Hw2_stat615

Venkata Dhanush Kikkisetti

2023-02-21

1 A chemist studied the concentraTion of a solution (Y) over time (X). Fifteen identical solutions were prepared. The IS. solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after I, 3, 5, 7, and 9 hours. The results follow.

a Fit a linear regression function.
b. Perform the F test to determine whether or not there is lack of fit of a linear regression function; use ex = .025. State the alternatives, decision rule, and conclusion.

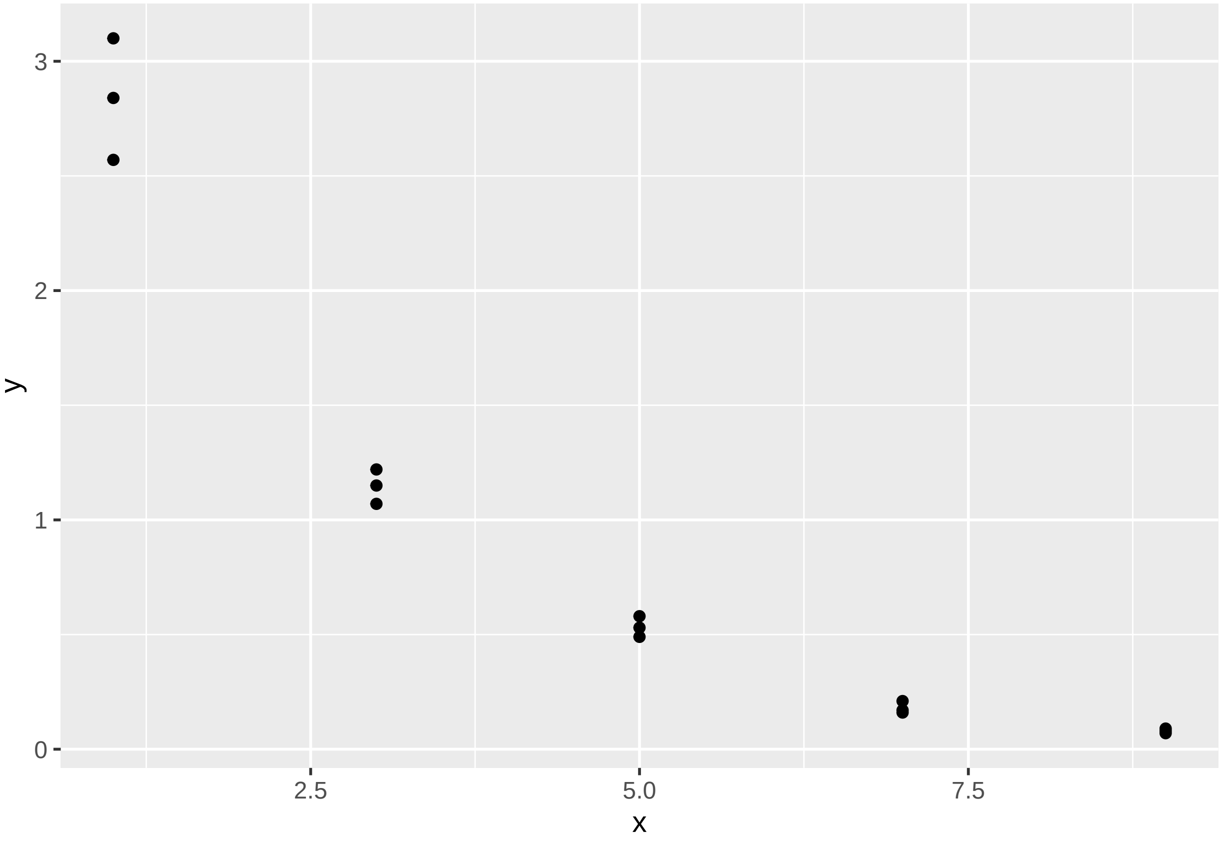
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
solutions=tribble(~y   , ~x,

                0.07 , 9.0,
                0.09  ,9.0,
                0.08  ,9.0,
                0.16   ,7.0,
                0.17   ,7.0,
                0.21    ,7.0,
                0.49    ,5.0,
                0.58    ,5.0,
                0.53    ,5.0,
                1.22    ,3.0,
                1.15    ,3.0,
                1.07    ,3.0,
                2.84    ,1.0,
                2.57    ,1.0,
                3.10    ,1.0)
```

```r
ggplot(data = solutions,mapping = aes(x=x,y=y))+
  geom_point()
```



```r
lm<-lm(y~x,data=solutions)

lm
```

```
##
## Call:
## lm(formula = y ~ x, data = solutions)
##
## Coefficients:
## (Intercept)            x
##       2.575       -0.324
```

```r
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 12.5971  12.597  55.994 4.611e-06 ***
## Residuals 13  2.9247   0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The slope is negative, we can see a decrease trend between concentrated solution and the time. As time passes by the concentration in the liquid has been reducing. From the analysis of variance we have p-value far less than 0.05. We can conclude that variability of concentration solution is explained by variability of the time.

