MATH 6364: Statistical Methods

Mid term

Author: Md Salman Rahman

---

problem 1:

Solution:

Matrice $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n \times 2}$

and $x^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}_{2 \times n}$

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}_{2 \times n} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}_{n \times 2}$$

$$= \begin{bmatrix} 1 \times 1 + 1 \times 1 + \cdots + 1 \times 1 & x_1 + x_2 + \cdots + x_n \\ x_1 + x_2 + \cdots + x_n & x_1^2 + x_2^2 + \cdots + x_n^2 \end{bmatrix}_{2 \times 2}$$

$$= \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}_{2 \times 2}$$

Inverse:

$$(X^T X)^{-1} = \frac{1}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}$$

The linear regression model:

$$y = \beta_0 + \beta_1 x$$

Estimates of β_0 and β_1:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

The method of estimating β_0, β_1 by minimizing $S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$ is referred to as method of least square. There are other principle available but for this problem I will use least square estimates.

The error $\epsilon_i = y_i - \mu_i = y_i - \beta_0 - \beta_1 x_i$
$(i = 1, 2, \ldots . n)$

To obtain a line $\mu_i = \beta_0 + \beta_1 x_i$ that is closest to the point (x_i, y_i), the error ϵ_i should be as small as possible. Our aim is to minimize the function.

So, let us write,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2$$

$$= \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Now, taking derivative with respect to β_0 and β_1 and setting the derivatives to zero.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right)$$

$$\Rightarrow (+2)\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)(-1) = 0 \quad \text{------- (1)}$$

And, $$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right)$$

$$\Rightarrow (+2)\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) = 0 \quad \text{------- (2)}$$

From equation (1) and (2) $\Rightarrow$

Normal equation

$$\begin{cases} n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i \\ \left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 = \sum x_i y_i \end{cases}$$

Now solving, $n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i$

$$\Rightarrow n\beta_0 = \sum y_i - \left(\sum x_i\right)\beta_1$$

$$\Rightarrow \beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

where, $\bar{y} = \frac{\sum y_i}{n}$

$\bar{x} = \frac{\sum x_i}{n}$

$\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$

because, $\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$$= \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= \frac{\sum x_i y_i - \frac{\sum y_i}{n}\sum x_i - \frac{\sum x_i}{n}\sum y_i + n\frac{\sum x_i}{n}\frac{\sum y_i}{n}}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\frac{1}{n}\left(\sum x_i\right)^2}$$

$$= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$ (this is we got in previous page)

So, $\hat{\beta}_1 = \dfrac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$

and, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ where, $\bar{y} = \dfrac{\sum y_i}{n}$ & $\bar{x} = \dfrac{\sum x_i}{n}$

called least square estimates (LSE) of β_0 and β_1.

Finding the Variance of β_0 and β_1 :

In the previous page we find,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i - \bar{y}\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \left[\text{Note: } \sum(x_i-\bar{x}) = 0\right]$$

$$= \frac{1}{S_{xx}}\left[\sum_{i=1}^{n}(x_i-\bar{x})\,y_i\right] \quad \left[\because S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$\therefore \hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \left[\text{where, } c_i = \frac{\sum(x_i-\bar{x})}{S_{xx}}\right]$$

$$\text{Var}\left(\hat{\beta}_1\right) = \text{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2\, V(y_i)$$

$$= \sum_{i=1}^{n} c_i^2\, \sigma^2$$

Since, the y_i's are independent and $V(y_i) = \sigma^2$ is constant.

$$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \quad \left[\because \sum c_i^2 = \frac{1}{S_{xx}}\right]$$

From properties of c_i

We find, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$= \sum_{i=1}^{n} (y_i/n) - \bar{x} \sum_{i=1}^{n} c_i y_i$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} c_i \right) y_i$$

$$= \sum_{i=1}^{n} k_i y_i$$

where, $k_i = \frac{1}{n} - \bar{x} c_i$

$$= \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}\left(\sum_{i=1}^{n} k_i y_i \right)$$

$$= \sum_{i=1}^{n} k_i^2 \, \text{Var}(y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2$$

we have, $\sum_{i=1}^{n} k_i^2 = \sum \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right]^2$

$$= \sum \left[\frac{1}{n^2} - 2 \frac{1}{n} \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} + (\bar{x})^2 \left(\frac{x_i - \bar{x}}{S_{xx}} \right)^2 \right]$$

$$= \sum \left[\frac{1}{n^2} + (\bar{x})^2 \left(\frac{x_i - \bar{x}}{S_{xx}} \right)^2 \right]$$

$$\left[\because \sum (x_i - \bar{x}) = 0 \right]$$

$$= n \times \frac{1}{n^2} + (\bar{x})^2 \frac{\sum (x_i - \bar{x})^2}{S_{xx}^2}$$

$$= \frac{1}{n} + (\bar{x})^2 \frac{S_{xx}}{S_{xx}^2}$$

$$= \left[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right]$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}\right]$$

Entries of the H (hat) matrix:

We can write (i,j)th element of the hat matrix H as

$$h_{ij} = x_i^T (X^T X)^{-1} x_j$$

earlier we find,

$$(X^T X)^{-1} = \frac{1}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}\begin{bmatrix} \sum_i x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sum_i x_i^2}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} & -\frac{\sum x_i}{n\sum x_i^2 - \left(\sum_i x_i\right)^2} \\ \frac{-\sum x_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} & \frac{n}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \end{bmatrix}$$

---------- (1)

We can write this

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum(x_i-\bar{x})^2} & \frac{1}{\sum(x_i-\bar{x})^2} \end{pmatrix} = A$$

