Linear model: Yogurt Fermentation!

In order to see which is the **influence of the temperature in the yogurt's fermentation**, some yogurts have been studied during their commercialization period. The **experimental units are yogurts** that have been analyzed several days after their fermentation. The variables are:

Two explanatory variables that are:

- *Days*: days between the fermentation and the day when the analysis is performed.
- *Groups*: there are two groups of yogurts depending on their fermentation temperature, which are 42° and 43.5°.

Three response variables that have been rerecorded during the analysis, and that are:

- *pH*: the yogurt's pH
- *Strep*: the concentration of Streptococcus salivarius thermophilus.
- *Lactob*: the concentration of Lactobacillus delbrueckii bulgaricus.

```
setwd("~/Desktop/xperiments/Rii")
dat<-read.csv2("Iogurt.csv")
head(dat)
```

```
##   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
```

Shown above: some of the rows from the Iogurt.csv table.

With the *summary* command we calculate the basic descriptive statistics for each variable of the dataset:

```
summary(dat)
```

```
##     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
```

We have two explanatory variables: FERMENTATION which is a factor (categorical variable) with two levels corresponding to the two fermentation temperatures considered, and DIA (= DAY) which is another factor with five levels, each one corresponding to a determined day, after the fermentation started. Also, there are three response variables: pH, STREP and LACTOB.

The fact that the median and the mean of the numeric variables (the response) are similar is due to the fact that the variables are quite symmetric.

```
library(car)
library(tables)
```

(1) For each one of the response variables, we want to perform a dispersion diagram. Afterwards we will describe what do we deduce from it.

First of all, we are going to study how pH does change from one day to another due to the fermentation.

It is convenient to plot a *scatter plot* (also known as *dispersion diagram*). This can be done with the *scatter plot* command which is *sp*(...). Then we define the variables: we want to put the pH as a function of the days, meaning that the days form the independent variable (the x axis) and the pH - the dependent variable (on the y axis).

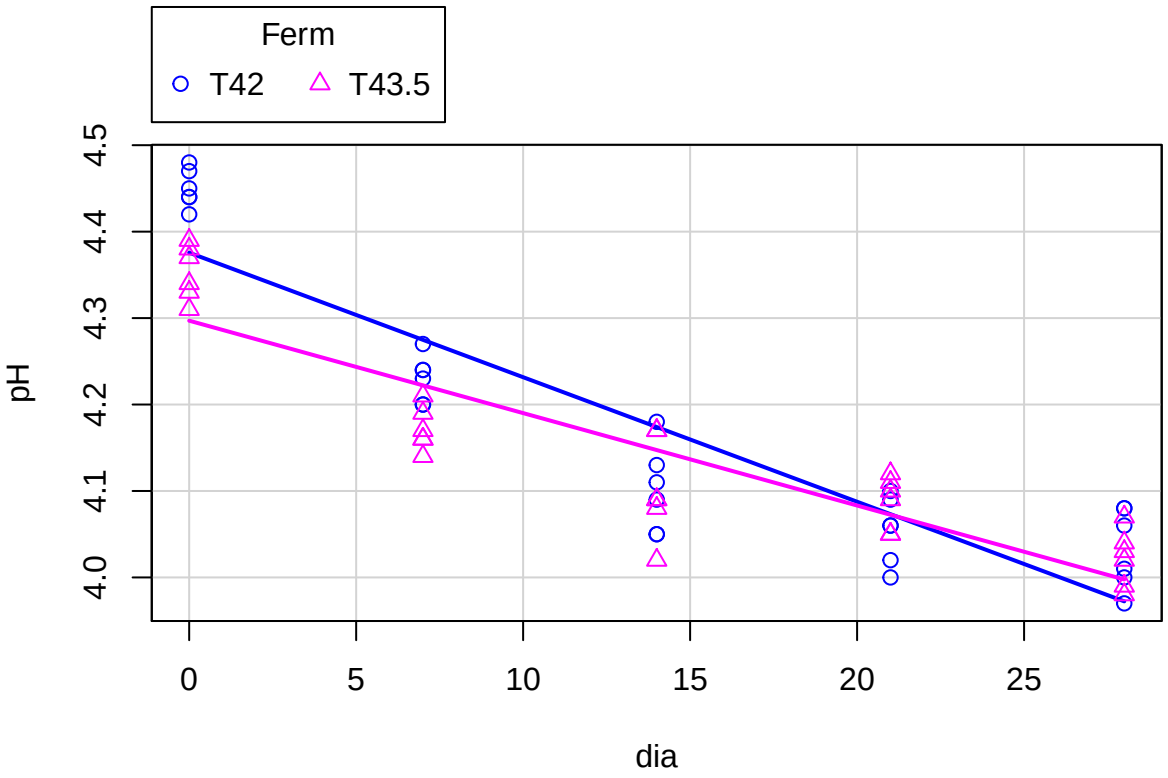
$$days = g(pH) \rightarrow pH \sim dia$$

This is expressed as $pH \sim dia$. If we want to distinguish between different fermentation temperatures (42° and 43.5°), we have to condition on the fermentation.

