**Linear Regression Assignment**

**Problem Statement: Salary\_hike -> Build a prediction model for Salary\_hike**

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. r= 0.97 it is strong.

4. Coefficient of determination is R2(0 to 1) thumb rule if R2>0.8 then it is strong. R2= 0.957 this is also strong.

I have done the analysis with R studio as follows:

R-script

# Load Salary\_Data.csv dataset

library(readr)

Salary\_Data <- read\_csv("Data Science Assignments/Linear Regression/Dataset/Salary\_Data.csv")

View(Salary\_Data)

attach(Salary\_Data)

# Exploratory data analysis

summary(Salary\_Data)

#Scatter plot

plot(Salary\_Data$YearsExperience, Salary\_Data$Salary) # plot(x,y)

#Correlation Coefficient (r)

cor(YearsExperience, Salary) # cor(X,Y)

# Simple Linear Regression model

reg <- lm(Salary\_Data$Salary ~ Salary\_Data$YearsExperience) # lm(Y ~ X)

summary(reg)

confint(reg,level=0.95)

predict(reg,interval="predict")

**Console Window**

|  |  |  |  |
| --- | --- | --- | --- |
| > # Load Salary\_Data.csv dataset  > library(readr)  > Salary\_Data <- read\_csv("Data Science Assignments/Linear Regression/Dataset/Salary\_Data.csv")  Parsed with column specification:  cols(  YearsExperience = col\_double(),  Salary = col\_double()  )  > View(Salary\_Data)  > attach(Salary\_Data)  The following objects are masked from Salary\_Data (pos = 3):  Salary, YearsExperience  > # Exploratory data analysis  > summary(Salary\_Data)  YearsExperience Salary  Min. : 1.100 Min. : 37731  1st Qu.: 3.200 1st Qu.: 56721  Median : 4.700 Median : 65237  Mean : 5.313 Mean : 76003  3rd Qu.: 7.700 3rd Qu.:100545  Max. :10.500 Max. :122391  > #Scatter plot  > plot(Salary\_Data$YearsExperience, Salary\_Data$Salary) # plot(x,y)  > #Correlation Coefficient (r)  > cor(YearsExperience, Salary) # cor(X,Y)  [1] 0.9782416  > # Simple Linear Regression model  > reg <- lm(Salary\_Data$Salary ~ Salary\_Data$YearsExperience) # lm(Y ~ X)  > summary(reg)  Call:  lm(formula = Salary\_Data$Salary ~ Salary\_Data$YearsExperience)  Residuals:  Min 1Q Median 3Q Max  -7958.0 -4088.5 -459.9 3372.6 11448.0  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*  Salary\_Data$YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 5788 on 28 degrees of freedom  Multiple R-squared: 0.957, Adjusted R-squared: 0.9554  F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16  > confint(reg,level=0.95)  2.5 % 97.5 %  (Intercept) 21136.061 30448.34  Salary\_Data$YearsExperience 8674.119 10225.81  > predict(reg,interval="predict")  fit lwr upr  1 36187.16 23698.92 48675.40  2 38077.15 25628.63 50525.67  3 39967.14 27556.52 52377.76  4 44692.12 32368.22 57016.03  5 46582.12 34289.64 58874.59  6 53197.09 40999.70 65394.48  7 54142.09 41956.37 66327.80  8 56032.08 43868.25 68195.91  9 56032.08 43868.25 68195.91  10 60757.06 48639.42 72874.70  11 62647.05 50544.46 74749.65  12 63592.05 51496.24 75687.86  13 63592.05 51496.24 75687.86  14 64537.05 52447.52 76626.57  15 68317.03 56247.70 80386.36  16 72097.02 60039.93 84154.10  17 73987.01 61933.05 86040.96  18 75877.00 63824.18 87929.82  19 81546.98 69485.57 93608.39  20 82491.97 70427.39 94556.56  21 90051.94 77944.06 102159.83  22 92886.93 80754.66 105019.20  23 100446.90 88228.15 112665.65  24 103281.89 91022.76 115541.02  25 108006.87 95670.98 120342.77  26 110841.86 98454.30 123229.42  27 115566.84 103084.00 128049.68  28 116511.84 104008.59 129015.09  29 123126.81 110468.27 135785.35  30 125016.80 112309.98 137723.63  Warning message:  In predict.lm(reg, interval = "predict") :  predictions on current data refer to \_future\_ responses  Scatter Diagram     1. Direction-Positive 2. Strength of Correlation-Strong 3. Linear 4. Correlation Coefficient value is r=r>0.85 it is a good relationship.   I got r>0.97-It is a good relationship   1. Coefficient of Determination is R2>0.8-is Strong   I got R2>0.957 -It is a strong   1. Y=B0+B1X    1. Point Estimation is Salary=25792.2+9450.0(Years of Exp)    2. Lower Prediction Equation is Salary=21136.061+8674.119(Years of Exp)    3. Upper Prediction Equation is Salary=30448.34+10225.81(Years of exp)   **Analysis: Build a prediction model for Salary\_hike is good.**   |  | | --- | |  | |  | |  | |
|  |
| |  | | --- | |  | |