Linear model: Fattening of piglets

To see if the dose of a sweetener improves the fattening of piglets, one experiment was performed. A set of piglets with similar conditions where selected and 5 different sweetener doses where considered and have been randomly assigned to the piglets. The **response variable** is the *average daily gain*, ADG, and the **explanatory variable** is the *sweetener dose*:

(1) Define and fit the linear model appropriate to this situation.

$$Y_{ij} = \mu + \alpha_i + e_{ij} \quad e_{ij} \sim N(0, \sigma^2)$$

Where α_i is the factor and j are the repetitions. Each $\mu_i = \mu + \alpha_i$. μ of the complete model, the groups all together.

```
library(car)
library(emmeans)
library(tables)
library(RcmdrMisc)

dd<-read.csv2("ADG.csv")
head(dd)
```

```
##   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
```

We perform descriptive statistics:

```
summary(dd)
```

```
##       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
```

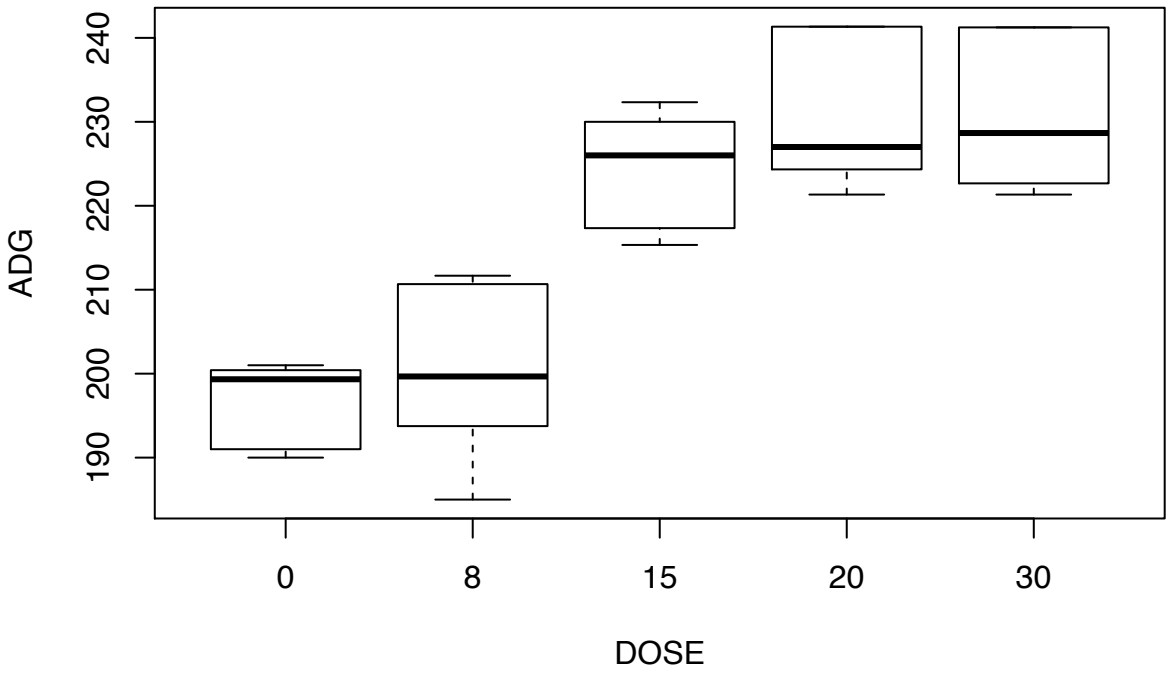
Observation: the dose has to be a factor, not a covariate! First of all, we check whether it is a factor in the data set:

```
is.factor(dd$DOSE)
```

```
## [1] FALSE
```

And if we plot a scatter plot, we would obtain a regression line, but we do not want that, so we factor the variable:

```
dd$DOSE<-as.factor(dd$DOSE)
sp(ADG~DOSE,dd,smooth=F)
```