We find this by resolving each element of matrix in eq (1), Let us proof this, by reverse approach:

element 11 of matrix $(X^TX)^{-1}$ $= \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2} = \frac{1}{n} + \frac{(\bar{x})^2}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - 2\frac{(\sum x_i)^2}{n} + \frac{(\sum x_i)^2}{n}}$$

$$= \frac{1}{n} + \frac{\left(\frac{\sum x_i}{n}\right)^2}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{1}{n} + \frac{(\sum x_i)^2}{n^2} \times \frac{n}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{n\sum x_i^2 - (\sum x_i)^2 + (\sum x_i)^2}{n\left(n\sum x_i^2 - (\sum x_i)^2\right)}$$

$$= \frac{n\sum x_i^2}{n\left(n\sum x_i^2 - (\sum x_i)^2\right)}$$

This is same as first element of matrix in eq (1) ←

Now coming into second element, (A_{12}), and 3rd element (A_{21}) (both are same)

$$= -\frac{\bar{x}}{\sum (x_i - \bar{x})^2}$$

$$= -\frac{\bar{x}}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= -\frac{\frac{\sum x_i}{n}}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= -\frac{\frac{\sum x_i}{n}}{\frac{n\sum x_i^2 - 2(\sum x_i)^2 + (\sum x_i)^2}{n}}$$

$$= -\frac{\sum x_i}{n} \times \frac{n}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= -\frac{\sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Now the last element (A_{22})

$$\frac{1}{\sum (x_i - \bar{x})^2}$$

$$= \frac{1}{\sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2}$$

$$= \frac{1}{\sum x_i^2 - 2\frac{\sum x_i}{n}\sum x_i + n\left(\frac{\sum x_i}{n}\right)^2}$$

$$= \frac{1}{\frac{n\sum x_i^2 - 2(\sum x_i)^2 + (\sum x_i)^2}{n}}$$

$$= \frac{n}{n\sum x_i^2 - (\sum x_i)^2}$$

So, we can say that element of both matrix are Same so,

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{pmatrix}$$

So, $h_{ij} = x_i^T (X^T X)^{-1} x_j$

$$= \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{pmatrix} \times \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \begin{pmatrix} \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_i}{\sum (x_i - \bar{x})^2}\right) & \left(-\frac{\bar{x}}{\sum (x_i - \bar{x})^2} + \frac{x_i}{\sum (x_i - \bar{x})^2}\right) \end{pmatrix} \times \begin{pmatrix} 1 \\ x_j \end{pmatrix}$$

$$= \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_i}{\sum (x_i - \bar{x})^2} - \frac{\bar{x} x_j}{\sum (x_i - \bar{x})^2} + \frac{x_i x_j}{\sum (x_i - \bar{x})^2}\right)$$

$$= \frac{1}{n} + \frac{\bar{x}^2 - \bar{x} x_i - \bar{x} x_j + x_i x_j}{\sum (x_i - \bar{x})^2}$$

$$= \frac{1}{n} + \frac{x_j (x_i - \bar{x}) - \bar{x} (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

So, $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2}$ [Answer: showed]

Problem 2 :

Solution:

Condition : intercept $\beta_0 = 0$

This is called Regression through the origin.

Here, the regression model :

$y = \beta_1 x$ (According to question & condition)

We will estimate the slope β_1 by least square estimates.

The error $\epsilon_i = y_i - \mu_1$

$= y_i - \beta_1 x_i \quad (i = 1, 2, \ldots n)$

Our aim is to minimize the error,

So, $S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$

Now, $\frac{\partial S(\beta_1)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (y_i - \beta_1 x_i)(-x_i) = 0$

$\Rightarrow \sum_{i=1}^{n} \beta_1 x_i^2 - \sum_{i=1}^{n} x_i y_i = 0$

$\Rightarrow \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

$\therefore \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ Answere [Showed]

Problem 3:

Solution:

Given three models

(i) $E(y_i) = \beta_0 + \beta_1 x_{1i}$

(ii) $E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

(iii) $E(y_i) = \beta_0 + \beta_2 x_{2i}$

where, x_{1i} = right leg strength

x_{2i} = left leg strength.

(a) - plot of residuals for these three models is given in next page.

* Note: In data cleaning step, I convert the response (y) into one unit feet

- plot aganist $\hat{y}$ is also shown in the next page.

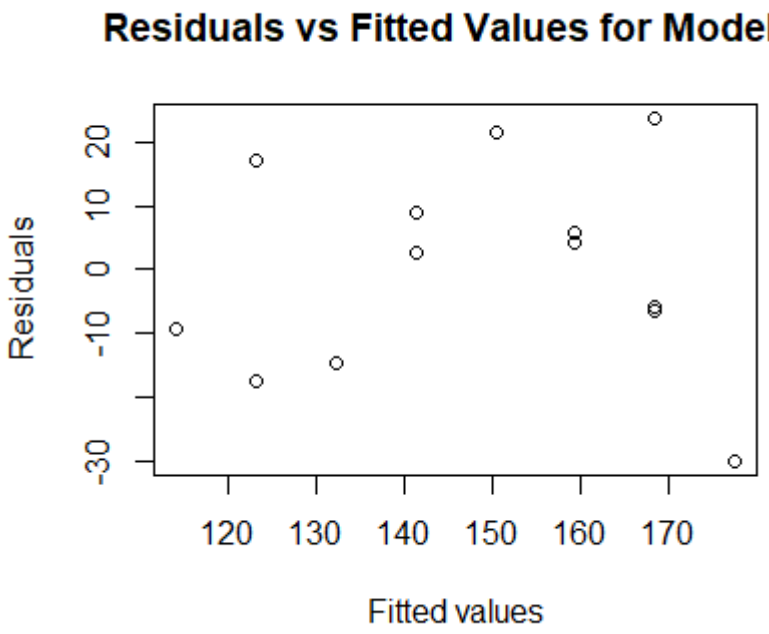
Comment on three models:

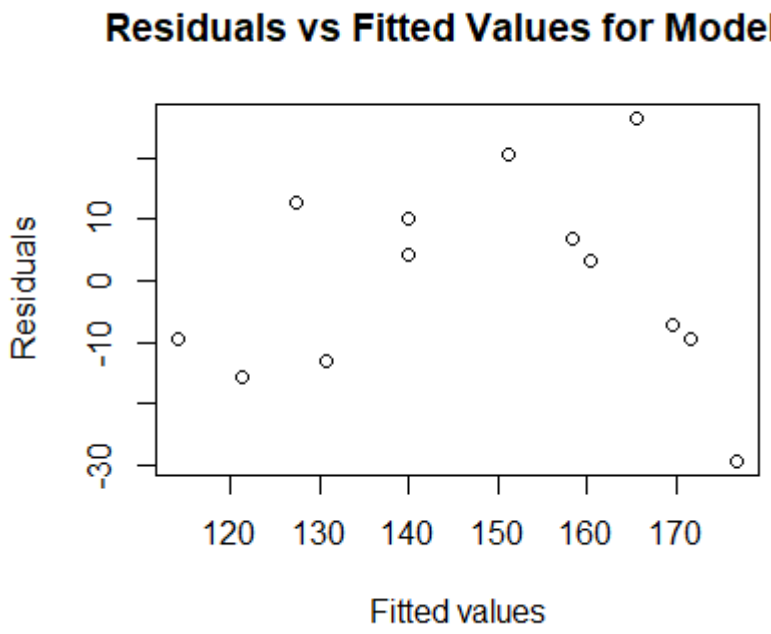
Using R programming (code is attached at the end of problem)

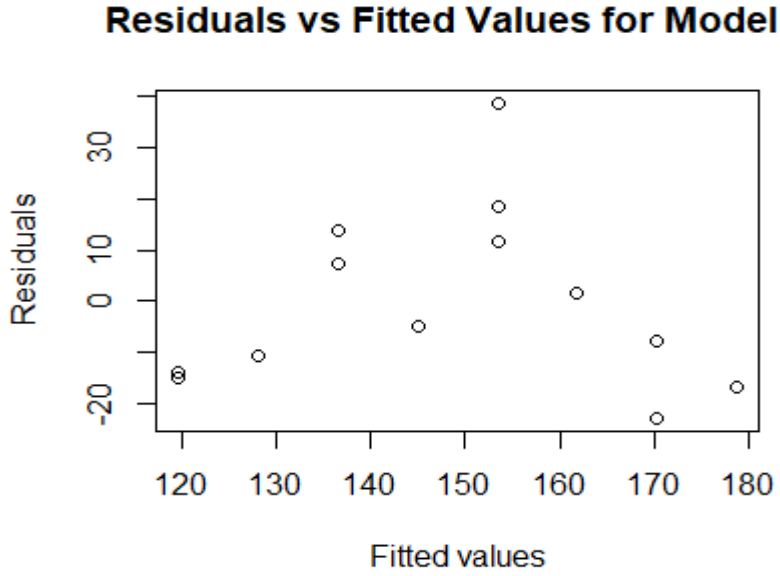
(i) For first model - Coefficient $\beta_0 = 14.9093$

