# A love story of LM and GLM: Vitamin C in Orange juice

One wants to compare the evolution in time of the Vitamin C level of an orange juice, as a function of: the type of container and the conservation temperature. To that end, three conservation methods where considered: "a", "b" and "c".

For each conservation method, and during 12 weeks, two units of orange juice where analyzed. The structure of the dataset is as follows: the first column corresponds to the Treatment: conservation method, second column corresponds to Week: and it indicates the time after packaging, the third column indicates corresponds to VitC: level of vitamin C that has been observed.

```
library(car)
library(tables)
library(emmeans)
dd <- read.csv2("vitc.csv")
head(dd)
```

```
##   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
```

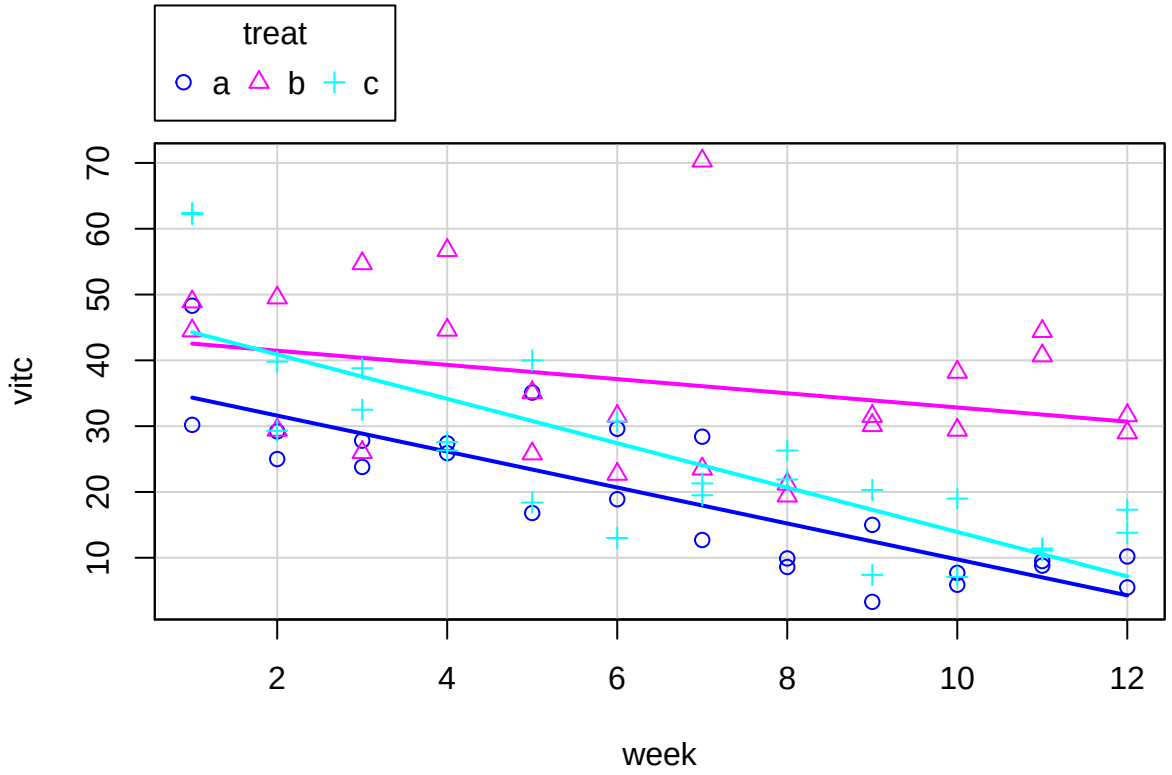
It is supposed that the Vitamin C level evolves following the exponential function:

