Iogurts

R Markdown

** PROBABILITAT I ESTADÍSTICA 2 (GCED) **

** EXERCICI 1 **

Treballem el conjunt de dades Iogurts.csv Comencem llegint el fitxer i guardant-lo en un dataset que es diu ** dd**. A continuació fem imprimir la capçalera del conjunt de dades i amb la comanda * summary* fem calcular els estadístics descriptius bàsics per a cada variable.

```
#setwd("~/Desktop/PiE2")
setwd("~/Documents/CURS 2018-2019/PIE2")
library(car)
```

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

```
library(tables)
```

```
## Loading required package: Hmisc
## Loading required package: lattice
## Loading required package: survival
## Warning: package 'survival' was built under R version 3.4.4
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
#library(Hmisc)
#library(lattice)
#library(survival)
#library(Formula)
#library(colorspace)
#library(ggplot2)

dd<-read.csv2("./Dades/Iogurt.csv")

head(dd)
```

```
##   Ferm dia   pH strep lactob
## 1  T42  21 4.10  7.43   7.46
## 2  T42   0 4.44  7.65   7.75
## 3  T42  21 4.02  7.10   7.35
## 4  T42   7 4.24  7.54   7.62
## 5  T42   7 4.27  7.54   7.66
## 6  T42  28 4.01  7.25   7.41
```

```
summary(dd)
```

```
##     Ferm         dia             pH            strep           lactob
##  T42  :30   Min.   : 0   Min.   :3.970   Min.   :6.990   Min.   :7.310
##  T43.5:30   1st Qu.: 7   1st Qu.:4.058   1st Qu.:7.237   1st Qu.:7.430
##             Median :14   Median :4.110   Median :7.335   Median :7.495
##             Mean   :14   Mean   :4.161   Mean   :7.374   Mean   :7.530
##             3rd Qu.:21   3rd Qu.:4.232   3rd Qu.:7.495   3rd Qu.:7.612
##             Max.   :28   Max.   :4.480   Max.   :7.820   Max.   :7.880
```

Del dataset es desprèn que tenim dues variables explicatives: FERMENTACIÓ que es un factor (variable categòrica) amb dos nivells corresponents a les dues temperatures de fermentació considerades, i DIA que es un altre factor amb cinc nivells corresponents als dies transcorreguts des de la fermentació. Iambe es tenen tres variables resposta: pH, STREP i LACTOB.

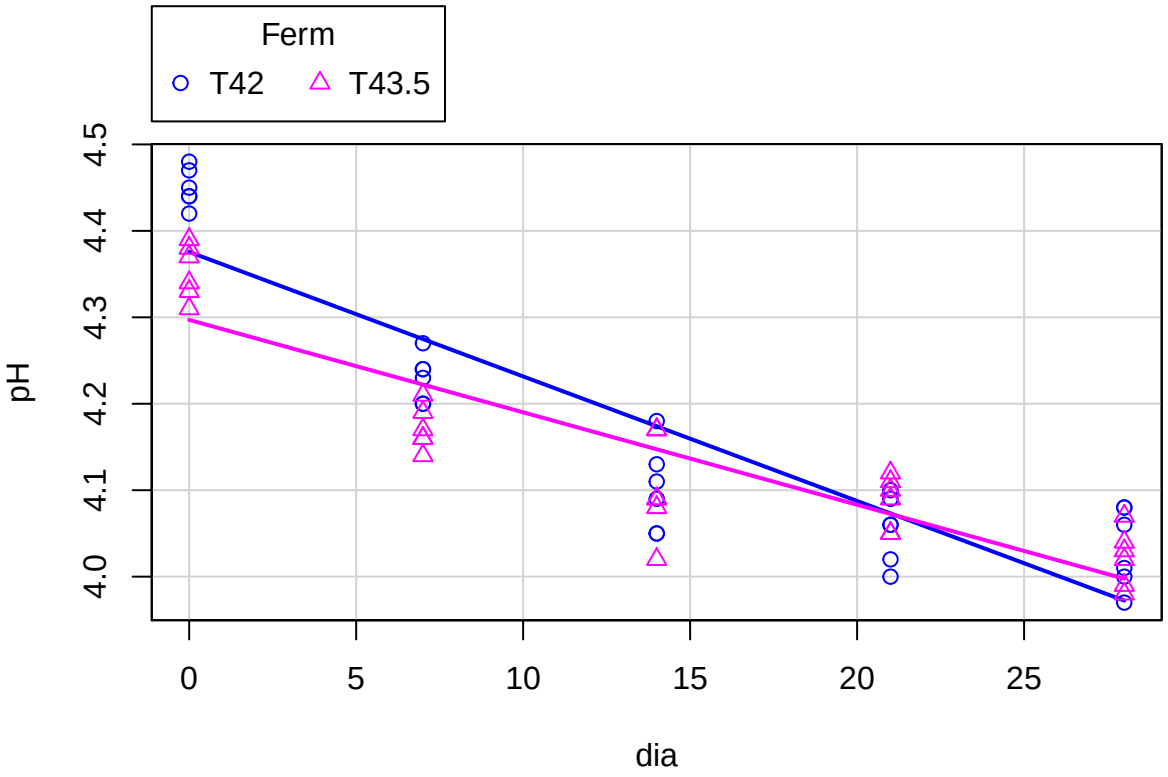
Del fet que la mediana i la mitjana aritmètica de les variables numèriques (les resposta) siguin sempre molt properes es desprèn que es tracta de variables força simètriques.

