Regression Analaysis

Assignment 1

Anirudhda Pardhi,s3807109

Q1

```
# This is an R chunk for QUESTION 1:
library(readr)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'

## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(car)
```

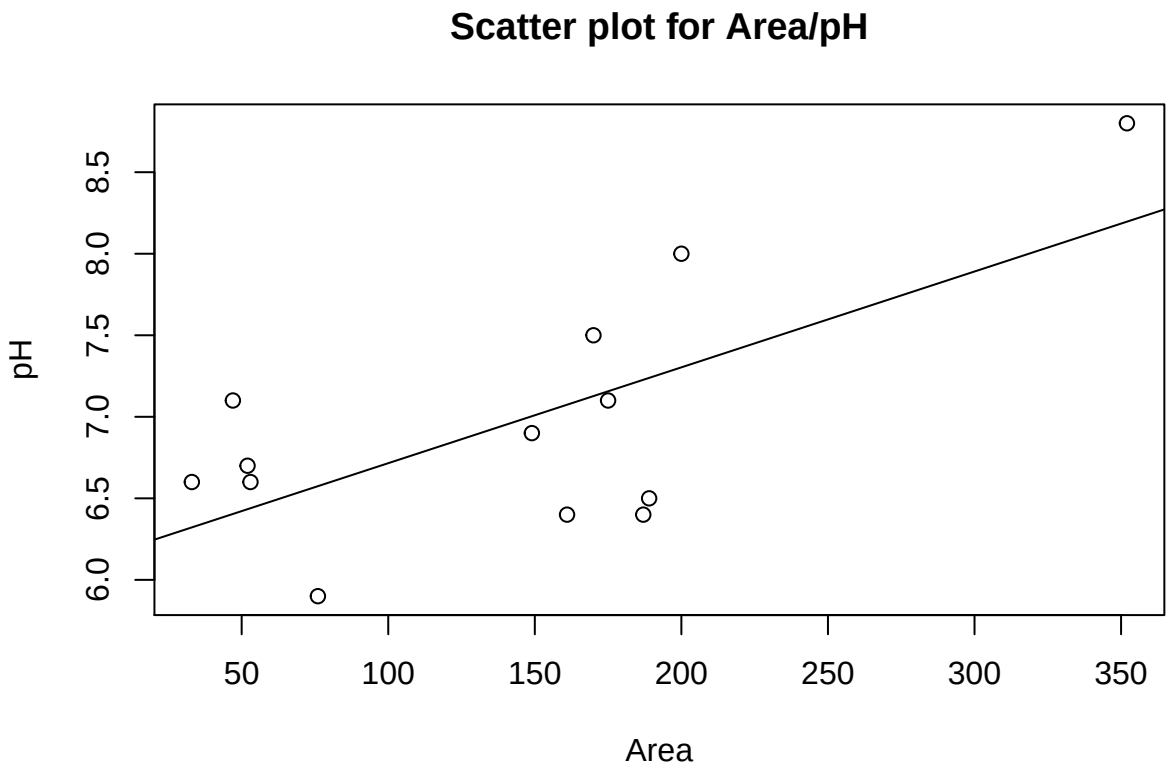
```
## Loading required package: carData

##
## Attaching package: 'car'

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

```r
library(ISwR)
```

```
##
## Attaching package: 'ISwR'

## The following object is masked from 'package:survival':
##
##     lung
```

```r
Lakes = data.frame(Area = c( 33, 161, 189, 149, 47, 170, 352, 187, 76, 52, 175, 53, 200),
                           pH = c(6.6, 6.4, 6.5, 6.9, 7.1, 7.5, 8.8, 6.4, 5.9, 6.7, 7.1, 6.6, 8.0))

x = Lakes[,1]
y = Lakes[,2]
n = length(x)
```

```r
plot(x,y,xlab="Area", ylab="pH", main="Scatter plot for Area/pH")

# a) As the value of Area is increasing we can see that
#value of pH is also increasing.Thus we can
#   interpret from this scatterplot that is is showing Linear trend

reg = lm(y~x)
abline(lm(y~x))
```



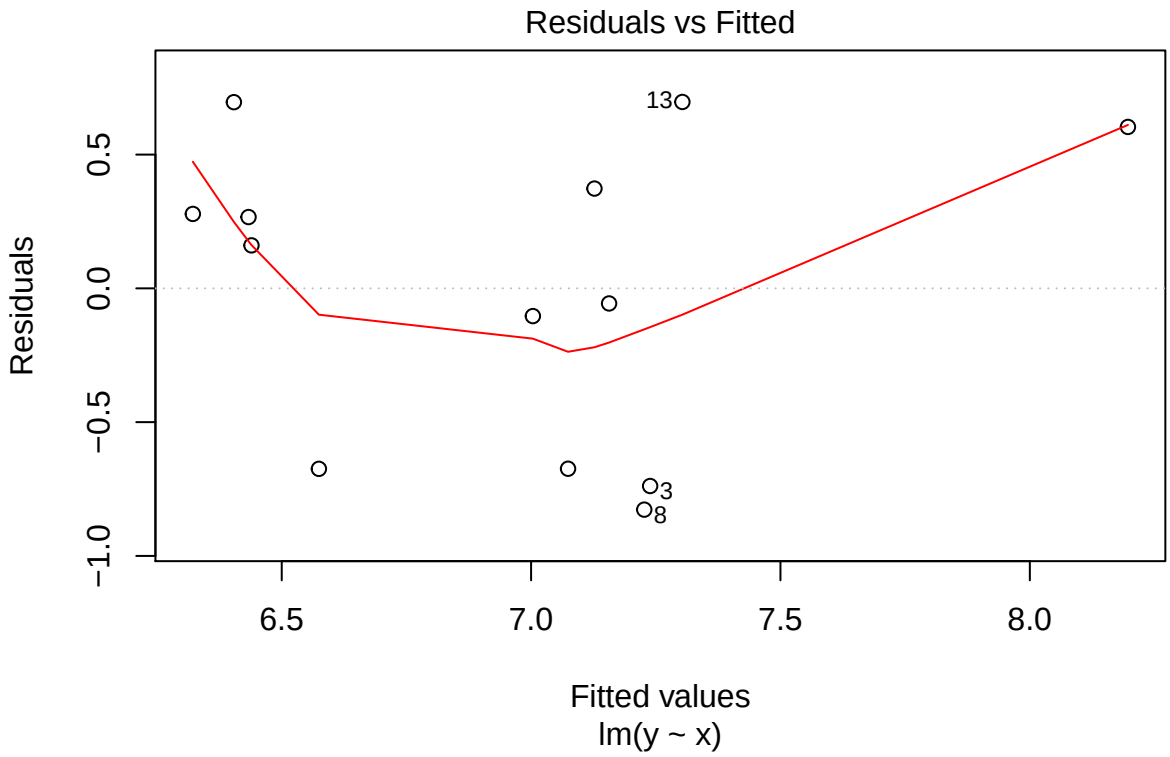

```r
#b) The above code is for line of best fit on the scatterplot.