$$VitC = \alpha_i e^{-\beta_i \cdot Week},$$

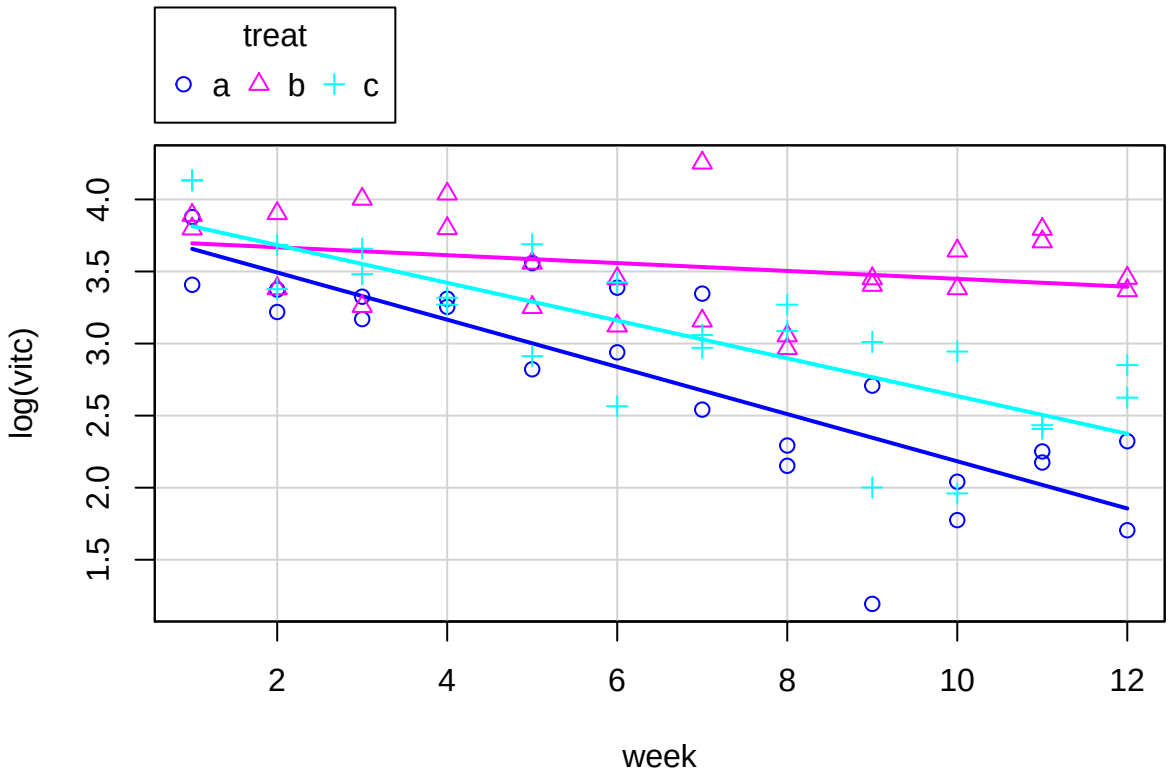
with $\alpha_i > 0$ and $\beta_i > 0$, and that these parameters may depend on the conservation method, indicated by the subscript $i$. Assuming that in the moment of packaging may exists differences between the levels of Vitamin C, and using a significance level equal to 5%, answer the following questions:

(a) Define a generalized linear model with the "gamma" family, use it to check whether the treatments lose Vitamin C at the same velocity, that is if $\beta_1 = \beta_2 = \beta_3$ or not, and also to see if the three values of $\alpha_i$ are or are not statistically equivalent. From this model, estimate $\alpha_i$. Are they statistically different? Estimate $\beta_i$. Are they statistically different?

```
sp(vitc~week|treat, smooth=F, data=dd)
```




```
scatterplot(log(vitc)~week|treat,smooth=F,data=dd)
```




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

## Model with different intercepts and 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 first model we fit contains the main effects as well as the interaction term.

```
summary(model.lm<-lm(log(vitc)~treat*week, data=dd))
```

```
##
## Call:
## lm(formula = log(vitc) ~ treat * week, data = dd)
##
## 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 ***
## treatb      -0.09785    0.22327  -0.438    0.663
## treatc       0.12472    0.22327   0.559    0.578
## week        -0.16373    0.02145  -7.632 1.20e-10 ***
## treatb:week  0.13636    0.03034   4.495 2.88e-05 ***
## treatc:week  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.

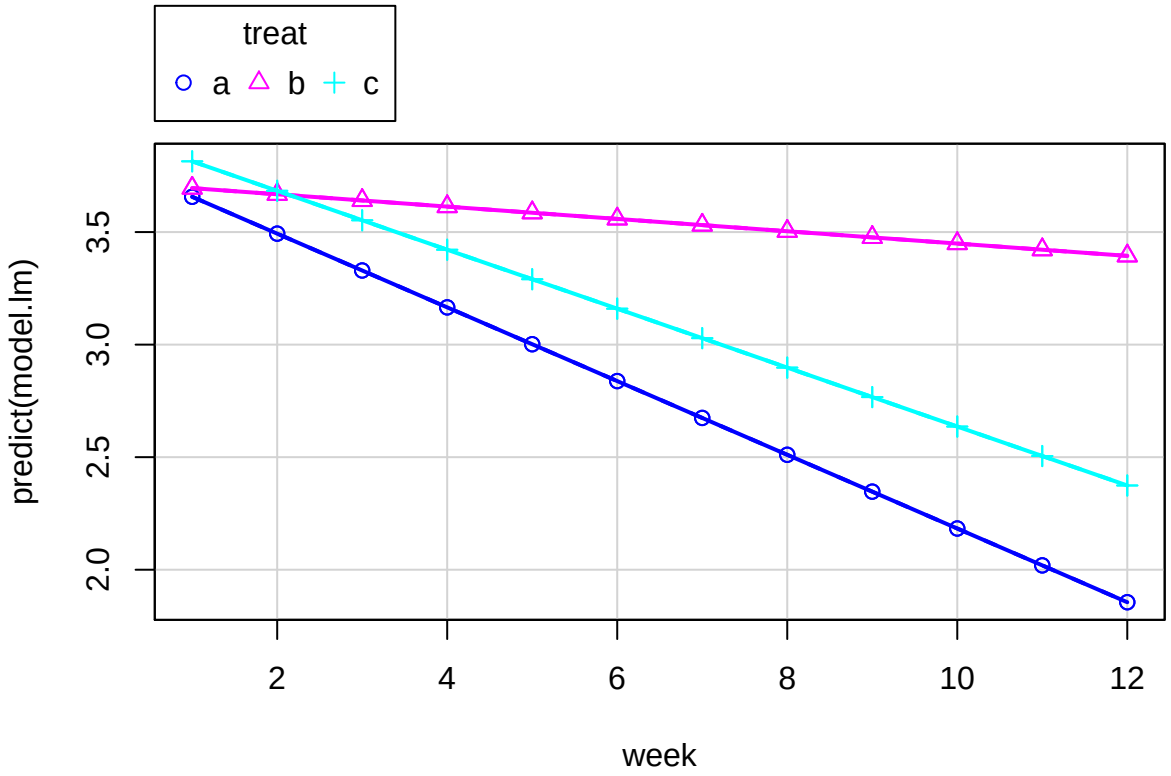
```
Anova(model.lm,ty=3)
```

