Project 2

Achraf cherkaoui

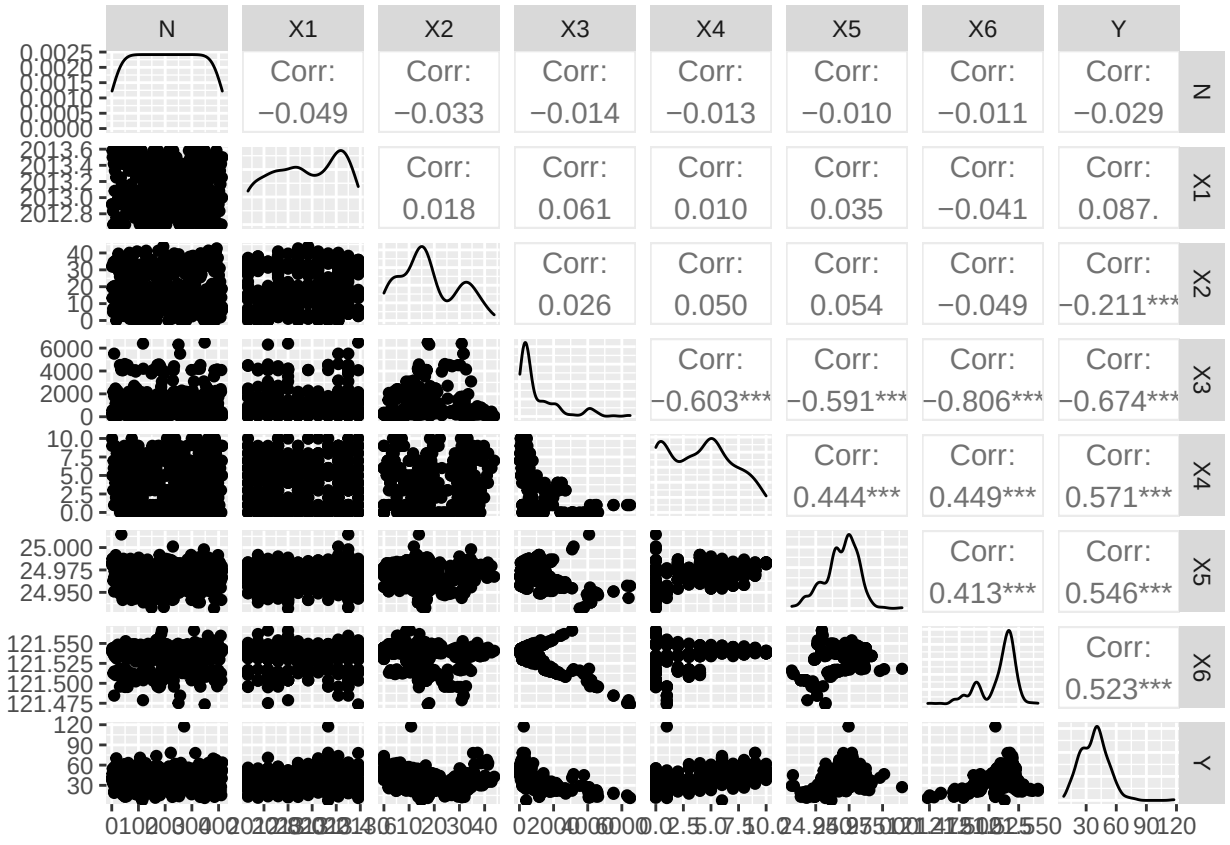
10/8/2021

```
knitr::opts_chunk$set(echo = TRUE)
```

```
library(ggplot2)
library(GGally)
```

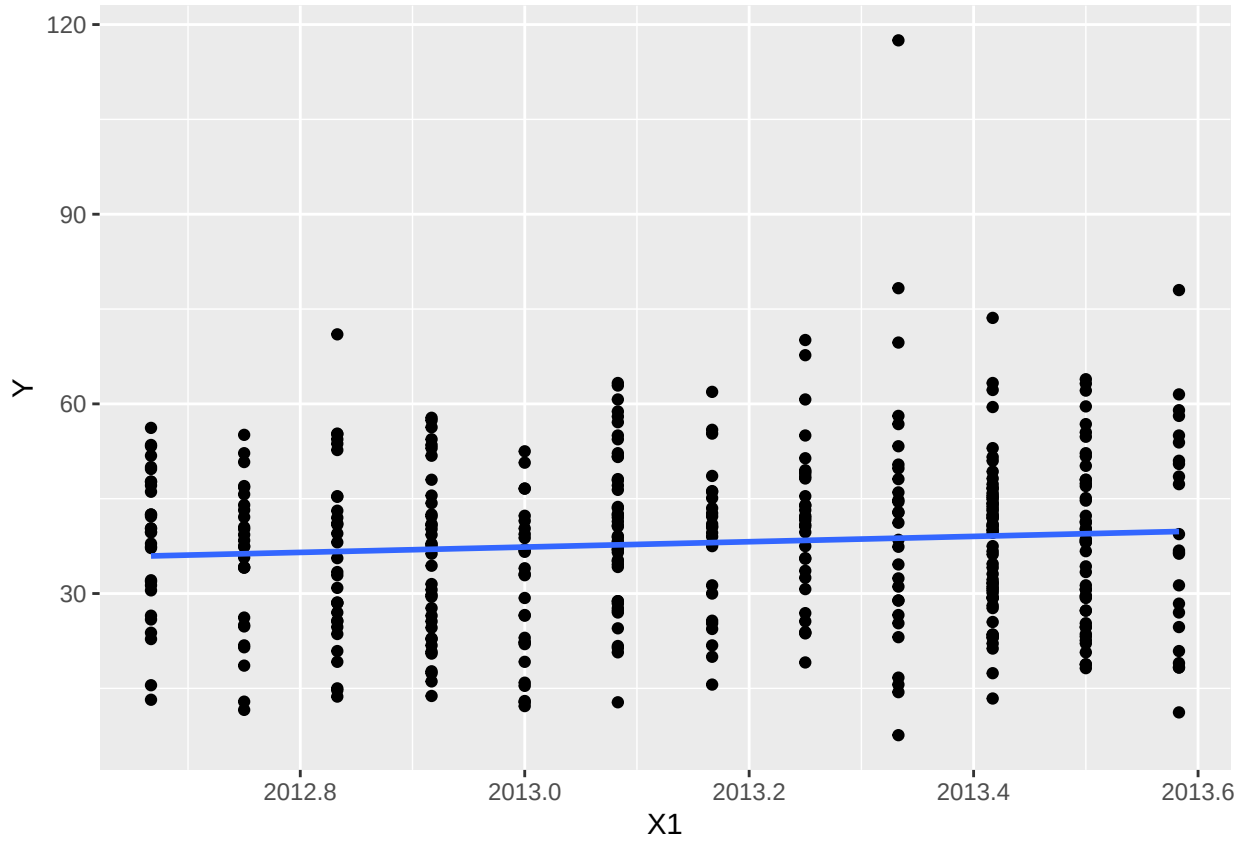
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(readxl)
house <- read_excel("Real Estate Price Prediction.xlsx")
colnames(house) <-c( "N","X1","X2","X3","X4","X5","X6","Y") # change the columns names
View(house)
attach(house)
ggpairs(house) # shows the correlation between all the variables
```



```
ggplot (house ,aes(X1, Y)) +
 geom_point() +
geom_smooth(method = lm ,se =F) # graph a linear model in the data
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(Y,X1) # corrolation between Y and X1
```