(a) Descriptiva

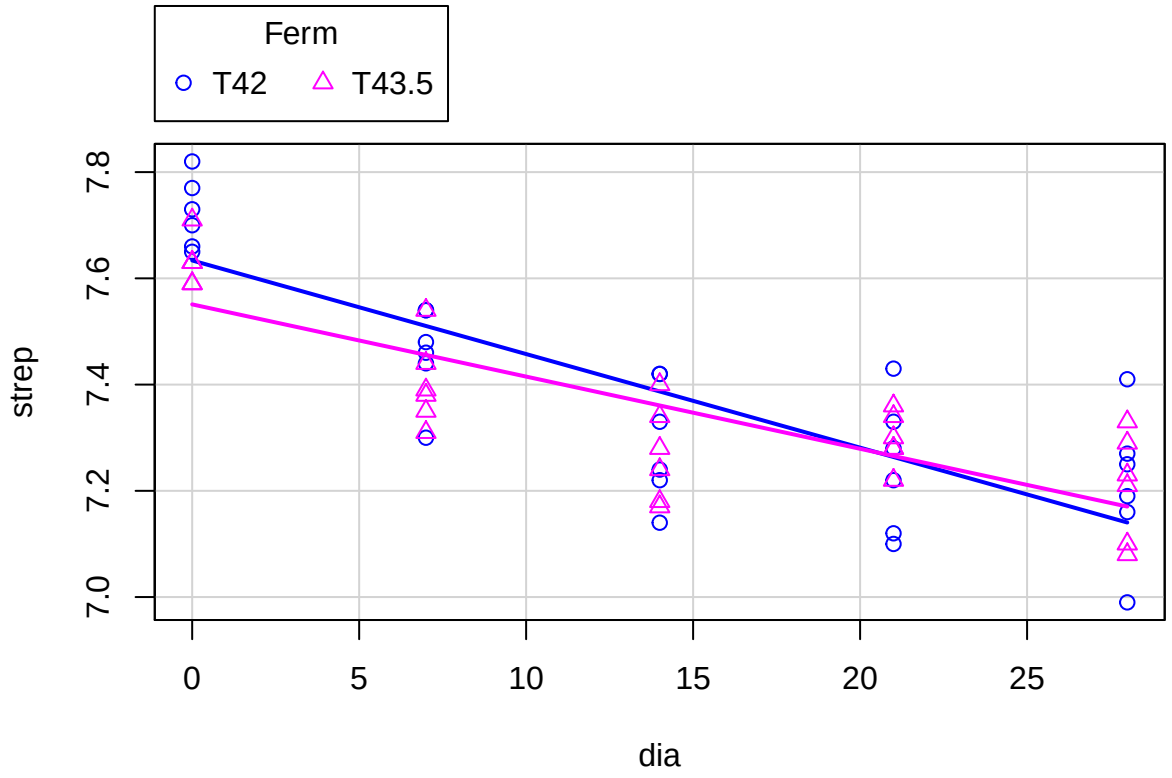
Gràfiques

Anem a realitzar els diagrames de dispersió per a les tres variables resposta en funció dels dies,utilitzant un símbol i color diferent per a cada temperatura.

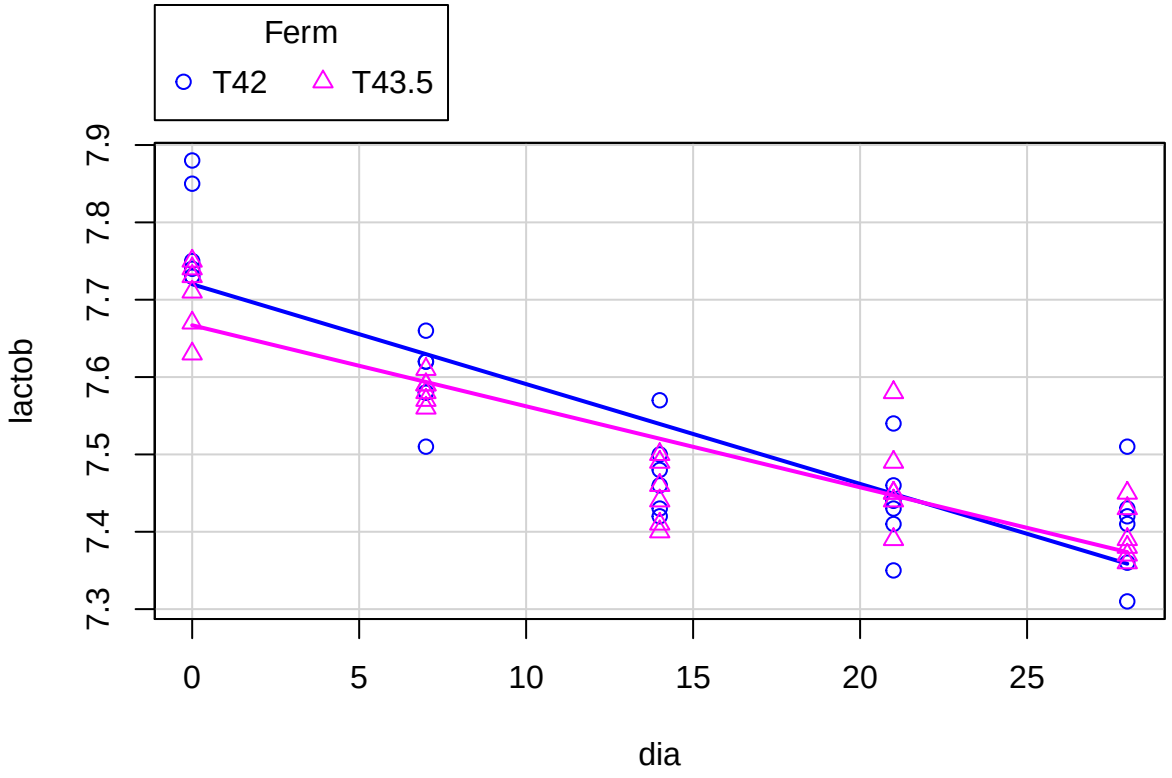
```
sp(pH~dia|Ferm,dd,smooth=F)
```



```
sp(strep~dia|Ferm,smooth=F,dd)
```




```
sp(lactob~dia|Ferm,smooth=F,dd)
```