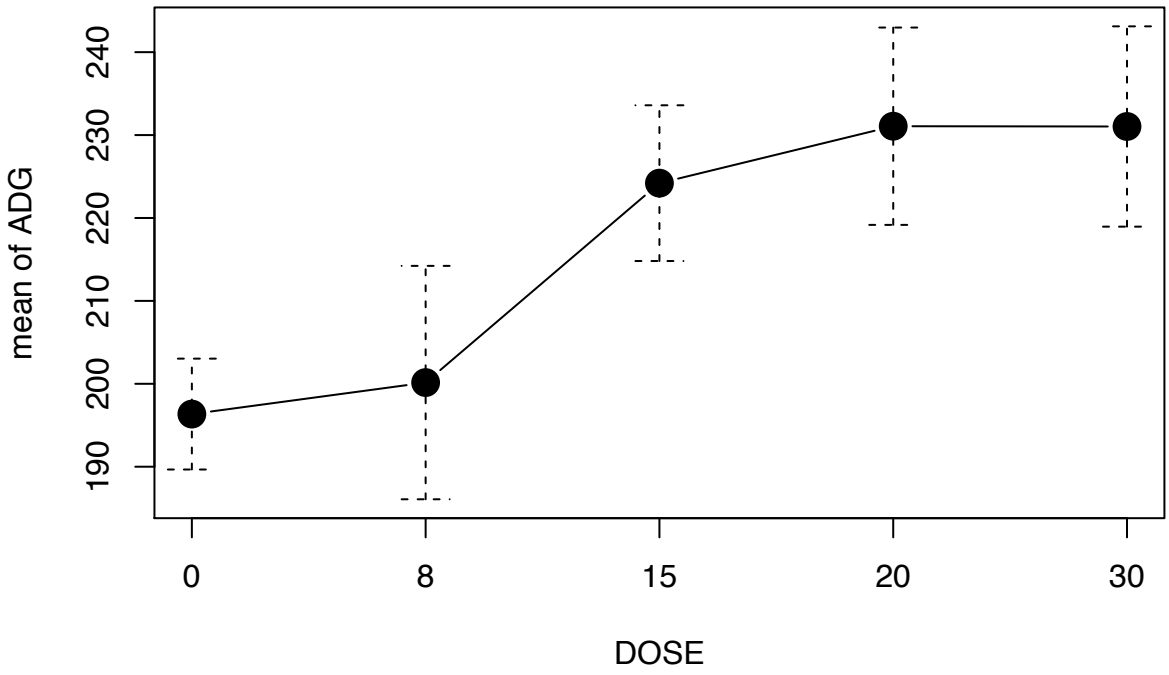
This is the boxplot of ADG for each dose.

```
tabular(DOSE~ADG*((n=1)+mean+sd),dd) # information for each dose group.
```

	ADG		
DOSE	n	mean	sd
0	5	196.3	5.385
8	5	200.2	11.335
15	5	224.2	7.563
20	5	231.1	9.584
30	5	231.0	9.727

And now, the plot of means with its confident intervals:

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



We clearly observe that:

(a) As the dose increases, the adg also increases. Nevertheless, the last two doses are very similarly.
(b) We do not see big differences in the variability within the doses (homocedasticity property), nevertheless 5 observations are very few, in general.
(c) The symmetry 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.

Com que el nostre objectiu és comparar si s'engreixen igual o no amb diferents dosis d'edulcorant, primer de tot haurem d'**ajustar el model**. Després haurem de contrastar emb el **test Omnibus**, per saber si el model explica la variabilitat o no (si hi ha efectes o no, per les dosis). Després, fer el **test Anova** per mirar efecte per efecte. I finalment, per veure les diferències entre les dosis i quina dosi va millor, haurem de fer les **comparacions múltiples** (de Tukey).

We define the linear model:

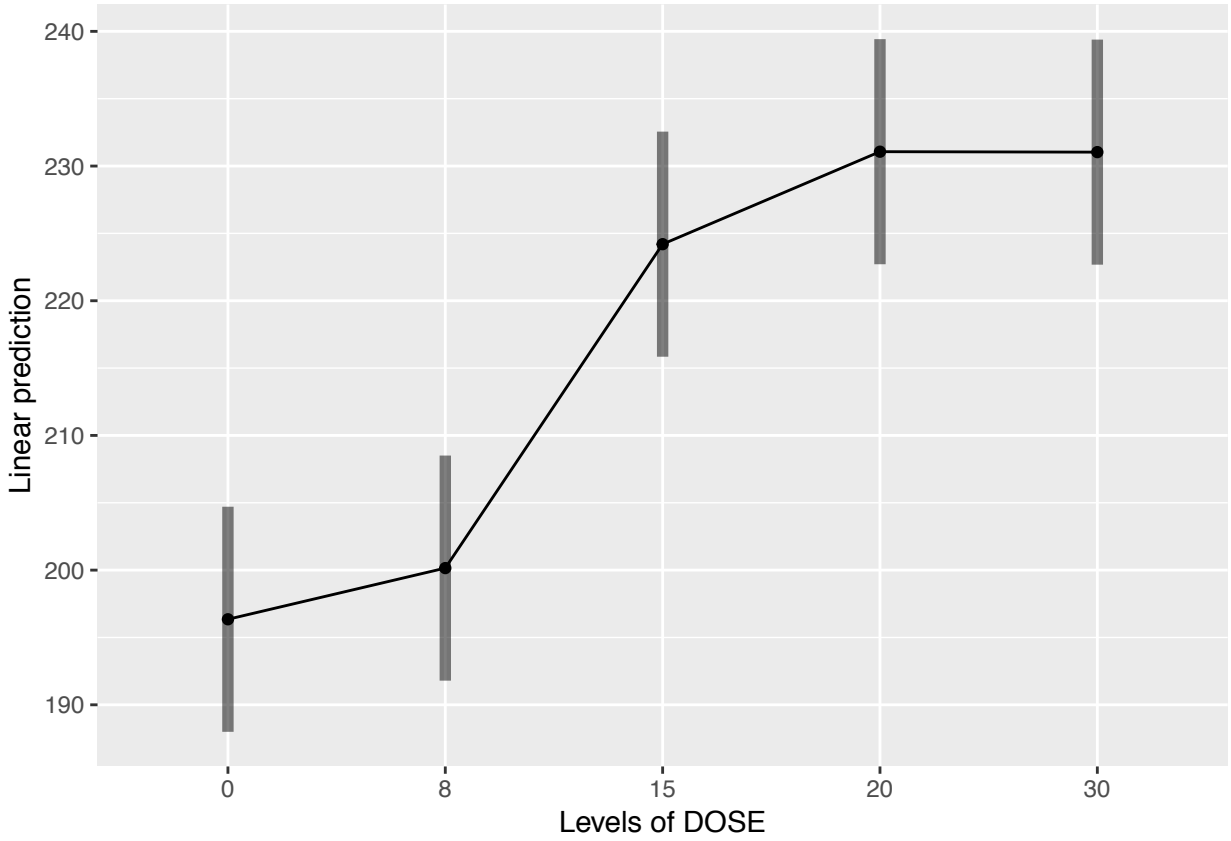
```
summary(mod<-lm(ADG~DOSE,dd))
```

```
##
## Call:
## lm(formula = ADG ~ DOSE, data = dd)
##
## 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 ***
## DOSE8          3.800      5.665   0.671     0.51
## DOSE15        27.850      5.665   4.916 8.34e-05 ***
## DOSE20        34.717      5.665   6.128 5.47e-06 ***
```

```
## DOSE30        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
```

El que fa és calcular per cada dosi - l'estimació de l'esperança. I per cada una, també podem visualitzar l'interval de confiança:

```
(emmip(mod,~DOSE,CIs=T))
```



Les linies verticals son les desviacions tipus. Són totes pràcticament de la mateixa longitud perquè n=5 en cada grup.

This model corresponds to what is called the **One-way ANOVA**. From the summary we can see:

(a) The first dose has been taken as baseline. In consequence the ADG estimation for the first does equal to the model intercept. That is to say, the intercept is the mean of the first dose.

Observavció: els paràmetres que obtenim al *summary* no ens interessen gens. Perquè ens interessa $\mu + \alpha_i$ i l'espreança del segon és el primer paràmetre més el segon, l'esperança del tercer és el primer paràmetre més el tercer i axió succesivament.

(b) There are not significative differences between the first and second doses since the parameter associated to the second dose is not significant.
(c) The last three doses give place to a ADG significatively different to the one of the first dose, since their parameters are significant.
(d) 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.

(e) The model explains 78% of the variability. Thus 78% of the differences observed in the ADG are a direct consequence of the sweetener dose.

(f) We reject the null hypothesis of the Omnibus test. Thus, the sweetener has a significant influence on the ADG.

(g) The error standard deviation is estimated by $\hat{\sigma} = 8.957$.

(2) What can we conclude from the ANOVA test?

