Problem 1

Consider the following dataset:

Y1	X1	Y2	X2	Y3	X3	Y4	X4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

For the above dataset:

a. Compute correlation between Yi and Xi where i = 1, 2,3,4. Comment on it.
b. Fit the simple linear regression for each pair of Yi and Xi (i = 1, 2,3,4) and find the following:
   i. $\widehat{\beta_0}\ and\ \widehat{\beta_1}$
   ii. R^2
   iii. $t - test$
c. Comment on the results which you get.

CODE:

```
#1)

#load the dataset first
#i)
y1=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)
y1
x1=c(10,8,13,9,11,14,6,4,12,7,5)
x1
#To find the corre1ation between the var
corr1=cor(y1,x1)
corr1
fit1=lm(y1~x1) #To fit the regression model to the given data
fit1
summary(fit1) #To get all that is required

#ii)
y2=c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.1,9.13,7.26,4.74)
y2
x2=c(10,8,13,9,11,14,6,4,12,7,5)
x2
#To find the corre1ation between the var
```

```
corr2=cor(y2,x2)
corr2
fit2=lm(y2~x2) #To fit the regression model to the given data
fit2
summary(fit2) #To get all that is required

#iii)
y3=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73)
y3
x3=c(10,8,13,9,11,14,6,4,12,7,5)
x3
#To find the corre1ation between the var
corr3=cor(y3,x3)
corr3
fit3=lm(y3~x3) #To fit the regression model to the given data
fit3
summary(fit3) #To get all that is required

#iv)
y4=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89)
y4
x4=c(8,8,8,8,8,8,8,19,8,8,8)
x4
#To find the corre1ation between the var
corr4=cor(y4,x4)
corr4
fit4=lm(y4~x4) #To fit the regression model to the given data
fit4
summary(fit4) #To get all that is required


#Comment -

#To plot the scatter plots
plot(y1,x1,col="red")
plot(y2,x2,col="blue")
plot(y3,x3,col="yellow")
plot(y4,x4,col="black")
```

OUTPUT:

```
> #load the dataset first
> #i)
> y1=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)
> y1
 [1]  8.04  6.95  7.58  8.81  8.33  9.96  7.24  4.26 10.84  4.82  5.68
> x1=c(10,8,13,9,11,14,6,4,12,7,5)
> x1
 [1] 10  8 13  9 11 14  6  4 12  7  5
> #To find the corre1ation between the var
> corr1=cor(y1,x1)
```

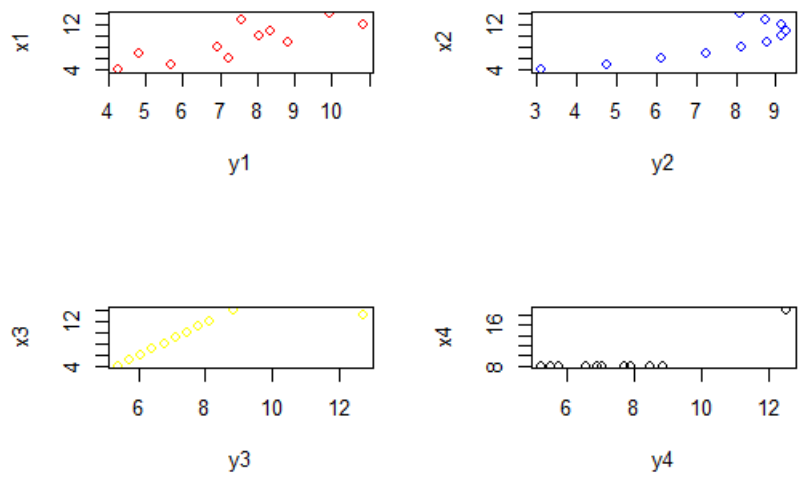
```
> corr1
[1] 0.8164205
> fit1=lm(y1~x1) #To fit the regression model to the given data
> fit1

Call:
lm(formula = y1 ~ x1)

Coefficients:
(Intercept)         x1
    3.0001      0.5001

> summary(fit1) #To get all that is required

Call:
lm(formula = y1 ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,      Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

> #ii)
> y2=c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.1,9.13,7.26,4.74)
> y2
 [1] 9.14 8.14 8.74 8.77 9.26 8.10 6.13 3.10 9.13 7.26 4.74
> x2=c(10,8,13,9,11,14,6,4,12,7,5)
> x2
 [1] 10  8 13  9 11 14  6  4 12  7  5
> #To find the corre1ation between the var
> corr2=cor(y2,x2)
> corr2
[1] 0.8162365
> fit2=lm(y2~x2) #To fit the regression model to the given data
> fit2

Call:
lm(formula = y2 ~ x2)

Coefficients:
(Intercept)         x2
```

```
    3.001       0.500

> summary(fit2) #To get all that is required

Call:
lm(formula = y2 ~ x2)

Residuals:
   Min     1Q  Median     3Q    Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,      Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179

> #iii)
> y3=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73)
> y3
 [1]  7.46  6.77 12.74  7.11  7.81  8.84  6.08  5.39  8.15  6.42  5.73
> x3=c(10,8,13,9,11,14,6,4,12,7,5)
> x3
 [1] 10  8 13  9 11 14  6  4 12  7  5
> #To find the corre1ation between the var
> corr3=cor(y3,x3)
> corr3
[1] 0.8162867
> fit3=lm(y3~x3) #To fit the regression model to the given data
> fit3

Call:
lm(formula = y3 ~ x3)

Coefficients:
(Intercept)         x3
    3.0025      0.4997

> summary(fit3) #To get all that is required

Call:
lm(formula = y3 ~ x3)

Residuals:
   Min     1Q  Median     3Q    Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,      Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

> #iv)
> y4=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89)
> y4
 [1]  6.58  5.76  7.71  8.84  8.47  7.04  5.25 12.50  5.56  7.91  6.89
> x4=c(8,8,8,8,8,8,8,19,8,8,8)
> x4
 [1]  8  8  8  8  8  8  8 19  8  8  8
> #To find the corre1ation between the var
> corr4=cor(y4,x4)
> corr4
[1] 0.8165214
> fit4=lm(y4~x4) #To fit the regression model to the given data
> fit4

Call:
lm(formula = y4 ~ x4)

Coefficients:
(Intercept)         x4
    3.0017       0.4999

> summary(fit4) #To get all that is required

Call:
lm(formula = y4 ~ x4)

Residuals:
   Min     1Q Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,      Adjusted R-squared:  0.6297
```

