Vellore Institute of Technology

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

ABHINAV VIJAYAKUMAR

19BCE1311

CSE3506 – ESSENTIALS OF DATA ANALYTICS

LAB-1

DR. LAKSHMI PATHI JAKKAMPUTI (L21 + L22)

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. Set the working directory

2. Read data into a variable as a dataframe

3. Take 75% of the data for training the model

4. Take 25% of the data for testing the data

5. Find correlation between the 2 variables for additional statistics

6. Plot the data points

7. Train the linear model

8. Plot the linear model in the same graph

9. Print the summary of the model

Statistics:

i) For mtcars:

Residuals:

Min	1Q	Median	3Q	Max
-4.6037	-2.6129	-0.1983	1.3715	6.5714

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.2943	2.2919	16.71	5.50e-14
wt	-5.6437	0.7171	-7.87	7.73e-08

Residual standard error: 3.336 on 22 **degrees of freedom**

Multiple R-squared: 0.7379, **Adjusted R-squared:** 0.726

F-statistic: 61.94 on 1 and 22 DF, **p-value:** 7.733e-08

ii) For data.csv:

Residuals:

Min	1Q	Median	3Q	Max
-30.307	-13.598	1.082	13.168	28.924

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	170.562451	2.772873	61.511	<2e-16
Weight	-0.004918	0.025536	-0.193	0.847

Residual standard error: 16.22 on 373 **degrees of freedom**

Multiple R-squared: 9.944e-05, **Adjusted R-squared:** -0.002581

F-statistic: 0.03709 on 1 and 373 DF**, p-value:** 0.8474

Inference:

mtcars: The linear model is accepted because p-value(7.733e-08) is less than 0.05

Data.csv: The linear model is rejected because p-value (0. 8474) is greater than 0.05

Program:

i) For mtcars:

```
rm(list=ls())

library(dplyr)
library(Metrics)
data1 <- mtcars
```

```
## 75% of the sample size
smp_size <- floor(0.75 * nrow(mtcars))

#setting  the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(mtcars)), size = smp_size)
train <- mtcars[train_ind, ]
test <- mtcars[-train_ind, ]



cr<-cor.test(train$wt,train$mpg)
print(cr)

plot(train$wt,train$mpg,xlab = "Wt",ylab = "mpg",main="mpg VS Wt")

## Linear model
lmodel<-lm(mpg~wt,data=train)
abline(lmodel,col="red")

summary(lmodel)

predicted<-predict(lmodel,data=test)
mae(test$mpg,predicted)
```

ii) For data.csv

```
rm(list=ls())
library(dplyr)
library(Metrics)

setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/Essentials of Data Analytics/LAB/LAB 1")
```

```
data<-read.csv('data.csv')

## 75% of the sample size
smp_size <- floor(0.75 * nrow(data))

#setting  the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)
train <- data[train_ind, ]
test <- data[-train_ind, ]
cr<-cor.test(train$Height,train$Weight)
print(cr)

plot(train$Weight,train$Height,xlab = "Weight",ylab = "Height",main="Height vs Weight")

##Linear model
lmodel<-lm(Height~Weight,data=train)
abline(lmodel,col="red")
summary(lmodel)

predicted<-predict(lmodel,data=test)
mae(test$Height,predicted)
```