Sweetener

Jordi Valero and Marta P?rez-Casany

31 de octubre de 2018

ONE WAY ANOVA

We want to model the influence of a sweetener in the piglets Average Daily Gain (ADG)

```
library(car)
```

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

```
library(emmeans)
library(tables)
```

```
## Loading required package: 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(RcmdrMisc)
```

```
## Loading required package: sandwich

##
## Attaching package: 'RcmdrMisc'

## The following object is masked from 'package:Hmisc':
##
##     Dotplot
```

```
library(multcompView)
```

Loading the data and printing the top part

```
setwd("G:/PiE2/2018")
adg<-read.csv2("./Dades/ADG.csv")
head(adg)
```

```
##   DOSE      ADG
## 1    0 200.4167
## 2    0 190.0000
## 3    0 199.3333
## 4    0 191.0000
## 5    0 201.0000
## 6    8 210.6667
```

```
dim(adg)
```

```
## [1] 25  2
```

The data set has observations of 25 piglets and 5 doses. It is a balanced experiment because we have the same number of observations for each sweetener.

We perform descriptive statistics (EDA)

```
summary(adg)
```

```
##       DOSE           ADG       
##  Min.   : 0.0   Min.   :185.0  
##  1st Qu.: 8.0   1st Qu.:200.4  
##  Median :15.0   Median :221.3  
##  Mean   :14.6   Mean   :216.6  
##  3rd Qu.:20.0   3rd Qu.:228.7  
##  Max.   :30.0   Max.   :241.3
```

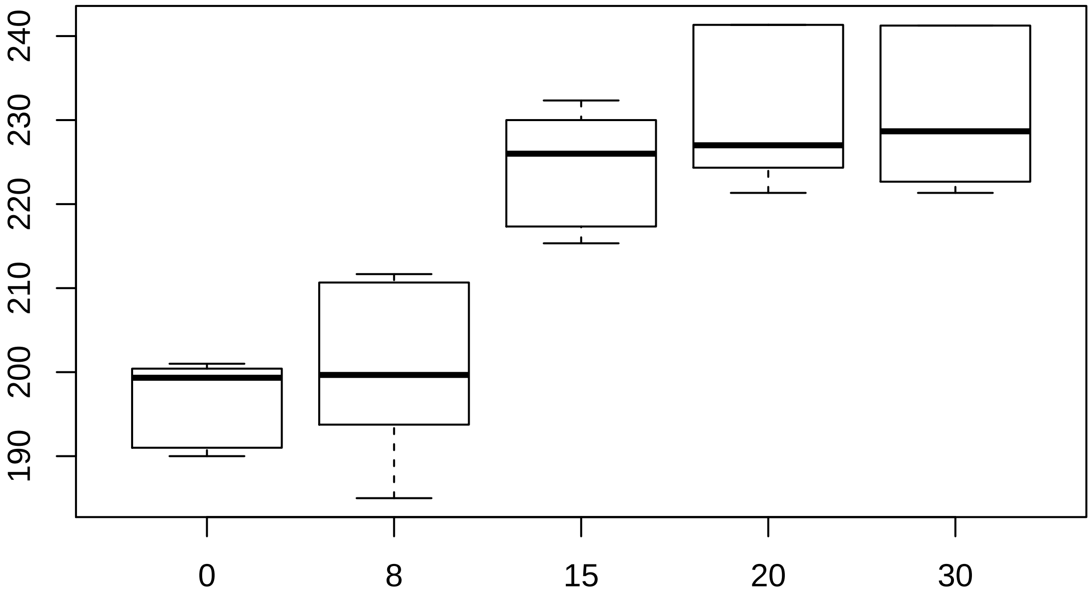
We define the dose as a factor

We perform the boxplot of ADG for eachd dose and the plot of means with its confident intervals.