$$days = g(pH)_{42^\circ} \vee g(pH)_{43.5^\circ} \rightarrow pH \sim dia|Ferm$$

This can be done by $|Ferm$. Next we define which dataframe to use to perform the *scatter plot*, which is *dat*.

```
sp(pH~dia|Ferm,dat,smooth=F) # sp stands for scatter plot
```

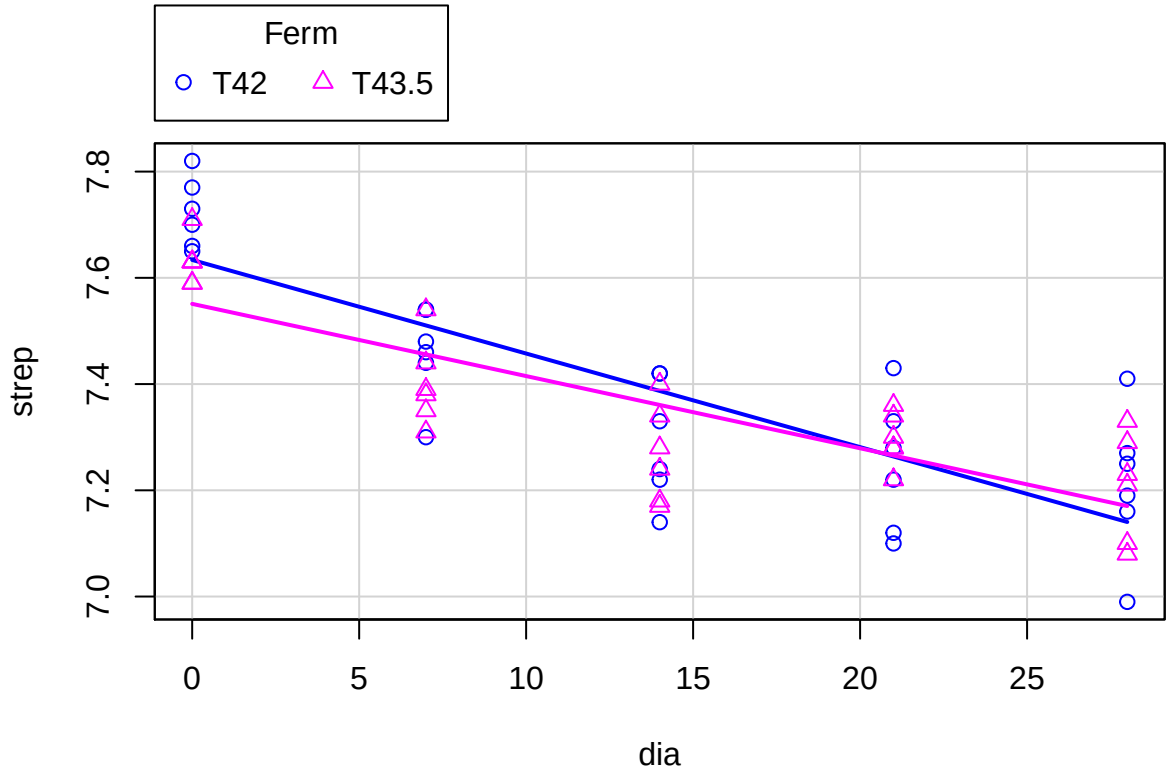


We see that with the course of the days, the pH is decreasing. When the pH of the yogurt is lower than 4, it means that the yogurt is past its use-by date. We can observe that:

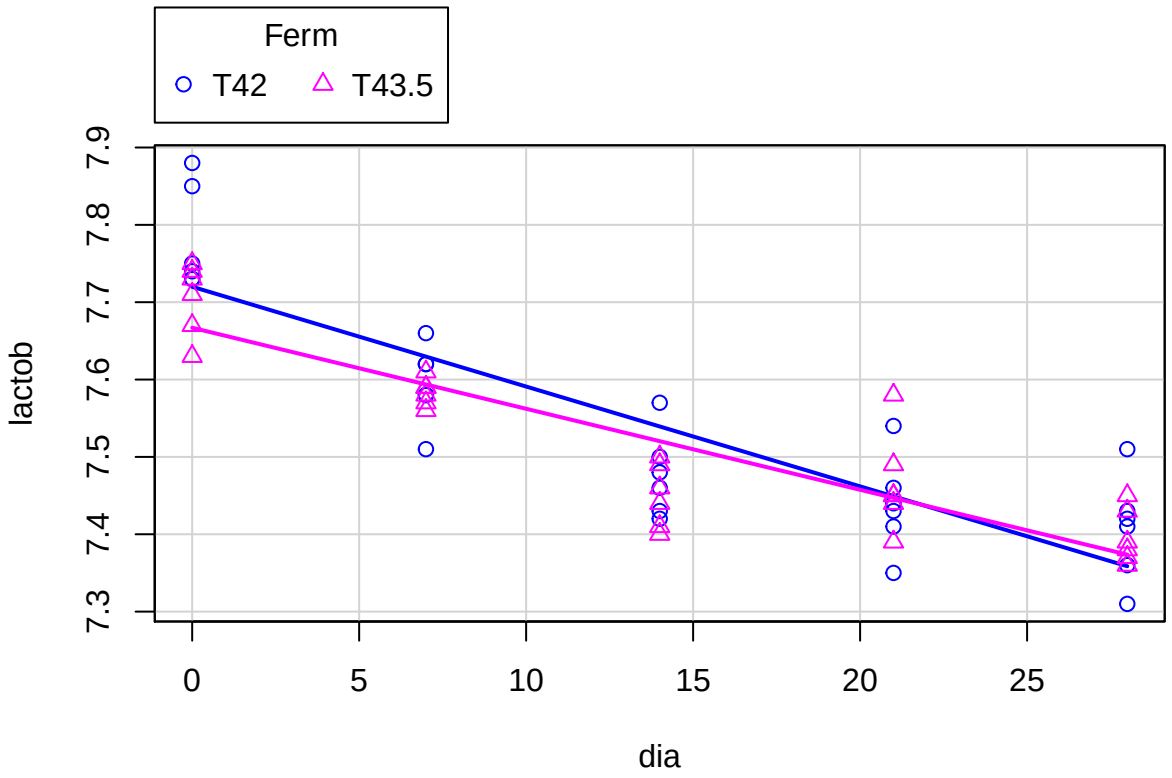
- The yogurts fermented at 42° on the first days have a higher pH than the others. However, with the course of the days, the values of the pH are practically indistinguishable between both temperatures.
- The pH decreases in a more drastic way in the yogurts fermented at 42° than the ones at 43.5°.
- We do not observe much variability between the groups of both temperatures.