F-statistic: 18 on 1 and 9 DF, p-value: 0.002165

```
> #To plot the scatter plots
> par(mfrow=c(2,2))
> plot(y1,x1,col="red")
> plot(y2,x2,col="blue")
> plot(y3,x3,col="yellow")
> plot(y4,x4,col="black")
```



INTERPRETATION:

#The data is different but the slope and intercept of all the
#regression line is same i.e. the regression lines are same.
#This does not make sense because the types of the data are different in
#Each case and it is not even linear in some. Therefore these are not good fits

Problem 2

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) (Source: Gale Directory of Publications, 1994)

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between Daily and Sunday circulation? Do you think this is a plausible relationship?

(b) Fit a regression line predicting Sunday circulation from Daily circulation.

(c) Obtain the 95% confidence intervals for PO and PI.

(d) Is there a significant relationship between Sunday circulation and Daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

(e) What proportion of the variability in Sunday circulation is accounted for by Daily circulation?

(f) Provide an interval estimate (based on 95% level) for the true average Sunday circulation of newspapers with Daily circulation of 500,000.

(g) The particular newspaper that is considering a Sunday edition has a Daily circulation of 500,000. Provide an interval estimate (based on 95% level) for the predicted Sunday circulation of this paper. How does this interval differ from that given in (f)?

(h) Another newspaper being considered as a candidate for a Sunday edition has a Daily circulation of 2,000,000. Provide an interval estimate for the predicted Sunday circulation for this paper? How does this interval compare with the one given in (g)? Do you think it is likely to be accurate?

Newspaper	Daily	Sunday
1	391.952	48 8.506
2	516.981	798.298
3	355.628	235.084
4	238.555	299.451
5	537.78	559.093
6	733.775	1133.249
7	198.832	348.744
8	252.624	417.779
9	206.204	344.522
10	231.177	323.084
11	449.755	620.752
12	288.571	423.305
13	185.736	202.614
14	1164.388	1531.527

15	444.581	553.479
16	412.871	685.975
17	272.28	324.241
18	781.796	983.24
19	1209.225	1762.015
20	825.512	960.308
21	223.748	284.611
22	354.843	407.76
23	515.523	982.663
24	220.465	557
25	337.672	440.923
26	197.12	268.06
27	133.239	262.048
28	374.009	432.502
29	273.844	338.355
30	570.364	704.322
31	391.286	585.681
32	201.86	267.781
33	321.626	408.343
34	838.902	1165.567