```
## Anova Table (Type III tests)
##
## Response: log(vitc)
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 77.063  1 585.5676 < 2.2e-16 ***
## treat        0.131  2   0.4992    0.6093
## week         7.667  1  58.2546 1.204e-10 ***
## treat:week   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 ensures that the treatment is not significatively different from zero and thus, 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 just fitted model allows different intercepts for the three groups, Thus the predicted value in the zero week (initial moment) will be different. This is appreciated in the following scatterplot:

```
scatterplot(predict(model.lm)~week|treat,dat=dd)
```




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(model.lm,~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 found in the lose of vitaminC between two conservation methods were due to the conservation method, or simply a consequence of the fact that we started with different vitaminC levels.

Multiple comparison of the three slopes:

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

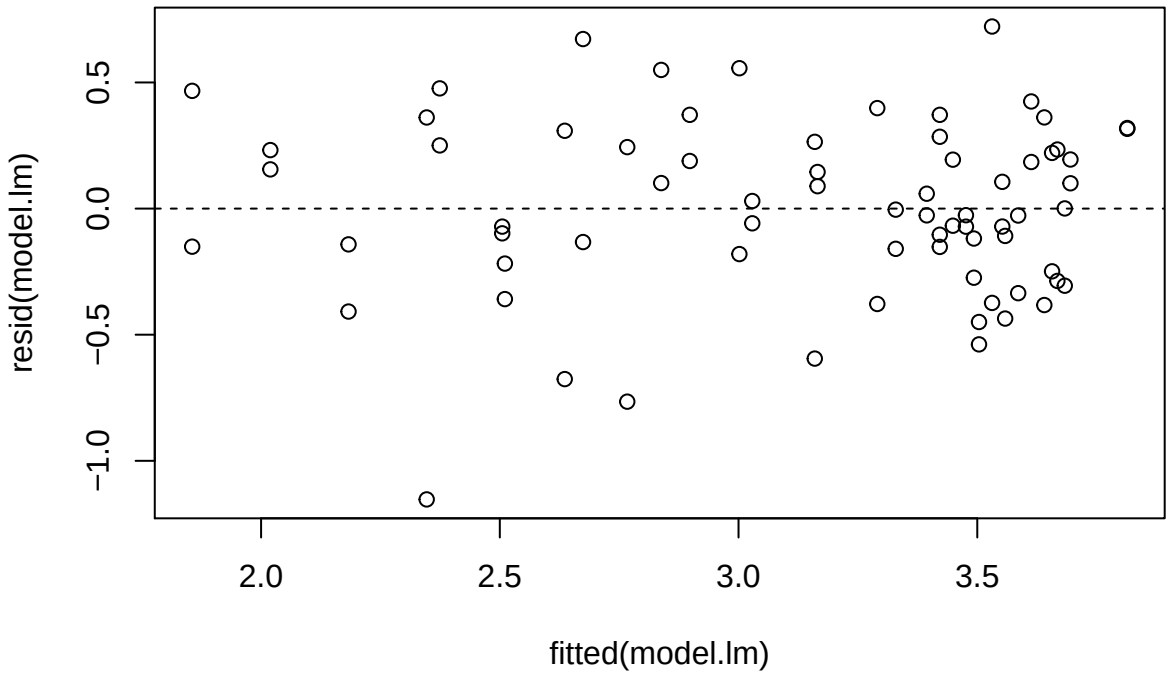
```
##  contrast    estimate         SE df t.ratio p.value
##  a - b    -0.13636085 0.03033663 66  -4.495  0.0001
```

```
##  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
```