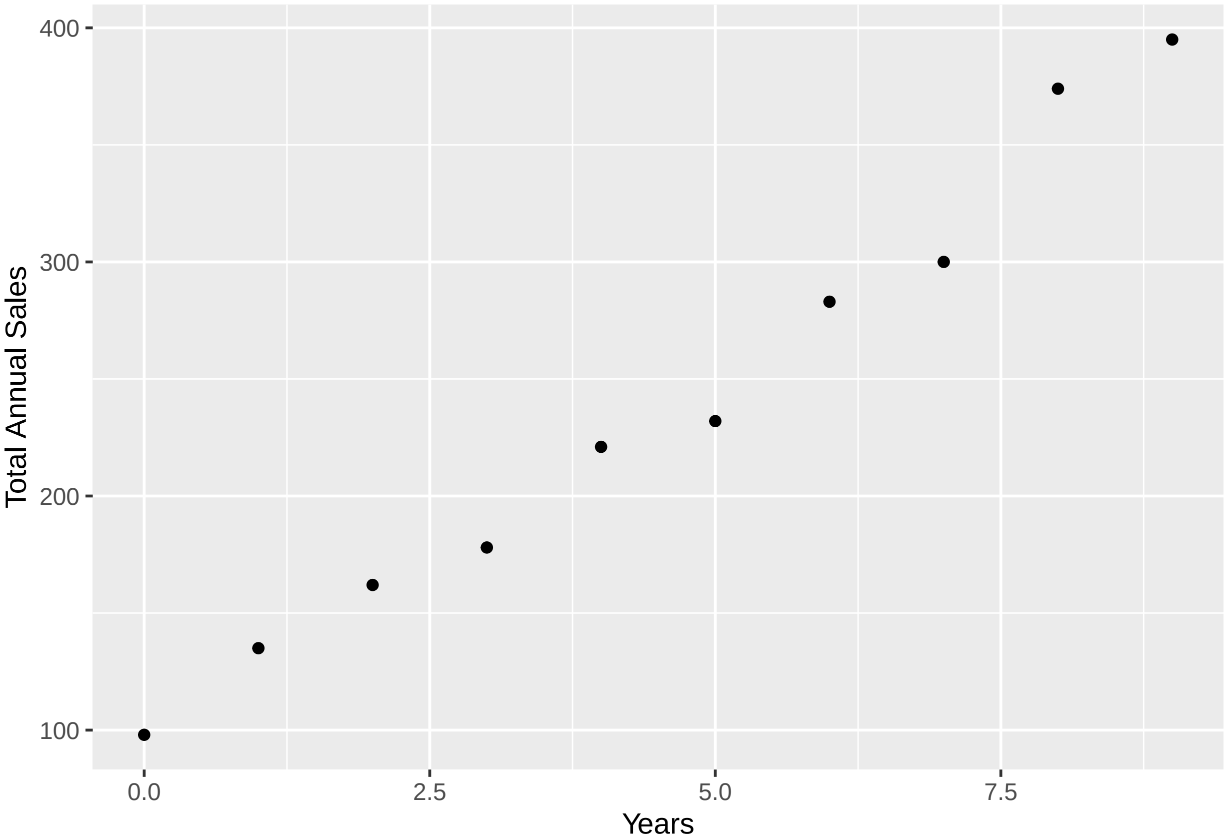
2 Sales growth. A marketing researcher studied annual sales ofa product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands

```
sales=data.frame(x=c(0:9),y=c(98,135,162,178,221,232,283,300,374,395))
sales
```

```
##    x   y
## 1  0  98
## 2  1 135
## 3  2 162
## 4  3 178
## 5  4 221
## 6  5 232
## 7  6 283
## 8  7 300
## 9  8 374
## 10 9 395
```

a Prepare a scatter plot of the data. Does a linear relation appear adequate here?

```
ggplot(data = sales,mapping = aes(x=x,y=y))+
  geom_point()+
  xlab("Years")+
  ylab("Total Annual Sales")
```