Let's have look on the plots of *bacteria* $\sim$ *dia*.

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




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



We see that if the concentration of the bacteria is low, it cannot be called yogurt anymore, because the pH level decreases and the yogurt goes bad.

From what we have seen, we can conclude that with the course of the days from the fermentation, the pH, the concentration of Streptococcus and Lactobacillus decreases. This decrease is more noticeable in yogurts fermented at 42° than the ones at 43.5°.

(2) We are going to create a *Table* containing the basic descriptive measures for each one of the groups and each one of the days.

Tables can be created by using a spreadsheet or using R, with the *tables* package (we have already downloaded it earlier). Also, before starting, observe that for accessing to a parameter of a dataframe, we use the dolar symbol, like $dat\$pH$. So, we have to distinguish between three components of table:

- *Factors*: these are categorical variables for grouping. *For example*: the temperature (42° or 43.5°) or the days (0 or 7 or 14 or 21 or 28). We will factor the numeric variable *dia* because otherwise it will be considered as a numeric non-catergorical variable. We can factor a numeric variable by:

```
dat$Fdia <- as.factor(dat$dia)
```

- *Variables*: we use one variable or more if we sum them (+), but we cannot combine variables (*).
- *Statistics*: examples of statistics are the *mean*, *standard deviation*... We cannot combine statistics with statistics, yet we can combine statistics with variables.

To create a table use $tabular(rows \sim columns)$:

```
tabular((pH+strep+lactob)*Ferm*((n=1)+mean+sd)~Fdia,dat)
```

			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

(3) Considering the pH as a response variable and fixing a particular value for the variable Days. Assuming that the the response variable follows a Normal Distribution with constant variance and mean value that depends on the *temperature group*, with an alpha value equal to 0.05 (by default), we can wonder about:

- Whether there do exist some differences between the mean values for the pH for the two groups?
- Is it correct to assume that the variances of the pH variable in the two groups are equal?

To be able to guarantee that the mean values for the pH for the groups, μ_1 and μ_2, are different we need to check that with a test. Taking $day = 0$ and assuming

$$X_1 = pH|_{42,0} \sim N(\mu_1, \sigma^2)$$

$$X_2 = pH|_{43.5,0} \sim N(\mu_2, \sigma^2)$$

In case the variances are the same we can use the T-Student test, Fisher and other exact distributions. However, when the variances are different, we can have just an approximation (using T-student with degrees of freedom). When we are asking about similar things (like the mean values for the pH with similar temperatures) we can assume same variance, since, in general, it varies just a bit.

How do we contrast a test (null hypothesis *vs* alternative hypothesis)?

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

We must compute the statistic and the p-value. α is the significance level. That is, the risk that we accept. Now, if the p-value $> \alpha$ (p-value is large), we accept H_0, though we have not proved it. We say that "we do not detect any difference". But if p-value $< \alpha$ (p-value is small), then for **sure** we accept H_1, because the probability of being mistaken is really low. Now, let's test whether there are differences in the pH in the two groups on day 0:

```
t.test(pH~Ferm,var.equal=T,dat[dat$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
```

We use T-test (= with equal variances). See that we have 12 data and 2 groups, therefore $12 - 2 = 10$ degrees of freedom (df). See that p-value $= 0.0001038 < \alpha = 0.05$, thus H_1 is true (the test is significative). Also, see that the confidence interval does not contain 0.

We can also check the test with different variances:

```
t.test(pH~Ferm,dat[dat$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
```

The degrees of freedom have decreased and consequently the p-value has increased, but still p-value $= 0.0001673 < \alpha = 0.05$ and thus the test is significative - H_1 is true, the mean values of pH at different temperatures on day 0 are different!

We can also test whether the variances are equal (with a Fisher test)...

