VitaminC

Jordi Valero and Marta Pérez-Casany

12 de noviembre de 2018

```
library(car)
```

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

```
library(HH)
```

```
## Loading required package: lattice
## Loading required package: grid
## Loading required package: latticeExtra
## Loading required package: RColorBrewer
## 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
```

```
library(car)
library(emmeans)
```

```
##
## Attaching package: 'emmeans'
## The following object is masked from 'package:HH':
##
##     as.glht
## The following object is masked from 'package:multcomp':
##
##     cld
```

```
library(tables)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
## The following object is masked from 'package:latticeExtra':
##
##     layer
##
## 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
```

ANCOVA

We want to COMPARE TWO PEDAGOGICAL METHODOLOGIES.

Loading the data and printing the top part

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

```
##   treat week vitc
## 1     a    1 30.2
## 2     a    2 29.2
## 3     a    3 23.8
## 4     a    4 27.4
## 5     a    5 16.8
## 6     a    6 29.6
```

```
dim(vitamincdata)
```

```
## [1] 72  3
```

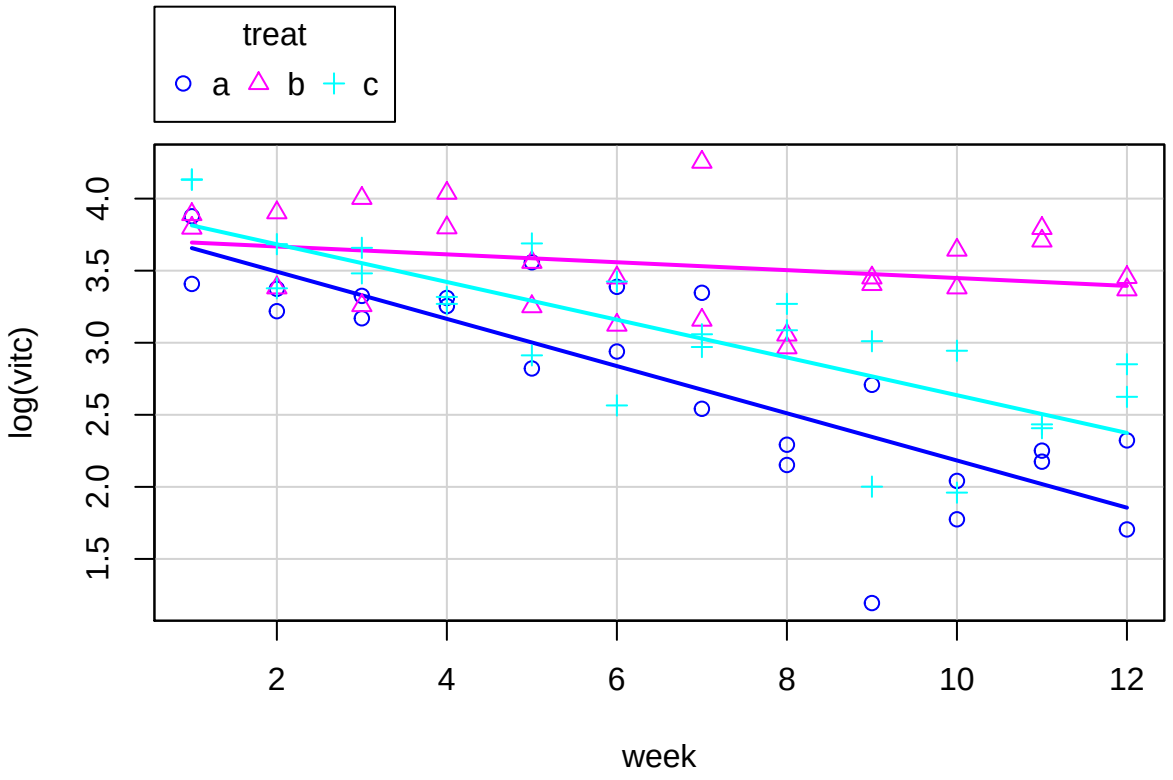
The dataset has 72 rows and three columns. Each row corresponds to a fix orange juice. The columns specify: the treatment or conservation method which is a categorical variable with three levels: a, b, and c, the week which is also a categorical variable with 12 levels, indicating the week in which the measure has been taken and finally, the vitaminC level of the orange juice.