CODE:

```
#Q2)
install.packages("readxl")
library("readxl")
data_q2=read_excel(file.choose())
View(data_q2)

#a)
plot(data_q2$Sunday,data_q2$Daily)
#The plot sugggests a linear relationship and there is positive relationship

#b)
y1=data_q2$Sunday
y1
class(y1)
length(y1)
x1=data_q2$Daily
x1
length(x1)
class(x1)
?lm
fit=lm(y1~x1)
fit
summary(fit)
coeff=fit$coefficients
coeff
```

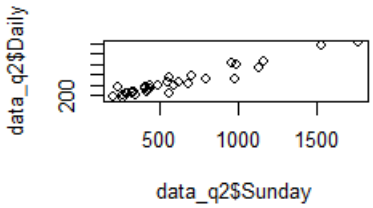
```
#the standard errors can be seen as 35.80401 and 0.07075 respectively in summary
se=c(35.80401, 0.07075)


#c)

#again we we t vaala since sig2 is unknown - Direct nikalo as Se same h rahega

#B1
model=lm(y1~x1)

ci_lower=coeff[2]-qt(0.975,32)*se[2]
ci_upper=coeff[2]+qt(0.975,32)*se[2]

CI=c(Lower=ci_lower,Upper=ci_upper)
CI

#B0 #Iska notes mai formula nahi hai so direct nikala estimate +- quantile*SE(Estimate se)
ci_lower=coeff[1]-qt(0.975,32)*se[1]
ci_upper=coeff[1]+qt(0.975,32)*se[1]
Ci_2=c(Lower=ci_lower,Upper=ci_upper)
Ci_2

#OR
confint(model,level=0.95)
#Match horahe hai

#d
cor(y1,x1)
#Positive correlation exists
#cor.test
cor.test(x1,y1)
#Correlation != 0 s 0 vala hypo i.ie null vala is rejected

#Also we can use test of significance for B1
#as B1 0 raha toh koi bhi var x ka value affect nahi karega y ko
#Notes mai dekh Anova and F test use kiya hai

model=lm(y1~x1)
summary(model)
anova(model)   #pvalue for reg vala<APLHA so rej Ho so model mai lin relationship hai as b0=0 ka
hypothesis is rejected
```

```
#e
summary(model)
#R2 value is 0.91 therefore 91% of the variability in Sunday circulation is accounted for by Daily circulation


#f
#X0=500 As voh thousands mai diya hai na'

estimate=coeff[1]+coeff[2]*500
estimate
t_quantile=qt(0.975,32)
sxx=sum(x1^2)-((sum(x1)^2)/34)
sxx
a=(500-mean(x1))^2
a
model=lm(y1~x1)
deviance(model)
ci_lower_meanresp=estimate-t_quantile*(sqrt((deviance(model)/32)*((1/34)+(a/sxx))))
ci_upper_meanresp=estimate+t_quantile*(sqrt((deviance(model)/32)*((1/34)+(a/sxx))))
ci_meanresp=c(ci_lower_meanresp,ci_upper_meanresp)
ci_meanresp


#g  #PRediction of actual value vala hai ye
x0=500
x0

model=lm(y1~x1)


#To calculate MSE - Same hai MSres ke no need of this but still did, Upar ka hi voh use kar sakte the

MSE=sum((model$residuals^2))/32
MSE
anova(model)
#MEan sum of squares hai na

yo_estimated=coeff[1]+coeff[2]*x0
ci_lower=yo_estimated-t_quantile*(sqrt(MSE*(1+(1/34)+(a/sxx))))
ci_upper=yo_estimated+t_quantile*(sqrt(MSE*(1+(1/34)+(a/sxx))))
Ci_g=c(ci_lower,ci_upper)
Ci_g
#Since in this interval the Se(estimate) is more than the previous interval due to the addition of 1,
#interval in g is bigger as its variance is larger


#h
x0=2000
b=(2000-mean(x1))^2
```

```
yo_estimated=coeff[1]+coeff[2]*x0
ci_lower=yo_estimated-t_quantile*(sqrt(MSE*(1+(1/34)+(b/sxx))))
ci_upper=yo_estimated+t_quantile*(sqrt(MSE*(1+(1/34)+(b/sxx))))
Ci_g=c(ci_lower,ci_upper)
Ci_g

#Yes the interval estimate is right.
#The size of both the intervals is same
```