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$
$$H_1: \quad \sigma_1^2 \neq \sigma_2^2$$

```r
var.test(pH~Ferm,dat[dat$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
```

Now, we have obtained p-value $= 0.4479 > \alpha = 0.05$, meaning that we do not dare to say that the variances are different. So, we just say that we do not detect any difference (accept H_0), that the variances of both groups are not significantly different. Also, see that the interval contains 1, meaning that the variances are not statistically different.

Look that if try to check the mean value of pH on the last day of both groups we see ...

```r
t.test(pH~Ferm,var.equal=T,dat[dat$dia==28,])
```

```
##
##  Two Sample t-test
##
## data:  pH by Ferm
## t = 0.5018, df = 10, p-value = 0.6267
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04013723  0.06347056
## sample estimates:
##   mean in group T42 mean in group T43.5
##            4.033333            4.021667
```

With p-value $= 0.6267 > \alpha = 0.05$ we accept H_0, we cannot say that they are different. We say that we do not detect any difference between the mean values of the pH of both groups. This proves our intuition from what we have seen at the beginning in the scatter plot.

(4) Given that it is easier to obtain the pH value than to obtain the Strep and Lactob values, one is interested in predicting the Strep and Lactob variable from the pH:

- Which simple expression can we propose to predict the Strep concentration from the pH? If one considers the values of pH smaller than 4 this give place to yogurts that have already expired, which is the limit value of the Strep variable to say that a yogurt has expired?
- Which simple expression can we propose to predict the Lactob variable from the pH?. If one considers the values of pH smaller than 4 give place to yogurts that have already expired, which is the limit value of the Lactob variable to say that a yogurt has expired?
- From the expressions we have, which one is more accurate and why?

We are going to predict the values of teh variables *Strep* and *Lactob* using pH if it is possible. If the results are good, just looking at the pH of the yogurt we will be able to predict the concentration of Streptococcus and Lactobacillus in the yogurts without any further analysis.

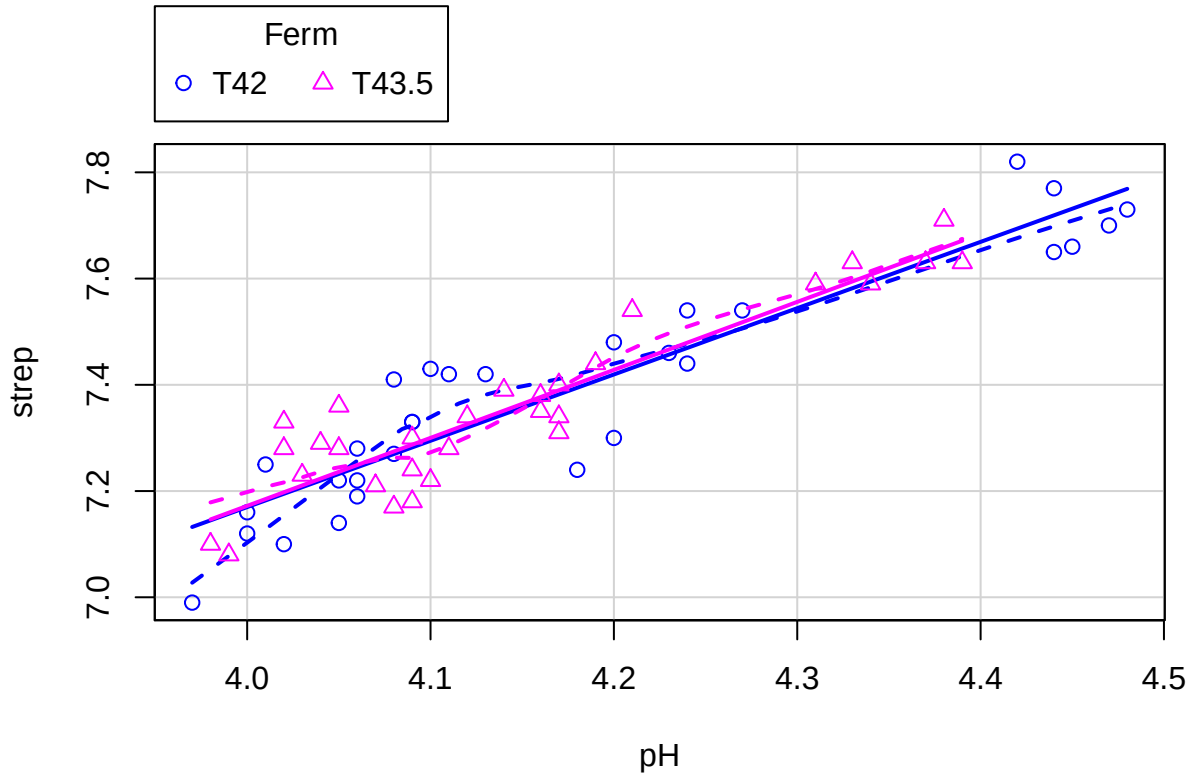
Let's start with the variable of *Streptoccocus*. Start off with a *linear model*:

$$Steptococcus = \alpha + \beta \cdot pH + e_i$$

Where e_i is some error (deviation) and we ask for $e_i \sim N(0,1)$, with mean 0 because we do not want e_i contribute anything to the expected value and with constant variance ($= 1$).

