2

Bernat Chiva

12/5/2020

```r
library(car)
```

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

```r
library(HH)
```

```
## Loading required package: lattice
```

```
## Loading required package: grid
```

```
## Loading required package: latticeExtra
```

```
## Loading required package: multcomp
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
## 
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
## 
##     geyser
```

```
## Loading required package: gridExtra
```

```
## 
## Attaching package: 'HH'
```

```
## The following objects are masked from 'package:car':
## 
##     logit, vif
```

```r
library(readr)
REG8 <- read_delim("C:/Users/berna/OneDrive/Desktop/UPC/S1/5. Models
Lineals/datasets/REG8.csv",
    ";", escape_double = FALSE, locale = locale(decimal_mark = ","),
    trim_ws = TRUE)
```

```
## 
## -- Column specification --------------------------------------------------
------
## cols(
##   REG = col_double(),
##   X = col_double(),
##   Y = col_double()
## )
```

We first subset the dataset into eight different datasets:

```
first <- subset(REG8, REG8$REG == 1)
second <- subset(REG8, REG8$REG == 2)
third <- subset(REG8, REG8$REG == 3)
fourth <- subset(REG8, REG8$REG == 4)
fifth <- subset(REG8, REG8$REG == 5)
sixth <- subset(REG8, REG8$REG == 6)
seventh <- subset(REG8, REG8$REG == 7)
eight <- subset(REG8, REG8$REG == 8)
```

1. Modeling:

1.1. First dataset

We fit a simple linear model for each one of the datasets and perform an anova analysis as well:

```
mod1 <- lm(Y~X, data = first)
summary(mod1)
```

```
## 
## Call:
## lm(formula = Y ~ X, data = first)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.384 -3.289 -1.409  3.176  7.434 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054 
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
anova(mod1)

## Analysis of Variance Table
## 
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.2. Second dataset

```
mod2 <- lm(Y~X, data = second)
summary(mod2)

## 
## Call:
## lm(formula = Y ~ X, data = second)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.7659 -2.2250  0.4169  2.6096  7.6157 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.0000     1.0594   9.439 1.32e-08 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054 
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11

anova(mod2)

## Analysis of Variance Table
## 
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.3. Third dataset

```
mod3 <- lm(Y~X, data = third)
summary(mod3)
```

```
## 
## Call:
## lm(formula = Y ~ X, data = third)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.5485 -3.0263 -0.0003  3.0596  7.5382 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054 
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
anova(mod3)
```

```
## Analysis of Variance Table
## 
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.4. Fourth dataset

```
mod4 <- lm(Y~X, data = fourth)
summary(mod4)
```

```
## 
## Call:
## lm(formula = Y ~ X, data = fourth)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.2433 -2.8824 -0.8368  2.8820  8.2657 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```r
anova(mod4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.5. Fifth dataset

```r
mod5 <- lm(Y~X, data = fifth)
summary(mod5)
```

```
##
## Call:
## lm(formula = Y ~ X, data = fifth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5558 -1.8347  0.5321  1.0613 13.4747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```r
anova(mod5)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6. Sixth dataset

```
mod6 <- lm(Y~X, data = sixth)
summary(mod6)
```

```
## 
## Call:
## lm(formula = Y ~ X, data = sixth)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.3797 -2.8178  0.7244  3.3092  7.7345 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054 
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
anova(mod6)
```

```
## Analysis of Variance Table
## 
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)    
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.7. Seventh dataset

```
mod7 <- lm(Y~X, data = seventh)
summary(mod7)
```

```
## 
## Call:
## lm(formula = Y ~ X, data = seventh)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.590 -2.782  0.563  2.858  7.044 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0000     1.9909   5.023 7.55e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
anova(mod7)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.8. Eight dataset

```
mod8 <- lm(Y~X, data = eight)
summary(mod8)
```

```
##
## Call:
## lm(formula = Y ~ X, data = eight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0982 -1.9827 -0.8758  0.4341 15.7201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0000     1.6852   5.934 1.03e-05 ***
## X             2.0000     0.1441  13.874 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4 on 19 degrees of freedom
## Multiple R-squared:  0.9102, Adjusted R-squared:  0.9054
## F-statistic: 192.5 on 1 and 19 DF,  p-value: 2.153e-11
```

```
anova(mod8)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X          1   3080    3080   192.5 2.153e-11 ***
## Residuals 19    304      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compute the 8 regression lines and compare them. Answer the following questions:

a) Do they have the same regression coefficients?

All of the eight models we have fitted with different data have the same regression coefficients. An intercept of 10 and a beta value of 2 for the X variable. All the coefficients are significantly different from 0, meaning that they are important to the model.

b) Do they have the same coefficient of determination?

Yes, we can see how all the eight models have an R squared of 0.9102. Meaning that all the models have the same efficiency, explaining a 91 per cent of the variability of our dependent variable Y.

c) Do they give the same conclusion in the ANOVA test?

The values are all the same upon the eight anova tables, the sum of squares of the coefficients and residuals as well as the significance of the model.