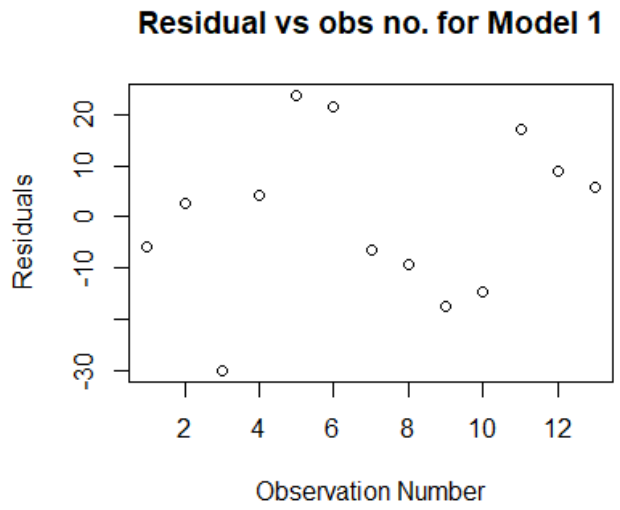
$\beta_1 = 0.9027$

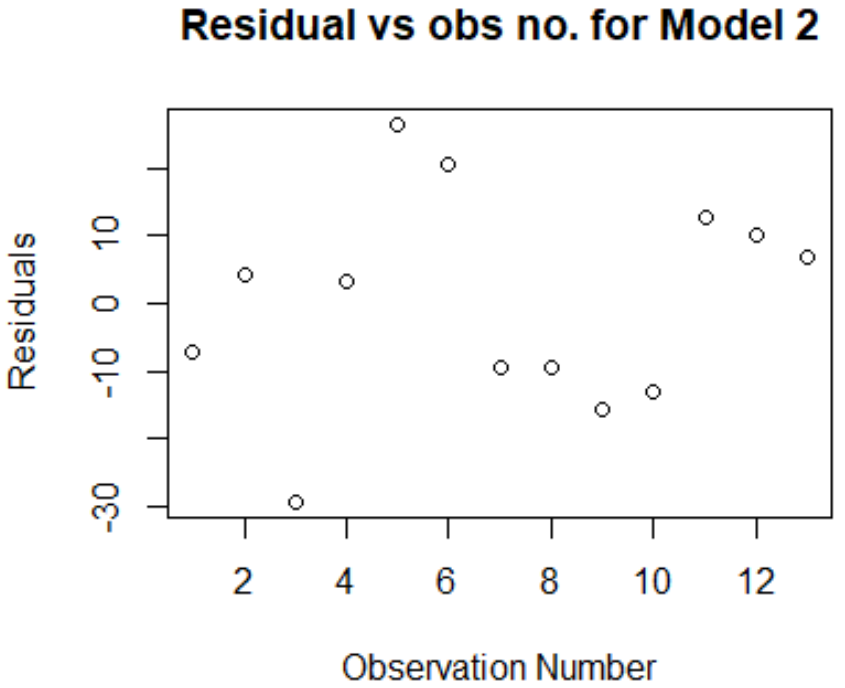
and Residual standard error = 16.58 on 11 DF (Degree of freedom)

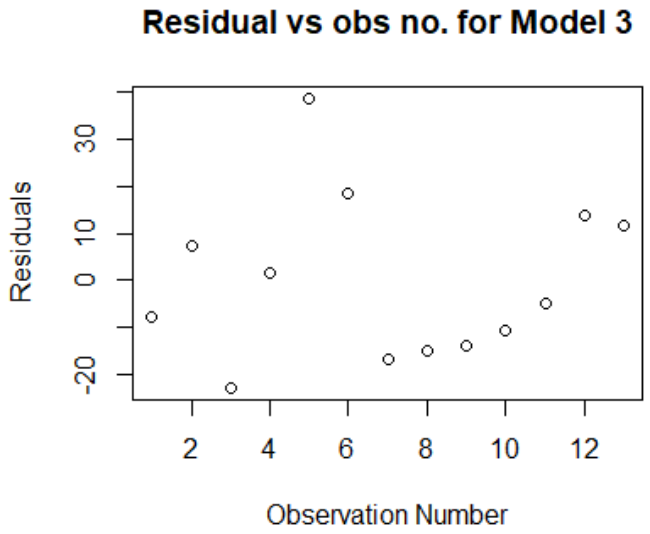

Residuals vs Fitted Values for Model 1
Residuals
-30
-10
0
10
20
120
130
140
150
160
170
Fitted values


Residuals vs Fitted Values for Model 2
Residuals
-30
-10
0
10
120
130
140
150
160
170
Fitted values


Residuals vs Fitted Values for Model 3
Residuals
-20
0
10
30
120
130
140
150
160
170
180
Fitted values


Residual vs obs no. for Model 1
Residuals
-30
-10
0
10
20
2
4
6
8
10
12
Observation Number


Residual vs obs no. for Model 2
Residuals
-30
-10
0
10
2
4
6
8
10
12
Observation Number


Residual vs obs no. for Model 3
Residuals
-20
0
10
30
2
4
6
8
10
12
Observation Number

Also for model (i),

Adjusted $R^2 = 0.5926$

Multiple $R^2 = 0.6266$ (multiple means multiple regression R^2)

F statistics $= 18.46$

p value $= 0.001264$ (from regression model)

(ii) For second model:

Coefficient of regression,

$\beta_0 = 12.7483$

$\beta_1 = 0.7213$

$\beta_2 = 0.2012$

And, Residual standard error $= 17.25$ on 10 DF

Multiple $R^2 = 0.6328$

Adjusted $R^2 = 0.5594$

F statistics $= 8.617$

p value $= 0.006675$

(iii) For third model:

Coefficient of Regression, $\beta_0 = 26.9112$

$\beta_1 = 0.8434$

Residual Standard error = 18.13

Multiple $R^2 = 0.5537$

Adjusted $R^2 = 0.5131$

F statistics = 13.65

P value = 0.003537

Comparison:

From the residual plot, we know that positive value for the residual (on y axis of Residual vs Fitted values plot) means the prediction was too low, and negative means prediction is too high. 0 means (in the residuals) the guess was exactly correct.

From the above data & residual plot the three model look alike having some outlier. There is no noticeable normality among them. Hence analyzing all the information I will say model 2 is more appropriate having largest R^2 and better fitting.

3 (b) :

We will test the hypothesis for model 2 & model 3 as β_2 is involved in this two model:

Here Null Hypothesis : $H_0 : \beta_2 = 0$

Alternative " $H_1 : \beta_2 = 0$

(i) Model 2 :

From 3(a) we find (using R) [All the R code attach at the end of problem]

$\hat{\beta}_2 = 0.2012$ with

t value = 0.412

and p value = 0.689 > 0.05

So, there is not enough evidence to reject the null hypothesis $\beta_2 = 0$.

Also, we can verify the result using ANOVA Table. (more details in R code)

Source	DF	SS	MS	F	P
$\hat{\beta}_1$	1	5076.7	5076.7	17.0644	0.002041
$\hat{\beta}_2$	1	50.5	50.5	0.1696	0.689115
Residuals	10	2975	297.5		

Also we can see that the confidence interval for β_2 is : $(-0.8871681, 1.289544)$

we see that this interval contains 0 inside it.

So, we can conclude, from p value, Anova table, and CI that, there is not enough evidence to reject null hypothesis for model 2.

