MET
Bhujbal Knowledge City
MET INSTITUTE OF COMPUTER SCIENCE

Advanced DBMS Lab
ACADEMIC YEAR 2025-2026

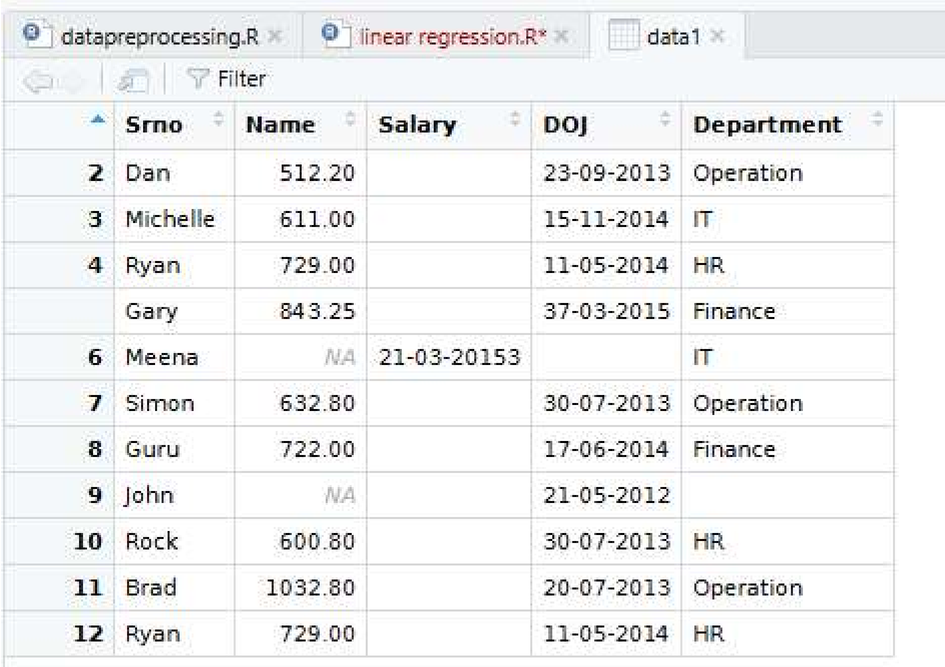
Assignment No	9
Title	Data Preprocessing & Regression
Objective	Implement data clearing using mean, median in vector and data frames
Roll No	MCA2511

```
> my_data <- mtcars
> head(mtcars, 5)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
> my_data <- my_data[1:6.1:5]

Warning message:
In 1:6.1:5 : numerical expression has 6 elements: only the first used

> install.packages("dplyr")

Restarting R session...
> install.packages("dplyr")

WARNING: Rtools is required to build R packages but is not currently
installed. Please download and install the appropriate version of Rtools
before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL
'https://cran.rstudio.com/bin/windows/contrib/4.5/dplyr_1.1.4.zip'
Content type 'application/zip' length 1593566 bytes (1.5 MB)
downloaded 1.5 MB

package ‘dplyr’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\MCA2511\AppData\Local\Temp\Rtmp84KAwx\downloaded_packages
Loading required namespace: XLConnect
```

```
> require(dplyr)

Loading required package: dplyr

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:


   intersect, setdiff, setequal, union
> my_data <- rename(my_data,horse_power=hp)
> my_data$new_hp <- my_data$horse_power*05
> colnames(my_data)
[1] "mpg"         "cyl"         "disp"        "horse_power" "drat"
[6] "new_hp"
> my_data
                     mpg cyl  disp horse_power drat new_hp
Mazda RX4           21.0   6 160.0         110 3.90    550
Mazda RX4 Wag       21.0   6 160.0         110 3.90    550
Datsun 710          22.8   4 108.0          93 3.85    465
Hornet 4 Drive      21.4   6 258.0         110 3.08    550
Hornet Sportabout   18.7   8 360.0         175 3.15    875
Valiant             18.1   6 225.0         105 2.76    525
Duster 360          14.3   8 360.0         245 3.21   1225
Merc 240D           24.4   4 146.7          62 3.69    310