phareamodel <- lm(y ~ x, data = Lakes)
phareamodel%>% anova()
```

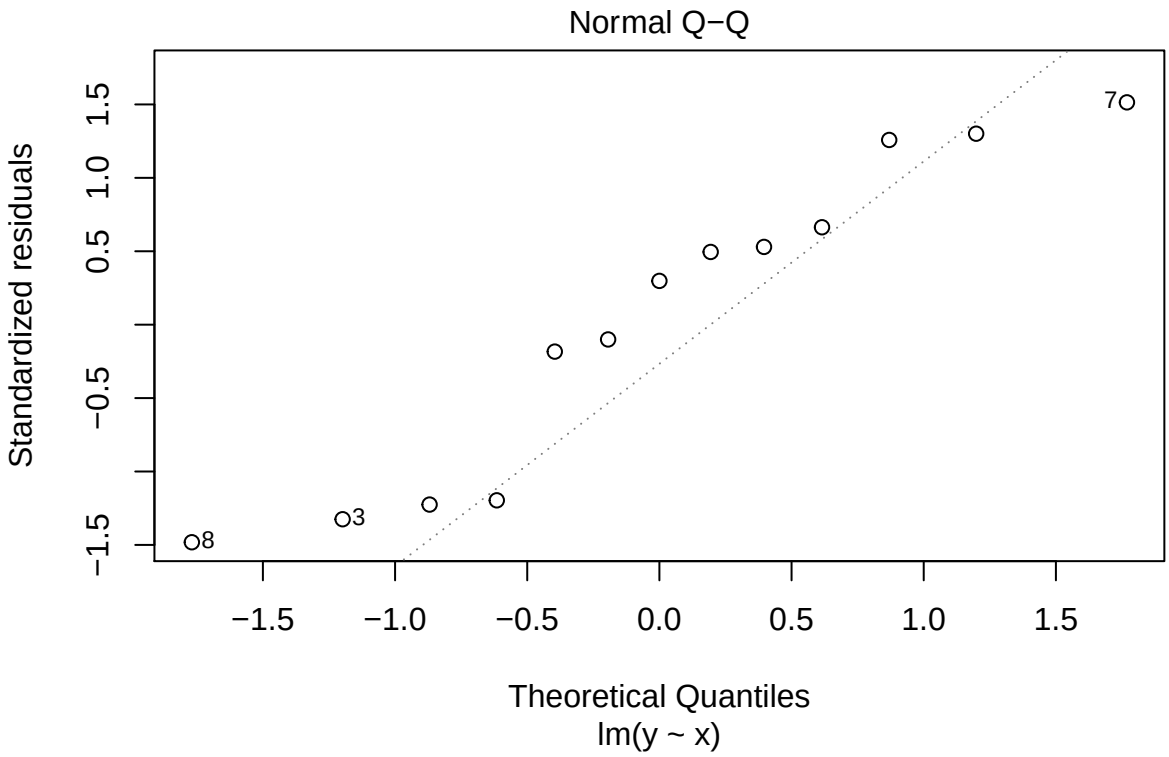
```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x          1 3.2910  3.2910  9.5272 0.01035 *
## Residuals 11 3.7998  0.3454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

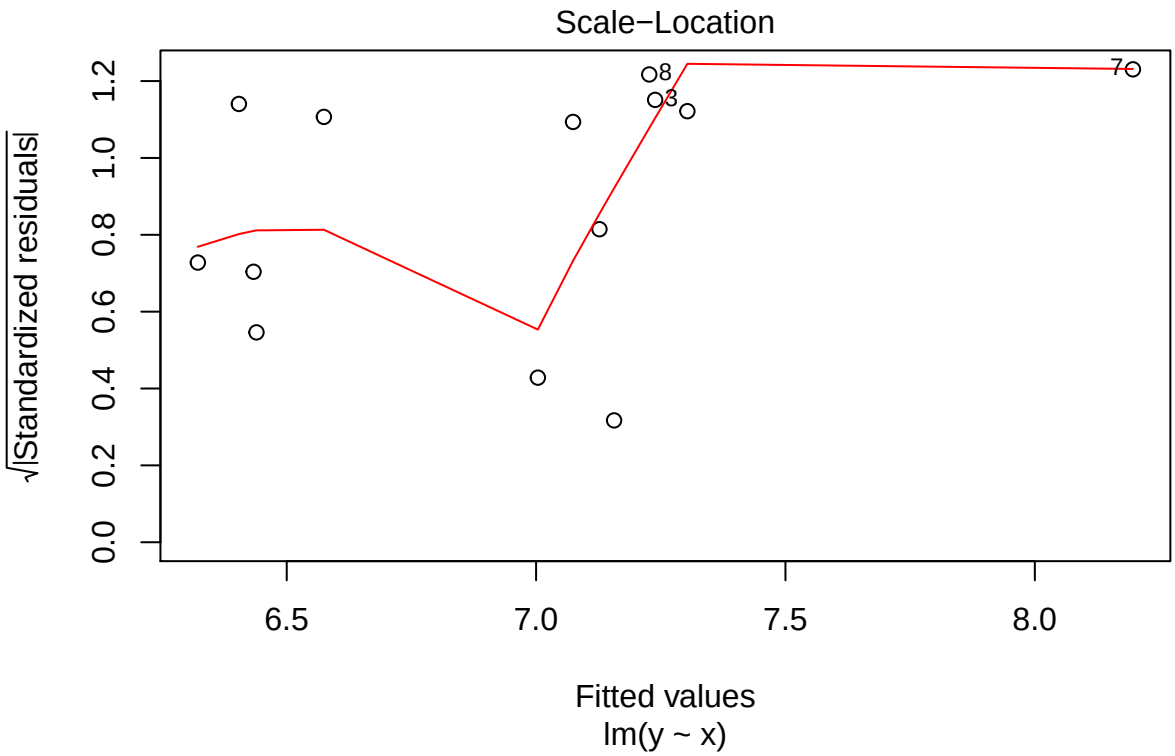
```r
#c) The above code is for Anova test and this is evident from
#high F value that there is Linear relationship between Area and pH.

reg$residuals
```