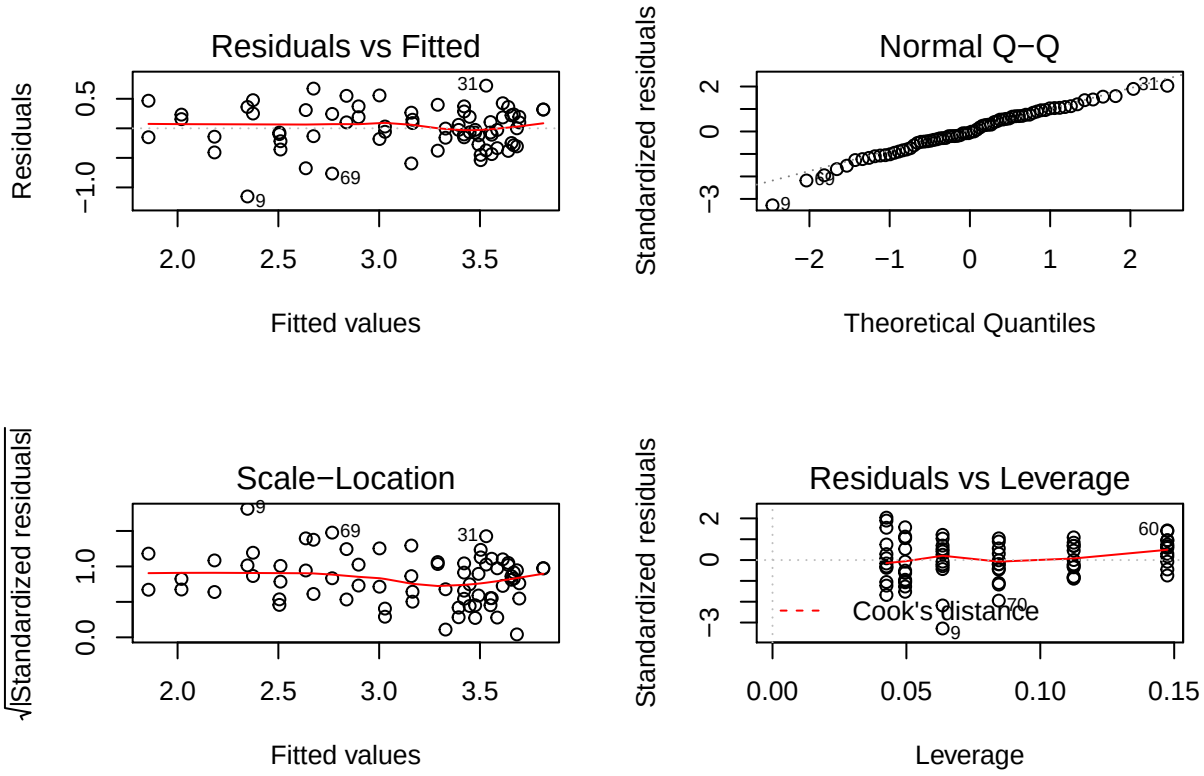
Slopes of threatments a and c are not statistically different while the other pairs are not.

### Residual analysis of the first model

```
plot(fitted(model.lm),resid(model.lm))
abline(h=0,lty=2)
```




```
oldpar<-par(mfrow=c(2,2))
plot(model.lm,ask=F)
```




```
par(oldpar)
```

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

### Model with the same intercepts and different slopes

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

```
summary(model.lm2<-lm(log(vitc)~week+treat:week, data=dd))
```

```
##
## Call:
## lm(formula = log(vitc) ~ week + treat:week, data = dd)
##
## 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:

$$VitaminC* = e^{-0.1648} \cdot VitaminC$$

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:

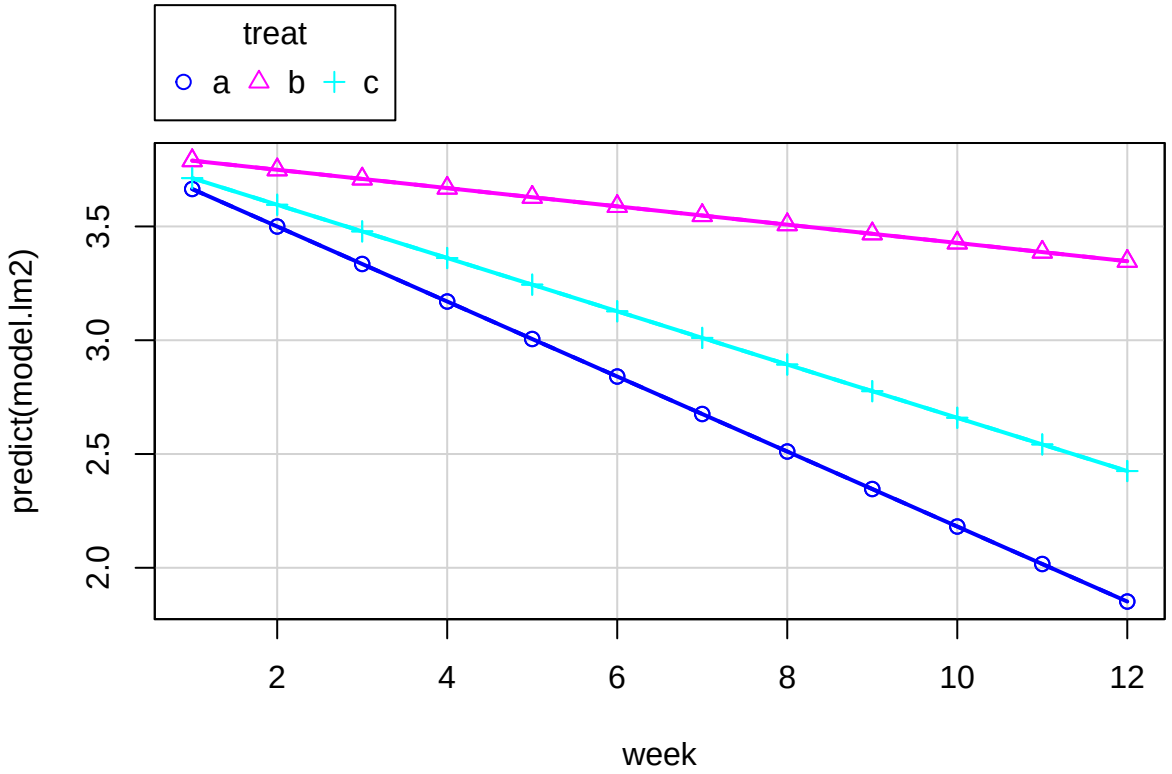
$$VitaminC* = e^{-0.1648+0.1246} \cdot VitaminC$$

if orange juice follows the conservation method b and:

$$VitaminC* = e^{-0.1648+0.04778} \cdot VitaminC$$

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(model.lm2)~week|treat,dat=dd)
```