Merc 230            22.8   4 140.8          95 3.92    475
Merc 280            19.2   6 167.6         123 3.92    615
Merc 280C           17.8   6 167.6         123 3.92    615
Merc 450SE          16.4   8 275.8         180 3.07    900
Merc 450SL          17.3   8 275.8         180 3.07    900
Merc 450SLC         15.2   8 275.8         180 3.07    900
Cadillac Fleetwood  10.4   8 472.0         205 2.93   1025
Lincoln Continental 10.4   8 460.0         215 3.00   1075
Chrysler Imperial   14.7   8 440.0         230 3.23   1150
Fiat 128            32.4   4  78.7          66 4.08    330
Honda Civic         30.4   4  75.7          52 4.93    260
Toyota Corolla      33.9   4  71.1          65 4.22    325
Toyota Corona       21.5   4 120.1          97 3.70    485
Dodge Challenger    15.5   8 318.0         150 2.76    750
AMC Javelin         15.2   8 304.0         150 3.15    750
Camaro Z28          13.3   8 350.0         245 3.73   1225
Pontiac Firebird    19.2   8 400.0         175 3.08    875
Fiat X1-9           27.3   4  79.0          66 4.08    330
Porsche 914-2       26.0   4 120.3          91 4.43    455
Lotus Europa        30.4   4  95.1         113 3.77    565
Ford Pantera L      15.8   8 351.0         264 4.22   1320
Ferrari Dino        19.7   6 145.0         175 3.62    875
```

```
Maserati Bora       15.0   8 301.0         335 3.54   1675
Volvo 142E          21.4   4 121.0         109 4.11    545
> V <- c(1,2,NA,3)
> V[complete.cases(V)]
[1] 1 2 3
> naVals <- is.na(V)
> install.packages("Hmisc")

WARNING: Rtools is required to build R packages but is not currently
installed. Please download and install the appropriate version of Rtools
before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
trying URL
'https://cran.rstudio.com/bin/windows/contrib/4.5/Hmisc_5.2-4.zip'
Content type 'application/zip' length 3847252 bytes (3.7 MB)
downloaded 3.7 MB

package 'Hmisc' successfully unpacked and MD5 sums checked

Warning: cannot remove prior installation of package 'Hmisc'
Warning: restored 'Hmisc'

