Lab experiment - 2

Name: Prithviraj Guntha

Reg. No.: 20BRS1188

Subject: Essentials of data analytics

Subject code: CSE3506

Professor: A Sheik Abdullah

Slot: L55+L56

1. **Perform multi linear regression on mtcars dataset.**

```
#Question 1: Performing Multiple Linear Regression on
mtcars dataset
data("mtcars")
dat=mtcars[,c("mpg","disp","hp","wt")]
head(dat)
install.packages("scatterplot3d")
library(scatterplot3d)
attach(mtcars)
scatterplot3d(wt,disp,mpg,main="3D Scatterplot")
mod=lm(mpg~wt+disp)
mod
scr = scatterplot3d(wt,disp,mpg,main="3D
scatterplot")
scr$plane3d(mod)
```

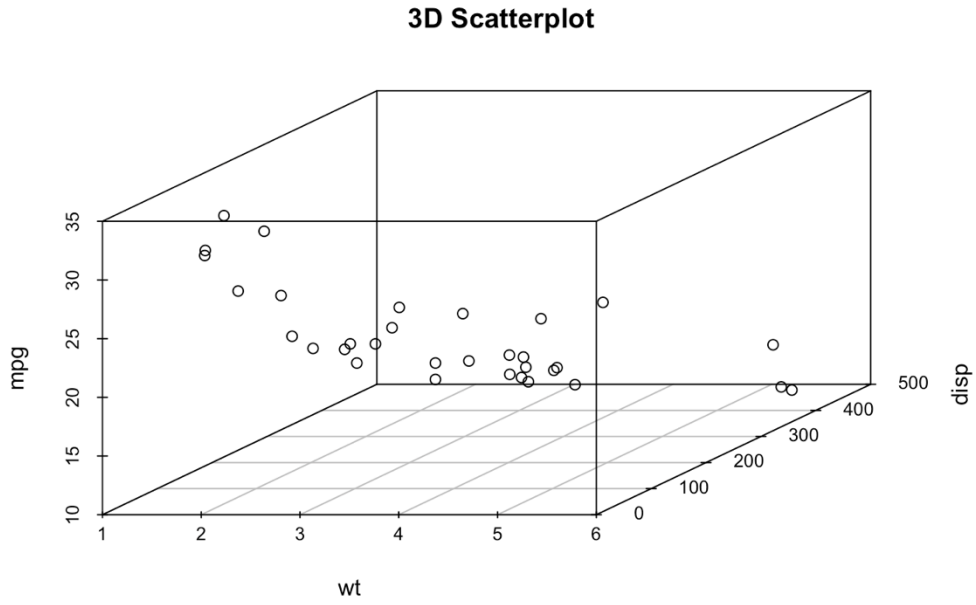
```
#Question 1: Performing Multiple Linear Regression on mtcars dataset
data("mtcars")
dat=mtcars[,c("mpg","disp","hp","wt")]
head(dat)


install.packages("scatterplot3d")
library(scatterplot3d)
attach(mtcars)
scatterplot3d(wt,disp,mpg,main="3D Scatterplot")


mod=lm(mpg~wt+disp)
mod
scr = scatterplot3d(wt,disp,mpg,main="3D scatterplot")
scr$plane3d(mod)
```

Output:

```
> head(dat)
                   mpg disp  hp    wt
Mazda RX4         21.0  160 110 2.620
Mazda RX4 Wag     21.0  160 110 2.875
Datsun 710        22.8  108  93 2.320
Hornet 4 Drive    21.4  258 110 3.215
Hornet Sportabout 18.7  360 175 3.440
Valiant           18.1  225 105 3.460
>
```

