IDA PROJECT REPORT

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

Data set - SALARY data

Date of Submission - 27/11/2021

Group Number - 18

UPPALA NISHITA nishita.u19@iiits.in S20190010184	**KRISHNA NIVEDA** niveda.k19@iiits.in S20190020224
SRI LAKSHMI PRASANNA KONERU srilakshmiprasanna.k19@iiits.in S20190010168	**SARABU ANUHYA** anuhya.s19@iiits.in S20190020251
PEDAPALLI SAAM PRASANTH DEEVEN saamprasanthdeevan.p19@iiits.in S20190010136	

Problem Statement :

Database contains salary information of different employees in different organisations. It is required to test whether Overtime Pay, Other Pay and benefits altogether increase with Basic Pay for the year 2014.

i. Simple linear regression analysis.

ii. Simple non-linear regression analysis with degree 2.

iii. Simple non-linear regression analysis with degree 3.

c) Calculate R^2 values in each of the above-mentioned cases and finally conclude your results precisely.

Understanding the theory to solve the project problem:

Regression Analysis

- Given the values of independent variables(Regressor- X), regression analysis is used to predict the values of dependent variables(Response variables - Y).
- In the case of Linear regression analysis,the relationship between Y and X is the linear relationship in the form of $Y = \alpha + \beta X$ where, α and β are the regression coefficients. If the equation is considered in terms of r degree, In this case r =1.
- We need to find the next relationship between X and Y meaning the best fitted values of α and β as well.
- We use the least square method to estimate α and β.This method uses SSE(sum of squared errors).

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

$$where\ a = \alpha\ and\ b = \beta$$

- Then we differentiate this equation twice separately, once w.r.t a and then w.r.t b and equate both the equations to 0, so that we get the minimum values of a and b.

- More minimal the values, we get more minimal errors which leads to much more accurate results.
- After differentiating the final equations to find a and b would be the following;

$$b = \frac{\Sigma^{n}_{i=1} \ (x_i - \bar{x}) \ (y_i - \bar{y})}{\Sigma^{n}_{i=1} \ (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- Regression equation in terms of r degree where r>1 we say it as non-linear regression.

Implementation of the project :

Step 1 - Analysing the Dataset

- In the dataset the salary is from the year 2011 to 2014, but according to the problem statement we are required to test only for the year 2014.
- So, for the year 2014 we should filter the given dataset accordingly.
- The dataset provided has 148654 records with 13 attributes in each record.
- Since the relation analysis only asked for the year 2014. We have sliced the data. Now the total records we have are 16506

Step2 - Dealing with missing value

- The mode of an object that holds character strings in R is "character".
- We will convert Character into Numerical Value and then for the missing values.
- We will assign NA for zero data then we will delete all the rows that have no data in specified columns.

Step 3 - Initializing Regressor and Response

- Regressor = OtherPay+ OvertimePay+ Benefits
- Response = BasePay

Step 4 - Dealing with Outliers

- Outlier is a value or an observation that is distant from other observations, that is, a data point that differs significantly from other data points.
- In most practical circumstances an outlier decreases the value of a correlation coefficient and weakens the regression relationship, but it's also possible that in some circumstances an outlier may increase a correlation value and improve regression.
- One of the easiest ways to identify outliers in R is by visualizing them in boxplots.
- Boxplots typically show the median of a dataset along with the first and third quartiles.
- They also show the limits beyond which all data values are considered as outliers.

Step 5 - Normalization

- We have used min max normalisation.
- The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.
- For x=regressor, regressor=((x- min(x)) /(max(x)-min(x)))
- For x=response, response=((x- min(x)) /(max(x)-min(x)))