Per que respecta al pH observem:

1) Els iogurts fermentats a T42 tenen els primers dies un pH superior al altres. Ara bé, a mida que passen els dies els valors de pH son pràcticament indistingibles entre les dues temperatures.

2) La manera en que el pH decreix a mida que passen els dies es mes dràstica pels iogurts fermentats a T42 que els fermentats a T43.5

3) Si considerem un * grup d'observacions * com les observacions que corresponen a un dia concret i a una temperatura de fermentació fixa, no es veuen moltes diferencies entre la variabilitat dels valors dels diferents grups d'observacions.

Pel que respecta a la variable Strep observem bàsicament el mateix que abans, però ara les diferencies entre les dues temperatures de Fermentació a l'inici no son tant acusades. Pels dies 21 i 28 i pel que respecte a la temperatura 42 de fermentació s'observen uns valors de pH que varien molt mes que per les altres temperatures i dies.

Pel que respecta a la variable Lactob observem: Hi ha dos iogurts fermentats a temperatura 42 que tenen una presencia de Lactobacilus molt mes superior que la resta de iogurts. Aquests dos continuen tenint valors mes alts que la resta a mida que passen els dies.

En general podem concloure que tant el pH com l'Strep com el Lactob van decreixent en el iogurt a mida que passen els dies des de la seva fermentació. Aquest decreixement sembla mes marcat pels iogurts fermentats a T42 que pels fermentats a T43.5. Podríem dir que quan han passat mes de 20 dies des de la fermentació, no sembla que hi hagi diferencies entre els dos grups de iogurts er a cap de les tres variables resposta.

Taules

Procedim ara a fer les taules de l'estadística descriptiva per a cada grup d'observacions. Primer li direm al R que consideri la variable dia com un Factor (com que es numèrica, si no li diem la prendrà com una variable numèrica no categòrica). A continuació hi figuren els resultats de associats a la variable pH, i de forma semblant es calcularien les taules de les altres variables.

```
dd$Fdia<-as.factor(dd$dia)
tabular((pH+strep+lactob)*Ferm*((n=1)+mean+sd)~Fdia,dd)
```

			Fdia				
	Ferm		0	7	14	21	28
pH	T42	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	4.45000	4.23000	4.10167	4.05500	4.03333
		sd	0.02191	0.02683	0.04997	0.03886	0.04633
	T43.5	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	4.35333	4.17167	4.10333	4.08667	4.02167
		sd	0.03141	0.02483	0.05785	0.03011	0.03312
strep	T42	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	7.72167	7.46000	7.29500	7.24667	7.21167
		sd	0.06555	0.08854	0.11415	0.12644	0.13891
	T43.5	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	7.63000	7.40167	7.26833	7.29667	7.20667
		sd	0.04382	0.08035	0.09042	0.04967	0.10013
lactob	T42	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	7.78000	7.59500	7.47667	7.43833	7.40667
		sd	0.06693	0.05128	0.05465	0.06242	0.06772
	T43.5	n	6.00000	6.00000	6.00000	6.00000	6.00000
		mean	7.70500	7.58333	7.45000	7.46667	7.39667
		sd	0.04637	0.01751	0.04099	0.06408	0.03559

(b) Comparacions de 2

Ara fixarem un dia, el zero en el nostre cas, i compararem els valors esperats del pH en els dos grups de iogurts, els que han fermentat a T42 i els que ho han fet a T43,5. Això ho farem amb un test t-d'Student de comparació de dos valors esperats sota hipòtesi de normalitat. Aquest test el farem primer suposant variàncies diferents i desconegudes i després suposant variàncies iguals i desconeguda. A continuació farem el test de Fisher per a comparar les variàncies de dos grups. Observeu que li diem que del dataset dd, agafi només les dades corresponents l dia zero.