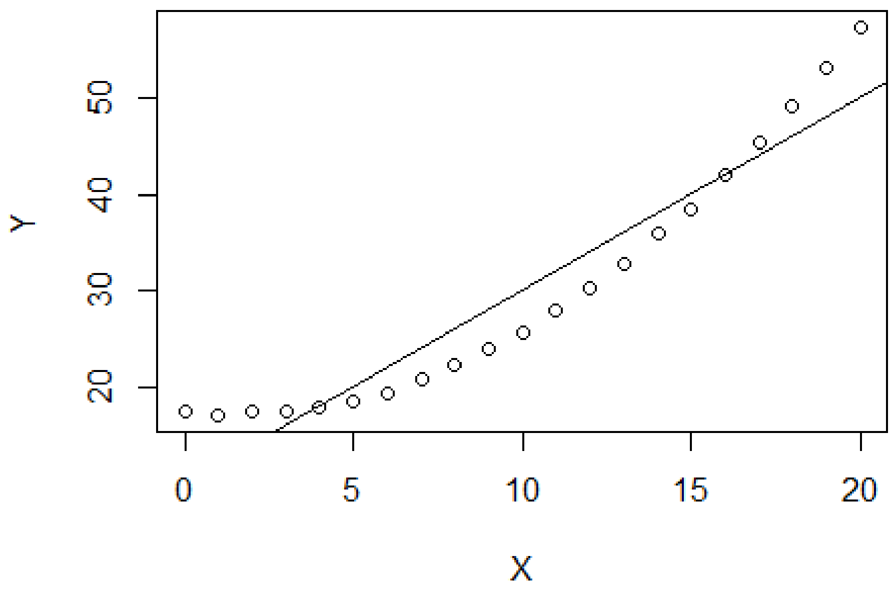
d) Are all the lines reasonable? Which problems do you observe?

In order to respond to this question we will plot the observations and the regression line to each model and comment it.

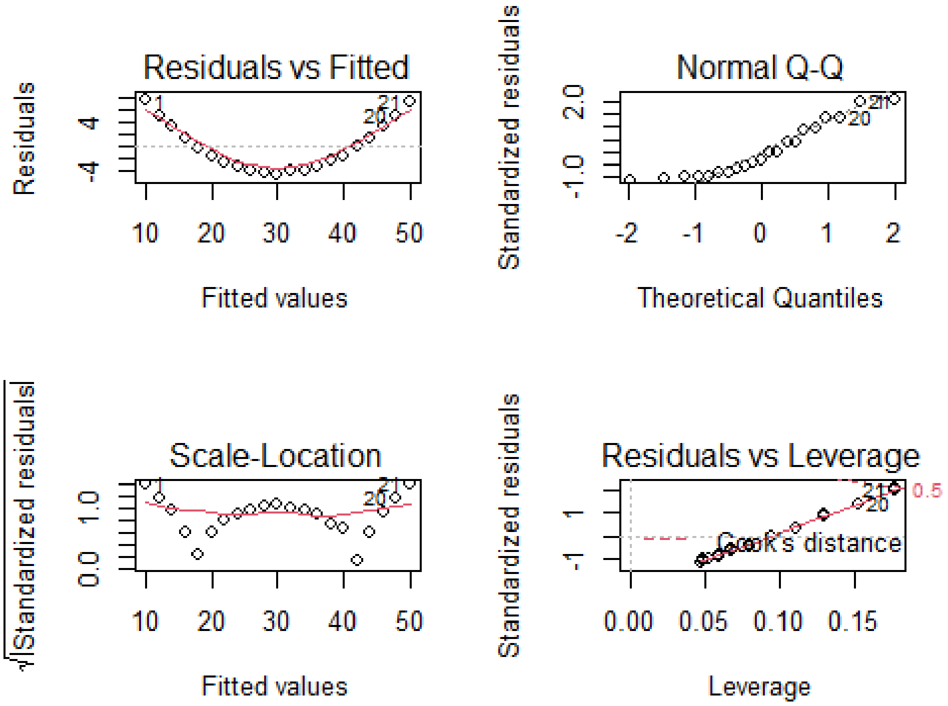
2. Interpretation:

2.1. First dataset

```
with(first, plot(X,Y))
abline(lm(Y~X, data = first))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod1, ask=F)
```