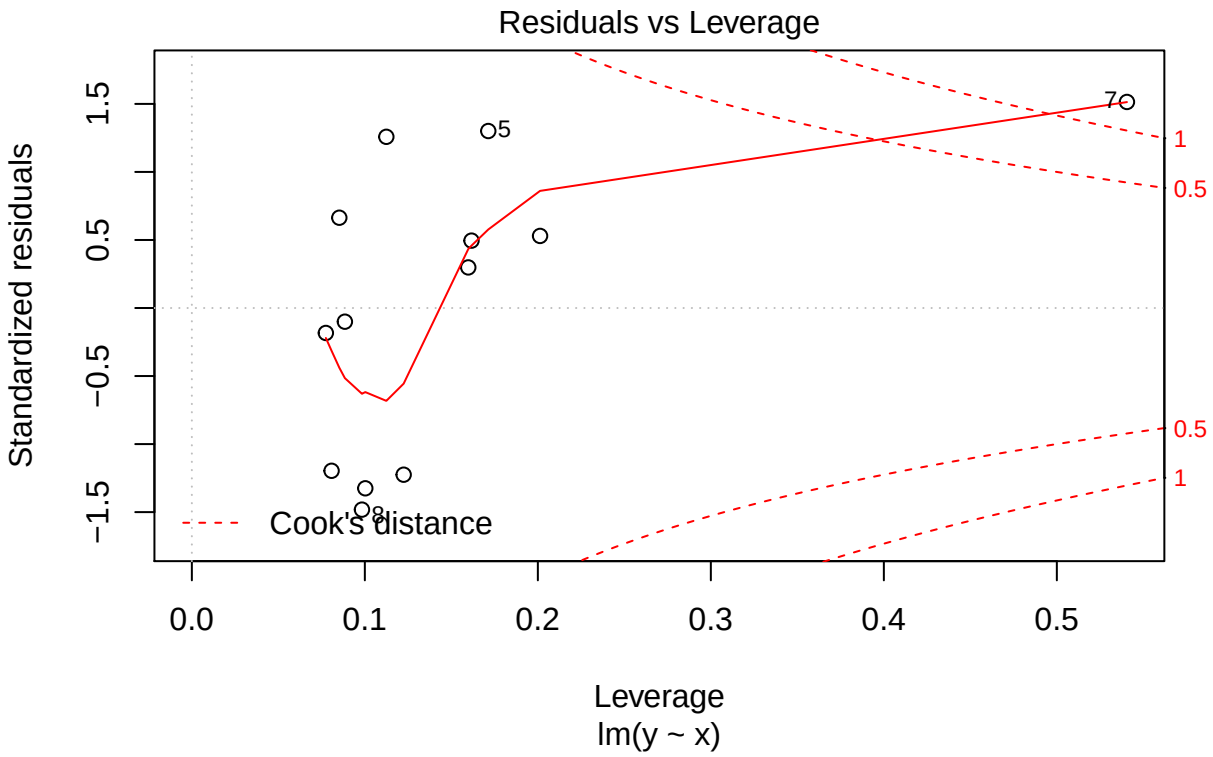
```
##           1           2           3           4           5           6
##  0.27821706 -0.67411735 -0.73869050 -0.10358600  0.69593049  0.37298414
##           7           8           9          10          11          12
##  0.60325865 -0.82693528 -0.67452028  0.26654242 -0.05640392  0.16066481
##          13
##  0.69665576
```

```r
plot(phareamodel)
```




Normal Q−Q
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
8
3
7
−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
8
3
7
6.5
7.0
7.5
8.0
Fitted values
lm(y ~ x)




```
ncvTest(reg)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4042468, Df = 1, p = 0.5249
```

```
shapiro.test(reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.89077, p-value = 0.09995
```

```
durbinWatsonTest(reg)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2016949      1.448513   0.296
##  Alternative hypothesis: rho != 0
```

```
#d) In residual vs fitted graph we can see that the red line is slightly curved
# and the residuals seem to increase as the fitted Y values increase so there
#may be heteroscedasticity exists.So we do "NCV test"
```