(ii) Model 3 :

For model 3, $\hat{\beta}_2 = 0.8434$

t value = 3.694

p value = $0.00354 < 0.05$

So, we can say that, the null hypothesis is rejected. $\hat{\beta}_2$ is statistically different from 0.

Verifying the result using ANOVA Table.

Source	DF	SS	MS	F	P
$\hat{\beta}_2$	1	4486.3	4486.3	13.648	0.003537
Residuals	11	3615.9	328.7		

Also, the confidence interval for β_2 is (0.34090, 1.345807) and we see that CI doesn't contain 0.

From Anova table (with small p value), CI we can conclude that we reject the null hypothesis as there is not enough evidence to support it.

Final Conclusion:

we fit three model for problem 3(a) and analyze the result (All result are provided) previous, we find in problem (a) model seems most appropriate

And we test the hypothesis for β_2 in problem 3(b) and for model 3 the hypothesis is rejected and for model 2, we donot have enough evidence to reject the null hypothesis.

```
############################### Problem 3
###################################


# Author: Md Salman Rahman
# Course: MATH 6364 Statistical Methods
# Course Instructor: Dr. George Yanev


library("readxl")
library(ggplot2)
data_3<- read_excel("C:/Users/User/OneDrive - The University of Texas-Rio
Grande Valley/Course_video/Statistical Methods/HW_and R/Midterm Exam/
Exercise_2_11_data.xlsx")



########################  data Preparation and Cleaning  #################
library(tidyr)
library(dplyr)


split_data <- data_3 %>% separate(`Ave Punting Distance`, c("Ave Punting
Distance in ft","Unit_Ft", "Ave Punting Distance in Inch", "Unit_Inch"))


# removing unit column
new_data<- subset(split_data, select = -c(Unit_Ft,Unit_Inch))

final_data <- lapply(new_data,as.numeric)

df<- as.data.frame(final_data)

# converting inch into feet

inch_to_feet <- (df[,5])/12

df$Ave.Punting.Distance.in.ft = df$Ave.Punting.Distance.in.ft + inch_to_feet


# final data


data<- subset(df, select = -c(Ave.Punting.Distance.in.Inch))



## first Regression model

reg_1 <- lm(data$Ave.Punting.Distance.in.ft~ data$Right.Leg..lb.,data=data)
summary(reg_1)
```

```
## second Regression model

reg_2<-lm(data$Ave.Punting.Distance.in.ft~data$Right.Leg..lb. +
data$Left.Leg..lb.,data=data)
summary(reg_2)



## third Regression model

reg_3<-lm(data$Ave.Punting.Distance.in.ft~data$Left.Leg..lb.,data=data)
summary(reg_3)


# first residuals plot

residuals(reg_1)
plot(residuals(reg_1),xlab="Observation Number",
ylab="Residuals",main="Residual vs obs no. for Model 1")



plot(reg_1$fitted.values,reg_1$residuals,xlab="Fitted values",
ylab="Residuals",main="Residuals vs Fitted Values for Model 1")


#plot(data$Ave.Punting.Distance.in.ft,residuals(reg_1), main="First model")


# second residuals plot

plot(residuals(reg_2),xlab="Observation Number",
ylab="Residuals",main="Residual vs obs no. for Model 2")



residuals(reg_2)
plot(reg_2$fitted.values,reg_2$residuals,xlab="Fitted values",
ylab="Residuals",main="Residuals vs Fitted Values for Model 2")


# third residuals plot

plot(residuals(reg_3),xlab="Observation Number",
ylab="Residuals",main="Residual vs obs no. for Model 3")

residuals(reg_3)
plot(reg_3$fitted.values,reg_3$residuals,xlab="Fitted values",
ylab="Residuals",main="Residuals vs Fitted Values for Model 3")
```

```
# comparison

summary(reg_1)$r.squared
summary(reg_2)$r.squared
summary(reg_3)$r.squared

# anova table:
anova(reg_1)
anova(reg_2)
anova(reg_3)


# confidence interval

confint(reg_2)

confint(reg_3)
```

problem 4:

Solution:

(a) Multiple Regression Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

The residual for this multiple regression is plotted in the next page:

— The value of Residual (given in next page)

— Also, For more easy visualization I provide a Residual Vs No of obs. plot.