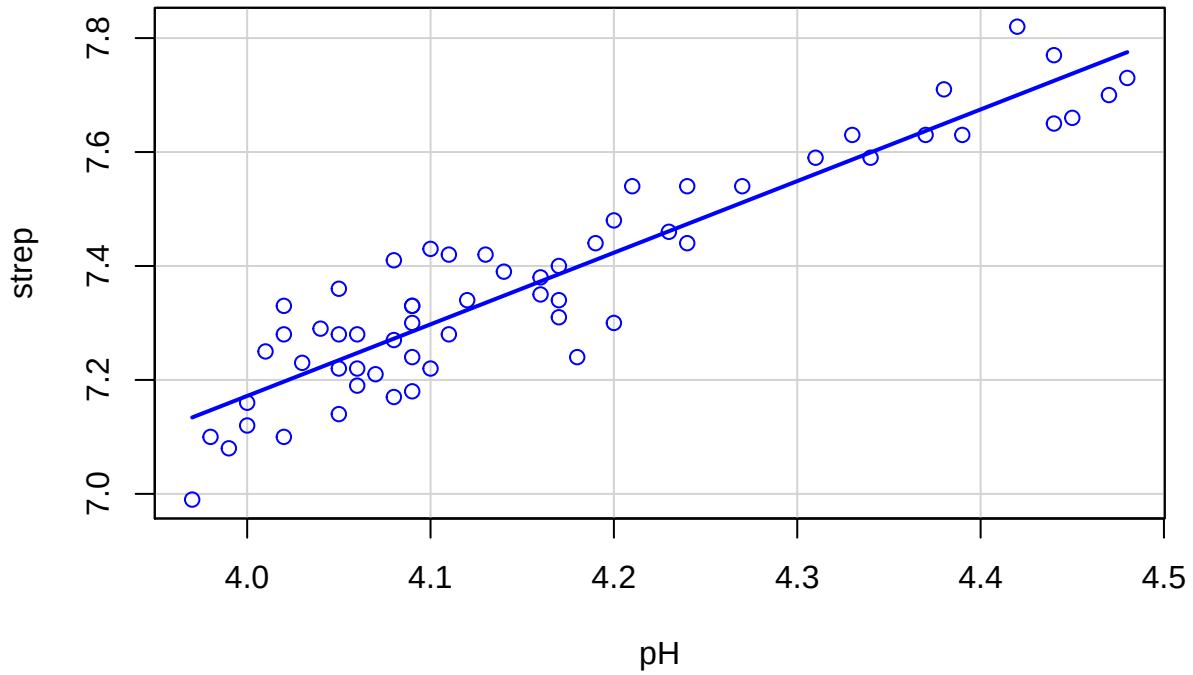
We start plotting the cloud of points, with values of pH at the x axis and values of *Strep* at the y axis. Just as before, each symbol corresponds to a group (42° or 43.5°). The clouds of points are accompanied by a line that "better" fits them, in terms that minimize the mean of the discrepancies (variance) between the value of *Strep* observed and the predicted one by the line, at the power of two (least squares).

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



We do not see the lines of both groups really different. Next, we are going to fit an only one line in the whole cloud of points. We do this because in that the way our model is going to be much simpler and also, this we can predict the concentration of *Streptococcus* just by the pH, we do not have to keep in mind the temperature at which the yogurt has been fermented.

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



It seems that an only one line is acceptable for both groups (the temperature of fermentation does not affect). To obtain explicitly the coefficients of the line, we shall invoke the *procedure* of R which implements the linear models, called *lm* (=*linear model*). Firstly, we define which linear model we want to fit our data and then we ask to summarize the main inormation obtained from this model. If we do not put the parenthesis around the *model_strep* assignment, it will save the results into the linear model *model_strep* and we will not see the output:

```
(model_strep<-lm(strep~pH,dat))
```

```
##
## Call:
## lm(formula = strep ~ pH, data = dat)
##
## Coefficients:
## (Intercept)           pH
##       2.143        1.257
```

We obtain the coefficients of the line: Intercept: $\alpha = 2.143$

The coefficient that multiplies pH: $\beta = 1.257$

The fact that the slope is positive implies that if we increase the pH, the number of *Streptococcus* also increases. The model that we propose to predict *Strep* as a function of pH independently of the fermentation temperature is the following:

$$Strep = 2.143 + 1.257 \cdot pH$$

```
summary(model_strep)
```

```
##
## Call:
## lm(formula = strep ~ pH, data = dat)
##
## 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
```

The *summary* resolves the tests upon the coefficients (we can see that at the **Coefficients** section). The first one is about α:

$$H_0: \quad \alpha = 0$$
$$H_1: \quad \alpha \neq 0$$

See that p-value $= 2.89e-10 < \alpha = 0.05$, for sure H_1! α is not zero meaning that the line does not cross (0,0). Next test about β:

$$H_0: \quad \beta = 0$$
$$H_1: \quad \beta \neq 0$$

If we accept H_0 it would mean that *Strep* does not depend upon pH, however, if we accept H_1 it would mean that β is significative and that the pH has some effect upon the concentration of *Streptococcus*. We can clearly see that p-value $= 2e-16 < \alpha = 0.05$, for sure we accept H_1!