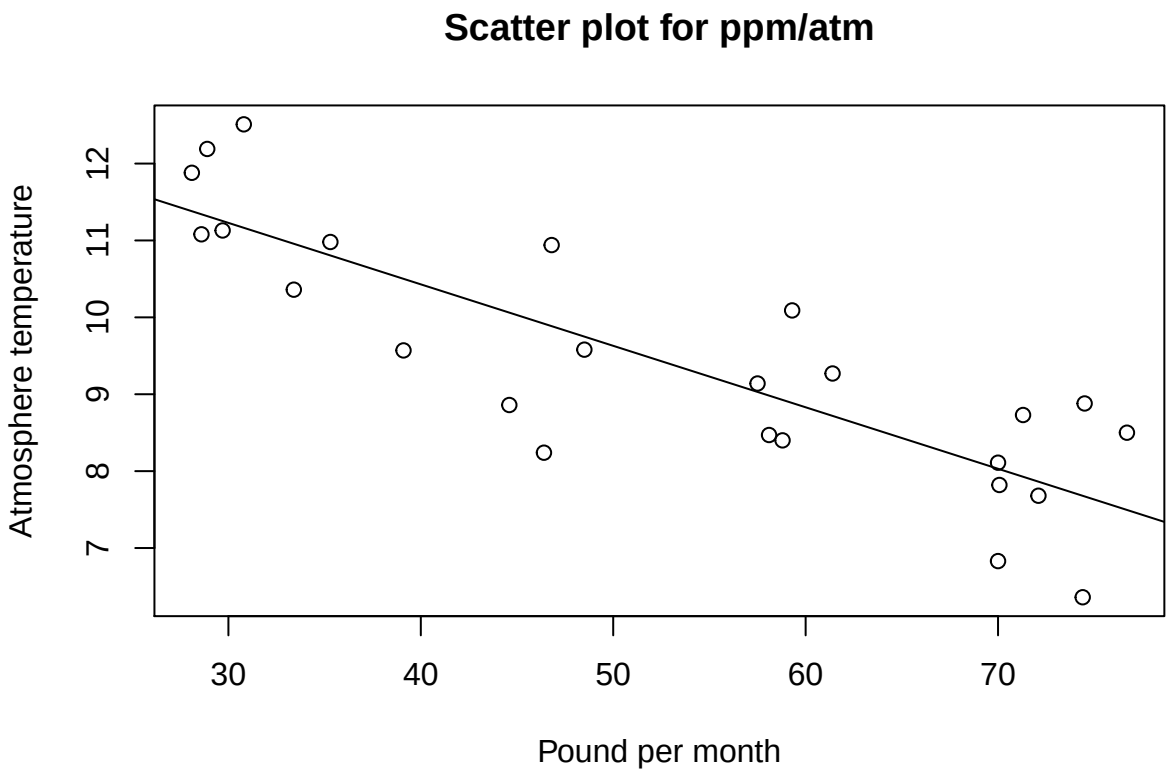
```r
#and p = 0.5249 which is more than significance level 0.05 we we can reject
#the null hypothesis that the variance of the residuals is constant and can say
#that Homoscedasticity is present.


# To test the Normality we can see the QQ plot and can say that
#there is not any gross deviations from  normality.But since the
#number of observations are less than 30 it is safe to do "Shapiro test".
# Obtained p value = 0.099 which is greater than 0.05 which is implying that
#the distribution of the data are not significantly different from normal distribution.
#Thus we can assume the normality.



conf_interval_pH <- predict(reg,
                            newdata=data.frame(x=2050),
                            interval="confidence", level = 0.99)
conf_interval_pH
```

```
##        fit      lwr      upr
## 1 18.17693 6.880469 29.47339
```

```r
# 1:e) CI range is (6.880469 , 29.47339)
```

Q3

```r
# This is an R chunk for QUESTION 3

tempound = data.frame(ppm = c( 35.3, 29.7, 30.8, 58.8, 61.4, 71.3, 74.4, 76.7, 70.07, 57.5, 46.4, 28.9,
                      atm = c(10.98, 11.13, 12.51, 8.40, 9.27, 8.73, 6.36, 8.50, 7.82, 9.14, 8.24, 12.1
X = tempound[,1]
Y = tempound[,2]
N = length(X)
plot(X,Y,xlab="Pound per month", ylab="Atmosphere temperature", main="Scatter plot for ppm/atm")
#As the value of Pound per month increasing we can see that
#value of Atmosphere temperature is also decreasing so we can interpret
#from this scatterplot that is is showing Linear trend.

REG = lm(Y~X)
abline(lm(Y~X))
```




```
tempoundmodel <- lm(Y ~ X, data = tempound)
tempoundmodel%>% anova()
```

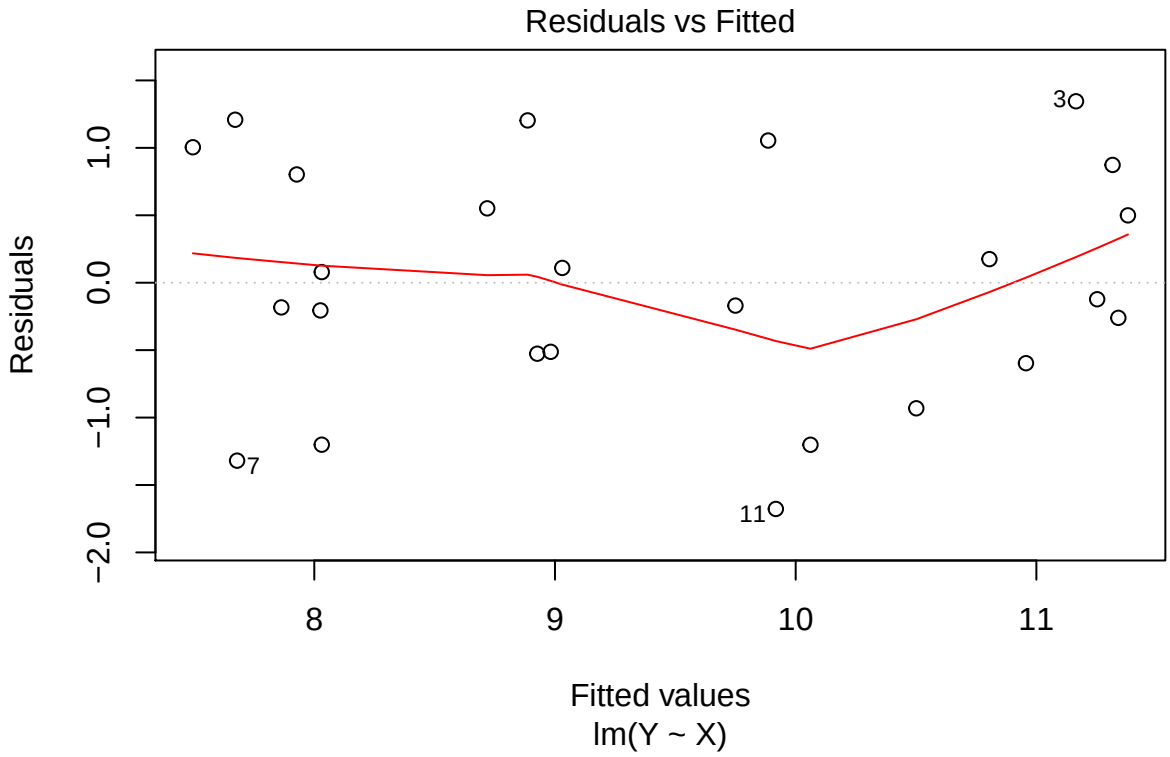
```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1 45.574  45.574  57.462 1.067e-07 ***
## Residuals 23 18.242   0.793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

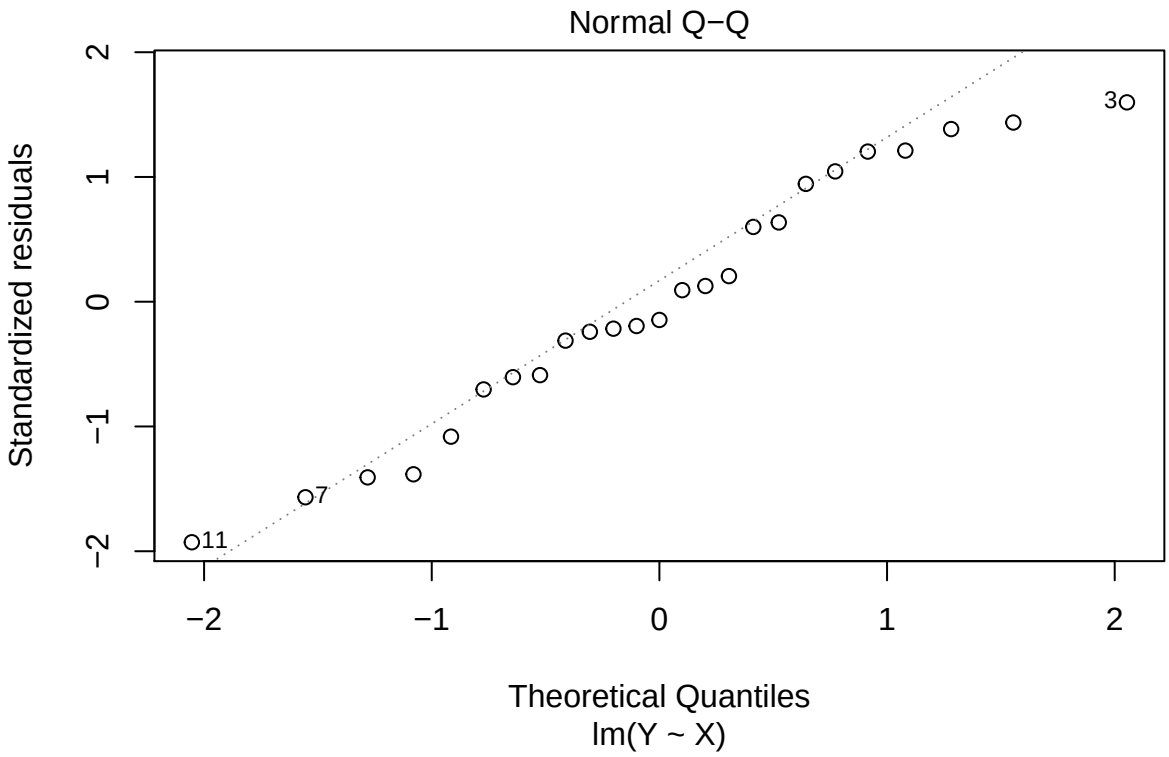
```
#Anova table is a statastical method by which we can compare variability
#between groups.In our case high F -value = 57.46 > Fc(Critical F value= 4.28)
#which indicates that Regression is significant.