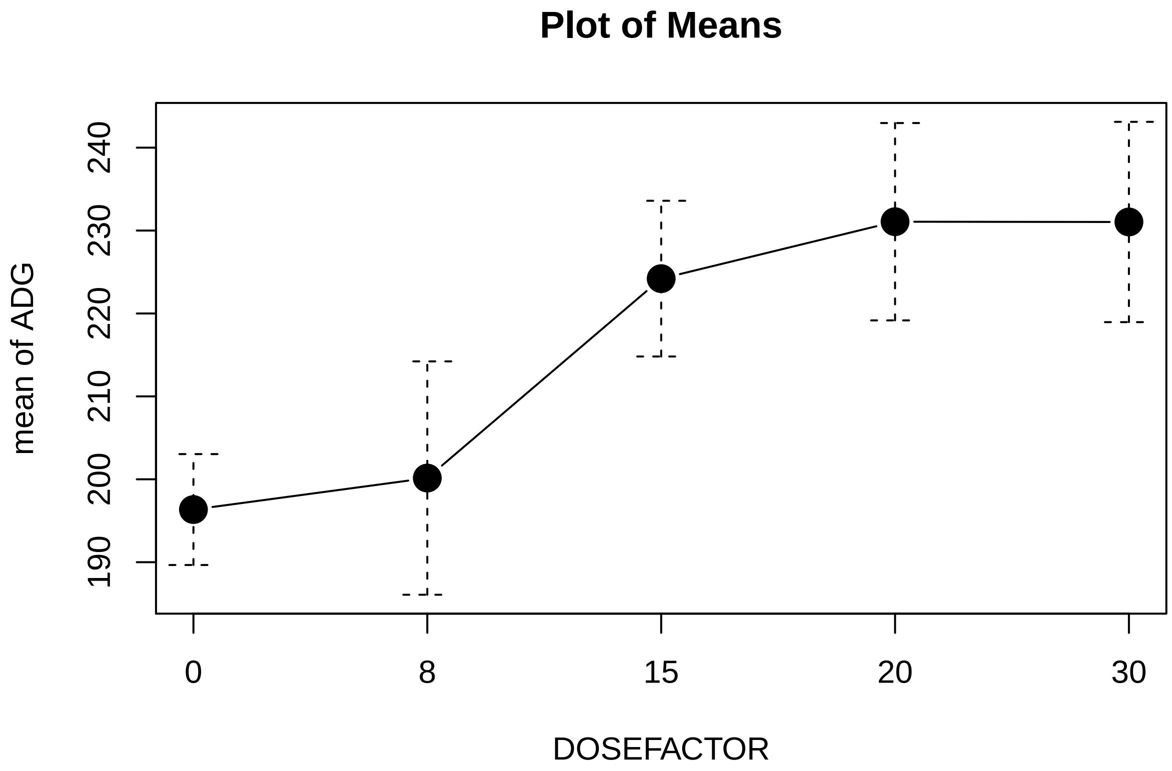
```
adg$DOSEFACTOR<-as.factor(adg$DOSE)
head(adg)
```

```
##   DOSE      ADG DOSEFACTOR
## 1    0 200.4167          0
## 2    0 190.0000          0
## 3    0 199.3333          0
## 4    0 191.0000          0
## 5    0 201.0000          0
## 6    8 210.6667          8
```

```
boxplot(adg$ADG~adg$DOSEFACTOR)
```



```r
with(adg, plotMeans(ADG, DOSEFACTOR, error.bars="conf.int",level=0.95, connect=TRUE))
```



We clearly observe that:

1) as the dose increases, the adg also increases. Nevertheless the last two doses perfomr very similarly.
2) We do not see big differences in the variability within the doses (homocedasticity property), nevertheless 5 observations are very few in general.
3) The simmetry of the adg distribution depends on the dose level, the first and the fourth doses are the ones that clearly have a lack of simmetry.

tabular(DOSE~ADG*((n=1)+mean+sd),adg)

with(adg, plotMeans(ADG, DOSE, error.bars="conf.int",level=0.95, connect=TRUE))

The MODEL

```
model1<-lm(adg$ADG~adg$DOSEFACTOR, data=adg)
summary(model1)
```

```
##
## Call:
## lm(formula = adg$ADG ~ adg$DOSEFACTOR, data = adg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1500  -6.7333  -0.4833   8.1333  11.5167
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        196.350      4.006  49.017  < 2e-16 ***
## adg$DOSEFACTOR8      3.800      5.665   0.671     0.51    
## adg$DOSEFACTOR15    27.850      5.665   4.916 8.34e-05 ***
## adg$DOSEFACTOR20    34.717      5.665   6.128 5.47e-06 ***
## adg$DOSEFACTOR30    34.683      5.665   6.122 5.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.957 on 20 degrees of freedom
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7393 
## F-statistic: 18.01 on 4 and 20 DF,  p-value: 2.071e-06
```

This model corresponds to what is called the One-way ANOVA

From the summary we can see:

1) The first dose has been taken as baseline. In consequence the ADG estimation for the first does is equal to the model intercept.
2) There are not significative differences between the first and second doses since the parameter associated to the second dose is not significant.
3) The last three doses give place to a ADG significatively different to the one of the first dose, since their parameters are significant.
4) To give a dose of 15 instead of 0 or 8 increases the ADG in 27.85 units. if the dose is 20, the increment is of 34.717 units and if we administrate a dose of 30, then the increment is of 34.68 units, always with respect to doses 0 and 8 which are not statistically different.
5) The model explains 78% of the variability. Thus 78% of the differences observed in the ADG are a direct consequence of the sweetener dose.
6) We reject the null hypothesis of the Omnibus test. Thus, the sweetener has a significant influence on the ADG.
7) The error standard deviation is estimated by $\hat{\sigma} = 8.957$.