OUTPUT-

```
> library("readxl")
> data_q2=read_excel(file.choose())
> #a)
> plot(data_q2$Sunday,data_q2$Daily)
```




```
#b)
> y1=data_q2$Sunday
> y1
 [1]  488.506  798.298  235.084  299.451  559.093 1133.249  348.744  417.779  344.522  323.084
[11]  620.752  423.305  202.614 1531.527  553.479  685.975  324.241  983.240 1762.015  960.308
[21]  284.611  407.760  982.663  557.000  440.923  268.060  262.048  432.502  338.355  704.322
[31]  585.681  267.781  408.343 1165.567
> class(y1)
[1] "numeric"
> length(y1)
[1] 34
> x1=data_q2$Daily
> x1
 [1]  391.952  516.981  355.628  238.555  537.780  733.775  198.832  252.624  206.204  231.177
[11]  449.755  288.571  185.736 1164.388  444.581  412.871  272.280  781.796 1209.225  825.512
[21]  223.748  354.843  515.523  220.465  337.672  197.120  133.239  374.009  273.844  570.364
[31]  391.286  201.860  321.626  838.902
> length(x1)
```

```
[1] 34
> class(x1)
[1] "numeric"
>
?lm
> fit=lm(y1~x1)
> fit

Call:
lm(formula = y1 ~ x1)

Coefficients:
(Intercept)         x1
     13.84        1.34

> summary(fit)

Call:
lm(formula = y1 ~ x1)

Residuals:
   Min     1Q  Median     3Q    Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
x1          1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,      Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16

> coeff=fit$coefficients
> coeff
(Intercept)         x1
  13.835630    1.339715
> #the standard errors can be seen as 35.80401 and 0.07075 respectively in summary
> se=c(35.80401, 0.07075)



> fit=lm(y1~x1)
> fit

Call:
lm(formula = y1 ~ x1)
```

```
Coefficients:
(Intercept)         x1
     13.84        1.34

> summary(fit)

Call:
lm(formula = y1 ~ x1)

Residuals:
   Min     1Q  Median     3Q    Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
x1          1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,      Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16

> coeff=fit$coefficients
> coeff
(Intercept)         x1
  13.835630    1.339715
> #the standard errors can be seen as 35.80401 and 0.07075 respectively in summary
> se=c(35.80401, 0.07075)
>
>
> #B1
> model=lm(y1~x1)
> ci_lower=coeff[2]-qt(0.975,32)*se[2]
> ci_upper=coeff[2]+qt(0.975,32)*se[2]
> CI=c(Lower=ci_lower,Upper=ci_upper)
> CI
Lower.x1 Upper.x1
1.195602 1.483828
> #B0 #Iska notes mai formula nahi hai so direct nikala estimate +- quantile*SE(Estimate se)
> ci_lower=coeff[1]-qt(0.975,32)*se[1]
> ci_upper=coeff[1]+qt(0.975,32)*se[1]
> Ci_2=c(Lower=ci_lower,Upper=ci_upper)
> Ci_2
Lower.(Intercept) Upper.(Intercept)
       -59.09475          86.76601
> #OR
> confint(model,level=0.95)
                2.5 %    97.5 %
(Intercept) -59.094743 86.766003
```

```
x1          1.195594  1.483836
> #d
> cor(y1,x1)
[1] 0.9581543
> #Positive correlation exists
> #cor.test
> cor.test(x1,y1)

           Pearson's product-moment correlation

data:  x1 and y1
t = 18.935, df = 32, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9171632 0.9790826
sample estimates:
     cor
0.9581543

> model=lm(y1~x1)
> summary(model)

Call:
lm(formula = y1 ~ x1)

Residuals:
   Min     1Q  Median     3Q    Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
x1          1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,      Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16

> anova(model)   #pvalue for reg vala<APLHA so rej Ho so model mai lin relationship hai as b0=0
ka hypothesis is rejected
Analysis of Variance Table

Response: y1
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 4292653 4292653  358.53 < 2.2e-16 ***
Residuals 32  383136   11973
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #e
> summary(model)

Call:
lm(formula = y1 ~ x1)

Residuals:
   Min     1Q  Median     3Q    Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
x1           1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,      Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16

> estimate=coeff[1]+coeff[2]*500
> estimate
(Intercept)
   683.693
> t_quantile=qt(0.975,32)
> sxx=sum(x1^2)-((sum(x1)^2)/34)
> sxx
[1] 2391669
> a=(500-mean(x1))^2
> a
[1] 4766.18
> model=lm(y1~x1)
> deviance(model)
[1] 383135.5
> ci_lower_meanresp=estimate-t_quantile*(sqrt((deviance(model)/32)*((1/34)+(a/sxx))))
> ci_upper_meanresp=estimate+t_quantile*(sqrt((deviance(model)/32)*((1/34)+(a/sxx))))
> ci_meanresp=c(ci_lower_meanresp,ci_upper_meanresp)
> ci_meanresp
(Intercept) (Intercept)
  644.1951    723.1910
> #g  #PRediction of actual value vala hai ye
> x0=500
> x0
[1] 500
> model=lm(y1~x1)
> MSE=sum((model$residuals^2))/32
> MSE
[1] 11972.99
> anova(model)
Analysis of Variance Table
```

```
Response: y1
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 4292653 4292653  358.53 < 2.2e-16 ***
Residuals 32  383136   11973
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> yo_estimated=coeff[1]+coeff[2]*x0
> ci_lower=yo_estimated-t_quantile*(sqrt(MSE*(1+(1/34)+(a/sxx))))
> ci_upper=yo_estimated+t_quantile*(sqrt(MSE*(1+(1/34)+(a/sxx))))
> Ci_g=c(ci_lower,ci_upper)
> Ci_g
(Intercept) (Intercept)
  457.3367    910.0493
> #h
> x0=2000
> b=(2000-mean(x1))^2
> yo_estimated=coeff[1]+coeff[2]*x0
> ci_lower=yo_estimated-t_quantile*(sqrt(MSE*(1+(1/34)+(b/sxx))))
> ci_upper=yo_estimated+t_quantile*(sqrt(MSE*(1+(1/34)+(b/sxx))))
> Ci_g=c(ci_lower,ci_upper)
> Ci_g
(Intercept) (Intercept)
  2373.463    3013.068
```