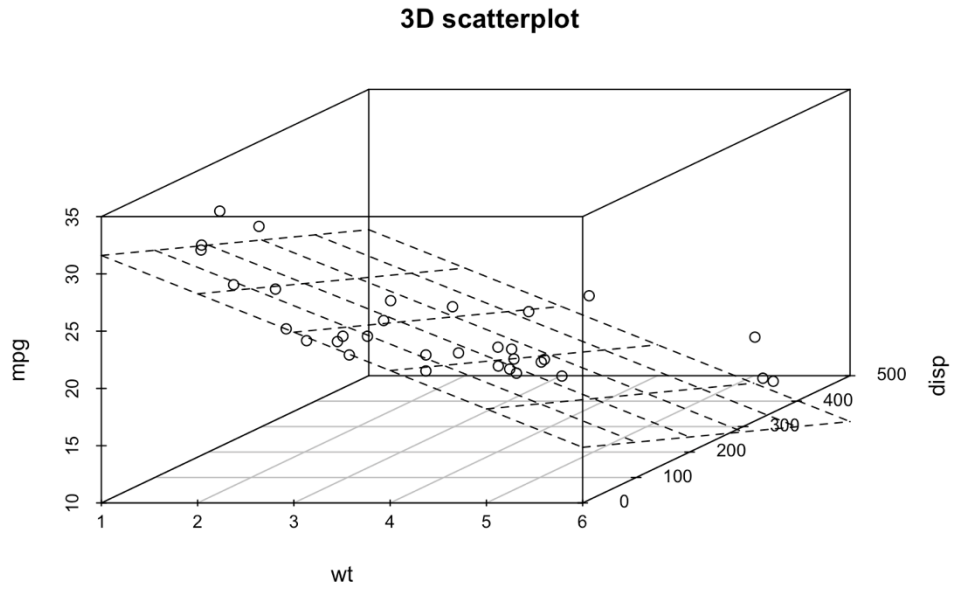


After applying the regression model,

```
> mod

Call:
lm(formula = mpg ~ wt + disp)

Coefficients:
(Intercept)           wt         disp
   34.96055     -3.35083     -0.01772
```



2. **In a random sample of 6 students each from the CSE major branch of 2014 and 2020, each student was asked about their salary package after they completed graduation. We wanted to know if the typical salary offered after graduation had changed over the past 6 years.**

```
b2014=c(667,859,1129,500,1098,1036)
b2020=c(920,1060,800,645,869,1101)

x1b=mean(b2014)
x2b=mean(b2020)
s1=sd(b2014)
s2=sd(b2020)
n1=length(b2014)
n2=length(b2020)

diff_in_means=x1b-x2b
SE_diff_mean=sqrt((s1^2/n1)+(s2^2/n2))
t_stat=diff_in_means/SE_diff_mean
t_stat

pval=2*pt(t_stat,df=n1+n2-2)
pval
```

```
#Question 2 : Testing of Hypothesis
#In a random sample of 6 students each from the CSE major branch of 2014 and
#2020, each student was asked about their salary package after they completed
#graduation. We wanted to know if the typical salary offered after graduation
#had changed over the past 6 years

b2014=c(667,859,1129,500,1098,1036)
b2020=c(920,1060,800,645,869,1101)

x1b=mean(b2014)
x2b=mean(b2020)
s1=sd(b2014)
s2=sd(b2020)
n1=length(b2014)
n2=length(b2020)

diff_in_means=x1b-x2b
SE_diff_mean=sqrt((s1^2/n1)+(s2^2/n2))
t_stat=diff_in_means/SE_diff_mean
t_stat

pval=2*pt(t_stat,df=n1+n2-2)
pval

```

Output:

```
>
> diff_in_means=x1b-x2b
> SE_diff_mean=sqrt((s1^2/n1)+(s2^2/n2))
> t_stat=diff_in_means/SE_diff_mean
> t_stat
[1] -0.1416828
>
> pval=2*pt(t_stat,df=n1+n2-2)
> pval
[1] 0.8901443
```

Interpretation:

Hence, we got the p value as 0.89 which is greater than the default assumption. By this we can say that there is not much difference between the mean pay offered to students in 2014 and 2020.