Let us perform the anova/ANOVA test

```
anova(model1)
```

```
## Analysis of Variance Table
## 
## Response: adg$ADG
##                Df Sum Sq Mean Sq F value    Pr(>F)    
## adg$DOSEFACTOR  4 5780.1 1445.03  18.011 2.071e-06 ***
## Residuals      20 1604.6   80.23                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The type of sums of squares computed by the anova sentence are the type I sums of squares.

We can see by means of the F test that the factor has a significant influence of the response variable. The sum of squares that corresponds to the factor is equal to 5780.1 while the sum of squares devoted to the error is equal to 1604.6.

In what follows, by means of the Anova sentence, we are going to compute the type III sums of squares that, in this case will be equal to the type I because we just have a factor.

```r
Anova(model1)
```

```
## Anova Table (Type II tests)
## 
## Response: adg$ADG
##                 Sum Sq Df F value    Pr(>F)    
## adg$DOSEFACTOR 5780.1  4  18.011 2.071e-06 ***
## Residuals      1604.6 20                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey method for comparing the pairs of means

```r
emm<-emmeans(model1,~DOSEFACTOR)
emm
```

```
##  DOSEFACTOR   emmean       SE df lower.CL upper.CL
##  0          196.3500 4.005784 20 187.9941 204.7059
##  8          200.1500 4.005784 20 191.7941 208.5059
##  15         224.2000 4.005784 20 215.8441 232.5559
##  20         231.0667 4.005784 20 222.7107 239.4226
##  30         231.0333 4.005784 20 222.6774 239.3893
## 
## Confidence level used: 0.95
```

