ADEI Lab 2 (Josep)

Josep Franquet

2025-03-06

Load Data

```
# Clear plots
if(!is.null(dev.list())) dev.off()
```

```
## null device
##           1
```

```
# Clean workspace
rm(list=ls())

load("Anscombe73raw.RData")

ls()
```

```
## [1] "anscombe"     "last.warning"
```

```
anscombe
```

```
##    XA    YA XB   YB XC    YC XD    YD
## 1  10  8.04 10 9.14 10  7.46  8  6.58
## 2   8  6.95  8 8.14  8  6.77  8  5.76
## 3  13  7.58 13 8.74 13 12.74  8  7.71
## 4   9  8.81  9 8.77  9  7.11  8  8.84
## 5  11  8.33 11 9.26 11  7.81  8  8.47
## 6  14  9.96 14 8.10 14  8.84  8  7.04
## 7   6  7.24  6 6.13  6  6.08  8  5.25
## 8   4  4.26  4 3.10  4  5.39 19 12.50
## 9  12 10.84 12 9.13 12  8.15  8  5.56
## 10  7  4.82  7 7.26  7  6.42  8  7.91
## 11  5  5.68  5 4.74  5  5.73  8  6.89
```

```
attach(anscombe) #Thus, we will not have to write anscombe$var when accessing a variable
summary(anscombe) #Summary of the whole data (at a variable level)
```

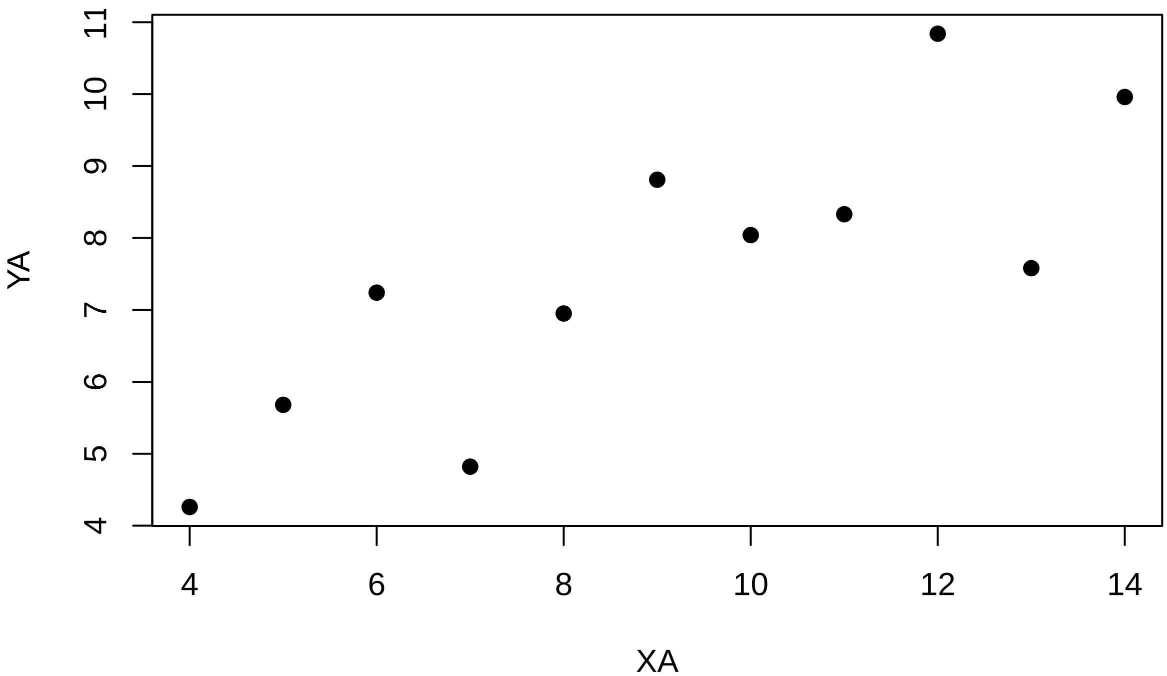
```
##        XA              YA               XB             YB              XC
##  Min.   : 4.0   Min.   : 4.260   Min.   : 4.0   Min.   :3.100   Min.   : 4.0
##  1st Qu.: 6.5   1st Qu.: 6.315   1st Qu.: 6.5   1st Qu.:6.695   1st Qu.: 6.5
##  Median : 9.0   Median : 7.580   Median : 9.0   Median :8.140   Median : 9.0
##  Mean   : 9.0   Mean   : 7.501   Mean   : 9.0   Mean   :7.501   Mean   : 9.0
##  3rd Qu.:11.5   3rd Qu.: 8.570   3rd Qu.:11.5   3rd Qu.:8.950   3rd Qu.:11.5
##  Max.   :14.0   Max.   :10.840   Max.   :14.0   Max.   :9.260   Max.   :14.0
##        YC              XD           YD
##  Min.   : 5.39   Min.   : 8   Min.   : 5.250
```

```
##   1st Qu.: 6.25   1st Qu.: 8   1st Qu.: 6.170
##   Median : 7.11   Median : 8   Median : 7.040
##   Mean   : 7.50   Mean   : 9   Mean   : 7.501
##   3rd Qu.: 7.98   3rd Qu.: 8   3rd Qu.: 8.190
##   Max.   :12.74   Max.   :19   Max.   :12.500
```