```
## [1] 0.08749061
```

```
hp1 <- lm(data = house , Y~X1) # linear regression
summary(hp1)
```

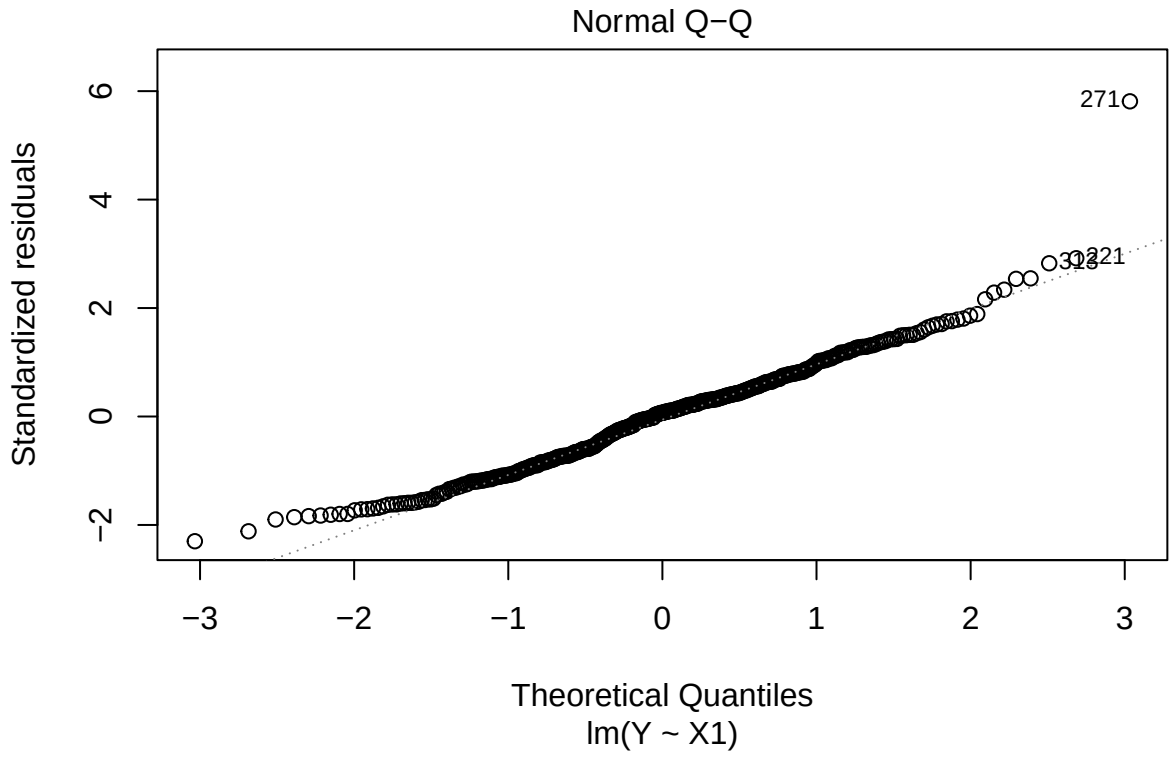
```
##
## Call:
## lm(formula = Y ~ X1, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.157 -10.083   0.921   8.528  78.743
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8461.350   4767.669  -1.775   0.0767 .
## X1              4.222      2.368   1.783   0.0754 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.57 on 412 degrees of freedom
## Multiple R-squared:  0.007655,   Adjusted R-squared:  0.005246
## F-statistic: 3.178 on 1 and 412 DF,  p-value: 0.07537
```

```
pE <- sigma(hp1)*100 /mean(Y)# percentage rate
pE
```

```
## [1] 35.73113
```

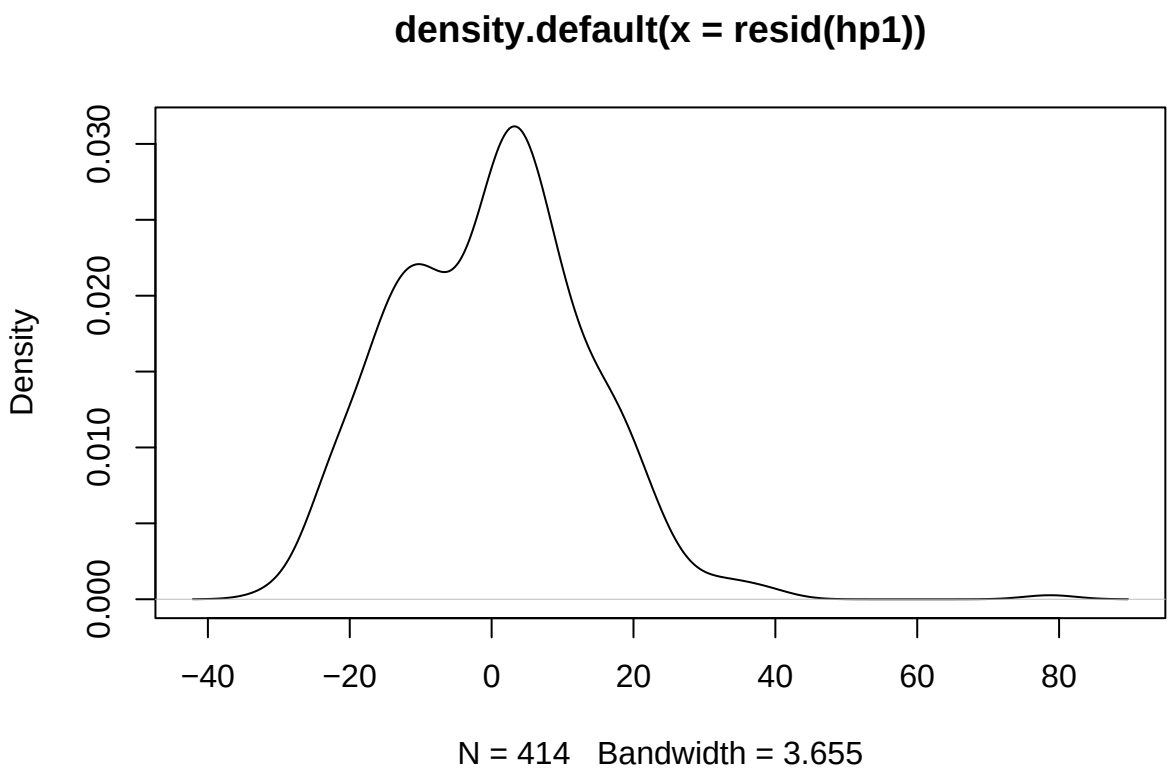
```
plot(hp1 , which = 2) #Q-QPlot to test normality
```