Comencem fent el test assumint variàncies diferents:

```
t.test(pH~Ferm,dd[dd$dia==0,])
```

```
##
##  Welch Two Sample t-test
##
## data:  pH by Ferm
## t = 6.1828, df = 8.9338, p-value = 0.0001673
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06125851 0.13207482
## sample estimates:
##   mean in group T42 mean in group T43.5
##            4.450000            4.353333
```

Observem que es rebutja la igualtat del pH en els dos grups en el moment d'inici. Això ho podem deduir del fet que el p-valor es molt menor que 0.05 i també del fet que l'interval de confiança no conte el valor zero. això es el que esperàvem un cop vists els plots de dispersió, ates que hem vist gràficament que en l'instant de fermentació era quan les diferencies entre els dos grups, pel que respecte al pH, eren mes evidents i semblaven importants.

Repetim la mateixa comparació però ara assumint que les variàncies en els dos grups son iguals.

```
t.test(pH~Ferm,var.equal=T,dd[dd$dia==0,])
```

```
##
##  Two Sample t-test
##
## data:  pH by Ferm
## t = 6.1828, df = 10, p-value = 0.0001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06183034 0.13150299
## sample estimates:
##   mean in group T42 mean in group T43.5
##            4.450000            4.353333
```

Novament concluim que els valors esperats en els dos grups son diferents. A continuació comparem les variàncies en els dos grups.

```
var.test(pH~Ferm,dd[dd$dia==0,])
```

```
##
##  F test to compare two variances
##
## data:  pH by Ferm
## F = 0.48649, num df = 5, denom df = 5, p-value = 0.4479
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.06807452 3.47661819
## sample estimates:
## ratio of variances
##          0.4864865
```

Com que el p-valor es superior a 0.05 concluim que les variàncies dels dos grups no son significativament diferents. Iambe concluim el mateix si calculem l'interval de confiança pel quocient de les dues variàncies. Com que l'interval conte la unitat, es conclou que les variàncies no son estadísticament diferents.

En aquest document no apareix la resolució del punt (3) perquè es deixa al lector.

==========

(c) Predicció a partir del pH

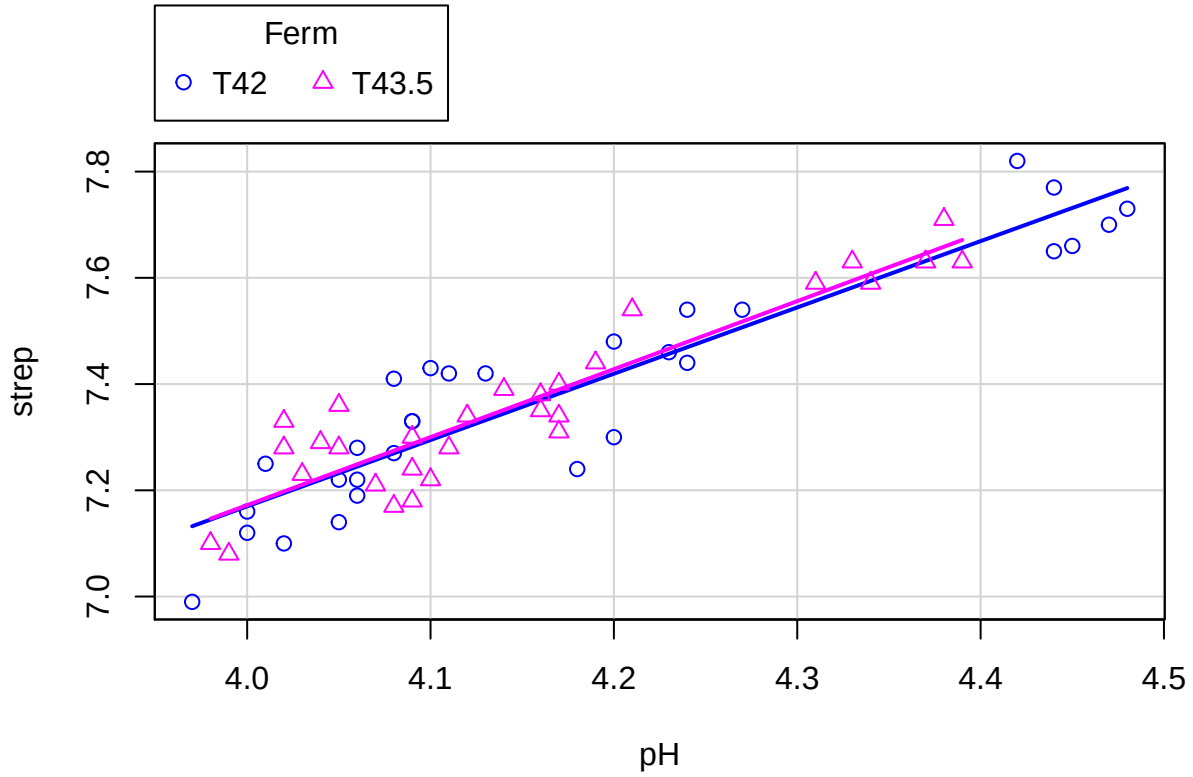
Anem a predir ara les variables Strep i Lactob a partir del pH en cas que sigui possible. si els resultats son bons, al mirar el pH d'un iogurts serem capaços d'estimar la quantitat d'Streptococcus i de Lactobacilus sense necessitat de fer les corresponents anàlisis.