SET A

- Do a scatterplot between XA and YA

```r
plot(YA~XA, pch=19)
```



- Run a linear model where you regress YA ~ XA

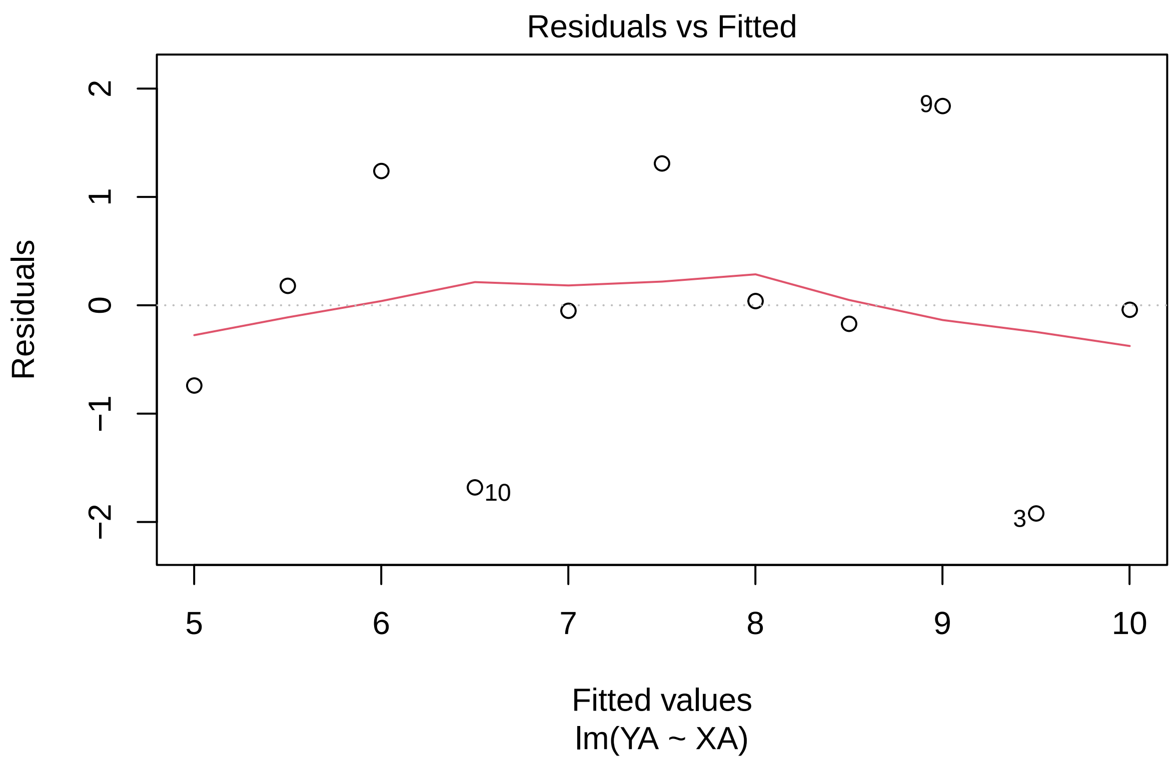
```r
ma <- lm(YA~XA)
summary(ma)
```

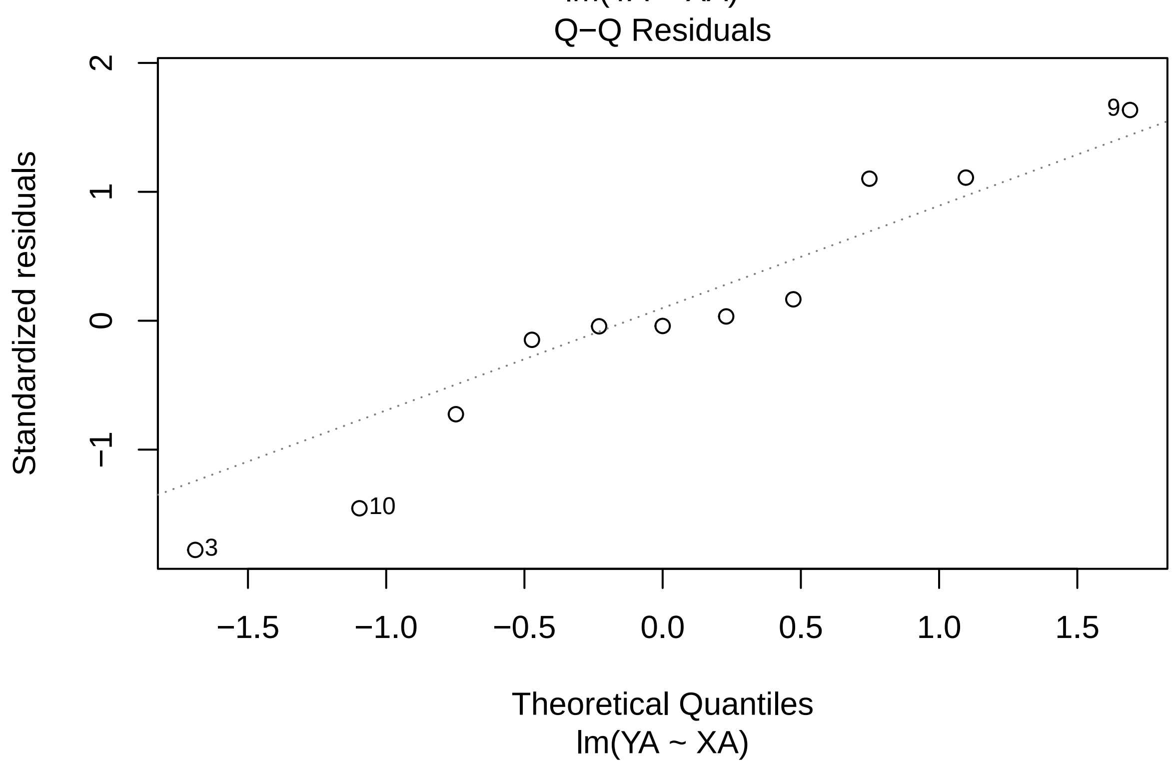
```
##
## Call:
## lm(formula = YA ~ XA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## XA            0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

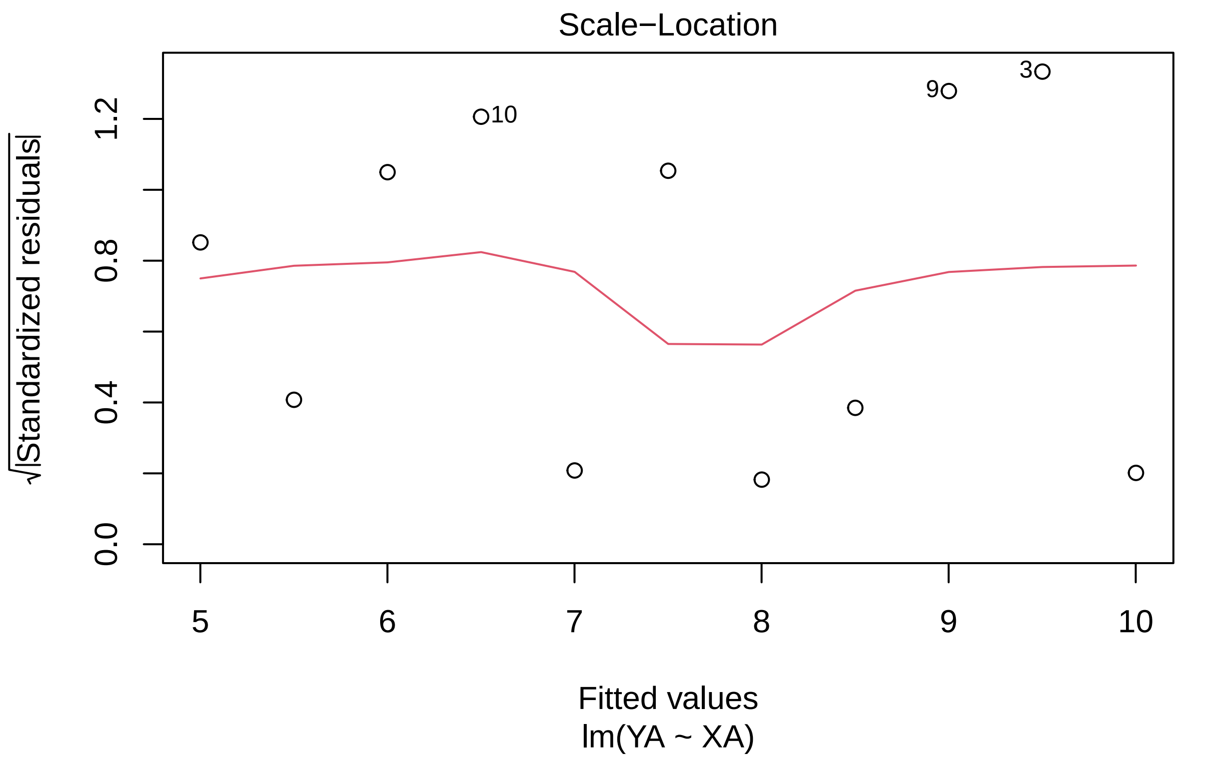
```
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