See that the *Residual standard error* is small, meaning the model fits well the data. Also, the determination coefficient (R-squared) is near 1 (= that is good, the model fits well).

Next, we calculate the value of *Strep* associated to $pH = 4$. Values of *Strep* lower than the obtained will be associated to expired yogurts:

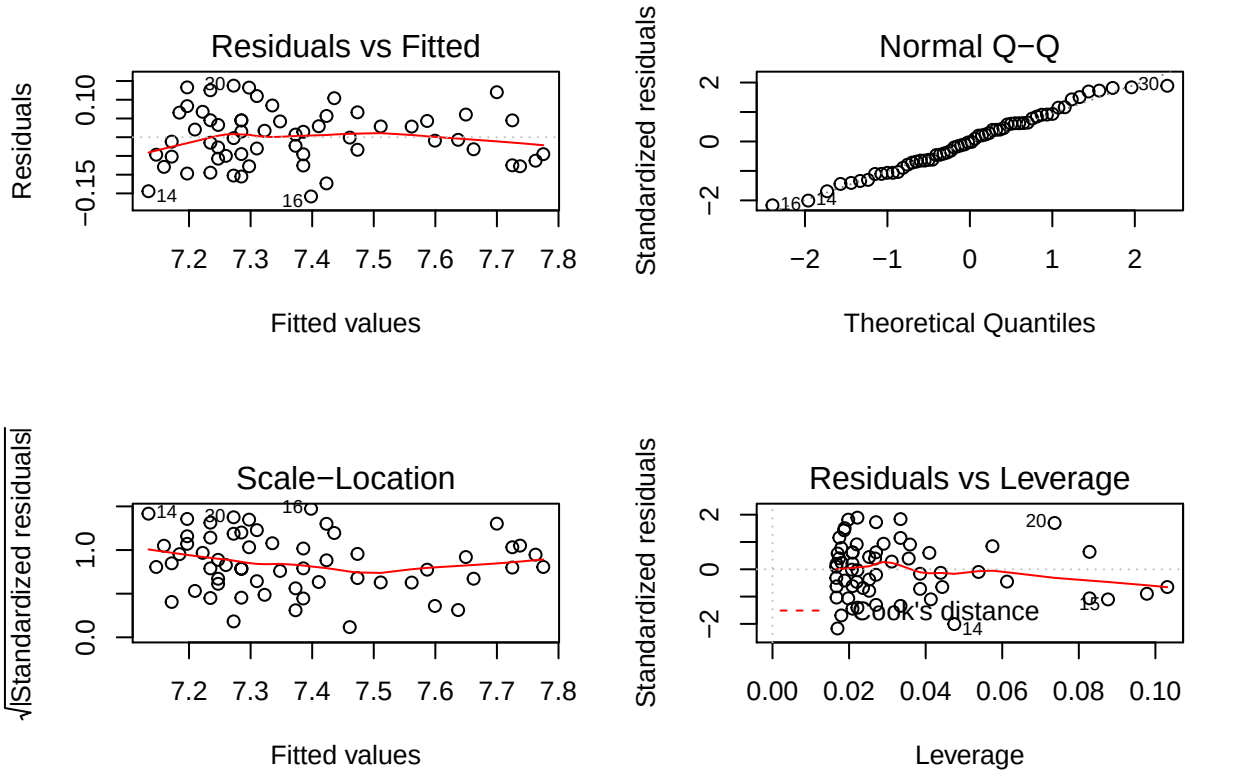
$$Strep_{lim} = 2.143 + 1.257 \cdot 4 = 7.171852$$

```
(limit_strep <- model_strep$coef[1]+model_strep$coef[2]*4)
```

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

In order to consider the linear model that we have fitted to our data **appropriate** we have to check that the *errors* or *residues* obtained from the model follow a Normal distribution of expected value 0 and constant variance. Also, there should not be any patterns in the picture of the residues as function of the expected values. A residue is understood as the difference between the real value of *Strep* for a determined pH and the estimated by the regression line. The following R commands plot these graphics of residues needed to verify the hypothesis of the linear model:

```
oldpar <- par(mfrow=c(2,2))
# c(2,2) is for 2 rows and 2 columns.
plot(model_strep,ask=F) # ask=F is for plotting everything at once.
```