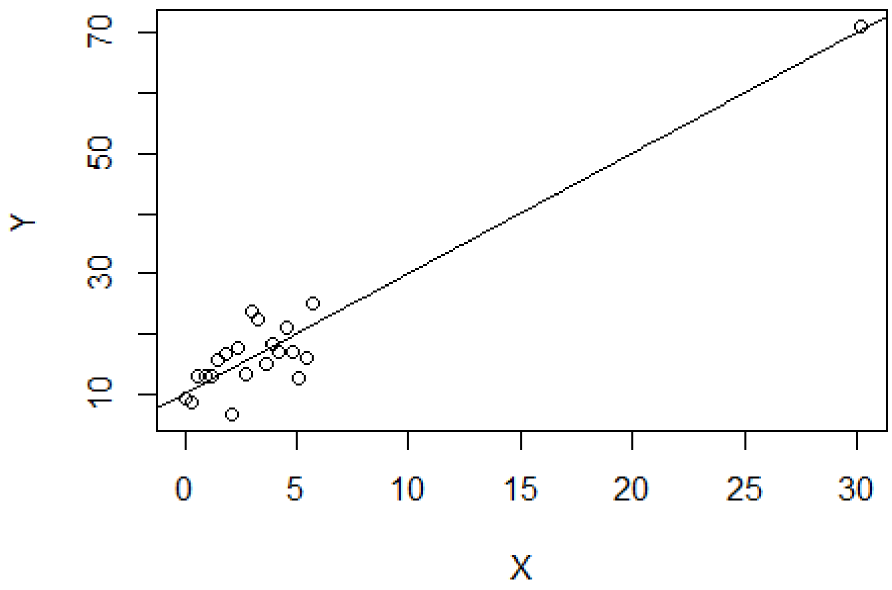


```
par(oldpar)
```

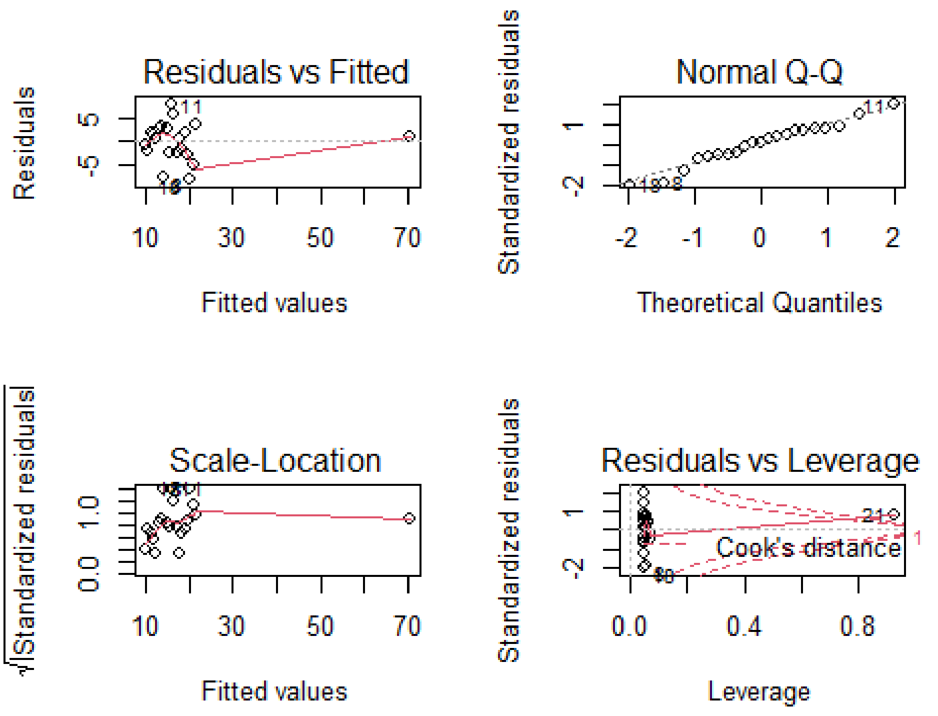
In the first model the issue we can observe in the plot is the non-linearity of the observations, they draw a parabola. So a linear model might not be the best fir for these variables. If we check the residuals the shapes are weird and non-linear, although we could assume homoscedasticity and normality. If we adjust the squared root might work better.

2.2. Second dataset

```
with(second, plot(X,Y))
abline(lm(Y~X, data = second))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod2, ask=F)
```

