Marc Lavielle
February 21st, 2017

1 Introduction

1.1 The orthodont data

The Orthodont data has 108 rows and 4 columns of the change in an orthdontic measurement over time for several young subjects.

Here, distance is a numeric vector of distances from the pituitary to the pterygomaxillary fissure (mm). These distances are measured on x-ray images of the skull.

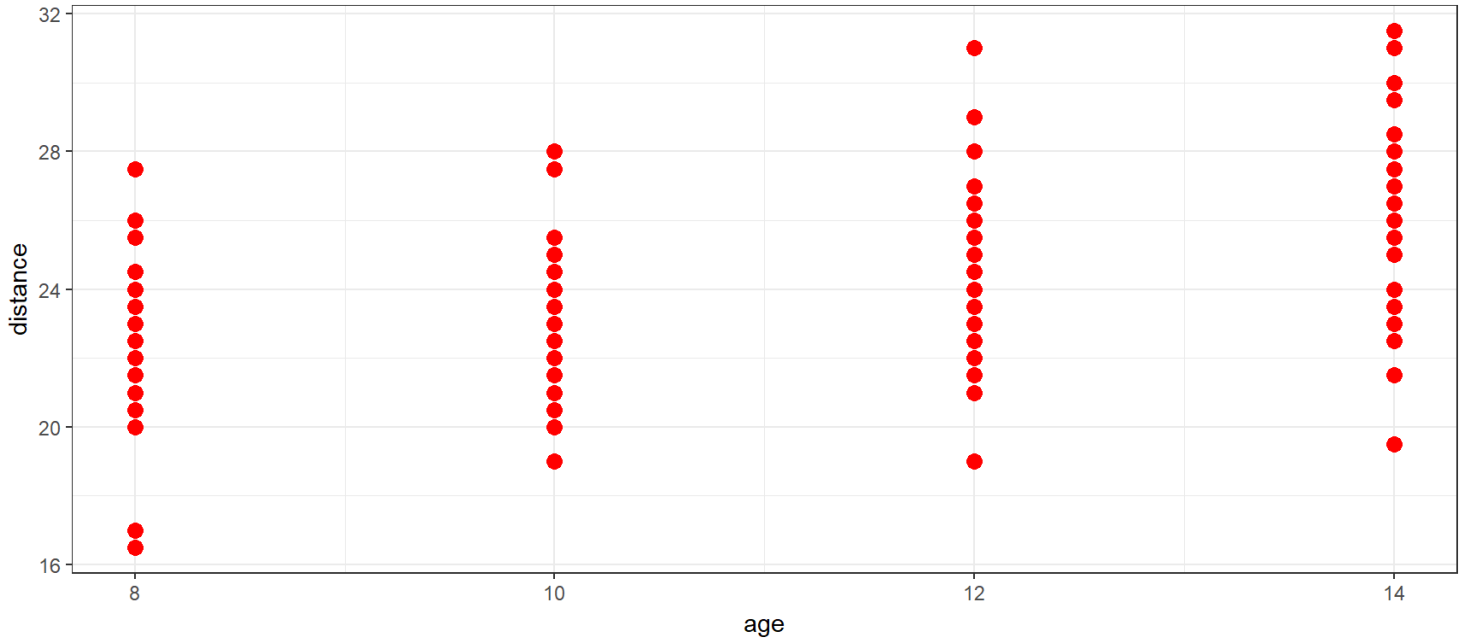
```
data("Orthodont", package="nlme")
head(Orthodont)
```

```
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
```

```
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

Let us plot the data, i.e. the distance versus age:

```
library(ggplot2)
theme_set(theme_bw())
pl <- ggplot(data=Orthodont) + geom_point(aes(x=age,y=distance), color="red", size=3)
pl
```



1.2 Fitting linear models to the data

A linear model by definition assumes there is a linear relationship between the observations $(y_j, 1 \le j \le n)$ and m series of variables $(x_j^{(1)}, \ldots, x_j^{(m)}, 1 \le j \le n)$:

$$y_j = c_0 + c_1 x_j^{(1)} + c_2 x_j^{(2)} + \cdots + c_m x_j^{(m)} + e_j, \qquad 1 \le j \le n,$$

where $(e_j, 1 \le j \le n)$ is a sequence of residual errors.

In our example, the observations $(y_j, 1 \le j \le n)$ are the $n = 108$ measured distances.

We can start by fitting a linear model to these data using age as a regression variable:

$$\text{linear model 1:} \qquad y_j = c_0 + c_1 \times \text{age}_j + e_j$$

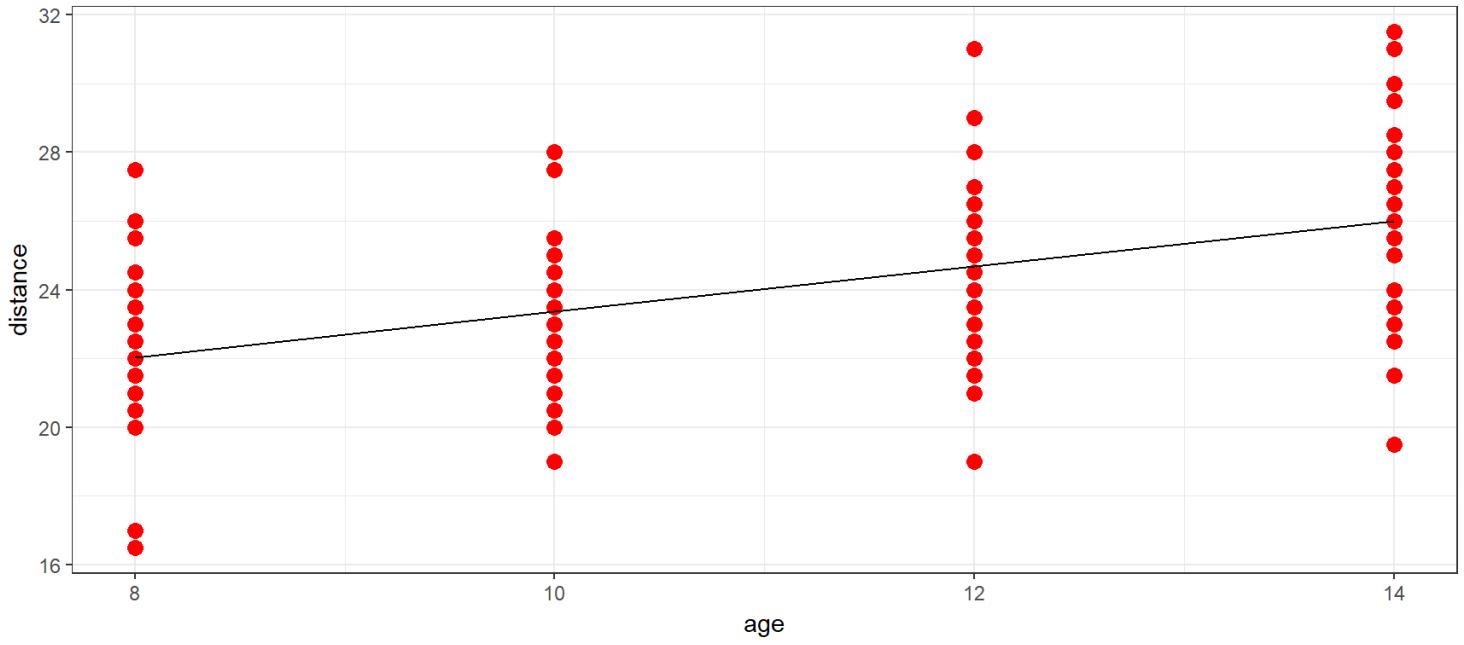
```
lm1 <- lm(distance~age, data=Orthodont)
summary(lm1)
```

```
##
## Call:
```

```
## lm(formula = distance ~ age, data = Orthodont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5037 -1.5778 -0.1833  1.3519  6.3167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.7611     1.2256  13.676  < 2e-16 ***
## age           0.6602     0.1092   6.047 2.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.537 on 106 degrees of freedom
## Multiple R-squared:  0.2565, Adjusted R-squared:  0.2495
## F-statistic: 36.56 on 1 and 106 DF,  p-value: 2.248e-08
```

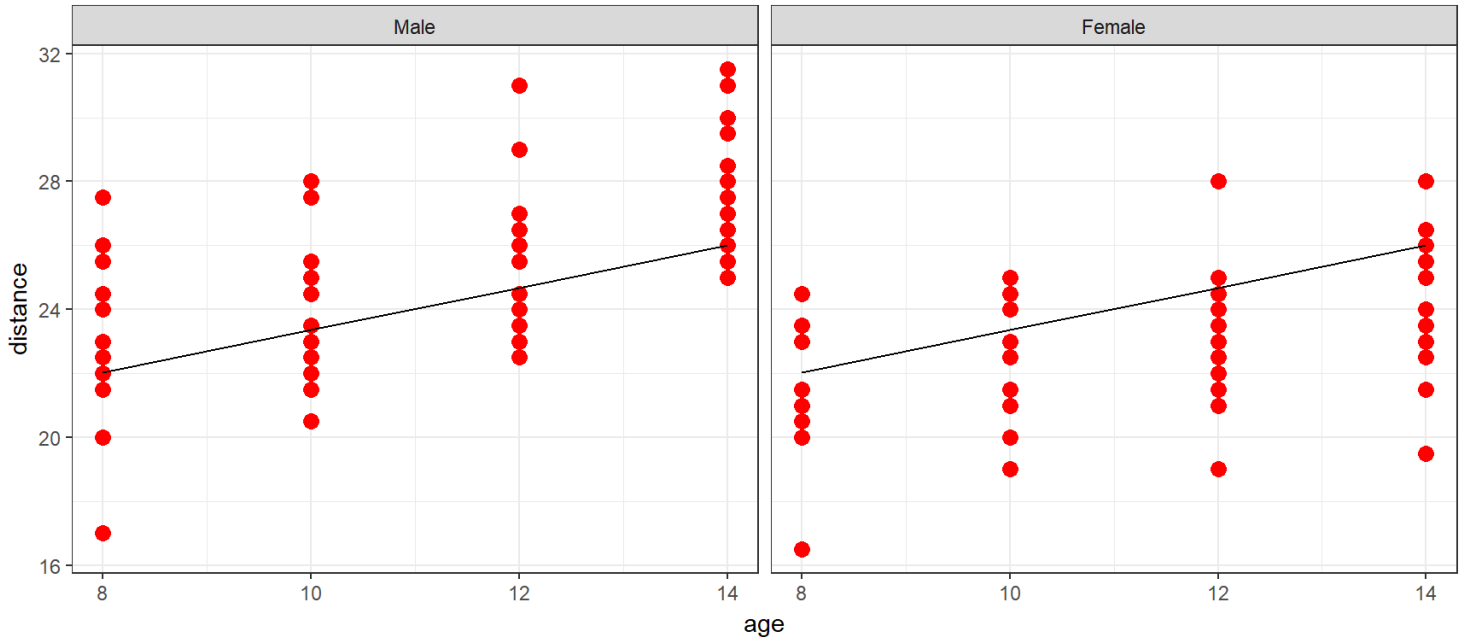
Let us plot the predicted distance $\hat{\alpha}_0 + \hat{\alpha}_1 \times \text{age}$ together with the observed distances

```
pl + geom_line(aes(x=age,y=predict(lm1)))
```



if we now display separately the boys and girls, we see that we are missing something: we underestimate the distance for the boys and overestimate it for the girls:

```
pl + geom_line(aes(x=age,y=predict(lm1))) + facet_grid(.~ Sex )
```



We can then assume the same slope but different intercepts for boys and girls,

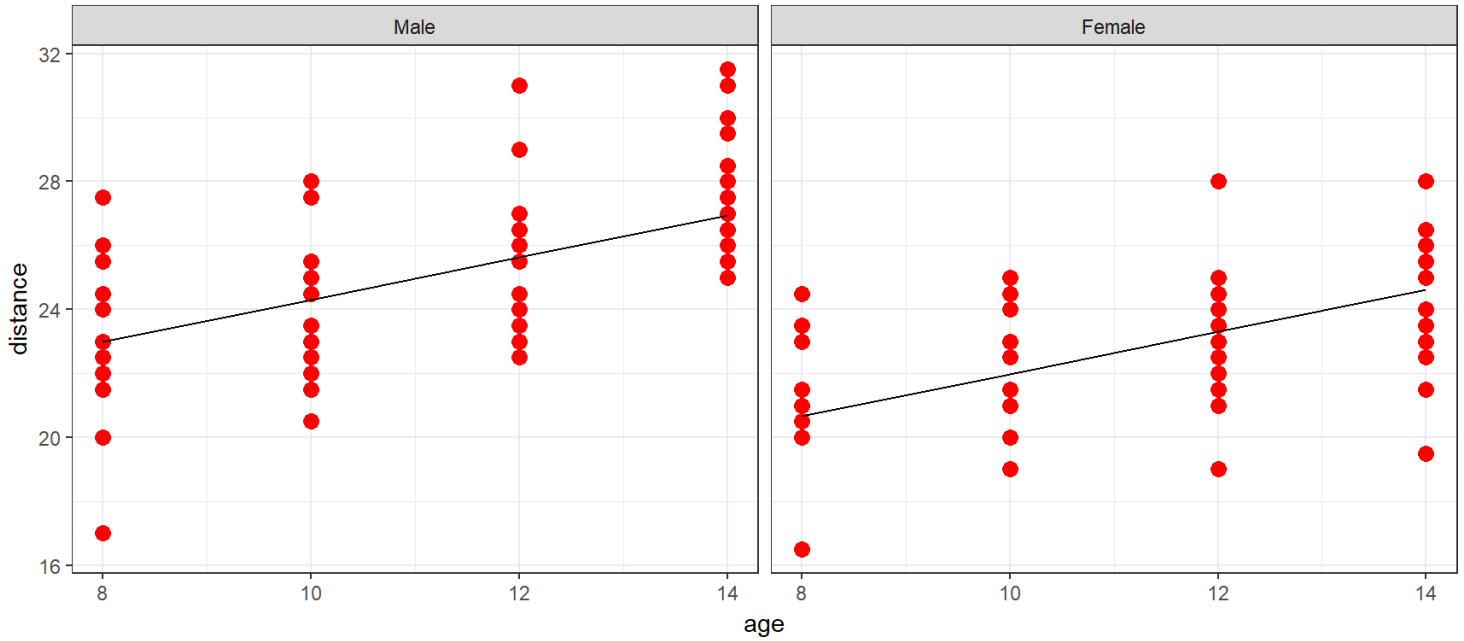
$$\text{linear model 2:} \qquad y_j = c_0 + \delta_{0F} \times 1\mathrm{I}_{\text{Sex}_j=\text{F}} + c_1 \times \text{age}_j + e_j$$

Here, $c_0 = c_{0M}$ is the intercept for the boys and $c_0 + \delta_{0F} = c_{0F}$ the intercept for the girls.

```
lm2 <- lm(distance~age+Sex, data=Orthodont)
summary(lm2)
```

```
##
## Call:
## lm(formula = distance ~ age + Sex, data = Orthodont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9882 -1.4882 -0.0586  1.1916  5.3711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.70671    1.11221  15.920  < 2e-16 ***
## age          0.66019    0.09776   6.753 8.25e-10 ***
## SexFemale   -2.32102    0.44489  -5.217 9.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.272 on 105 degrees of freedom
## Multiple R-squared:  0.4095, Adjusted R-squared:  0.3983
## F-statistic: 36.41 on 2 and 105 DF,  p-value: 9.726e-13
```

```
Orthodont$pred.lm2 <- predict(lm2)
pl + geom_line(data=Orthodont,aes(x=age,y=pred.lm2)) + facet_grid(.~ Sex )
```



We could instead assume the same intercept but different slopes for boys and girls:

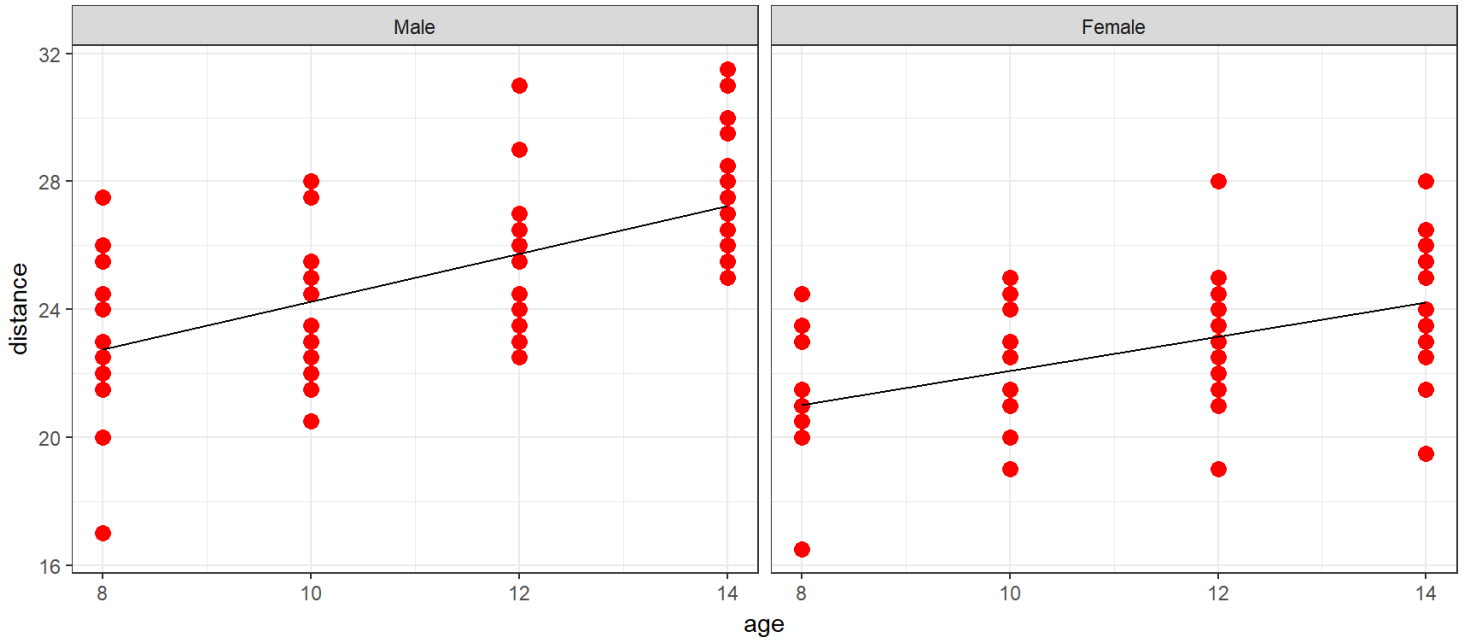
$$\text{linear model 3:} \qquad y_j = c_0 + c_{1M} \times \text{age}_j \times 1\text{I}_{\text{Sex}_j=\text{M}} + c_{1F} \times \text{age}_j \times 1\text{I}_{\text{Sex}_j=\text{F}} + e_j$$

Here, c_{1M} is the slope for the boys and c_{1F} the slope for the girls.