The response of dependent variable is the vitaminC level and the covariates or explanatory variables are the conservation method and the week.

The most important question is to known if the orange juices lose the vitaminC in a similar way for the different consevation methods.

Descriptive statistics

```
scatterplot(log(vitc)~week|treat,smooth=F,dat=vitamincdata)
```



From the scattered plot we see a clear influence of the week in the lose of vitaminC specially in conservation methods different from b. Conservation method a is the one that seems to lose vitaminC faster.

Model with different intercepts ans slopes

Given that the VitaminC of an orange juice is an exponential fucntion of the Week, in order to fit a linear model, it is necessary to apply a logarithmic transformation to the response variable.

Important to observe that, assuming that $\log(VitaminC)$ is normal distributed is equivalent, by definition, to assume that $VitaminC$ follows a log-normal distribution. So, by doing that we are changing the distribution of the response variable.

The fisrt model we fit contains the main effects as well as the interaction term.

```
mvitaminc1<-lm(log(vitc)~week+treat+treat:week,vitamincdata)

summary(mvitaminc1)
```

```
##
```

```
## Call:
## lm(formula = log(vitc) ~ week + treat + treat:week, data = vitamincdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15293 -0.18979 -0.01522  0.24540  0.72179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.82038    0.15788  24.199  < 2e-16 ***
## week        -0.16373    0.02145  -7.632 1.20e-10 ***
## treatb      -0.09785    0.22327  -0.438    0.663
## treatc       0.12472    0.22327   0.559    0.578
## week:treatb  0.13636    0.03034   4.495 2.88e-05 ***
## week:treatc  0.03282    0.03034   1.082    0.283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3628 on 66 degrees of freedom
## Multiple R-squared:  0.7003, Adjusted R-squared:  0.6776
## F-statistic: 30.84 on 5 and 66 DF,  p-value: 4.858e-16
```

We do not observe differences statistically significatives between the different levels of treatment (conservation methods). To be sure about the fact that the treatment is not significant, we compute the type III sums of squares.

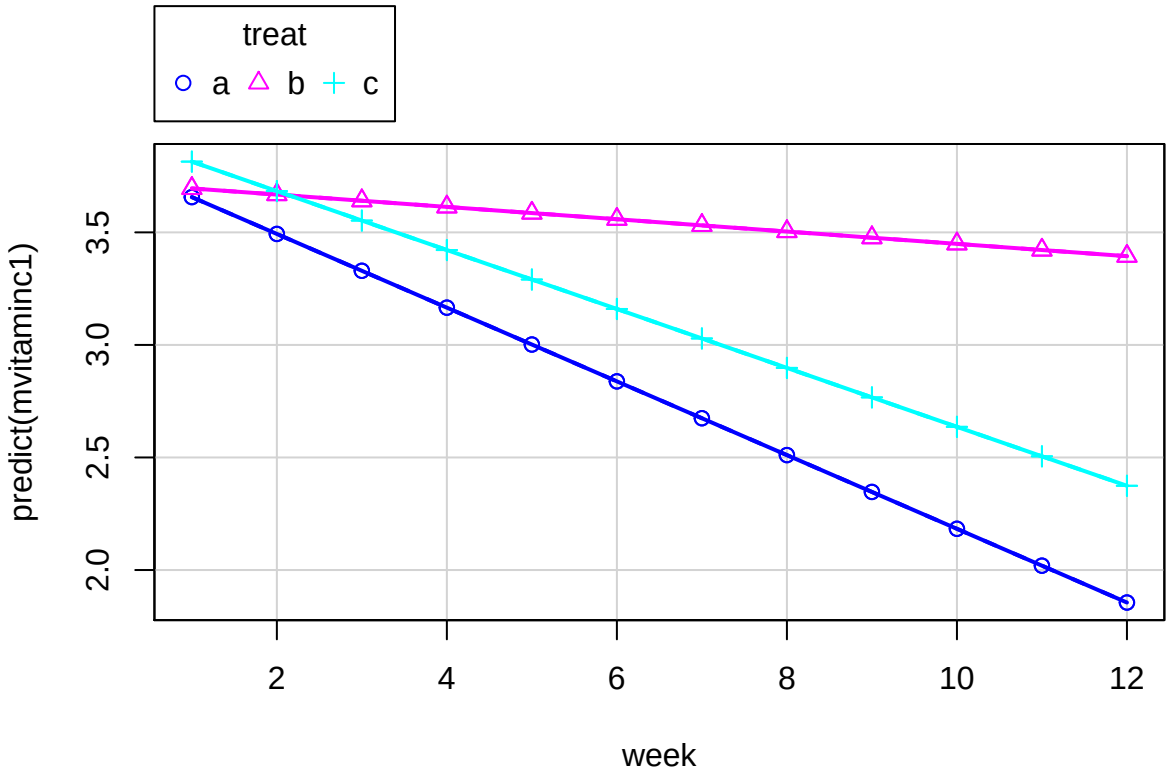
```
Anova(mvitaminc1,ty=3)
```

```
## Anova Table (Type III tests)
##
## Response: log(vitc)
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 77.063  1 585.5676 < 2.2e-16 ***
## week         7.667  1  58.2546 1.204e-10 ***
## treat        0.131  2   0.4992    0.6093
## week:treat   2.897  2  11.0081 7.488e-05 ***
## Residuals    8.686 66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The type III sums of squares ensure that the treatment is not significatively different from zero and thus, that we can remove it from the model. Important to know that sometimes if the interaction is significant and one of the main effects is not, one may prefer to leave in the model the main effect term of the not significant factor.

The model just fitted allows for different intercepts in the three groups, Thus the predicted value in the zero week (initial moment) will be different. This is appreciate in the following scatterplot:

```
scatterplot(predict(mvitaminc1)~week|treat,dat=vitamincdata)
```



In what follows we estimate the marginal means (emm) and we compare them in pairs using the Tukey method, at week zero. To do that at week zero is very important, because it will allow us to conclude if at the initial moment, all the orange juices had the same vitaminc level.

```
emmt<-emmeans(mvitaminc1,~treat|week,at=list(week=c(0)))
print(pairs(emmt))
```

```
## week = 0:
##  contrast    estimate        SE df t.ratio p.value
##  a - b     0.09784614 0.2232716 66   0.438  0.8997
##  a - c    -0.12471756 0.2232716 66  -0.559  0.8424
##  b - c    -0.22256370 0.2232716 66  -0.997  0.5815
##
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 3 estimates
```

We see that the means are not statistically different from zero at week zero. This allows us to say the the vitaminC level at the initial point (week zero) is the same for all conservation methods, and it is estimated by the model intercept 3.82038. Observe that in the case where two conservation methods differ in the vitaminC level at the origin, then we could not be able to ensure if the differences founded in the lose of vitaminC between two conservation methods were due to the conservation method, or simply a consequence of the fact that the started with different vitaminC levels

Multiple comparison of the three slopes

```
emmm<-emtrends(mvitaminc1,~treat,var="week")
print(pairs(emmm))
```

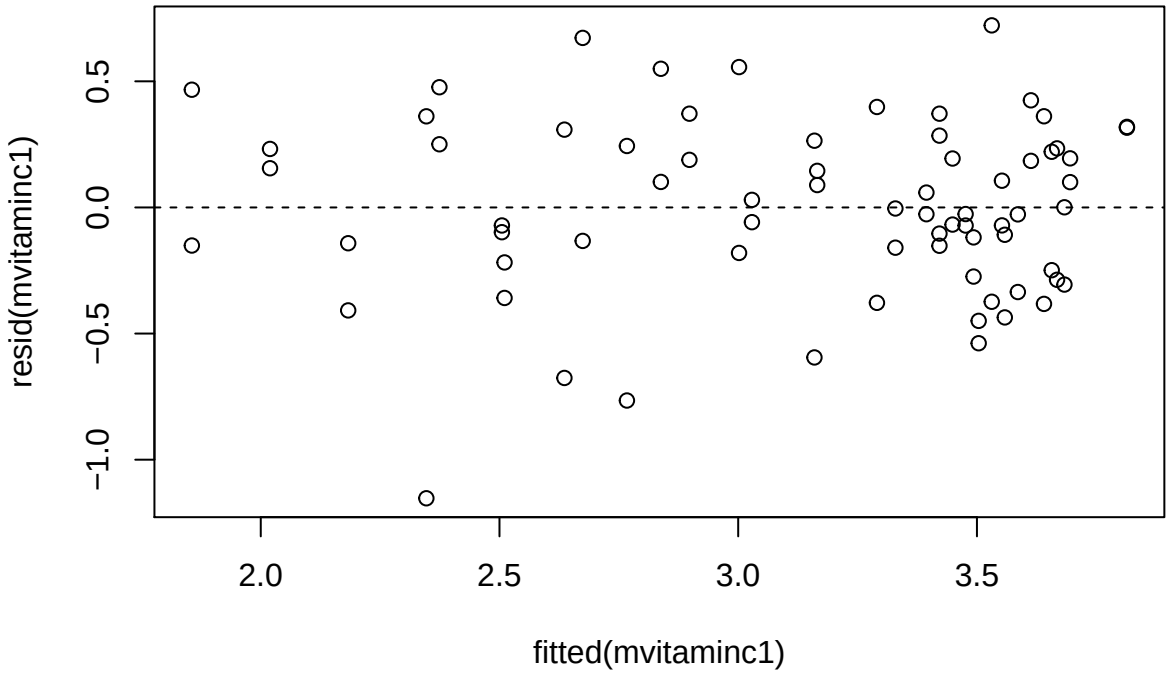
```
##  contrast    estimate         SE df t.ratio p.value
```