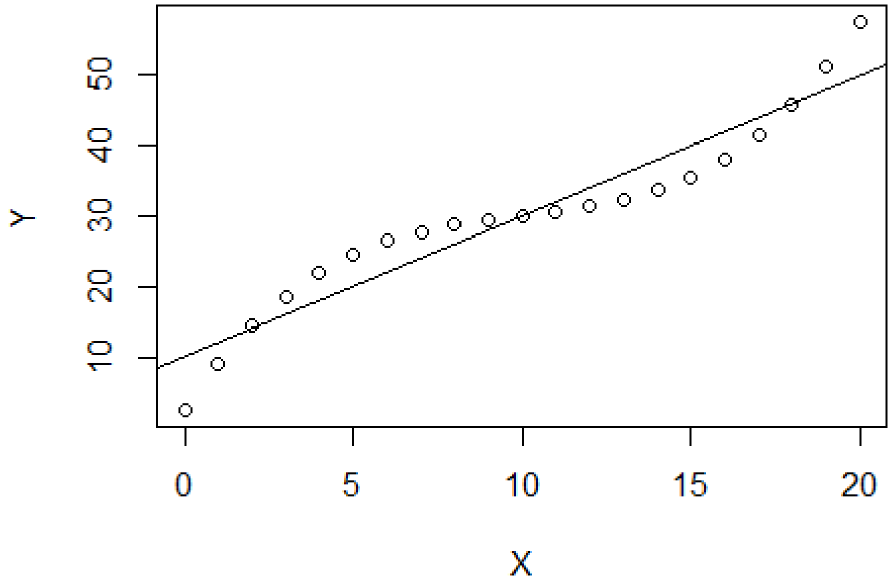


```
par(oldpar)
```

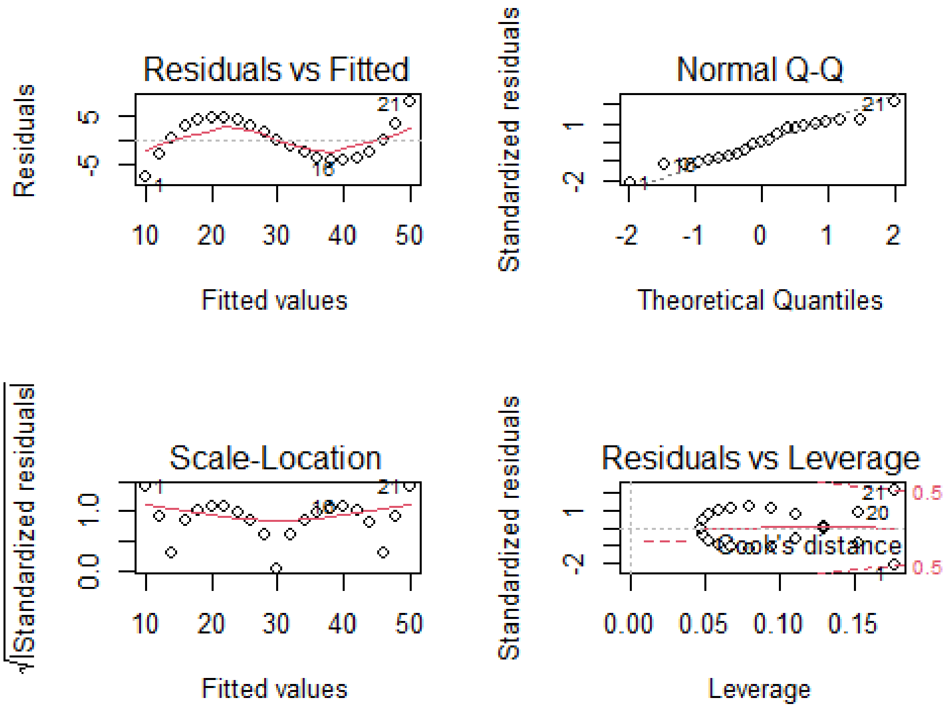
In the second model we can see an observation far from the others, this observations, though it has a large leverage, its residual is small, as it is close to the regression line drawn from the model.

2.3. Third dataset

```
with(third, plot(X,Y))
abline(lm(Y~X, data = third))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod3, ask=F)
```

