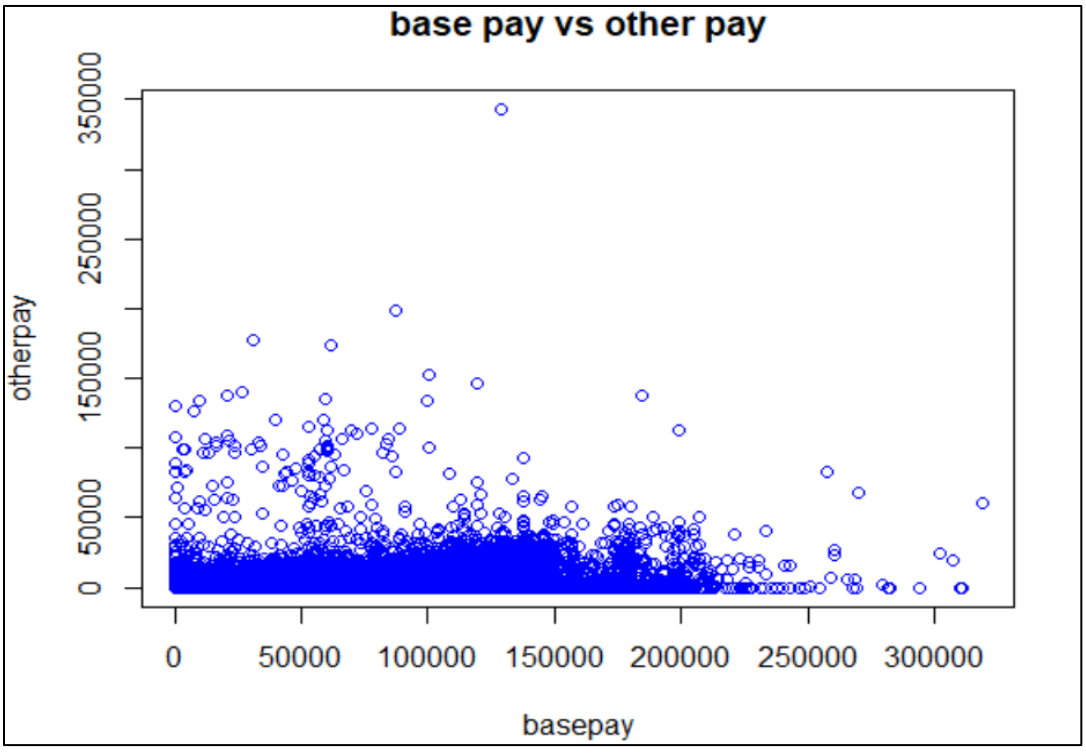
4. *Drop unnecessary columns and rows. Separate the data into X (independent) and y (dependent). Add Overtime Pay, Other Pay and benefits together to create only one new column (let it be X).*
5. *For linear regression consider only X and y. For degree 2 and 3 consider X^2 and X^3 respectively.*
6. *In order to find the appropriate values of the coefficients required to fit the corresponding model onto the data, a function has been defined which returns the optimal values for the coefficients.*
7. *The values are calculated by applying derivatives to the sum of squared losses function, with respect to the coefficients.*
8. *After these, the values of the coefficients are used to make predictions. Plots are drawn for original salary and our predicted salary against the X.*
9. *Coefficient of determination i.e., R^2 is calculated for the fitted model, to determine the quality of the fit.*

DATA VISUALIZATON:

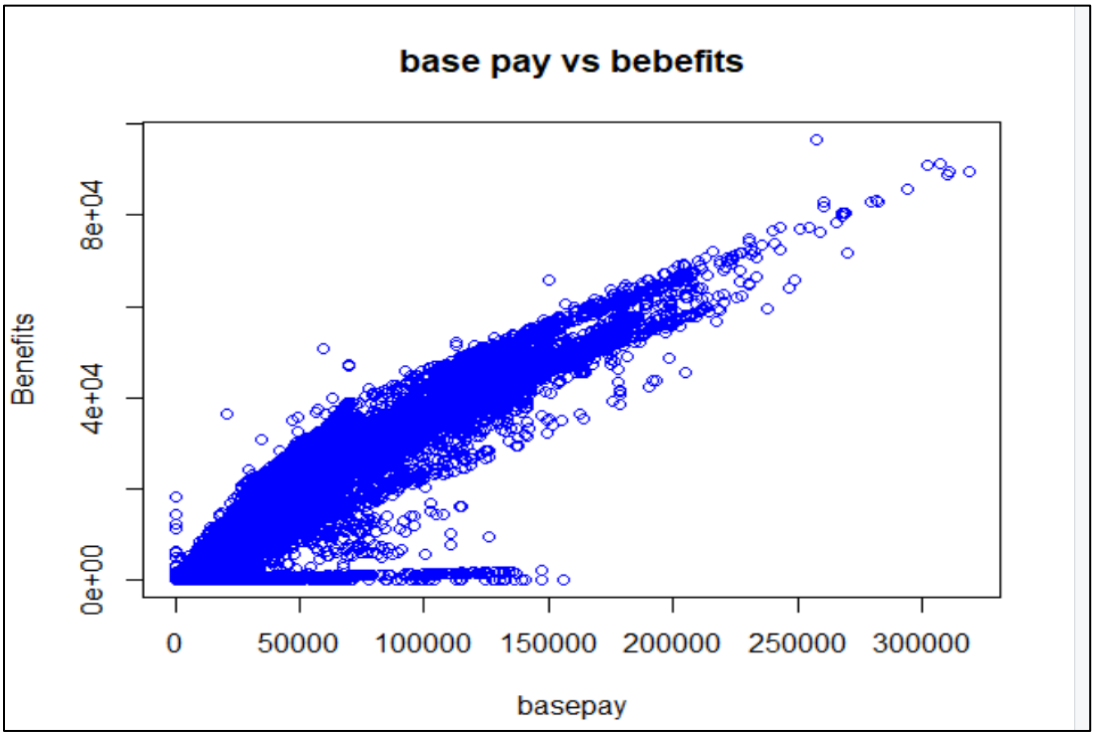
Overtime Pay vs Base Pay



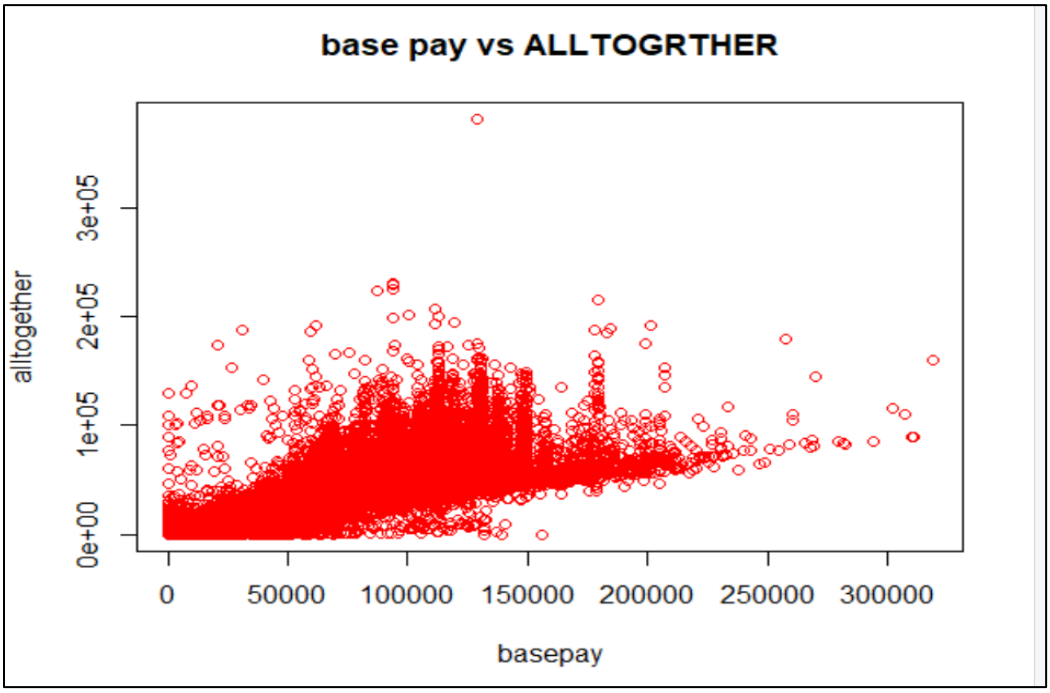
Other Pay vs Base Pay



Benefits vs Base Pay