```
##  a - b    -0.13636085 0.03033663 66  -4.495  0.0001
##  a - c    -0.03281687 0.03033663 66  -1.082  0.5287
##  b - c     0.10354399 0.03033663 66   3.413  0.0031
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

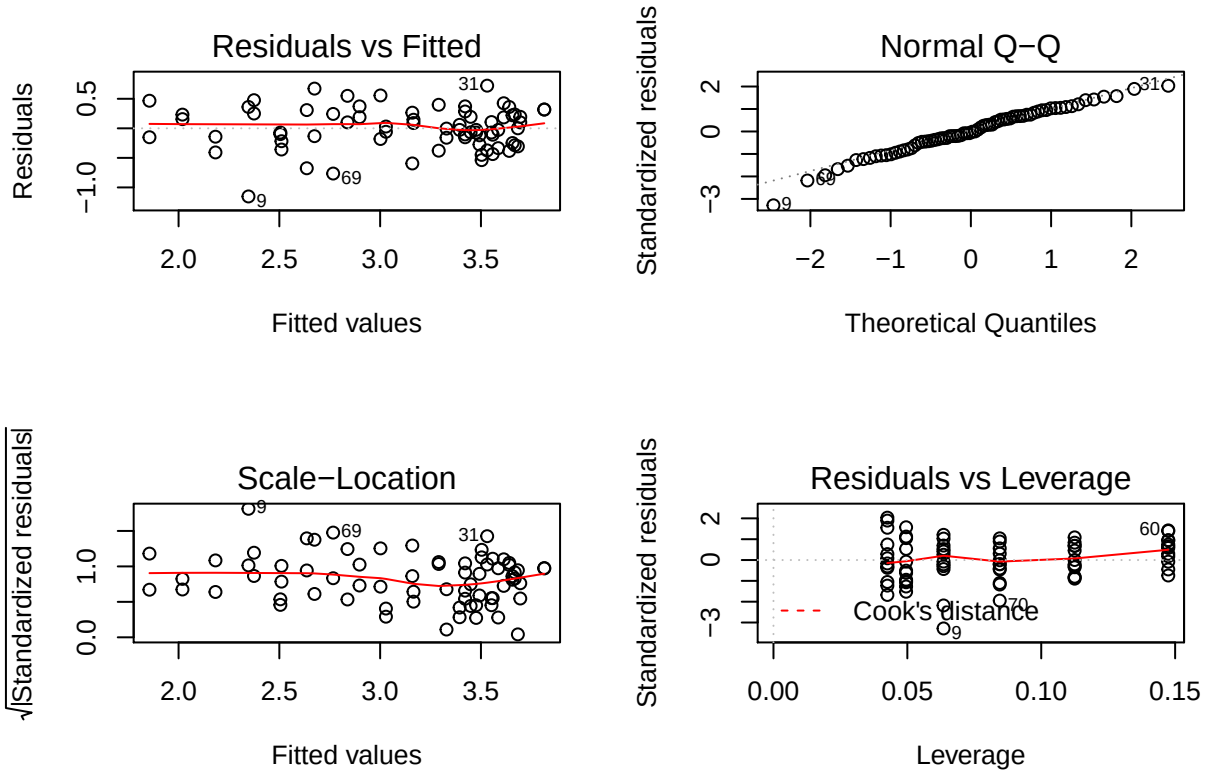
Slopes of threatments a and c are not statistically different while the other pairs are not.