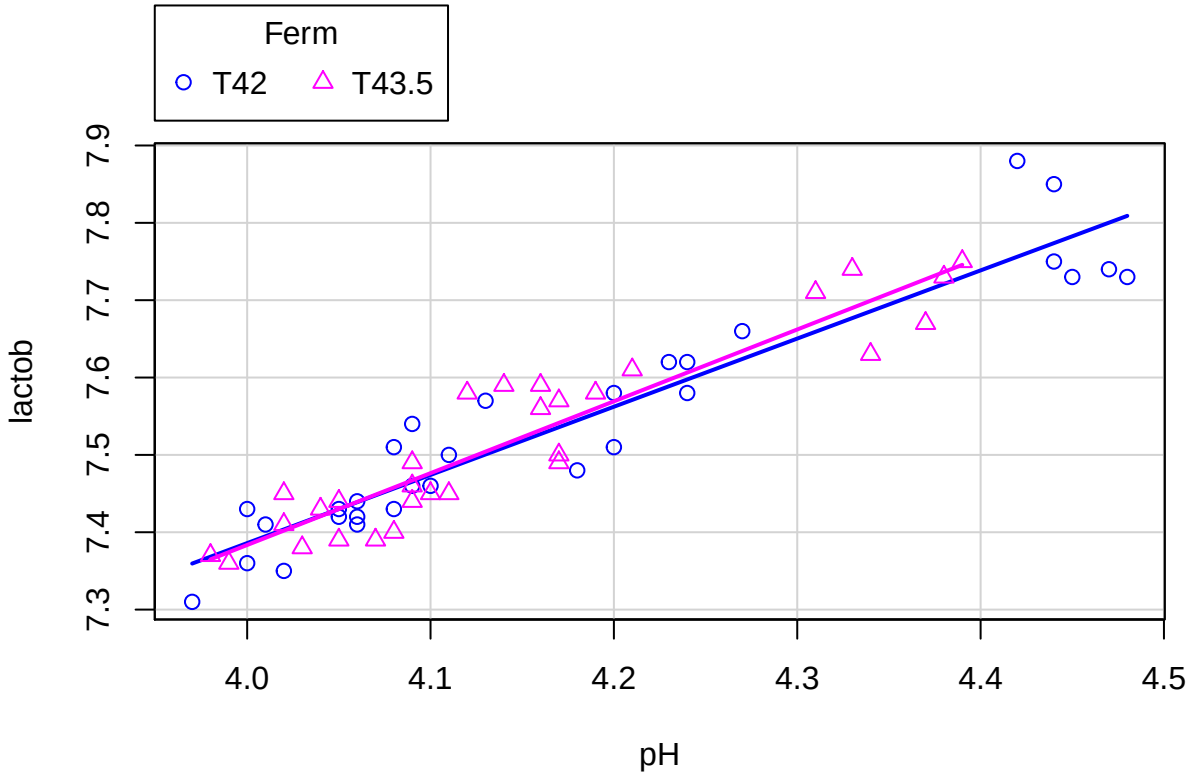

```
par(oldpar) # anula la primera comanda
```

We see that the residues do not follow any pattern. Also in the normality plot we can consider them Normal, because if the data is near the line, we can consider the data Normal. We do not see differences between the variances in the whole range of the fitted values. It seems that the hypothesis of the linear model accomplish.

And now we will do the same analysis for the variable *Lactob*, to be able to predict the level of concentration of *Lactobacillus* knowing just the *pH*.

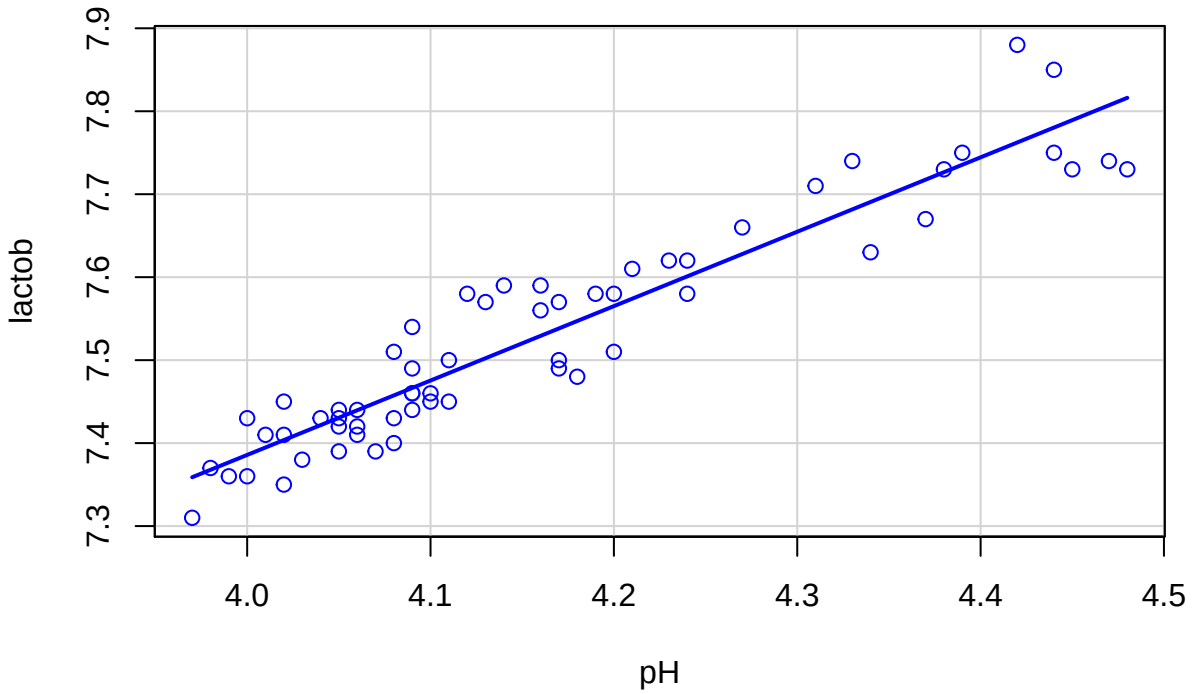
Fit the cloud of points with two lines (each one for each fermentation temperature):