INTERPRETATION:

a) The plot sugggests a linear relationship and there is positive relationship
b) Model parameter estimates are B0= 13.835630, B1= 1.339715
c) 95% CI for B1 is (1.195602, 1.483828) and 95%c CI for B0 is (-59.09475, 86.76601)
d) Positive correlation exists and it can be confirmed using cor.test and also using ANOVA and testing for B1=0
e) R^2 value is 0.91 therefore 91% of the variability in Sunday circulation is accounted for by Daily circulation
f) Required 95% CI is (644.1951 ,723.1910)
g) Required 95% CI is (457.3367, 910.0493). #Since in this interval the Se(estimate) is more than the previous interval due to the addition of 1, interval in g is bigger as its variance is larger.
h) Required 95% CI is (2373.463 , 3013.068). #Yes the interval estimate is right.

Problem 3

Consider the simple linear regression model $y = 50 + 10\,x + \varepsilon$ where ε is NID $(0, 16)$. Suppose that $n = 20$ pairs of observations are used to fi t this model. Generate 500 samples of 20 observations, drawing one observation for each level of $x = 1, 1.5, 2, \ldots,$ 10 for each sample.

a. For each sample compute the least - squares estimates of the slope and intercept. Construct histograms of the sample values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$. Discuss the shape of these histograms.

b. For each sample, compute an estimate of $E\,(y \mid x = 5)$. Construct a histogram of the estimates you obtained. Discuss the shape of the histogram.

c. For each sample, compute a 95% CI on the slope. How many of these intervals contain the true value $\beta_1 = 10$? Is this what you would expect?

d. For each estimate of $E\,(y \mid x = 5)$ in part b, compute the 95% CI. How many of these intervals contain the true value of $E\,(y \mid x = 5) = 100$? Is this what you would expect?