Step 6 - Implementation

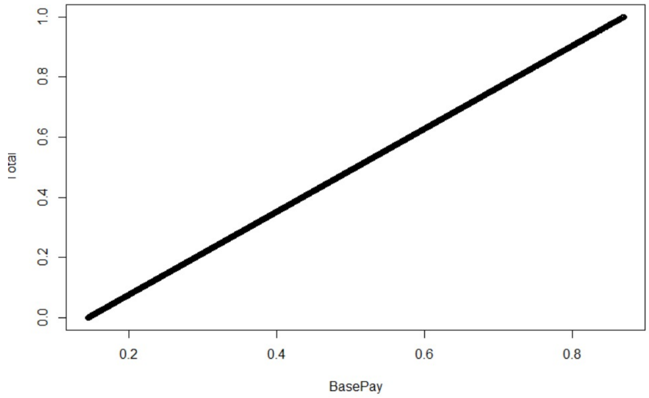
- With respect to our problem statement we don't require the whole dataset.
- We want BasePay , OvertimePay, OtherPay, Benefits, year columns.
- From the data info provided above, we can see that all required 4 columns are in object type and we have to change datatype to float.
- As we know we have to analyse how overtime pay + other pay + benefits increases with BasePay.
- Let's make a new column which is the sum of OvertimePay, OtherPay, Benefits as Response and BasePay column as Regressor.
- Dropped all other columns except Response, Regressor, Year.

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^k \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^k & \sum_{i=1}^{n} x_i^{k+1} & \cdots & \sum_{i=1}^{n} x_i^{2k} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^k y_i \end{bmatrix}.$$

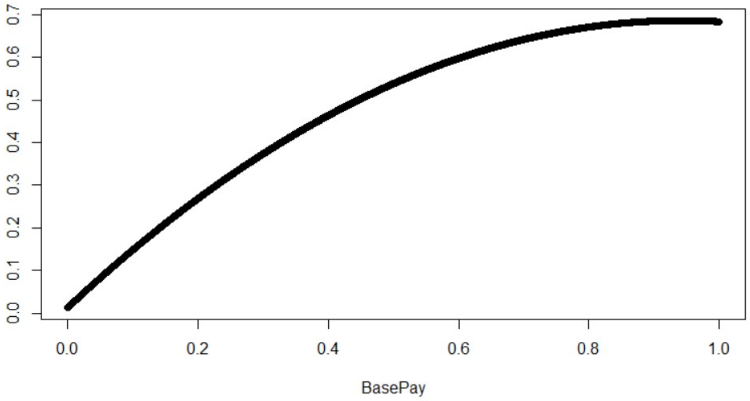
- Using this matrix, shown in the above figure we have found the coefficients for 2nd and 3rd degree.

Experimental results :