Also, we find the coefficient,

$\hat{\beta}_0 = -6.5122$

$\hat{\beta}_1 = 1.9994$

$\hat{\beta}_2 = -3.6751$

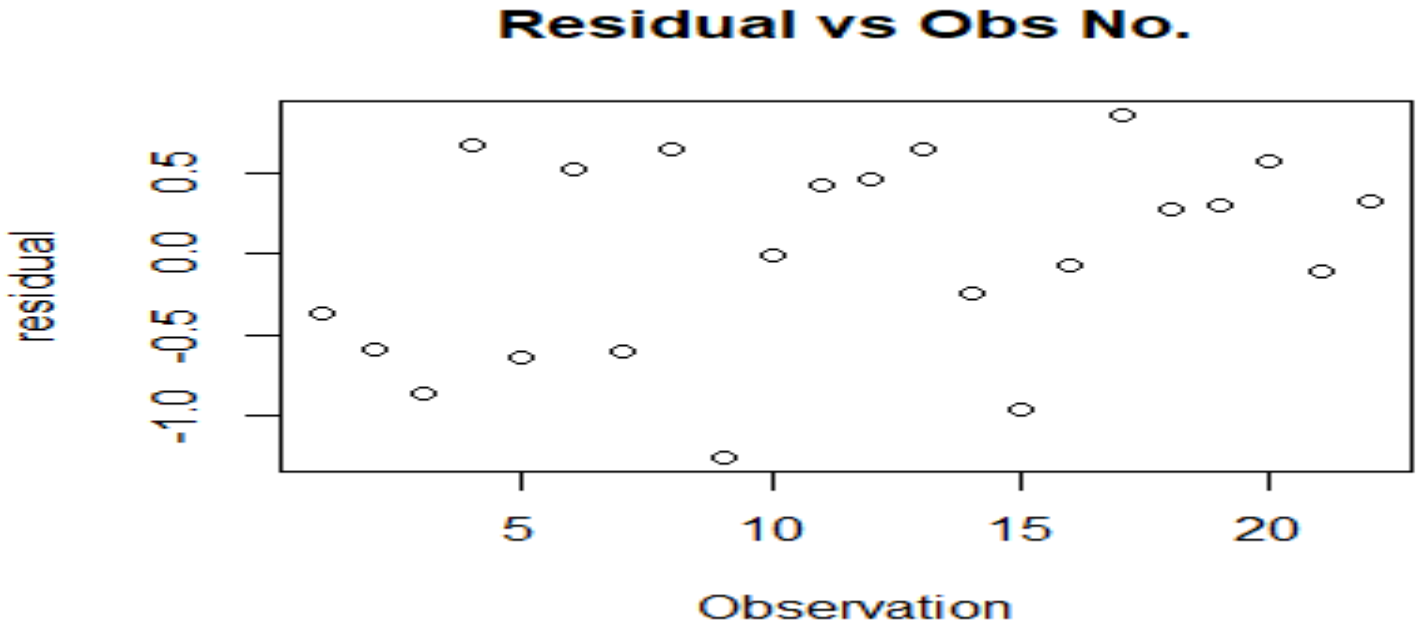
$\hat{\beta}_3 = 2.5245$

$\hat{\beta}_4 = 5.1581$

$\hat{\beta}_5 = 14.4012$

Value of Residual and Plot

1	2	3	4	5	6	7	8
-0.370312100	-0.593019254	-0.860317051	0.670587693	-0.645720945	0.528983945	-0.607669105	0.650472341
9	10	11	12	13	14	15	16
-1.261035297	-0.009077407	0.420544952	0.460955709	0.654116797	-0.244405724	-0.959780045	-0.066927381
17	18	19	20	21	22		
0.861122223	0.280081389	0.299619407	0.575658550	-0.106828243	0.322949545		



(b) 95% Confidence intervals of the mean response given in the next page, with lower & upper limit

95% prediction interval on a new observation given in the next page, (for 22 specimen)

problem 4 (a)

(i) New multiple regression using regressor x_2, x_4, x_5:

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Using R [Code is attached at the end of problem] the coefficient of regression:

$$\hat{\beta}_0 = -5.9804$$

$$\hat{\beta}_2 = -2.8898$$

$$\hat{\beta}_4 = 6.8793$$

$$\hat{\beta}_5 = 17.1090$$

s^2: From R code, Residual standard error, = 0.6783

$s^2 = 0.4600908$

Standard error of predictions:

Standard error, intercept $\hat{\beta}_0 = 0.6591$

Standard error, $\hat{\beta}_2 = 2.3465$

Standard error $\hat{\beta}_4 = 2.7828$

Standard error, $\hat{\beta}_5 = 3.0819$

[R code is given at the end]

95% confidence intervals on the mean response at the regressor location for the 22 specimen:

is shown in the next page.

Left: Confidence Interval for 22 Specimen and

Right: Confidence Interval for the five parameter

	fit	lwr	upr
1	2.45031210	1.98249330	2.9181309
2	2.57301925	2.18948436	2.9565541
3	2.76031705	2.18338303	3.3372511
4	7.88941231	6.77566738	9.0031572
5	5.13572095	4.51178948	5.7596524
6	7.96101605	6.95000340	8.9720287
7	6.77766910	6.14341040	7.4119278
8	6.88952766	6.28157215	7.4974832
9	7.62103530	6.79497800	8.4470926
10	7.63907741	6.47331169	8.8048431
11	7.35945505	6.67869945	8.0402106
12	9.68904429	8.77606341	10.6020252
13	6.22588320	5.22879401	7.2229724
14	2.19440572	1.65261926	2.7361922
15	3.85978005	2.73918687	4.9803732
16	0.78692738	-0.04314064	1.6169954
17	-0.05112222	-0.61728110	0.5150367
18	0.80991861	0.16766513	1.4521721
19	0.92038059	0.26202366	1.5787375
20	0.44434145	-0.48546460	1.3741475
21	2.03682824	1.48142716	2.5922293
22	0.31705046	-0.23987182	0.8739727

	2.5 %	97.5 %
(Intercept)	-8.491275	-4.533154
x1	-3.455820	7.454647
x2	-9.554992	2.204800
x3	-10.931602	15.980574
x4	-2.601372	12.917535
x5	4.106915	24.695409

- 95% Prediction Interval of the 22 specimen

	fit	lwr	upr
1	2.45031210	0.8874020	4.013222
2	2.57301925	1.0332360	4.112802
3	2.76031705	1.1613529	4.359281
4	7.88941231	6.0281597	9.750665
5	5.13572095	3.5192050	6.752237
6	7.96101605	6.1593557	9.762676
7	6.77766910	5.1571391	8.398199
8	6.88952766	5.2791105	8.499945
9	7.62103530	5.9162761	9.325795
10	7.63907741	5.7462374	9.531917
11	7.35945505	5.7201681	8.998742
12	9.68904429	7.9405113	11.437577
13	6.22588320	4.4319991	8.019767
14	2.19440572	0.6077848	3.781027
15	3.85978005	1.9944215	5.725139
16	0.78692738	-0.9197789	2.493634
17	-0.05112222	-1.6462301	1.543986
18	0.80991861	-0.8137571	2.433594
19	0.92038059	-0.7097321	2.550493
20	0.44434145	-1.3130352	2.201718
21	2.03682824	0.4455068	3.628150
22	0.31705046	-1.2748025	1.908903

