VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Essentials of Data Analytics - (CSE3506)

Faculty – Lakshmi Pathi Jakkamputi Sir

Lab-1

Harshanth k Prakash

19BCE1293

L21-22 Slot

Tasks for Week-1: Regression

Understand the following operations/functions on random dataset and perform similar operations on mtcars and ‘data.csv’ dataset based on given instructions.

AIM

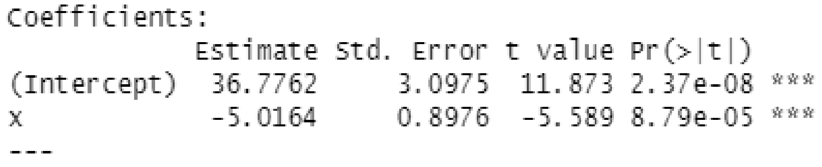
To develop linear regression model for the given data using R programming and to verify the null hypothesis.

Algorithm

1. Start
2. Read data and save to a variable.
3. Take random 50 rows of data using sample_n.
4. Store independent column data to variable ‘x’.
5. Store dependent column data to variable ‘y’.
6. Using lm function create a linear regression model between x and y
7. Summary of the model
8. Plot the linear regression line using abline.
9. Stop.

Statistic

Case 1: Mtcars dataset Linear Model.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.7762     3.0975  11.873 2.37e-08 ***
x            -5.0164     0.8976  -5.589 8.79e-05 ***
---
```

Residual standard error: 3.317 on 13 degrees of freedom

Multiple R-squared: 0.7061,   Adjusted R-squared: 0.6835

F-statistic: 31.24 on 1 and 13 DF   p-value: 8.789e-05

Pearson's product-moment correlation

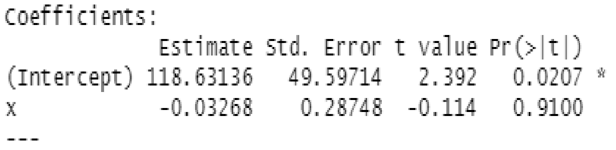
data: x and y

t = -5.5888, df = 13, p-value = 8.789e-05 alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval: -0.9455500 -0.5759782

sample estimates: cor -0.8403067

Case 2: Data.csv dataset Linear Model.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.63136   49.59714   2.392   0.0207 *
x            -0.03268    0.28748  -0.114   0.9100
---
```

Residual standard error: 34.64 on 48 degrees of freedom

Multiple R-squared: 0.0002692,   Adjusted R-squared: -0.02056

F-statistic: 0.01293 on 1 and 48 DF   p-value: 0.91

Pearson's product-moment correlation

data: x and y

t = -0.11369, df = 48, p-value = 0.91

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval: -0.2934159 0.2631413

sample estimates: cor -0.01640823

Inference

Case 1: Mtcars dataset Linear Model.

From the statistics of first case, since the p-value is less than 0.05 that is 8.789e-05, there is a significant relation between the variable, hence this model is accepted.

Case 2: Data.csv dataset Linear Model.

From the statistics of second case, since the p-value is greater than 0.05 that is 0.91, hence this model is not accepted.

Program

Case 1: Mtcars dataset Linear Model.

```
library(dplyr)

data <- mtcars

dataset <- sample_n(data,15)

dataset

x <- dataset$wt

y <- dataset$mpg

model <- lm(y~x,dataset)

summary(model)

cor.test(x,y)

par(mar=c(1,1,1,1))

plot(x,y,main = "Scatter",xlab = "weight",ylab = "mpg")

abline(model,col='red')
```

Case 2: Data.csv dataset Linear Model.

```
setwd("D:/6th Sem Works/A2- EDA")

data1 <- read.csv("data.csv")

head(data1)

data1 <- sample_n(data1,100)

x <- as.numeric(data1$Height)

y <- as.numeric(data1$Weight)

plot(x,y)

model1 <- lm(y~x)

summary(model1)

cor.test(x,y)

abline(model1,col="red")
```