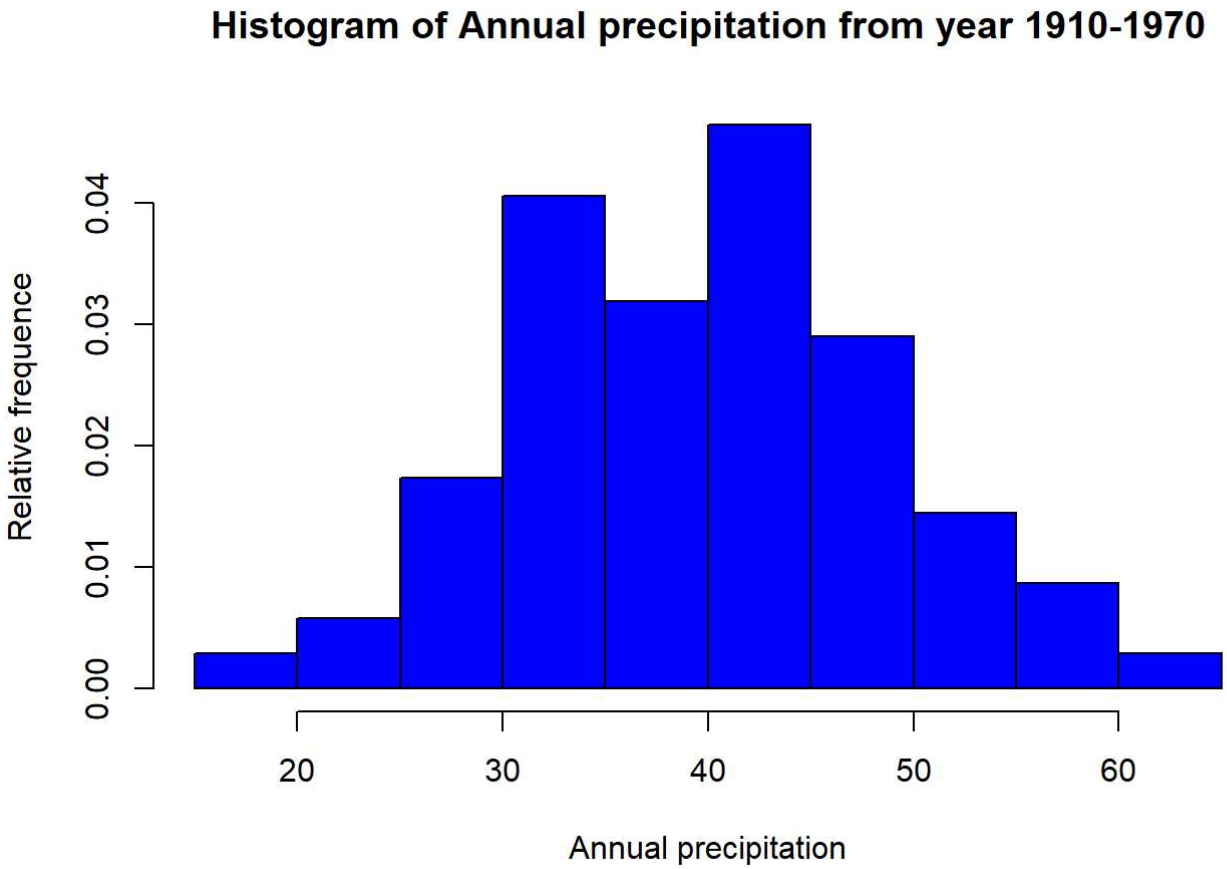
Kaustubh_203350013_RExercise-1.R

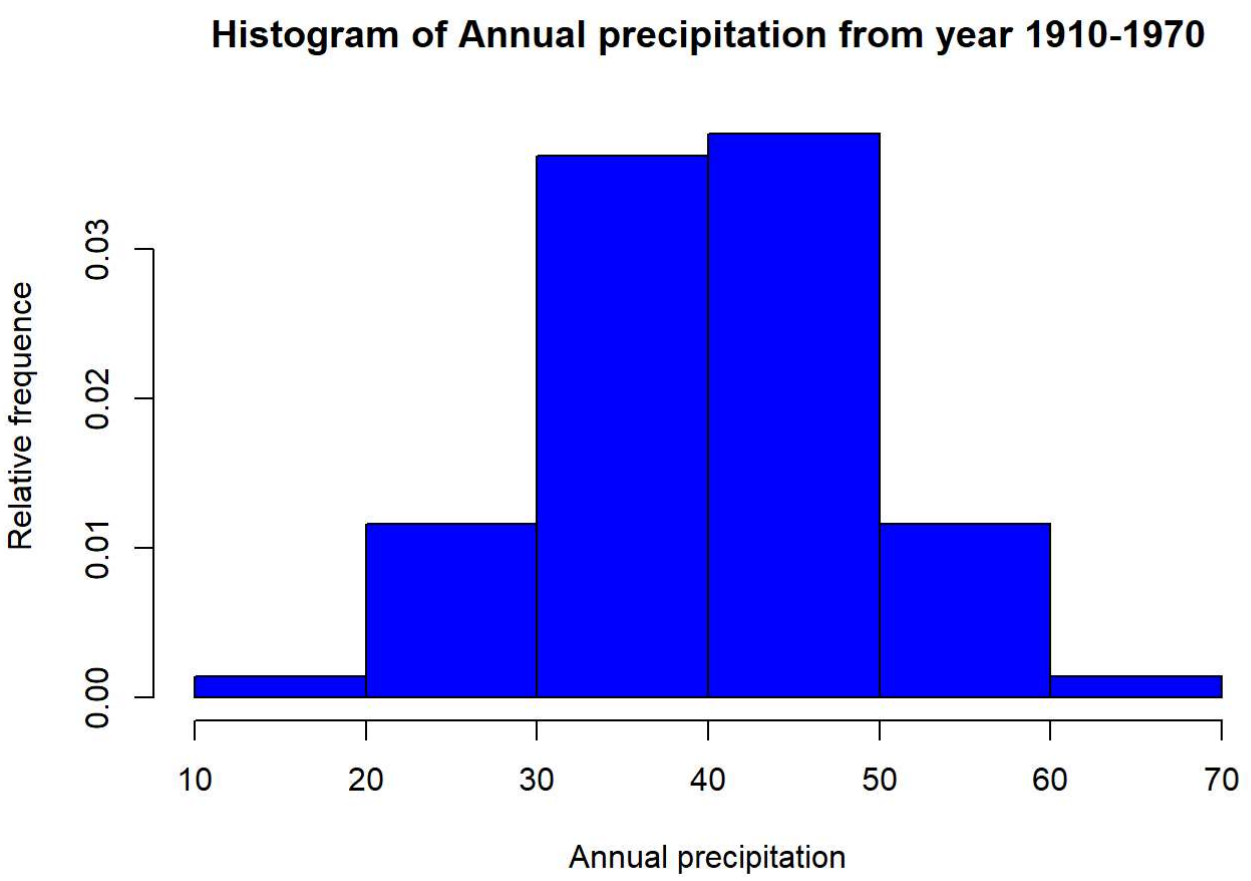
Kaustubh Patil

2021-02-16

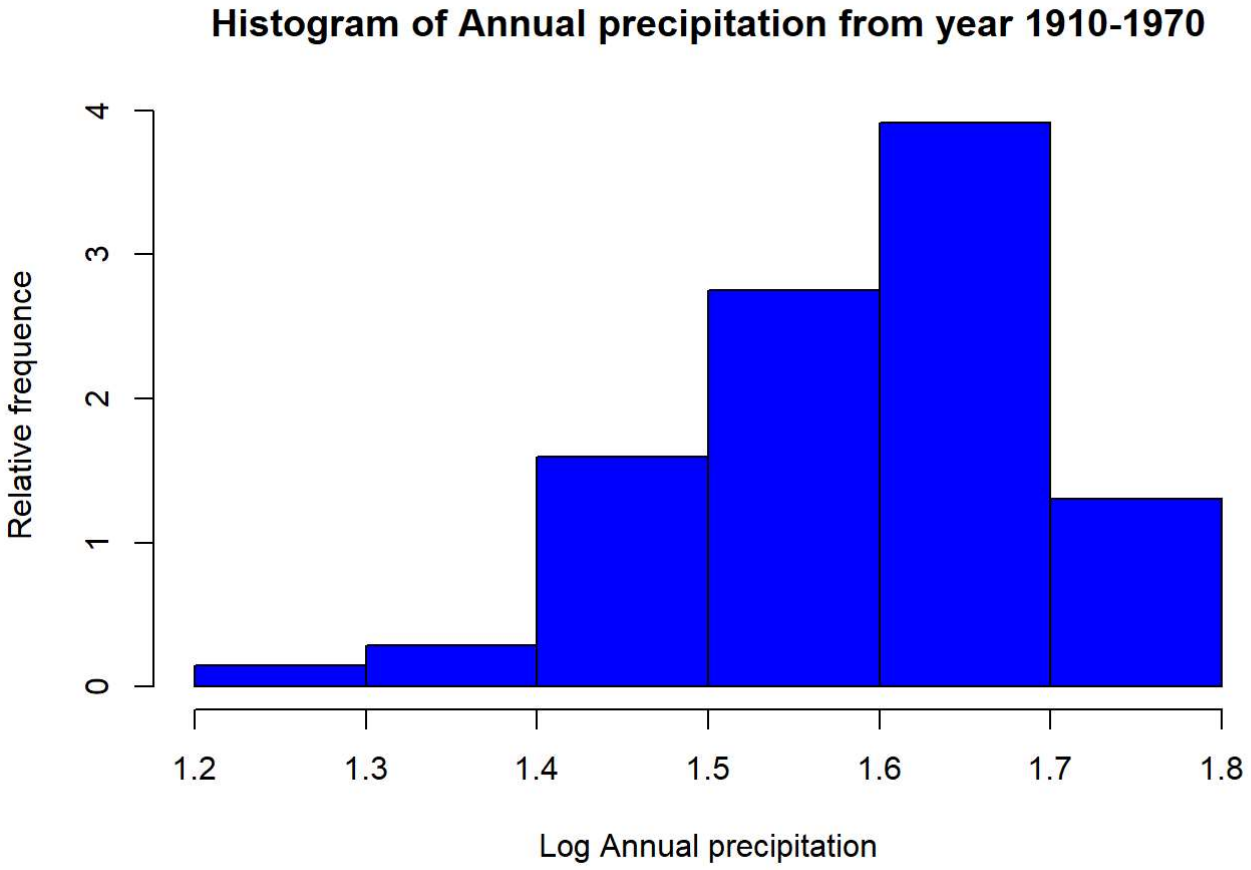
```
#Name: Kaustubh Patil; Roll. no.: 203350013
# Set the working directory
setwd("C:\\Users\\Kaustubh Patil\\OneDrive\\Desktop\\GNR640\\R")

#import the libraries
library(readxl)
library(ggplot2)

#Read the data
rain = read_excel('data_annual_precipitation.xlsx')

#Convert data into vectior format

temp = unlist(rain[,-1])
rain_vec = matrix(temp, ncol=1)

#1) Plot histogram and add title in histogram

hist(rain_vec, freq=FALSE, main = 'Histogram of Annual precipitation from year 1910-1970', col='blue',
 xlab= 'Annual precipitation', ylab='Relative frequence')
```