REG$residuals
```

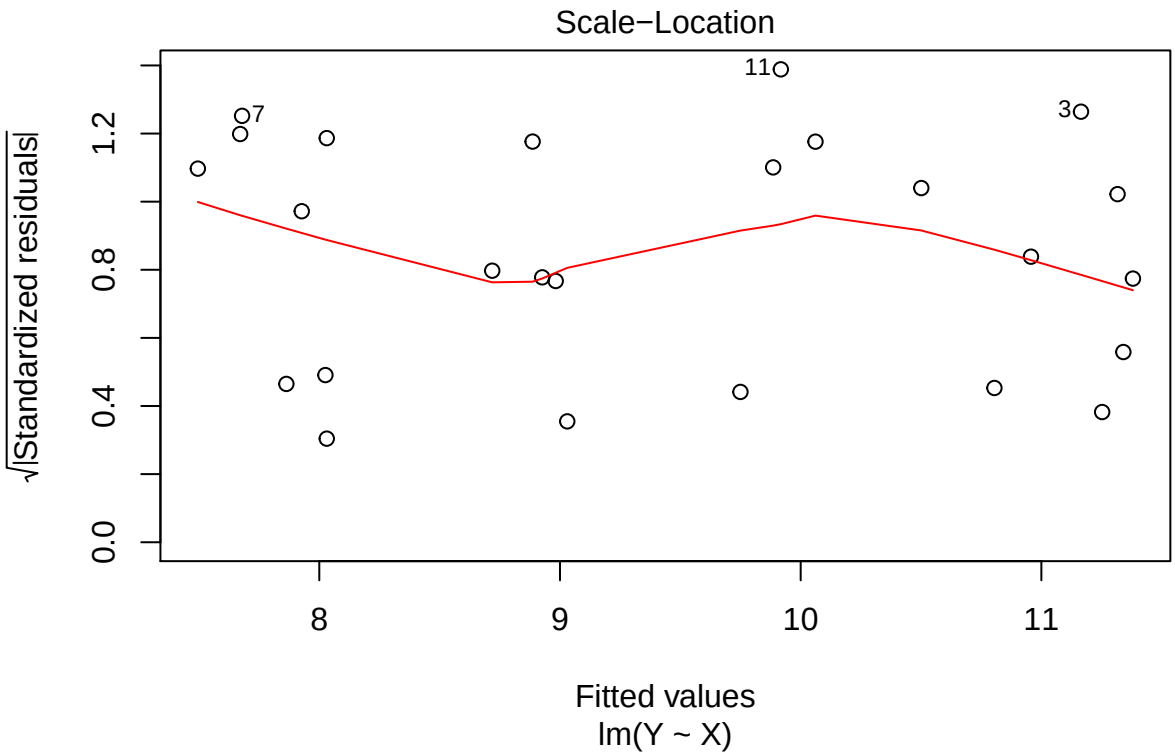
```
##           1           2           3           4           5           6
##  0.17508689 -0.12256596  1.34536585 -0.52636991  0.55146891  0.80285520
##           7           8           9          10          11          12
## -1.31933697  1.00452045 -0.20546855  0.10971068 -1.67760122  0.87348363
##          13          14          15          16          17          18
##  0.49953323 -0.93114868  1.05437399 -0.16973140  1.20359910  0.07893579
##          19          20          21          22          23          24
```

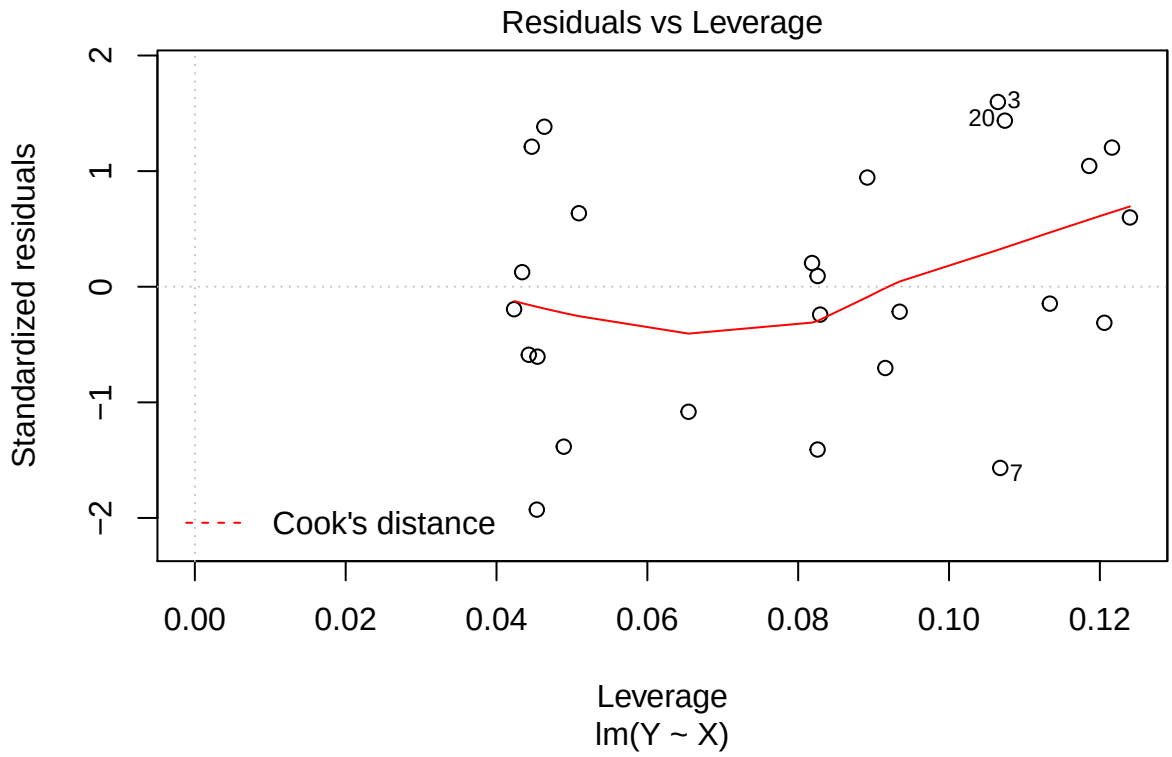
```
## -1.20106421  1.20865683 -0.18319439 -0.51232651 -1.20148963 -0.59679533
##          25
## -0.26049777
```