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(model.lm2,~treat,var="week")
pairs(emmm)
```

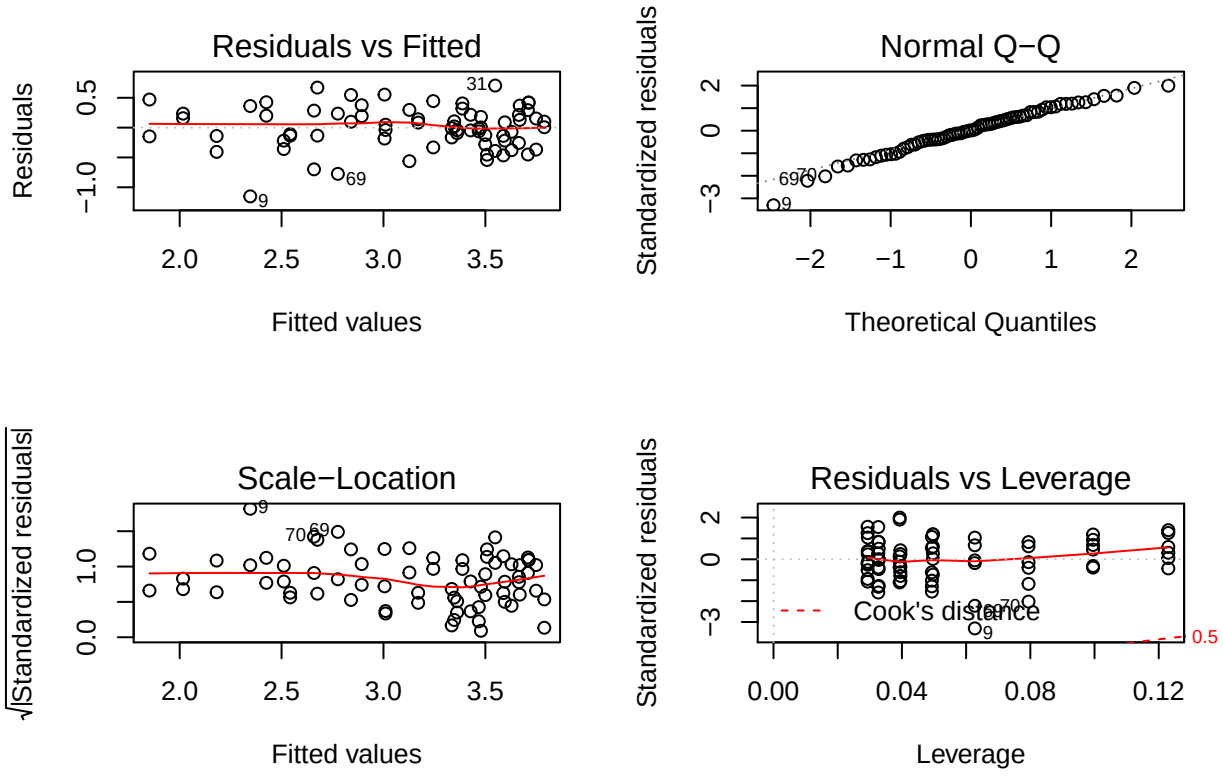
```
##  contrast    estimate         SE df t.ratio p.value
##  a - b    -0.12461932 0.01412397 68  -8.823  <.0001
##  a - c    -0.04778297 0.01412397 68  -3.383  0.0034
##  b - c     0.07683634 0.01412397 68   5.440  <.0001
##
## P value adjustment: tukey method for comparing a family of 3 estimates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

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 of the second model**

```
oldpar<-par(mfrow=c(2,2))
plot(model.lm2,ask=F)
```




```
par(oldpar)
```

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

In order 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 in the second model, thus, we consider the second model as the more appropiate one.