Comencem per la variable Streptococcus

strep

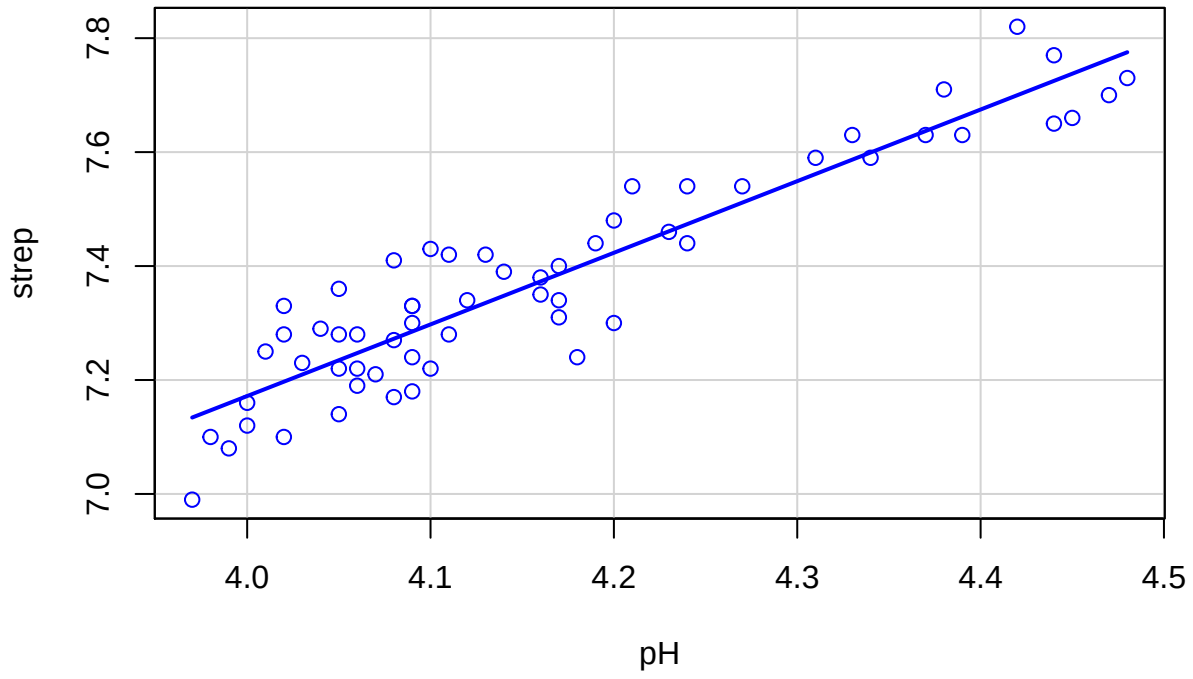
Comencem dibuixant el nuvol de punts que te a l'eix de les x el pH i al de les y's el valor de la variable Strep. Això ho fem, a l'igual que hem fet abans, utilitzant un símbol diferent per a cada grup. Els núvols de punts van acompanyats de la recta que "millor" els ajusta, en termes de que minimitza la mitjana de les discrepàncies entre el valor de l'Strep observat i el predit per la recta al quadrat (mínims quadrats).

```
sp(strep~pH|Ferm,dd,smooth=F,boxplot=F)
```



Com que no veiem (a ull nu) que les rectes siguin molt diferents en els dos grups, a continuació ajustarem una única recta a tot el nuvol de punts. Això ho fem perquè el model es molt mes simple, i perquè així per a predir Strep només cal tenir en compte el valor del pH i no a quina temperatura ha fermentat el iogurt.

```
sp(strep~pH,dd,smooth=F,boxplot=F)
```



Sembla que una única recta pels dos grups es acceptable. Per tal d'obtindre explícitament els coeficients de la recta, cridarem el * procedure* de R que implementa els models lineals i que es diu lm. Primer explicitem

quin model lineal volem ajustar a les nostres dades i després li direm que ens resumeixi la principal informació obtinguda d'ajustar aquest model. Si no invoquem la comanda * summary* guardara els resultats del model lineal a mstrep i no veurem la sortida.

```
mstrep<-lm(strep~pH,dd)
summary(mstrep)
```

```
##
## Call:
## lm(formula = strep ~ pH, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15814 -0.05035 -0.00171  0.04508  0.13758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14327    0.28205   7.599 2.89e-10 ***
## pH           1.25715    0.06775  18.556  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07359 on 58 degrees of freedom
## Multiple R-squared:  0.8558, Adjusted R-squared:  0.8533
## F-statistic: 344.3 on 1 and 58 DF,  p-value: < 2.2e-16
```

