Probability and Stadistics 2

Sonia Castro

18/1/2021

```
setwd("/Users/sonia/Desktop")
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default
```

```
library(survival)
library(Formula)
library(colorspace)
library(ggplot2)
library(carData)
library(Hmisc)
```

```
## Loading required package: lattice

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:EnvStats':
##
##     stripChart

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(car)
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:EnvStats':
##
##     qqPlot
```

```
library(tables)
library(lattice)
library(grid)
library(gridExtra)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:EnvStats':
##
##     boxcox
```

```
library(latticeExtra)
```

```
##
## Attaching package: 'latticeExtra'

## The following object is masked from 'package:ggplot2':
##
##     layer
```

```
library(RColorBrewer)
library(multcompView)
library(mvtnorm)
library(emmeans)
library(RcmdrMisc)
```

```
## Loading required package: sandwich

##
## Attaching package: 'RcmdrMisc'

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

```
library(multcomp)
```

```
## Loading required package: TH.data
```

```
##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
library(HH)
```

```
##
## Attaching package: 'HH'

## The following object is masked from 'package:emmeans':
##
##     as.glht

## The following objects are masked from 'package:car':
##
##     logit, vif
```

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. A combination of container and conservation temperature is denoted as conservation method. Three conservation methods have been considered and denoted by: "a", "b" y "c". For each conservation method, and during 12 consecutive weeks, the level of Vitamin C of two units of orange juice has been analyzed.

It is supposed that the Vitamin C level evolves following the exponential function: (hem canviat la lletra β de l'enunciat per γ per no confondrela amb les que utilitzem per definir els models més endevant.)

$$VitaminaC = \alpha_i e^{-\gamma_i \mathring{u} setmana},$$

with $\alpha_i > 0$ and $\gamma_i > 0$, and that these parameters may depend on the conservation method, indicated by the supscript i.

1)Perform the descriptive statistics of the dataset. Which are the main conclusions?

Per entendre millor les dades aplicarem log a la fórmula de la Vitamina C de forma que ens quedi: $log(VitaminaC) = log(\alpha_i) - \gamma_i setmana$.

I afegirem la variable Tipus_tractament com a factor.

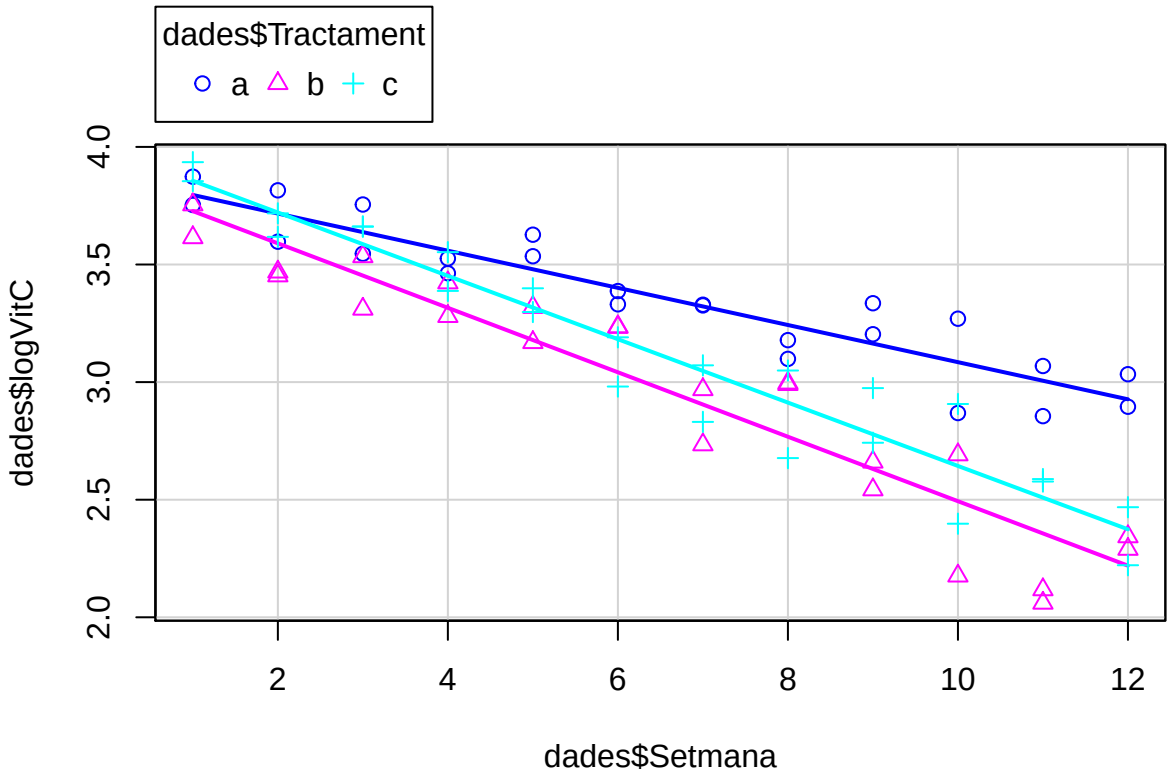
```
dades<-read.csv2("/Users/sonia/Desktop/ViTCGroup26.csv")
head(dades)
```