```r
# Residual values gives the difference between trend value and observed value.
plot(tempoundmodel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X)
3
7
11
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
11
7
3
8
9
10
11
Fitted values
lm(Y ~ X)




```
sumX = sum(X)
sumX
```

```
## [1] 1314.37
```

```
sumY = sum(Y)
sumY
```

```
## [1] 235.6
```

```
sumX2 = sum(X^2)
sumX2
```

```
## [1] 76234.73
```

```
sumY2 = sum(Y^2)
sumY2
```

```
## [1] 2284.11
```

```
sumXY = sum(X*Y)
sumXY
```

```
## [1] 11816.51
```

```
s_xx = sumX2 - sumX^2/N
s_xx
```

```
## [1] 7131.995
```

```
s_yy = sumY2 - sumY^2/N
s_yy
```

```
## [1] 63.8158
```

```
s_xy = sumXY - sumX*sumY/N
s_xy
```

```
## [1] -570.1175
```

```
beta = s_xy/s_xx
beta
```

```
## [1] -0.07993801
```

```
# Slope calculation
alpha = mean(Y) - beta* mean(X)
alpha
```

```
## [1] 13.62672
```

```
# intercept calculation
Rsq = s_xy^2/(s_xx*s_yy)
Rsq
```

```
## [1] 0.71415
```

```
#Coefficient of determination
#71.4% value indicates good regression prediction.
s = sqrt((s_yy - beta*s_xy)/(N-2))
s
```

```
## [1] 0.8905725
```

```
# Standard deviation from error
s_alpha = s*sqrt(sumX2/(N*s_xx))
s_alpha
```

```
## [1] 0.5823314
```

```
# Standard deviation for intercept
s_beta = s/sqrt(s_xx)
s_beta
```

```
## [1] 0.01054542
```