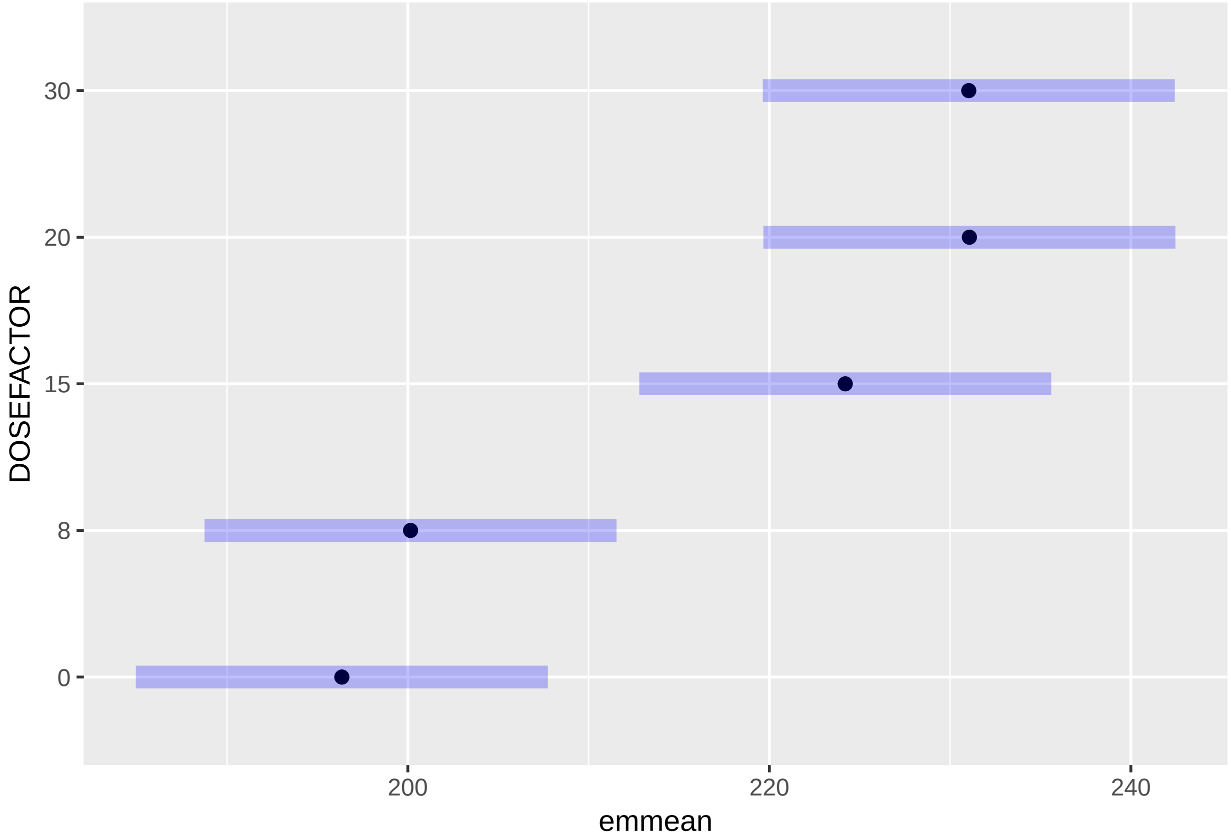
```r
pairs(emm)
```

```
##  contrast   estimate       SE df t.ratio p.value
##  0 - 8     -3.800000 5.665034 20  -0.671  0.9605
##  0 - 15   -27.850000 5.665034 20  -4.916  0.0007
##  0 - 20   -34.716666 5.665034 20  -6.128  <.0001
##  0 - 30   -34.683333 5.665034 20  -6.122  <.0001
##  8 - 15   -24.050000 5.665034 20  -4.245  0.0032
##  8 - 20   -30.916666 5.665034 20  -5.457  0.0002
##  8 - 30   -30.883333 5.665034 20  -5.452  0.0002
##  15 - 20   -6.866667 5.665034 20  -1.212  0.7446
##  15 - 30   -6.833334 5.665034 20  -1.206  0.7479
##  20 - 30    0.033333 5.665034 20   0.006  1.0000
## 
## P value adjustment: tukey method for comparing a family of 5 estimates
```

```r
CLD(emm,alpha=0.05)
```

```
##  DOSEFACTOR   emmean       SE df lower.CL upper.CL .group
##  0          196.3500 4.005784 20 187.9941 204.7059  1    
##  8          200.1500 4.005784 20 191.7941 208.5059  1    
##  15         224.2000 4.005784 20 215.8441 232.5559   2   
##  30         231.0333 4.005784 20 222.6774 239.3893   2   
##  20         231.0667 4.005784 20 222.7107 239.4226   2   
## 
## Confidence level used: 0.95 
## P value adjustment: tukey method for comparing a family of 5 estimates 
## significance level used: alpha = 0.05
```

```
plot(emm,level=0.99,adjust="tukey")
```



```
confint(emm,level=0.99,adjust="tukey")
```

```
##  DOSEFACTOR   emmean       SE df lower.CL upper.CL
##  0          196.3500 4.005784 20 182.1292 210.5708
##  8          200.1500 4.005784 20 185.9292 214.3708
##  15         224.2000 4.005784 20 209.9792 238.4208
##  20         231.0667 4.005784 20 216.8458 245.2875
##  30         231.0333 4.005784 20 216.8125 245.2542
##
## Confidence level used: 0.99
## Conf-level adjustment: sidak method for 5 estimates
```

The variable emm contains the five marginal means (emmeans: estimated marginal means), jointly with their corresponding standard error and confidence intervals computed from the student t-distribution.

The command pairs allows to perform two by two comparisons with several methods. By default the choosen method is the Tukey.

The CLD command (compact letter display) joint with the same number the means that are not statistically different with the Tukey method, unless another method is specified.

The plot and confint commands also compute the confidence intervals for each mean but they are computed based on the student rang distribution.

We can deduce by means of looking at the p-values that:

1) the dose zero is not statistically different from the dose 8,

2) the dose 20 and 15 are not statistically different

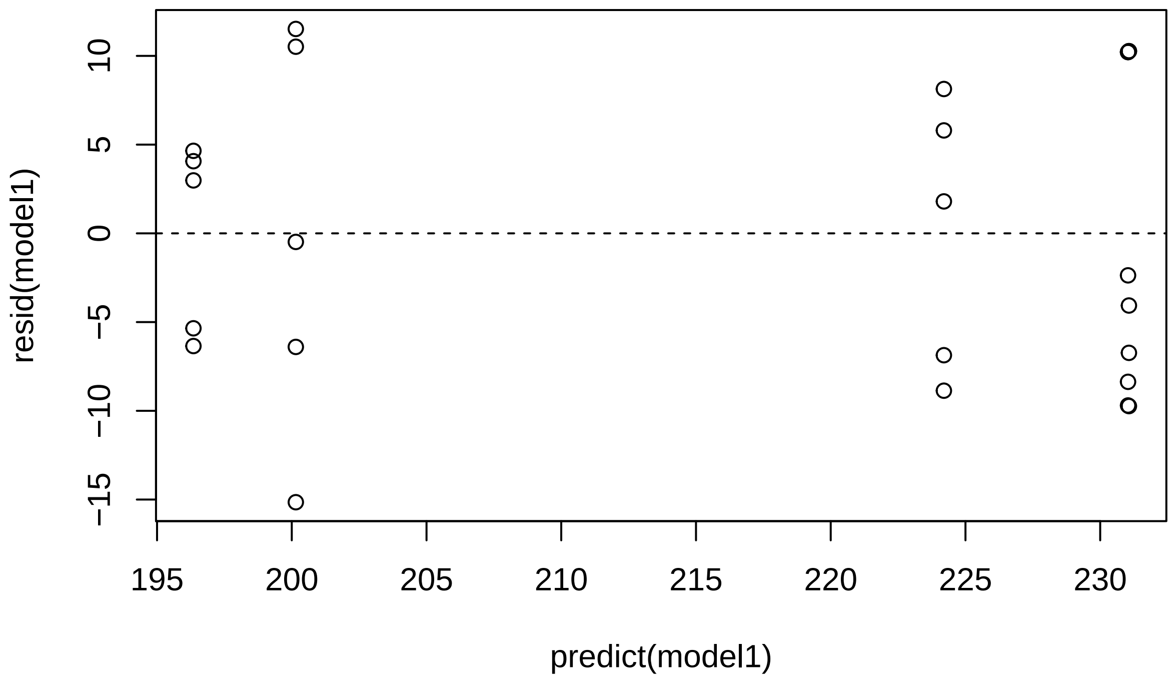
3) the dose 30 and 15 are not statistically different

4) the dose 30 and 20 are not statistically different

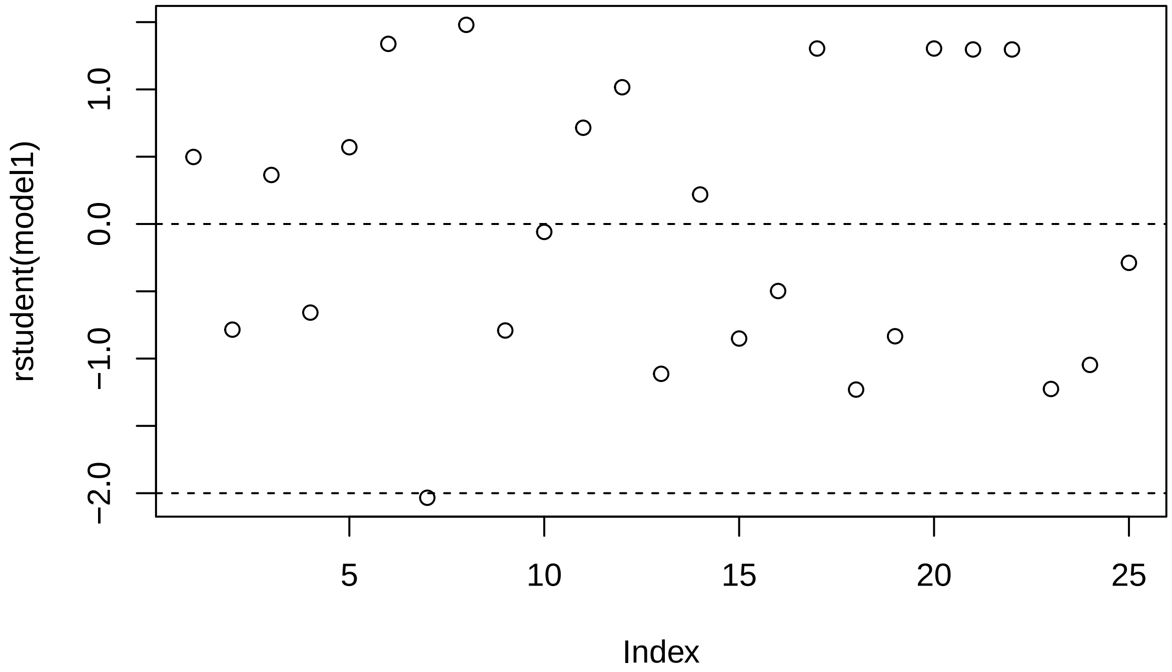
Thus, we can conclude that in terms of ADG we distinguish two dose groups: group1: contains dose 0 and 8 and group2: contains doses 15, 20 and 30. The effect of the doses in the ADG are not distinguishable between doses of the same group, but they are different for the two groups.