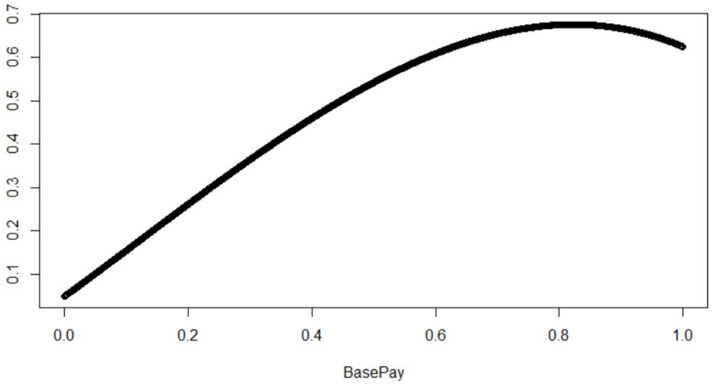
Predicted response graph for degree 1



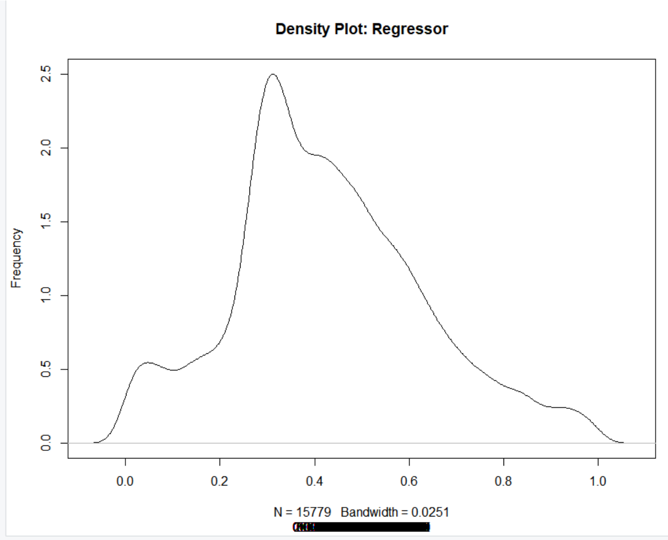
Predicted response graph for degree 2



Predicted response graph for degree 3

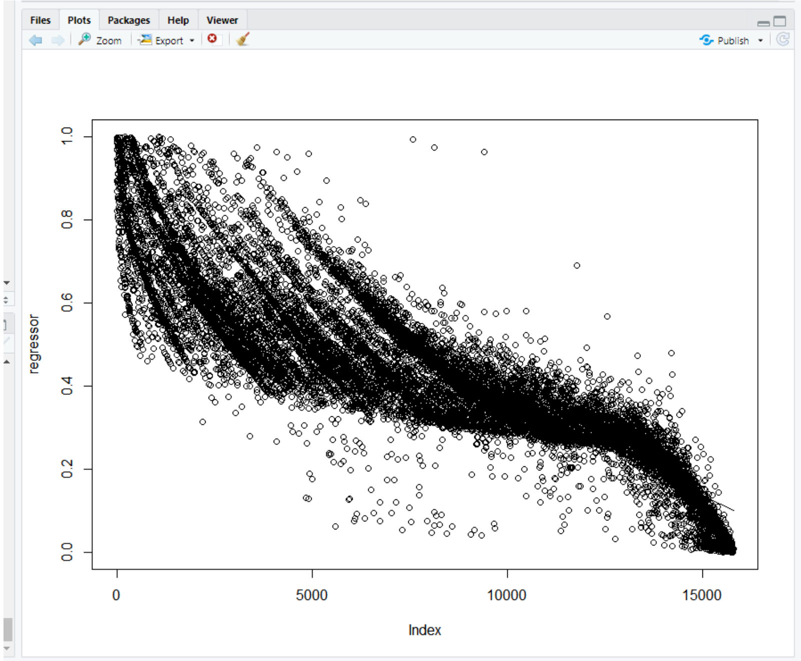


Density Plot : Regression

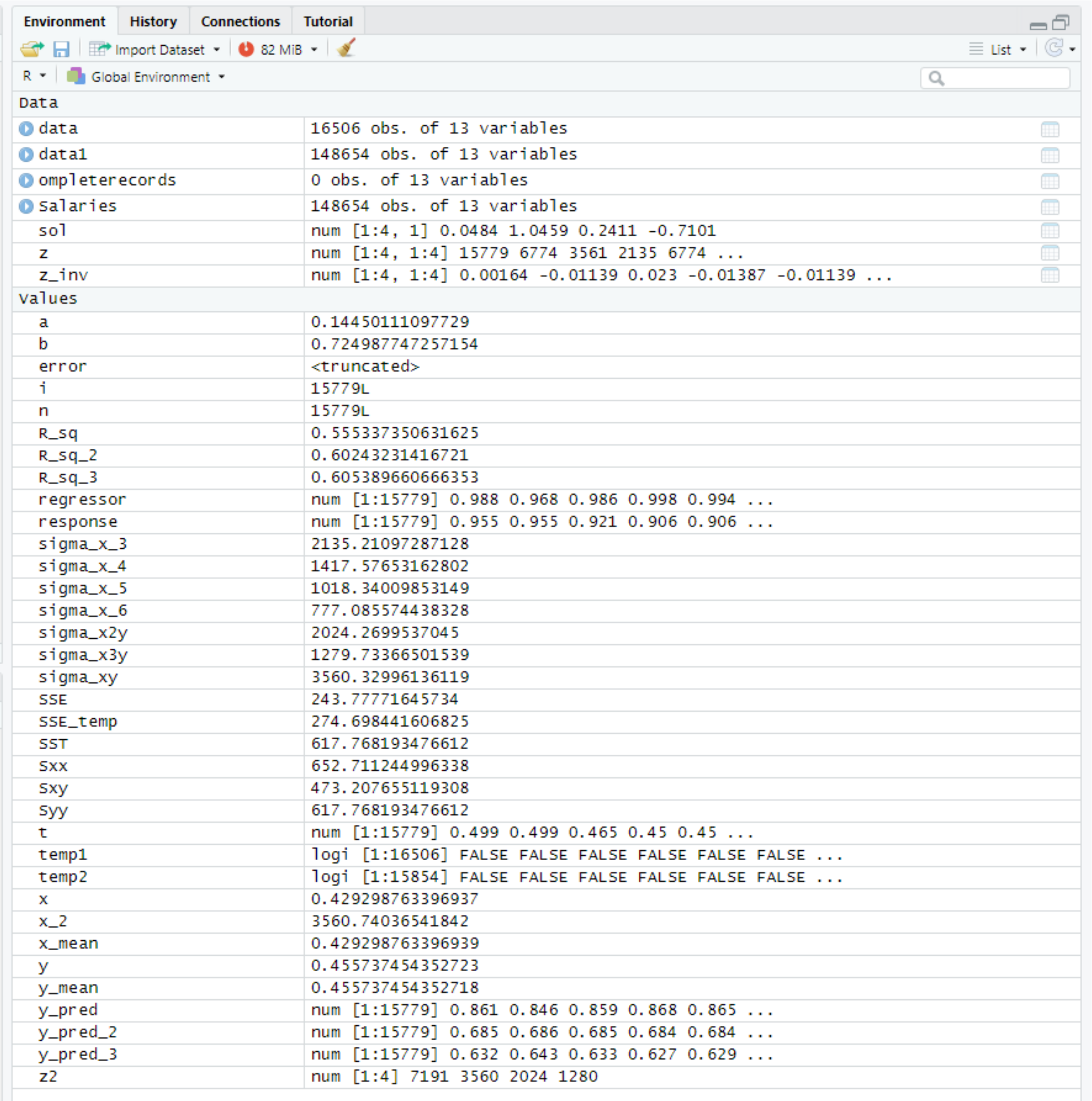


[redacted]

Scatter plot for regressor

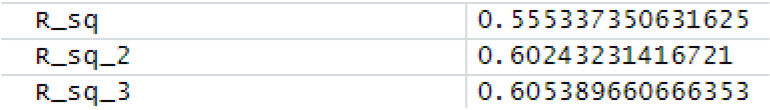


Data and Values :



```
Data
data              16506 obs. of 13 variables
data1             148654 obs. of 13 variables
ompleterecords    0 obs. of 13 variables
salaries          148654 obs. of 13 variables
sol               num [1:4, 1] 0.0484 1.0459 0.2411 -0.7101
z                 num [1:4, 1:4] 15779 6774 3561 2135 6774 ...
z_inv             num [1:4, 1:4] 0.00164 -0.01139 0.023 -0.01387 -0.01139 ...
Values
a                 0.14450111097729
b                 0.724987747257154
error             <truncated>
i                 15779L
n                 15779L
R_sq              0.555337350631625
R_sq_2            0.60243231416721
R_sq_3            0.605389660666353
regressor         num [1:15779] 0.988 0.968 0.986 0.998 0.994 ...
response          num [1:15779] 0.955 0.955 0.921 0.906 0.906 ...
sigma_x_3         2135.21097287128
sigma_x_4         1417.57653162802
sigma_x_5         1018.34009853149
sigma_x_6         777.085574438328
sigma_x2y         2024.2699537045
sigma_x3y         1279.73366501539
sigma_xy          3560.32996136119