```
lm3 <- lm(distance~age:Sex , data=Orthodont)
summary(lm3)
```

```
##
## Call:
## lm(formula = distance ~ age:Sex, data = Orthodont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7424 -1.2424 -0.1893  1.2681  5.2669
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.76111    1.08613  15.432  < 2e-16 ***
## age:SexMale    0.74767    0.09807   7.624 1.16e-11 ***
## age:SexFemale  0.53294    0.09951   5.355 5.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.249 on 105 degrees of freedom
## Multiple R-squared:  0.4215, Adjusted R-squared:  0.4105
## F-statistic: 38.26 on 2 and 105 DF,  p-value: 3.31e-13
```

```
Orthodont$pred.lm3 <- predict(lm3)
pl + geom_line(data=Orthodont,aes(x=age,y=pred.lm3)) + facet_grid(.~ Sex )
```



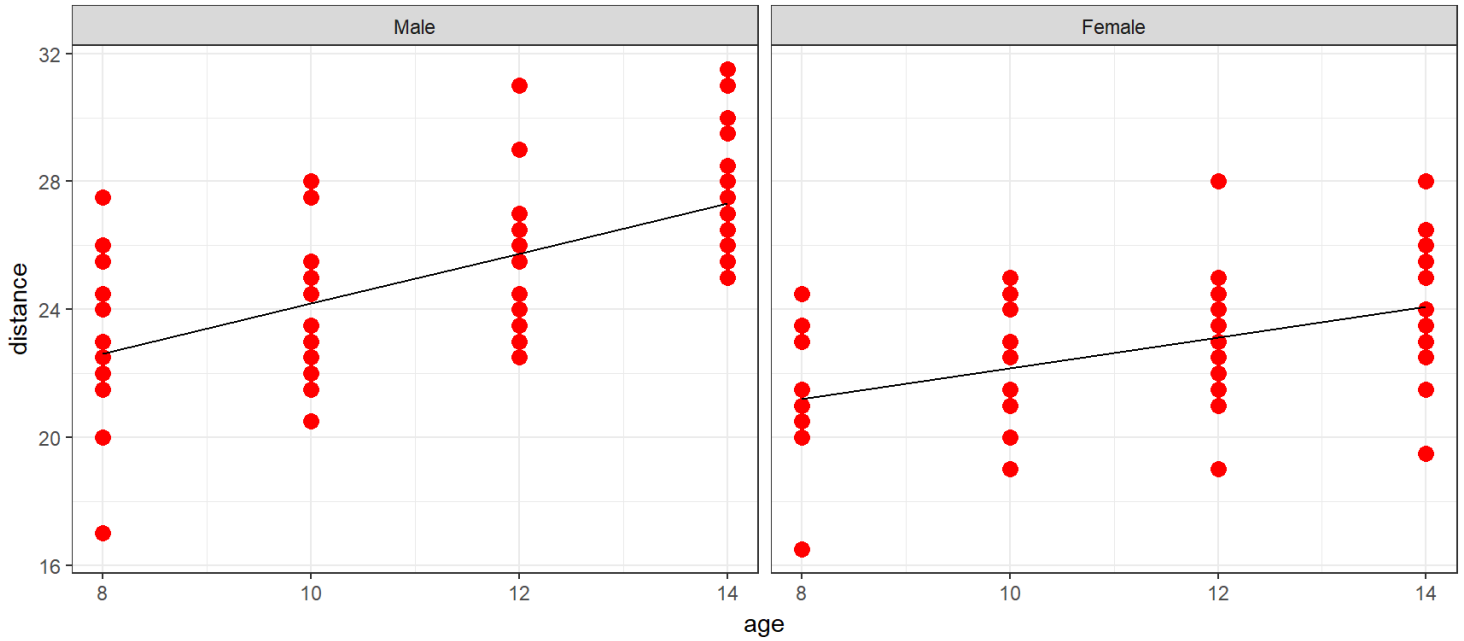
We can also combine these two models by assuming different intercepts and different slopes:

linear model 4: $$y_j = c_0 + \delta_{0F} \times 1\mathrm{I}_{\mathrm{Sex}_j=\mathrm{F}} + c_{1M} \times \mathrm{age}_j \times 1\mathrm{I}_{\mathrm{Sex}_j=\mathrm{M}} + c_{1F} \times \mathrm{age}_j \times 1\mathrm{I}_{\mathrm{Sex}_j=\mathrm{F}} + e_j$$

```
lm4 <- lm(distance~age:Sex+Sex, data=Orthodont)
summary(lm4)
```

```
##
## Call:
## lm(formula = distance ~ age:Sex + Sex, data = Orthodont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6156 -1.3219 -0.1682  1.3299  5.2469
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.3406     1.4162  11.538  < 2e-16 ***
## SexFemale       1.0321     2.2188   0.465  0.64279
## age:SexMale     0.7844     0.1262   6.217 1.07e-08 ***
## age:SexFemale   0.4795     0.1522   3.152  0.00212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.257 on 104 degrees of freedom
## Multiple R-squared:  0.4227, Adjusted R-squared:  0.4061
## F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

```
Orthodont$pred.lm4 <- predict(lm4)
pl + geom_line(data=Orthodont,aes(x=age,y=pred.lm4)) + facet_grid(.~ Sex )
```



Remarks:

- 1. The p-value cannot be used as such since the deign matrix is not orthogonal

```
C <- t(model.matrix(lm4))%*%model.matrix(lm4)
C/sqrt(diag(C)%*%t(diag(C)))
```

```
##              (Intercept) SexFemale age:SexMale age:SexFemale
## (Intercept)    1.0000000 0.6382847   0.7543719     0.6254922
## SexFemale      0.6382847 1.0000000   0.0000000     0.9799579
## age:SexMale    0.7543719 0.0000000   1.0000000     0.0000000
## age:SexFemale  0.6254922 0.9799579   0.0000000     1.0000000
```

```
summary(lm(distance ~ age , data=subset(Orthodont, Sex=="Male")))
```

```
##
## Call:
## lm(formula = distance ~ age, data = subset(Orthodont, Sex ==
##     "Male"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6156 -1.6844 -0.2875  1.2641  5.2469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.3406     1.4544  11.236  < 2e-16 ***
## age           0.7844     0.1296   6.054 9.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.318 on 62 degrees of freedom
```

```
## Multiple R-squared:  0.3715, Adjusted R-squared:  0.3614
## F-statistic: 36.65 on 1 and 62 DF,  p-value: 9.024e-08
```

- 2. A different parameterization for this model could be used in an equivalent way:

$$\text{linear model 4':} \qquad y_j = c_0 + \delta_{0F} \times \mathbb{1}_{\text{Sex}_j=\text{F}} + (c_1 + \delta_{1F} \times \text{age}_j \times \mathbb{1}_{\text{Sex}_j=\text{F}}) \times \text{age}_j + e_j$$

where $c_1 = c_{1M}$ and $c_1 + \delta_{1F} = c_{1F}$. This model is implemented by defining the interaction between Sex and age as age*Sex:

```
lm4b <- lm(distance~age*Sex, data=Orthodont)
summary(lm4b)
```

```
##
## Call:
## lm(formula = distance ~ age * Sex, data = Orthodont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6156 -1.3219 -0.1682  1.3299  5.2469
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.3406     1.4162  11.538  < 2e-16 ***
## age             0.7844     0.1262   6.217 1.07e-08 ***
## SexFemale       1.0321     2.2188   0.465    0.643
## age:SexFemale  -0.3048     0.1977  -1.542    0.126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.257 on 104 degrees of freedom
## Multiple R-squared:  0.4227, Adjusted R-squared:  0.4061
## F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

Both prametrizations give the same predictions

```
cbind(head(predict(lm4)), head(predict(lm4b)))
```

```
##       [,1]     [,2]
## 1 22.61562 22.61563
## 2 24.18437 24.18437
## 3 25.75312 25.75312
## 4 27.32187 27.32187
## 5 22.61562 22.61563
## 6 24.18437 24.18437
```

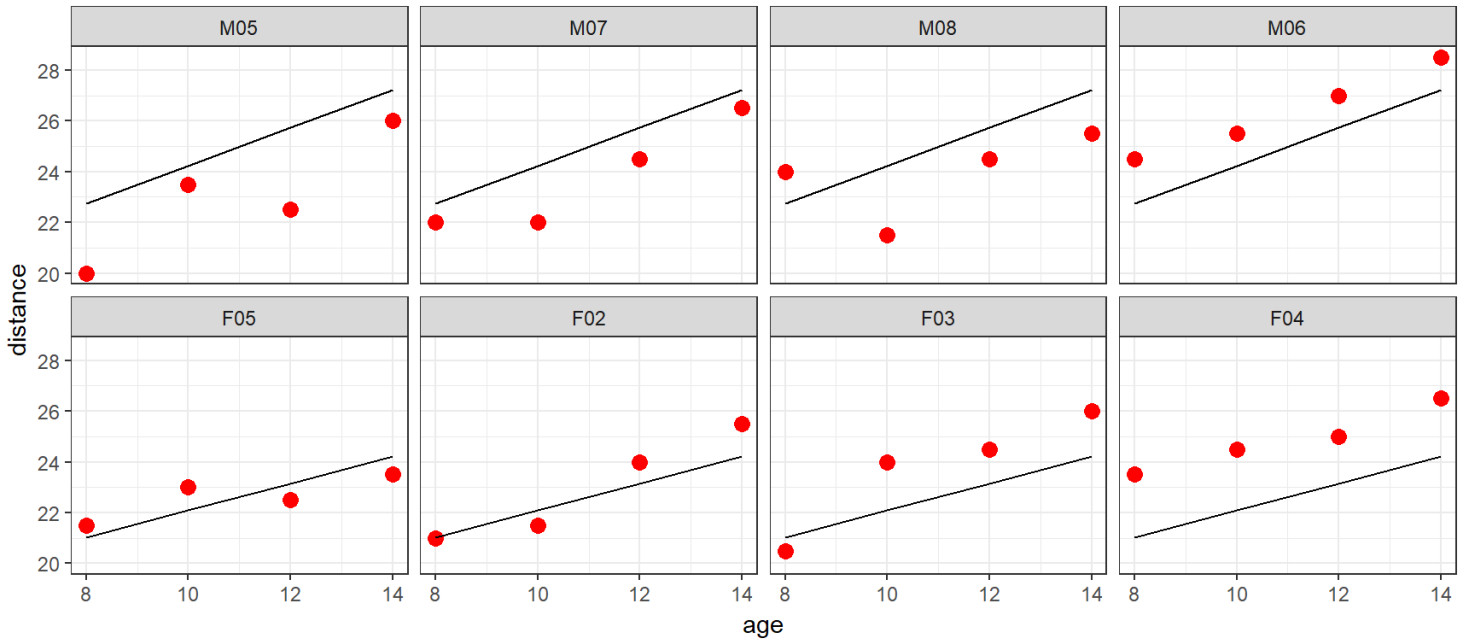
Different criteria for model selection, including BIC, seem to prefer `lm3`.

```
BIC(lm1,lm2,lm3, lm4)
```

```
##      df      BIC
## lm1   3 519.6234
## lm2   4 499.4121
## lm3   4 497.1948
## lm4   5 501.6524
```

Should we then consider that `lm3` is our final model? Let us look at the *individual fits* for 8 subjects,

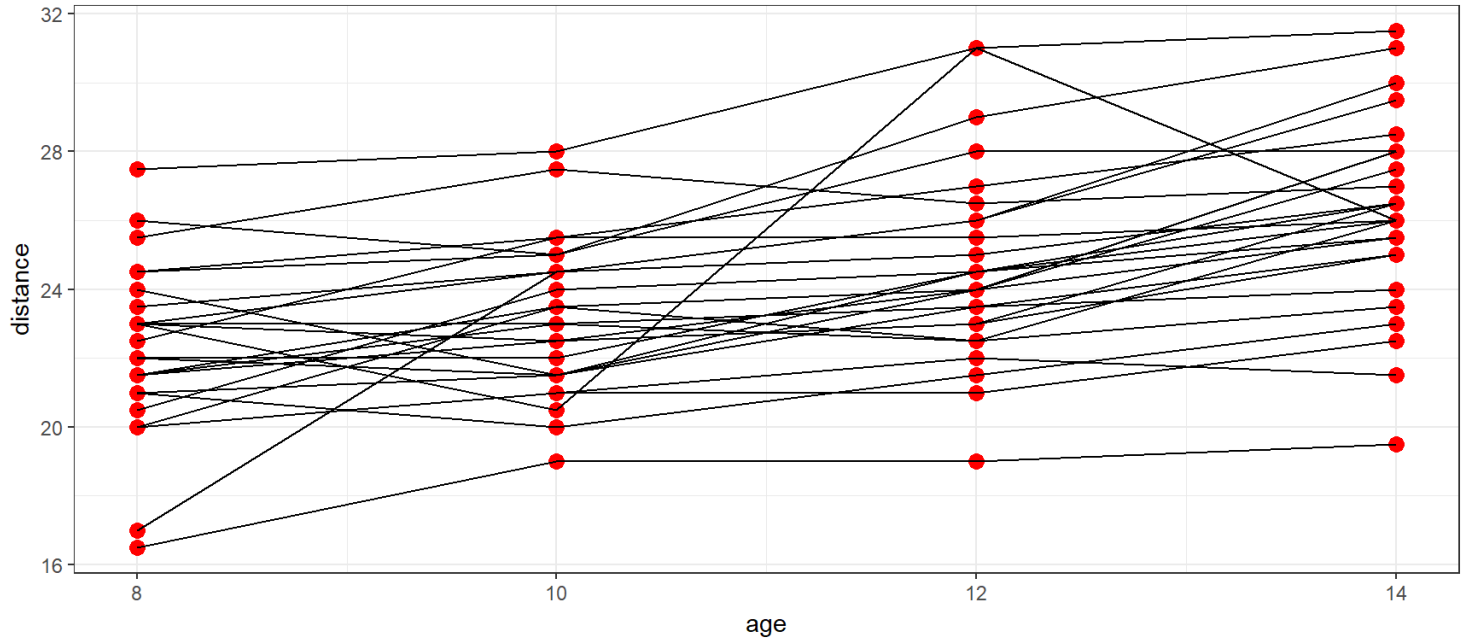
```
Subject.select <- c(paste0("M0",5:8),paste0("F0",2:5))
Orthodont.select <- subset(Orthodont,Subject %in% Subject.select)
ggplot(data=Orthodont.select) + geom_point(aes(x=age,y=distance), color="red", size=3) +
  geom_line(aes(x=age,y=predict(lm3,newdata=Orthodont.select))) + facet_wrap(~Subject, nrow=2)
```



We see that the model for the boys, respectively for the girls, seems to underestimate or overestimate the individual data of the four boys, respectively the four girls.

Indeed, we didn't take into account the fact that the data are *repeated measurements* made on the same subjects. A more convenient plot for this type of data consists in joining the data of a same individual:

```
library(ggplot2)
theme_set(theme_bw())
ggplot(data=Orthodont) + geom_point(aes(x=age,y=distance), color="red", size=3)  +
  geom_line(aes(x=age,y=distance,group=Subject)) # +  facet_grid(~Sex)
```



We see on this plot, that even if the distance seems to increase linearly for each individual, the intercept and the slope may change from a subject to another one, including within the same Sex group.

We therefore need to extend our linear model in order to take into account this *inter-individual* variability.

2 Mathematical definition of a linear mixed effects models

The linear model introduced above concerns a single individual. Suppose now that a study is based on N individuals and that we seek to build a global model for all the collected observations for the N individuals. We will denote y_{ij} the jth observation taken of individual i and $x_{ij}^{(1)}, \ldots, x_{ij}^{(m)}$ the values of the m explanatory variables for individual i. If we assume the parameters of the model can vary from one individual to another, then for any subject i, $1 \leq i \leq N$, the linear model becomes

$$y_{ij} = c_{i0} + c_{i1} x_{ij}^{(1)} + c_{i2} x_{ij}^{(2)} + \cdots + c_{im} x_{ij}^{(m)} + e_{ij}, \quad 1 \leq j \leq n_i.$$

Suppose to begin with that each individual parameter c_{ik} can be additively broken down into a fixed component β_k and an individual component η_{ik}, i.e.,

$$c_{ik} = \beta_k + \eta_{ik}$$

where η_{ik} represents the deviation of c_{ik} from the ``typical'' value β_k in the population for individual i.

where η_{ik} is a normally distributed random variable with mean 0.

Using this parametrization, the model becomes

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^{(1)} + \cdots + \beta_m x_{ij}^{(m)} + \eta_{i0} + \eta_{i1} x_{ij}^{(1)} + \ldots + \eta_{im} x_{ij}^{(m)} + e_{ij}.$$

We can then rewrite the model in matrix form:

$$y_i = X_i \beta + X_i \eta_i + e_i,$$

where

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} \quad , \quad X_i = \begin{pmatrix} 1 & x_{i1}^{(1)} & \cdots & x_{i1}^{(m)} \\ 1 & x_{i2}^{(1)} & \cdots & x_{i2}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{in}^{(1)} & \cdots & x_{in}^{(m)} \end{pmatrix} \quad , \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \quad , \quad \eta_i = \begin{pmatrix} \eta_{i0} \\ \eta_{i1} \\ \vdots \\ \eta_{im} \end{pmatrix} \quad , \quad e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{pmatrix}$$

Here, y_i is the n_i vector of observations for individual i, X_i is the $n_i \times d$ design matrix (with $d = m + 1$), β is a d-vector of **fixed effects** (i.e. common to all individuals of the population), η_i is a d-vector of **random effects** (i.e. specific to each individual) and e_i is a n_i-vector of residual errors.

The model is called **linear mixed effects model** because it is a linear combination of fixed and random effects.

The random effects are assumed to be normally distributed in a linear mixed effects model

$$\eta_i \underset{\text{i.i.d.}}{\sim} \mathrm{N}(0_d \ , \ \Omega)$$

Ω is the $d \times d$ variance-covariance matrix of the random effects. This matrix is diagonal if the components of η_i are independent.

The vector of residual errors e_i is also normally distributed:

$$e_i \underset{\text{i.i.d.}}{\sim} \mathrm{N}(0_{n_i} \ , \ \Sigma_i)$$

The particular case of a diagonal matrix with constant diagonal terms, i.e. $\Sigma_i = \sigma^2 I_{n_i}$, means that, for any individual i, the residual errors $(e_{ij}, 1 \leq j \leq n_i)$ are independent and identically distributed:

$$e_{ij} \underset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$$

We can extend this model to models invoking more complicated design matrices that may even differ for fixed and random effects:

$$y_i = X_i \beta + A_i \eta_i + e_i$$

As an example, consider the following model

$$\begin{aligned} y_{ij} &= c_{i0} + c_{i1} x_{ij}^{(1)} + c_{i2} x_{ij}^{(2)} + e_{ij} \\ &= \beta_0 + \beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \eta_{i0} + \eta_{i1} x_{ij}^{(1)} + \eta_{i2} x_{ij}^{(2)} + e_{ij} \end{aligned}$$

The variance covariance matrix Ω of the vector of random effects $(\eta_{i0}, \eta_{i1}, \eta_{i2})$ is a 3×3 matrix.

Assume now that parameter $c_i 2$ does not vary from one individual to another. Then $c_i 2 = \beta_2$ for all i which means that $\eta_{i2} = 0$ for all i. A null variance for η_{i2} means that Ω_{33}, the third diagonal term of Ω is 0.

Instead of considering a variance-covariance matrix Ω with null diagonal terms, it is more convenient to rewrite the model a follows

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + \eta_{i0} + \eta_{i1} x_{ij}^{(1)} + e_{ij} \quad , \quad 1 \leq j \leq n_i$$

or, in a matricial form, as

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & x_{i1}^{(1)} & x_{i1}^{(2)} \\ 1 & x_{i2}^{(1)} & x_{i2}^{(m)} \\ \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & x_{i1}^{(1)} \\ 1 & x_{i2}^{(1)} \\ \vdots & \vdots \end{pmatrix} \begin{pmatrix} \eta_{i0} \\ \eta_{i1} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \end{pmatrix}$$

$$\begin{pmatrix} y_{in_i} \end{pmatrix} \quad \begin{pmatrix} 1 & x_{in}^{(1)} & x_{in}^{(m)} \end{pmatrix} \quad \beta_2 \quad \begin{pmatrix} 1 & x_{in}^{(1)} \end{pmatrix} \quad \begin{pmatrix} e_{in} \end{pmatrix}$$

Ω is now the 2×2 variance-covariance matrix of $(\eta_{i0}, \eta_{i1})'$.

3 Statistical inference in linear mixed effects models

3.1 Estimation of the population parameters

The model parameters are the vector of fixed effects β, the variance-covariance matrix Ω of the random effects and the variance σ^2 of the residual errors (assuming *i.i.d.* residual errors).

Let $\theta = (\beta, \Omega, \sigma^2)$ be the set of model parameters.

We easily deduce from the matricial representation of the model $y_i = X_i \beta + A_i \eta_i + e_i$, that y_i is normally distributed:

$$y_i \sim \mathrm{N}(X_i \beta, A_i \Omega A_i' + \sigma^2 I_{n_i})$$

Let $y = (y_i, 1 \leq i \leq N)$ be the set of observations for the N individuals. The **maximum likelihood (ML)** estimator of θ maximizes the log-likelihood function defined as

$$\begin{aligned} \mathrm{LL}(\theta) &= \log(\mathrm{p}(y\,;\,\theta)) \\ &= \sum_{i=1}^{N} \log(\mathrm{p}(y_i\,;\,\theta)) \\ &= \sum_{i=1}^{N} \{-\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(|A_i \Omega A_i' + \sigma^2 I_{n_i}|) - \frac{1}{2}(y_i - X_i\beta)'(A_i \Omega A_i' + \sigma^2 I_{n_i})^{-1}(y_i - X_i\beta)\} \end{aligned}$$

There is no analytical solution to this maximization problem. Nevertheless, numerical methods such as the Newton-Raphson and the EM algorithms, can be used for maximizing $\mathrm{LL}(\theta)$.

The **restricted maximum likelihood (REML)** approach is a variant of the ML approach. In contrast to the earlier maximum likelihood estimation, REML can produce unbiased estimates of variance and covariance parameters.

Consider the linear model $y = X\beta + e$ as an example where β is a d-vector of unknown coefficients and where $e_j \sim_{\text{i.i.d.}} \mathrm{N}(0, \sigma^2)$ for $1 \leq j \leq n$. Both the ML and the REML estimators of β reduce to the least-squares estimator $\hat{\beta} = (X'X)^{-1}X'y$, but the estimator of the variance component σ^2 differs according to the method:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\|y - X\hat{\beta}\|^2 \quad ; \quad \hat{\sigma}^2_{\mathrm{REML}} = \frac{1}{n-d}\|y - X\hat{\beta}\|^2$$

Standard errors (se) of the parameter estimate $\hat{\theta}$ can be obtained by computing the Fisher information matrix

$$I(\hat{\theta}) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log \mathrm{p}(y;\hat{\theta})\right)$$

Then, the standard errors are the square roots of the diagonal elements of the inverse matrix of $I(\hat{\theta})$.

3.2 Estimation of the individual parameters

3.2.1 Estimation of the random effects

Individual parameters for individual i are the individual coefficients $(c_{ik}, 0 \leq k \leq m)$.

Once the set of population parameters $\theta = (\beta, \Omega, \sigma^2)$ has been estimated, the ℓ-vector of nonzero random effects η_i can be estimated using the conditional distribution $\mathrm{p}(\eta_i \mid y_i \,;\, \hat{\theta})$.

Since the marginal distributions of y_i and η_i are both Gaussian, this conditional distribution is also Gaussian with a mean and a variance that can be computed. Indeed, from Bayes Theorem,

$$\begin{aligned}\mathrm{p}(\eta_i \mid y_i \,;\, \theta) &= \frac{\mathrm{p}(y_i \mid \eta_i \,;\, \theta)\mathrm{p}(\eta_i \,;\, \theta)}{\mathrm{p}(y_i \,;\, \theta)} \\ &= \frac{(2\pi\sigma^2)^{-\frac{n_i}{2}}\,(2\pi)^{-\frac{\ell}{2}}\,|\Omega|^{-\frac{1}{2}}}{(2\pi)^{-\frac{n_i}{2}}\,|A_i\,\Omega A_i' + \sigma^2 I_{n_i}|^{-\frac{1}{2}}}\;\frac{\exp\{-\frac{1}{2\sigma^2}\|y_i - X_i\beta - A_i\eta_i\|^2 - \frac{1}{2}\eta_i'\Omega^{-1}\eta_i\}}{\exp\{-\frac{1}{2}(y_i - X_i\beta)'(A\Omega A' + \Sigma)^{-1}(y_i - X_i\beta)\}}\end{aligned}$$

Then, we can show that

$$\mathrm{p}(\eta_i \mid y_i \,;\, \theta) = (2\pi)^{-\frac{\ell}{2}}\,|\Gamma_i|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(\eta_i - \mu_i)'\Gamma_i^{-1}(\eta_i - \mu_i)}$$

where

$$\Gamma_i = \left(\frac{A_i'A_i}{\sigma^2} + \Omega^{-1}\right)^{-1} \quad ; \quad \mu_i = \frac{\Gamma_i A_i'(y_i - X_i\beta)}{\sigma^2}$$

We can therefore estimate the conditional mean μ_i and the conditional variance Γ_i of η_i using these formulas and the estimated parameters $\hat{\beta}$, $\hat{\Omega}$ and $\hat{\sigma}^2$:

$$\hat{\Gamma}_i = \left(\frac{A_i'A_i}{\hat{\sigma}^2} + \hat{\Omega}^{-1}\right)^{-1} \quad ; \quad \hat{\mu}_i = \frac{\hat{\Gamma}_i A_i'(y_i - X_i\hat{\beta})}{\hat{\sigma}^2}$$

Since the conditional distribution of η_i is Gaussian, $\hat{\mu}_i$ is also the conditional mode of this distribution. This estimator of η_i is the so-called **maximum a posteriori** (MAP) estimator of η_i . It is also called **empirical Bayes estimator** (EBE).

3.2.2 Deriving individual parameter estimates and individual predictions

Estimation of the d individual parameters is straightforward once the ℓ nonzero random effects have been estimated:

$$\hat{c}_{ik} = \begin{cases} \hat{\beta}_k & \text{if } \eta_{ik} \equiv 0 \\ \hat{\beta}_k + \hat{\eta}_{ik} & \text{otherwise} \end{cases}$$

We see that, for a parameter c_{ik} with no random component (i.e. $\eta_{ik} \equiv 0$, $\hat{c}_{ik} = \hat{\beta}_k$ is the maximum likelihood estimator of c_{ik} , i.e. the parameter value that maximizes the likelihood of making the observations.

On the other hand, if c_{ik} is a random parameter ($\eta_{ik} \neq 0$), then $\hat{c}_{ik} = \hat{\beta}_k + \hat{\eta}_{ik}$ is the MAP estimator of c_{ik} , i.e. the most likely value of c_{ik} , given the observations y_i and its estimated prior distribution.

For any set of explanatory variable $(x^{(1)}, x^{(2)}, \ldots x^{(m)})$, individual prediction of the response variable is then obtained using the individual estimated parameters:

$$\hat{y}_i = \hat{c}_{i0} + \hat{c}_{i1}x^{(1)} + \hat{c}_{i2}x^{(2)} + \cdots + \hat{c}_{im}x^{(m)}$$

3.2.3 About the MAP estimator in a linear mixed effects model

As an example, consider a model where all the individual parameters are random parameters:

$$y_i = X_i c_i + e_i$$

where $c_i = \beta + \eta_i \sim \mathrm{N}(\beta\,,\,\Omega)$.

Then, the conditional distribution of c_i given y_i is also a normal distribution:

$$c_i | y_i \sim \mathrm{N}(m_i, \Gamma_i)$$

where

$$\begin{aligned} m_i &= \mu_i + \beta \\ &= \Gamma_i \left(\frac{X_i'}{\sigma^2} y_i + \Omega^{-1}\beta\right) \\ &= \left(\frac{X_i' X_i}{\sigma^2} + \Omega^{-1}\right)^{-1} \left(\frac{X_i' X_i}{\sigma^2} (X_i' X_i)^{-1} X_i' y_i + \Omega^{-1}\beta\right) \end{aligned}$$

We see that the MAP estimator of c_i is a weighted average of the least square estimator of c_i, $(X_i' X_i)^{-1} X_i' y_i$, which maximizes the conditional distribution of the observations $\mathrm{p}(y_i | c_i, \theta)$, and β which maximizes the prior distribution of c_i. The relative weights of these two terms depend on the design and the parameters of the model:

- A lot of information about c_i in the data and small residual errors will make $(X_i' X_i)/\sigma^2$ large: the estimate of c_i will be close to the least-square estimate which only depends on the observations.
- A very informative prior will make Ω^{-1} large: the estimate of c_i will be close to the prior mean β.

4 Fitting linear mixed effects models to the orthodont data

4.1 Fitting a first model

A first linear mixed effects model assumes that the birth distance and the growth rate (i.e. the intercept and the slope) may depend on the individual:

$$\begin{aligned} \text{lmem:} \qquad y_{ij} &= c_{i0} + c_{i1} \times \text{age}_{ij} + e_{ij} \\ &= \beta_0 + \beta_1 \times \text{age}_{ij} + \eta_{i0} + \eta_{i1} \times \text{age}_{ij} + e_{ij} \end{aligned}$$

We can use the function `lmer` for fitting this model. By default, the restricted mximum likelihood (REML) method is used.

```
library(lme4)
lmem <- lmer(distance ~ age + (age|Subject), data = Orthodont)
summary(lmem)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + (age | Subject)
##    Data: Orthodont
##
```

```
## REML criterion at convergence: 442.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2231 -0.4938  0.0073  0.4722  3.9160
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 5.41509  2.3270
##           age         0.05127  0.2264   -0.61
##  Residual             1.71620  1.3100
## Number of obs: 108, groups:  Subject, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 16.76111    0.77525  21.620
## age          0.66019    0.07125   9.265
##
## Correlation of Fixed Effects:
##     (Intr)
## age -0.848
```

The estimated fixed effects are

$$\hat{\beta}_0 = 16.7611 \quad , \quad \hat{\beta}_1 = 0.66019$$

The standard errors and correlation of these estimates are

$$\text{se}(\hat{\beta}_0) = 0.77525 \quad , \quad \text{se}(\hat{\beta}_1) = 0.07125 \quad , \quad \text{corr}(\hat{\beta}_0, \hat{\beta}_1) = -0.848$$

The estimated standard deviations and correlation of the random effects are

$$\hat{\text{sd}}(\eta_{i0}) = 2.3270 \quad , \quad \hat{\text{sd}}(\eta_{i1}) = 0.2264 \quad , \quad \hat{\text{corr}}(\eta_{i0}, \eta_{i1}) = -0.61$$

The estimated variance-covariance matrix of the random effects is therefore

$$\hat{\Omega} = \begin{pmatrix} 5.41509 & -0.32137 \\ -0.32137 & 0.05127 \end{pmatrix}$$

Finally, the estimated variance of the residual errors is

$$\hat{\sigma}^2 = 1.71620$$

Note that functions `fixef` and `VarCorr` return these estimated parameters:

```
(psi.pop <- fixef(lmem))
```

```
## (Intercept)         age
##  16.7611111   0.6601852
```

```
(Omega <- VarCorr(lmem)$Subject[,])
```

```
##             (Intercept)         age
## (Intercept)    5.415091 -0.32106096
## age           -0.321061  0.05126957
```

```
(sigma2 <- attr(VarCorr(lmem), "sc")^2)
```

```
## [1] 1.716204
```

The estimated individual parameters for our 8 selected individuals can be obtained using function `coef`

```
coef(lmem)$Subject[Subject.select,]
```

```
##     (Intercept)       age
## M05    15.58444 0.6857855
## M06    17.97875 0.7433764
## M07    16.15314 0.6950852
## M08    17.62141 0.5654489
## F02    15.74926 0.6700432
## F03    15.98832 0.7108276
## F04    17.83027 0.6303230
## F05    17.27792 0.4922275
```

using the formula obtained in the previous section, we can check that these estimated parameters are the empirical Bayes estimates, i.e. the conditional means of the individual parameters,

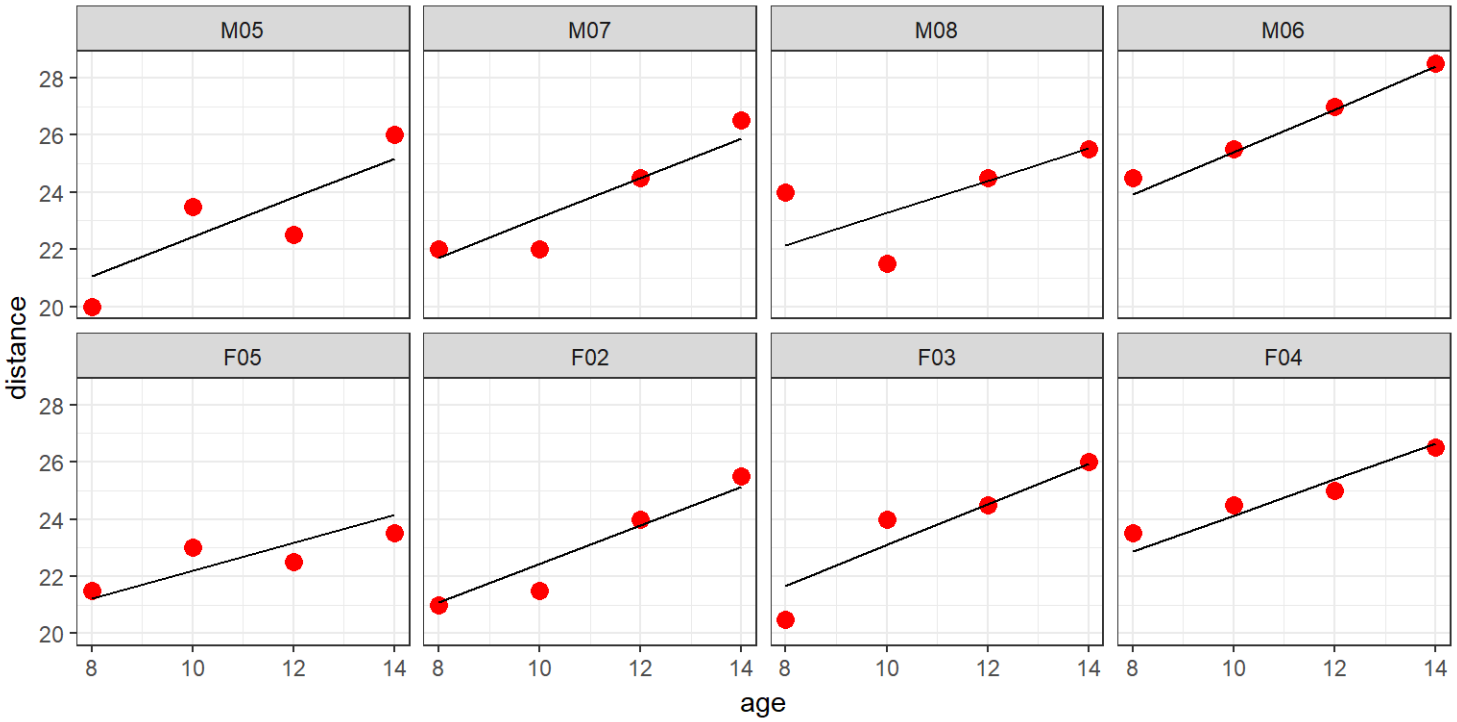
```
Orthodont.i <- Orthodont[Orthodont$Subject=="M05",]
yi <- Orthodont.i$distance
Ai <- cbind(1,Orthodont.i$age)
iO <- solve(Omega)
Gammai <- solve(t(Ai)%*%Ai/sigma2 + iO)
mui <- Gammai%*%(t(Ai)%*%yi/sigma2 + iO%*%psi.pop)
mui
```