```
##   Tractament Setmana VitaminaC
## 1          a       1     42.66
## 2          a       2     45.40
## 3          a       3     42.74
## 4          a       4     31.91
## 5          a       5     37.59
## 6          a       6     27.96
```

```
#añadimos variable categòrica como factor
dades$Tipus_tractament<-as.factor(dades$Tractament)
#canviem la variabe VitaminaC pel seu logarítme
dades$logVitC<-log(dades$VitaminaC)
head(dades)
```

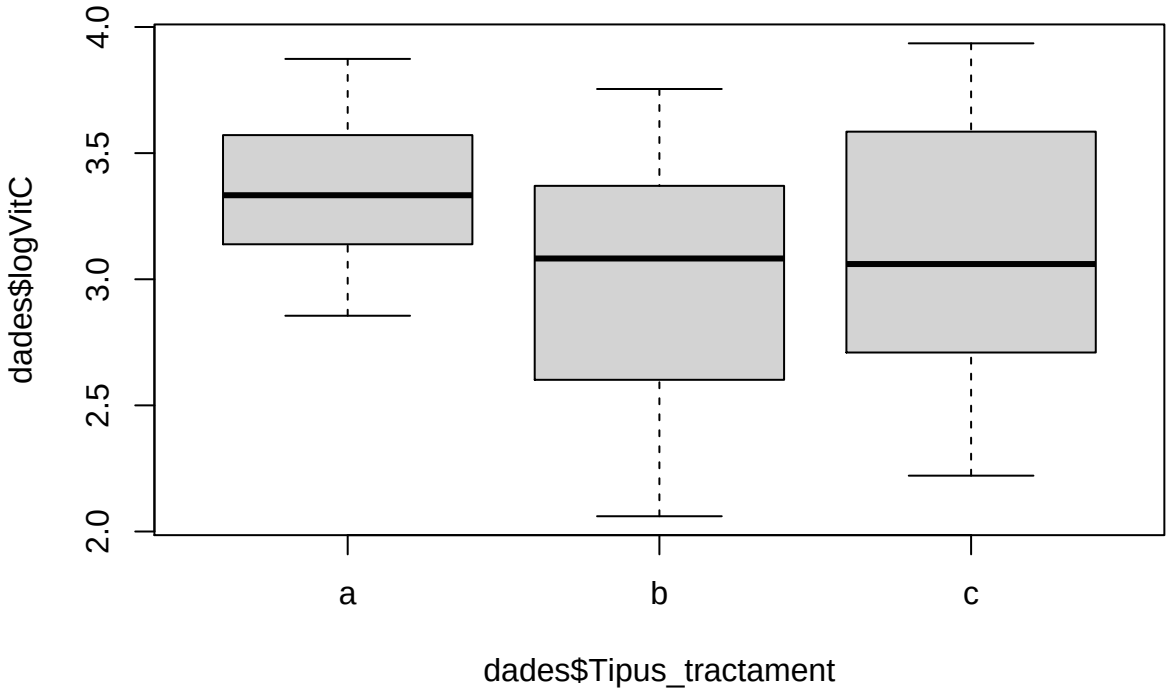
```
##   Tractament Setmana VitaminaC Tipus_tractament  logVitC
## 1          a       1     42.66                 a 3.753262
## 2          a       2     45.40                 a 3.815512
## 3          a       3     42.74                 a 3.755135
## 4          a       4     31.91                 a 3.462919
## 5          a       5     37.59                 a 3.626738
## 6          a       6     27.96                 a 3.330775
```