SSE               243.77771645734
SSE_temp          274.698441606825
SST               617.768193476612
Sxx               652.711244996338
Sxy               473.207655119308
Syy               617.768193476612
t                 num [1:15779] 0.499 0.499 0.465 0.45 0.45 ...
temp1             logi [1:16506] FALSE FALSE FALSE FALSE FALSE FALSE ...
temp2             logi [1:15854] FALSE FALSE FALSE FALSE FALSE FALSE ...
x                 0.429298763396937
x_2               3560.74036541842
x_mean            0.429298763396939
y                 0.455737454352723
y_mean            0.455737454352718
y_pred            num [1:15779] 0.861 0.846 0.859 0.868 0.865 ...
y_pred_2          num [1:15779] 0.685 0.686 0.685 0.684 0.684 ...
y_pred_3          num [1:15779] 0.632 0.643 0.633 0.627 0.629 ...
z2                num [1:4] 7191 3560 2024 1280
```

Output :



R_sq	0.555337350631625
R_sq_2	0.60243231416721
R_sq_3	0.605389660666353

R^2 for Simple Linear regression analysis =0.555

R^2 for Simple non linear analysis with Degree 2 = 0.6024

R^2 for Simple non linear analysis with Degree 3 = 0.60538.

Conclusion :

By comparing all the three R^2 values from different cases we can say that the fit is more for non-linear regression with degree 3. All the other pays increase with basic pay till some certain point and later start decreasing for people with higher basic pay. People in the mid range get a good amount of other pays.