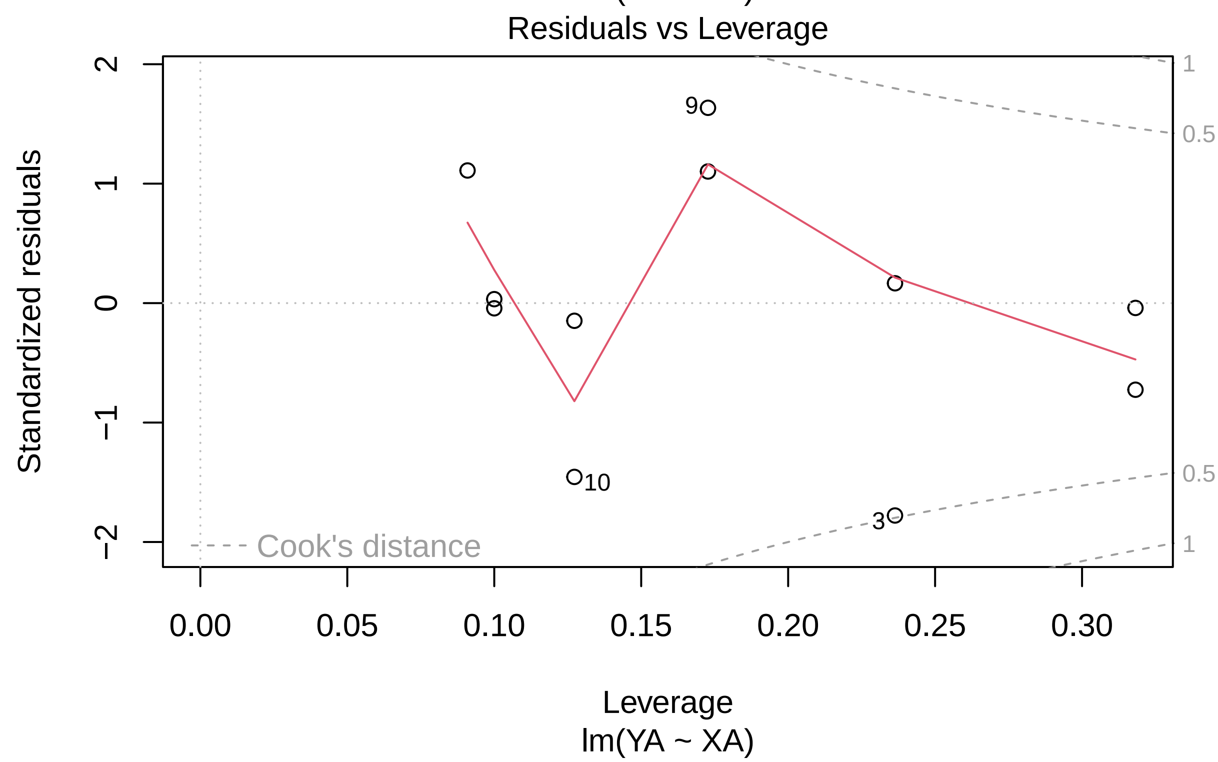
- Validate the basic hypothesis of the linear model graphically.

```r
plot(ma)
```









Basic hypothesis:

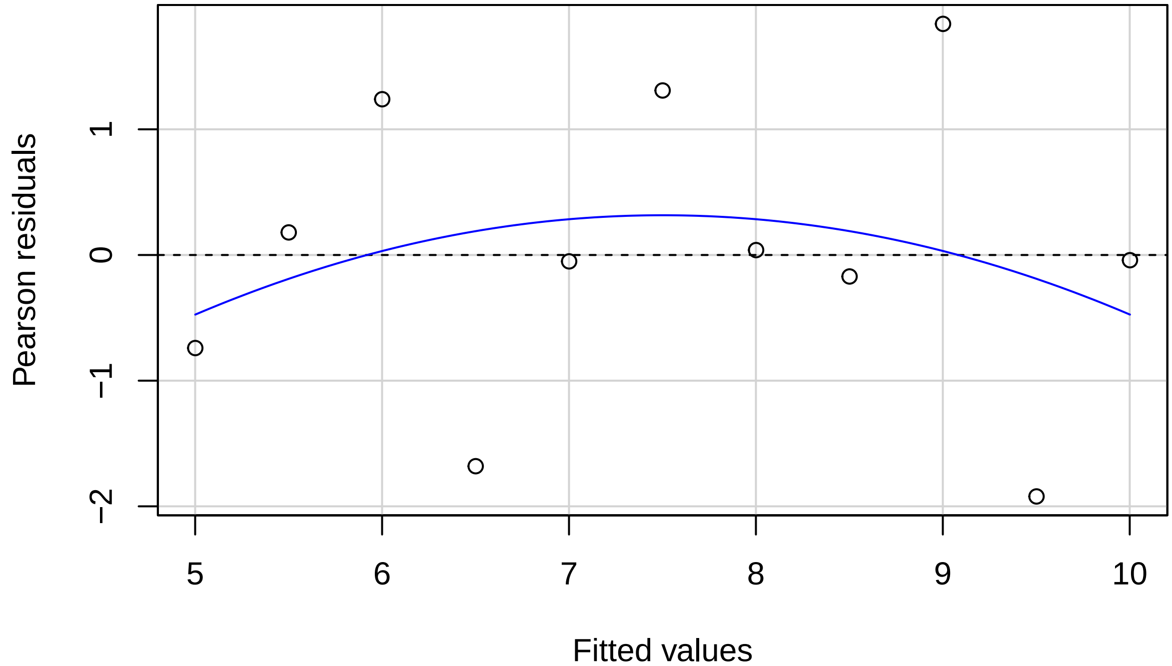
- Homoskedasticity: bptest
- Autocorrelation: acf (plot with significance peaks or not) and dwtest (durbin watson test)
- Normality: Shapiro wilk test

Other options to check linearity:

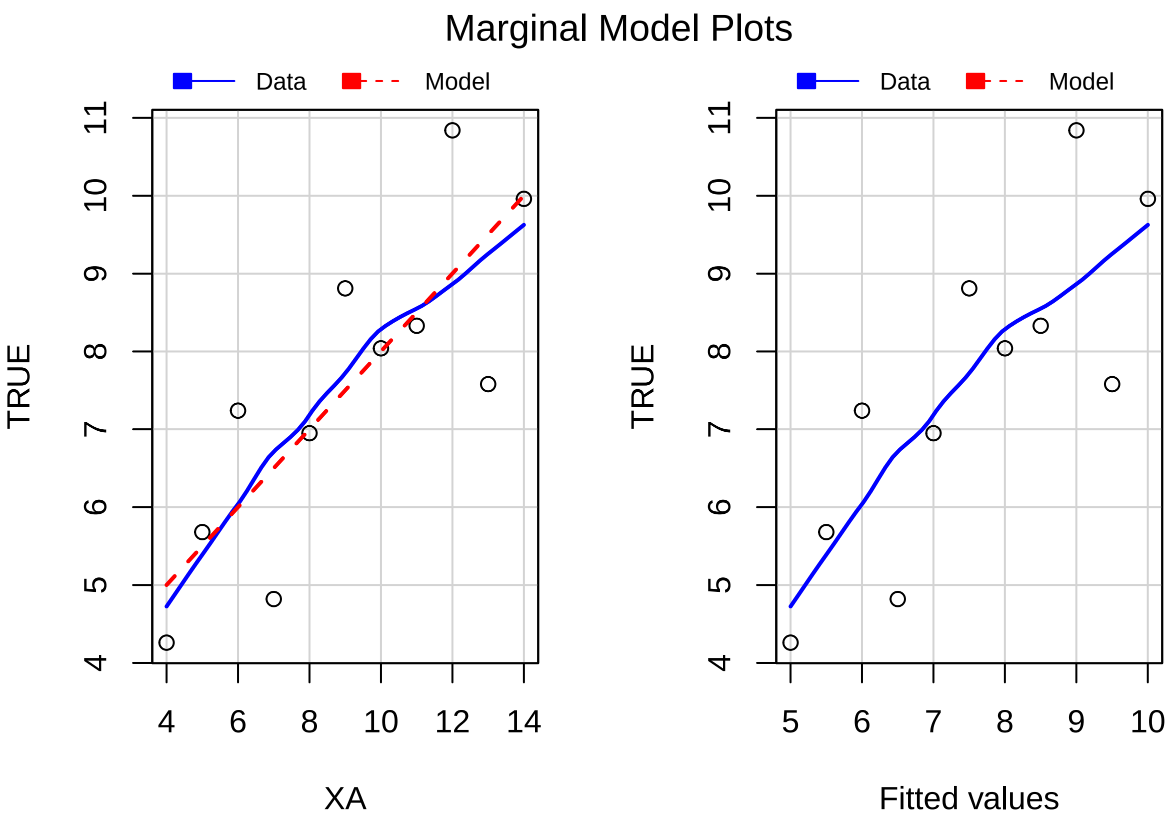
```r
library(car)
```

```
## S'està carregant el paquet requerit: carData
```