```
##                  [,1]
## (Intercept) 15.5844433
## age          0.6857855
```

Individual predicted distances can also be computed and plotted with the observed distances

```
Orthodont$pred.lmem <- fitted(lmem)
ggplot(data=subset(Orthodont,Subject %in% Subject.select)) + geom_point(aes(x=age,y=distance),
```

```
color="red", size=3) +
 geom_line(aes(x=age,y=pred.lmem)) + facet_wrap(~Subject, ncol=4)
```



We can check that the predicted distances for a given individual ("M05" for instance)

```
subset(Orthodont,Subject == "M05")
```

```
## Grouped Data: distance ~ age | Subject
##    distance age Subject  Sex pred.lm2 pred.lm3 pred.lm4 pred.lmem
## 17     20.0   8     M05 Male 22.98819 22.74244 22.61562  21.07073
## 18     23.5  10     M05 Male 24.30856 24.23777 24.18437  22.44230
## 19     22.5  12     M05 Male 25.62894 25.73310 25.75312  23.81387
## 20     26.0  14     M05 Male 26.94931 27.22843 27.32187  25.18544
```

are given by the linear model $c_0 + c_1 \text{age}$ using the individual estimated parameters

$$\hat{\text{distance}}_i = \hat{c}_{i0} + \hat{c}_{i1} \times \text{age}$$

```
mui[1] + mui[2]*c(8,10,12,14)
```

```
## [1] 21.07073 22.44230 23.81387 25.18544
```

4.2 Some extensions of this first model

- We can fit the same model to the same data via maximum likelihood (ML) instead of REML

```
lmer(distance ~ age + (age|Subject), data = Orthodont, REML=FALSE)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: distance ~ age + (age | Subject)
##    Data: Orthodont
##       AIC       BIC    logLik  deviance  df.resid
##  451.2116  467.3044 -219.6058  439.2116       102
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 2.1941
##           age         0.2149   -0.58
##  Residual             1.3100
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
## (Intercept)          age
##     16.7611       0.6602
```

The estimated fixed effects are the same with the two methods. The variance components slightly differ since REML provides an unbiased estimate of Ω and σ^2.

- By default, the variance-covariance matrix Ω is estimated as a full matrix, assuming that the random effects are correlated. It is possible with `lmer` to constrain Ω to be a diagonal matrix by defining the random effects model using `||` instead of `|`

```
lmer(distance ~ age + (age||Subject), data = Orthodont)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + ((1 | Subject) + (0 + age | Subject))
##    Data: Orthodont
## REML criterion at convergence: 443.3146
## Random effects:
##  Groups    Name        Std.Dev.
##  Subject   (Intercept) 1.3860
##  Subject.1 age         0.1493
##  Residual              1.3706
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
## (Intercept)          age
##     16.7611       0.6602
```

4.3 Fitting other models

The mixed effects model combines a model for the fixed effects and a model for the random effects. Let us see some possible combinations.

- In this model, we assume that *i)* the birth distance, is the same in average for boys and girls but randomly varies between individuals, *ii)* the distance increases with the same rate for all the individuals. Here is the mathematical representation of this model:

$$\begin{aligned} y_{ij} &= c_{i0} + \beta_1 \times \text{age}_{ij} + e_{ij} \\ &= \beta_0 + \beta_1 \times \text{age}_{ij} + \eta_{i0} + e_{ij} \end{aligned}$$

```
lmer(distance ~ age + (1|Subject), data = Orthodont)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + (1 | Subject)
##    Data: Orthodont
## REML criterion at convergence: 447.0025
## Random effects:
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 2.115
##  Residual             1.432
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
## (Intercept)          age
##     16.7611       0.6602
```

- We extend the previous model, assuming now different mean birth distances and different growth rates for boys and girls. The growth rate remains the same for individuals of same Sex,

$$y_{ij} = \beta_0 + \beta_{0M} \times \mathbb{1}_{\text{Sex}_i=\text{M}} + \beta_{1M} \times \text{age}_{ij} \times \mathbb{1}_{\text{Sex}_i=\text{M}} + \beta_{1F} \times \text{age}_{ij} \times \mathbb{1}_{\text{Sex}_i=\text{F}} + \eta_{i0} + e_{ij}$$

```
lmer(distance ~ Sex+age:Sex + (1|Subject), data = Orthodont)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ Sex + age:Sex + (1 | Subject)
##    Data: Orthodont
## REML criterion at convergence: 433.7572
## Random effects:
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 1.816
##  Residual             1.386
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
##   (Intercept)      SexFemale    SexMale:age  SexFemale:age
##       16.3406         1.0321         0.7844         0.4795
```

- We can instead assume the same birth distance for all the individuals, but different growth rates for individuals of same Sex,

$$y_{ij} = \beta_0 + \beta_{1M} \times \text{age}_{ij} \times \mathbb{1}_{\text{Sex}_i=\text{M}} + \beta_{1F} \times \text{age}_{ij} \times \mathbb{1}_{\text{Sex}_i=\text{F}} + \eta_{i1} \times \text{age}_{ij} + e_{ij}$$

```
lmer(distance ~ age:Sex  + (-1+age|Subject), data = Orthodont)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age:Sex + (-1 + age | Subject)
##    Data: Orthodont
## REML criterion at convergence: 439.7694
## Random effects:
##  Groups   Name Std.Dev.
##  Subject  age  0.1597
##  Residual      1.4126
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
##   (Intercept)    age:SexMale  age:SexFemale
##       16.7611         0.7477         0.5329
```

Remark: By default, the standard deviation of a random effect (η_{i0} or η_{i1}) is the same for all the individuals. If we put a random effect on the intercept, for instance, it is then possible to consider different variances for males and females:

$$y_{ij} = \beta_0 + \beta_1 \times \text{age}_{ij} + \eta_{i0}^{\text{F}} 1\!\text{I}_{\text{Sex}_i=\text{F}} + \eta_{i0}^{\text{M}} 1\!\text{I}_{\text{Sex}_i=\text{M}} + e_{ij}$$

where $\eta_{i0}^{\text{F}} \sim \text{N}(0, \omega_{0F}^2)$ and $\eta_{i0}^{\text{M}} \sim \text{N}(0, \omega_{0M}^2)$

```
lmer(distance ~ age + (-1 + Sex|Subject), data = Orthodont)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nea
rly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + (-1 + Sex | Subject)
##    Data: Orthodont
## REML criterion at convergence: 446.152
## Random effects:
##  Groups   Name      Std.Dev. Corr
##  Subject  SexMale   1.778
##           SexFemale 2.574    -0.39
##  Residual           1.432
## Number of obs: 108, groups:  Subject, 27
## Fixed Effects:
## (Intercept)          age
##     17.1000       0.6602
## convergence code 0; 1 optimizer warnings; 0 lme4 warnings
```

In this example, $\hat{\omega}_{0F} = 2.574$ and $\hat{\omega}_{0M} = 1.778$.

4.4 Comparing linear mixed effects models

If we want to compare all the possible linear mixed effect models, we need to fit all these models and use some information criteria in order to select the ``best one''.

In our model $y_{ij} = c_{i0} + c_{i1} \times \text{age}_{ij} + e_{ij}$, each of the two individual coefficients c_{i0} and c_{i1}

- may depend on the explanatory variable Sex or not,
- may include a random component or not

Furthermore,

- when the model includes two random effects (one for the intercept and one for the slope), these two random effects may be either correlated or independent,
- the variance of a random effect may depend on the variable Sex or not.

At the end, there would be a very large number of models to fit and compare...

Let us restrict ourselves to models with correlated random effects, with the same variance for males and females. We therefore have $2 \times 2 \times 2 \times 2 = 16$ models to fit and compare if we want to perform an exhaustive comparison.

For a sake of simplicity in the notations, let us define the 2 numerical explanatory variables $s_i = \mathbb{1}_{\text{Sex}_i=M}$ and $a_i = \text{age}_i$.

$$
\begin{array}{ll}
\text{M1} & y_{ij} = \beta_0 + \beta_1 a_{ij} + e_{ij} \\
\text{M2} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_1 a_{ij} + e_{ij} \\
\text{M3} & y_{ij} = \beta_0 + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + e_{ij} \\
\text{M4} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + e_{ij} \\
\text{M5} & y_{ij} = \beta_0 + \beta_1 a_{ij} + \eta_{i0} + e_{ij} \\
\text{M6} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_1 a_{ij} + \eta_{i0} + e_{ij} \\
\text{M7} & y_{ij} = \beta_0 + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i0} + e_{ij} \\
\text{M8} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i0} + e_{ij} \\
\text{M9} & y_{ij} = \beta_0 + \beta_1 a_{ij} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M10} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_1 a_{ij} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M11} & y_{ij} = \beta_0 + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M12} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M13} & y_{ij} = \beta_0 + \beta_1 a_{ij} + \eta_{i0} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M14} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_1 a_{ij} + \eta_{i0} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M15} & y_{ij} = \beta_0 + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i0} + \eta_{i1} a_{ij} + e_{ij} \\
\text{M16} & y_{ij} = \beta_0 + \beta_{0\text{M}} s_i + \beta_{1\text{M}} s_i a_{ij} + \beta_{1\text{F}} (1 - s_i) a_{ij} + \eta_{i0} + \eta_{i1} a_{ij} + e_{ij}
\end{array}
$$

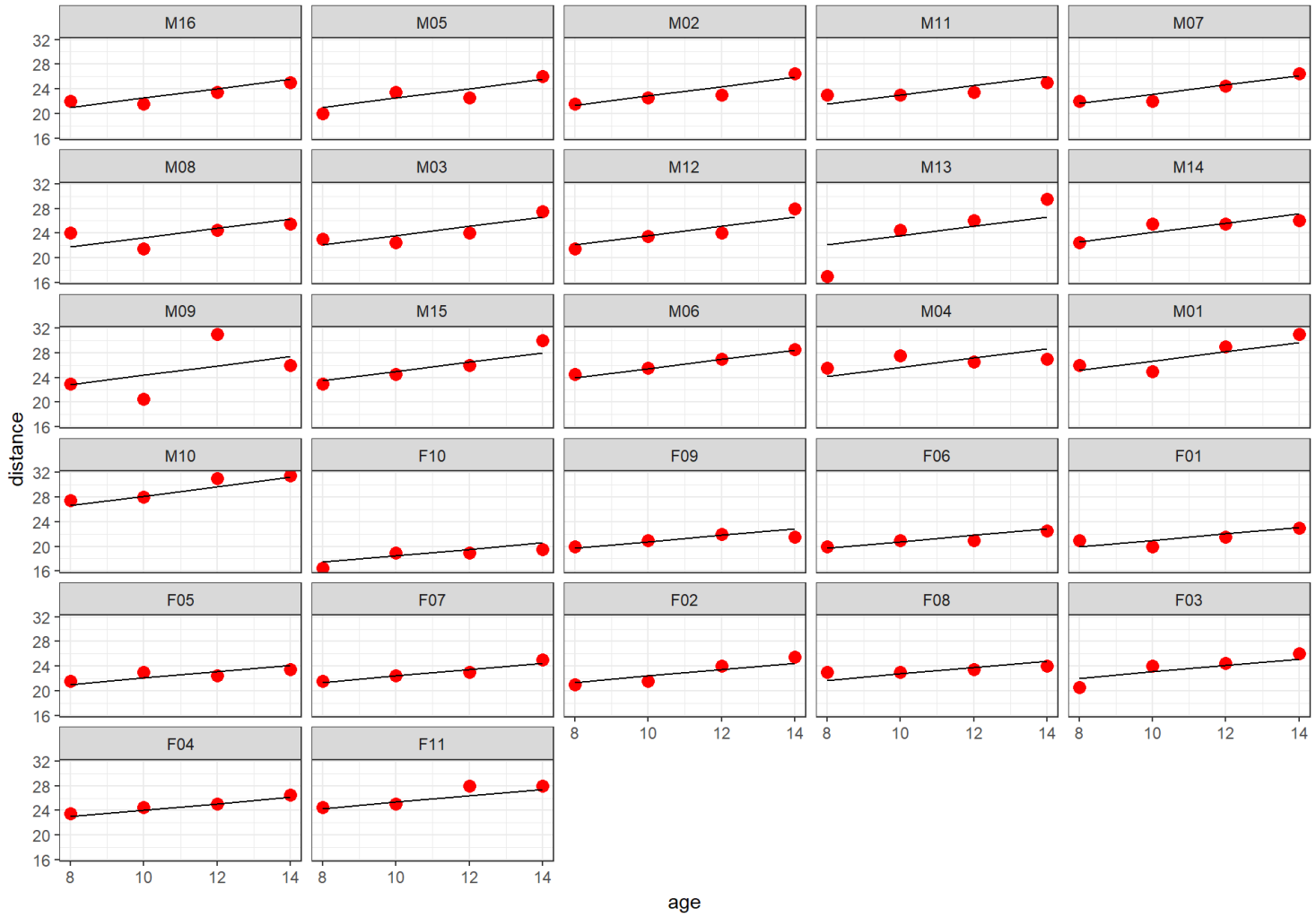
```
m1  <- lm(distance ~ age , data=Orthodont)
m2  <- lm(distance ~ Sex + age , data=Orthodont)
m3  <- lm(distance ~ 1 + age:Sex , data=Orthodont)
m4  <- lm(distance ~ Sex + age:Sex , data=Orthodont)
m5  <- lmer(distance ~ age + (1|Subject) , data=Orthodont)
m6  <- lmer(distance ~ Sex + age + (1|Subject) , data=Orthodont)
m7  <- lmer(distance ~ 1 + age:Sex + (1|Subject) , data=Orthodont)
m8  <- lmer(distance ~ Sex + age:Sex + (1|Subject) , data=Orthodont)
m9  <- lmer(distance ~ age + (-1+age|Subject) , data=Orthodont)
m10 <- lmer(distance ~ Sex + age + (-1+age|Subject) , data=Orthodont)
m11 <- lmer(distance ~ 1 + age:Sex + (-1+age|Subject) , data=Orthodont)
m12 <- lmer(distance ~ Sex + age:Sex + (-1+age|Subject) , data=Orthodont)
m13 <- lmer(distance ~ age + (age|Subject) , data=Orthodont)
m14 <- lmer(distance ~ Sex + age + (age|Subject) , data=Orthodont)
m15 <- lmer(distance ~ 1 + age:Sex + (age|Subject) , data=Orthodont)
m16 <- lmer(distance ~ Sex + age:Sex + (age|Subject) , data=Orthodont)
```

```
BIC(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11,m12,m13,m14,m15,m16)
```

```
##     df      BIC
## m1   3 519.6234
## m2   4 499.4121
## m3   4 497.1948
## m4   5 501.6524
## m5   4 465.7310
## m6   5 460.9232
## m7   5 460.3152
## m8   6 461.8500
## m9   4 463.8142
## m10  5 462.7684
## m11  5 463.1800
## m12  6 464.8139
## m13  6 470.7295
## m14  7 468.0088
## m15  7 468.5409
## m16  8 470.0387
```

The best model, according to BIC, is model M7 that assumes different fixed slopes for males and females and a random intercept.

```
Orthodont$pred.final <- fitted(m7)
ggplot(data=Orthodont) + geom_point(aes(x=age,y=distance), color="red", size=3) +
  geom_line(aes(x=age,y=pred.final)) + facet_wrap(~Subject, ncol=5)
```



We can compute 95% profile-based confidence intervals for the parameters of the model:

```
confint(pr)
```

```
##                   2.5 %     97.5 %
## .sig01         1.2911611  2.4196733
## .sigma         1.1850214  1.6142498
## (Intercept)   15.2918183 18.2304039
## age:SexMale    0.6288955  0.8810452
## age:SexFemale  0.3866637  0.6576257
```

Parametric bootstrap can also be used for computing confidence intervals:

```
confint(m7, method="boot")
```

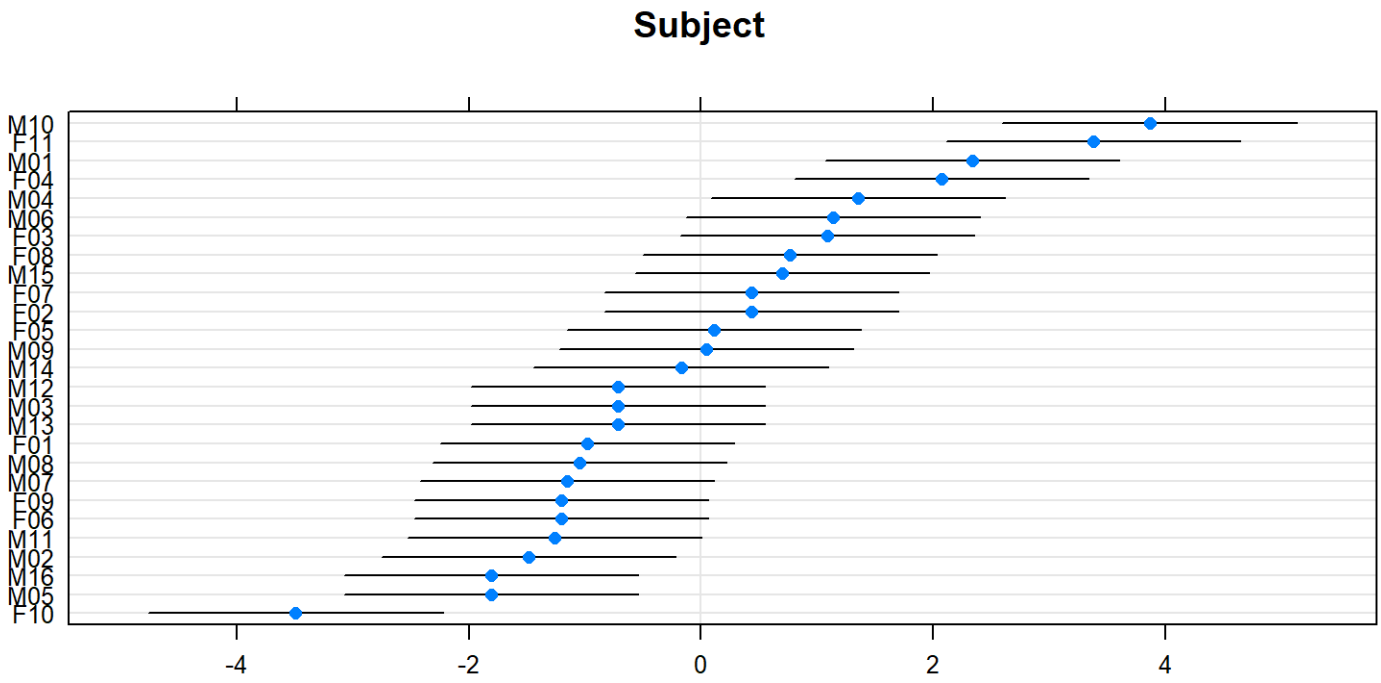
```
## Computing bootstrap confidence intervals ...
```

```
##                   2.5 %     97.5 %
## .sig01         1.2832787  2.3928604
## .sigma         1.1609884  1.5791560
## (Intercept)   15.2074278 18.2450180
```

```
## age:SexMale    0.6378229  0.8931138
## age:SexFemale  0.3736038  0.6650938
```

There is only one random effect in the final model. We can plot 95% prediction intervals on the random effects (η_i):

```
library(lattice)
d = dotplot(ranef(m7, condVar=TRUE), strip=FALSE)
print(d[[1]])
```



5 Some examples of models and designs

5.1 One factor (or one-way) classification

A “one-way classification” of data refers to data sets that are grouped according to one criterion. It can result from designed experiments, sample surveys, or observational studies.

5.1.1 Repeated measures

dataset: Rail (package: nlme)

Experiment: Six rails chosen at random, three measurements of travel time of a ultrasonic wave through each rail.