CODE:

```
#3)
#i)
#For one sample:
x=seq(1,10,0.5)
x
length(x)
#y=50+10x+Epsilon

?rnorm
#We know that y~N(50+10x,16)
y=rnorm(19,50+10*x,sd=4)
y
length(y)
#so y is a sample from the given normal distribution
data=data.frame(x,y)
fit=lm(y~x)
fit
summary(fit)
coef=fit$coefficients #extract the coeff for the given reg
coef
class(coef)

#To generate the 500 samples
#i) and iii) combined hai ye

#Since variance sig2 is unknown we use t vaala CI
```

```
x=seq(1,10,0.5)
sxx=sum(x^2)-((sum(x)^2)/19)
sxx
#Formimg empty vectors to fill them
ci_b1_lower = c()
ci_b1_upper = c()
?qt
qt(0.975,19-2) #t value

sam=500 #No. of samples to be generated
coef=matrix(nrow=sam,ncol=2)
y=c()
y_data=matrix(nrow=500,ncol=19) #next vale mai bhi same sample use karunga
dim(y_data)
for (i in 1:sam){
  y_data[i,]=rnorm(19,50+10*x,sd=4)
  fit=lm(y_data[i,]~x)
  coef[i,]=fit$coefficients   #Issey coeff milenge
  deviance(fit)     #Sum of squares od residuals dega ye
  ci_b1_lower[i]=coef[i,2] - qt(0.975,17)*(sqrt((deviance(fit)/(17*sxx))))
  ci_b1_upper[i] = coef[i,2] + qt(0.975,17)*(sqrt((deviance(fit)/(17*sxx))))

}
#Checking for coefficients
head(coef)
dim(coef)

ci=data.frame(ci_b1_lower,ci_b1_upper)
head(ci)
dim(ci)

count=0
for (i in 1:500){
  if (ci[i,1]>10 | ci[i,2]<10){
    count=count+1
  }
}
print(count)
y_data
View(y_data)

#There are 36 such cases where B1 is out of the CI and this might be because it is a 95% CI
#36/500 -5% error hai approx hai
#This is what we would expect since 5% error is allowed

#Abhi histogram banana hai

hist(coef[,1]) #B0 ka
hist(coef[,2]) #B1 ka

#ii) and iv) combined hai ye
```

```
#Finf the estimates and har ek uske liye CI nikalo
#We need to find the confidence interval for har ek sample now for mean  response

x=seq(1,10,0.5)
sxx=sum(x^2)-((sum(x)^2)/19)
sxx
xbar=mean(x)
(5-xbar)^2
#Formimg empty vectors to fill them
ci_lower = c()
ci_upper = c()

sam=500 #No. of samples to be generated
Avg_y_givenx5=matrix(nrow=sam,ncol=1)
y_data
for (i in 1:sam){
  fit=lm(y_data[i,]~x)
  Avg_y_givenx5[i]=coef[i,1]+coef[i,2]*5
  deviance(fit)     #Sum of squares od residuals dega ye
  ci_lower[i] = Avg_y_givenx5[i] - qt(0.975,17)*(sqrt((deviance(fit)/17)*((1/19)+((5-xbar)^2/sxx))))
  ci_upper[i] = Avg_y_givenx5[i] + qt(0.975,17)*(sqrt((deviance(fit)/17)*((1/19)+((5-xbar)^2/sxx))))
}



#Checking for estimates of E(y|x=5)

head(Avg_y_givenx5)
dim(Avg_y_givenx5)

#FOr histogram
hist(Avg_y_givenx5)
#lOOKS NORMALLY DISTRIBUTED



ci_2 = data.frame(ci_lower,ci_upper)
head(ci_2)
dim(ci_2)


count=0
for (i in 1:500){
  if (ci_2[i,1]>100 | ci_2[i,2]<100){
    count=count+1
  }
}
print(count)
24/500
#There are 24 such cases where B1 is out of the CI and this might be because it is a 95% CI
#24/500- 5% error hai approx hai
```

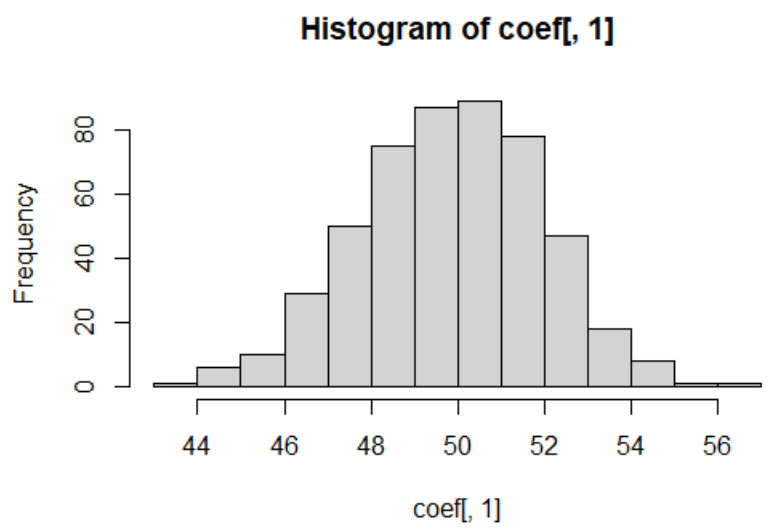
#This is what we would expect since 5% error is allowed