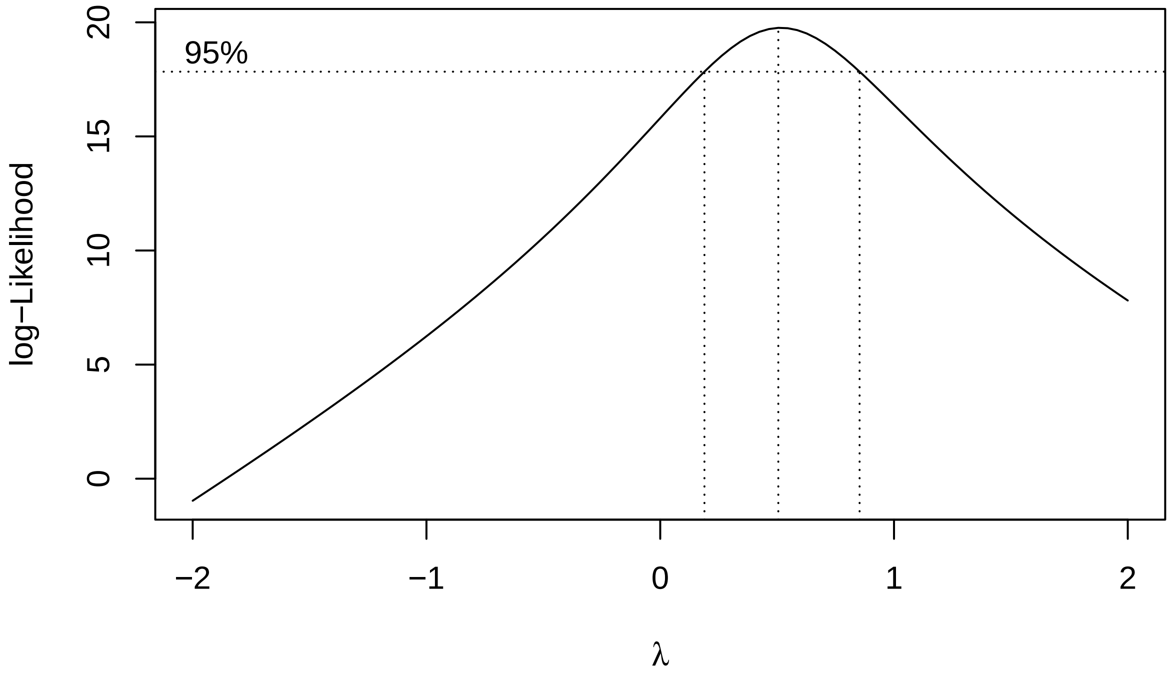
## b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transfor- mation of Y. Evaluate SSE for A = .3, .4, .5, .6, .7. What transformation of Y is suggested?

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
box_cox=boxcox(y ~ x,data=sales)
```



```
lambda=box_cox$x[which.max(box_cox$y)]
lambda
```

```
## [1] 0.5050505
```

```
anova(lm((y^0.3)~x,data=sales))
```

```
## Analysis of Variance Table
##
## Response: (y^0.3)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 3.9836  3.9836  605.92 7.926e-09 ***
## Residuals  8 0.0526  0.0066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm((y^0.4)~x,data=sales))
```

```
## Analysis of Variance Table
##
## Response: (y^0.4)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 20.7888 20.7888  688.23 4.789e-09 ***
## Residuals  8  0.2416  0.0302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm((y^0.5)~x,data=sales))
```

```
## Analysis of Variance Table
##
## Response: (y^0.5)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 95.568  95.568  728.37 3.826e-09 ***
## Residuals  8  1.050   0.131
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(lm((y^0.6)~x,data=sales))
```

```
## Analysis of Variance Table
##
## Response: (y^0.6)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 405.81  405.81  711.26 4.203e-09 ***
## Residuals  8   4.56    0.57
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(lm((y^0.7)~x,data=sales))
```