```
library(nlme)
data(Rail)
head(Rail)
```

```
## Grouped Data: travel ~ 1 | Rail
##   Rail travel
## 1    1     55
## 2    1     53
## 3    1     54
```

```
## 4    2     26
## 5    2     37
## 6    2     32
```

Linear model:

$$y_{ij} = \mu_1 + \beta_i + e_{ij} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3$$

where $\beta_1 = 0$

The `lm` function returns the estimated intercept $\hat{\mu}_1$ and the estimated effects $(\hat{\beta}_i)$ for $i = 2, 3, \ldots 6$:

```
#define Rail as factor using the original levels 1, 2, ... 6
Rail$Rail <- factor(Rail$Rail, levels=unique(Rail$Rail))
Rail$Rail <- factor(unclass(Rail$Rail))
(lm.rail <- lm(travel ~ Rail, data = Rail))
```

```
##
## Call:
## lm(formula = travel ~ Rail, data = Rail)
##
## Coefficients:
## (Intercept)        Rail2        Rail3        Rail4        Rail5
##       54.00       -22.33        30.67        42.00        -4.00
##       Rail6
##       28.67
```

The estimated intercepts $(\hat{\mu}_i = \hat{\mu}_1 + \hat{\beta}_i)$ for the 6 rails are therefore

```
cf <- coef(lm.rail)
c <- c(cf[1], cf[1]+cf[2:6])
c
```

```
## (Intercept)       Rail2       Rail3       Rail4       Rail5       Rail6
##    54.00000    31.66667    84.66667    96.00000    50.00000    82.66667
```

These intercepts are the 6 empirical means of the travel times for the 6 rails $(\bar{y}_i, 1 \leq i \leq 6)$:

```
aggregate(Rail$travel, list(Rail$Rail), mean)
```

```
##   Group.1        x
## 1       1 54.00000
## 2       2 31.66667
## 3       3 84.66667
## 4       4 96.00000
## 5       5 50.00000
```

```
## 6       6 82.66667
```

A linear mixed effects model considers that the 6 rails were randomly selected from a ``population'' of rails. The rail effect is therefore treated as a **random effect**:

$$y_{ij} = \mu + \eta_i + e_{ij} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3$$

where η_i is the deviation from the population intercept μ for the i-th rail: $\eta_i \sim_{\text{i.i.d.}} \mathrm{N}(0, \omega^2)$.

```
(lme.rail <- lmer(travel ~ 1 + (1|Rail), data = Rail))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: travel ~ 1 + (1 | Rail)
##    Data: Rail
## REML criterion at convergence: 122.177
## Random effects:
##  Groups   Name        Std.Dev.
##  Rail     (Intercept) 24.805
##  Residual              4.021
## Number of obs: 18, groups:  Rail, 6
## Fixed Effects:
## (Intercept)
##        66.5
```

The population intercept μ is estimated by the empirical mean of the 18 travel times

```
mean(Rail$travel)
```

```
## [1] 66.5
```

The estimated individual predicted travel times $(\hat{\mu}_i)$ are

```
coef(lme.rail)
```

```
## $Rail
##   (Intercept)
## 1    54.10852
## 2    31.96909
## 3    84.50894
## 4    95.74388
## 5    50.14325
## 6    82.52631
##
## attr(,"class")
## [1] "coef.mer"
```

These individual parameter estimates are not anymore the empirical means $(\bar{y}_i, 1 \leq i \leq 6)$. Indeed, the MAP estimate of μ_i combines the least square estimate $\bar{y}_i$ and the estimated population intercept $\hat{\mu}$:

$$\hat{\mu}_i = \frac{n_i \hat{\omega}^2}{n_i \hat{\omega}^2 + \hat{\sigma}^2} \bar{y}_i + \frac{\hat{\sigma}^2}{n_i \hat{\omega}^2 + \hat{\sigma}^2} \hat{\mu}$$

where n_i is the number of observations for rail i (here, $n_i = 3$).

```
ni <- 3
omega2.est <- VarCorr(lme.rail)$Rail
sigma2.est <- attr(VarCorr(lme.rail)[],"sc")^2
mu.est <- fixed.effects(lme.rail)
yi.est <- aggregate(Rail$travel, list(Rail$Rail), mean)
ni*omega2.est/(ni*omega2.est+sigma2.est)*yi.est + sigma2.est/(ni*omega2.est+sigma2.est)*mu.est
```

```
## Warning in Ops.factor(left, right): '*' not meaningful for factors
```

```
## Warning in FUN(left, right): Recycling array of length 1 in array-vector arithmetic is depr
ecated.
##   Use c() or as.vector() instead.
```

```
## Warning in FUN(left, right): Recycling array of length 1 in vector-array arithmetic is depr
ecated.
##   Use c() or as.vector() instead.

## Warning in FUN(left, right): Recycling array of length 1 in vector-array arithmetic is depr
ecated.
##   Use c() or as.vector() instead.
```

```
##   Group.1        x
## 1      NA 54.10852
## 2      NA 31.96909
## 3      NA 84.50894
## 4      NA 95.74388
## 5      NA 50.14325
## 6      NA 82.52631
```

We can also check that $\hat{\mu}_i = \hat{\mu} + \hat{\eta}_i$, where $(\hat{\eta}_i)$ are the estimated random effects

```
ranef(lme.rail)
```

```
## $Rail
##   (Intercept)
## 1   -12.39148
## 2   -34.53091
## 3    18.00894
```

```
## 4    29.24388
## 5   -16.35675
## 6    16.02631
```

5.2 Two factors block design

5.2.1 Design with no replications

dataset: ergoStool (package: nlme)

Experiment: Nine testers had to sit in four different ergonomic stools and their effort to raise was measured once.

```
data(ergoStool)
# define "Subject" as a factor with unorderedlevels
ergoStool$Subject <- factor(unclass(ergoStool$Subject))
head(ergoStool)
```

```
## Grouped Data: effort ~ Type | Subject
##   effort Type Subject
## 1     12   T1       8
## 2     15   T2       8
## 3     12   T3       8
## 4     10   T4       8
## 5     10   T1       9
## 6     14   T2       9
```

```
xtabs(~ Type + Subject, ergoStool)
```

```
##     Subject
## Type 1 2 3 4 5 6 7 8 9
##   T1 1 1 1 1 1 1 1 1 1
##   T2 1 1 1 1 1 1 1 1 1
##   T3 1 1 1 1 1 1 1 1 1
##   T4 1 1 1 1 1 1 1 1 1
```

5.2.1.1 Model with one fixed and one random factor

In this model, the stool type is considered as a fixed effect (β_j) while the testing subject is treated as a random effect (η_i)

$$y_{ij} = \mu + \beta_j + \eta_i + e_{ij} \quad , \quad i = 1, \ldots, 9 \, , \quad j = 1, \ldots, 4$$

where $\beta_1 = 0$.

```
(lme.ergo1 <- lmer(effort ~ Type + (1|Subject), data = ergoStool))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: effort ~ Type + (1 | Subject)
##    Data: ergoStool
## REML criterion at convergence: 121.1308
## Random effects:
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 1.332
##  Residual             1.100
## Number of obs: 36, groups:  Subject, 9
## Fixed Effects:
## (Intercept)       TypeT2       TypeT3       TypeT4
##      8.5556       3.8889       2.2222       0.6667
```

Even if it is of very little interest, we could instead consider the stool type as a random effect (η_j) and the testing subject as a fixed effect (β_i)

$$y_{ij} = \mu + \beta_i + \eta_j + e_{ij} \quad , \quad i = 1, \ldots, 9 \, , \quad j = 1, \ldots, 4$$

where $\beta_1 = 0$.

```
(lme.ergo2 <- lmer(effort ~ Subject + (1|Type) , data = ergoStool))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: effort ~ Subject + (1 | Type)
##    Data: ergoStool
## REML criterion at convergence: 103.5818
## Random effects:
##  Groups   Name        Std.Dev.
##  Type     (Intercept) 1.695
##  Residual             1.100
## Number of obs: 36, groups:  Type, 4
## Fixed Effects:
## (Intercept)     Subject2     Subject3     Subject4     Subject5
##        8.25         0.25         1.00         1.75         2.00
##    Subject6     Subject7     Subject8     Subject9
##        2.50         2.50         4.00         4.00
```

5.2.1.2 Model with two random factors

Both effects (stool type and testing subject) can be treated as random effects:

$$y_{ij} = \mu + \eta_i + \eta_j + e_{ij} \quad , \quad i = 1, \ldots, 9 \, , \quad j = 1, \ldots, 4$$

```
(lme.ergo3 <- lmer(effort ~ (1|Subject) + (1|Type) , data = ergoStool))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: effort ~ (1 | Subject) + (1 | Type)
##    Data: ergoStool
## REML criterion at convergence: 134.3337
```

```
## Random effects:
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 1.332
##  Type     (Intercept) 1.695
##  Residual             1.100
## Number of obs: 36, groups:  Subject, 9; Type, 4
## Fixed Effects:
## (Intercept)
##       10.25
```

5.2.1.3 Comparison between these models

We can compare these 3 models with a linear model assuming only one fixed factor

$$y_{ij} = \mu + \beta_j + e_{ij} \quad , \quad i = 1, \dots, 9 \, , \quad j = 1, \dots, 4$$

where $\beta_1 = 0$.

```
(lm.ergo <- lm(effort ~ Type , data = ergoStool))
##
## Call:
## lm(formula = effort ~ Type, data = ergoStool)
##
## Coefficients:
## (Intercept)       TypeT2       TypeT3       TypeT4
##      8.5556       3.8889       2.2222       0.6667
cat("Residual standard error: ",summary(lm.ergo)$sigma)
## Residual standard error:  1.728037
```

```
BIC(lm.ergo, lme.ergo1, lme.ergo2, lme.ergo3)
```

```
##           df      BIC
## lm.ergo    5 155.2240
## lme.ergo1  6 142.6319
## lme.ergo2 11 143.0005
## lme.ergo3  4 148.6678
```

Remark: The interaction between the testing subject and the stool type cannot be taken into account with this design as there is no replication.

5.2.2 Design with replications

dataset: Machines (package: nlme)

Experiment: Six workers were chosen randomly among the employees of a factory to operate each machine three times. The response is an overall productivity score taking into account the number and quality of components produced.

```
data(Machines)
```

```
Machines$Worker <- factor(Machines$Worker, levels=unique(Machines$Worker))

head(Machines)
```

```
## Grouped Data: score ~ Machine | Worker
##   Worker Machine score
## 1      1       A  52.0
## 2      1       A  52.8
## 3      1       A  53.1
## 4      2       A  51.8
## 5      2       A  52.8
## 6      2       A  53.1
```

```
xtabs(~ Machine + Worker, Machines)
```

```
##        Worker
## Machine 1 2 3 4 5 6
##       A 3 3 3 3 3 3
##       B 3 3 3 3 3 3
##       C 3 3 3 3 3 3
```

5.2.2.1 Model with one fixed and one random factor, without interaction

Although the operators represent a sample from the population of potential operators, the three machines are the specific machines of interest. That is, we regard the levels of Machine as fixed levels and the levels of Worker as a random sample from a population.

A first model considers therefore the machine as a fixed effect (β_j) and the subject (or worker in this example) as a random effect (η_i). We don't assume any interaction between the worker and the machine in this first model.

$$y_{ijk} = \mu + \beta_j + \eta_i + e_{ijk} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3 \, , \quad k = 1, 2, 3 \text{ replications}$$

where $\beta_1 = 0$.

```
(lme.machine1 <- lmer(score ~ Machine + (1|Worker), data = Machines))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: score ~ Machine + (1 | Worker)
##    Data: Machines
## REML criterion at convergence: 286.8782
## Random effects:
##  Groups   Name        Std.Dev.
##  Worker   (Intercept) 5.147
##  Residual             3.162
## Number of obs: 54, groups:  Worker, 6
## Fixed Effects:
## (Intercept)     MachineB     MachineC
##      52.356        7.967       13.917
```

5.2.2.2 Model with one fixed and one random factor, with interaction

We can furthermore assume that there exists an interaction between the worker and the machine. This interaction is treated as a random effect (η_{ij}):

$$y_{ijk} = \mu + \beta_j + \eta_i + \eta_{ij} + e_{ijk} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3 \, , \quad k = 1, 2, 3 \text{ replications}$$

```
(lme.machine2 <- lmer(score ~ Machine +  (1|Worker) + (1|Worker:Machine), data = Machines))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: score ~ Machine + (1 | Worker) + (1 | Worker:Machine)
##    Data: Machines
## REML criterion at convergence: 215.6876
## Random effects:
##  Groups         Name        Std.Dev.
##  Worker:Machine (Intercept) 3.7295
##  Worker         (Intercept) 4.7811
##  Residual                   0.9616
## Number of obs: 54, groups:  Worker:Machine, 18; Worker, 6
## Fixed Effects:
## (Intercept)     MachineB     MachineC
##      52.356        7.967       13.917
```

5.2.2.3 Model with two random factors without interaction

The effect of the machine could be considered as a random effect, instead of a fixed one:

$$y_{ijk} = \mu + \eta_i + \eta_j + e_{ijk} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3 \, , \quad k = 1, 2, 3 \text{ replications}$$

```
(lme.machine3 <- lmer(score ~ (1|Machine) + (1|Worker), data = Machines))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: score ~ (1 | Machine) + (1 | Worker)
##    Data: Machines
## REML criterion at convergence: 301.4263
## Random effects:
##  Groups   Name        Std.Dev.
##  Worker   (Intercept) 5.147
##  Machine  (Intercept) 6.943
##  Residual             3.162
## Number of obs: 54, groups:  Worker, 6; Machine, 3
## Fixed Effects:
## (Intercept)
##       59.65
```

5.2.2.4 Model with two random factors with interaction

$$y_{ijk} = \mu + \eta_i + \eta_j + \eta_{ij} + e_{ijk} \quad , \quad i = 1, \ldots, 6 \, , \quad j = 1, 2, 3 \, , \quad k = 1, 2, 3 \text{ replications}$$

```
(lme.machine4 <- lmer(score ~ (1|Machine) + (1|Worker) + (1|Worker:Machine), data = Machines))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: score ~ (1 | Machine) + (1 | Worker) + (1 | Worker:Machine)
##    Data: Machines
## REML criterion at convergence: 230.2356
## Random effects:
##  Groups         Name        Std.Dev.
##  Worker:Machine (Intercept) 3.7295
##  Worker         (Intercept) 4.7811
##  Machine        (Intercept) 6.8109
##  Residual                   0.9616
## Number of obs: 54, groups:  Worker:Machine, 18; Worker, 6; Machine, 3
## Fixed Effects:
## (Intercept)
##       59.65
```

Model comparison:

```
BIC(lme.machine1, lme.machine2, lme.machine3, lme.machine4)
```

```
##              df      BIC
## lme.machine1  5 306.8231
## lme.machine2  6 239.6215
## lme.machine3  4 317.3822
## lme.machine4  5 250.1806
```

Statistical tests: single comparison

Marc Lavielle
January 20th, 2018

1 Introduction

A feeding study served to authorize the MON863 maize, a genetically modified organism (GMO) developed by the Monsanto company, by the European and American authorities. It included male and female rats. For each sex, one group was fed with GMOs in the equilibrated diet, and one with the closest control regimen without GMOs.

We are interested in the weight of the rats after a period of 14 weeks.

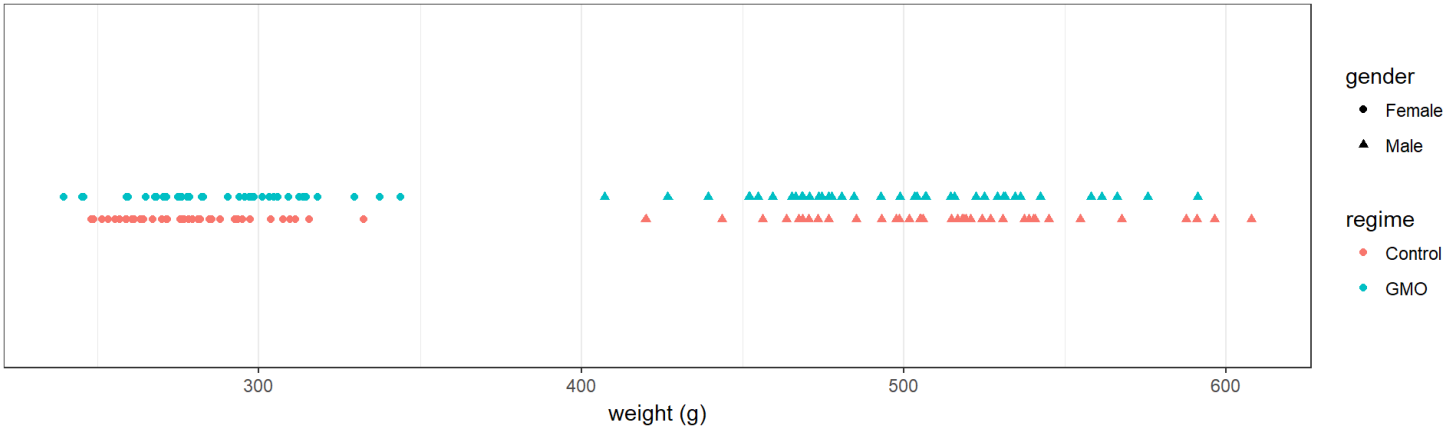
```
ratWeight <- read.csv("ratWeight.csv")
data <- subset(ratWeight, week==14)
head(data)
```