```
shapiro.test(resid(hp1)) # Shapiro test for testing normality
```

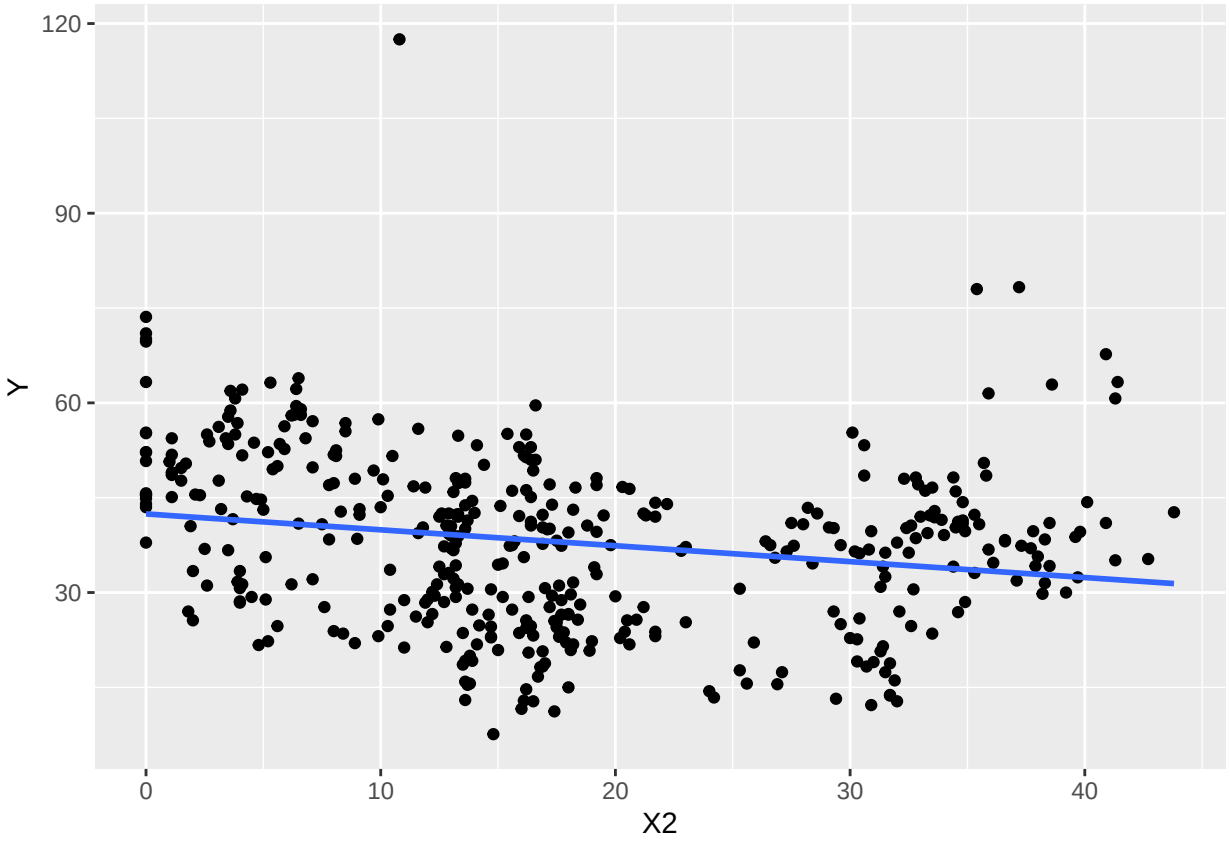
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp1)
## W = 0.97309, p-value = 6.284e-07
```

```
plot(density(resid(hp1)))# to test normality
```




```
ggplot (house ,aes(X2, Y)) +
 geom_point() +
stat_smooth(  method = lm ,se= F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



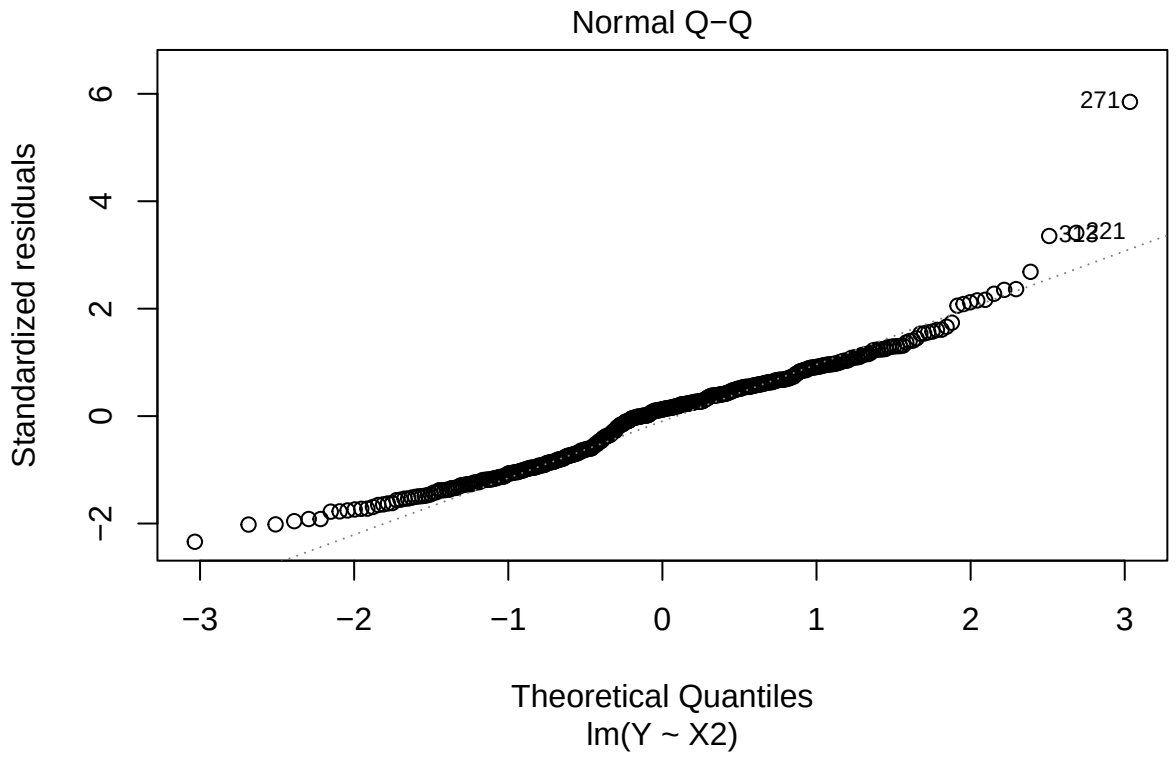

```
cor(Y,X2)
```

```
## [1] -0.210567
```

```
hp2 <- lm(data = house , Y~X2)
summary(hp2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.113 -10.738   1.626   8.199  77.781
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.43470    1.21098  35.042  < 2e-16 ***
## X2          -0.25149    0.05752  -4.372 1.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.32 on 412 degrees of freedom
## Multiple R-squared:  0.04434,    Adjusted R-squared:  0.04202
## F-statistic: 19.11 on 1 and 412 DF,  p-value: 1.56e-05
```

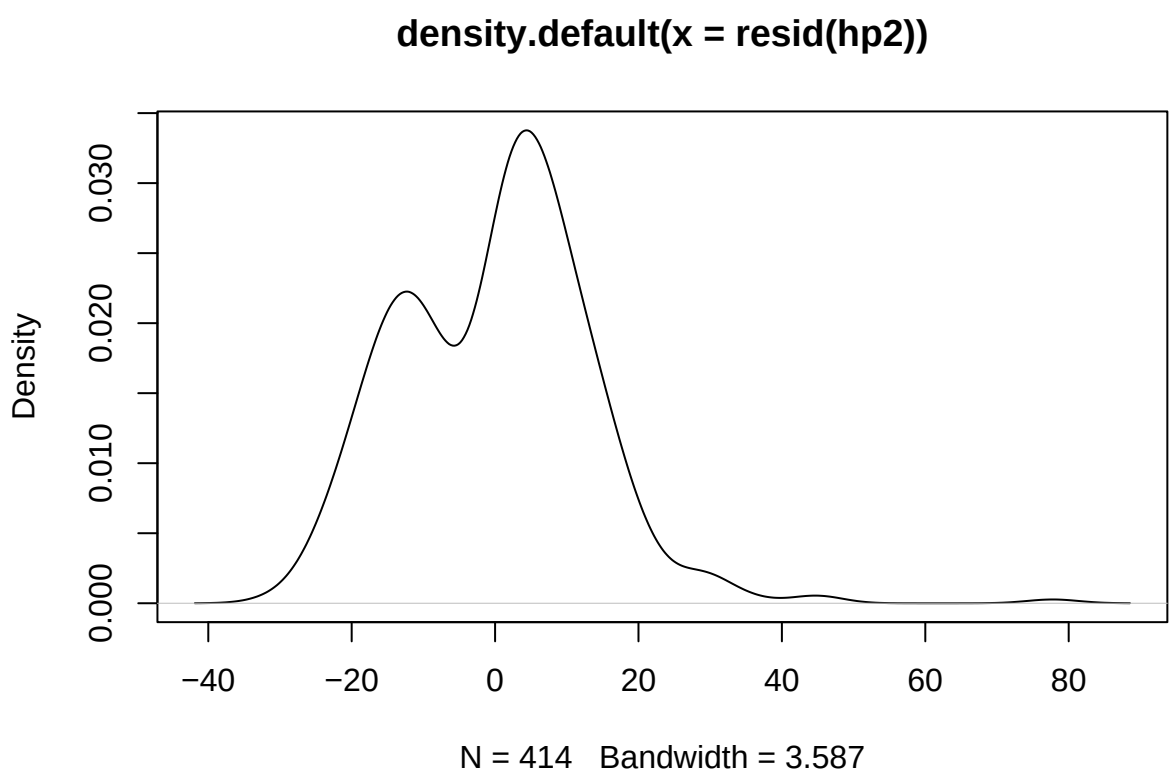
```
plot(hp2 , which = 2)
```




```
shapiro.test(resid(hp2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp2)
## W = 0.96353, p-value = 1.267e-08
```