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



Both lines are very similar and it takes us to reconsidering to use just one line, for both fermentation temperatures:

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



Again, positive slope implies that when we increment the pH, the number of *Lactobacillus* that we can find there also increases. Next, we apply a linear model - that way we can obtain the coefficients of the lines explictly:

```
(model_lactob<-lm(lactob~pH,dat))
```

```
##
## Call:
## lm(formula = lactob ~ pH, data = dat)
##
## Coefficients:
## (Intercept)           pH
##      3.7989       0.8967
```

We obtain the coefficients of the line: Intercept: $\alpha = 3.7989$

The coefficient that multiplies pH: $\beta = 0.8967$

$$Lactob = 3.7989 + 0.8967 \cdot pH$$

```
summary(model_lactob)
```

```
##
## Call:
## lm(formula = lactob ~ pH, data = dat)
##
## 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
```

The *summary* resolves the tests upon the coefficients (we can see that at the **Coefficients** section). The first one is about α:

$$H_0: \quad \alpha = 0$$
$$H_1: \quad \alpha \neq 0$$

See that p-value $= 2e-16 < \alpha = 0.05$, for sure H_1! α is not zero meaning that the line does not cross (0,0).

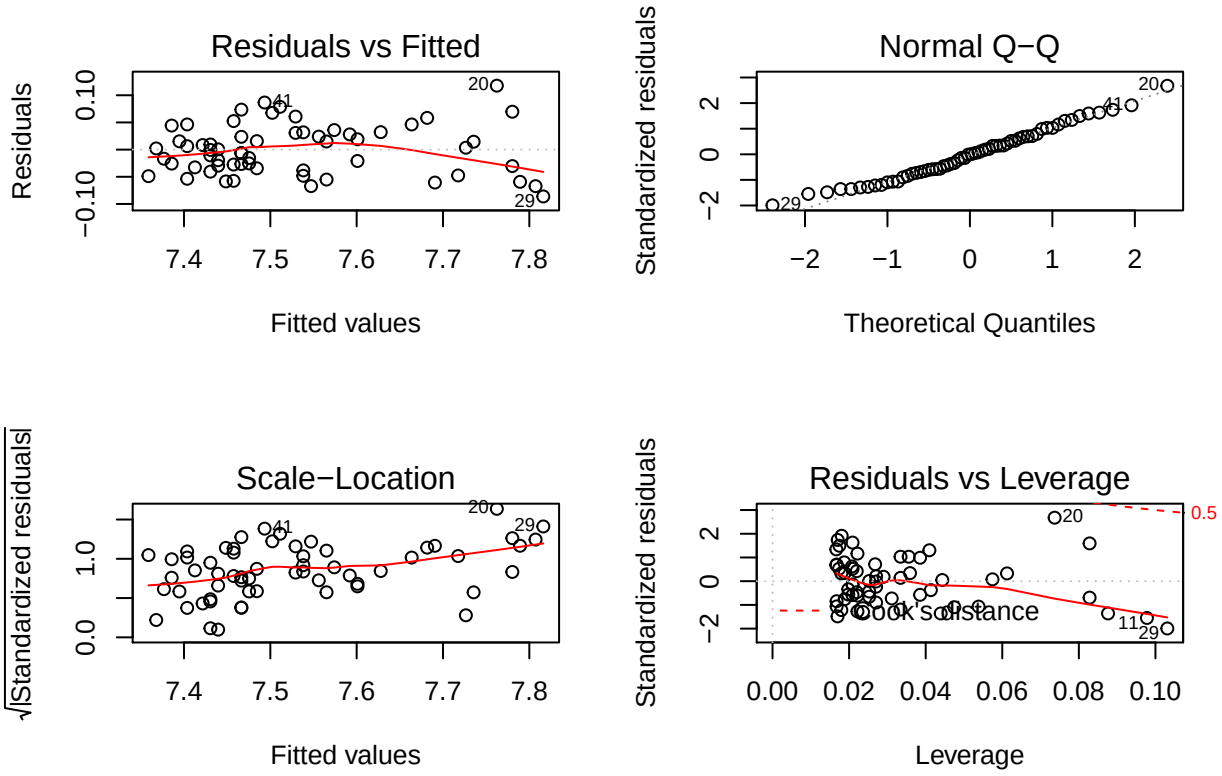
Next text about β:

$$H_0: \quad \beta = 0$$
$$H_1: \quad \beta \neq 0$$

We can see that p-value $= 2e-16 < \alpha = 0.05$, for sure we accept H_1!

As before, we are going to test the hypotheses of Normality, independence and equality of variances of the residues:

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




```
par(oldpar) # anula la primera comanda
```

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.

```
c(strep=summary(model_strep)$r.squared, lactob=summary(model_lactob)$r.squared) ## strep lactob
```

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

```
c(strep=logLik(model_strep),lactob=logLik(model_lactob)) ## strep lactob
```

```
##     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'R2) es lleugerament superior en el segon model indicant que ajustem una mica millor el *Lactob* que l'*Strep*.