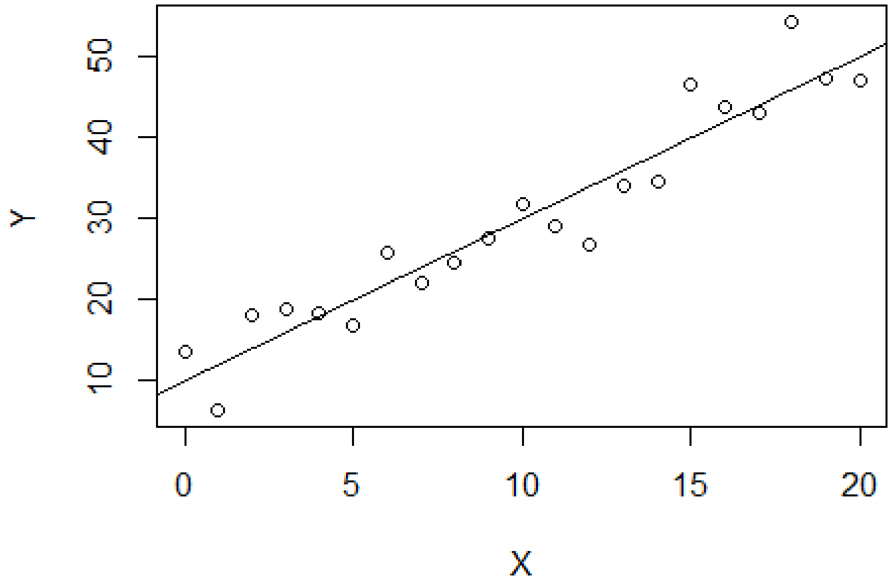


```
par(oldpar)
```

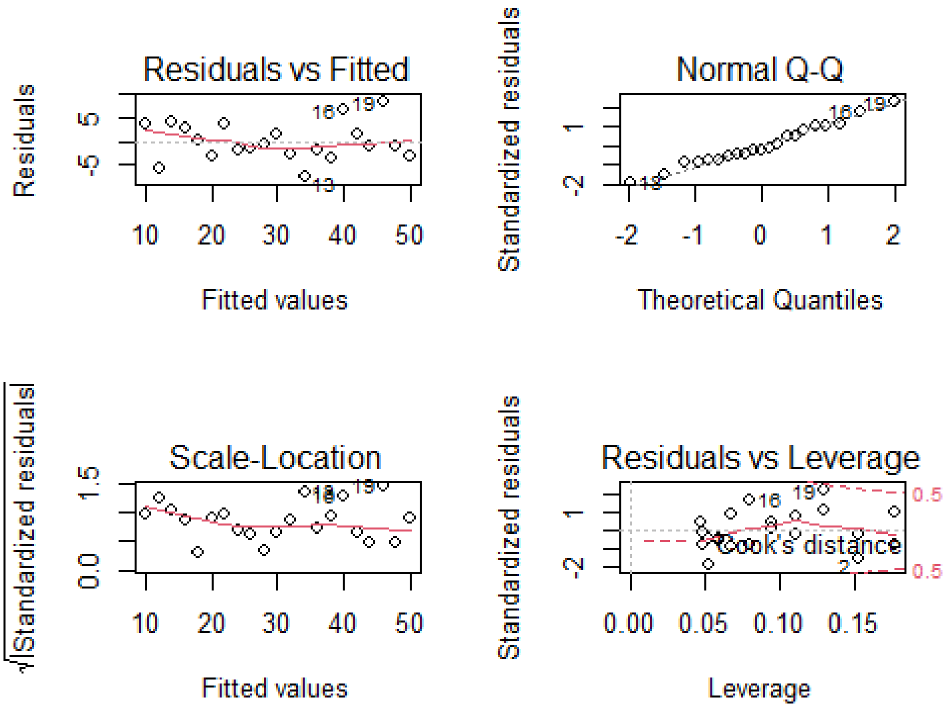
In the third model, the issue is similar to the first one, as the observations do not follow a straight line. In this case they follow a S shape, so again, plotting a linear model is not the best solution although the residuals are small. We could assume normality with some transformation of X.

2.4. Fourth dataset

```
with(fourth, plot(X,Y))
abline(lm(Y~X, data = fourth))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod4, ask=F)
```