The downloaded binary packages are in
      C:\Users\MCA2511\AppData\Local\Temp\Rtmp84KAwx\downloaded_packages
Warning message:
In file.copy(savedcopy, lib, recursive = TRUE) :
  problem copying
C:\Users\MCA2511\AppData\Local\Programs\R\R-4.5.2\library\00LOCK\Hmisc\libs
\x64\Hmisc.dll to
C:\Users\MCA2511\AppData\Local\Programs\R\R-4.5.2\library\Hmisc\libs\x64\Hm
isc.dll: Permission denied


> library(Hmisc)


Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':


    format.pval, units
> x = c(1,2,3,NA,4,4,NA)
> v <-impute(x, fun=mean)
> v
```

```
   1    2    3    4    5    6    7
 1.0  2.0  3.0 2.8*  4.0  4.0 2.8*
> v<-impute(x,fun=median)
> v
 1  2  3  4  5  6  7
 1  2  3 3*  4  4 3*
> data1<-data.frame(Srno = c(1,2,3,NA,4,4,NA),
+                  Name = c("a","b","c","d","e","f","g"),
+                  Salary = c(400,200,NA,500,NA,800,900)
+                  )
> v <-impute(data1$rno, fun=mean)
> v
NULL
> v <- impute(data1$Salary, fun=median)
> v
   1    2    3    4    5    6    7
 400  200 500*  500 500*  800  900
> c1 <-c("low","medium","high","low")
> c1 <-factor(c1,levels=c("low","medium","high"))
> c1
[1] low    medium high   low
Levels: low medium high
> data1<-read.csv("missing_col.csv",sep=",",
+                col.names=c("Srno","Name","Salary","DOJ","Department"))

Warning message:
In read.table(file = file, header = header, sep = sep, quote = quote,  :

  header and 'col.names' are of different lengths


> View(data1)
```



	Srno	Name	Salary	DOJ	Department
2	Dan	512.20		23-09-2013	Operation
3	Michelle	611.00		15-11-2014	IT
4	Ryan	729.00		11-05-2014	HR
	Gary	843.25		37-03-2015	Finance
6	Meena	NA	21-03-20153		IT
7	Simon	632.80		30-07-2013	Operation
8	Guru	722.00		17-06-2014	Finance
9	John	NA		21-05-2012	
10	Rock	600.80		30-07-2013	HR
11	Brad	1032.80		20-07-2013	Operation
12	Ryan	729.00		11-05-2014	HR

```
> x <- c(1,2,3,NA,4,NA,5)
> x
[1]  1  2  3 NA  4 NA  5
> #Indicates which elements are missing
> xn <-is.na(x)
> x[!xn]
[1] 1 2 3 4 5
> NA+4
[1] NA