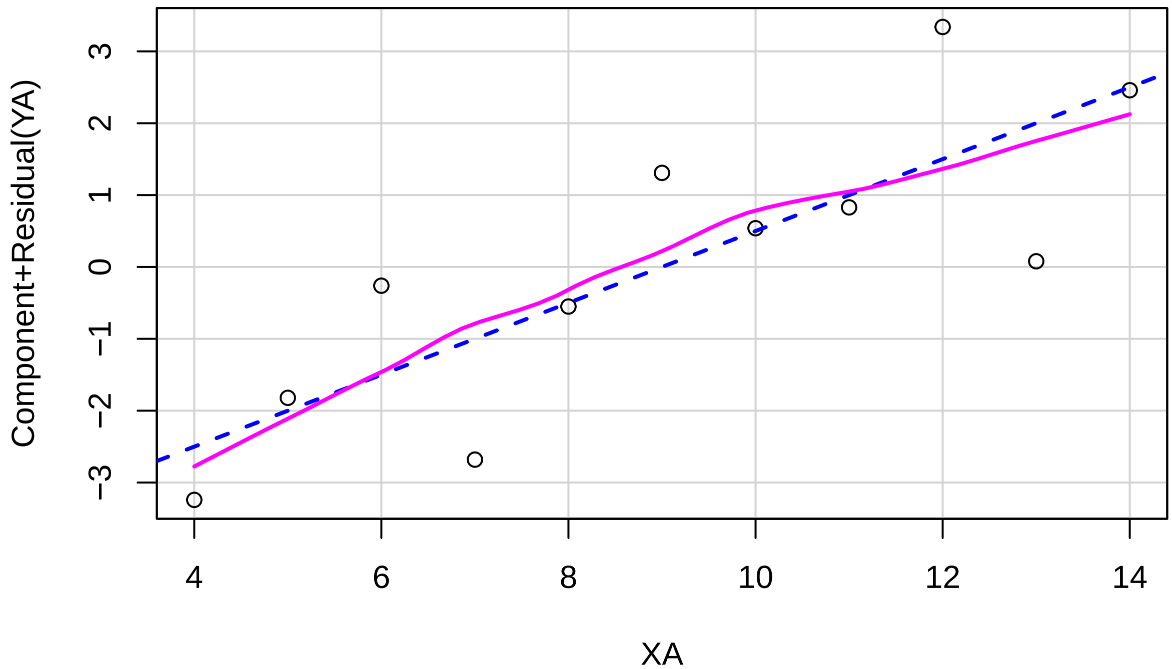
```r
residualPlot(ma)
```




```r
marginalModelPlots(ma)
```



```r
crPlot(ma,"XA")
```



Calculate the AIC and BIC value from the linear model we just estimated:

```r
AIC(ma) #AIC
```

```
## [1] 39.68137
```

```r
AIC(ma, k = log(nrow(anscombe))) #BIC
```

```
## [1] 40.87506
```

Let's run the stepwise regression algorithm to determine the best linear model:

```r
ma_0 <- lm(YA ~ 1, data=anscombe) # Null model
step(ma_0, ~XA, direction="forward", data=anscombe)
```

```
## Start:  AIC=16.55
## YA ~ 1
##
##        Df Sum of Sq    RSS     AIC
## + XA    1     27.51 13.763  6.4647
## <none>              41.273 16.5454
##
## Step:  AIC=6.46
## YA ~ XA

##
## Call:
## lm(formula = YA ~ XA, data = anscombe)
##
## Coefficients:
## (Intercept)           XA
##      3.0001       0.5001
```

Backward direction:

```r
step(ma, direction = "backward", data=anscombe)
```

```
## Start:  AIC=6.46
## YA ~ XA
##
##        Df Sum of Sq    RSS     AIC
## <none>              13.763  6.4647
## - XA    1     27.51 41.273 16.5454
```

```
##
## Call:
## lm(formula = YA ~ XA)
##
## Coefficients:
## (Intercept)           XA
##      3.0001       0.5001
```

SET B:

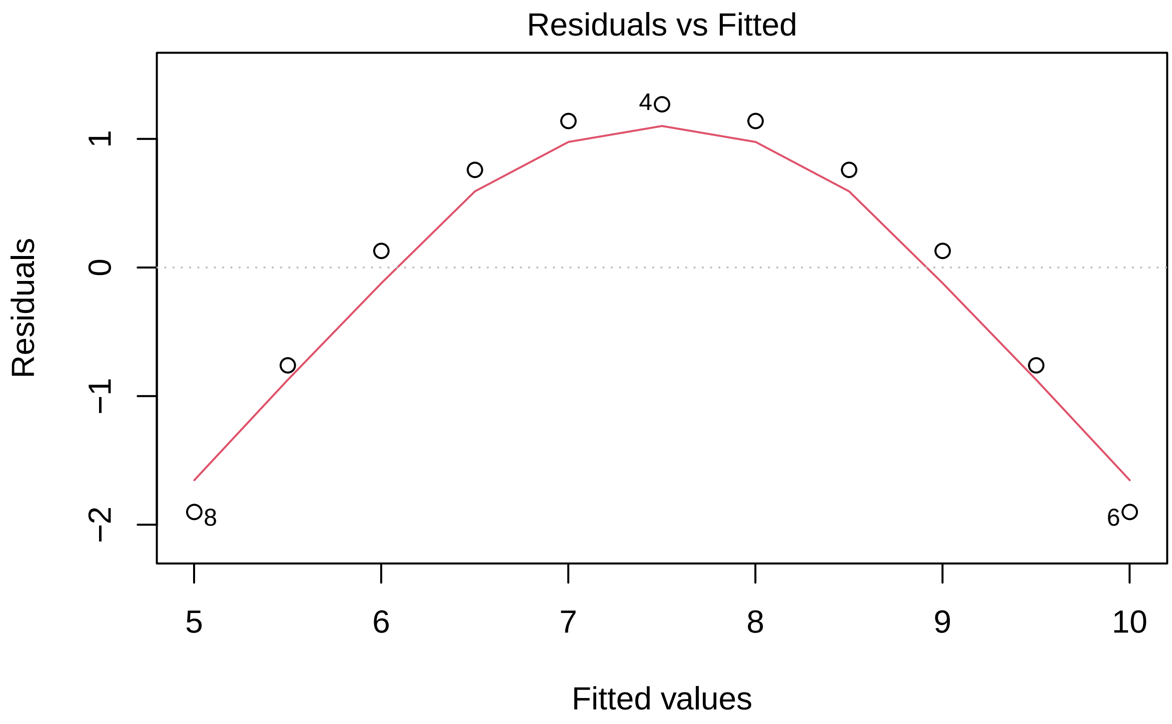
- Estimate a new model by regressing YB ~ XB