Model adequance checking

```
plot(predict(model1),resid(model1))
abline(h=0,lty=2)
```



```
plot(rstudent(model1))
abline(h=c(-2,0,2),lty=2)
```



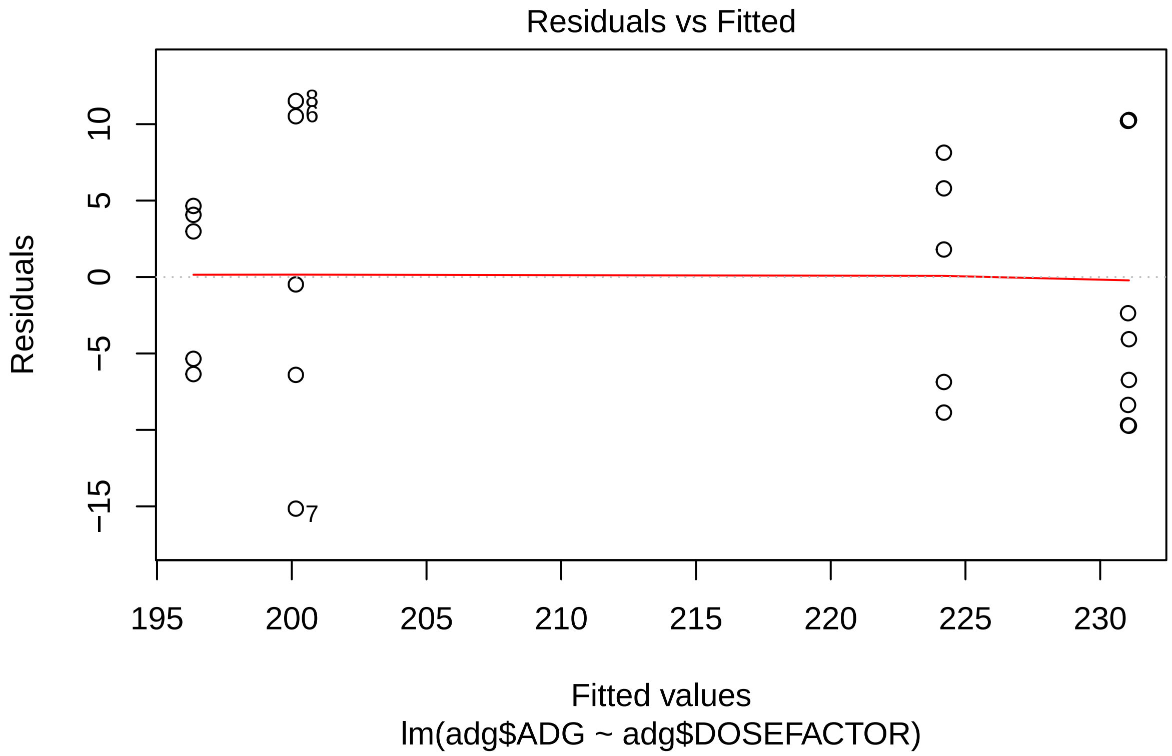
Looking at the residuals vs predicted plot we do not observe any pattern.