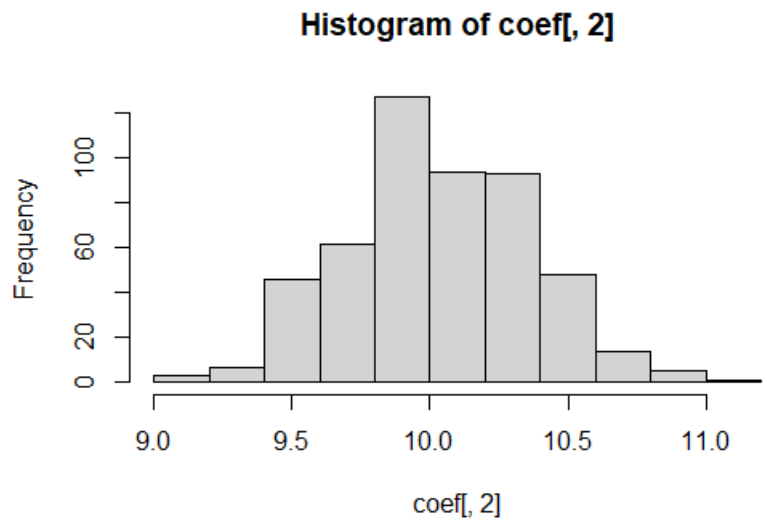
OUTPUT:

```
> #3)
> #i)
> #For one sample:
> x=seq(1,10,0.5)
> x
 [1]  1.0  1.5  2.0  2.5  3.0  3.5  4.0  4.5  5.0  5.5  6.0  6.5  7.0  7.5  8.0  8.5  9.0  9.5
[19] 10.0
> length(x)
[1] 19
> ?rnorm
> #We know that y~N(50+10x,16)
> y=rnorm(19,50+10*x,sd=4)
> y
 [1]  56.30308  64.88367  71.76312  74.96518  84.28239  89.63704  96.16953  96.64161 103.65810
[10] 108.55649 110.42261 117.25518 126.67396 123.82199 131.57705 137.59056 143.50384
147.09503
[19] 148.44940
> length(y)
[1] 19
> #so y is a sample from the given normal distribution
> data=data.frame(x,y)
> fit=lm(y~x)
> fit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
     51.48       10.10

> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q  Median     3Q    Max
-5.2733 -1.6901  0.0897  1.6184  4.5152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.4793     1.3931   36.95   <2e-16 ***
x            10.0971     0.2267   44.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

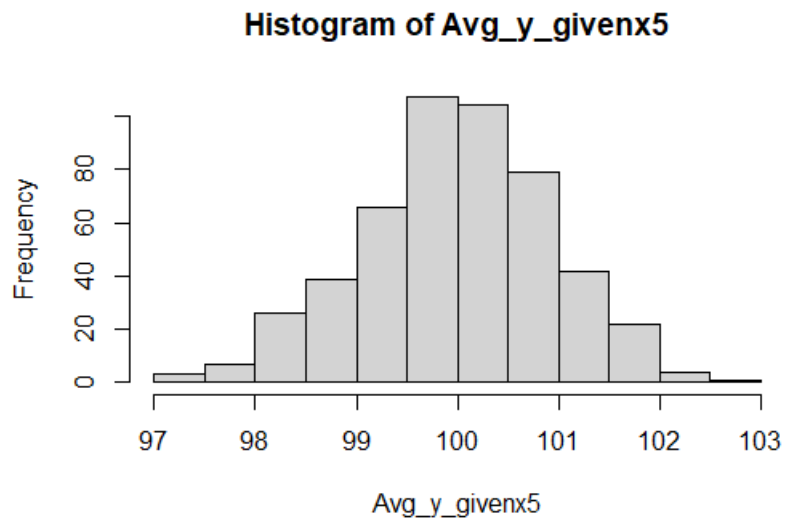
```
Residual standard error: 2.707 on 17 degrees of freedom
Multiple R-squared:  0.9915,      Adjusted R-squared:  0.991
F-statistic:  1983 on 1 and 17 DF,  p-value: < 2.2e-16