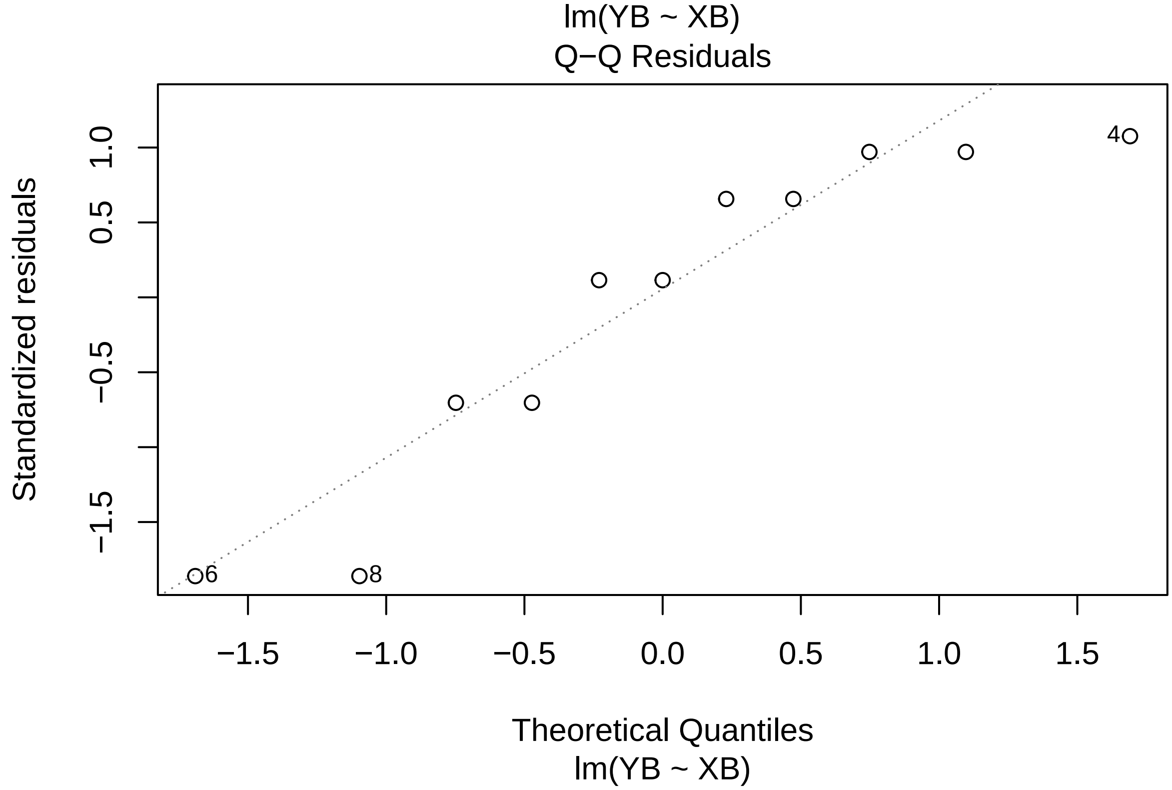
```r
mb <- lm(YB~XB, data = anscombe)
summary(mb)
```

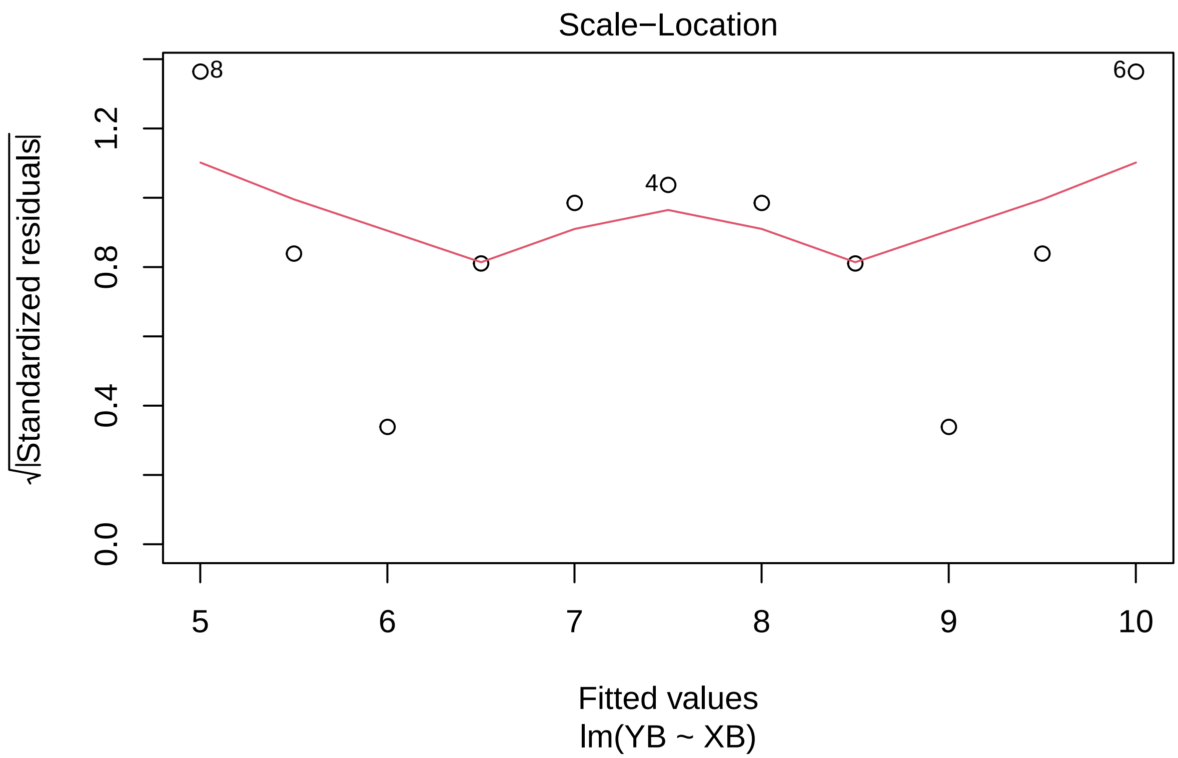
```
##
## Call:
## lm(formula = YB ~ XB, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## XB             0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

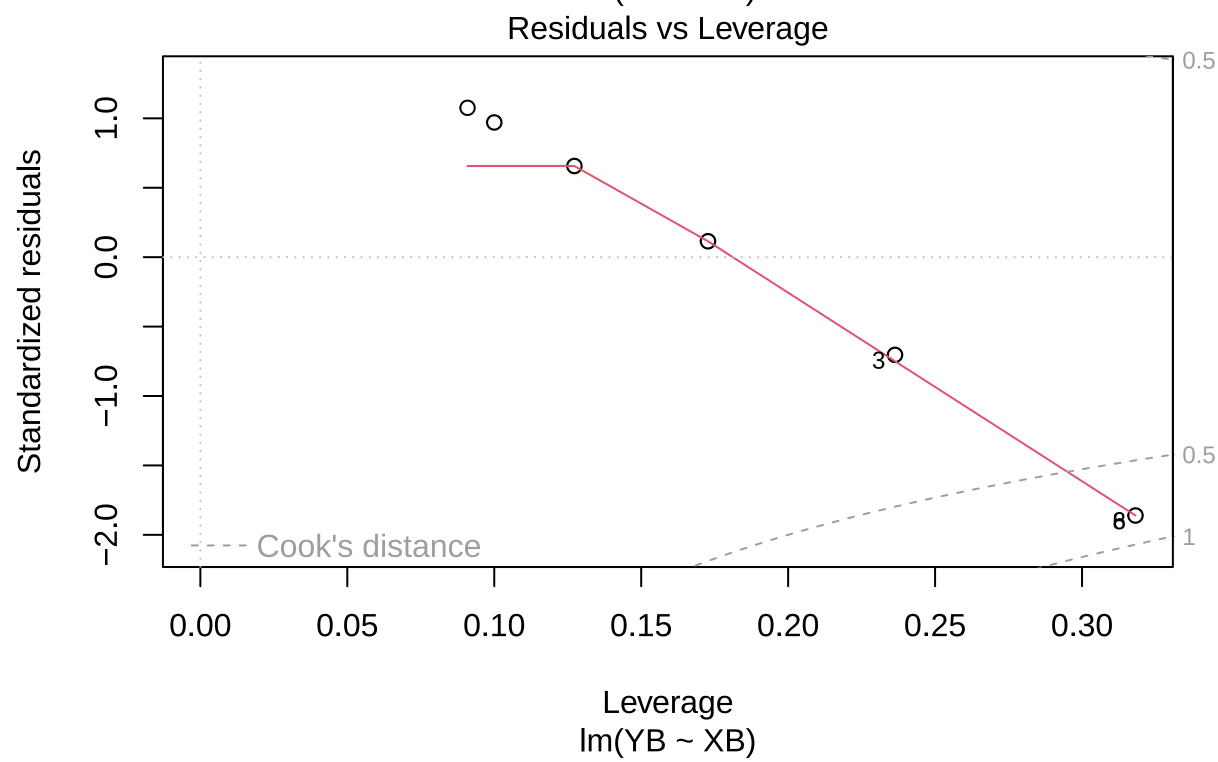
- Validate the basic hypothesis of the model