```
##        id week weight  regime gender dosage
## 14 B38602   14  514.9 Control   Male    11%
## 28 B38603   14  505.0 Control   Male    11%
```

```
## 42 B38604   14  545.1 Control   Male    11%
## 56 B38605   14  596.6 Control   Male    11%
## 70 B38606   14  516.8 Control   Male    11%
## 84 B38607   14  518.1 Control   Male    11%
```

The data per gender and regime is displayed below

```
library(ggplot2)
theme_set(theme_bw())
ggplot(data=data) + geom_point(aes(x=weight,y=as.numeric(regime),colour=regime, shape=gender)) +
  ylab(NULL) + scale_y_continuous(breaks=NULL, limits=c(-5,10)) + xlab("weight (g)")
```



The following table provides the mean weight in each group

```
aggregate(weight ~ regime+gender, data=data, FUN= "mean" )
```

```
##    regime gender   weight
## 1 Control Female 278.2825
## 2     GMO Female 287.3225
## 3 Control   Male 513.7077
## 4     GMO   Male 498.7359
```

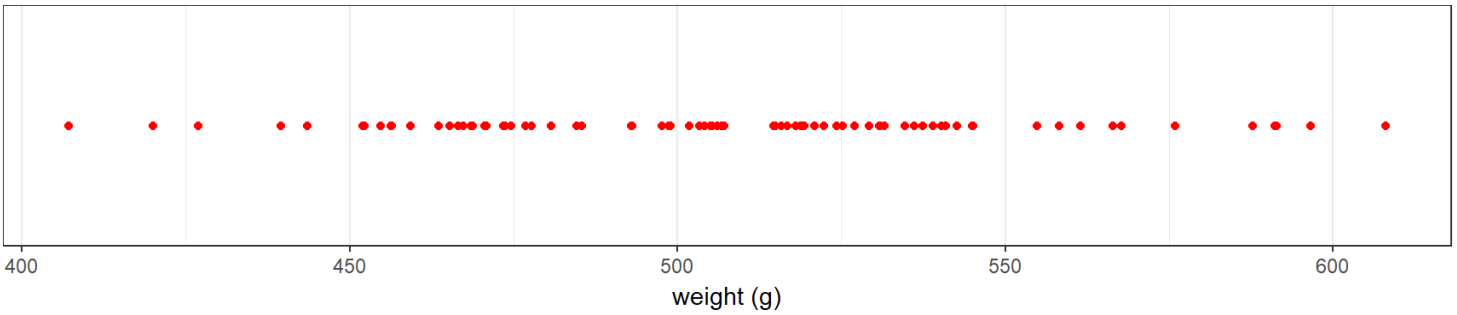
Our main objective is to detect some possible effect of the diet on the weight. More precisly, we would like to know if the differences observed in the data are due to random fluctuations in sampling or to differences in diet.

2 Student's t-test

2.1 One sample t-test

Before considering the problem of comparing two groups, let us start looking at the weight of the male rats only:

```
ggplot(data=subset(data,gender=="Male")) + geom_point(aes(x=weight,y=0), colour="red") +
  ylab(NULL) + scale_y_continuous(breaks=NULL) + xlab("weight (g)")
```



Let x_1, x_2, x_n the weights of the n male rats. We will assume that the x_i's are independent and normally distributed with mean μ and variance σ^2:

$$x_i \underset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu\,,\,\sigma^2)$$

2.1.1 One sided test

We want to test

$$H_0 : \text{``}\mu \leq \mu_0\text{''} \quad \text{versus} \quad H_1 : \text{``}\mu > \mu_0\text{''}$$

Function `t.test` can be used for performing this test:

```
x <- data[data$gender=="Male","weight"]
mu0 <- 500
t.test(x, alternative="greater", mu=mu0)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.2708, df = 77, p-value = 0.1038
## alternative hypothesis: true mean is greater than 500
## 95 percent confidence interval:
##  498.0706      Inf
## sample estimates:
## mean of x
##  506.2218
```

Let us see what these outputs are and how they are computed.

Let $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ be the empirical mean of the data.

$$\overline{x} \sim \mathrm{N}(\mu\,,\,\frac{\sigma^2}{n})$$

Then,

$$\frac{\sqrt{n}\,(\overline{x}-\mu)}{\frac{\sigma}{\sqrt{n}\,(\overline{x}-\mu)}} \sim \mathrm{N}(0\,,\,1)$$

$$\frac{}{s} \sim t_{n-1}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is the empirical variance of (x_i).

The statistic used for the test should be a function of the data whose distribution under H_0 is known, and whose expected behavior under H_1 allows one to define a *rejection region* (or *critical region*) for the null hypothesis.

Here, the test statistic is

$$T_{\text{stat}} = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$$

which follows a t-distribution with $n-1$ degrees of freedom when $\mu = \mu_0$.

$\bar{x}$ is expected to be less than or equal to μ_0 under the null hypothesis, and greater than μ_0 under the alternative hypothesis, Hence, T_{stat} is expected to be less than or equal to 0 under H_0 and greater than 0 under H_1. We then reject the null hypothesis H_0 if T_{stat} is greater than some threshold q.

Such decision rule may lead to two kinds of error:

- The **type I error** is the incorrect rejection of null hypothesis when it is true,
- The **type II error** is the failure to reject the null hypothesis when it is false.

The type I error rate or *significance level* is therefore the probability of rejecting the null hypothesis given that it is true.

In our case, for a given significance level α, we will reject H_0 if $T_{\text{stat}} > qt_{1-\alpha,n-1}$, where $qt_{1-\alpha,n-1}$ is the quantile of order $1-\alpha$ for a t-distribution with $n-1$ degrees of freedom.

Indeed, by definition,

$$\begin{aligned}
\text{P}(\text{reject } H_0 \mid H_0 \text{ true}) &= \text{P}(T_{\text{stat}} > qt_{1-\alpha,n-1} \mid \mu \leq \mu_0) \\
&\leq \text{P}(T_{\text{stat}} > qt_{1-\alpha,n-1} \mid \mu = \mu_0) \\
&\leq \text{P}(t_{n-1} > qt_{1-\alpha,n-1}) \\
&\leq \alpha
\end{aligned}$$

```
alpha <- 0.05
x.mean <- mean(x)
x.sd <- sd(x)
n <- length(x)
df <- n-1
t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
c(t.stat,qt(1-alpha, df))
```

```
## [1] 1.270806 1.664885
```

We therefore don't reject H_0 in our example since $T_{\text{stat}} < qt_{1-\alpha,n-1}$.

We can equivalently compute the significance level for which the test becomes significant. This value is called the *p-value*:

$$\begin{aligned} p_{\text{value}} &= \max \mathrm{P}_{H_0}(T_{\text{stat}} > T_{\text{stat}}^{\text{obs}}) \\ &= \mathrm{P}(T_{\text{stat}} > T_{\text{stat}}^{\text{obs}} \mid \mu = \mu_0) \\ &= 1 - \mathrm{P}(t_{n-1} \leq T_{\text{stat}}^{\text{obs}}) \end{aligned}$$

Now, $T_{\text{stat}} > qt_{1-\alpha,n-1}$ under H_0 if and only if $\mathrm{P}(t_{n-1} \leq T_{\text{stat}}^{\text{obs}}) \geq 1-\alpha$.$ Then, the test is significant at the level α if and only if $p_{\text{value}} \leq \alpha$.

```
( p.value <- 1 - pt(t.stat,df) )
```

```
## [1] 0.1038119
```

Here, we would reject H_0 for any significance level $\alpha \geq$ 0.104.

Important: The fact that the test is not significant at the level α does not allow us to conclude that H_0 is true, i.e. that μ is less than or equal to 500. We can only say that the data does not allow us to conclude that $\mu > 500$.

Imagine now that we want to test if $\mu \geq 515$ for instance. The alternative here is H_1 : ‘‘$\mu < 515$.

```
mu0 <- 515
t.test(x, alternative="less", mu=mu0)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.793, df = 77, p-value = 0.03845
## alternative hypothesis: true mean is less than 515
## 95 percent confidence interval:
##     -Inf 514.373
## sample estimates:
## mean of x
##  506.2218
```

More generally, we may want to test

$$H_0 : \text{‘‘}\mu \geq \mu_0 \text{''} \quad \text{versus} \quad H_1 : \text{‘‘}\mu < \mu_0 \text{''}$$

We still use the statistic $T_{\text{stat}} = \sqrt{n}(\bar{x} - \mu_0)/s$ for this test, but the rejection region is now the area that lies to the left of the critical value $qt_{\alpha,n-1}$ since

$$\begin{aligned} \mathrm{P}(\text{reject } H_0 \mid H_0 \text{ true}) &= \mathrm{P}(T_{\text{stat}} < qt_{\alpha,n-1} \mid \mu \geq \mu_0) \\ &\leq \mathrm{P}(T_{\text{stat}} < qt_{\alpha,n-1} \mid \mu = \mu_0) \\ &\leq \alpha \end{aligned}$$

```
t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
p.value <- pt(t.stat,df)
c(t.stat, df, p.value)
```

```
## [1] -1.79295428 77.00000000  0.03845364
```

Here, the p-value is less than $\alpha = 0.05$: we then reject the null hypothesis at the 5% level and conclude that $\mu < 515$.

2.1.2 Two sided test

A two sided test (or two tailed test) can be used to test if $\mu = 500$ for instance

```
mu0 = 500
t.test(x, alternative="two.sided", mu=mu0)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.2708, df = 77, p-value = 0.2076
## alternative hypothesis: true mean is not equal to 500
## 95 percent confidence interval:
##  496.4727 515.9709
## sample estimates:
## mean of x
##  506.2218
```

More generally, we can test

$$H_0 : \text{``}\mu = \mu_0\text{''} \quad \text{versus} \quad H_1 : \text{``}\mu \neq \mu_0\text{''}$$

The test also uses the statistic $T_{\text{stat}} = \sqrt{n}(\overline{x} - \mu_0)/s$, but the rejection region has now two parts: we reject H_0 if $|T_{\text{stat}}| > qt_{1-\alpha/2}$. Indeed,

$$\begin{aligned}
\text{P}(\text{reject } H_0 \mid H_0 \text{ true}) &= \text{P}(|T_{\text{stat}}| > qt_{1-\frac{\alpha}{2},n-1} \mid \mu = \mu_0) \\
&= \text{P}(T_{\text{stat}} < qt_{\frac{\alpha}{2},n-1} \mid \mu = \mu_0) + \text{P}(T_{\text{stat}} > qt_{1-\frac{\alpha}{2},n-1} \mid \mu = \mu_0) \\
&= \text{P}(t_{n-1} \leq qt_{\frac{\alpha}{2},n-1}) + \text{P}(t_{n-1} \geq qt_{1-\frac{\alpha}{2},n-1}) \\
&= \frac{\alpha}{2} + \frac{\alpha}{2} \\
&= \alpha
\end{aligned}$$

The p-value of the test is now

$$\begin{aligned}
p_{\text{value}} &= \text{P}_{H_0}(|T_{\text{stat}}| > |T_{\text{stat}}^{\text{obs}}|) \\
&= \text{P}_{H_0}(T_{\text{stat}} < -|T_{\text{stat}}^{\text{obs}}|) + \text{P}_{H_0}(T_{\text{stat}} > |T_{\text{stat}}^{\text{obs}}|)
\end{aligned}$$

$$= \mathrm{P}(t_{n-1} \leq -|T_{\text{stat}}^{\text{obs}}|) + \mathrm{P}(t_{n-1} \geq |T_{\text{stat}}^{\text{obs}}|)$$
$$= 2\,\mathrm{P}(t_{n-1} \leq -|T_{\text{stat}}^{\text{obs}}|)$$

```
t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
p.value <- 2*pt(-abs(t.stat),df)
c(t.stat, df, p.value)
```

```
## [1]  1.2708058 77.0000000  0.2076238
```

Here, $p_{\text{value}} = 0.208$. Then, for any significance level less than 0.208, we cannot reject the hypothesis that $\mu = 500$.

2.1.3 Confidence interval for the mean

We have just seen that the data doesn't allow us to reject the hypothesis that $\mu = 500$. But we would come to the same conclusion with other values of μ_0. In particular, we will never reject the hypothesis that $\mu = \bar{x}$:

```
t.test(x, mu=x.mean, conf.level=1-alpha)$p.value
```

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

For a given significance level ($\alpha = 0.05$ for instance), we will not reject the null hypothesis for values of μ_0 close enough to $\bar{x}$.

```
pv.510 <- t.test(x, mu=510, conf.level=1-alpha)$p.value
pv.497 <- t.test(x, mu=497, conf.level=1-alpha)$p.value
c(pv.510, pv.497)
```

```
## [1] 0.44265350 0.06340045
```

On the other hand, we will reject H_0 for values of μ_0 far enough from $\bar{x}$:

```
pv.520 <- t.test(x, mu=520, conf.level=1-alpha)$p.value
pv.490 <- t.test(x, mu=490, conf.level=1-alpha)$p.value
c(pv.520, pv.490)
```

```
## [1] 0.006204188 0.001406681
```

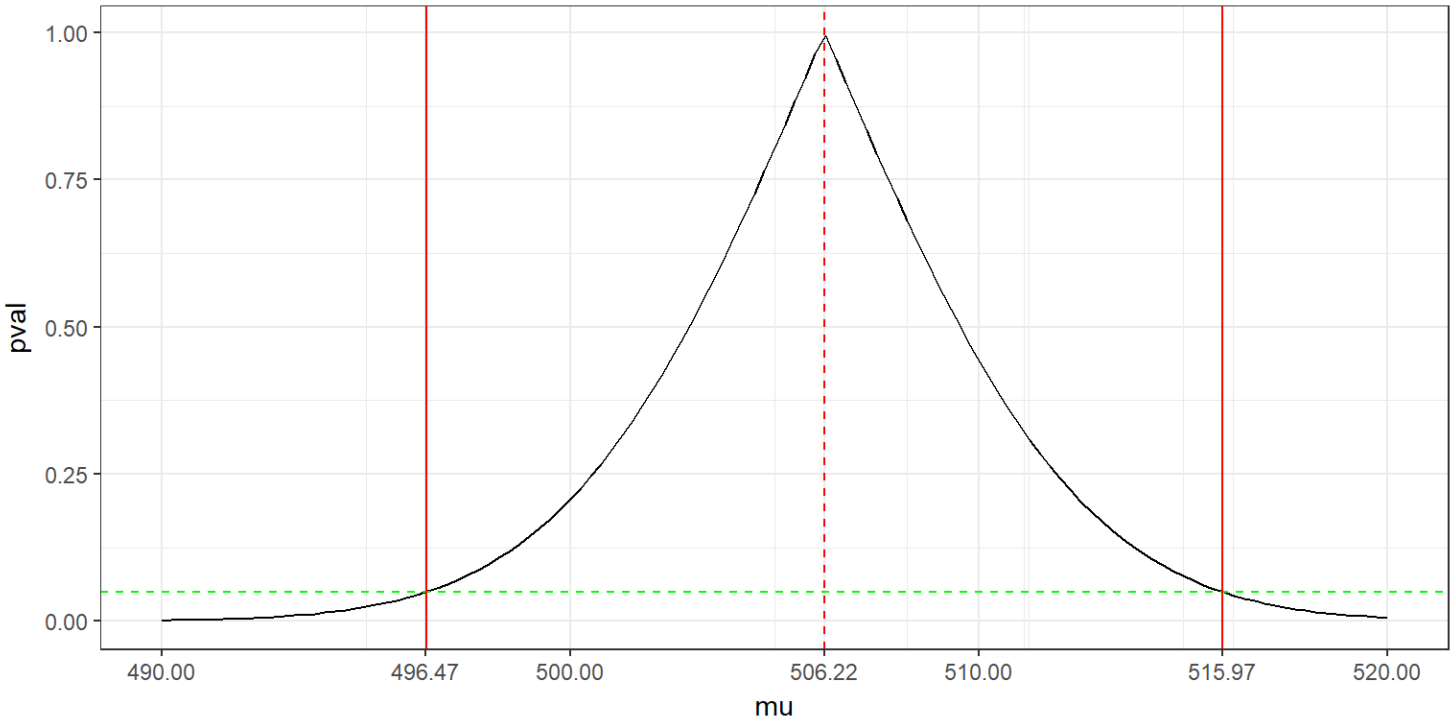
There exist two values of μ_0 for which the decision is borderline

```
pv1 <- t.test(x, mu=495.892, conf.level=1-alpha)$p.value
pv2 <- t.test(x, mu=515.5443, conf.level=1-alpha)$p.value
c(pv1,pv2)
```

```
## [1] 0.03811820 0.06062926
```

In fact, for a given α, these two values $\mu_{\alpha,\text{lower}}$ and $\mu_{\alpha,\text{upper}}$ define a *confidence interval* for μ: We are ``confident'' at the level $1-\alpha$ that any value between $\mu_{\alpha,\text{lower}}$ and $\mu_{\alpha,\text{upper}}$ is a possible value for μ.