Residual analysis of the first model

```
plot(fitted(mvitaminc1),resid(mvitaminc1))
abline(h=0,lty=2)
```




```
oldpar<-par(mfrow=c(2,2))
plot(mvitaminc1,ask=F)
```




```
par(oldpar)
```

We can accept the normality, independence and homocedasticity properties of the errors.

Model with same intercepts and different slopes

The second model we fit is the one without the treatment (conservation method) as main effect.

```
mvitaminc2<-lm(log(vitc)~week+treat:week,vitamincdata)

summary(mvitaminc2)
```

```
##
## Call:
## lm(formula = log(vitc) ~ week + treat:week, data = vitamincdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15221 -0.19240  0.00464  0.23452  0.70470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.82934    0.09048  42.324  < 2e-16 ***
## week        -0.16480    0.01475 -11.171  < 2e-16 ***
## week:treatb  0.12462    0.01412   8.823 7.04e-13 ***
## week:treatc  0.04778    0.01412   3.383  0.00119 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3601 on 68 degrees of freedom
## Multiple R-squared:  0.6957, Adjusted R-squared:  0.6823
## F-statistic: 51.83 on 3 and 68 DF,  p-value: < 2.2e-16
```

Now we clearly see that the week and the interaction are clearly significant.

The week coefficient is equal to -0.1648 which may be interpreted as the decrese in log(vitaminC) by increasing one unit the week if the orange juice comes from hte conservation method a. Thus, if we denote by $vitaminC$ the level of vitaminC in a given week of an orange juice of conservation method a, and by $vitaminc^*$ the corresponding level one week later, we have that

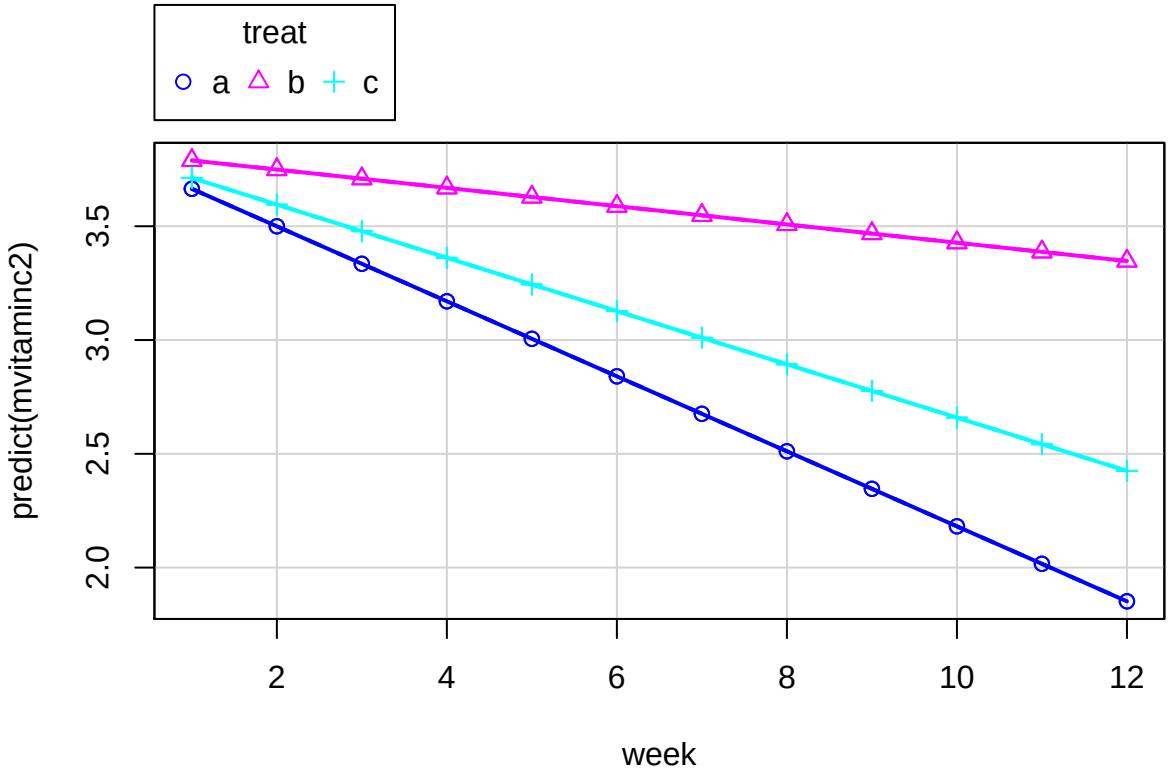
$$VitaminC^* = e^{-0.1648} * vitaminC \text{ if orange juice follows the conservation method } a$$