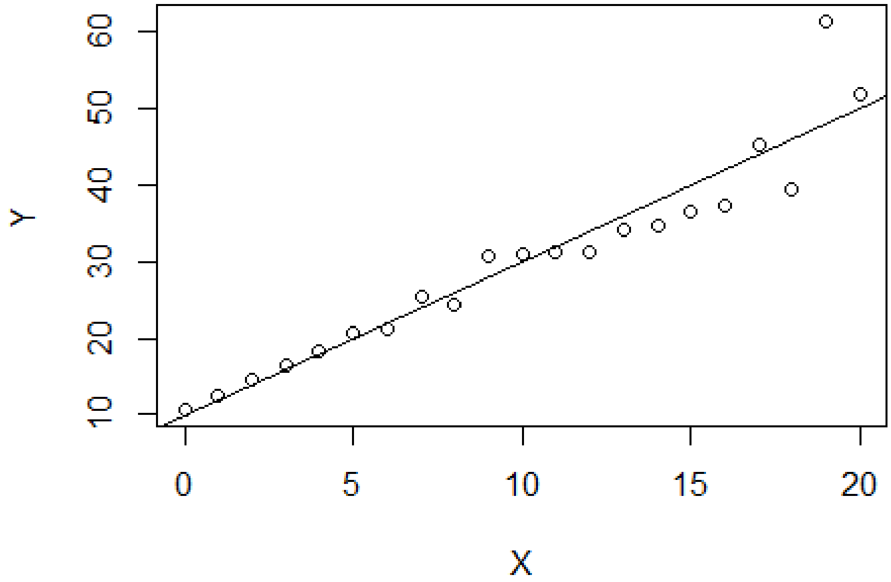


```
par(oldpar)
```

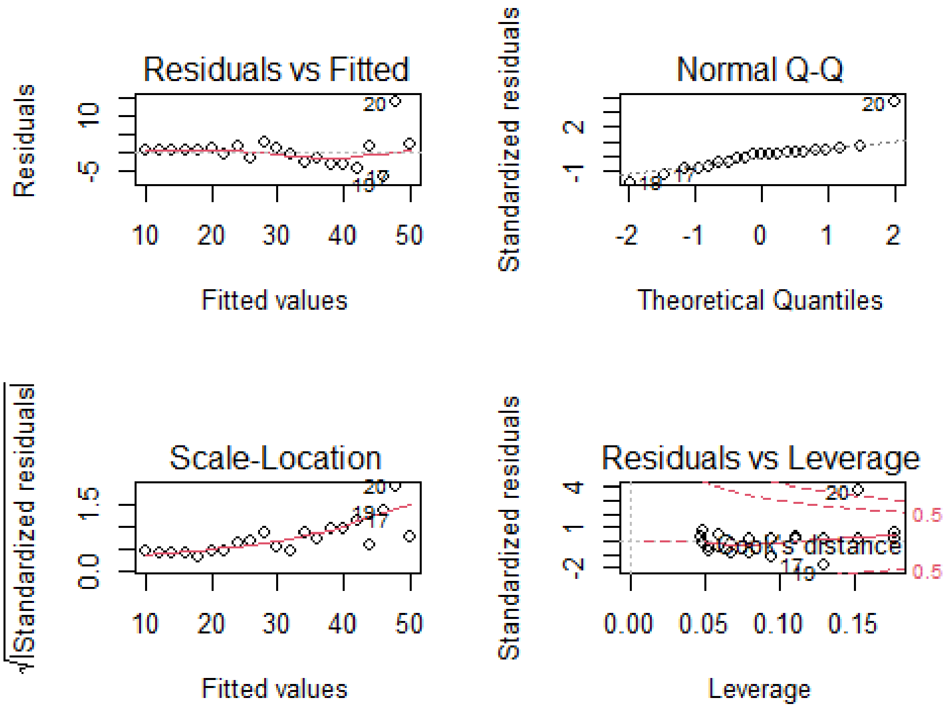
In the fourth model the line is very reasonable, data seem to be linear with no outliers, observations with high residual and leverage. In this case our model would be the best fit. Normality should be checked in order to verify if the model is good.

2.5. Fifth dataset

```
with(fifth, plot(X,Y))
abline(lm(Y~X, data = fifth))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod5, ask=F)
```

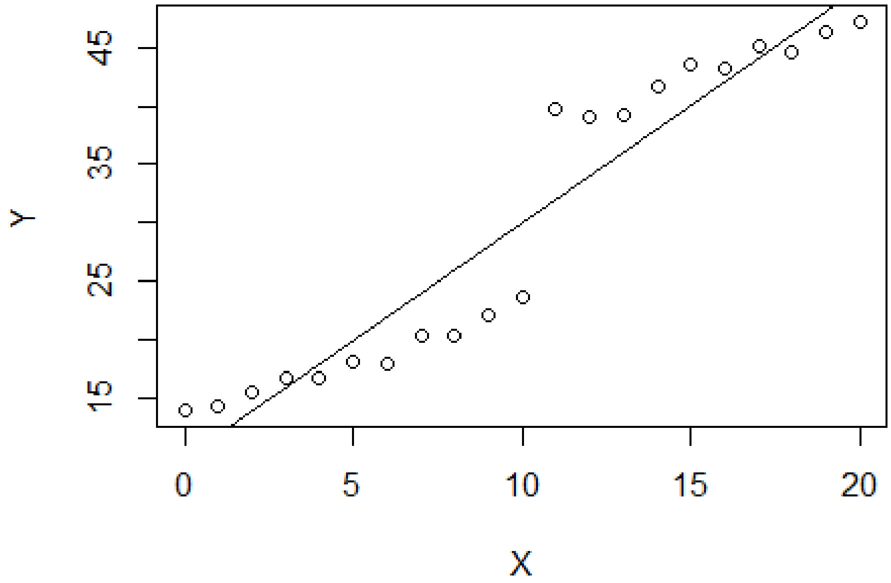


```
par(oldpar)
```

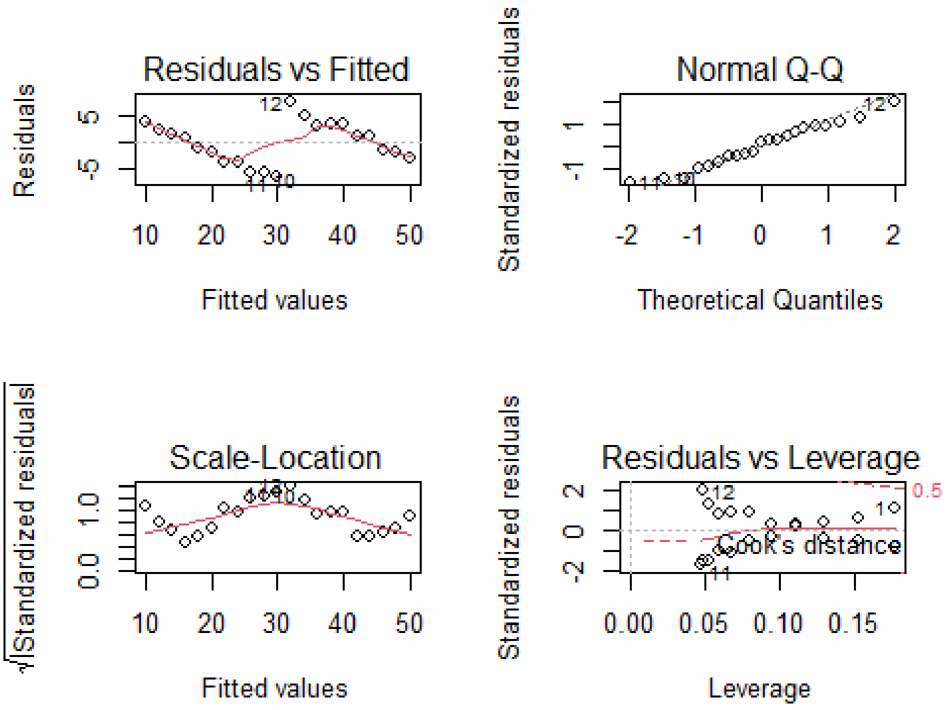
In the fifth model the data is linear as well, but we observe how the linearity decays as X and Y increases, meaning that the variability increases as well and we might not have homoscedasticity. If we check the variance of the residuals we can see how we cannot assume homoscedasticity and as well we have influential observations that makes our model worst, the observation 20 has a high value of Cook’s distance.