```
plot(density(resid(hp2)))
```

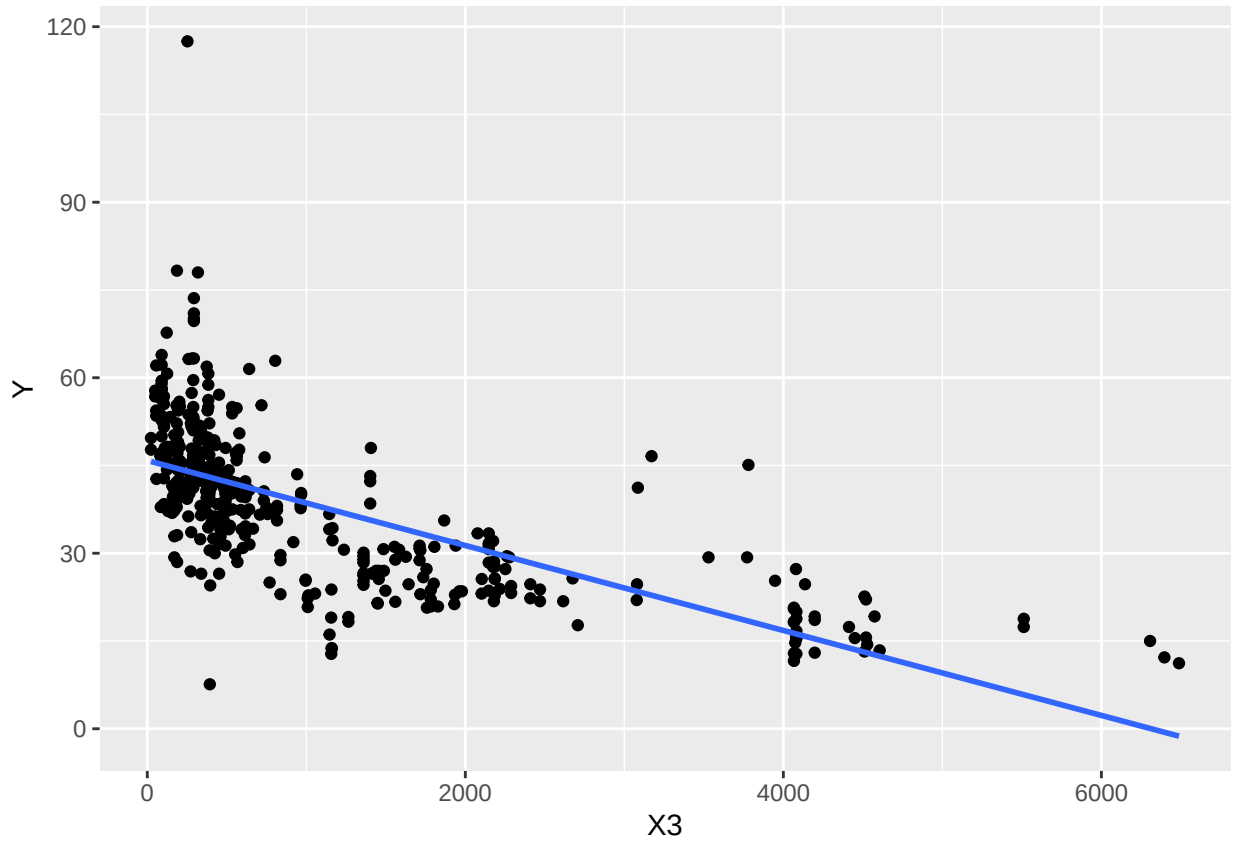



```
library(car)
```

```
## Loading required package: carData
```

```
ggplot (house ,aes(X3, Y)) +
 geom_point() +
stat_smooth(method = lm , se =F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



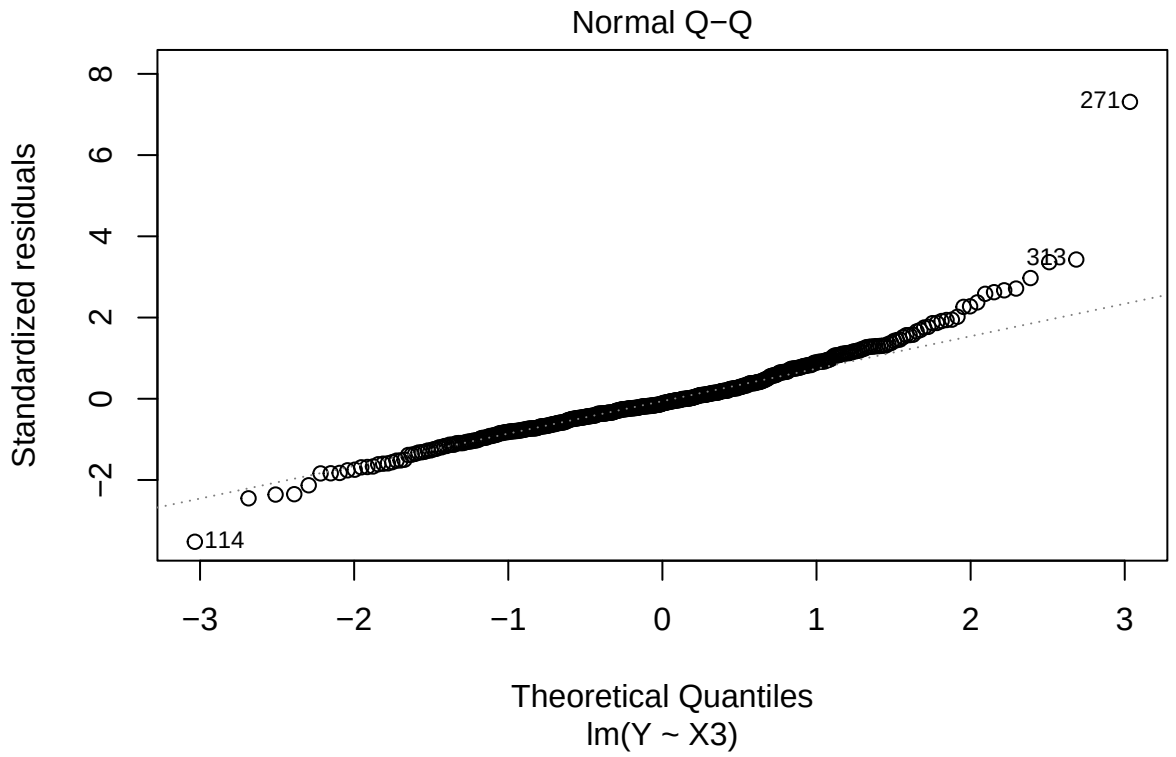

```
cor(Y,X3)
```

```
## [1] -0.6736129
```

```
hp3 <- lm(data = house , Y~X3)
summary(hp3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.396  -6.007  -1.195   4.831  73.483
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.8514271  0.6526105   70.26   <2e-16 ***
## X3          -0.0072621  0.0003925  -18.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.07 on 412 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4524
## F-statistic: 342.2 on 1 and 412 DF,  p-value: < 2.2e-16
```