> #This will keep NA rows in data while removes them during calculate
> median(x,na.rm=T)
[1] 3
> #Return a logical vector indicating
> complete.cases(x)
[1]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
> is.na(data1)
    Srno  Name Salary   DOJ Department
2  FALSE FALSE  FALSE FALSE      FALSE
3  FALSE FALSE  FALSE FALSE      FALSE
4  FALSE FALSE  FALSE FALSE      FALSE
   FALSE FALSE  FALSE FALSE      FALSE
6  FALSE  TRUE  FALSE FALSE      FALSE
7  FALSE FALSE  FALSE FALSE      FALSE
8  FALSE FALSE  FALSE FALSE      FALSE
9  FALSE  TRUE  FALSE FALSE      FALSE
10 FALSE FALSE  FALSE FALSE      FALSE
11 FALSE FALSE  FALSE FALSE      FALSE
12 FALSE FALSE  FALSE FALSE      FALSE
> datacompletecases <- data1[complete.cases(data1),]
> datacompletecases
       Srno     Name Salary        DOJ Department
2        Dan  512.20       23-09-2013  Operation
3   Michelle  611.00       15-11-2014         IT
4       Ryan  729.00       11-05-2014         HR
        Gary  843.25       37-03-2015    Finance
7      Simon  632.80       30-07-2013  Operation
8       Guru  722.00       17-06-2014    Finance
10      Rock  600.80       30-07-2013         HR
11      Brad 1032.80       20-07-2013  Operation
12      Ryan  729.00       11-05-2014         HR
> #Detect if there are any NAs: any(is.na(datan))
> #Identify positions of NAs; which(is.na(data$v1))
> any(is.na(x))
[1] TRUE
> which(is.na(data1$Srno))
integer(0)
> na.omit(x)
[1] 1 2 3 4 5
attr(,"na.action")
[1] 4 6
attr(,"class")
```

```
[1] "omit"
```

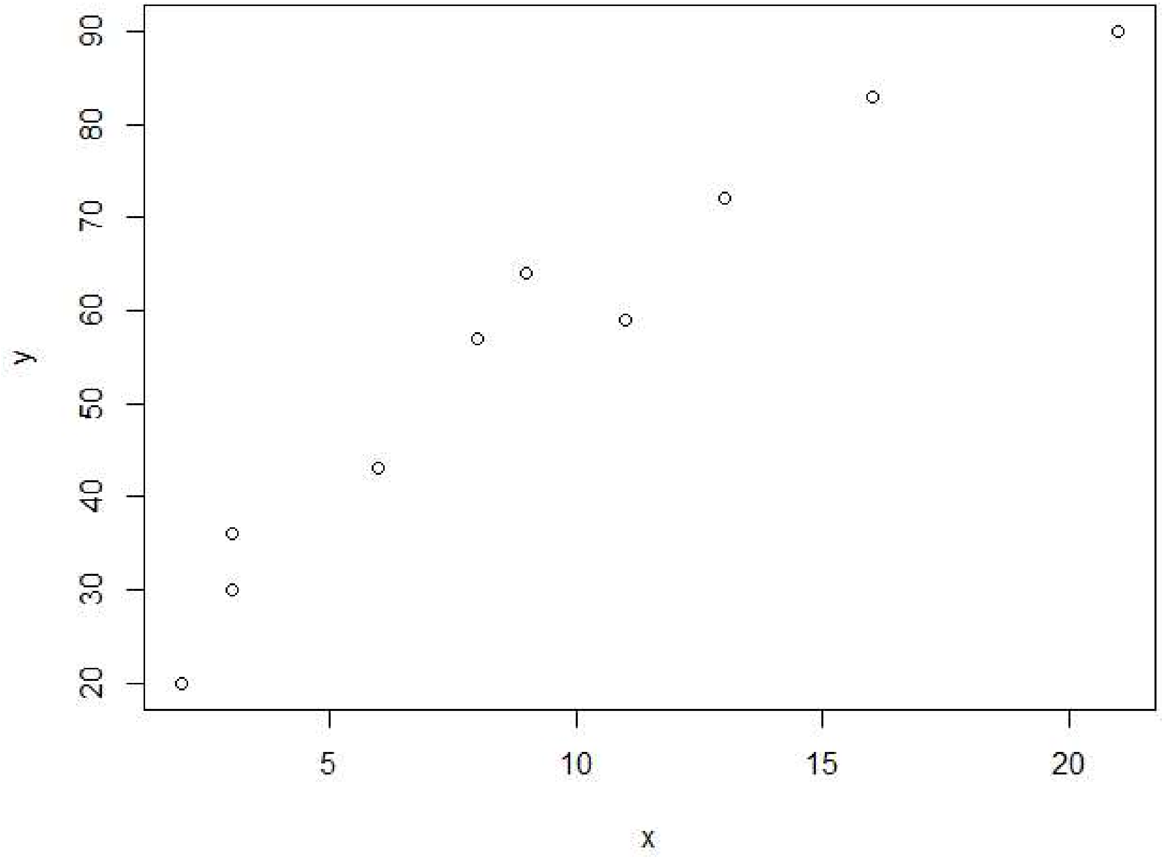
missing_col.csv

1	Rick	623.3		01-01-2012	IT
2	Dan	512.2		23-09-2013	Operation
3	Michelle	611		15-11-2014	IT
4	Ryan	729		11-05-2014	HR
	Gary	843.25		37-03-2015	Finance
6	Meena	NA	21-03-20153		IT
7	Simon	632.8		30-07-2013	Operation
8	Guru	722		17-06-2014	Finance
9	John	NA		21-05-2012	
10	Rock	600.8		30-07-2013	HR
11	Brad	1032.8		20-07-2013	Operation
12	Ryan	729		11-05-2014	HR

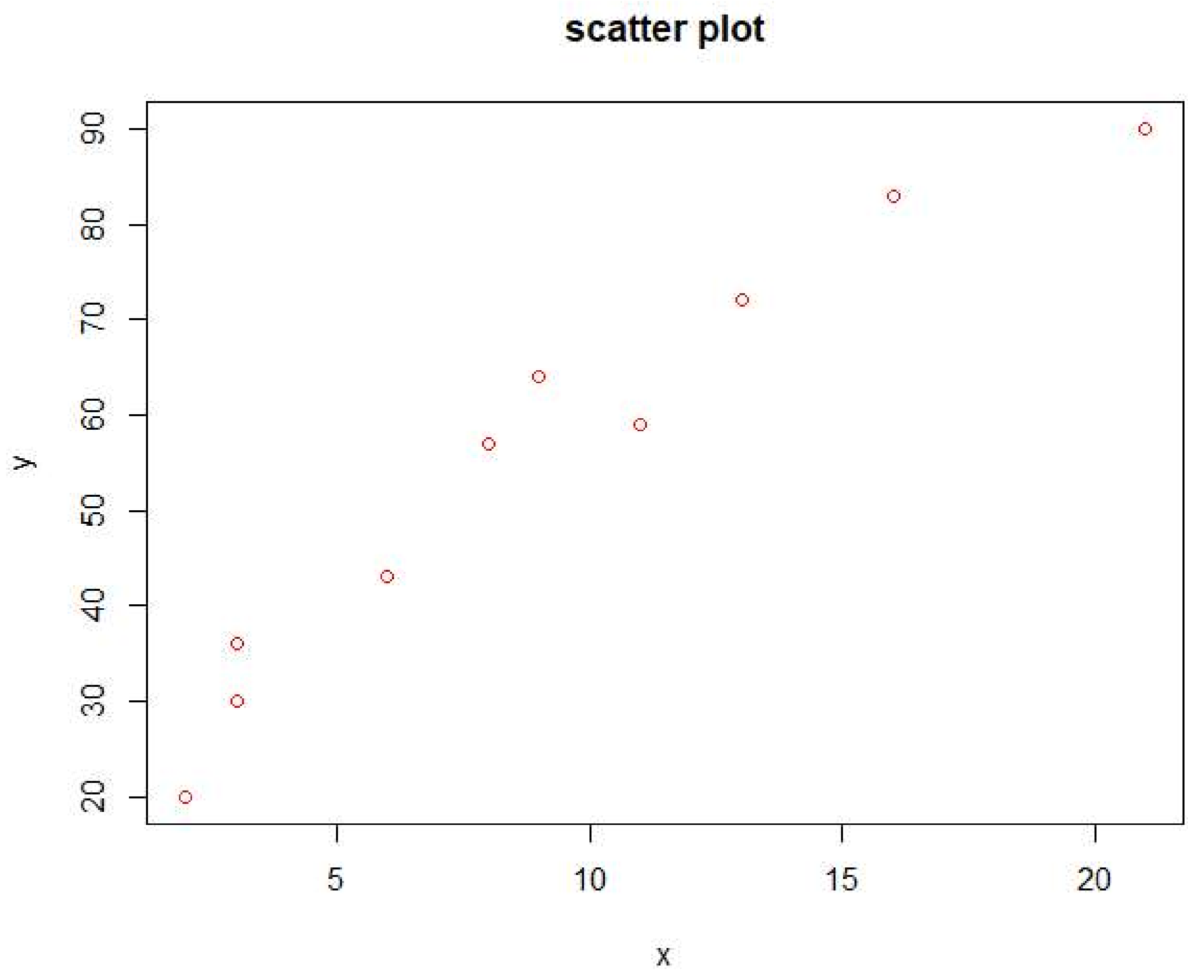
Practical 10 (half) :

Implement Linear Regression and create plot.