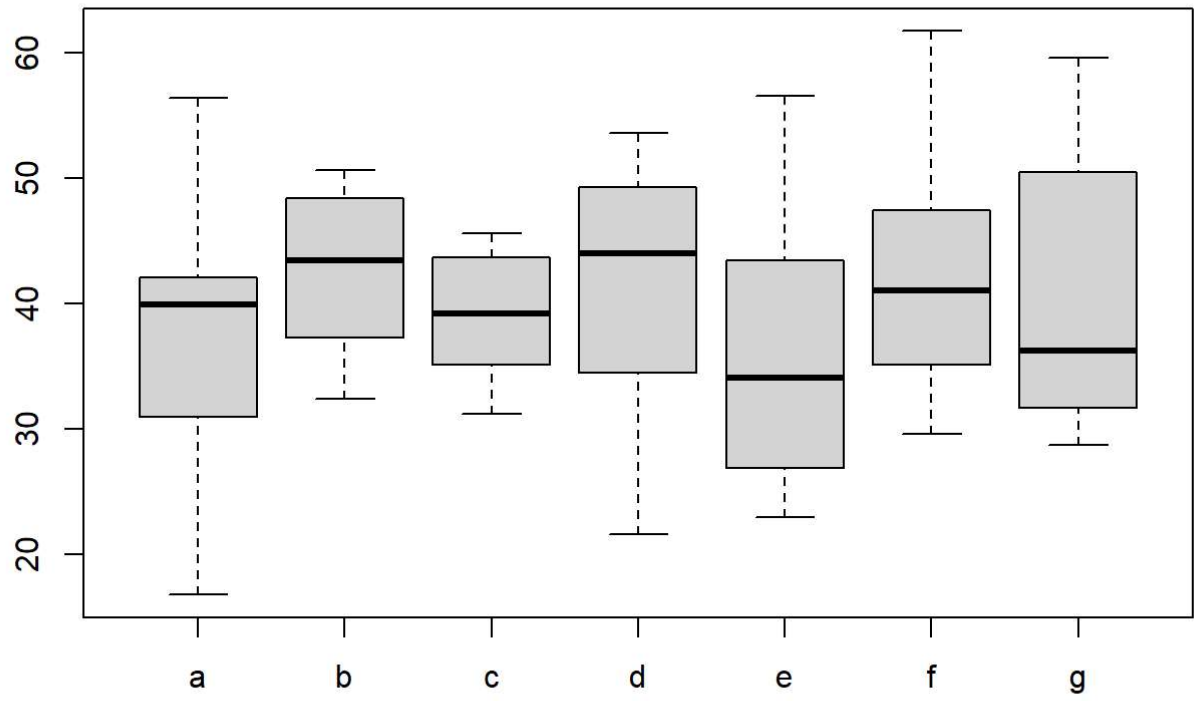
```
#2) Change the number of bins in histogram
hist(rain_vec, freq=FALSE,breaks= 5, main = 'Histogram of Annual precipitation from year 1910-1970', co
l='blue', xlab= 'Annual precipitation', ylab='Relative frequence')
```




```
#3) Plot histogram of logarithm of data
log_rain = log10(rain_vec)
hist(log_rain, freq=FALSE,breaks= 5, main = 'Histogram of Annual precipitation from year 1910-1970', co
l='blue', xlab= 'Log Annual precipitation', ylab='Relative frequence')
```




```
#4) Plot a boxplot. Change boxplot labels on x axis. Replace years with alphabets a, b, c, ...
boxplot(rain[,-1], names=c("a","b","c","d","e","f","g"))
```



```r
#5) Explore Chi Square test in R.
#Null hypothesis (H0): the row and the column variables are independent.
#(H1): row and column variables are dependent


chtest = chisq.test(rain[,-1:-2])
chtest
```

```
##
##  Pearson's Chi-squared test
##
## data:  rain[, -1:-2]
## X-squared = 103.57, df = 45, p-value = 1.632e-06
```

```r
# Observed counts
chtest$observed
```

```
##       1920 1930 1940 1950 1960 1970
##  [1,] 48.7 44.8 49.3 31.2 46.0 33.9
##  [2,] 44.1 34.0 44.2 27.0 44.3 31.7
##  [3,] 42.8 45.6 41.7 37.0 37.8 31.5
##  [4,] 48.4 37.3 30.8 46.8 29.6 59.6
##  [5,] 34.2 43.7 53.6 26.9 35.1 50.5
##  [6,] 32.4 41.8 34.5 25.4 49.7 38.6
##  [7,] 46.4 41.1 50.3 23.0 36.6 43.4
##  [8,] 38.9 31.2 43.8 56.5 32.5 28.7
##  [9,] 37.3 35.2 21.6 43.4 61.7 32.0
## [10,] 50.6 35.1 47.1 41.3 47.4 51.8
```

```r
# As p-value is close to 0,  the row and the column variables are significantly associated.

# Expected counts
round(chtest$expected,2)
```

```
##        1920  1930  1940  1950  1960  1970
##  [1,] 44.62 41.04 43.90 37.75 44.30 42.30
##  [2,] 39.60 36.42 38.95 33.50 39.31 37.53
##  [3,] 41.55 38.21 40.87 35.15 41.24 39.38
##  [4,] 44.38 40.82 43.65 37.54 44.05 42.06
##  [5,] 42.88 39.44 42.18 36.28 42.57 40.65
##  [6,] 39.09 35.95 38.45 33.06 38.80 37.05
##  [7,] 42.32 38.93 41.63 35.80 42.01 40.11
##  [8,] 40.70 37.44 40.04 34.43 40.41 38.58
##  [9,] 40.63 37.37 39.97 34.37 40.34 38.51
## [10,] 48.03 44.18 47.25 40.63 47.68 45.53
```