## GLM Gamma model

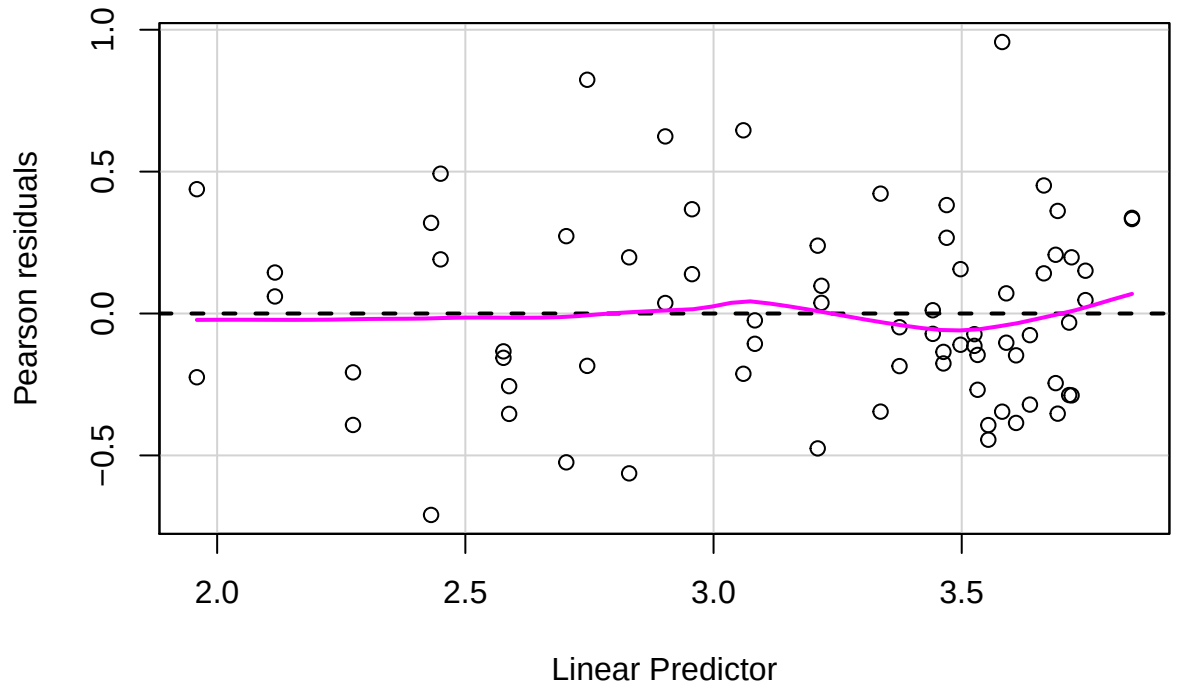
Quin link fem servir? EL logaritme, la part lineal està a l'exponent i per tant si volem aïllar la part lineal hauriem d'utilitzar el link log. El tractament és l'efecte que podriem eliminar si dóna el mateix en els tres grups. Però a la descriptiva es veu que hi ha diferències. Per això, ho inclourem.

```
summary(model<-glm(vitc~treat*week, family=Gamma(link="log"), data=dd))
```

```
##
## Call:
## glm(formula = vitc ~ treat * week, family = Gamma(link = "log"),
##     data = dd)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.02691  -0.25019  -0.06136   0.18828   0.75558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.84653    0.14949  25.731  < 2e-16 ***
## treatb      -0.06931    0.21141  -0.328    0.744
## treatc       0.12314    0.21141   0.582    0.562
## week        -0.15729    0.02031  -7.744 7.61e-11 ***
## treatb:week  0.12933    0.02872   4.502 2.80e-05 ***
## treatc:week  0.03066    0.02872   1.067    0.290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1179906)
##
##     Null deviance: 24.790  on 71  degrees of freedom
## Residual deviance:  8.217  on 66  degrees of freedom
## AIC: 514.38
##
## Number of Fisher Scoring iterations: 5
```

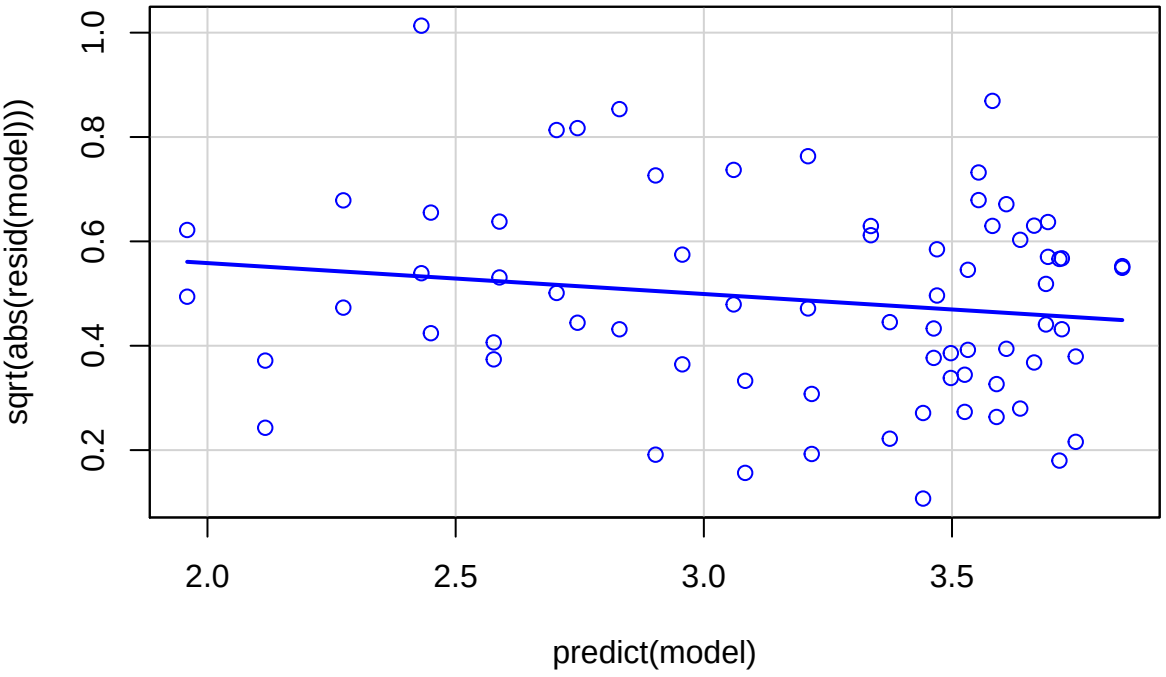
Hem de fer els diagnòstics i l'estudi dels efectes. Una visió descriptiva del models:

```
residualPlot(model, ty="pearson")
```