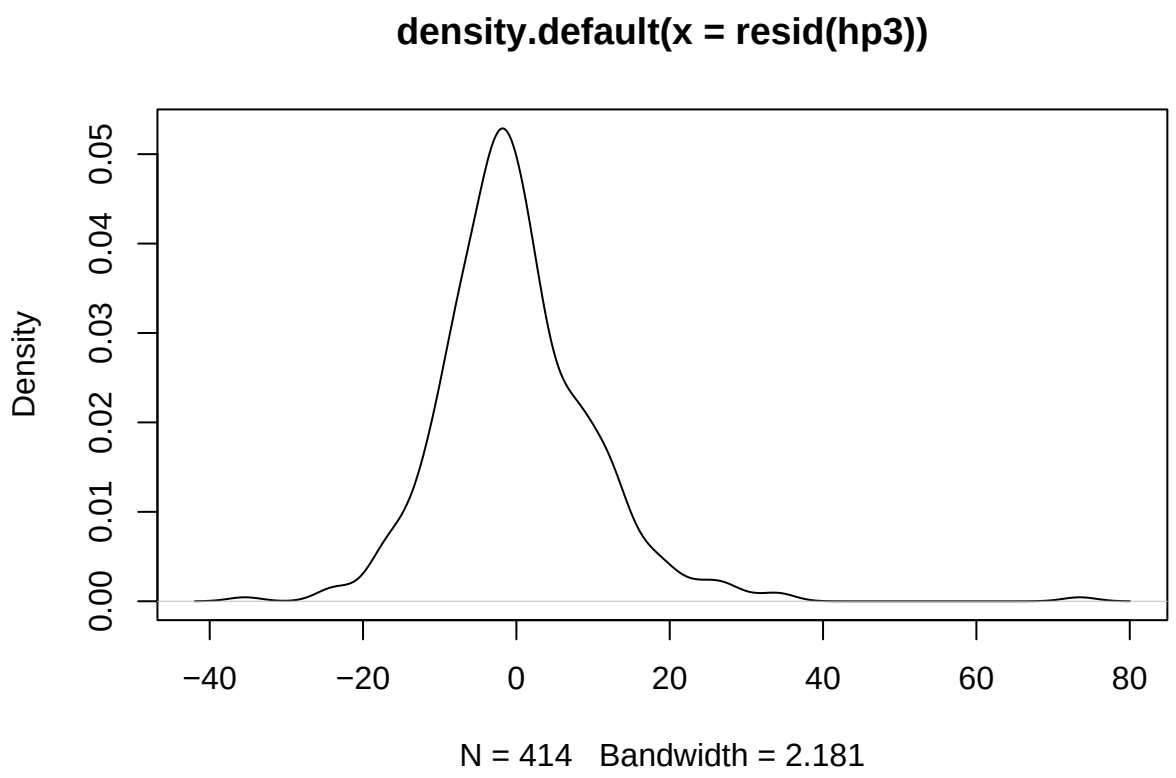
```r
plot(hp3 , which = 2)
```



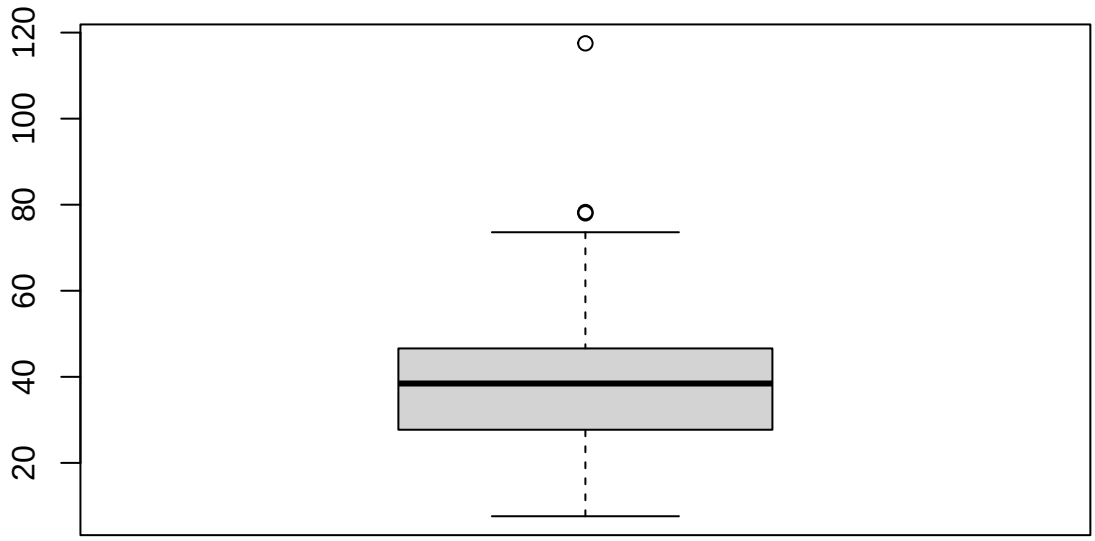

```r
shapiro.test(resid(hp3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp3)
## W = 0.93245, p-value = 9.639e-13
```

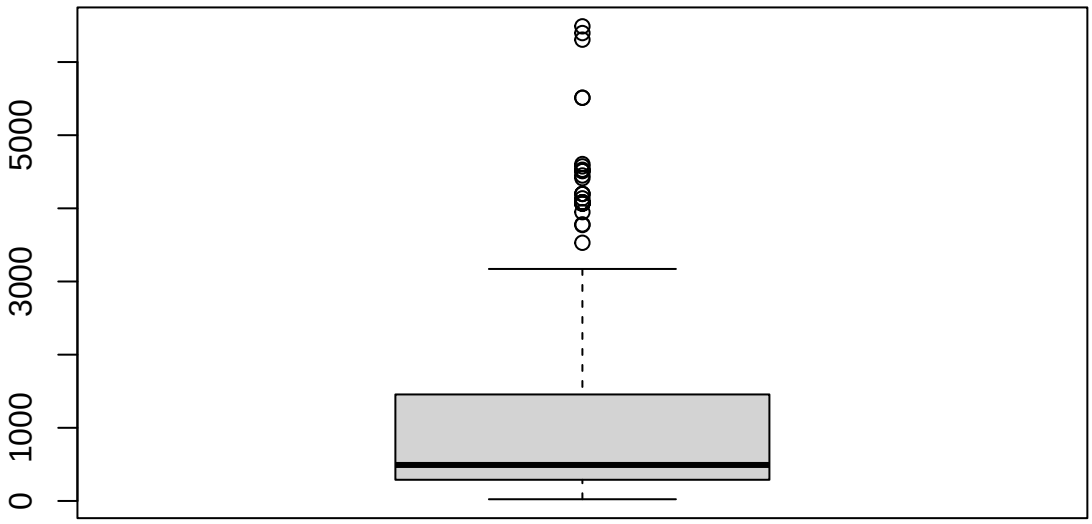
```r
plot(density(resid(hp3)))
```




```
boxplot(Y ) # to view outliers
```

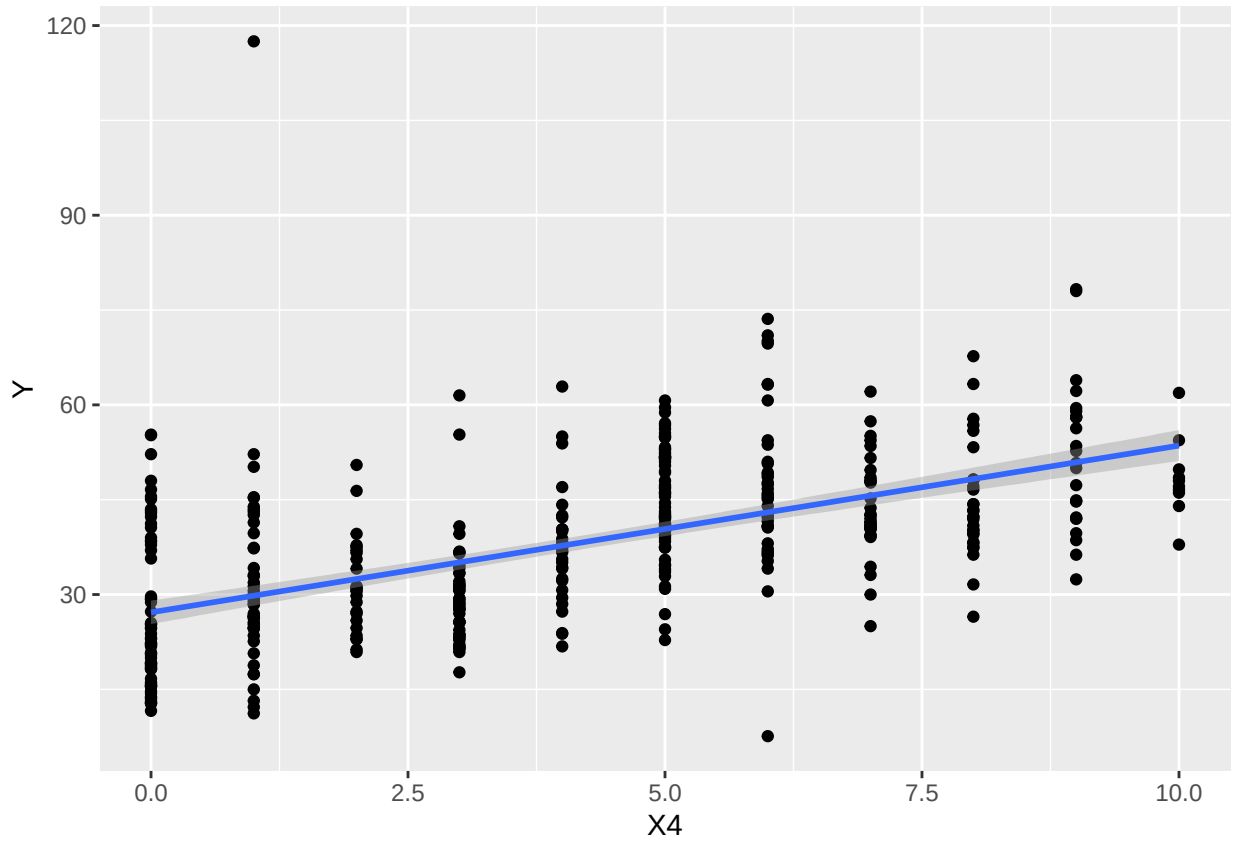



```
boxplot(X3) # to view ou tliers
```




```
ggplot (house ,aes(X4, Y)) +
 geom_point() +
stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



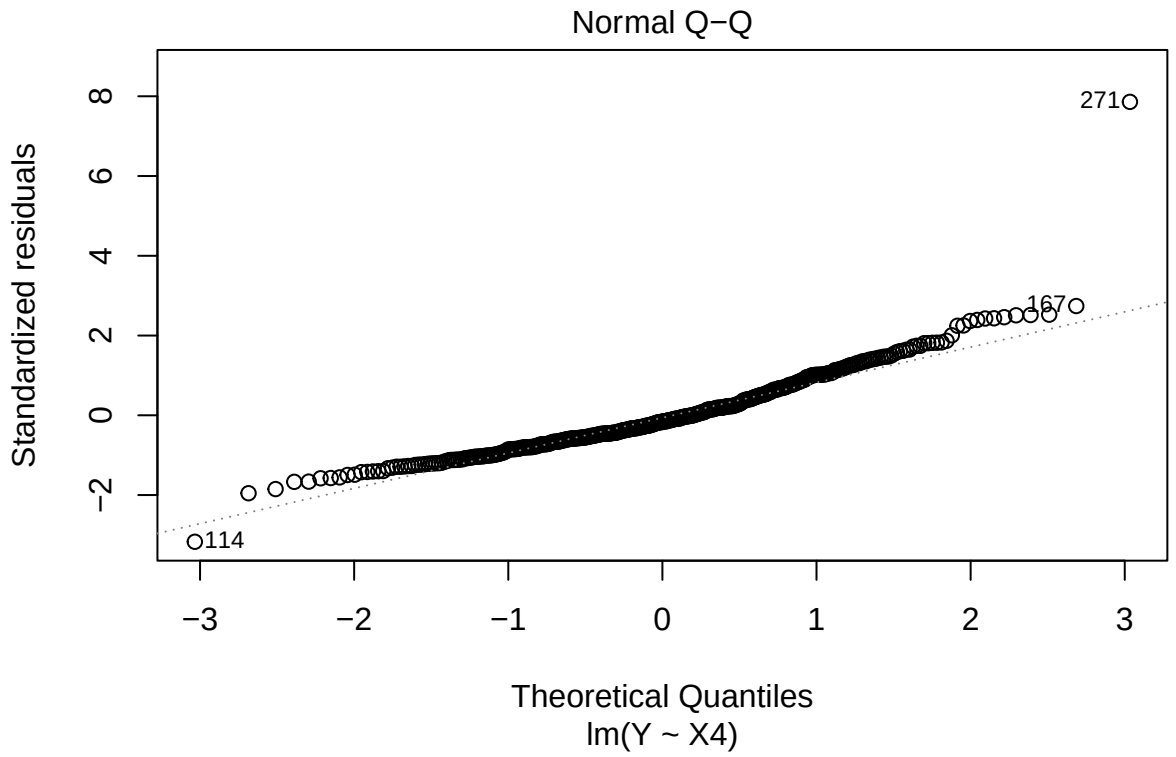