L'important que hem de destacar d'aquesta sortida en aquest moment es que ens dona els dos coeficients de la recta que es el que necessitem. El significat de la resta de valors que apareixen es veurà al llarg del curs. El fet que la recta tingui pendent positiva implica que a l'augmentar el pH augmenta també el nombre de Strepococcus. El model que proposem per a predir Strep com a funció del pH independentment de la temperatura de fermentació es el següent:

$$Strep = 2.14 + 1.25 * pH$$

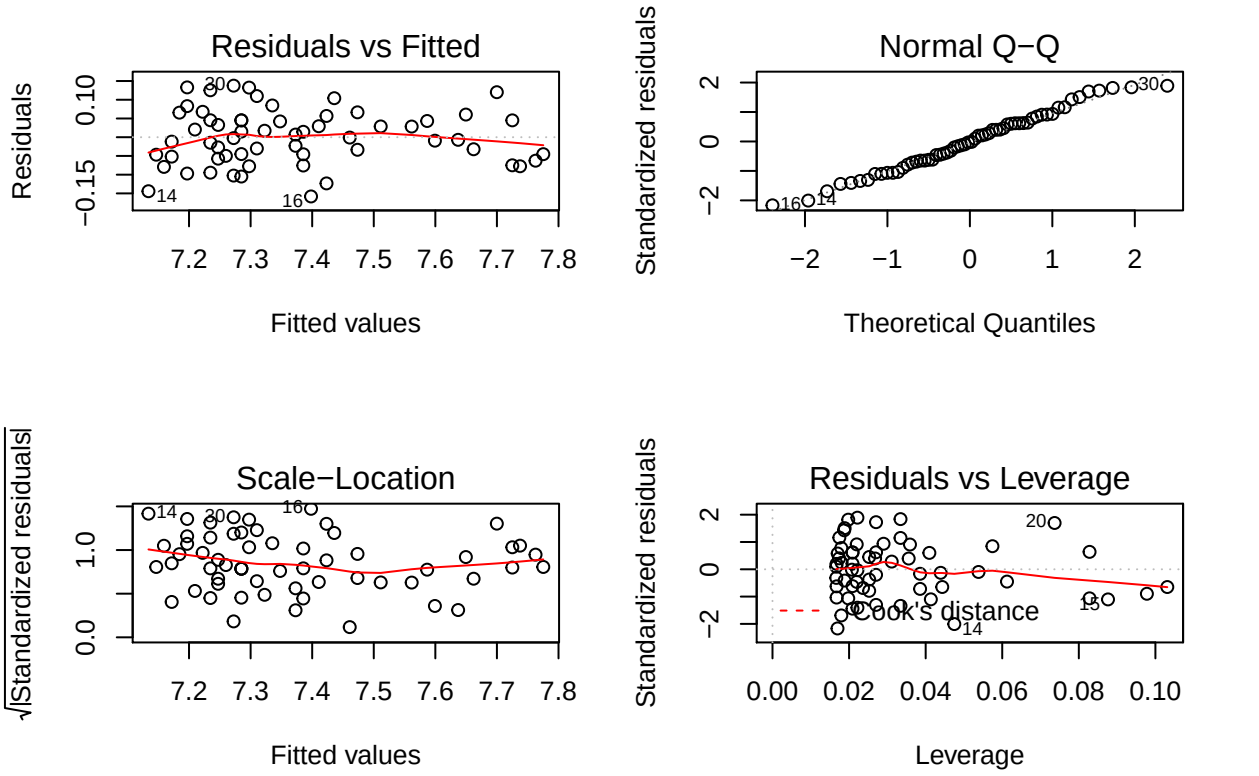
A continuació es calcula el valor de l'Strep associat a un pH de 4. Valors de Strep inferiors a l'obtingut aniran associats a iogurts caducats.

```
mstrep$coef[1]+mstrep$coef[2]*4
```

```
## (Intercept)
##    7.171852
```

Per tal que el model lineal que hem ajustat es pugui considerar un model apropiat per a les nostres dades cal, entre altres coses, comprovar que els * errors* o * residus * que s'obtenen a partir del model segueixin una distribució Normal de valor esperat zero i variància constant. Iambe cal que no es vegin patrons en el dibuix dels residus com a funció dels valors esperats. Per residu s'entén la diferencia entre el valor real de Strep per un pH concret i l'estimat mitjançant la recta. Les següents comandes de R porten a terme els plots de residus que son necessaris per a verificar les hipòtesis del model lineal.