```
> x <- c(3,8,9,13,3,6,11,21,2,16)
> #response variable
> y <-c(30,57,64,72,36,43,59,90,20,83)
> plot(x,y)
```

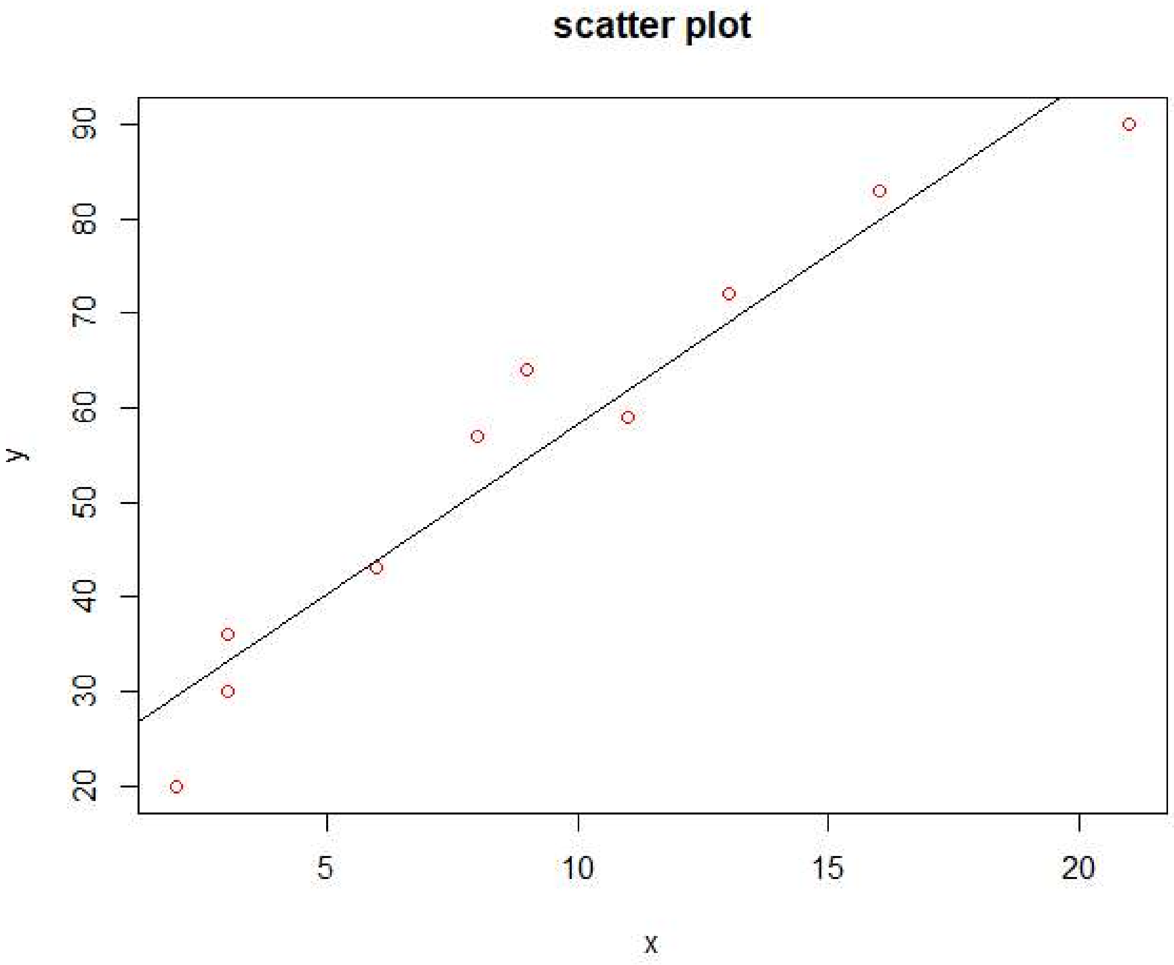


```
> plot(x,y,col="red",main="scatter plot")
```

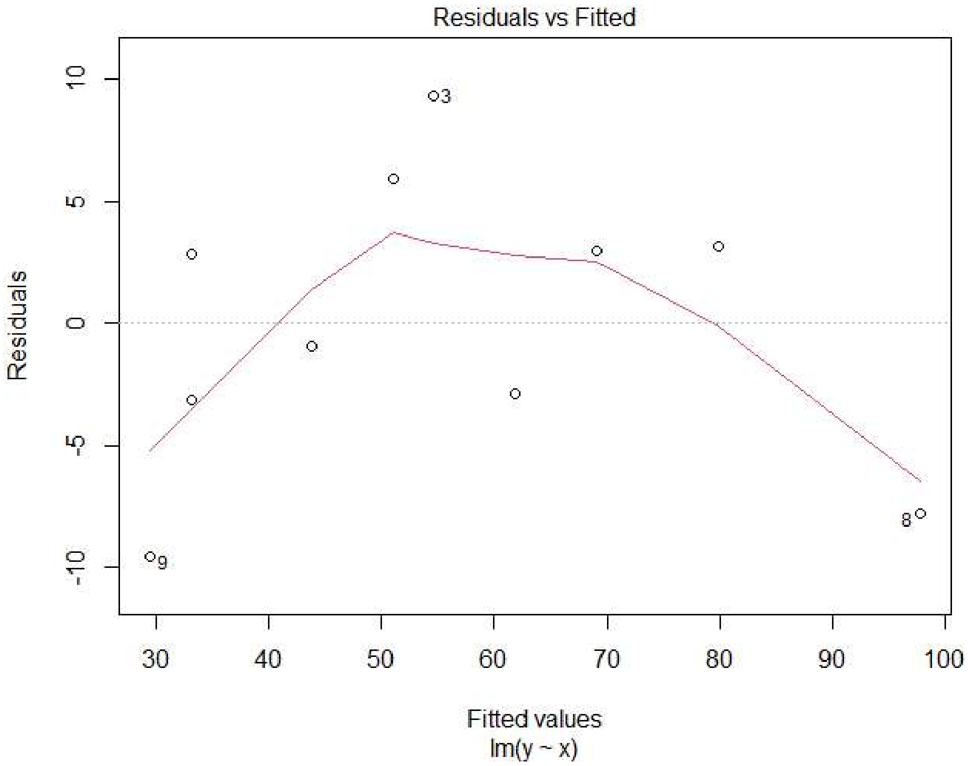


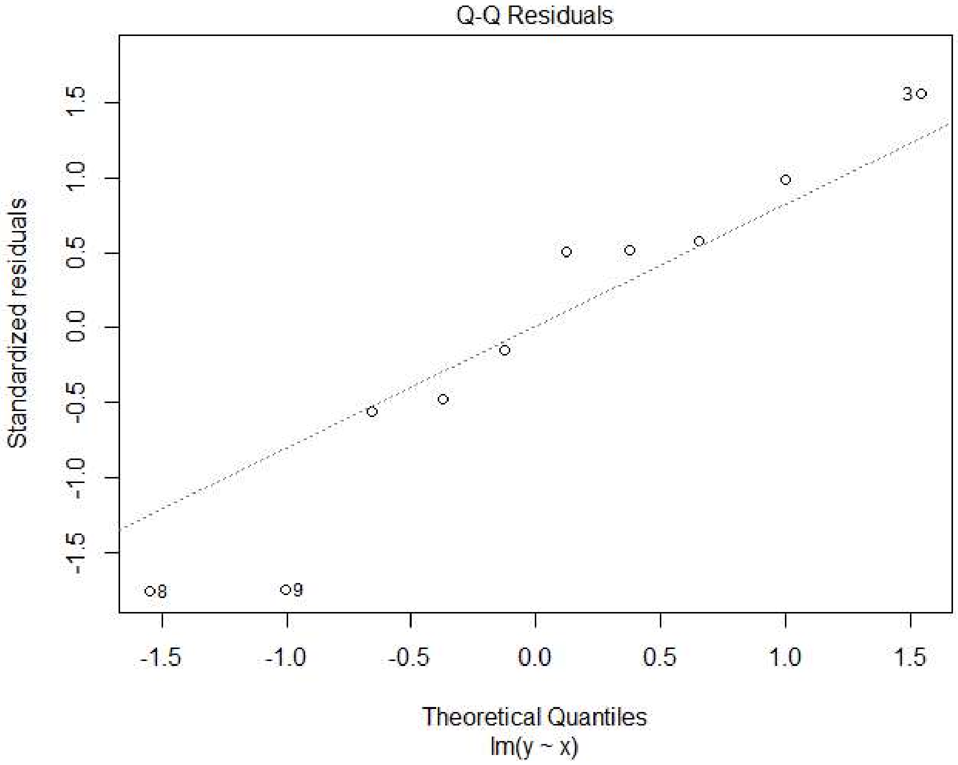
```
> model = lm(y~x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
     22.354        3.592  

> attributes(model)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"         
 [5] "fitted.values" "assign"        "qr"            "df.residual"  