The decrease in log(vitaminC) for an orange juice of conservation methods b and c will be estimated by $-0.1648 + 0.1246 = -0.04$ and $-0.1648 + 0.04778 = -0.117$ respectively. From where one has that:

$$VitaminC^* = e^{-0.1648+0.1246} * vitaminC \text{ if orange juice follows the conservation method } b$$
$$VitaminC^* = e^{-0.1648+0.04778} * vitaminC \text{ if orange juice follows the conservation method } c$$

Next it appears the scaterplot of the predicted values as a function of the week for the three conservation methods.

```
scatterplot(predict(mvitaminc2)~week|treat,dat=vitamincdata)
```



Important to know if the slopes of the predicted models are statistically different. The slopes correspond to the estimated trends of the model.

```
emmm<-emtrends(mvitaminc2,~treat,var="week")
pairs(emmm)
```

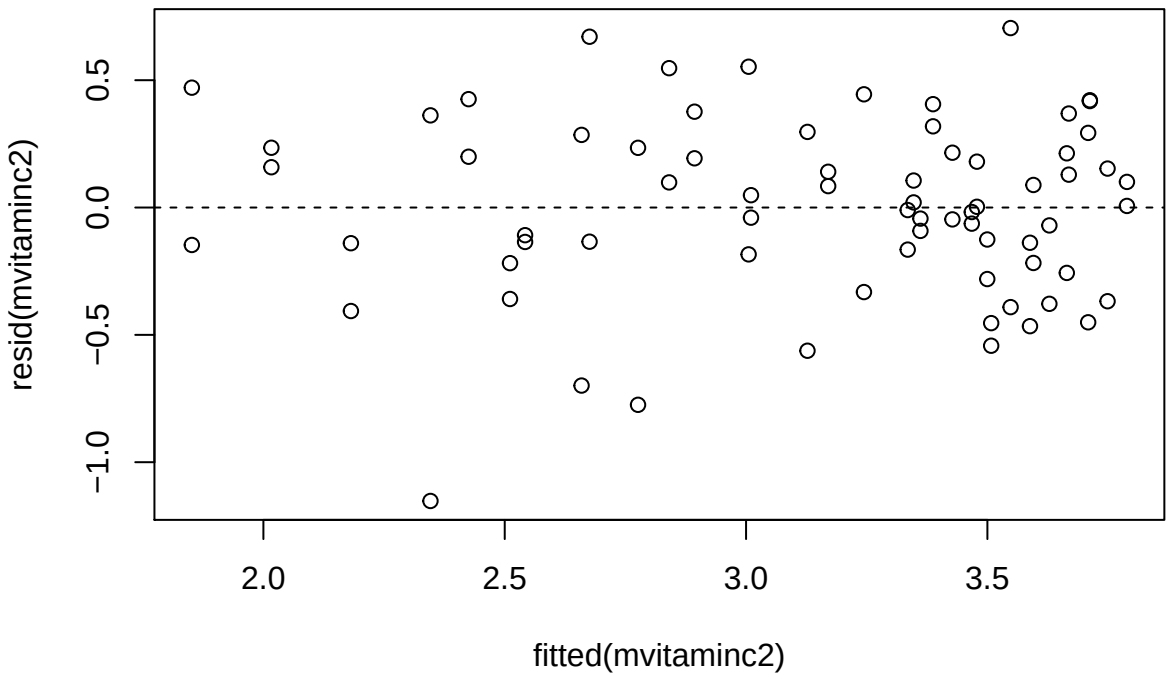
```
##  contrast    estimate         SE df t.ratio p.value
##  a - b    -0.12461932 0.01412397 68  -8.823  <.0001
##  a - c    -0.04778297 0.01412397 68  -3.383  0.0034
##  b - c     0.07683634 0.01412397 68   5.440  <.0001
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