```
cor(Y,X4)
```

```
## [1] 0.5710049
```

```
hp4 <- lm(data = house , Y~X4)
summary(hp4)
```

```
##
## Call:
## lm(formula = Y ~ X4, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.407  -7.341  -1.788   5.984  87.681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.1811     0.9419   28.86   <2e-16 ***
## X4            2.6377     0.1868   14.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 412 degrees of freedom
## Multiple R-squared:  0.326,  Adjusted R-squared:  0.3244
## F-statistic: 199.3 on 1 and 412 DF,  p-value: < 2.2e-16
```

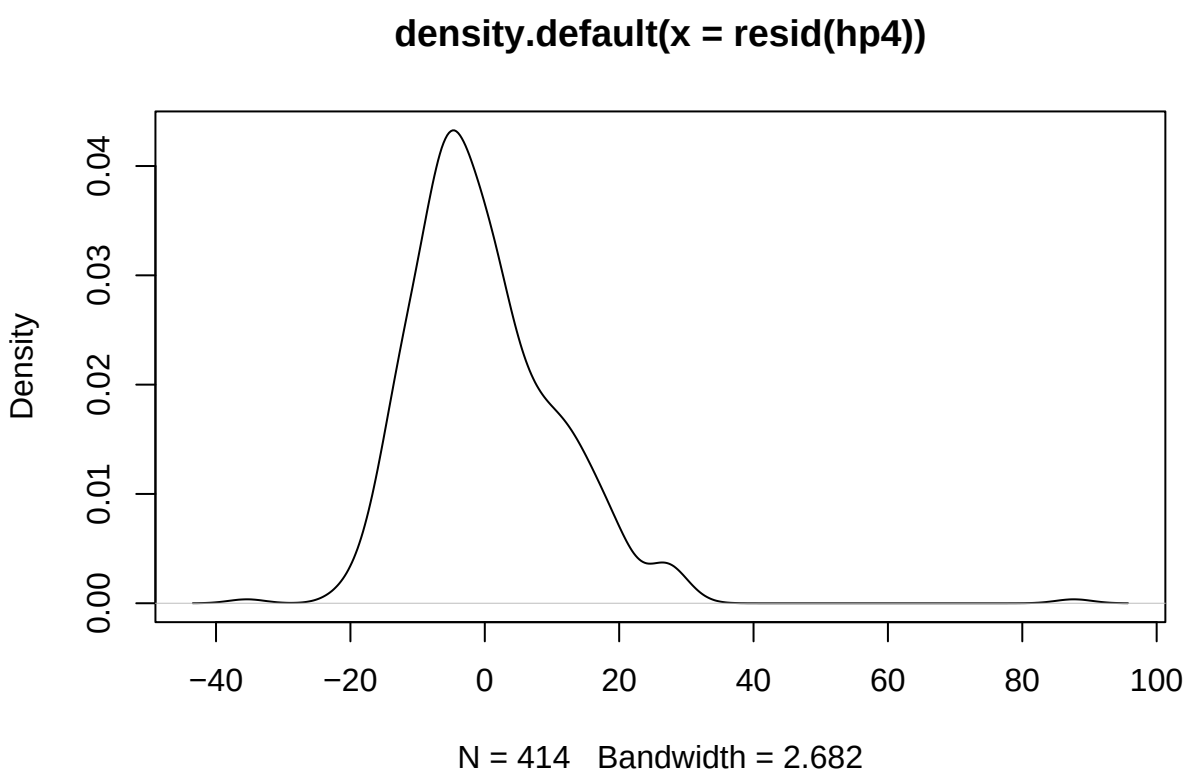
```
plot(hp4 , which = 2)
```




```
shapiro.test(resid(hp4))
```

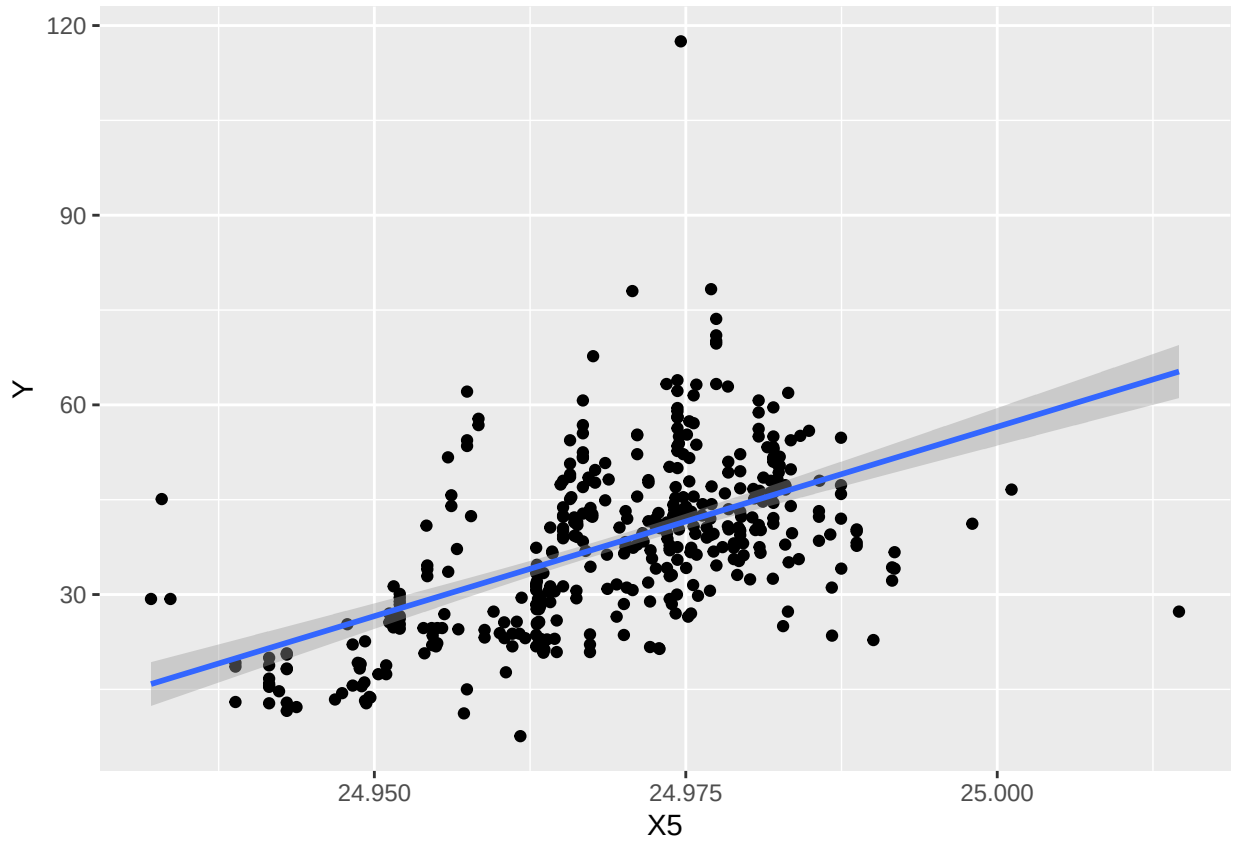
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp4)
## W = 0.91655, p-value = 2.291e-14
```

```
plot(density(resid(hp4)))
```




```
ggplot (house ,aes(X5, Y)) +
 geom_point() +
stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



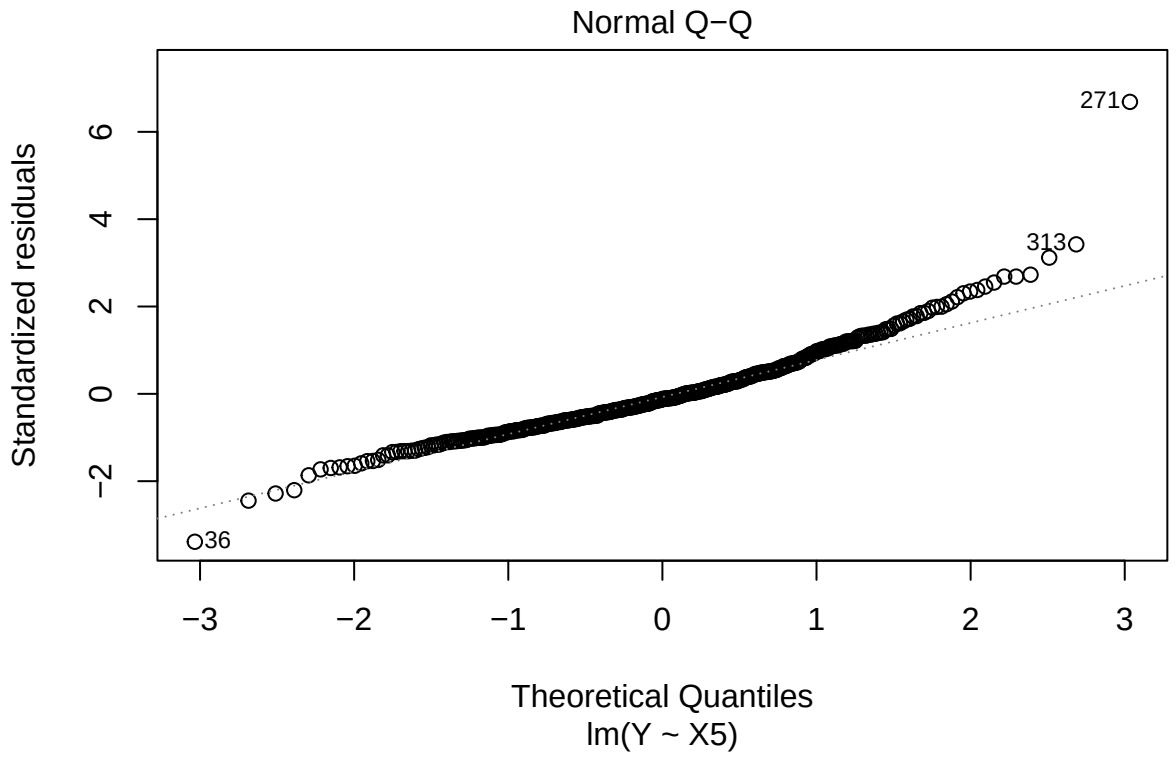