> coef=fit$coefficients #extract the coeff for the given reg
> coef
(Intercept)         x
  51.47930   10.09706
> class(coef)
[1] "numeric"
> x=seq(1,10,0.5)
> sxx=sum(x^2)-((sum(x)^2)/19)
> sxx
[1] 142.5
> #Formimg empty vectors to fill them
> ci_b1_lower = c()
> ci_b1_upper = c()
> qt(0.975,19-2) #t value
[1] 2.109816
> sam=500 #No. of samples to be generated
> coef=matrix(nrow=sam,ncol=2)
> y=c()
> y_data=matrix(nrow=500,ncol=19) #next vale mai bhi same sample use karunga
> dim(y_data)
[1] 500  19
> for (i in 1:sam){
+   y_data[i,]=rnorm(19,50+10*x,sd=4)
+   fit=lm(y_data[i,]~x)
+   coef[i,]=fit$coefficients   #Issey coeff milenge
+   deviance(fit)     #Sum of squares od residuals dega ye
+   ci_b1_lower[i]=coef[i,2] - qt(0.975,17)*(sqrt((deviance(fit)/(17*sxx))))
+   ci_b1_upper[i] = coef[i,2] + qt(0.975,17)*(sqrt((deviance(fit)/(17*sxx))))
+
+ }
> #Checking for coefficients
> head(coef)
        [,1]     [,2]
[1,] 49.71462  9.959250
[2,] 52.18389  9.715331
[3,] 47.43811 10.430403
[4,] 49.95620 10.220397
[5,] 49.01979 10.286074
[6,] 50.69933  9.907007
> dim(coef)
[1] 500   2
> ci=data.frame(ci_b1_lower,ci_b1_upper)
> head(ci)
  ci_b1_lower ci_b1_upper
1    9.253681    10.66482
2    9.061538    10.36912
3    9.890836    10.96997
```

```
4    9.550267    10.89053
5    9.577833    10.99432
6    9.079728    10.73429
> dim(ci)
[1] 500   2
> count=0
> for (i in 1:500){
+   if (ci[i,1]>10 | ci[i,2]<10){
+     count=count+1
+   }
+ }
> print(count)
[1] 36

#There are 36 such cases where B1 is out of the CI and this might be because it is a 95% CI
#36/500 -5% error hai approx hai
#This is what we would expect since 5% error is allowed

> hist(coef[,1]) #B0 ka
> hist(coef[,2]) #B1 ka
```






```
> x=seq(1,10,0.5)
> sxx=sum(x^2)-((sum(x)^2)/19)
> sxx
[1] 142.5
> xbar=mean(x)
> (5-xbar)^2
[1] 0.25
> #Formimg empty vectors to fill them
> ci_lower = c()
> ci_upper = c()
> sam=500 #No. of samples to be generated
> Avg_y_givenx5=matrix(nrow=sam,ncol=1)
> for (i in 1:sam){
+   fit=lm(y_data[i,]~x)
+   Avg_y_givenx5[i]=coef[i,1]+coef[i,2]*5
+   deviance(fit)     #Sum of squares od residuals dega ye
+   ci_lower[i] = Avg_y_givenx5[i] - qt(0.975,17)*(sqrt((deviance(fit)/17)*((1/19)+((5-
xbar)^2/sxx))))
+   ci_upper[i] = Avg_y_givenx5[i] + qt(0.975,17)*(sqrt((deviance(fit)/17)*((1/19)+((5-
xbar)^2/sxx))))
+ }
> head(Avg_y_givenx5)
         [,1]
[1,]  99.51087
[2,] 100.76055
[3,]  99.59012
[4,] 101.05818
[5,] 100.45016
[6,] 100.23437
> dim(Avg_y_givenx5)
[1] 500   1
> #FOr histogram
> hist(Avg_y_givenx5)
```




```
> ci_2 = data.frame(ci_lower,ci_upper)
> head(ci_2)
  ci_lower ci_upper
1 97.54665 101.4751
2 98.94047 102.5806
3 98.08803 101.0922
4 99.19262 102.9237
5 98.47850 102.4218
6 97.93132 102.5374
> dim(ci_2)
[1] 500   2
> count=0
> for (i in 1:500){
+   if (ci_2[i,1]>100 | ci_2[i,2]<100){
+     count=count+1
+   }
+ }
> print(count)
[1] 24
24/500
#There are 24 such cases where B1 is out of the CI and this might be because it is a 95% CI
#24/500- 5% error hai approx hai
#This is what we would expect since 5% error is allowed
```

INTERPRETATION:

a) The histograms are symmetric suggesting a normal distribution for B0 and B1.
b) The histograms is symmetric suggesting a normal distribution
c) #There are 36 such cases where B1 is out of the CI and this might be because it is a 95% CI ,36/500 -5% error hai approx hai, This is what we would expect since 5% error is allowed
d) There are 24 such cases where B1 is out of the CI and this might be because it is a 95% CI 24/500- 5% error hai approx hai, This is what we would expect since 5% error is allowed