```r
plot(mb)
```


Residuals vs Fitted
Residuals
Fitted values
lm(YB ~ XB)
4
8
6
1
0
−1
−2
5
6
7
8
9
10

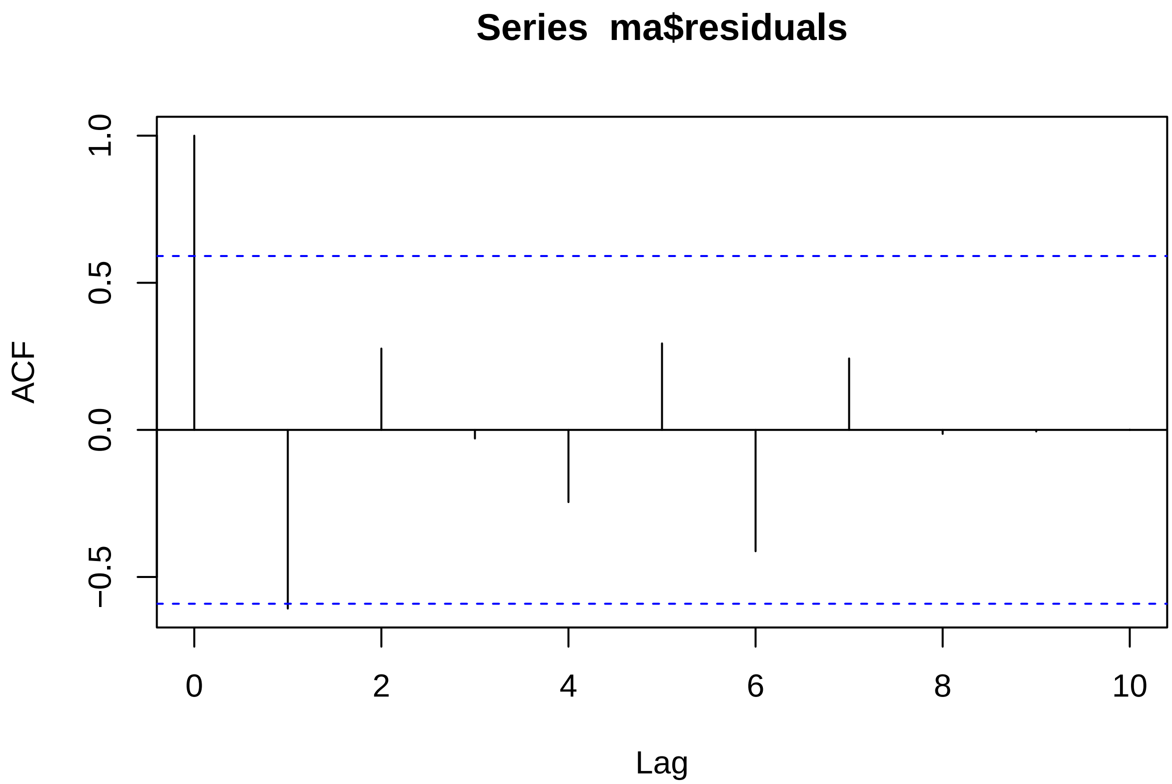

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(YB ~ XB)
4
6
8
1.0
0.5
−0.5
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5





Autocorrelation / Independence:

```
acf(ma$residuals)
```



Let's estimate a new linear model by adding the quadratic effect of the regressor XA:

Two ways of doing it:

```
mbb <- lm (YB~poly(XB,2), data = anscombe)
summary(mbb)
```

```
##
## Call:
## lm(formula = YB ~ poly(XB, 2), data = anscombe)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0013287 -0.0011888 -0.0006294  0.0008741  0.0023776
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5009091  0.0005043   14875   <2e-16 ***
## poly(XB, 2)1  5.2440442  0.0016725    3135   <2e-16 ***
## poly(XB, 2)2 -3.7116396  0.0016725   -2219   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001672 on 8 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.378e+06 on 2 and 8 DF,  p-value: < 2.2e-16
```

Second way of adding it:

```
mbb2 <- lm(YB~XB + I(XB^2), data = anscombe)
summary(mbb2)
```

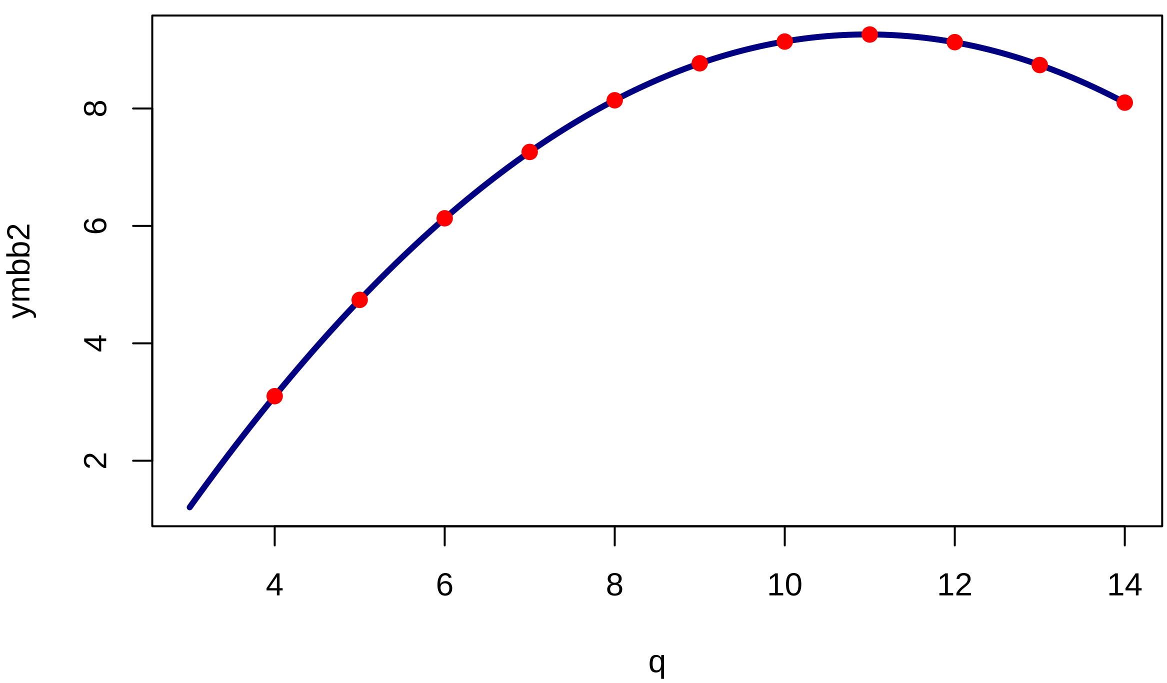
```
##
## Call:
## lm(formula = YB ~ XB + I(XB^2), data = anscombe)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0013287 -0.0011888 -0.0006294  0.0008741  0.0023776
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.9957343  0.0043299   -1385   <2e-16 ***
## XB           2.7808392  0.0010401    2674   <2e-16 ***
## I(XB^2)     -0.1267133  0.0000571   -2219   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001672 on 8 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.378e+06 on 2 and 8 DF,  p-value: < 2.2e-16
```

Let's regress a series of x on my model:

```r
q <- seq(3,14,0.01)
ymbb <- 7.5009091 + 5.2440442*q -3.7116396*q^2
ymbb2 <- -5.9957343 + 2.7808392*q -0.1267133*q^2
```

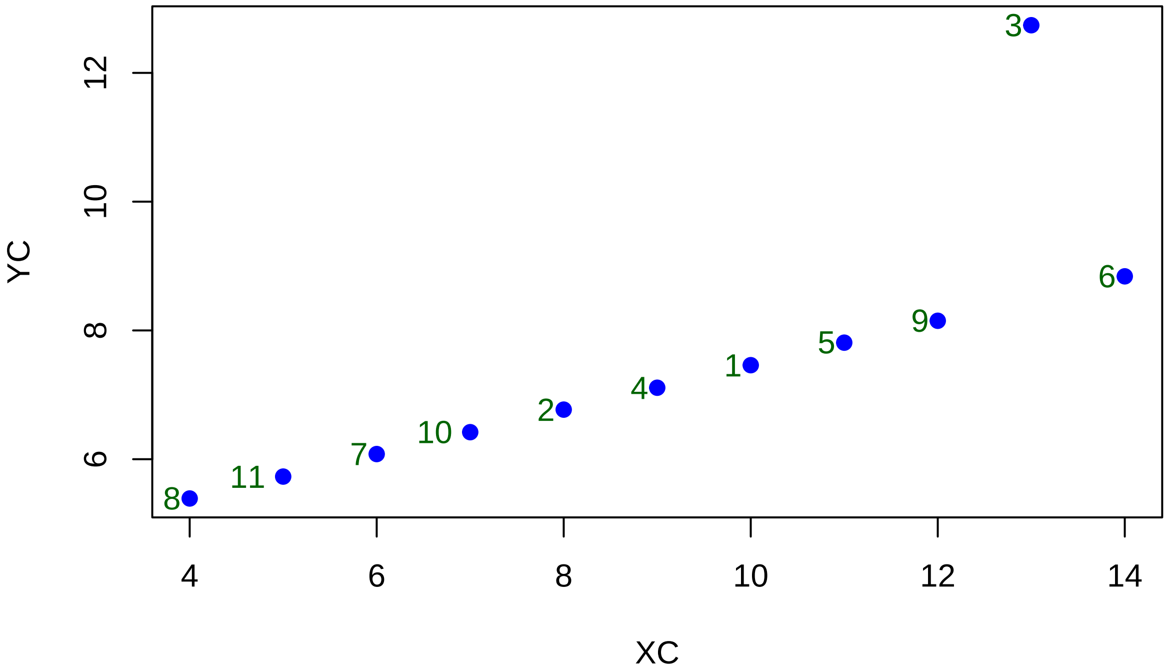
With the second approach:

```r
plot(q, ymbb2, type = 'l', col='navy', lwd=3)
points(YB~XB, col = "red", pch = 19)
```



SET C:

```r
plot(YC~XC, pch=19, col="blue")
text(XC,YC,label=row.names(anscombe), col = "darkgreen", adj=1.5)
```



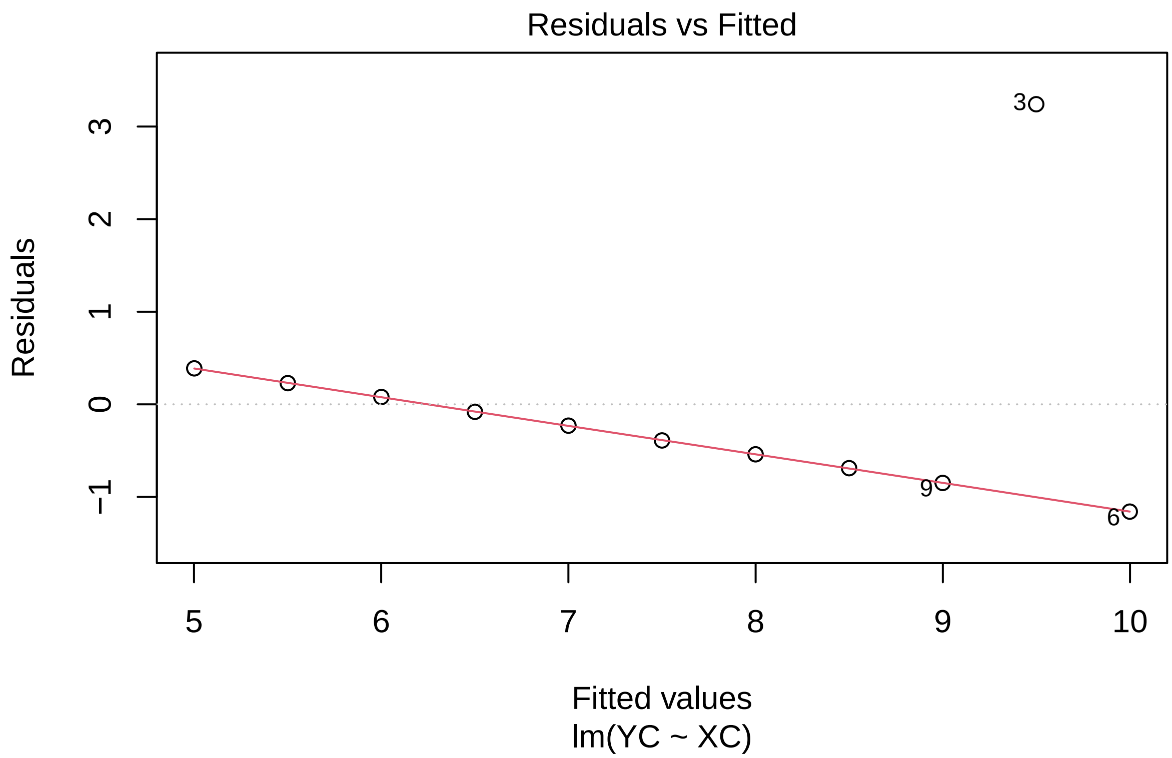
Let's forget about the influential observation we have detected on the data and estimate a linear model:

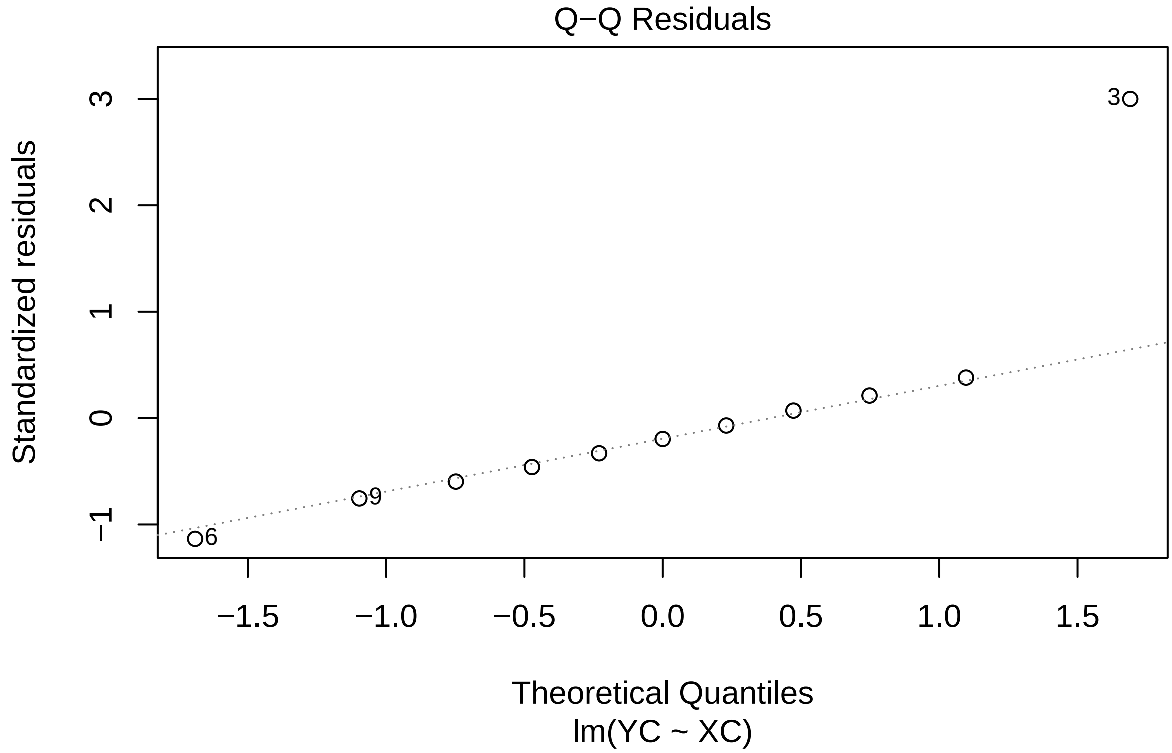
```r
mc <- lm(YC~XC, data=anscombe)
summary(mc)
```

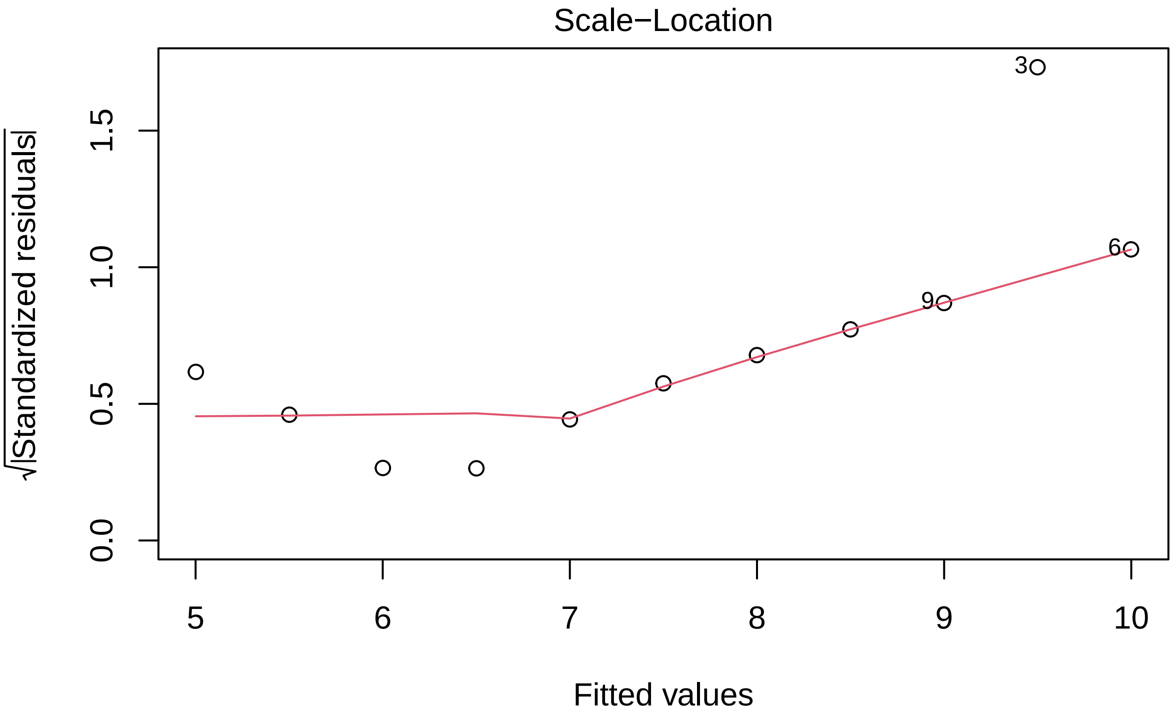
```
##
## Call:
## lm(formula = YC ~ XC, data = anscombe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## XC            0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