```
## Analysis of Variance Table
##
## Response: (y^0.7)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 1632.40 1632.40  646.11 6.148e-09 ***
## Residuals  8   20.21    2.53
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Sum of squared error after transforming y with lambda 0.3 is 0.0526
Sum of squared error after transforming y with lambda 0.4 is 0.2416
Sum of squared error after transforming y with lambda 0.5 is 1.050
Sum of squared error after transforming y with lambda 0.6 is 4.56
Sum of squared error after transforming y with lambda 0.7 is 20.21
I would suggest for lambda value 0.3 because the residual error is less for that compared to other values

c Use the transformation Y’ = ..JYand obtain the estimated linear regression function for the

transformed data.

```
sales%>%
  mutate(transformed_y=y^0.5)->sales

lm=lm(transformed_y~x,data=sales)
lm
```

```
##
## Call:
```

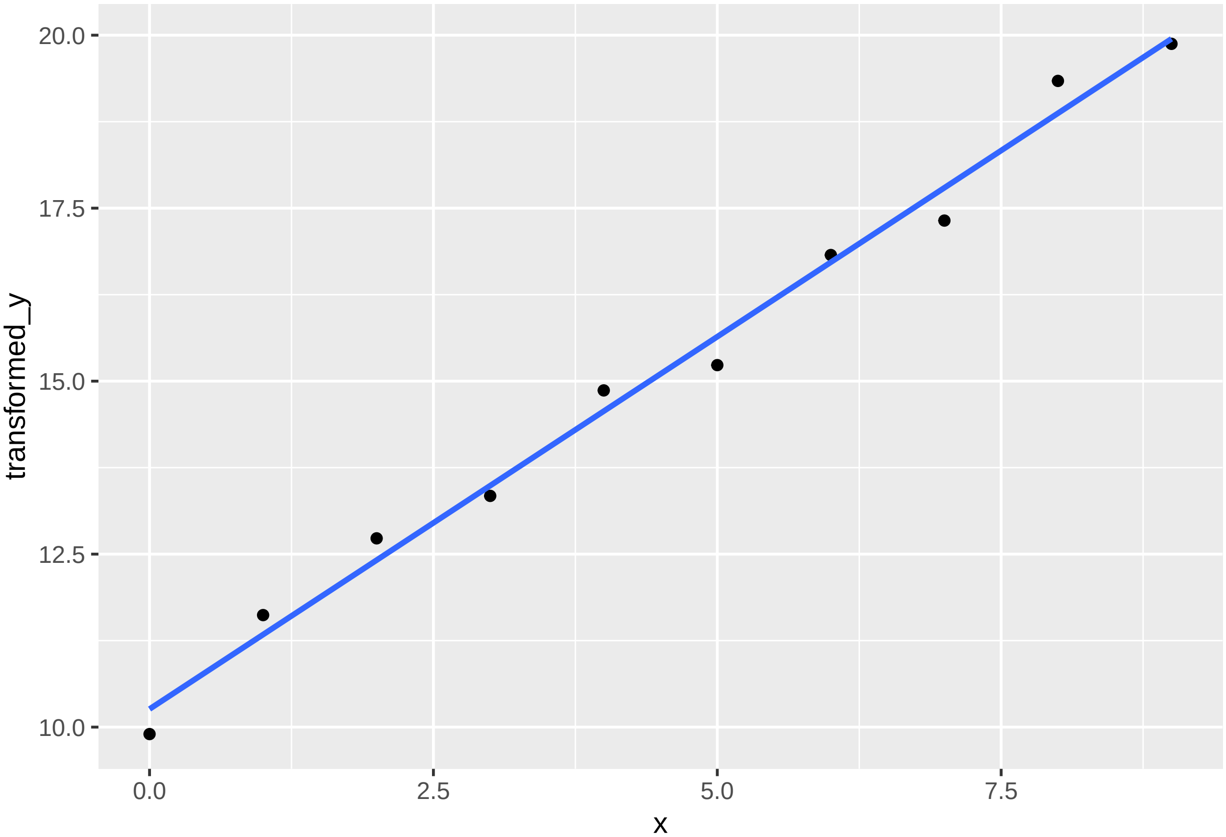
```
## lm(formula = transformed_y ~ x, data = sales)
##
## Coefficients:
## (Intercept)            x
##      10.261        1.076
```

d) Plot the estimated regression line and the transformed data. Does the regression line appear