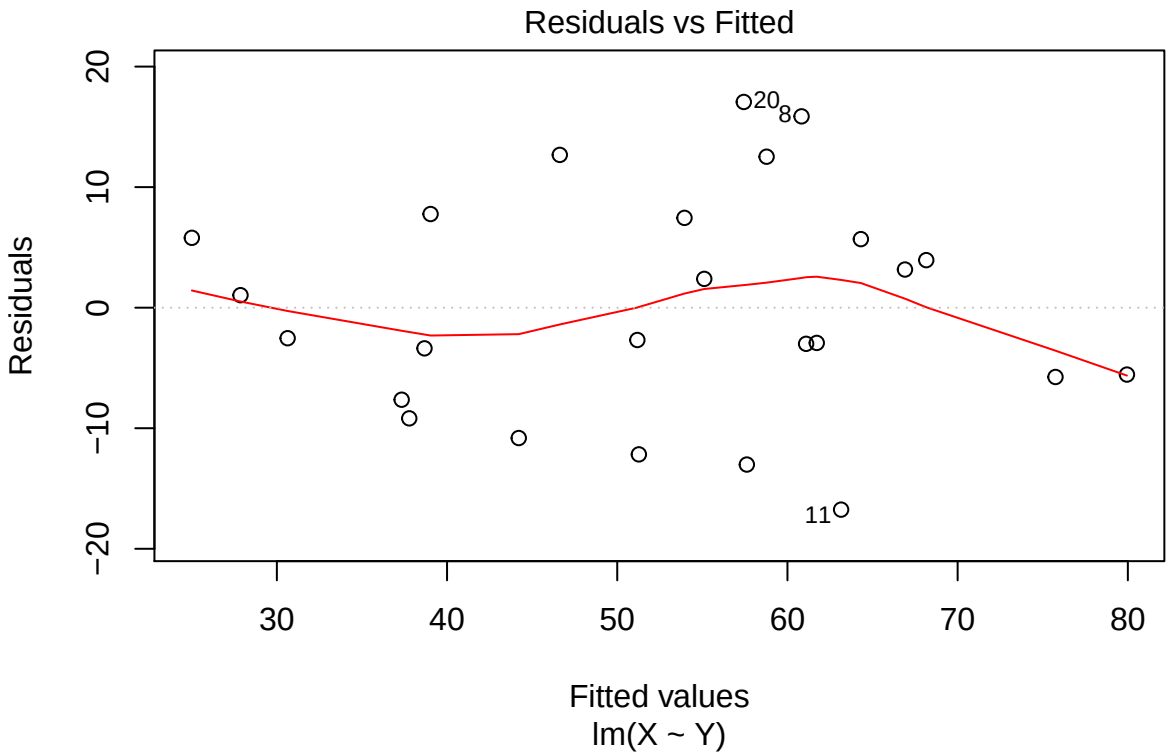
```r
# Standard deviation for slope
tempound.lm<-lm(X~Y, data= tempound)
summary(tempound.lm)
```

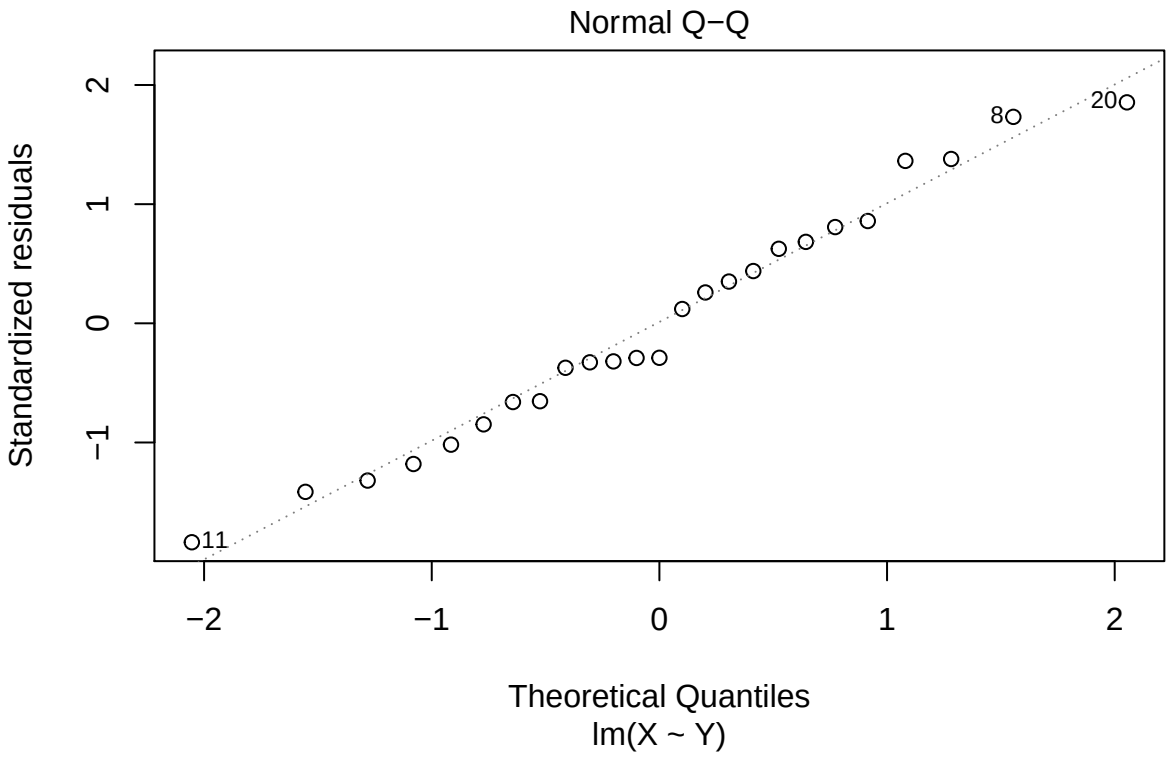
```
##
## Call:
## lm(formula = X ~ Y, data = tempound)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.752  -5.749  -2.533   5.795  17.065
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  136.767     11.265   12.14 1.75e-11 ***
## Y             -8.934      1.179   -7.58 1.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.415 on 23 degrees of freedom
## Multiple R-squared:  0.7142, Adjusted R-squared:  0.7017
## F-statistic: 57.46 on 1 and 23 DF,  p-value: 1.067e-07
```