 [9] "xlevels"       "call"          "terms"         "model"        

$class
[1] "lm"

> coef(model)
(Intercept)           x 
```

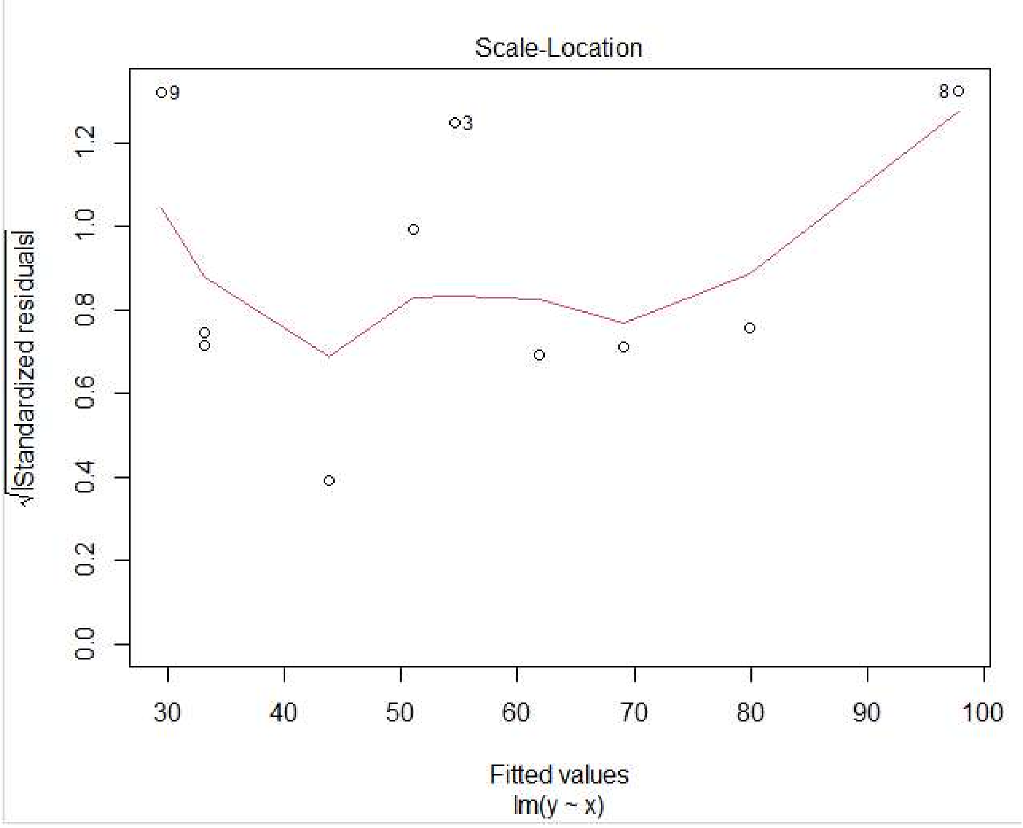
```
  22.353900    3.591967 
> residuals(model)
         1          2          3          4          5          6 
7 
-3.1298021  5.9103609  9.3183935  2.9505239  2.8701979 -0.9057043 
-2.8655413 
         8          9         10 
-7.7852154 -9.5378347  3.1746217 
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5378 -3.0637  0.9822  3.1186  9.3184 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.3539     3.7173   6.013 0.000319 ***