Anem a fer el test Anova. En aquest cas és redundant, perquè l'Omnibus ja ens dóna informació. Hi ha diferents formes de calcular la suma de quadrats. Hi ha dues comandes diferents una pel tipus I (minúscula) i l'altra es pel tipus II o III (majúscula, si no especifiquem res II, si especifiquem - III). El tipus II i III difereixen quan no hi ha igualtat de repeticions. En general, el més recomanable és del tipus II.

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: ADG
##           Df Sum Sq Mean Sq F value    Pr(>F)
## DOSE       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 upon 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 II sums of squares that, in this case will be equal to the type I because we have got just one factor.

```
Anova(mod)
```

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

Quina conclusió podem extreure: té la dosi cap efecte sobre el guany mig diari? Sí, té un efecte molt clar. En dir això, la probabilitat d'equivocar-me és petitíssima, pel test omnibus i pel test Anova. Si ja sabem que hi ha diferències, ens interessa saber quins són aquests paràmetres estimats, μ i α, i saber quines diferències hi ha entre aquests paràmetres. Per això, necessitem la comanda *emmeans*.

Tukey method for comparing the pairs of means

Ens calcula l'estimació de les mitjanes marginals (*emmean*): 1r+2n, 1r+3r, etc. I això, no depén de la parametrització que hem utilitzat. Qualsevol parametrització que utilitzem, canvia els paràmteres en el *summary*, però no l'estimació de les mitjanes marginals.

```
(emm<-emmeans(mod,~DOSE))
```

```
##  DOSE   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
```

La dosi 0 té una estimació de l'esperança de 196.35, la dosi 8 - 200.15. Utilitzant la T-Student, tenim calculats els intervals de confiança a la dreta (del 95%).

Ja tenim els parametres. Ara ens interessa saber quines diferències hi ha entre les dosis. Per aixo, farem molts tests:

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \mu_1 = \mu_3 \quad vs \quad H_1 : \mu_1 \neq \mu_3$$

...

$$H_0 : \mu_1 = \mu_j \quad vs \quad H_1 : \mu_1 \neq \mu_j$$

I farem tantes comparacions com grups hi hagi, en el nostre cas, $5 \cdot 4 = 20$, però com que l'ordre no importa, $= 10$. Clar, fer 10 tests simultàniament, si la probabilitat d'equivocar-me en un test és el 5%, la probabilitat d'equivocar-nos en algún - augmenta. Llavors haurem d'e fer'utilitzar algún sistema per tenir això en compte: Tukey! És un dels més equilibrats.

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 method.

Tests *parella per parella*:

```
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
```

Organitza els 10 tests. El primer i el segon: no ens atrevim a dir que són diferents (ens equivocariem el 96% de vegades). El primer i el tecer són diferents, el primer i el quart també són diferents, etc. S'ha d'anar un per un per veure si sí o si no.

- The dose zero is not statistically different from the dose 8.
- The dose 20 and 15 are not statistically different.
- The dose 30 and 15 are not statistically different.
- The dose 30 and 20 are not statistically different.