(d) choice between full model (a) and Reduced model:

In Full model, adjusted $R^2 = 0.9519$ (model in 4(a))

In Reduced model, adjusted $R^2 = 0.9553$ (model in 4c)

Also, we perform Anova Analysis (in R code) the reduced model have the slightly better adjusted R^2 than the full model.

we use adjusted R^2 as it slightly give better accuracy in terms of model selection.

Also, The Residual Standard error for full model = 0.7035, and residual Standard error for reduced model = 0.6783.

The reduce model have the smaller standard error than the full model.

Analyzing All this evidence, I choose reduce model as a better choice

(e) Testing

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 \qquad H_1 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \neq 0$$

Here the null hypothesis describe that the coefficient $\beta_1, \beta_2, \beta_3$ will be 0 at a time.

we will test this for two model

Case- (i) Restricted model: ($\beta_1, \beta_2, \beta_3$ will be 0) is,

$$y_i = \beta_0 + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Case- (ii) Full model (4@)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Hypothesis Testing of restricted and full model

```
> model_restricted <- lm(y~x4+x5, data=s_data_frame)
> summary(model_restricted)

Call:
lm(formula = y ~ x4 + x5, data = s_data_frame)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56123 -0.49604  0.09069  0.45717  0.98057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3351     0.6009 -10.543 2.23e-09 ***
x4            4.1542     1.7104   2.429   0.0252 *
x5           15.0160     2.6057   5.763 1.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6875 on 19 degrees of freedom
Multiple R-squared:  0.9584,    Adjusted R-squared:  0.954
F-statistic: 218.9 on 2 and 19 DF,  p-value: 7.584e-14
```

```
> regg_1<-lm(y~ x1+x2+x3+x4+x5, data=s_data_frame)
>
> summary(regg_1)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = s_data_frame)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2610 -0.5373  0.1355  0.5120  0.8611

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5122     0.9336  -6.976 3.13e-06 ***
x1            1.9994     2.5733   0.777  0.44851
x2           -3.6751     2.7737  -1.325  0.20378
x3            2.5245     6.3475   0.398  0.69610
x4            5.1581     3.6603   1.409  0.17791
x5           14.4012     4.8560   2.966  0.00911 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7035 on 16 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9519
F-statistic: 84.07 on 5 and 16 DF,  p-value: 6.575e-11
```

```
> anova(regg_1,model_restricted)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5
Model 2: y ~ x4 + x5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 7.9175
2     19 8.9793 -3   -1.0617 0.7152 0.5572
>
```

Case-(i) Restricted model:

$$y_i = \beta_0 + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Using R we perform regression analysis, of restricted model.

Case (ii) - Full model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

we also fit this using regression model.

Anova table for this two model

Model 1 : $y \sim x1 + x2 + x3 + x4 + x5$

Model 2 : $y \sim x4 + x5$

Res.DF	RSS	DF	Sum of sq	F	Pr(>F)
16	7.9175				
19	8.9793	−3	−1.0617	0.7152	0.5574

(ii) For second model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

From Anova table we see that, P value is 0.5572 which is greater than 0.05 (if we consider 95% significance level). Also, F statistics is 0.7152.

Hence, we don't have enough evidence to reject the null hypothesis.

Overall conclusion: In problem 4, we develop multiple regression model with full model in (a). In 4(a) the residual plot is not fully perfect as we see some high positive and negative y axis (residual) value. In (b) we show the 95% CI & PI. and as usual PI tells where we can see the next data point sampled. In 4(c) a new model is develop with x_2, x_4, x_5. And we choose new reduced model as better choice as it have smaller residual standard error and higher adjusted R^2. Finally we test the hypothesis using Anova approach and we don't able to reject the null hypothesis.

```
################## Problem 4 ################################



# Author: Md Salman Rahman
# Course: MATH 6364 Statistical Methods
# Course Instructor: Dr. George Yanev


library("readxl")
library(ggplot2)
squid_data<- read_excel("C:/Users/User/OneDrive - The University of Texas-Rio
Grande Valley/Course_video/Statistical Methods/HW_and R/Midterm Exam/
Squid_data.xlsx")


s_data_frame<- as.data.frame(squid_data)

attach(s_data_frame)




########### (a) generating the residual for the multiple regression model
###########

regg_1<-lm(y~ x1+x2+x3+x4+x5, data=s_data_frame)

summary(regg_1)




# residuals plot

residuals(regg_1)
plot(residuals(regg_1),xlab=" Observation",ylab="residual",main="Residual vs
Obs No.")






########### (b) computing the 95% confidence interval on the mean response
###########

# confidence interval
CI <- predict(regg_1, newdata = s_data_frame, interval = 'confidence')
```

```
CI_2 <- confint(regg_1, data=s_data_frame, interval ="confidence", level=0.95)

# prediction interval
PI <- predict(regg_1, newdata = s_data_frame, interval = 'prediction')




########### (c) new multiple regression using regressor x2,x4, and x5
###########

regg_2<-lm(y~ x2+x4+x5, data=s_data_frame)

summary(regg_2)


anova(regg_1,regg_2)


# confidence interval

CI_22 <- predict(regg_2, newdata = s_data_frame, interval = 'confidence')


# hypothesis testing


##########

model_restricted <- lm(y~x4+x5, data=s_data_frame)
summary(model_restricted)


anova(regg_1,model_restricted)


# hypothesis testing 2


n= 22

beta_hat_1 = 1.9994

beta_hat_2 = -3.6751

beta_hat_3= 2.5245

se_beta_hat_1 = 2.5733

se_beta_hat_2 = 2.7737

se_beta_hat_3 = 6.3475
```

```
t_statistics = qt(1-0.05/2, df=n-3)  # 95% CI

c(beta_hat_1-t_statistics*se_beta_hat_1,
beta_hat_1+t_statistics*se_beta_hat_1)

2 * pt(2.093024, df = n-3)
```