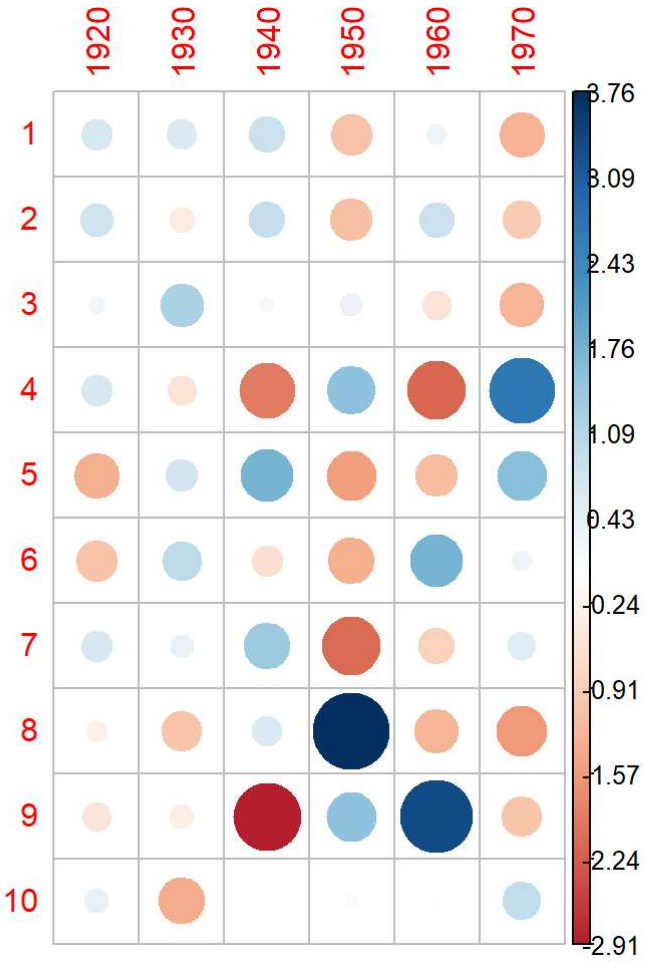
```r
#Pearson residuals
round(chtest$residuals, 3)
```

```
##         1920   1930   1940   1950   1960   1970
##  [1,]  0.610  0.586  0.816 -1.066  0.256 -1.291
##  [2,]  0.716 -0.401  0.841 -1.122  0.796 -0.952
##  [3,]  0.194  1.195  0.130  0.313 -0.536 -1.256
##  [4,]  0.604 -0.550 -1.945  1.512 -2.177  2.704
##  [5,] -1.326  0.678  1.758 -1.557 -1.145  1.546
##  [6,] -1.070  0.976 -0.637 -1.333  1.750  0.255
##  [7,]  0.627  0.349  1.344 -2.139 -0.835  0.519
##  [8,] -0.283 -1.019  0.594  3.761 -1.244 -1.591
##  [9,] -0.523 -0.355 -2.906  1.540  3.364 -1.050
## [10,]  0.371 -1.366 -0.022  0.105 -0.041  0.930
```

```
#visualize Pearson residuals
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(chtest$residuals, is.cor = FALSE)
```



```
#Plot empirical and theoretical CDFs. Both plots should be overlayed under single axes.
x=rgamma(rain_vec, 2, 1)
tiff("ecdf and cdf.tiff", width = 4, height = 4, pointsize = 1/300, units = 'in', res = 300)
plot(ecdf(x), xlab = 'Sample', ylab = '', main = 'Empirical Cumluative Distribution\n Precipitation')
curve(pgamma(x, shape = 2, scale = 1), -1, 10, add=TRUE, col="red")

#7) Slide 6 of Lecture 9
#1) Plot scatter diagram between Hydrocarbon level and purity
hc_purity = read_excel('data_linear_regression.xlsx')
#Assign column header
colnames(hc_purity) = c("o","hc","pu")
plot(hc_purity$hc,hc_purity$pu, main = "Scatter plot between Hydrocarbon level and purity",pch=19, fram
e = FALSE, xlab ="Hydrocarbon level(%)", ylab = "Purity (%)")

#2) Fit linear regression between Hydrocarbon level and purity
model.1 = lm( pu ~ hc, data = hc_purity)
summary(model.1)
```

```
##
## Call:
## lm(formula = pu ~ hc, data = hc_purity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83029 -0.73334  0.04497  0.69969  1.96809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.283      1.593   46.62  < 2e-16 ***
## hc            14.947      1.317   11.35 1.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.087 on 18 degrees of freedom
## Multiple R-squared:  0.8774, Adjusted R-squared:  0.8706
## F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