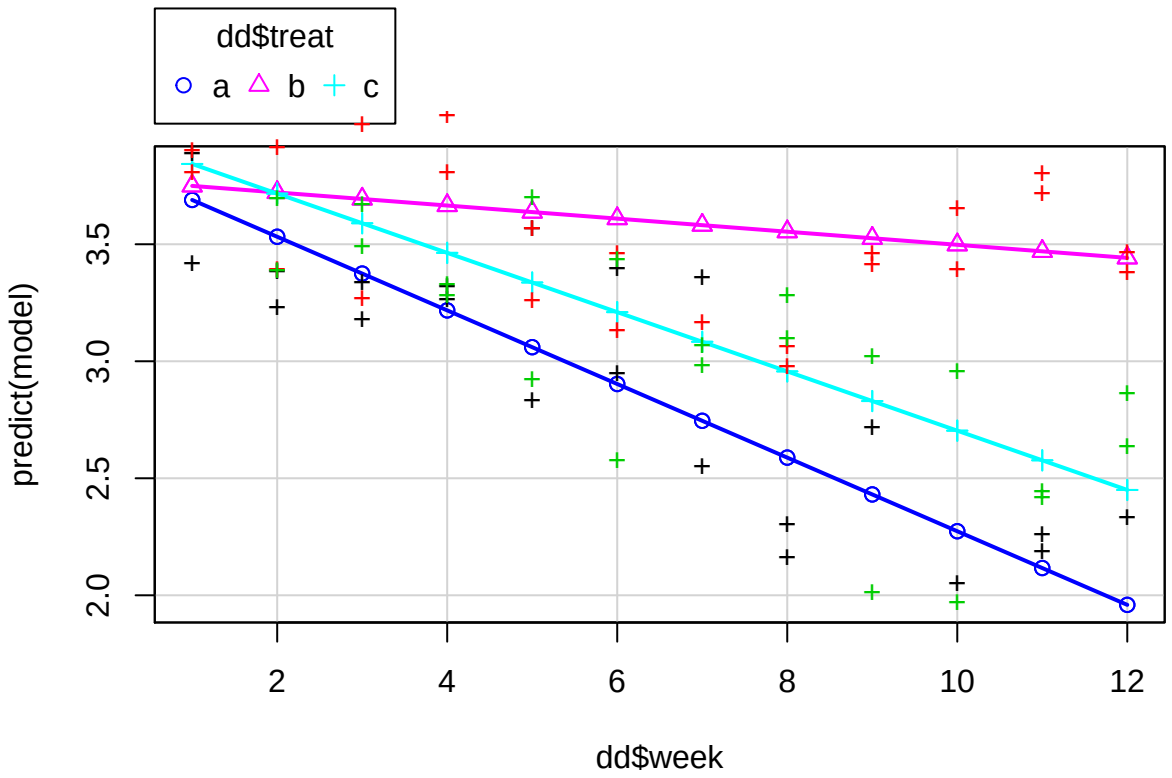

És igual si agafem el predictor lineal o la Vitamina C predita, o si agafem els residuals de Pearson o els de la Deviància. Han d'anar al voltant de zero i no hem de veure patrons. La concentració que veiem és simplement una qüestions de com han sortit aquestes dades. No veiem cap patró, la gràfica és acceptable. Podem tenir idea si les variàncies van canviant amb la següent gràfica:

```
scatterplot(predict(model), sqrt(abs(resid(model))), smooth=F, boxplot=F)
```




Les variàncies van disminuint una mica.

```
sp(predict(model)~dd$week|dd$treat, smooth=F)
points(dd$week, log(dd$vitc), pch="+", col=dd$treat)
```




Suposem que acceptem el model i que els diagnòstics són acceptables. EL primer pas és fer le test Anova (II).

```
Anova(model, test.statistic="F")
```