```
scatterplot(dades$logVitC~(dades$Setmana)|dades$Tractament,smooth=F,data=dades)
```



D'aquest plot es dedueix que entre els tractaments 'b' i 'c' no hi ha interacció, ja que les rectes són paral · leles(tenen el mateix pendent). En canvi si que hi ha interacció amb el model 'a', ja que la pendent és menor. Veiem com el pas de les setmanes sempre disminueix el nivell de Vitamina C. S'aconsegueix mantenir un nivell de Vitamina més alt a la llarga amb el tractament 'a'.

```
boxplot(dades$logVitC~dades$Tipus_tractament)
```



Sembla que no hi ha molta diferència entre els nivells de VitaminaC amb els tractaments 'b' i 'c'. Però si amb el nivell de Vitamina C del tractament 'a', que sembla que té una esperança superior i una variabilitat menor. En els tractamnets 'b' i 'c' semblen tenir una variancia semblant tot i que en el tractament b sembla que sigui més probable trobar-se per sota de l'esperança que per sobre i a l'inrevés en el c.

2) If you want to fit a linear model to the data, which is the reasonable response variable? Which are the explanatory variables? Which is the type of the explanatory variables?

Per crear un model lineal amb les dades aplicarem log a la fórmula de la Vitamina C de forma que ens quedi:

$$log(VitaminaC) = log(\alpha_i) - \gamma_i setmana$$

De forma que la nostra variable resposta i és el logaritme del nivell de Vitamina C i les variables explicatòries són: el tipus de tractament aplicat (α_i i γ_i depenen d'aquest) i la setmana en la que es mesura. El tractament és una variable categòrica de 3 nivells i la setmana és una variable numèrica.

3) Specify the questions that has sense to answer by means of analyzing the data (at least three).

- Té alguna influència el tipus de tractament en el nivell de Vitamina C mesurat?
- Té alguna influència la setmana en que es mesura en el nivell de Vitamina C?
- Es possible que l'efecte de la setmana en que es pren la mesura tipus de tractament en el nivell de Vitamina C canvii depenent del tipus de tractament?
- Quin tractament manté un nivell de Vitamina C més alt a la llarga?

4) Assuming that at the moment of packaging all the juices had the same Vitamin C level, define a linear model to see if the three conservation methods lose the vitamin C in a similar way. That is if statistically $\gamma_1 = \gamma_2 = \gamma_3$. From this model,

- **Compute the γ_i estimations.**
- **Are the three γ_i statically different or not?**

Tenim que $log(VitaminaC) = log(\alpha_i) - \gamma_i setmana$, i definim el model lineal ANCOVA amb interacció entre les variables tractament i setmana com:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 setmana + (\tau_i \sim \beta_1) setmana + e_{ij}$$

Tenint en compte que agafem com a baseline el tractament 'a' i per tant $\tau_0 = 0$.

Tenim $y_{ij} = \beta_0 + \tau_i + (\beta_1 + (\tau_i \sim \beta_1)) setmana + e_{ij}$, Per tant a l'hora d'interpretar-ho i relacionar-ho amb la fórmula $log(VitaminaC) = log(\alpha_i) - \gamma_i setmana$:

y_{ij} és el log(VitaminaC) de la mostra j del tractament i. Llavors $\beta_0 + \tau_i$ correspon a $log(\alpha_i)$ i $\beta_1 + (\tau_i \sim \beta_1)$ correspon a $-\gamma_i$.

Hem de veure si amb els tres tractaments es perd la Vitamina C de forma similar, és a dir, si no hi ha interacció (el pas de les setmanes afecta igual en qualsevol dels tractaments i les rectes són paral · leles).

En el nostre cas si no hi ha interacció $y_{ij} = \beta_0 + \tau_i + \beta_1 setmana + e_{ij}$ i per tant $-\gamma_i = \beta_1, \forall i$, com diu l'enunciat serien totes iguals.

Primer comparem els tractaments b i c, ja que sembla pel que hem vist a l'estadística descriptiva que la interacció será insignificativa:

```
m1 <- lm(logVitC~Tipus_tractament + Setmana + Tipus_tractament:Setmana, dades[dades$Tipus_tractament!=
summary(m1)
```

```
##
## Call:
## lm(formula = logVitC ~ Tipus_tractament + Setmana + Tipus_tractament:Setmana,
##     data = dades[dades$Tipus_tractament != "a", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31777 -0.11481  0.02541  0.09569  0.26339
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.864061   0.067338   57.38   <2e-16 ***
## Tipus_tractamentc          0.126613   0.095231    1.33    0.191
## Setmana                   -0.137040   0.009149  -14.98   <2e-16 ***
## Tipus_tractamentc:Setmana  0.002324   0.012939    0.18    0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1547 on 44 degrees of freedom
## Multiple R-squared:  0.9111, Adjusted R-squared:  0.9051
## F-statistic: 150.4 on 3 and 44 DF,  p-value: < 2.2e-16
```

```
Anova(m1,ty=3)
```

```
## Anova Table (Type III tests)
##
## Response: logVitC
##                           Sum Sq Df   F value Pr(>F)
## (Intercept)               78.836  1 3292.7959 <2e-16 ***
## Tipus_tractament           0.042  1    1.7677 0.1905
## Setmana                    5.371  1  224.3394 <2e-16 ***
## Tipus_tractament:Setmana   0.001  1    0.0323 0.8583
## Residuals                  1.053 44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tant mirant el summary com la taula Anova ens surt que la interacció i el tractament no són significants, probablement el tractament no ho és pel terme d'interacció. Si la interacció no és significant i l'estimador ($\tau_i \sim \beta_1$) és 0, això vol dir que el pas de les setmanes afecta igual als tractaments b i c.

Tornem a fer el model però sense interacció:

```
m2 <- lm(logVitC~Tipus_tractament + Setmana, dades[dades$Tipus_tractament!= 'a',])
summary(m2)
```

```
##
## Call:
## lm(formula = logVitC ~ Tipus_tractament + Setmana, data = dades[dades$Tipus_tractament !=
##     "a", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32184 -0.11067  0.02544  0.09843  0.26746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.85651    0.05202  74.129  < 2e-16 ***
## Tipus_tractamentc  0.14172    0.04418   3.207  0.00247 **
## Setmana           -0.13588    0.00640 -21.232  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1531 on 45 degrees of freedom
## Multiple R-squared:  0.9111, Adjusted R-squared:  0.9071
## F-statistic: 230.5 on 2 and 45 DF,  p-value: < 2.2e-16
```

Ara ens surt que amdues variables explicatòries són sifnificants i l'estimador del $\gamma_i = -\beta_1 = 0.13588$ $i \in \{2, 3\}$. I són estadísticament iguals.