```
#3) Check residuals and verify if the linear regression model is adequate or not
model1.res = resid(model.1)
qqnorm(model1.res, main = "Model.1 Residuals")
qqline(model1.res, col="red")
hist(model1.res, breaks=13, col='red')
shapiro.test(model1.res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1.res
## W = 0.97964, p-value = 0.9293
```

```
#Ans: Residual is approximately normal from qqplot and histogram

#Graph between fitted values and residuals
plot(model.1$fitted.values, model1.res, ylab="Residuals", xlab="Fitted values", main="Graph between fit
ted values and residuals")
abline(0, 0)
#Ans: There is no significant pattern observed (error normally distributed) as residuals and the fitted
values are uncorrelated

#8) Fit linear regression model with N, E, N^2, E^2, E*N as inputs and Aquifer height as output. 1) Com
ment if this model resulted in any improvement in R^2. 2) Conduct residual analysis (as in Q. 7)
aq = read.table("AQUIFER.txt", skip=1)
##Assign column header
colnames(aq) = c("E","N","wt")
#convert water table height in ft to m
aq$wt=aq$wt*0.3048

##sumarise data
summary(aq)
```

```
##        E                N                 wt
##  Min.   :500361   Min.   :4150248   Min.   :475.5
##  1st Qu.:518465   1st Qu.:4176120   1st Qu.:524.6
##  Median :533366   Median :4197238   Median :552.8
##  Mean   :535668   Mean   :4198439   Mean   :551.0
##  3rd Qu.:553569   3rd Qu.:4220405   3rd Qu.:579.4
##  Max.   :574430   Max.   :4248312   Max.   :623.2
```

```
##Fit linear regression
N_sq = aq$N * aq$N
E_sq = aq$E * aq$E
ExN = aq$E * aq$N
model.2 = lm(wt ~ N+E+N_sq +E_sq+ ExN, data = aq)
summary((model.2))
```

```
##
## Call:
## lm(formula = wt ~ N + E + N_sq + E_sq + ExN, data = aq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.847  -3.366   0.822   3.538  14.807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.161e+05  1.156e+04 -10.043  < 2e-16 ***
## N            5.937e-02  5.439e-03  10.914  < 2e-16 ***
## E           -2.798e-02  3.387e-03  -8.263 5.92e-14 ***
## N_sq        -7.500e-09  6.435e-10 -11.655  < 2e-16 ***
## E_sq        -1.648e-09  1.074e-09  -1.534    0.127
## ExN          6.700e-09  7.781e-10   8.610 7.74e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.598 on 155 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9751
## F-statistic:  1256 on 5 and 155 DF,  p-value: < 2.2e-16
```

```
#Check residuals and verify if the linear regression model is adequate or not
model2.res = resid(model.2)
qqnorm(model2.res, main = "Model.2 Residuals")
qqline(model2.res, col="red")
hist(model2.res, breaks=15, col='red')
shapiro.test(model2.res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2.res
## W = 0.97307, p-value = 0.003096
```

```
#Ans: p-value from shaprio test is 0.003 which is less than 0.05, thus error seems not normaly dist. fr
om qqplot and skewed from histogram

#Graph between fitted values and residuals
plot(model.2$fitted.values, model2.res, ylab="Residuals", xlab="Fitted values", main="Graph between fit
ted values and residuals")
abline(0, 0)
#Ans: As the error was not normally distributed, the points are denser at some location as data is skew
ed.
```