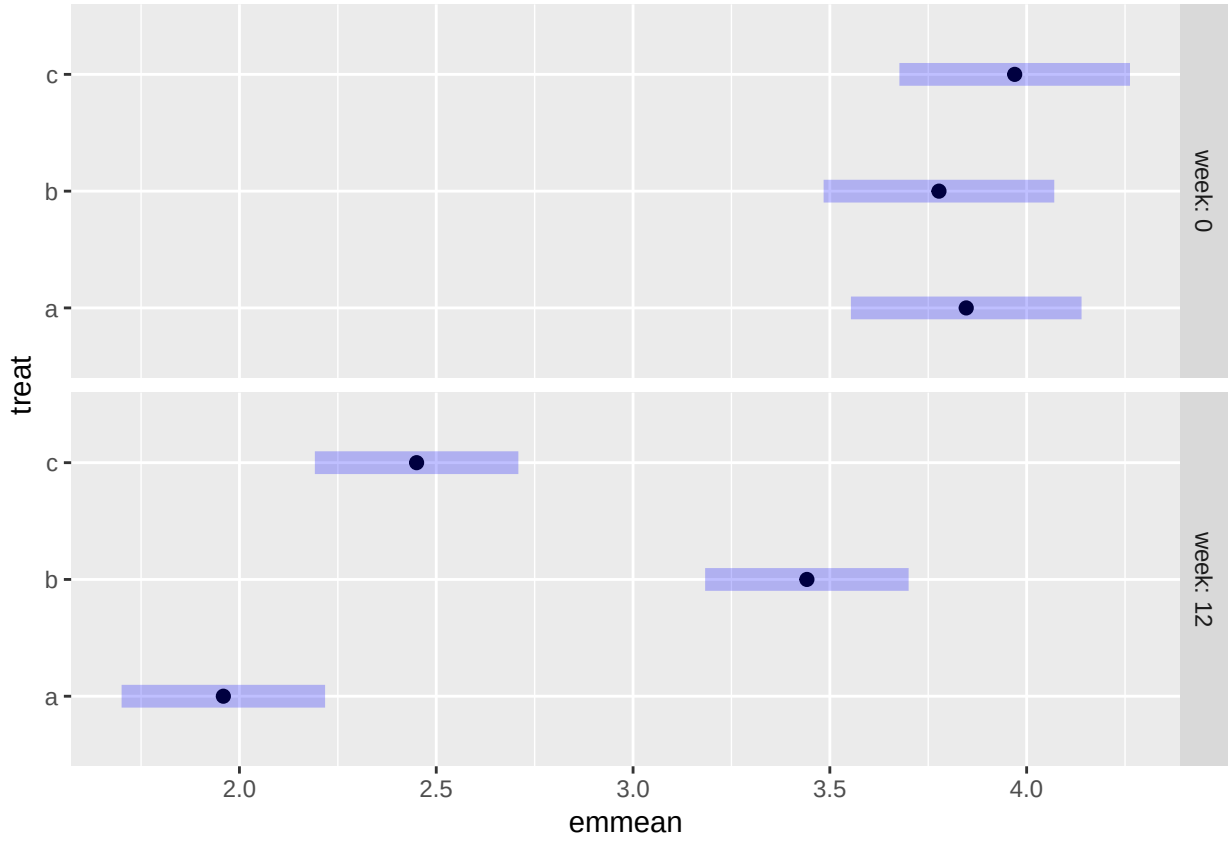
```
## Analysis of Deviance Table (Type II tests)
##
## Response: vitc
## Error estimate based on Pearson residuals
##
##             Sum Sq Df F value    Pr(>F)
## treat       7.4194  2  31.441 2.563e-10 ***
## week        8.9695  1  76.019 1.375e-12 ***
## treat:week 2.6655  2  11.296 6.040e-05 ***
## Residuals  7.7874 66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
CLD(emmg<-emmeans(model, ~treat|week, at=list(week=c(0,12))), reversed = TRUE)
```

```
## week =  0:
##  treat   emmean        SE  df asymp.LCL asymp.UCL .group
##  c     3.969668 0.1494881 Inf  3.676677  4.262660  1
##  a     3.846529 0.1494881 Inf  3.553537  4.139520  1
##  b     3.777219 0.1494881 Inf  3.484228  4.070211  1
##
## week = 12:
##  treat   emmean        SE  df asymp.LCL asymp.UCL .group
##  b     3.441699 0.1318941 Inf  3.183192  3.700207  1
##  c     2.450102 0.1318941 Inf  2.191595  2.708610   2
##  a     1.959079 0.1318941 Inf  1.700571  2.217587    3
##
```

```
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
```

```
plot(emmg)
```




```
(emmt<-emtrends(model, ~treat, var="week"))
```

```
##  treat  week.trend         SE  df  asymp.LCL   asymp.UCL
##  a     -0.15728747 0.02031144 Inf -0.1970972 -0.11747778
##  b     -0.02796001 0.02031144 Inf -0.0677697  0.01184968
##  c     -0.12663049 0.02031144 Inf -0.1664402 -0.08682080
##
## Trends are based on the log (transformed) scale
## Confidence level used: 0.95
```

```
print(pairs(emmt))
```

```
##  contrast    estimate         SE  df z.ratio p.value
##  a - b    -0.12932746 0.02872471 Inf  -4.502  <.0001
##  a - c    -0.03065697 0.02872471 Inf  -1.067  0.5345
##  b - c     0.09867048 0.02872471 Inf   3.435  0.0017
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

The means are not statistically different on day 0, the slopes of treatment a and c are not statistically different. The one that differs is c.