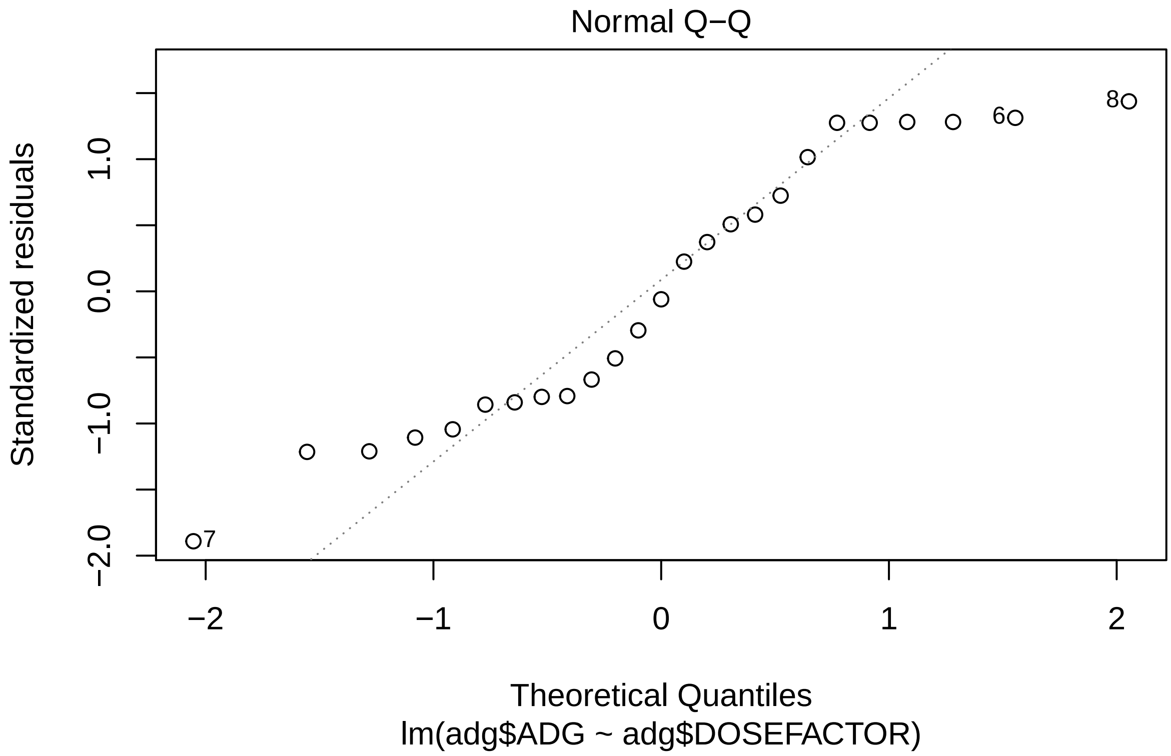
The variances of the residuals are quite similar in the fifth groups.

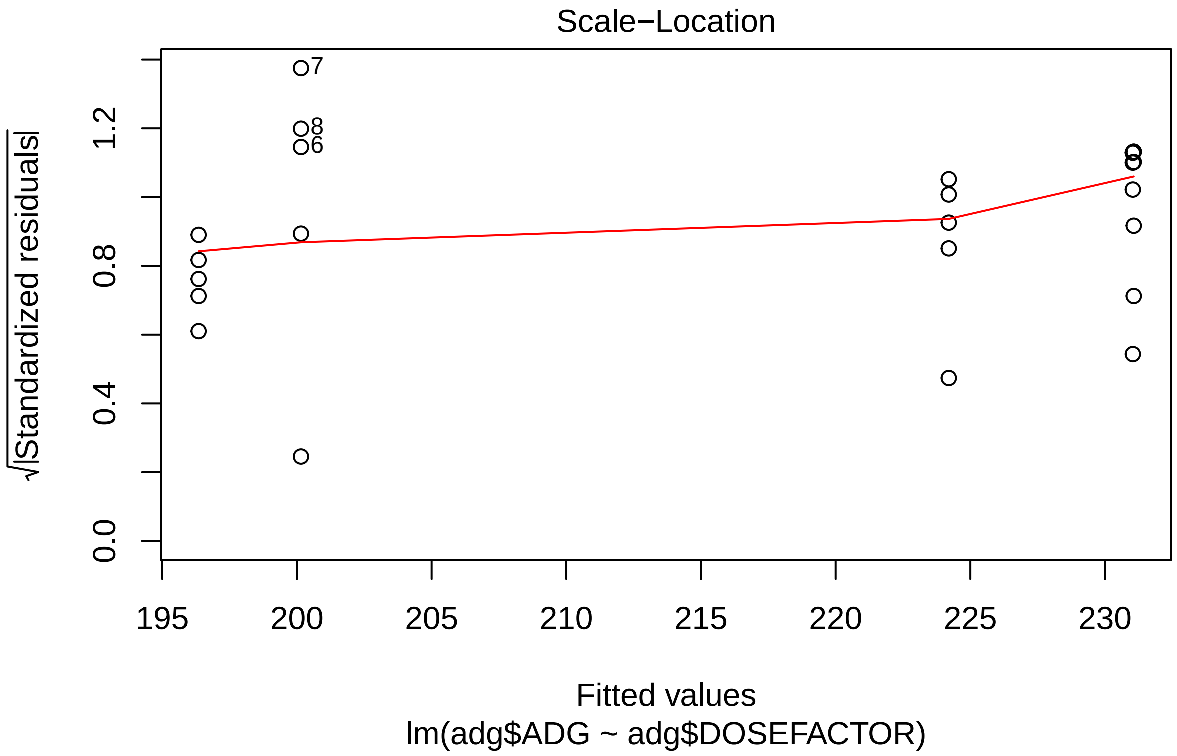
The standarized residuals do not show patterns neither.