Ara comparem els 3 tractaments per veure si la interacció també és insignificant quan també tenim en compte le tractament 'a'.

```
m3 <- lm(logVitC~Tipus_tractament + Setmana + Tipus_tractament:Setmana, dades)
summary(m3)
```

```
##
## Call:
## lm(formula = logVitC ~ Tipus_tractament + Setmana + Tipus_tractament:Setmana,
##     data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31777 -0.09776  0.00765  0.09522  0.26339
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.874678   0.061628  62.872  < 2e-16 ***
## Tipus_tractamentb           -0.010617   0.087155  -0.122    0.903
## Tipus_tractamentc            0.115996   0.087155   1.331    0.188
## Setmana                     -0.078986   0.008374  -9.433 7.46e-14 ***
## Tipus_tractamentb:Setmana -0.058055   0.011842  -4.902 6.47e-06 ***
## Tipus_tractamentc:Setmana -0.055731   0.011842  -4.706 1.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1416 on 66 degrees of freedom
## Multiple R-squared:  0.9147, Adjusted R-squared:  0.9083
## F-statistic: 141.6 on 5 and 66 DF,  p-value: < 2.2e-16
```

```
Anova(m3,ty=3)
```

```
## Anova Table (Type III tests)
##
## Response: logVitC
##                            Sum Sq Df   F value    Pr(>F)
## (Intercept)                79.269  1 3952.9103 < 2.2e-16 ***
## Tipus_tractament            0.052  2    1.2989    0.2797
## Setmana                     1.784  1   88.9765 7.457e-14 ***
## Tipus_tractament:Setmana    0.618  2   15.4068 3.229e-06 ***
## Residuals                   1.324 66
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En aquest cas ens surt que la interacció és significativa, és a dir que en el tractament a el pas de les setmanes te un efecte diferent en el tractament a, la recta no és paral · lela a les altres dues. Per tant no es compleix que $\gamma_1 = \gamma_2 = \gamma_3$, ja que la γ_1 és estadísticament diferent.

En conclusió, tenim que l'estimador de $\gamma_1 = -\beta_1 = 0.078986$ i en canvi els altres aquí ens donen $\gamma_2 = -(\beta_1 + (\tau_2 \sim \beta_1)) = 0.078986 + 0.058055 = 0.137041$ i $\gamma_3 = -(\beta_1 + (\tau_3 \sim \beta_1 9) = 0.078986 + 0.055731 = 0.134717$.

Aquestes dues últimes com hem vist fent el model anterior són estadísticament iguals. La única estadísticament diferent és la γ_1.

5)Define a linear model appropiate to check if at the moment of packaging, the juices of the three treatments had the same vitamin C level. From this model:

- **for each treatment, estimate the vitamin C level at the packaging moment, that is at Setmana = 0.**
- **Are they statistically different at Setmana = 0 or not?**

Si la setmana és 0, $log(VitaminaC) = log(\alpha_i)$, i el model lineal ANCOVA definit anteriorment:$y_{ij} = \beta_0 + \tau_i + e_{ij}$.

Com hem vist en l'apartat anterior les τ_i no són significants, són estadísticament iguals a 0, per tant $y_{ij} = \beta_0 + e_{ij}$ Per tant l'estimació del logVitC per a tots els models és estadísticament igual i és $\beta_0 = 3.874678$ i per tant:

```
exp(3.874678)
```

```
## [1] 48.16719
```

Per a comprovar-ho mirem:

```
emm3<- emmeans(m3,~Tipus_tractament|Setmana, at=list(Setmana=c(0)))
emm3
```

```
## Setmana = 0:
##  Tipus_tractament emmean     SE df lower.CL upper.CL
##  a                3.87 0.0616 66     3.75     4.00
##  b                3.86 0.0616 66     3.74     3.99
##  c                3.99 0.0616 66     3.87     4.11
##
## Confidence level used: 0.95
```

L'estimació de log(Vitamina C) a la setmana 0 del tractament 'a' és 3.87, del 'b' és 3.86 i del 'c' és 3.99. Per tant els nivells de Vitamina son en aquest ordre:

```
exp(3.87)
```

```
## [1] 47.94239
```

```
exp(3.86)
```

```
## [1] 47.46535
```

```
exp(3.99)
```

```
## [1] 54.05489
```

```
cld(emm3,Letters=letters, reversed=T)
```

```
## Setmana = 0:
##  Tipus_tractament emmean     SE df lower.CL upper.CL .group
##  c                3.99 0.0616 66     3.87     4.11  a
##  a                3.87 0.0616 66     3.75     4.00  a
##  b                3.86 0.0616 66     3.74     3.99  a
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
## NOTE: Compact letter displays can be misleading
##       because they show NON-findings rather than findings.
##       Consider using 'pairs()', 'pwpp()', or 'pwpm()' instead.
```

Tenint en compte això confirmem que el valor de Vitamina C a la setmana 0 en les mostres tractades amb diferents mètodes son estadísticament iguals.

6)Find out a linear model useful to fit the data and that verifies the general assumptions of the linear model, and answer the following questions: i) which is the variance estimation? ii) which is the amount of variability explained by your model?

El model que triem és el model lineal ANCOVA amb interacció entre les variables tractament i setmana:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 setmana + (\tau_i \sim \beta_1) setmana + e_{ij}$$

```
m3 <- lm(logVitC~ Tipus_tractament + Setmana + Tipus_tractament:Setmana, dades)
summary(m3)
```