to be a good fit to the transformed data?

```
ggplot(data = sales,mapping = aes(x=x,y=transformed_y))+
  geom_point()+
  geom_smooth(method = 'lm',se=FALSE)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

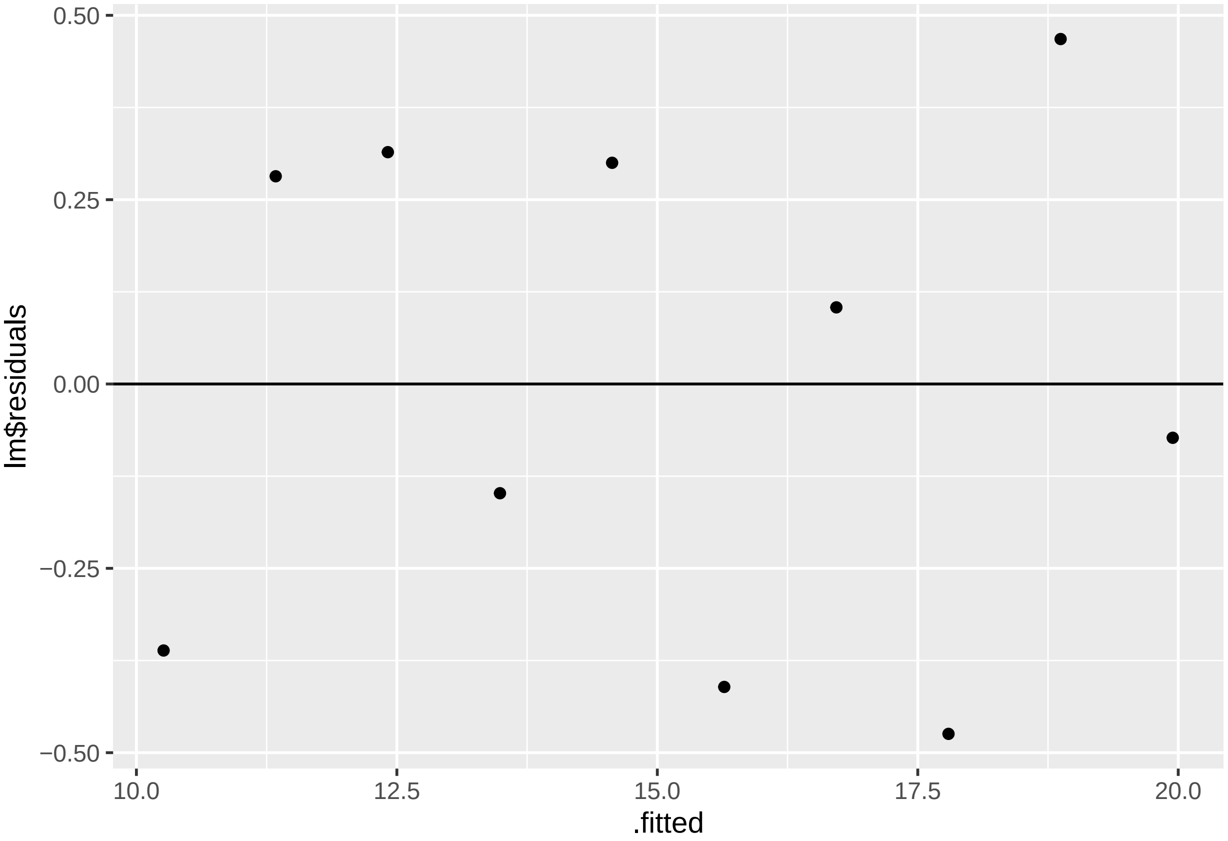


Yes, the regression line almost fits the points and has a very good relationship between sales and year.

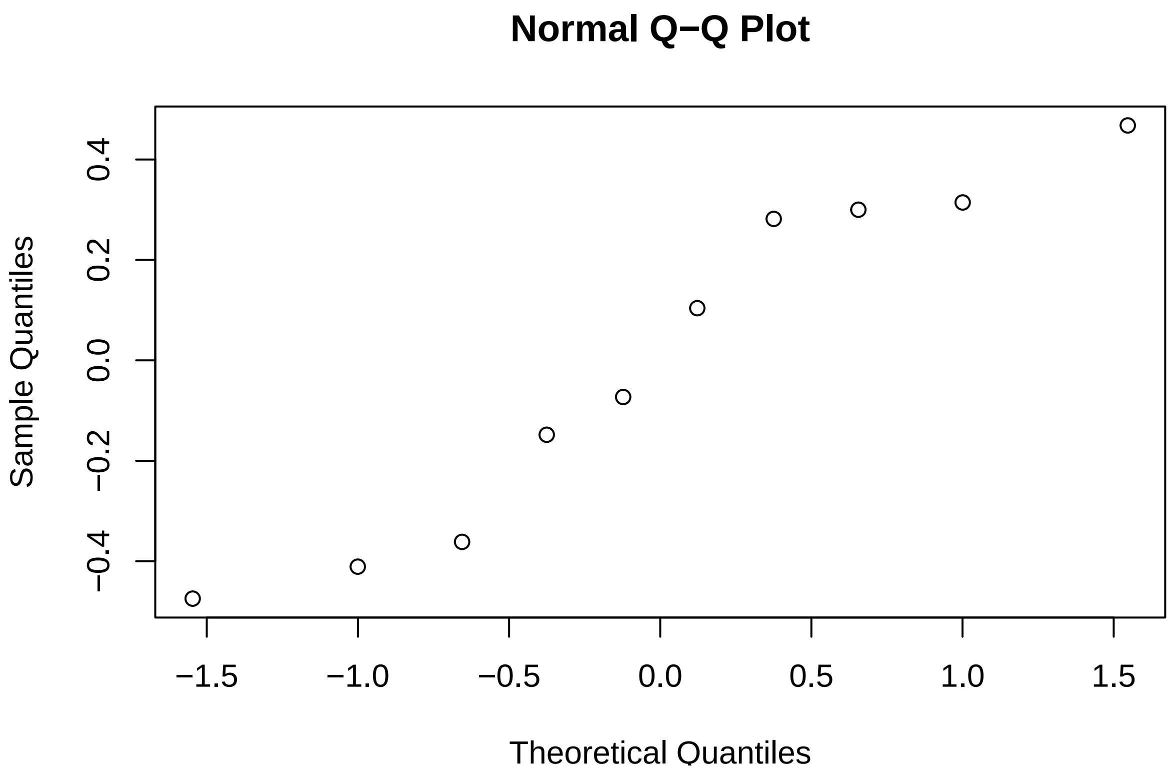
e)Obtain the residuals and plot them against the fitted values. Also prepare a normal probability

plot. What do your plots show?