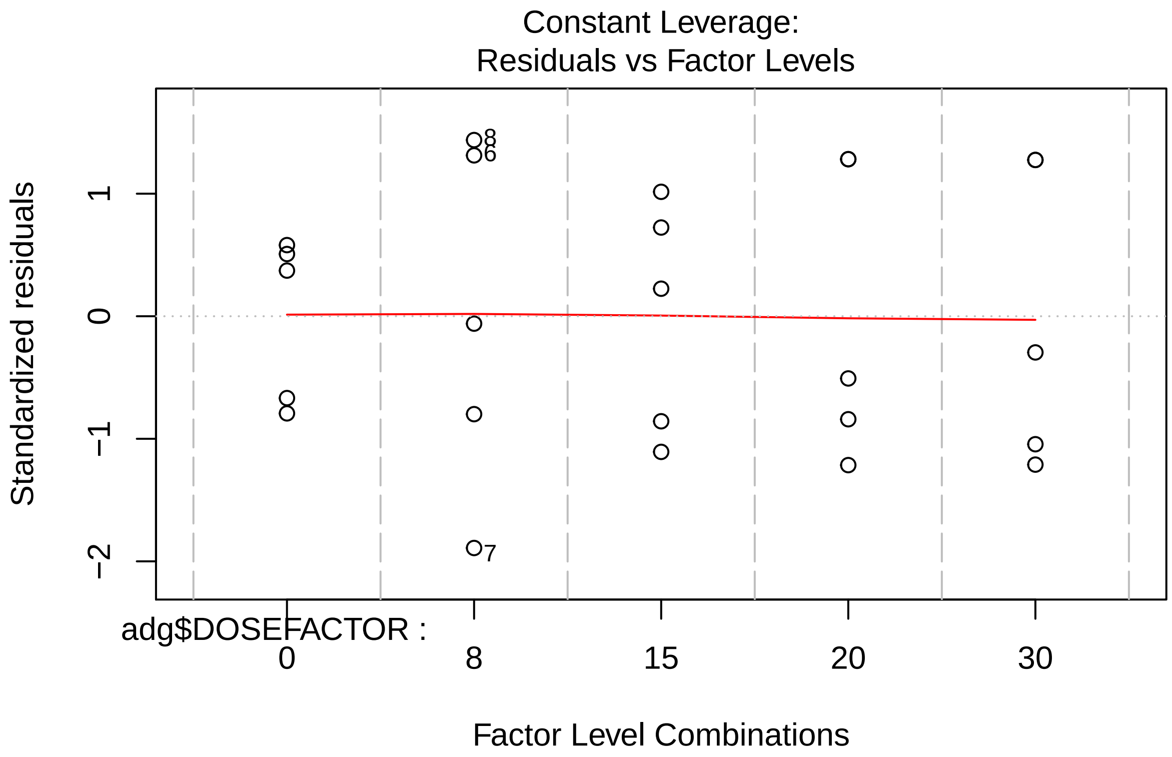
The four plots of R model adquance cheking

```r
plot(model1,ask=F)
```


Residuals vs Fitted
Residuals
10
5
0
−5
−15
8
6
7
195
200
205
210
215
220
225
230
Fitted values
lm(adg$ADG ~ adg$DOSEFACTOR)


Normal Q−Q
Standardized residuals
−2.0
−1.0
0.0
1.0
−2
−1
0
1
2
Theoretical Quantiles
lm(adg$ADG ~ adg$DOSEFACTOR)
7
6
8


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
195
200
205
210
215
220
225
230
7
8
6
Fitted values
lm(adg$ADG ~ adg$DOSEFACTOR)



Again we do not observe patterns and neither influcential points or outliers.

The normality assumption is not very clear, but this is probably a consequence of having an small set of data (just 25 observations).

We are not going to be very strict with this assumption and small discrepances from this assumption will be accepted.

Test for homogenity of variances

In what follows we are going to perform two test for checking the homocedasticity hypothesis

```
leveneTest(model1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  0.5712 0.6866
##       20
```

```
bartlett.test(ADG~DOSE,adg)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  ADG by DOSE
## Bartlett's K-squared = 2.1267, df = 4, p-value = 0.7125
```

In both cases it is not rejected.