2.6. Sixth dataset

```
with(sixth, plot(X,Y))
abline(lm(Y~X, data = sixth))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod6, ask=F)
```

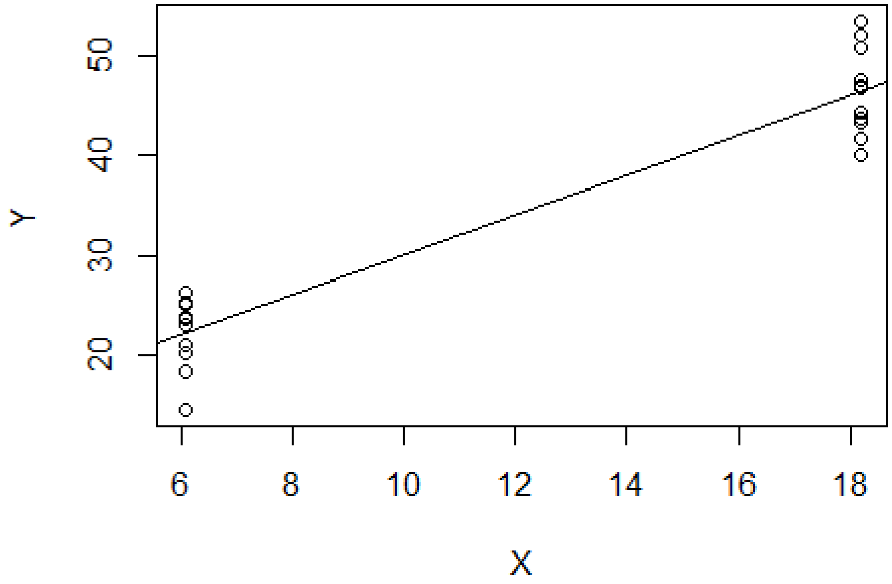


```
par(oldpar)
```

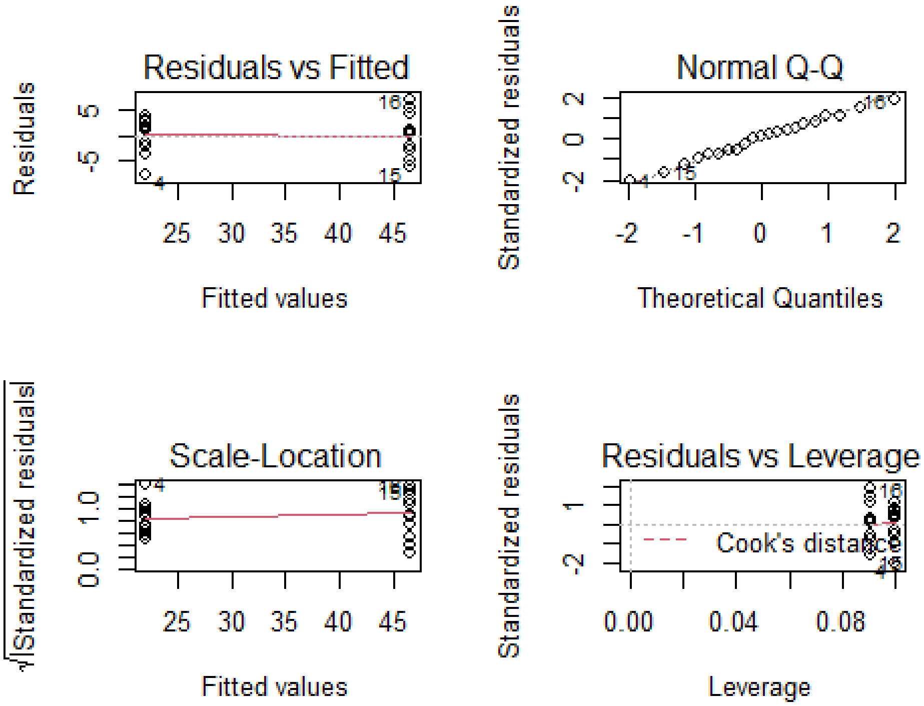
In the sixth model we can see the peculiarity that the dataset seem to come from two different variables, as we see a big change from the values of the first half of the observations and the second half. In this case we should check if dividing the dataset makes sense and perform two different models.