```
ggplot(data =lm,mapping = aes(x=.fitted,y=lm$residuals))+
  geom_point()+
  geom_hline(yintercept = 0)
```



```
qqnorm(lm$residuals)
```



We can observe that the residuals are normally distributed and residuals are spread above and below the intercept line. This model best describes the data.

3 For the matrices below, obtain (1) A +B, (2) A - B, (3) AC, (4) AB', (5) B'A.

```
A=matrix(c(1,4,2,6,3,4),ncol = 2,byrow = TRUE)
B=matrix(c(1,3,1,4,2,5),ncol = 2,byrow = TRUE)
C=matrix(c(3,8,1,5,4,0),ncol = 3,byrow = TRUE)

#1
A+B
```

```
##      [,1] [,2]
## [1,]    2    7
## [2,]    3   10
## [3,]    5    9
```

```
#2
A-B
```

```
##      [,1] [,2]
## [1,]    0    1
## [2,]    1    2
## [3,]    1   -1
```

```
#3
A%*%C
```

```
##      [,1] [,2] [,3]
## [1,]   23   24    1
## [2,]   36   40    2
## [3,]   29   40    3
```

```
#4
A%*%t(B)
```

```
##      [,1] [,2] [,3]
## [1,]   13   17   22
## [2,]   20   26   34
## [3,]   15   19   26
```

```
#5
t(B)%*%A
```

```
##      [,1] [,2]
## [1,]    9   18
## [2,]   26   56
```

##4 Call the mtcars data set and produce a linear model that uses hp (horse power) to predict mpg (miles per gallon). Using techniques and R code demonstrated in class, produce a prediction interval, and a confidence interval for the response variable mpg at a fixed mpg value of .21. (follow the coding sequence closely of the examples shown in class.)

```
data(mtcars)
lm=lm(mpg~hp,data=mtcars)
lm
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Coefficients:
## (Intercept)           hp
##    30.09886     -0.06823
```

```
new<-data.frame(hp=c(62))

predict(lm,newdata =new,interval = 'prediction' )
```

```
##        fit      lwr      upr
## 1 25.86871 17.66822 34.06919
```

```
predict(lm,newdata =new,interval = 'confidence' )
```

```
##        fit      lwr      upr
## 1 25.86871 23.63082 28.10659
```