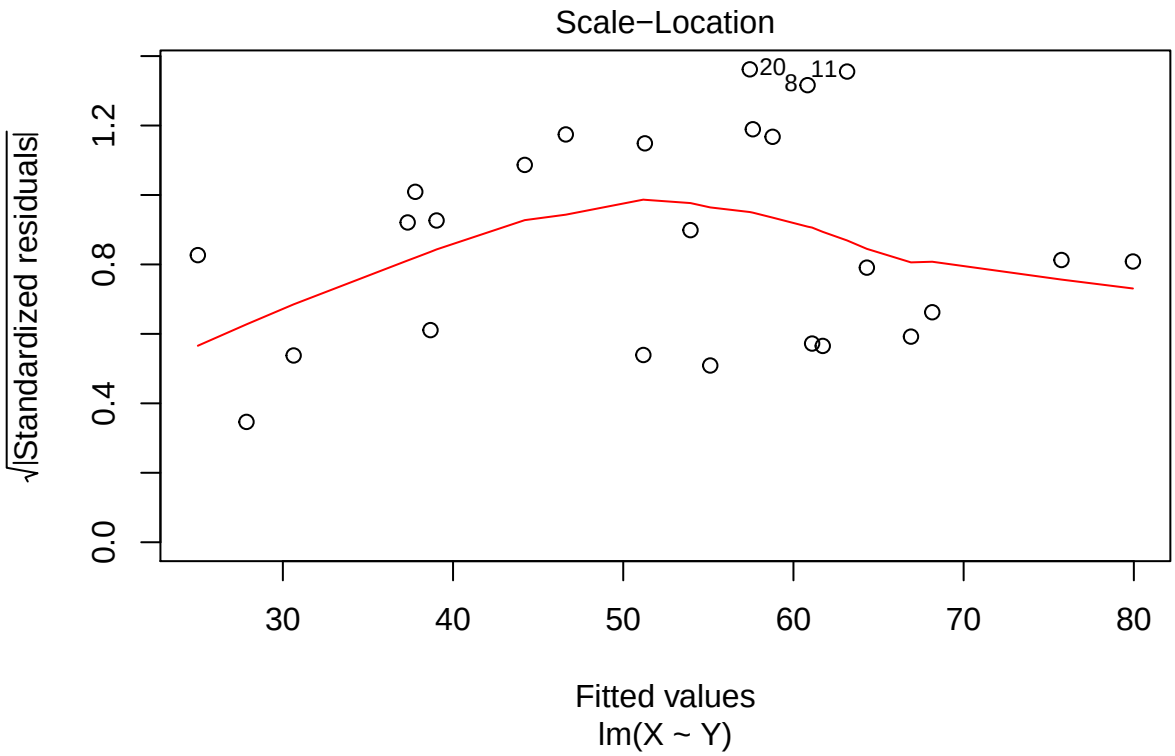
```r
# From the above p values we can conclude that both slope and intercept are significant.
fit <- lm(Y ~ X, data = tempound)
confint(fit, level=0.95)
```

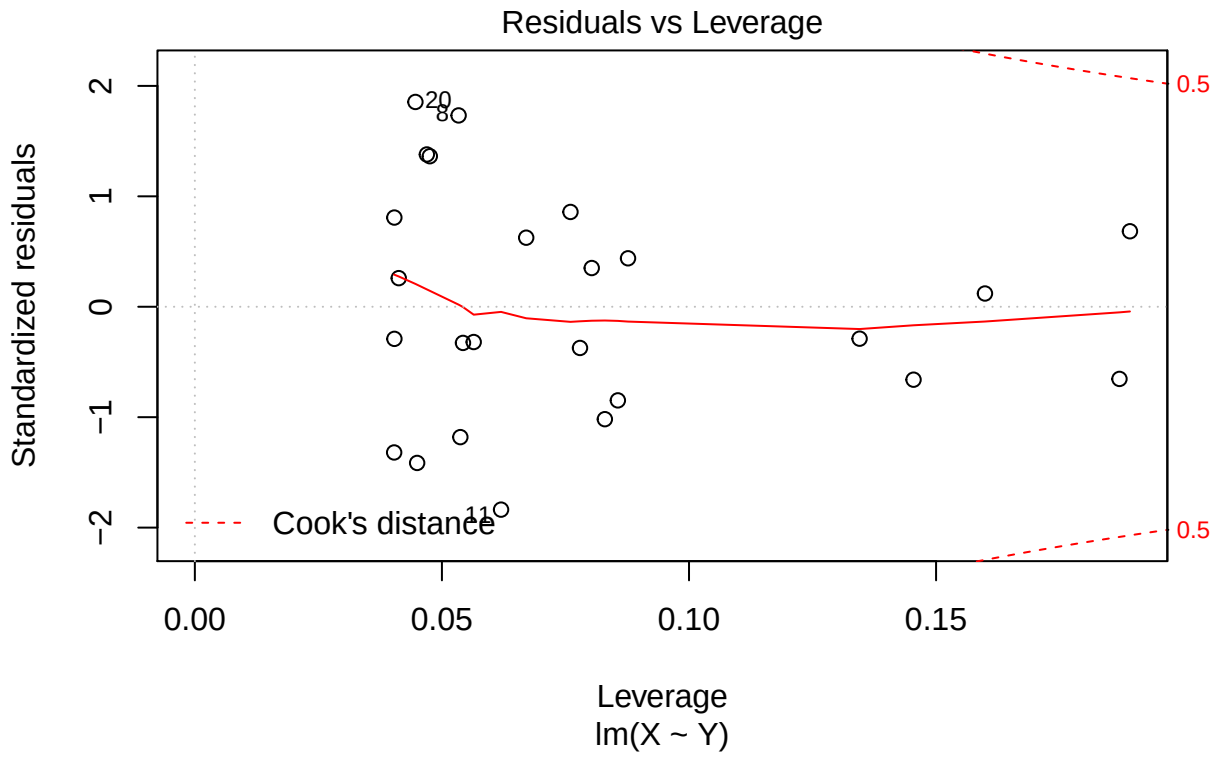
```
##                  2.5 %      97.5 %
## (Intercept) 12.4220806 14.83136904
## X           -0.1017529 -0.05812315
```

```r
# CI level for slope with significance level 95% = (-0.101, -0.05)
# CI level for intercept with significance level 95% = (12.42, 14.83)

plot(tempound.lm)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
30
40
50
60
70
80
Fitted values
lm(X ~ Y)
20
8
11


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(X ~ Y)
11
8
20
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
20
8
11
30
40
50
60
70
80
Fitted values
lm(X ~ Y)




```
ncvTest(tempound.lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3080335, Df = 1, p = 0.57889
```

```
shapiro.test(REG$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  REG$residuals
## W = 0.9595, p-value = 0.4046
```

```
durbinWatsonTest(tempound.lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01648246      1.920111   0.726
##  Alternative hypothesis: rho != 0
```

```
# In residual vs fitted graph we can see that the red line is slightly curved
# and the residuals seem to increase as the fitted Y values increase so
# there may be heteroscedasticity.
```

```
# So we do "NCV test" and p = 0.5788 which is more than significance level 0.05
# we we can reject the null hypothesis that the variance of the residuals is
# constant and can say that  Homoscedasticity is present.


# To test the Normality we can see the QQ plot and can say that
# there is not any gross deviations from normality.
# But since the number of observations are less than 30
# it is safe to do "Shapiro test".
# Obtained p value = 0.4046 which is greater than 0.05 which is implying
# that the distribution of the data are not significantly different
# from normal distribution. Thus we can assume the normality.
```