2.7. Seventh dataset

```
with(seventh, plot(X,Y))
abline(lm(Y~X, data = seventh))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod7, ask=F)
```

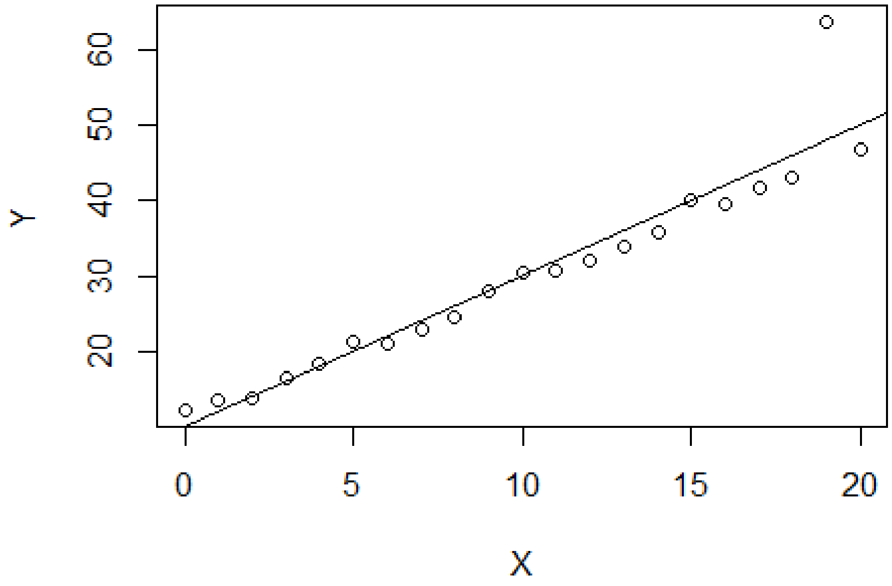


```
par(oldpar)
```

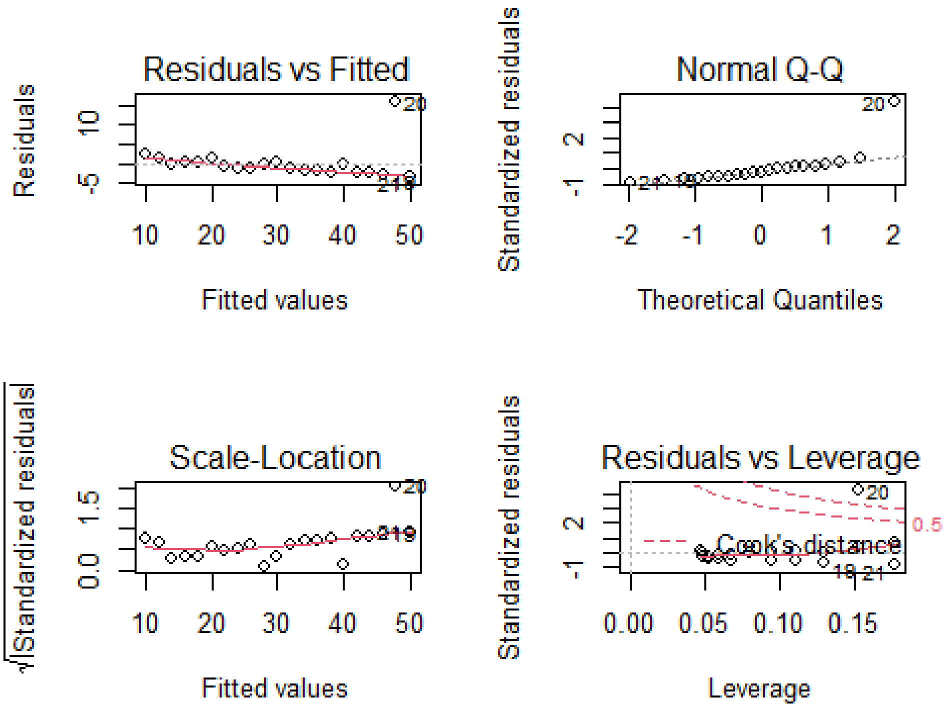
In the seventh model, the variable X only takes two different values throughout all the dataset. This makes us think that X might be a factor with two levels, then we should change the category of the variable as the analysis would be different.

2.8. Eight dataset

```
with(eight, plot(X,Y))
abline(lm(Y~X, data = eight))
```



```
oldpar <- par( mfrow=c(2,2))
plot(mod8, ask=F)
```



```
par(oldpar)
```

In the eight model, the regression line seems to fit very well, but we can appreciate an observation that has a larger residual than the other observations, making our model worst, it is an outlier as we can check by its Cook’s distance. In this case would make sense to get rid of that observation and fit the model again, our coefficient of determination would be greater.

Conclusions: Once we have computed and analyzed the eight models, although all of them have the same coefficients and coefficient of determination, very high, meaning that the model explains a big part of the variability of the response variable, the observations are different, and the model could be improved.

As well these exercises allows us to understand that performing ana analysis and getting a high R squared is not enough in order to understand the relation of two variables, more in depth analysis is needed, such as the analysis of residuals to verify our models assumptions.