But we need to validate the basic hypothesis from the model:

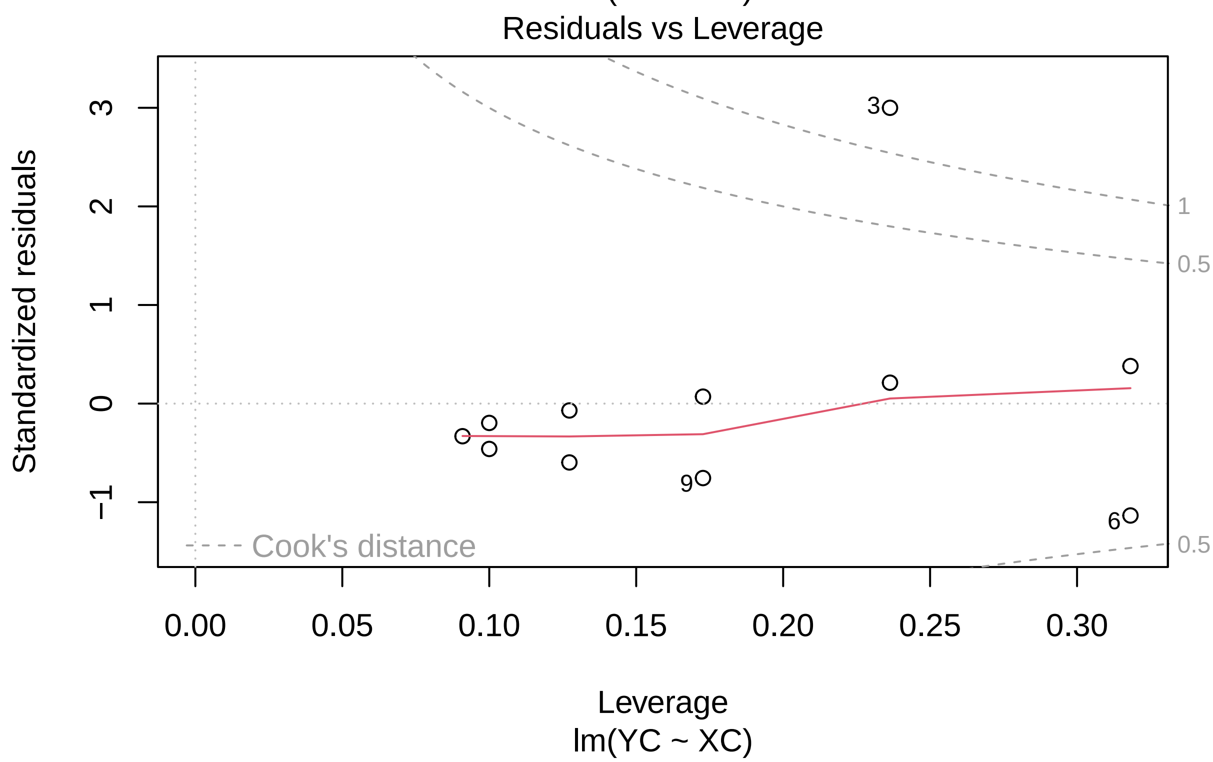
```r
plot(mc)
```


Residuals vs Fitted
Residuals
Fitted values
lm(YC ~ XC)
3
9
6
5
6
7
8
9
10
−1
0
1
2
3


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(YC ~ XC)
3
9
6
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1
0
1
2
3



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
6
9
5
6
7
8
9
10
Fitted values
lm(YC ~ XC)



Residuals vs Leverage
Standardized residuals
−1
0
1
2
3
3
9
6
1
0.5
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(YC ~ XC)