```
cor(Y,X5)
```

```
## [1] 0.5463067
```

```
hp5 <- lm(data =house , Y~X5)
summary(hp5)
```

```
##
## Call:
## lm(formula = Y ~ X5, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.969  -7.347  -1.392   5.685  76.184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14917.68    1129.66  -13.21   <2e-16 ***
## X5             598.97      45.24   13.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.41 on 412 degrees of freedom
## Multiple R-squared:  0.2985, Adjusted R-squared:  0.2967
## F-statistic: 175.3 on 1 and 412 DF,  p-value: < 2.2e-16
```

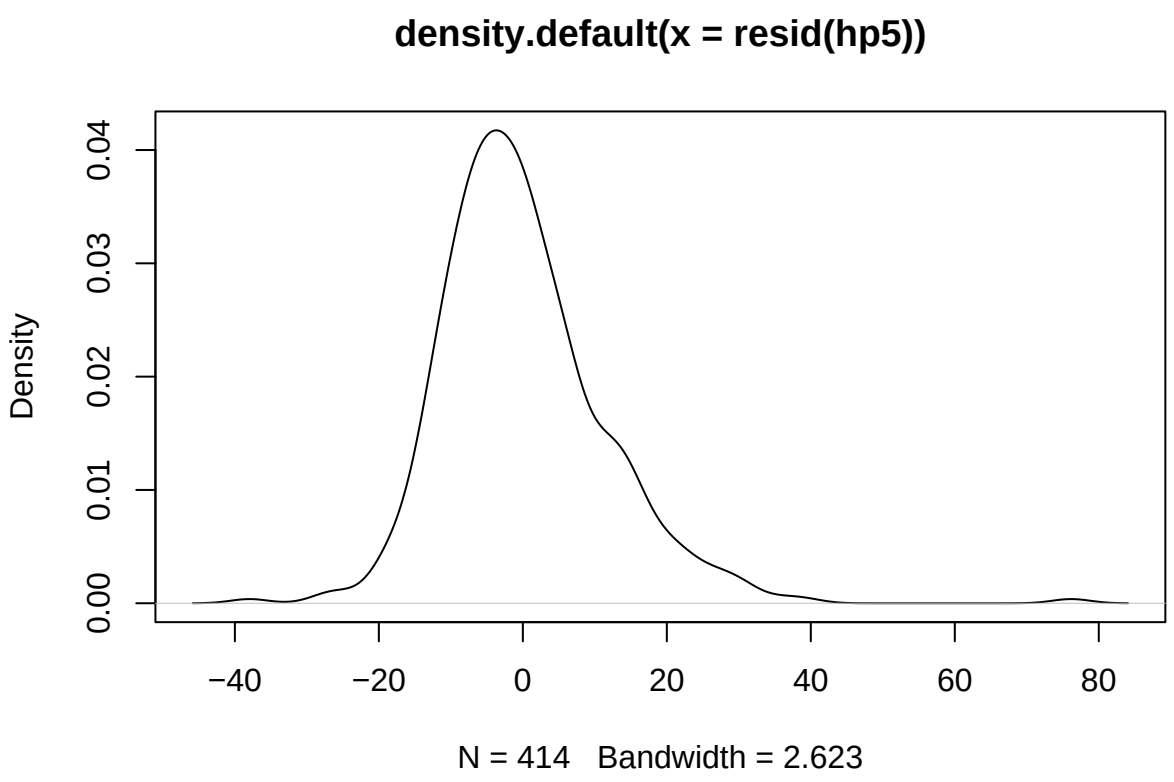
```r
plot(hp5 , which = 2)
```