```
mu <- seq(490,520,by=0.25)
t.stat <- (x.mean-mu)/x.sd*sqrt(n)
pval <- pmin(pt(-t.stat,df) + (1- pt(t.stat,df)),pt(t.stat,df) + (1- pt(-t.stat,df)))
dd <- data.frame(mu=mu, v.pval=pval)
CI <- x.mean + x.sd/sqrt(n)*qt(c(alpha/2,1-alpha/2), df)
ggplot(data=dd) + geom_line(aes(x=mu,y=pval)) +
  geom_vline(xintercept=x.mean,colour="red", linetype=2)+
  geom_hline(yintercept=alpha,colour="green", linetype=2)+
  geom_vline(xintercept=CI,colour="red") +
  scale_x_continuous(breaks=round(c(490,500,510,520,CI,x.mean),2))
```



By construction,

$$\begin{aligned} 1-\alpha &= \mathrm{P}(qt_{\frac{\alpha}{2},n-1} < T_{\text{stat}} < qt_{1-\frac{\alpha}{2},n-1}) \\ &= \mathrm{P}(qt_{\frac{\alpha}{2},n-1} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < qt_{1-\frac{\alpha}{2},n-1}) \\ &= \mathrm{P}(\bar{x} + \frac{s}{\sqrt{n}} qt_{\frac{\alpha}{2},n-1} < \mu < \bar{x} + \frac{s}{\sqrt{n}} qt_{1-\frac{\alpha}{2},n-1}) \end{aligned}$$

The confidence interval of level $1-\alpha$ for μ is therefore the interval

$$\mathrm{CI}_{1-\alpha} = [\bar{x} + \frac{s}{\sqrt{n}} qt_{\frac{\alpha}{2},n-1} \ , \ \bar{x} + \frac{s}{\sqrt{n}} qt_{1-\frac{\alpha}{2},n-1} \ ]$$

```
(CI <- x.mean + x.sd/sqrt(n)*qt(c(alpha/2,1-alpha/2), df))
```

```
## [1] 496.4727 515.9709
```

Remark 1: The fact that $\mathrm{P}(\mu \in \mathrm{CI}_{1-\alpha}) = 1-\alpha$ does not mean that μ is a random variable! It is the bounds of the confidence interval that are random because they are function of the data.

A confidence interval of level $1-\alpha$ should be interpreted like this: imagine that we repeat the same experiment many times, with the same experimental conditions, and that we build a confidence interval for μ for each of these replicate. Then, the true mean μ will lie in the confidence interval $(1-\alpha)100\%$ of the times.

Let us check this property with a Monte Carlo simulation.

```
L <- 100000
n <- 100
mu <- 500
sd <- 40
R <- vector(length=L)
for (l in (1:L)) {
  x <- rnorm(n,mu,sd)
  ci.l <- mean(x) + sd(x)/sqrt(n)*qt(c(alpha/2, 1-alpha/2),n-1)
  R[l] <- (mu > ci.l[1]  && mu < ci.l[2])
}
mean(R)
```

```
## [1] 0.94804
```

Remark 2:

The decision rule to reject or not the null hypothesis can be derived from the confidence interval. Indeed, the confidence interval plays the role of an acceptance region: we reject H_0 if μ_0 does not belong to $\mathrm{CI}_{1-\alpha}$.

In the case of a one sided test, the output of `t.test` called confidence interval is indeed an acceptance region for μ, but not a ``confidence interval'' (we cannot seriouly consider that μ can take any value above 500 for instance)

```
rbind(
c(x.mean + x.sd/sqrt(n)*qt(alpha,df) , Inf),
c(-Inf, x.mean + x.sd/sqrt(n)*qt(1-alpha,df)))
```

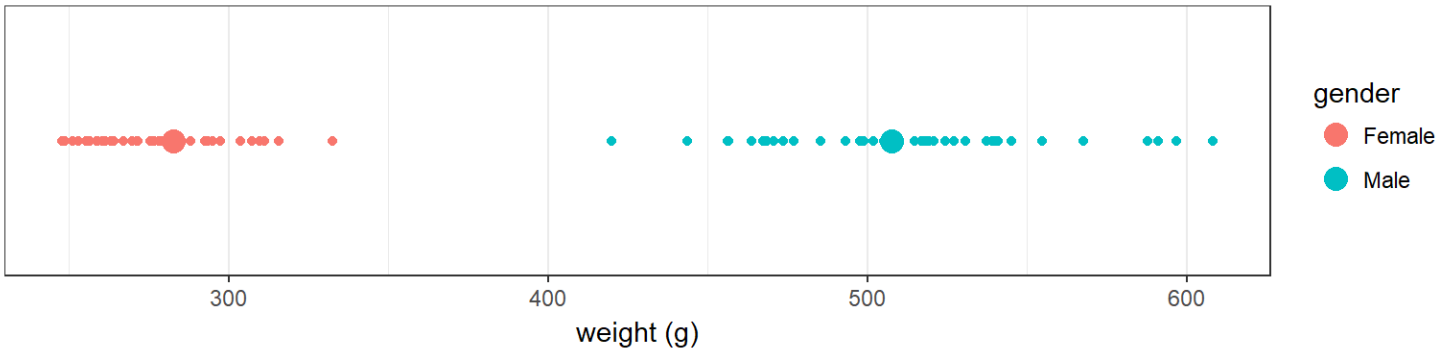
```
##          [,1]     [,2]
## [1,] 499.0229      Inf
## [2,]     -Inf 513.4207
```

2.2 Two samples t-test

2.2.1 What should we test?

Let us now compare the weights of the male and female rats.

```
y <- data[data$gender=="Female" ,"weight"]

dmean <- data.frame(x=c(mean(x),mean(y)),gender=c("Male","Female"))

ggplot(data=subset(data,regime=="Control")) + geom_point(aes(x=weight,y=0,colour=gender)) +

  geom_point(data=dmean, aes(x,y=0,colour=gender), size=4) +
  ylab(NULL) + scale_y_continuous(breaks=NULL) + xlab("weight (g)")
```



Looking at the data is more than enough for concluding that the mean weight of the males is (much) larger than the mean weight of the females Computing a p-value here is of little interest

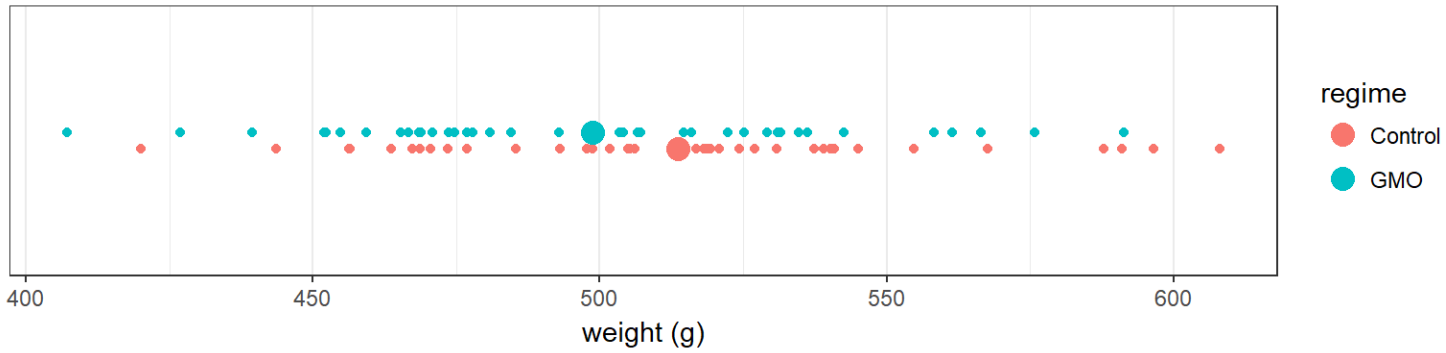
```
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 46.912, df = 166.81, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  207.0330 225.2244
## sample estimates:
## mean of x mean of y
##  498.9312  282.8025
```

Let us see now what happens if we compare the control and GMO groups for the male rats.

```
x <- data[data$gender=="Male" & data$regime=="Control","weight"]
y <- data[data$gender=="Male" & data$regime=="GMO","weight"]
```

```
dmean <- data.frame(x=c(mean(x),mean(y)),regime=c("Control","GMO"))
ggplot(data=data[data$gender=="Male",]) +  geom_point(aes(x=weight,y=as.numeric(regime),col
our=regime)) +
  geom_point(data=dmean, aes(x,y=as.numeric(regime),colour=regime), size=4) +
  ylab(NULL) + scale_y_continuous(breaks=NULL, limits=c(-6,9)) + xlab("weight (g)")
```



We observe a difference between the two empirical means (the mean weight after 14 weeks is greater in the control group), but we cannot say how significant this difference is by simply looking at the data. Performing a statistical test is now necessary.

Let $x_1, x_2, \ldots, x_{n_x}$ be the weights of the n_x male rats of the control group and $y_1, y_2, \ldots, y_{n_x}$ the weights of the n_y male rats of the GMO group. We will assume normal distributions for both (x_i) and (y_i):

$$x_i \underset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_x\ ,\ \sigma_x^2) \quad ; \quad y_i \underset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_y\ ,\ \sigma_y^2)$$

We want to test

$$H_0 : \text{``}\mu_x = \mu_y\text{''} \quad \text{versus} \quad H_1 : \text{``}\mu_x \neq \mu_y\text{''}$$

2.2.2 Assuming equal variances

We can use the function `t.test` assuming first equal variances ($\sigma_x^2 = \sigma_y^2$)

```
alpha <- 0.05
t.test(x, y, conf.level=1-alpha, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = 1.5426, df = 76, p-value = 0.1271
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.358031 34.301621
## sample estimates:
## mean of x mean of y
##  513.7077  498.7359
```

The test statistic is

$$T_{\text{stat}} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

where s_p^2 is the *pooled variance*:

$$s_p^2 = \frac{1}{n_x + n_y - 2}(\sum_{i=1}^{n_x} (x_i - \bar{x})^2 + \sum_{i=1}^{n_y} (y_i - \bar{y})^2)$$

Under the null hypothesis, T_{stat} follows a t-distribution with $n_x + n_y - 2$ degree of freedom. The p-value is therefore

$$\begin{aligned} p_{\text{value}} &= \mathrm{P}_{H_0}(|T_{\text{stat}}| > |T_{\text{stat}}^{\text{obs}}|) \\ &= \mathrm{P}(t_{n_x+n_y-2} \leq -T_{\text{stat}}^{\text{obs}}) + 1 - \mathrm{P}(t_{n_x+n_y-2} \leq T_{\text{stat}}^{\text{obs}}) \end{aligned}$$

```
nx <- length(x)
ny <- length(y)
x.mean <- mean(x)
y.mean <- mean(y)
x.sc <- sum((x-x.mean)^2)
y.sc <- sum((y-y.mean)^2)
xy.sd <- sqrt((x.sc+y.sc)/(nx+ny-2))
t.stat <- (x.mean-y.mean)/xy.sd/sqrt(1/nx+1/ny)
df <- nx + ny -2
p.value <- pt(-t.stat,df) + (1- pt(t.stat,df))
c(t.stat, df, p.value)
```

```
## [1]  1.5426375 76.0000000  0.1270726
```

The confidence interval for the mean difference $\mu_x - \mu_y$ is computed as

$$\text{CI}_{1-\alpha} = [\bar{x} - \bar{y} + s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\, qt_{\frac{\alpha}{2}, n_x+n_y-2} \;,\; \bar{x} - \bar{y} + s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\, qt_{1-\frac{\alpha}{2}, n_x+n_y-2}\ ]$$

```
x.mean-y.mean  + xy.sd*sqrt(1/nx+1/ny)*qt(c(alpha/2,1-alpha/2),df)
```

```
## [1] -4.358031 34.301621
```

2.2.3 Assuming different variances

Assuming equal variances for the two groups may be disputable.

```
aggregate(data$weight ~ data$regime, FUN= "sd" )
```

```
##   data$regime data$weight
## 1     Control    123.0689
## 2         GMO    111.8559
```

We can then use the `t.test` function with different variances (which is the default)

```
t.test(x, y, conf.level=1-alpha)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 1.5426, df = 75.976, p-value = 0.1271
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.358129 34.301719
## sample estimates:
## mean of x mean of y
##  513.7077  498.7359
```

The Welch (or Satterthwaite) approximation to the degrees of freedom is used instead of $n_x + n_y - 2 =$ 76:

$$\mathrm{df}_W = \frac{(c_x + c_y)^2}{c_x^2/(n_x - 1) + c_y^2/(n_y - 1)}$$

where $c_x = \sum(x_i - \bar{x})^2/(n_x(n_x - 1))$ and $c_y = \sum(y_i - \bar{y})^2/(n_y(n_y - 1))$.

Furthermore, unlike in Student's t-test with equal variances, the denominator is not based on a pooled variance estimate:

$$T_{\mathrm{stat}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$$

where s_x^2 and s_y^2 are the empirical variances of (x_i) and (y_i):

$$s_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \bar{x})^2 \quad ; \quad s_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \bar{y})^2$$

```
sbar.xy <- sqrt(var(x)/nx+var(y)/ny)
t.stat <- (x.mean-y.mean)/sbar.xy
cx <- x.sc/(nx-1)/nx
cy <- y.sc/(ny-1)/ny
dfw <- (cx + cy)^2 / (cx^2/(nx-1) + cy^2/(ny-1))
p.value <- pt(-t.stat,dfw) + (1- pt(t.stat,dfw))
c(t.stat, dfw, p.value)
```

```
## [1]  1.5426375 75.9760868  0.1270739
```

The confidence interval for $\mu_x - \mu_y$ is now computed as

$$\mathrm{CI}_{1-\alpha} = [\bar{x} - \bar{y} + \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\ qt_{\frac{\alpha}{2},\mathrm{df}_W}\ ,\ \bar{x} - \bar{y} + \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\ qt_{1-\frac{\alpha}{2},\mathrm{df}_W}\ ]$$

```
x.mean-y.mean  + sbar.xy*qt(c(alpha/2,1-alpha/2),dfw)
```

```
## [1] -4.358129 34.301719
```

2.3 Power of a t-test

Until now, we have demonstrated that the experimental data does not highlight any significant difference in weight between the control group and the GMO group.

Of course, that does not mean that there is no difference between the two groups. Indeed, **absence of evidence is not evidence of absence**. In fact, no experimental study would be able to demonstrate the absence of effect of the diet on the weight.

Now, the appropriate question is rather to evaluate what the experimental study can detect. If feeding a population of rats with GMOs has a signicant biological effect on the weight, can we ensure with a reasonable level of confidence that our statistical test will reject the null hypothesis and conclude that there is indeed a difference in weight between the two groups?

A *power analysis* allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence. Conversely, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints.

For a given $\delta \in \mathrm{R}$, let $\beta(\delta)$ be the type II error rate, i.e. the probability to fail rejecting H_0 when $\mu_x - \mu_y = \delta$, with $\delta \neq 0$.

The power of the test is the probability to reject the null hypothesis when it is false. It is also a function of $\delta = \mu_x - \mu_y$ defined as

$$\begin{aligned}\eta(\delta) &= 1 - \beta(\delta) \\ &= \mathrm{P}(\text{reject } H_0 \mid \mu_x - \mu_y = \delta)\end{aligned}$$

Remember that, for a two sided test, we reject the null hypothesis when $|T_{\text{stat}}| > qt_{1-\alpha/2,\mathrm{df}}$, where df is the appropriate degree of freedom.

On the other hand, $(\bar{x} - \bar{y} - \delta)/s_{xy}$, where $s_{xy} = \sqrt{s_x^2/n_x + s_y^2/n_y}$, follows a t-distribution with df degrees of freedom. Thus,

$$\begin{aligned}\eta(\delta) &= 1 - \mathrm{P}(qt_{\frac{\alpha}{2},\mathrm{df}} < T_{\text{stat}} < qt_{1-\frac{\alpha}{2},\mathrm{df}} \mid \mu_x - \mu_y = \delta) \\ &= 1 - \mathrm{P}(qt_{\frac{\alpha}{2},\mathrm{df}} < \frac{\bar{x} - \bar{y}}{s} < qt_{1-\frac{\alpha}{2},\mathrm{df}} \mid \mu_x - \mu_y = \delta)\end{aligned}$$

$$= 1 - \mathrm{P}(qt_{\frac{\alpha}{2},\mathrm{df}} - \frac{\delta}{s_{xy}} < \frac{\bar{x} - \bar{y} - \delta}{s_{xy}} < qt_{1-\frac{\alpha}{2},\mathrm{df}} - \frac{\delta}{s_{xy}} \mid \mu_x - \mu_y = \delta)$$

$$= 1 - Ft_{\mathrm{df}}(qt_{1-\frac{\alpha}{2},\mathrm{df}} - \frac{\delta}{s_{xy}}) + Ft_{\mathrm{df}}(qt_{\frac{\alpha}{2},\mathrm{df}} - \frac{\delta}{s_{xy}})$$

As an example, let us compute the probability to detect a difference in weight of 10g with two groups of 80 rats each and assuming that the standard deviation is 30g in each group.

```
alpha=0.05
nx.new <- ny.new <- 80
delta.mu <- 10
x.sd <- 30
df <- nx.new+ny.new-2
dt <- delta.mu/x.sd/sqrt(1/nx.new+1/ny.new)
1-pt(qt(1-alpha/2,df)-dt,df) + pt(qt(alpha/2,df)-dt,df)
```

```
## [1] 0.5528906
```

The function `pwr.t.test` allows to compute this power:

```
library(pwr)
pwr.t.test(n=nx.new, d=delta.mu/x.sd, type="two.sample", alternative="two.sided", sig.level
=alpha)
```

```
##
##      Two-sample t test power calculation
##
##               n = 80
##               d = 0.3333333
##       sig.level = 0.05
##           power = 0.5538758
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Let us perform a Monte Carlo simulation, to check this result and better understand what it means. Imagine that the ``true'' difference in weight is $\delta = 10$ g. Then, if could repeat the same experiment a (very) large number of times, we would reject the null hypothesis in 55% of cases.

```
L <- 100000
mux <- 500
muy <- mux + delta.mu
Rt <- vector(length=L)
for (l in (1:L)) {
  x.sim <- rnorm(nx.new,mux,x.sd)
```

```
  y.sim <- rnorm(ny.new,muy,x.sd)
  Rt[l] <- t.test(x.sim, y.sim, alternative="two.sided")$p.value < alpha
}
mean(Rt)
```

```
## [1] 0.55311
```

We may consider this probability as too small. If our objective is a power of 80% at least, with the same significance level, we need to increase the sample size.

```
pwr.t.test(power=0.8, d=delta.mu/x.sd, sig.level=alpha)
```

```
##
##      Two-sample t test power calculation
##
##               n = 142.2462
##               d = 0.3333333
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Indeed, we see that $n \geq$ 143 animals per group are required in order to reach a power of 80%.