problem 5:

Solution:

(a) Simple linear regression:

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

where y = Analytical Results.

x = quantities of Calcium in carefully prepared solution.

where $\hat{\beta}_0 = -0.228090$

$\hat{\beta}_1 = 0.994757$ [Details are in R code]

Assumption of linear regression:

(1) Regressor Variable $x_1, x_2, \ldots x_n$ are not random Variables.

(2) $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ are random Variable

and $E(\epsilon_i) = 0$, $i = 1, 2, \ldots n$

(3) $V(\epsilon_i) = \sigma^2$ is constant for all $i = 1, 2, \ldots n$

This means the variance $V(y_i) = \sigma^2$ are all the same and all observation have the same precision.

(4) ε_i and ε_j different errors, hence response y_i & y_j are independent.

This means $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$

Also, response variable (y_i) drawn from probability distribution with means $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i$ for constant variance σ^2

Besides, two observations y_i and y_j ($i \neq j$) are independent.

(b) 95% CI for the intercept :

95% confidence for the intercept is : (-0.5459503, 0.08977054)

[R code is given]

This means 95% confident interval is a range of value that we are 95% confident it contains the true value of the intercept.

5 (c) 95% Confidence Interval of the Slope :

95% CI for slope = (0.9827204, 1.006792)

This also mean, 95% CI for slope gives us range of value that we are 95% confident it contains the true value of the slope.

5(d):

(i) When $x=0$, then $y=0$ if there is no calcium present, our technique should not find any : we are looking for whether the intercept is zero or not.

From (a) we find,

$$\hat{\beta}_0 = -0.228090$$

with t value $= -1.655$

p value $= 0.137 > 0.05$

Also, From this we can say that we have not enough evidence to reject the null hypothesis ($\hat{\beta}_0 = 0$).

Also, From 5(b) we see that Confidence interval

of the intercept is $(-0.5459503, 0.0897705)$

which means that the CI contains the value 0.

So, we can conclude that if $x = 0$ then $y = 0$ as the intercept close to 0.

(ii) Here its say that if the empirical technique is any good then, the slope of regression = 1.

Let us formulate a hypothesis.

$[\beta_1$ is slope$]$

$$H_0 : \beta_1 = 1$$

$$H_A : \beta_1 \neq 1$$

From (a) we know, $\hat{\beta}_1 = 0.994757$

So, Test statistics $= \dfrac{0.994757 - 1}{S.e.(\hat{\beta}_1)}$

$= \dfrac{0.994757 - 1}{0.005219}$ $\left[Se(\hat{\beta}_1) = 0.005219 \text{ from R}\right]$

$= -1.00459$

and the pvalue is < 0.05,

So, we cannot able to reject the null hypothesis.

Also, from (a) we see that $\hat{\beta}_1 = 0.994757$ which is close to 1.

Also the CI of slope : (0.9827204, 1.006792) contains the value 1.

So, we show the evidence for both d(i) & d(ii).

5(e): if we accept (i) of 5(d) then the model become:

$$y = \beta_1 x + \varepsilon$$

So For this model, $\hat{\beta}_1 = 0.987153$ [R code is given at the end)

Re doing part (c)

95% CI is = (0.9810362, 0.9932693)

Re examine property (ii)

$H_0 : \beta_1 = 1$ (previous)

$H_A : \beta_1 \neq 1$

we see that the CI = (0.9810362, 0.9932693) which doesn't contain the value 1. So, we reject the null hypothesis.

So, the property (ii) doesn't hold in 5(e).

(f) Reason why the result for (d) & (e) are different.

The main reason is that in problem 5(d) we don't have enough evidence to reject the null hypothesis. Also we see that the CI of slope also contains the value 1 in 5(d).

But in 5(e) the CI is (0.9810362, 0.9932693) which doesn't contain the value 1.

As in the problem 5(e) we exclude the intercept which leads to poor fit of the model, which ultimately make the two result different.

```
############################ Problem 5 ########################



# Author: Md Salman Rahman
# Course: MATH 6364 Statistical Methods
# Course Instructor: Dr. George Yanev


x <-c(4,8,12.5,16,20,25,31,36,40,40)
y<-c(3.7,7.8,12.1,15.6,19.8,24.5,31.1,35.5,39.4,39.5)

data<-cbind(x,y)
d_frame <- as.data.frame(data)
########### (a) simple linear regression  ###########

reg5<-lm(y~ x)

summary(reg5)

plot(reg5)



########### (b) and (c) computing the 95% confidence interval and slope for
intercept ###########

# confidence interval
confint(reg5, level=0.95)


## (d),(e), and (f)

model_e<- lm(y~0+x, data=d_frame)

summary(model_e)

confint(model_e, level=0.95)
```