```r
shapiro.test(resid(hp5))
```

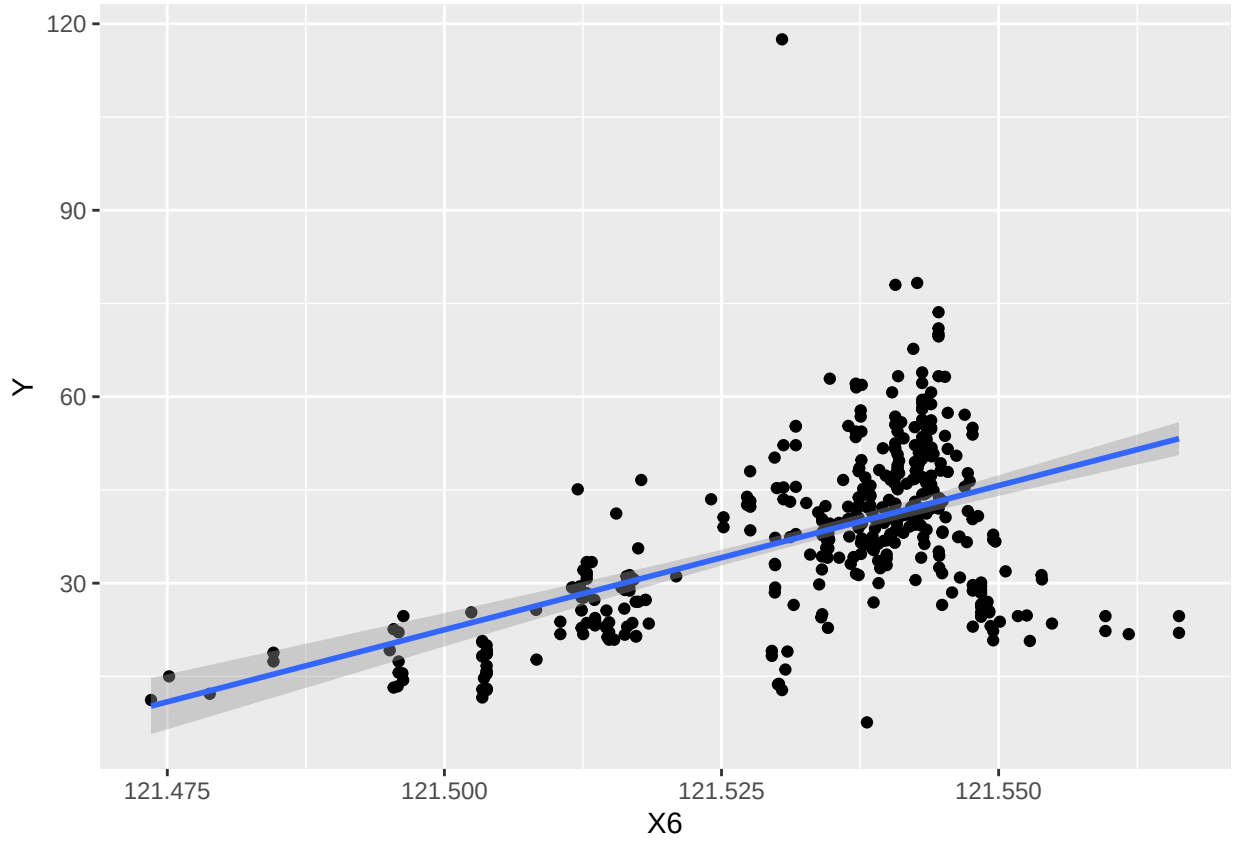
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp5)
## W = 0.9411, p-value = 9.531e-12
```

```r
plot(density(resid(hp5)))
```




```
ggplot (house ,aes(X6, Y)) +
 geom_point() +
stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



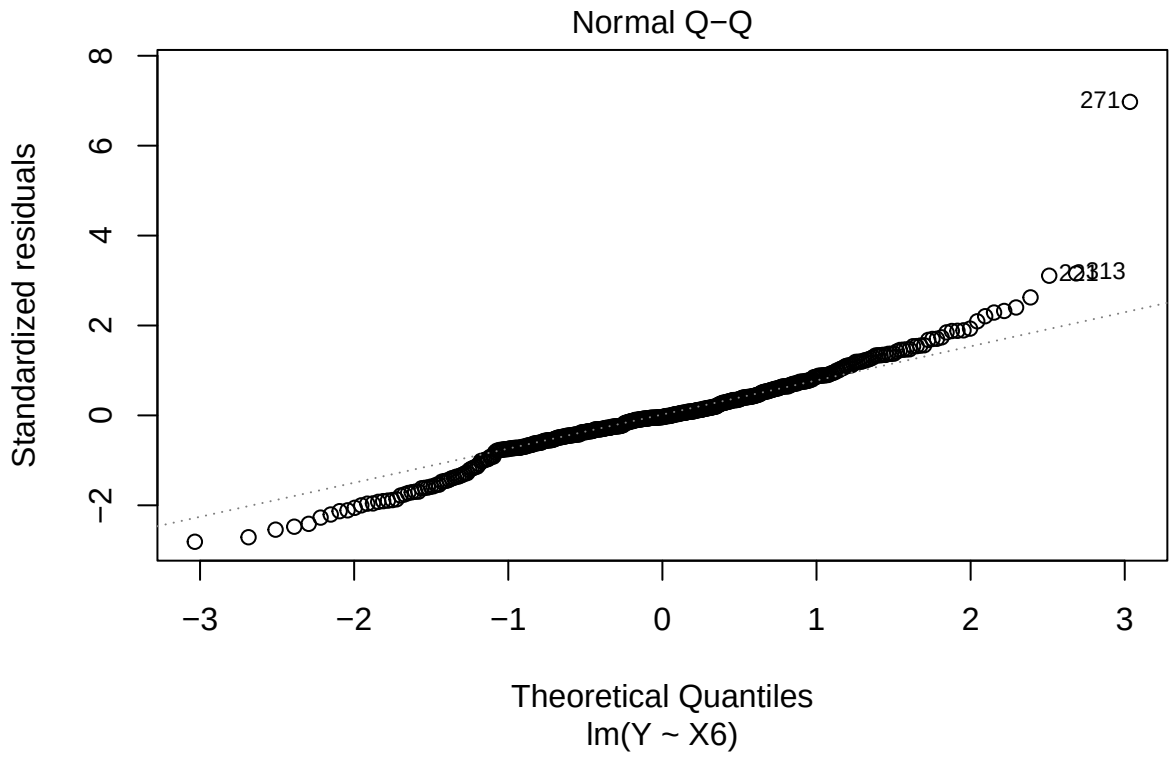

```
cor(Y,X6)
```

```
## [1] 0.5232865
```

```
hp6 <- lm(data = house , Y~X6)
summary(hp6)
```

```
##
## Call:
## lm(formula = Y ~ X6, data = house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.588  -5.693  -0.417   6.157  80.866
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56345.57    4523.60  -12.46   <2e-16 ***
## X6             463.93      37.22   12.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.61 on 412 degrees of freedom
## Multiple R-squared:  0.2738, Adjusted R-squared:  0.2721
## F-statistic: 155.4 on 1 and 412 DF,  p-value: < 2.2e-16
```

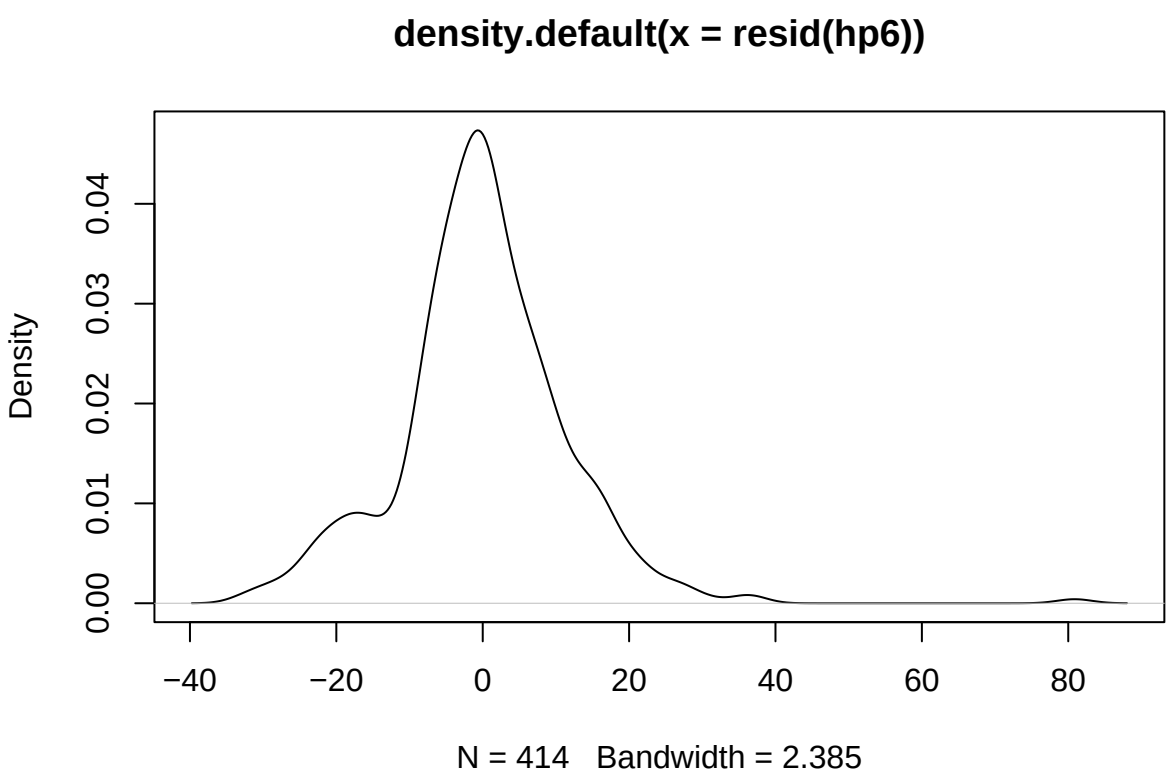
```
plot(hp6 , which = 2)
```




```
shapiro.test(resid(hp6))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(hp6)
## W = 0.9478, p-value = 6.586e-11
```

```
plot(density(resid(hp6)))
```


density.default(x = resid(hp6))
Density
0.00
0.01
0.02
0.03
0.04
−40
−20
0
20
40
60
80
N = 414 Bandwidth = 2.385