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




```
par(oldpar)
```

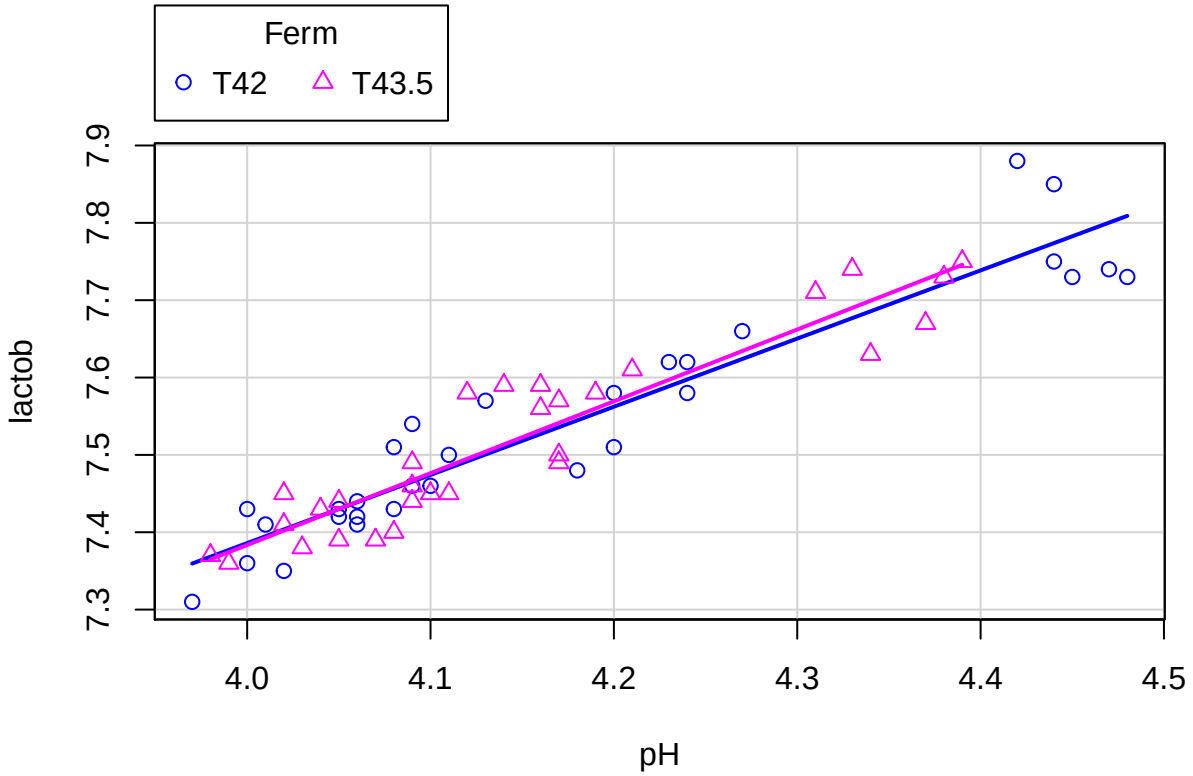
Veiem que els residus no presenten patrons. Iambe veiem en el plot de normalitat que es poden considerar normals. No veiem diferencies entre les variàncies en tot el rang dels valors ajustats. Les hipòtesis del model lineal semblen complir-se.

A continuació portarem a terme la mateixa anàlisi però ara per a la variable Lactob, per tal de poder predir el nivell de Lactobacilus un cop conegut el pH.

lactob

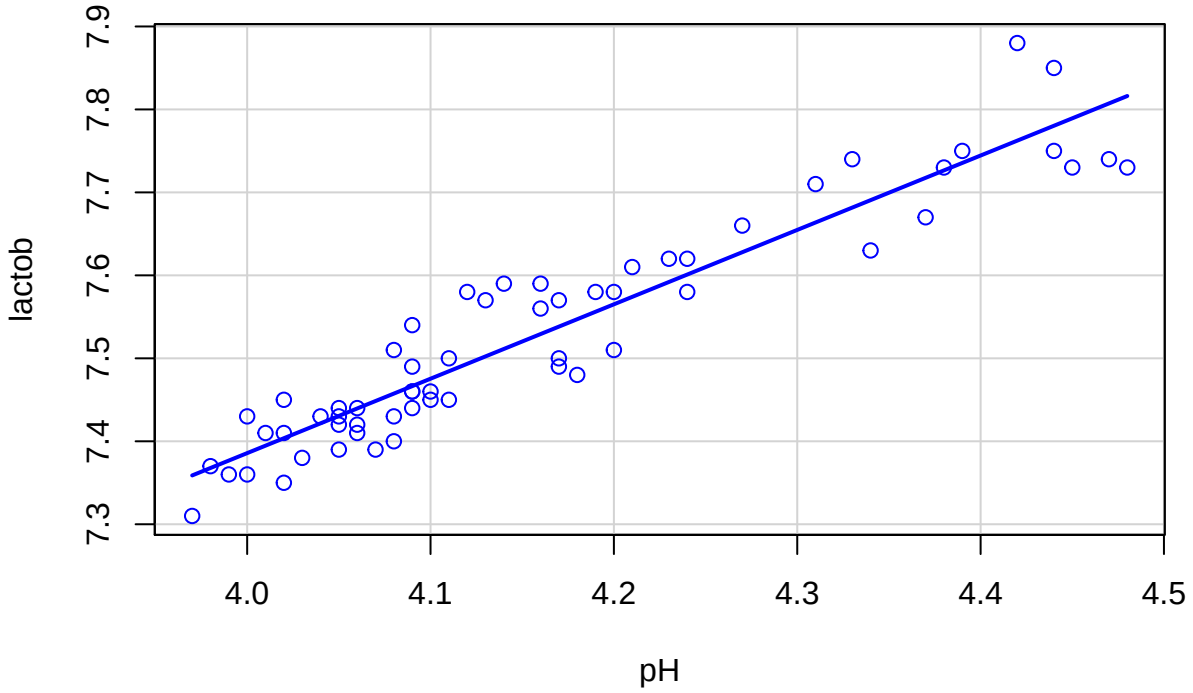
Ajustem el nuvol de punts per a dues rectes (un per a cada temperatura de fermentació)

```
sp(lactob~pH|Ferm,dd,smooth=F,boxplot=F)
```