---

3. **Answer the following questions:**

a. **Out of a sample of 1000 people, 300 watched movies in theatres before the pandemic. In a sample of 1200 persons, it was discovered that 350 of them watched movies in theatres after the outbreak. We wish to investigate whether there is a decline in theatre attendance following the pandemic at the 5% level of significance.**

```
prop.test(x=c(300,350), n=c(1000,1200),,alternative =
"two.sided")
```

```
# Problem 3.1 : testing of hypothesis (z proportion-test)
#Out of a sample of 1000 people, 300 watched movies in theatres before the
#pandemic. In a sample of 1200 persons, it was discovered that 350 of them
#watched movies in theatres after the outbreak. We wish to investigate whether
#there is a decline in theater attendance following the pandemic at the 5%
#level of significance
prop.test(x=c(300,350), n=c(1000,1200),,alternative = "two.sided")
```

```
> prop.test(x=c(300,350), n=c(1000,1200),,alternative = "two.sided")

        2-sample test for equality of proportions with continuity correction

data:  c(300, 350) out of c(1000, 1200)
X-squared = 0.14414, df = 1, p-value = 0.7042
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03089872  0.04756538
sample estimates:
   prop 1    prop 2
0.3000000 0.2916667
```

Interpretation:

x-squared values is 0.144 and p-values is 0.7042. the alternate hypothesis is that there is no decline in theatre attendance. Since the obtained p-value is 0.7042 is less than the assumed level of significance 0.05, we can reject the hypothesis.

b. Perform two-proportion z-test

```
prop.test(x=c(800,900),n=c(1000,1200),alternative = "greater")
```

```
# Problem 3.2 : Perform two-proportion z-test
prop.test(x=c(800,900),n=c(1000,1200),alternative = "greater")
```

```
>
> # Problem 3.2 : Perform two-proportion z-test
> prop.test(x=c(800,900),n=c(1000,1200),alternative = "greater")

        2-sample test for equality of proportions with continuity correction

data:  c(800, 900) out of c(1000, 1200)
X-squared = 7.4826, df = 1, p-value = 0.003115
alternative hypothesis: greater
95 percent confidence interval:
 0.01983221 1.00000000
sample estimates:
prop 1 prop 2
  0.80   0.75
```

---

4. Estimate the correlation coefficient using pearson correlation and spearman rank correlation. Solve using manual method and R code.

```
x <- c(45,50,53,58,60)
y <- c(9,8,8,7,5)

#Pearson Method
cor(x,y,method = "pearson")

#Spearman Method
cor(x,y,method = "spearman")
```

```
56  #Question 4 :
57  #Estimate the correlation coefficient using
58  #pearson correlation and spearman rank correlation
59  #(solve using manual method and R code)
60  x <- c(45,50,53,58,60)
61  y <- c(9,8,8,7,5)
62
63  #Pearson Method
64  cor(x,y,method = "pearson")
65
66  #Spearman Method
67  cor(x,y,method = "spearman")
```

Output:

```
> #Question 4 :
> #Estimate the correlation coefficient using
> #pearson correlation and spearman rank correlation
> #(solve using manual method and R code)
> x <- c(45,50,53,58,60)
> y <- c(9,8,8,7,5)
>
> #Pearson Method
> cor(x,y,method = "pearson")
[1] -0.9088444
>
> #Spearman Method
> cor(x,y,method = "spearman")
[1] -0.9746794
>
```

Manual calculation:

Pearson Method:

x	y	xy	x^2	y^2
45	9	405	2025	81
50	8	400	2500	64
53	8	425	2809	64
58	7	406	3364	49
60	5	300	3600	25
266	37	1935	14298	283

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}}$$

$$= \frac{5(1935) - (266)(37)}{\sqrt{(5(14298) - (266)^2)(5(283) - (37)^2)}}$$

$$= \frac{-167}{183.74} = -0.908$$

Spearman method:

x	y	rank(x)	rank(y)	d	d^2
45	9	5	1	4	16
50	8	4	2	2	4
53	8	3	3	0	0
58	7	2	4	-2	4
60	5	1	5	-4	16
					40

$$r = 1 - \frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6(40)}{125 - 5} = 1 - \frac{240}{} = 1 - 2$$

$$= -1$$