**Linear Regression Assignment**

**3. Problem statement: Emp\_data -> Build a prediction model for Churn\_out\_rate.**

1.Perform EDA-1st Moment,2nd Moment,3rd Moment & 4th Moment.

2. Scatter Diagram-Direction, Strength & Linear.

3. Correlation Coefficient Value =r (-1 to +1) thumb rule if r >0.85 then it is strong. I got r= -0.91. This is a good relationship.

4. Coefficient of determination is R2(0 to 1) thumb rule if R2>0.8 then it is strong. R2 =0.83. This is Strong.

I have done the analysis with R studio as follows:

R-script

# **Load emp\_data.csv dataset**

library(readr)

emp\_data <- read\_csv("C:Data Science Assignments/Linear Regression/Dataset/emp\_data.csv")

View(emp\_data)

attach(emp\_data)

# **Exploratory data analysis**

summary(emp\_data)

#**Scatter plot**

plot(emp\_data$Salary\_hike, emp\_data$Churn\_out\_rate) # plot(X,Y)

?plot

#Correlation Coefficient (r)

cor(Salary\_hike,Churn\_out\_rate )# cor(X,Y)

# Simple Linear Regression model

reg <- lm(emp\_data$Churn\_out\_rate ~ emp\_data$Salary\_hike) # lm(Y ~ X)

summary(reg)

pred <- predict(reg)

reg$residuals

sum(reg$residuals)

mean(reg$residuals)

sqrt(sum(reg$residuals^2)/nrow(emp\_data)) #RMSE

sqrt(mean(reg$residuals^2))

confint(reg,level=0.95)

predict(reg,interval="predict")

**Console Window of R**

> # Load emp\_data.csv dataset

> library(readr)

> emp\_data <- read\_csv("C:Data Science Assignments/Linear Regression/Dataset/emp\_data.csv")

Parsed with column specification:

cols(

Salary\_hike = col\_double(),

Churn\_out\_rate = col\_double()

)

> View(emp\_data)

> attach(emp\_data)

The following objects are masked from emp\_data (pos = 3):

Churn\_out\_rate, Salary\_hike

> # Exploratory data analysis

> summary(emp\_data)

Salary\_hike Churn\_out\_rate

Min. :1580 Min. :60.00

1st Qu.:1618 1st Qu.:65.75

Median :1675 Median :71.00

Mean :1689 Mean :72.90

3rd Qu.:1724 3rd Qu.:78.75

Max. :1870 Max. :92.00

> #Scatter plot

> plot(emp\_data$Salary\_hike, emp\_data$Churn\_out\_rate) # plot(X,Y)

> ?plot

> #Correlation Coefficient (r)

> cor(Salary\_hike,Churn\_out\_rate )# cor(X,Y)

[1] -0.9117216

> # Simple Linear Regression model

> reg <- lm(emp\_data$Churn\_out\_rate ~ emp\_data$Salary\_hike) # lm(Y ~ X)

> summary(reg)

Call:

lm(formula = emp\_data$Churn\_out\_rate ~ emp\_data$Salary\_hike)

Residuals:

Min 1Q Median 3Q Max

-3.804 -3.059 -1.819 2.430 8.072

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\*

emp\_data$Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.469 on 8 degrees of freedom

Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101

F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

> pred <- predict(reg)

> reg$residuals

1 2 3 4 5 6 7

8.0724687 3.1033216 -0.8812519 -2.8349726 -3.8041197 -2.7578403 -3.1331580

8 9 10

-3.6961345 0.4118507 5.5198359

> sum(reg$residuals)

[1] 4.440892e-16

> mean(reg$residuals)

[1] 4.449566e-17

> sqrt(sum(reg$residuals^2)/nrow(emp\_data)) #RMSE

[1] 3.997528

> sqrt(mean(reg$residuals^2))

[1] 3.997528

> confint(reg,level=0.95)

2.5 % 97.5 %

(Intercept) 181.2912317 307.4385905

emp\_data$Salary\_hike -0.1388454 -0.0642399

> predict(reg,interval="predict")

fit lwr upr

1 83.92753 72.38391 95.47115

2 81.89668 70.59327 93.20009

3 80.88125 69.68123 92.08127

4 77.83497 66.87456 88.79538

5 75.80412 64.94216 86.66607

6 72.75784 61.94828 83.56740

7 71.13316 60.30425 81.96206

8 68.69613 57.77694 79.61533

9 61.58815 50.00746 73.16884

10 54.48016 41.72742 67.23290

Warning message:

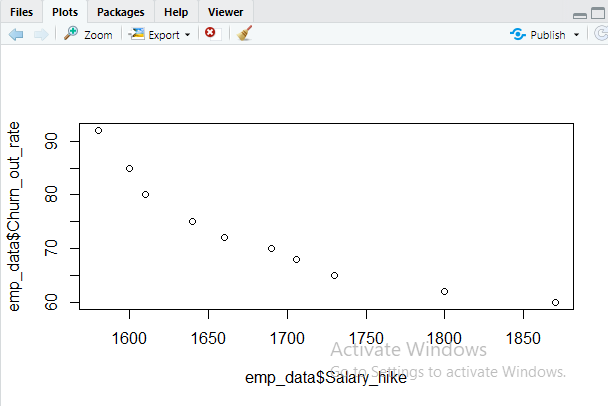
In predict.lm(reg, interval = "predict") :

predictions on current data refer to \_future\_ responses

Plot Observation: 1. Direction-Negative

2.Strength of Correlation -Moderate Negative Correlation

3.Linear



Y=B0+B1X, Here, B0+B1=Coefficient or Estimation or Parameter we can say. And B0 is intercept.

**Fitted Prediction Equation:**

Churn out Rate= 244.36491+(-0.10154)(Salary hike)

Multiple R2>0.8 this is thumb rule

R2=0.83

F-statistic:39.4 on 1 and 8 DF,p-value: 0.0002386

confint(reg,level=0.95)

**Lower Limit-2.5% Prediction Equation**

Churn out Rate=181.29+(-0.1388454) (Salary Hike)

**Upper Limit-97.5% Prediction Equation**

Churn out Rate=307.438+(-0.0642399) (Salary Hike)

predict(reg,interval="predict")Note: Here as per formula I have taken Predict so values will slight change because we need to put predict(reg,interval=”confint”)

**fit**  **lwr upr**

1 83.92753 72.38391 95.47115

2 81.89668 70.59327 93.20009

3 80.88125 69.68123 92.08127

4 77.83497 66.87456 88.79538

5 75.80412 64.94216 86.66607

6 72.75784 61.94828 83.56740

7 71.13316 60.30425 81.96206

8 68.69613 57.77694 79.61533

9 61.58815 50.00746 73.16884

10 54.48016 41.72742 67.23290

**Analysis: Churn out rate Prediction model is good**