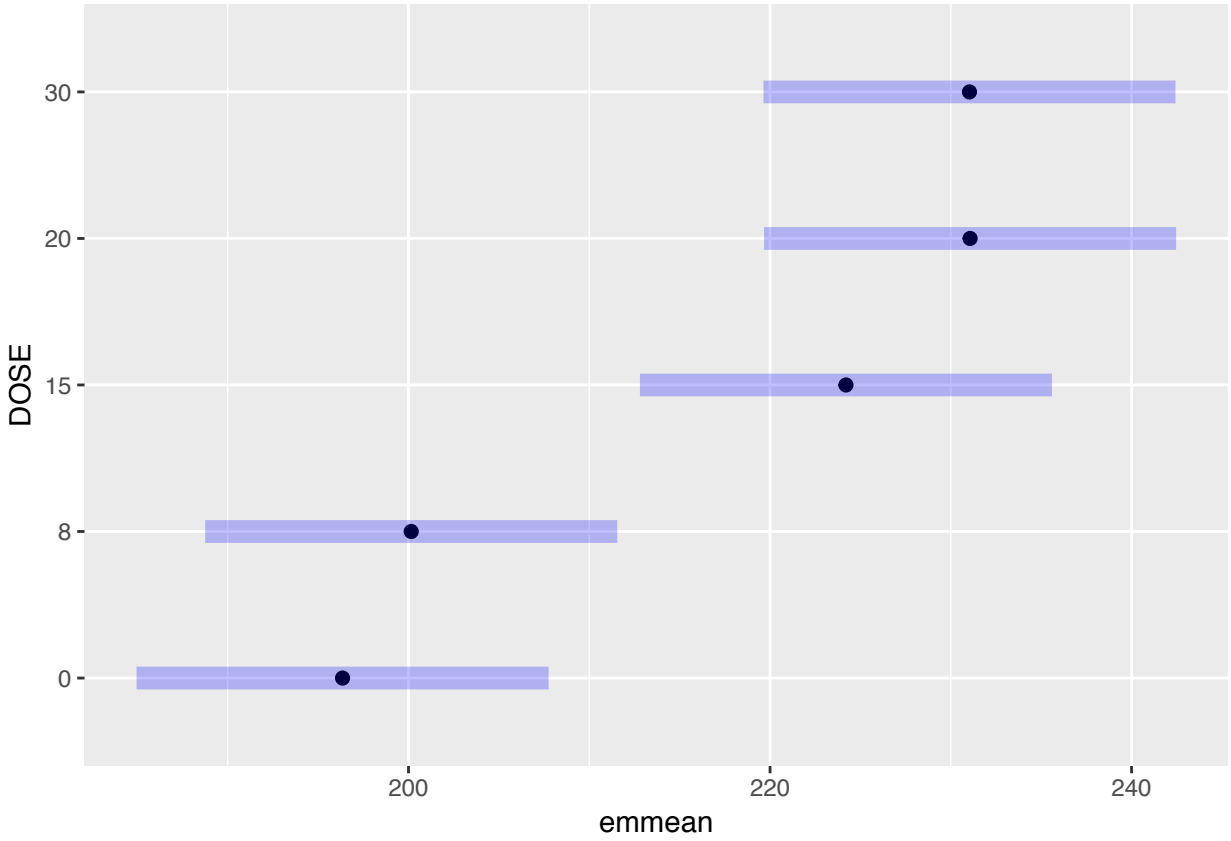
En base d'això, podem fer CLD (*Compact letter display*), que és ensenyar el test de forma compacta, i aquí hem de dir quina α definim (nivell de confiança), si no diem res, sera $\alpha = 0.05$. Ho posarà d'una forma més fàcil per a la lectura.

```
CLD(emm,alpha=0.01)
```

```
##  DOSE   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 = 0.01
```

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.

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




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

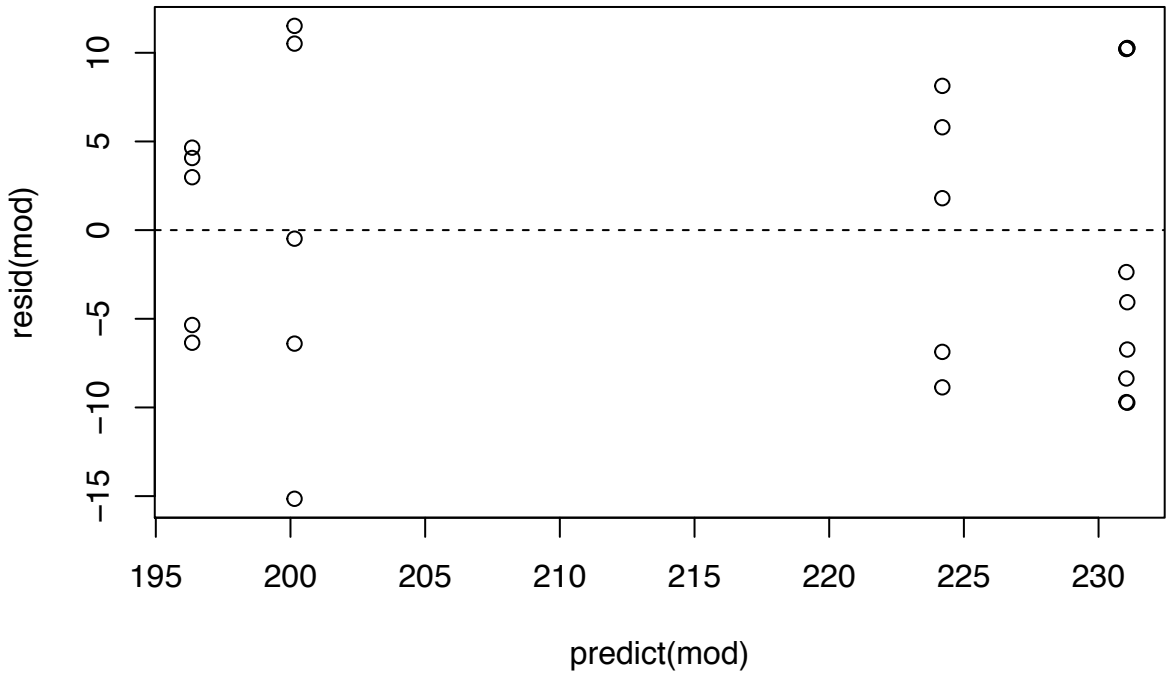
```
##  DOSE   emmean       SE df lower.CL upper.CL
##  0    196.3500 4.005784 20 182.1292 210.5708
##  8    200.1500 4.005784 20 185.9292 214.3708
##  15   224.2000 4.005784 20 209.9792 238.4208
```