For each treatment we obtain the slope, its standard deviation and the corresponding confidence interval. Observe that the slope stimation for treatment a corresponds to the coefficient of the week in the model. And the other two estimations correspond to the values that we have computed before.

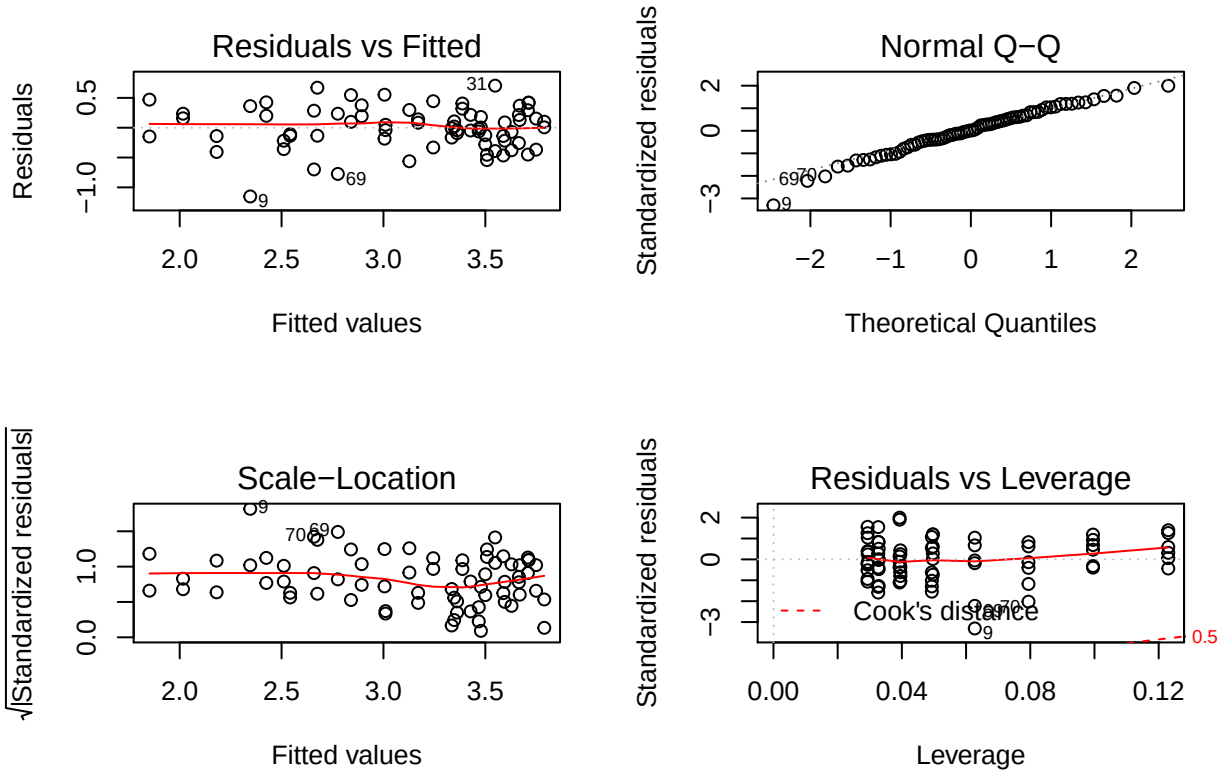
With the sentence pairs, we perform the two by two comparison of the slopes. The consequence is to reject all the null hypothesis and to conclude that the slopes between conservation methods are statistically different.

Residual analysis

```
plot(fitted(mvitaminc2),resid(mvitaminc2))
abline(h=0,lty=2)
```




```
oldpar<-par(mfrow=c(2,2))
plot(mvitaminc2,ask=F)
```




```
par(oldpar)
```

Again the residual analysis allows us to accpet the linear model assumptions, and to conclude that this secon model is also satisfactory.

In orther to choose one of the two models, we can use the adjusted R^2. As it can be seen, the adjusted R^2 is a little bit larger for the second model, and thus we consider the second model as the more appropiate one.