Novament les rectes son molt semblants i això ens porta a considerar una única recta.

```
sp(lactob~pH,dd,smooth=F,boxplot=F)
```



el fet que la recta tingui pendent positiva indica novament que a l'augmentar el pH del iogurt augmenta també el nombre de Lactobacilus que hi trobem. A continuació apliquem un model lineal, perquè així podrem obtindre els coeficients de la recta de forma explicita.

```
mlactob<-lm(lactob~pH,dd)
summary(mlactob)
```

```
##
## Call:
## lm(formula = lactob ~ pH, data = dd)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086181 -0.033098 -0.000082  0.031023  0.117622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.79895    0.17497   21.71   <2e-16 ***
## pH           0.89670    0.04203   21.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04565 on 58 degrees of freedom
## Multiple R-squared:  0.887,  Adjusted R-squared:  0.885
## F-statistic: 455.2 on 1 and 58 DF,  p-value: < 2.2e-16
```

La recta que ens descriu el nivell de lactobacilus en funció del pH es la següent:

$$Lactob = 3.79 + 0.89 * pH$$

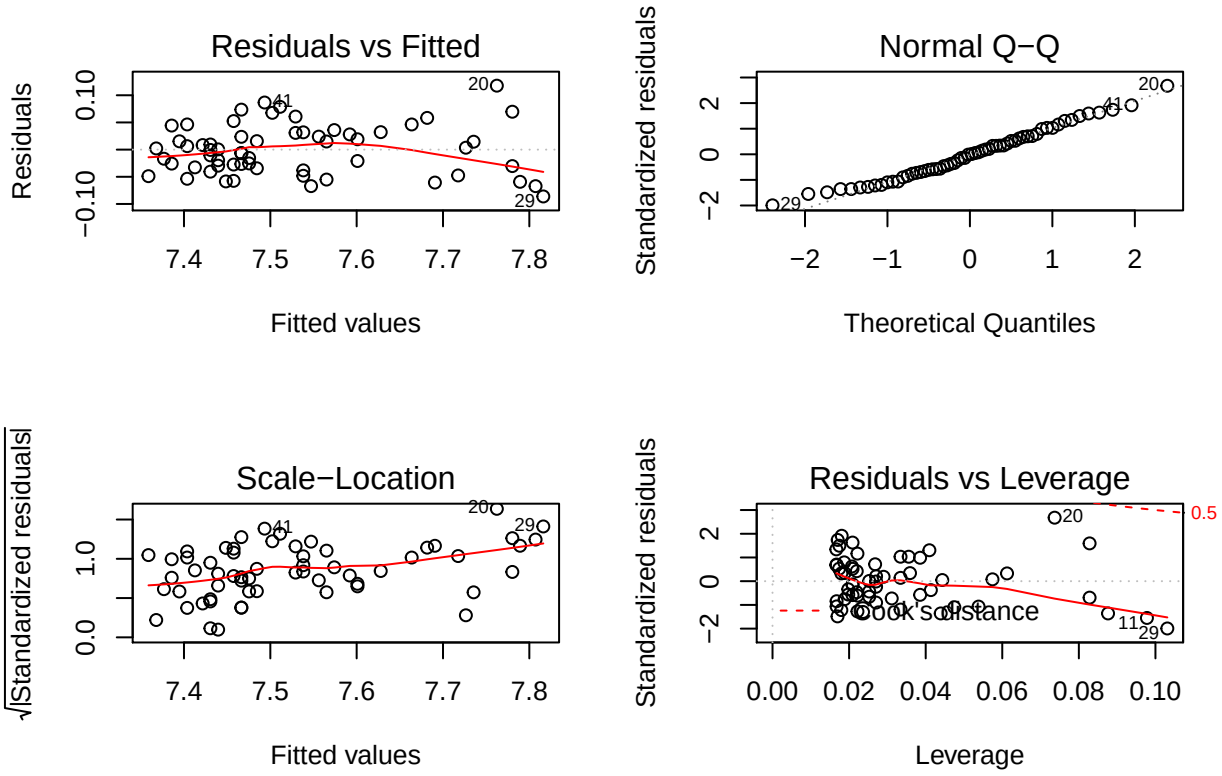
Nivells de Lactob inferiors al valor següent aniran associats a iogurts caducats:

```
mlactob$coef[1]+mlactob$coef[2]*4
```

```
## (Intercept)
##    7.385763
```

A l'igual que abans anem a efectuar els dibuixos corresponents sobre els residus del model per tal de veure que les hipòtesis de Normalitat, independència i igualtat de variàncies es compleixen.

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




```
par(oldpar)
```

Per acabar anem a calcular dues mesures que van associades a la *bondat d'ajust* del nostre model. La primera es el R^2 i s'interpreta com el grau de variabilitat en la variable Strep (o Lactob) que es deguda a iogurts amb diferent valor de pH. Com mes proper al 100% sigui aquesta mesura, mes satisfactòriament ajustara les dades el nostre model. La segona mesura es el valor de la log-versemblança.

R2 i lv

```
c(strep=summary(mstrep)$r.squared,lactob=summary(mlactob)$r.squared)
```

```
##     strep    lactob
## 0.8558319 0.8869845
```

```
c(strep=logLik(mstrep),lactob=logLik(mlactob))
```

```
##     strep    lactob
##  72.43656 101.08524
```

Veiem que en els dos casos mes d'un 80% de la variabilitat que observem en els iogurts en la variable Strep i Lactob es deguda a que els iogurts tenen diferent pH. Això ens diu que els dos models ajustats van molt be. El valor de la logversemblança (a l'igual que l'R^2) es lleugerament superior en el segon model indicant que ajustem una mica millor el Lactob que l'Strep.