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

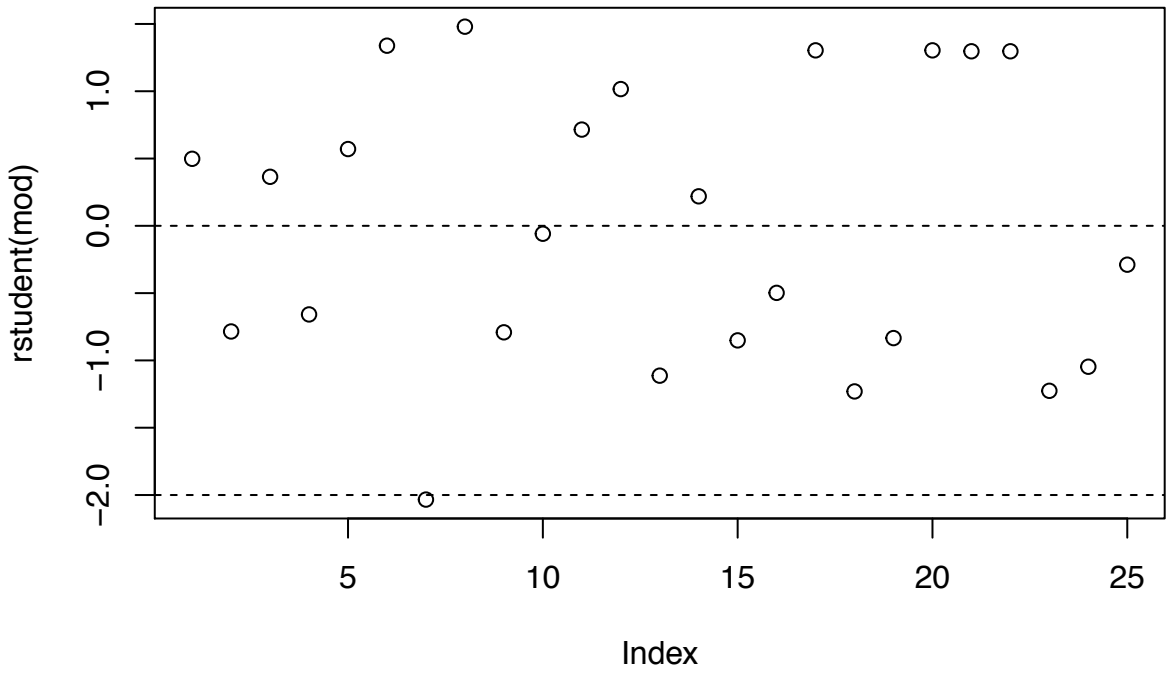
Ara, igual que en la regressio, ens quedaria fer els diagnòstics. Amb analisi de variàncies és més senzills encara:

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



Surten 4 columnes, perquè tenim 5 grups. Sí, hi ha diferència entre les grups, però no hi ha cap tendència a que vagin augmentant o coses semblants. En aquest sentit, no hi ha problema.

```
plot(rstudent(mod))
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 five groups. The standarized residuals do not show patterns neither.

We can test for the homogenity of the variances; two tests for checking the homocedasticity hypothesis:

```
leveneTest(mod)
```

```
## 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,dd)
```

```
##
##  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.