```
##
## Call:
## lm(formula = logVitC ~ Tipus_tractament + Setmana + Tipus_tractament:Setmana,
##     data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31777 -0.09776  0.00765  0.09522  0.26339
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.874678   0.061628  62.872  < 2e-16 ***
## Tipus_tractamentb          -0.010617   0.087155  -0.122    0.903
## Tipus_tractamentc           0.115996   0.087155   1.331    0.188
## Setmana                    -0.078986   0.008374  -9.433 7.46e-14 ***
## Tipus_tractamentb:Setmana -0.058055   0.011842  -4.902 6.47e-06 ***
## Tipus_tractamentc:Setmana -0.055731   0.011842  -4.706 1.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1416 on 66 degrees of freedom
## Multiple R-squared:  0.9147, Adjusted R-squared:  0.9083
## F-statistic: 141.6 on 5 and 66 DF,  p-value: < 2.2e-16
```

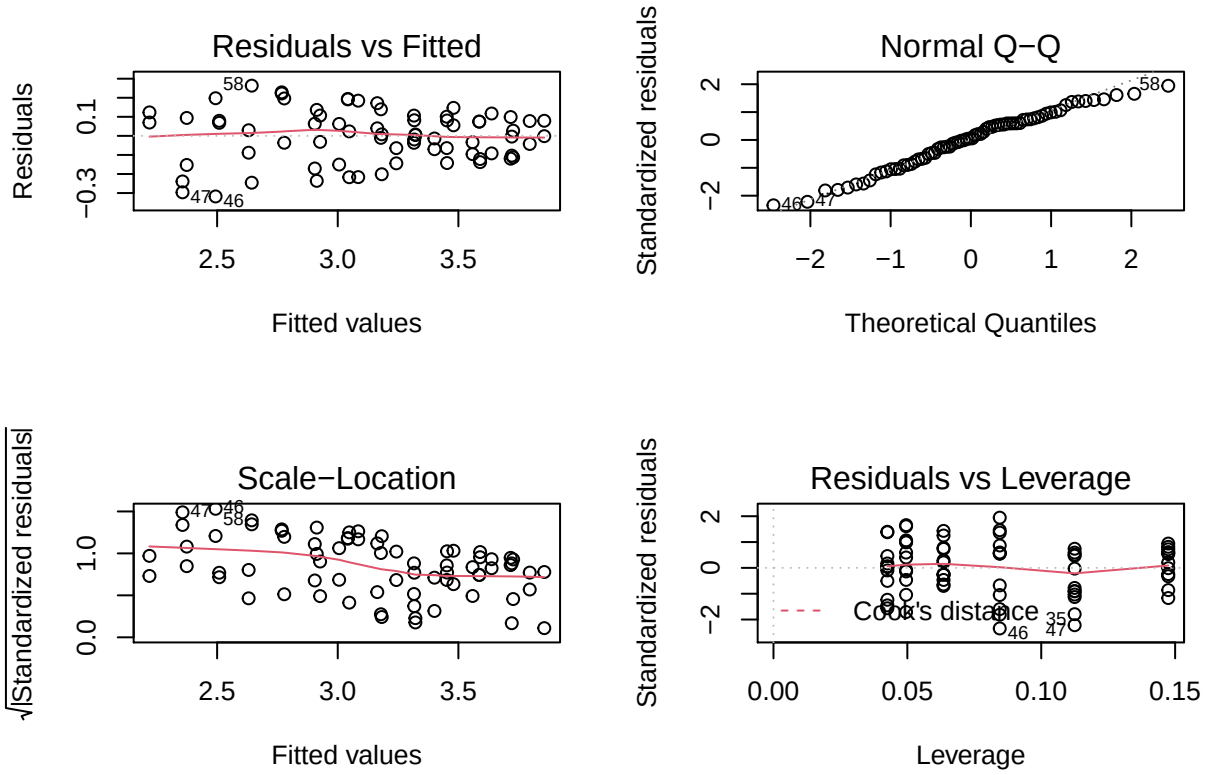
Tot i que el tipus de tractament sigui insignificatiu, si que és significatiu com depenent d'aquest la setmana en que es mesura fa variar el nivell de Vitamina C. Aquest model té una Multiple R-squared i una Adjusted R-squared més altes que el model sense aquesta variable i per tant hem decidit que és millor deixar-la.

L'estimació de la variància és $\hat{\sigma^2} = (0.1416)^2 = 0.02$.

La quantitat de variabilitat explicada pel model és el 91.47% d'aquesta.

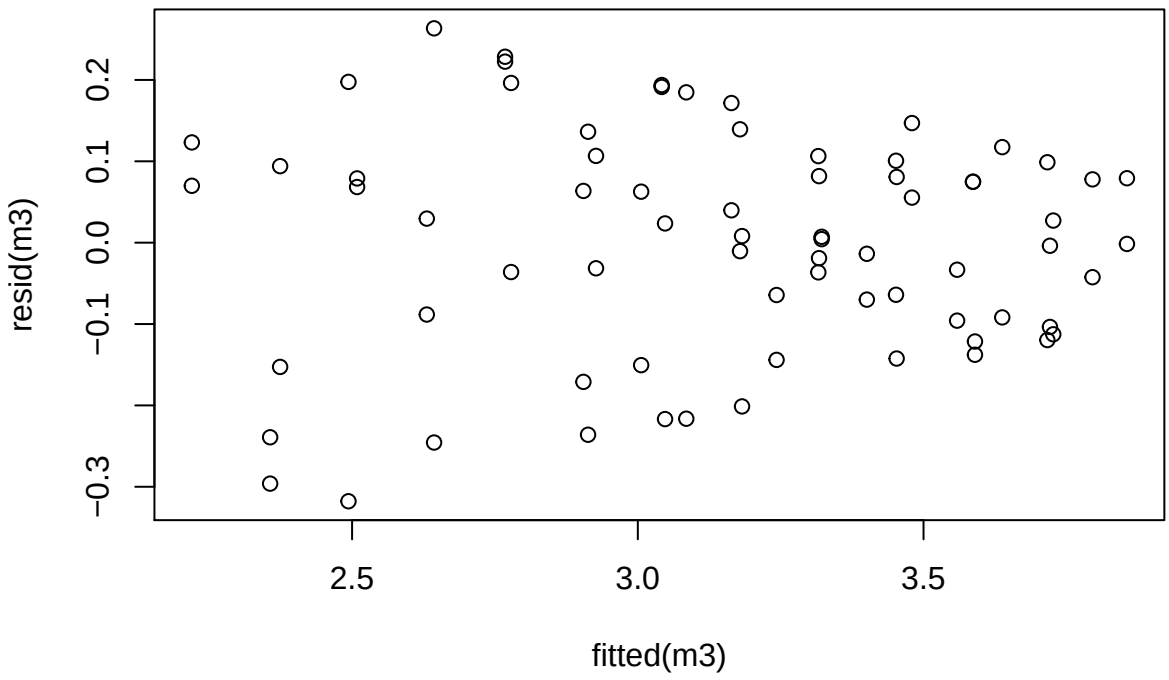
Ara comprovem que es verifiquen les hipòtesis generals dels models lineals:

```
par(mfrow = c(2,2))
plot(m3,ask=F)
```



Acceptem la hipòtesi de normalitat, de linealitat i independència. No veiem cap observació que superi la distància de cook, per tant no hi ha outliers.