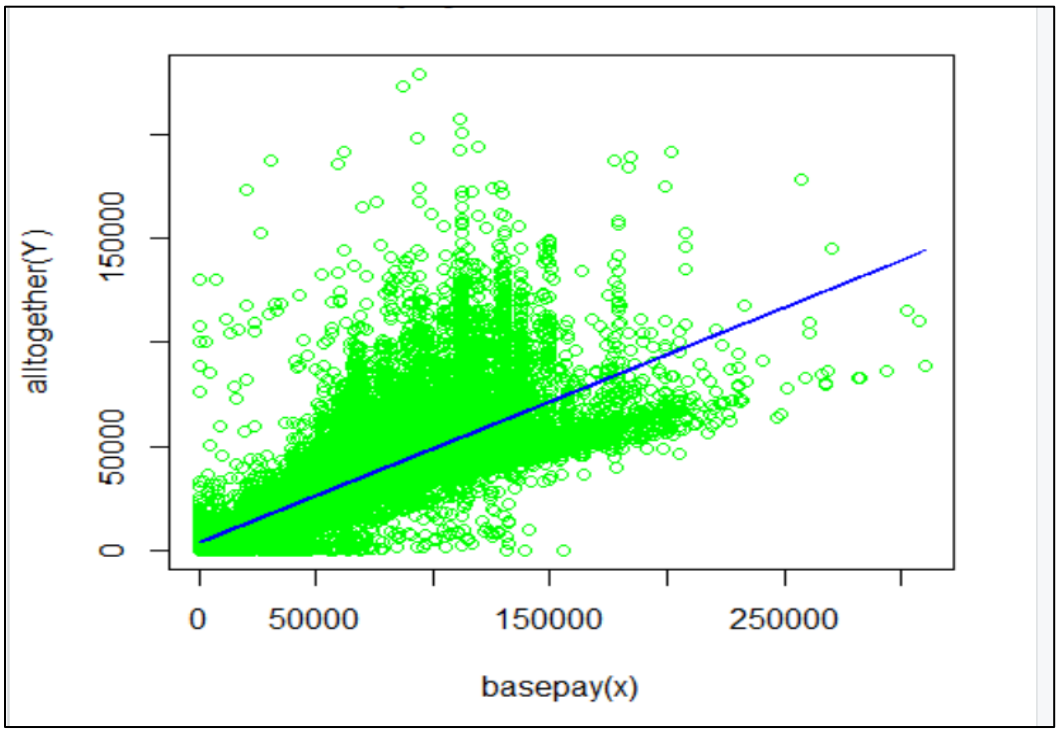


Altogether vs Base Pay



RESULTS:

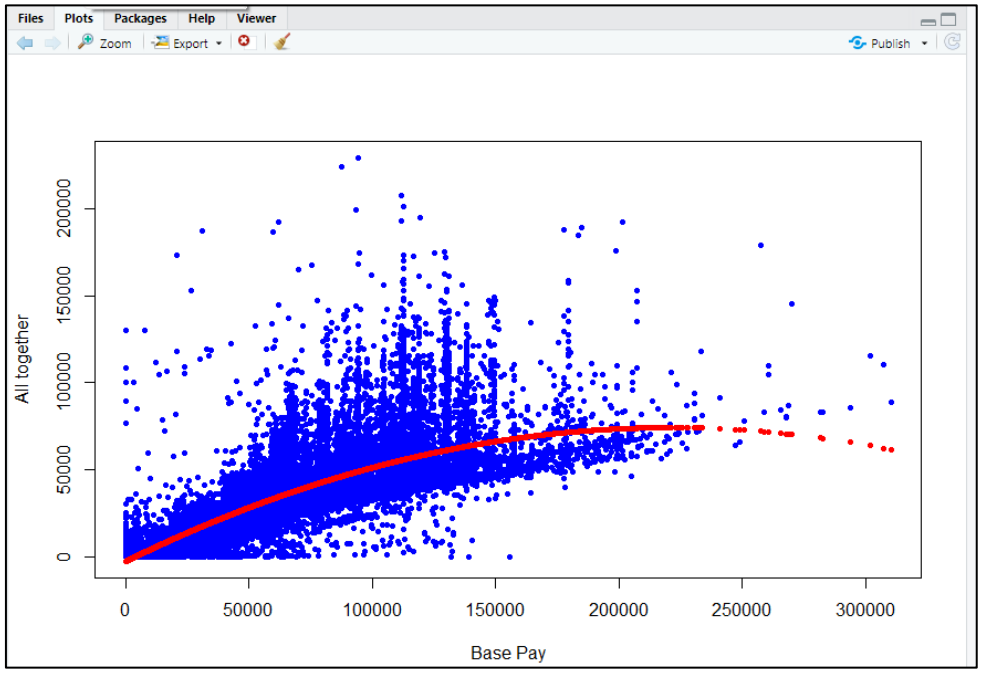
LINEAR REGRESSION ANALYSIS:



R2 value = 0.594216 in linear regression

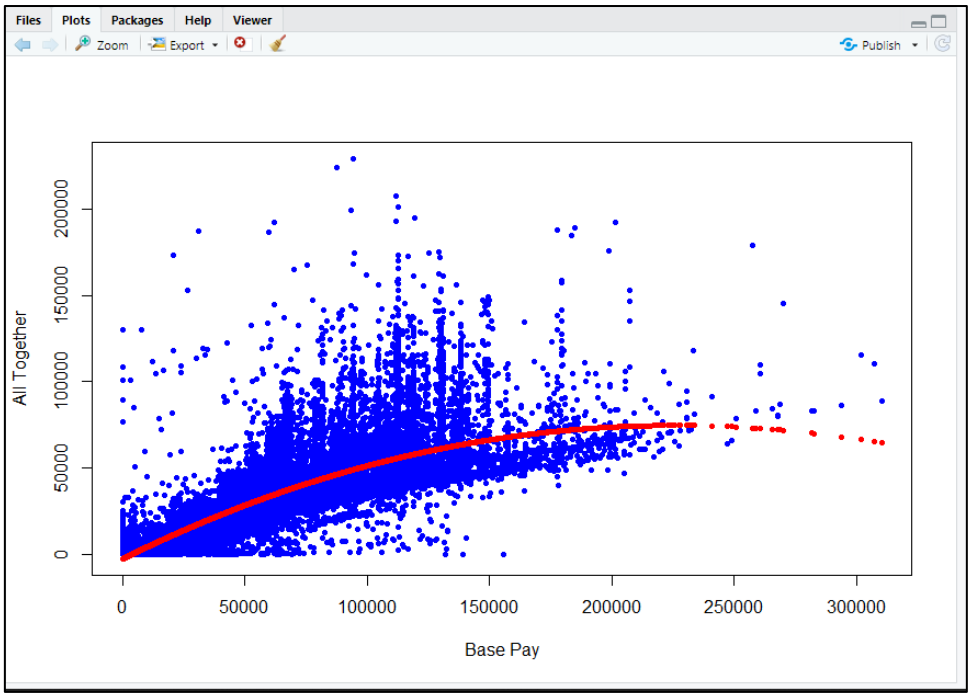
```
> R2_linear
[1] 0.5942168
```

NON-LINEAR REGRESSION ANALYSIS WITH DEGREE 2



R^2 value: 0.621044

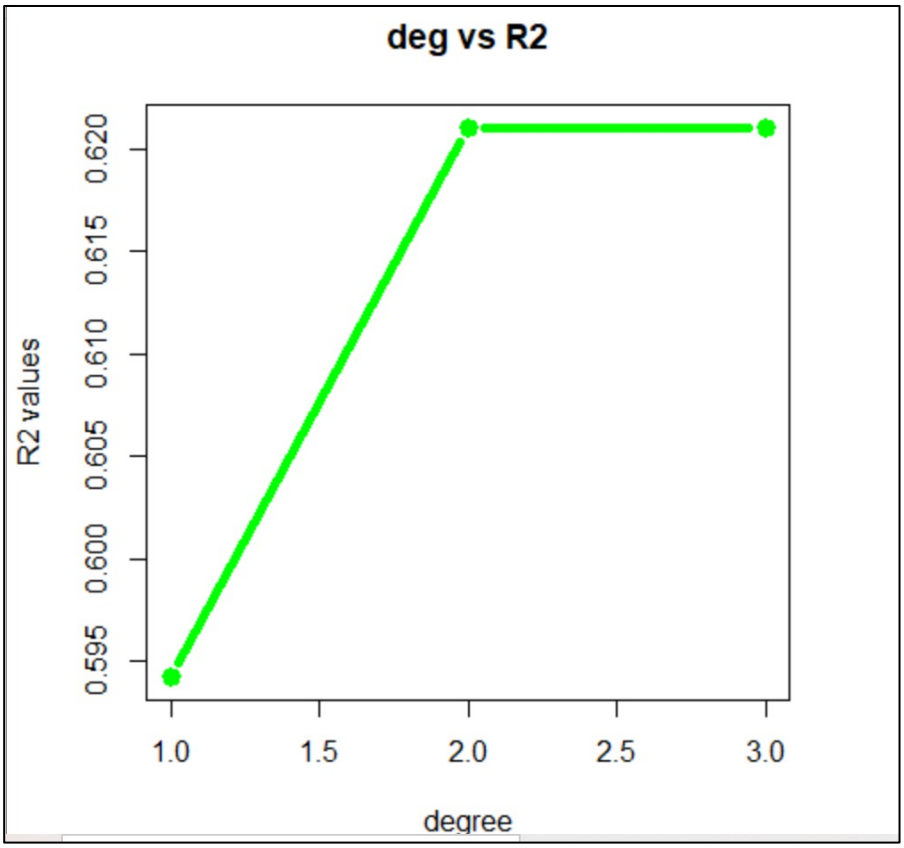
NON-LINEAR REGRESSION ANALYSIS WITH DEGREE 3:



R^2 value: 0.6210517

CONCLUSION:

Plot shows the R^2 value for the linear regression , non-linear with degree 2 and 3



R^2 Value in simple nonlinear regression = 0.594216

R^2 Value in nonlinear regression with degree 2 = 0.621044

R^2 Value in nonlinear regression with degree 3 = 0.6210517

So, we conclude that Polynomial model with degree 3 gives best R^2 score

Y=(x^3)*3.558125e-13 – (x^2)*1.679151e-06 + (x)*0.7034769 - 2737.416 *is the best fitted curve compared to linear and non linear with degree 2*