x             3.5920     0.3408  10.541 5.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.317 on 8 degrees of freedom
Multiple R-squared:  0.9328,	Adjusted R-squared:  0.9244 
F-statistic: 111.1 on 1 and 8 DF,  p-value: 5.721e-06

> abline(model)
```

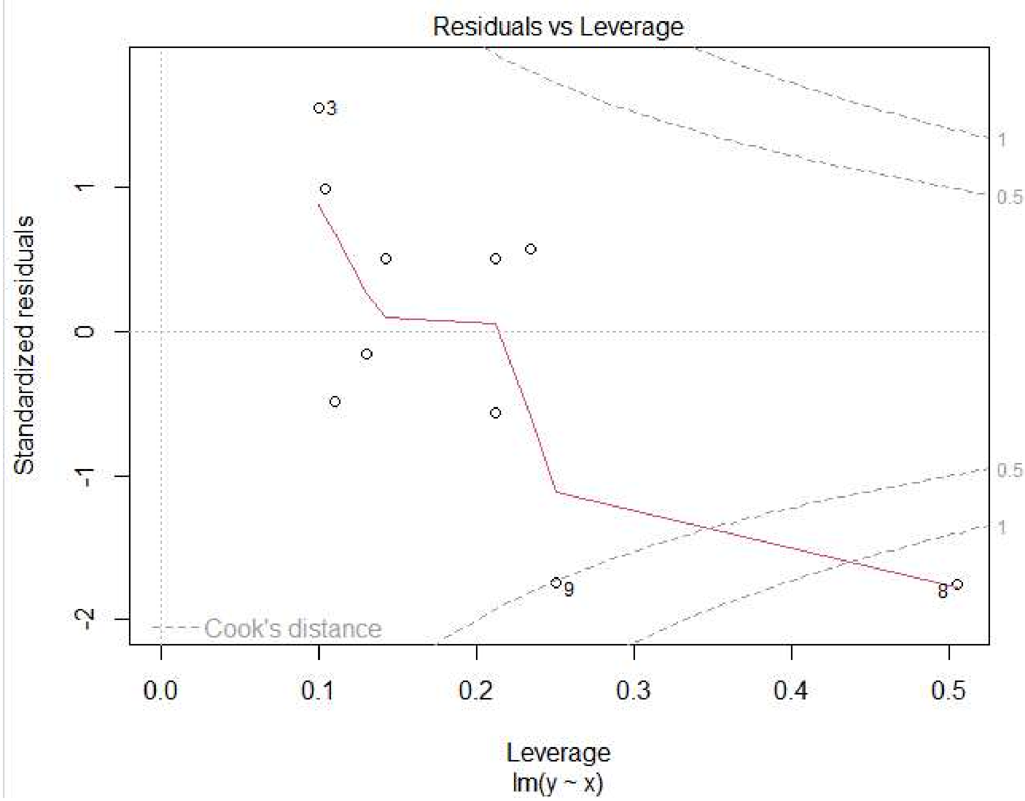


```
> #predicting values manually y = a + bx
> x10 <- model$coefficients[[1]]+model$coefficients[[2]]*10
> x10
[1] 58.27357
> #using predict()
> a <- data.frame(x=10)
> a
   x
1 10
> pred <- predict(model,a)
> pred
       1
58.27357
> plot(model)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```


Residuals vs Fitted
Residuals
10
5
0
-5
-10
3
9
8
30
40
50
60
70
80
90
100
Fitted values
lm(y ~ x)


Q-Q Residuals
Standardized residuals
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
3
8
9
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(y ~ x)


Scale-Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
9
3
8
30
40
50
60
70
80
90
100
Fitted values
lm(y ~ x)


Residuals vs Leverage
Standardized residuals
-2
-1
0
1
3
9
8
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(y ~ x)