```
plot(fitted(m3),resid(m3))
```



No es veu clar que la variança sigui constant, ja que veiem que com més nivell de Vitamina C menys variança. Hem fet més tests per comprovar-ho.

```
shapiro.test(residuals(m3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m3)
## W = 0.97934, p-value = 0.2833
```

Aquest test comprova que les dades provenen d'una normal, la hipòtesi nula és que les dades són normals. Com que p-valor>0.05 no la rebutgem.

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(m3)
```

```
##
##  RESET test
##
## data:  m3
## RESET = 0.4802, df1 = 2, df2 = 64, p-value = 0.6209
```

La hipòtesi nula és que les variables es relacionen de forma lineal. Com que p-valor>0.05 no la rebutgem.

```
bptest(m3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m3
## BP = 18.051, df = 5, p-value = 0.002883
```

Aquets test ens demostra que la variança no és constant en el nostre model, no hi ha homoscedasticitat. Hem intentat canviar la variable resposta (log,loglog,sqrt,invers,exp...) però no hem trobat un model on es compleixin totes les hipòtesis amb aquests canvis.

7) Write down the main conclusions that you may deduce from your model (aprox. half a page). In particular answer the questions specified in point3).

Del nostre model extraiem les següents conlcusions:

La setmana en que es fa la mesura si que té una influencia sobre el nivell de VitaminaC resultant, amb el pas de les setmanes en nivell disminueix.

El tipus de tractament no té una influència directament en el Nivell de vitaminaC mesurat , sino que fa canviar l'efecte del pas de les setmanes en el nivell de Vitamina C.

Hem comprovat que tots els experiments van començar amb el mateix nivell de VitaminaC, en concret tenint en compte que hem fet el model sobre la fórmula $log(VitaminaC) = log(\alpha_i) + \gamma_i setmana$ sabem que $\alpha_i$és estadísitcament igual per tots els tractaments i val 48.167.

Per cada setmana que passa en aquell model el logarítme de la Vitamina C disminueix γ_i unitats. En termes de Vitamina C, el nivell en una setmana és igual al nivell de la setmana anterior multiplicat per $e^{-\gamma_i}$. Tenim $\gamma_1 = 0.0789$ i $\gamma_i = 0.1350$ $i \in \{2, 3\}$.Per tant concluim que amb el tractament 'b' i 'c' el nivell de vitamina C descendeix més ràpid amb el pás del temps.

Al cap de les dotze setmanes els nivells de Vitamina C amb els tractaments 'b' i 'c' són estadísticament iguals (aprox 9.92). En canvi amb el tractament 'a' queda un nivell de Vitamina C superior (18.72).

Ho comprovem:

```
emm3<- emmeans(m3,~Tipus_tractament|Setmana, at=list(Setmana=c(12)))
emm3
```

```
## Setmana = 12:
##  Tipus_tractament emmean     SE df lower.CL upper.CL
##  a                2.93 0.0544 66     2.82     3.04
##  b                2.22 0.0544 66     2.11     2.33
##  c                2.37 0.0544 66     2.27     2.48
##
## Confidence level used: 0.95
```

```
cld(emm3,Letters=letters, reversed=T)
```

```
## Setmana = 12:
##  Tipus_tractament emmean     SE df lower.CL upper.CL .group
##  a                  2.93 0.0544 66     2.82     3.04  a
##  c                  2.37 0.0544 66     2.27     2.48   b
##  b                  2.22 0.0544 66     2.11     2.33   b
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
## NOTE: Compact letter displays can be misleading
##       because they show NON-findings rather than findings.
##       Consider using 'pairs()', 'pwpp()', or 'pwpm()' instead.
```