```
nx.new <- ny.new <- ceiling(pwr.t.test(power=0.8, d=delta.mu/x.sd, sig.level=alpha)$n)
df <- nx.new+ny.new-2
dt <- delta.mu/x.sd/sqrt(1/nx.new+1/ny.new)
1-pt(qt(1-alpha/2,df)-dt,df) + pt(qt(alpha/2,df)-dt,df)
```

```
## [1] 0.8020466
```

An alternative for increasing the power consists in increasing the type I error rate

```
pwr.t.test(power=0.8, d=delta.mu/x.sd, n=80, sig.level=NULL)
```

```
##
##      Two-sample t test power calculation
##
##               n = 80
##               d = 0.3333333
##       sig.level = 0.2067337
##           power = 0.8
```

```
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

If we accept a significance level of about 20%, then we will be less demanding for rejecting H_0: we will reject the null hypothesis when $|T_{\text{stat}}| > qt_{0.9,158} = 1.29$, instead of $|T_{\text{stat}}| > qt_{0.975,158} = 1.98$. This strategy will therefore increase the power, but also the type I error rate.

3 Mann-Whitney-Wilcoxon test

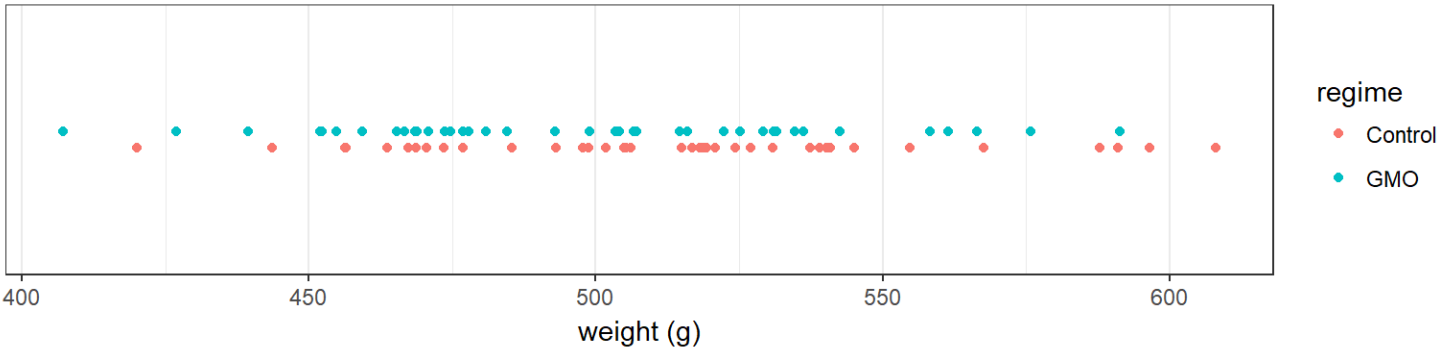
The Mann-Whitney-Wilcoxon test, or Wilcoxon rank sum test, can be used to test if the weight in one of the two groups tends to be greater than in the other group.

The Mann-Whitney-Wilcoxon test is a *non parametric test*: we don't make the assumption that the distribution of the data belongs to a family of parametric ditributions.

The logic behind the Wilcoxon test is quite simple. The data are ranked to produce two rank totals, one for each group. If there is a systematic difference between the two groups, then most of the high ranks will belong to one group and most of the low ranks will belong to the other one. As a result, the rank totals will be quite different and one of the rank totals will be quite small. On the other hand, if the two groups are similar, then high and low ranks will be distributed fairly evenly between the two groups and the rank totals will be fairly similar.

In our example, we don't clearly see any of the two groups on the right or on the left of the scatter plot

```
ggplot(data=data[data$gender=="Male",]) +  geom_point(aes(x=weight,y=as.numeric(regime),col
our=regime)) +
  ylab(NULL) + scale_y_continuous(breaks=NULL, limits=c(-6,9)) + xlab("weight (g)")
```



We can check that the Mann-Whitney-Wilcoxon test is not significant (at the level 0.05)

```
wilcox.test(x, y, alternative="two.sided", conf.level=1-alpha)
```

```
## Warning in wilcox.test.default(x, y, alternative = "two.sided", conf.level
## = 1 - : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  x and y
## W = 904.5, p-value = 0.1516
## alternative hypothesis: true location shift is not equal to 0
```

The test statistic W_x is computed a follows:

- Assign numeric ranks to all the observations, beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings
- define R_x (resp. R_y) as the sum of the ranks for the observations which came from sample x (resp. y)
- Let $W_x = R_x - n_x(n_x+1)/2$ and $W_y = R_y - n_y(n_y+1)/2$

```
nx <- length(x)
ny <- length(y)
Wx=sum(rank(c(x,y))[1:nx]) - nx*(nx+1)/2
Wy=sum(rank(c(y,x))[1:ny]) - ny*(ny+1)/2
c(Wx, Wy)
```

```
## [1] 904.5 616.5
```

For a two sided tests and assuming that $W_x^{\text{obs}} > W_y^{\text{obs}}$, the p-value is

$$p_{\text{value}} = \mathrm{P}(W_y \leq W_y^{\text{obs}}) + \mathrm{P}(W_x \geq W_x^{\text{obs}})$$

The distribution of W_x and W_y are tabulated and this p-value can then be computed

```
pwilcox(Wy,ny,nx)+ 1 - pwilcox(Wx,nx,ny)
```

```
## [1] 0.1508831
```

We could of course exchange the roles of x and y. In this case the test statistic would be W_y but the p-value would be the same.

```
wilcox.test(y, x, alternative="two.sided", conf.level=1-alpha)
```

```
## Warning in wilcox.test.default(y, x, alternative = "two.sided", conf.level
## = 1 - : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  y and x
## W = 616.5, p-value = 0.1516
## alternative hypothesis: true location shift is not equal to 0
```

Remark: It is easy to show that $W_x + W_y = n_x n_y$

```
c(Wx+Wy, nx*ny)
```

```
## [1] 1521 1521
```

Unlike the t-test, the Mann-Whitney-Wilcoxon does not require the assumption of normal distributions. However, it is nearly as efficient as the t-test on normal distributions. That means that both tests have similar power.

This important property can easily be checked by Monte Carlo simulation. Let us simulate L replicates of the experiments under H_1, assuming that $\mu_y = \mu_x + 15$. We can then compare the power of both tests by comparing the rajection rates of the null hypothesis.

```
L <- 10000
alpha <- 0.05
mux <- 500
muy <- 520
sdx <- sdy <- 30
nx <- ny <- 40
Rt <- vector(length=L)
Rw <- vector(length=L)
for (l in (1:L)) {
  x.sim <- rnorm(nx,mux,sdx)
  y.sim <- rnorm(ny,muy,sdy)
  Rt[l] <- t.test(x.sim, y.sim)$p.value < alpha
  Rw[l] <- wilcox.test(x.sim, y.sim)$p.value < alpha
}
c(mean(Rt), mean(Rw))
```

```
## [1] 0.8392 0.8230
```

On the other hand, the Wilcoxon test may be much more powerful than the t-test for non normal distributions when the empirical mean converges lowly in distribution to the normal distribution. Such is the case, for instance, of the log-normal ditribution which is strongly skewed for large variances.

```
mux <- 5
muy <- 6
sdx <- sdy <- 1
nx <- ny <- 20
Rt <- vector(length=L)
Rw <- vector(length=L)
for (l in (1:L)) {
  x.sim <- exp(rnorm(nx,mux,sdx))
  y.sim <- exp(rnorm(ny,muy,sdy))
  Rt[l] <- t.test(x.sim, y.sim, alternative="two.sided")$p.value < alpha
  Rw[l] <- wilcox.test(x.sim, y.sim, alternative="two.sided")$p.value < alpha
```

```
}
c(mean(Rt), mean(Rw))
```

```
## [1] 0.6679 0.8487
```

4 The limited role of the p-value

First of all it is important to emphasis that statistics is a tool for supporting decision-making. It is not a decision tool that can be used blindly.

A p-value below the sacrosanct 0.05 threshold does not mean that GMOs have some negative impacts on human health, or that a drug is better than another one. On the other hand, a p-value above 0.05 does not mean that GMOs are safe or that a drug has no effect.

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and P-Values" with six principles underlying the proper use and interpretation of the p-value.

"The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post p<0.05 era.'" "Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as 'p-hacking' and 'data dredging' that emphasize the search for small p-values over other statistical and scientific reasoning."

The statement's six principles, many of which address misconceptions and misuse of the pvalue, are the following:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

As an illustration, assume that the diet has a real impact on the weight: after 14 weeks, rats fed with GMOs weigh in average 15g less than control rats.

It will be extremly unlikely to conclude that there is an effect with only 10 rats per group, even if we observe a difference of 15g in the two samples.

```
n <- 10
mu <- 500
delta <- 15
sd <- 30
x.sim <- rnorm(n,mu,sd)
y.sim <- x.sim + delta
```

```
t.test(x.sim, y.sim)
```

```
##
##  Welch Two Sample t-test
##
## data:  x.sim and y.sim
## t = -0.77216, df = 18, p-value = 0.45
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -55.8125  25.8125
## sample estimates:
## mean of x mean of y
##  504.9729  519.9729
```

This basic example shows that a difference considered as biologically significant may not be statistically significant.

On the other hand, a small difference considered as not biologically significant (1g for instance) may be considered as statistically significant if the group sizes are large enough.

```
n <- 10000
delta <- 1
x.sim <- rnorm(n,mu,sd)
y.sim <- x.sim + delta
t.test(x.sim, y.sim)
```

```
##
##  Welch Two Sample t-test
##
## data:  x.sim and y.sim
## t = -2.3407, df = 19998, p-value = 0.01926
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.8373819 -0.1626181
## sample estimates:
## mean of x mean of y
##  499.9298  500.9298
```

This example confirms the need for a power analysis as a complement of the comparison test. We will see that equivalence testing may also be relevant for evaluating what the data allows to say.

5 Equivalence tests

5.1 Introduction

Traditional hypothesis testing seeks to determine if means are the same or different. Such approach has several drawbacks:

1. Testing that two means are exactly the same is usually of little interest. A very small difference may exist, even if it is not biologically, or physically significant. It is much more meaningful in such situation to test if some significant difference exists, i.e. if this difference may have some concrete impact.
2. When testing difference between means, the null hypothesis considers that there is no differences: it thus falls to the data to demonstrate the converse. In the absence of enough data, we may fail to detect a significant difference, because of the lack of power. An opposite point of view consists of applying the basic hypothesis that a significant difference exists: it now falls to the data to demonstrate the converse.

On the other hand, equivalence testing determines an interval where the means can be considered equivalent. In other words, equivalence does not mean that two means μ_x and μ_y are equal, but rather that they are ``close enough", i.e. $|\mu_x - \mu_y| < \delta$ for some equivalence limit δ that should be chosen according to the problem under study.

Thus, in terms of hypothesis testing, we want to test

$$H_0 : \text{``}|\mu_x - \mu_y| \geq \delta\text{''} \quad \text{versus} \quad H_1 : \text{``}|\mu_x - \mu_y| < \delta\text{''}$$

These tests are currently used in the field of medication, to put a generic drug on the market. These are bioequivalence tests. Equivalence testing may also be used to determine if new therapies have equivalent or noninferior efficacies to the ones currently in use. These studies are called equivalence/noninferiority studies.

In the field of GMO risk assessment, equivalence testing is required to demonstrate that a GMO crop is compositionally equivalent and as safe as a conventional crop.

5.2 Two samples test

5.2.1 The TOST procedure

The simplest and most widely used approach to test equivalence is the two one-sided test (TOST) procedure. Let $\mu_d = \mu_x - \mu_y$, then TOST consists in performing the two tests:

$$\begin{aligned} H_0^{(+)} &: \text{``}\mu_d \geq \delta\text{''} \quad \text{versus} \quad H_1^{(+)} : \text{``}\mu_d < \delta\text{''} \\ H_0^{(-)} &: \text{``}\mu_d \leq -\delta\text{''} \quad \text{versus} \quad H_1^{(-)} : \text{``}\mu_d > -\delta\text{''} \end{aligned}$$

which is equivalent to test

$$\begin{aligned} H_0^{(+)} &: \text{``}\mu_d = \delta\text{''} \quad \text{versus} \quad H_1^{(+)} : \text{``}\mu_d < \delta\text{''} \\ H_0^{(-)} &: \text{``}\mu_d = -\delta\text{''} \quad \text{versus} \quad H_1^{(-)} : \text{``}\mu_d > -\delta\text{''} \end{aligned}$$

Proposition: Using TOST, equivalence is established at the α significance level if a $(1-2\alpha)\times 100\%$ confidence interval for the difference $\mu_x - \mu_y$ is contained within the interval $(-\delta, \delta)$.

Proof: Let $d_i = x_i - y_i$ and $\bar{d} = \bar{x} - \bar{y}$. Let $s_{\bar{d}}$ be the estimated standard deviation of $\bar{d}$.

$$s_{\bar{d}} = \begin{cases} s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} & \text{if } \sigma_x^2 = \sigma_y^2 \\ \sqrt{s_x^2/n_x + s_y^2/n_y} & \text{otherwise} \end{cases}$$

Then, we

- reject $H_0^{(+)}$ if $T_{\text{stat}}^{(+)} = (\bar{d} - \delta)/s_{\bar{d}} < qt_{\alpha,v}$
  $H_0^{(-)}$ $T_{\text{stat}}^{(-)}$

- reject H_0 if $T_{\text{stat}} = (\bar{d} + \delta)/s_{\bar{d}} > qt_{1-\alpha,v}$

where v is the appropriate degree of freedom. Using the fact that $qt_{\alpha,v} = -qt_{1-\alpha,v}$, these decision rules are equivalent to:

- reject $H_0^{(+)}$ if $\bar{d} + qt_{1-\alpha,v}\, s_{\bar{d}} < \delta$
- reject $H_0^{(-)}$ if $\bar{d} + qt_{\alpha,v}\, s_{\bar{d}} > -\delta$

By definition, $\bar{d} + qt_{\alpha,v}\, s_{\bar{d}}$ and $\bar{d} + qt_{1-\alpha,v}\, s_{\bar{d}}$ are the bounds of a $(1-2\alpha)\times 100\%$ confidence interval for the difference mean μ_d.

We therefore reject the null hypothesis H_0 when both $H_0^{(+)}$ and $H_0^{(-)}$ are rejected, i.e. when a $(1-2\alpha)\times 100\%$ confidence interval for μ_d is contained within the equivalence limits $(-\delta, \delta)$. □

In our example, assuming unequal variance, the 90% confidence interval for $\mu_d = \mu_x - \mu_y$ is now

```
d.mean <- x.mean-y.mean
s.d.mean <- sqrt(var(x)/length(x) + var(y)/length(y))
d.mean + s.d.mean*qt(c(alpha,1-alpha),dfw)
```

```
## [1] -1.189099 31.132689
```

The calculated confidence interval can be plotted together with the value 0 (for difference testing) and the equivalence limits $-\delta$, δ. Such a plot will immediately reveal whether the GMO is significantly different from the conventional counterpart (at the $2(1-\alpha)$ confidence level), and/or equivalence can be claimed or denied (at the $1-\alpha$ confidence level).

Considering for instance that a difference of weight larger than 20g is biologically significant leads to choose $\delta = 20$. Since the 90% confidence interval is not contained in the equivalence interval $(-\delta, \delta)$, we don't reject the null hypothesis for a significance level of 0.05. We therefore cannot conclude to equivalence, even if the difference in mean is not statistically significant.

On the other hand, we will conclude that the two regimens are nutritionally equivalent if we choose 40g as a limit since the new equivalence interval $(-40, 40)$ contains the CI.

When it is considered useful to have results also in the form of a p-value from a statistical significance test, then it can be easily calculated as follows:

$$\begin{aligned} p_{\text{value}} &= p_{\text{value}}^{(+)} + p_{\text{value}}^{(-)} \\ &= \mathrm{P}\left(\frac{\bar{d} - \delta}{s_{\bar{d}}} < \frac{\bar{d}^{\text{obs}} - \delta}{s_{\bar{d}}} \mid \mu_d = \delta\right) + \mathrm{P}\left(\frac{\bar{d} + \delta}{s_{\bar{d}}} > \frac{\bar{d}^{\text{obs}} + \delta}{s_{\bar{d}}} \mid \mu_d = -\delta\right) \\ &= Ft_{\text{df}\,w}\left(\frac{\bar{d}^{\text{obs}} - \delta}{s_{\bar{d}}}\right) + 1 - Ft_{\text{df}\,w}\left(\frac{\bar{d}^{\text{obs}} + \delta}{s_{\bar{d}}}\right) \end{aligned}$$

Using $\delta = 20$, the p-value is greater than the significance level $\alpha = 0.05$:

```
delta = 20
pt((d.mean-delta)/s.d.mean, dfw) + 1 - pt((d.mean+delta)/s.d.mean, dfw)
```

```
## [1] 0.3032307
```

Then, we don't reject the null hypothesis. On the contrary, the p-value obtained with $\delta = 40$ is smaller than the significance level and the null hypothesis of non equivalence can be rejected:

```
delta = 40
pt((d.mean-delta)/s.d.mean, dfw) + 1 - pt((d.mean+delta)/s.d.mean, dfw)
```

```
## [1] 0.005923576
```

Function tost {equivanece} computes a TOST for equivalence from paired or unpaired data

```
library(equivalence)
tost(x, y, alpha=0.05, epsilon=20)
```

```
##
##  Welch Two Sample TOST
##
## data:  x and y
## df = 75.976
## sample estimates:
## mean of x mean of y
##  513.7077  498.7359
##
## Epsilon: 20
## 95 percent two one-sided confidence interval (TOST interval):
##  -1.189099 31.132689
## Null hypothesis of statistical difference is: not rejected
## TOST p-value: 0.3029514
```

```
tost(x, y, alpha=0.05, epsilon=40)
```

```
##
##  Welch Two Sample TOST
##
## data:  x and y
## df = 75.976
## sample estimates:
## mean of x mean of y
##  513.7077  498.7359
##
## Epsilon: 40
## 95 percent two one-sided confidence interval (TOST interval):
```

```
##  -1.189099 31.132689
## Null hypothesis of statistical difference is: rejected
## TOST p-value: 0.005923451
```

5.2.2 Difference testing versus equivalence testing

The choice of the test mainly depends on how the conclusion is formulated.

As an example, AFSSA (currently ANSES) was committing a fairly serious error in this area by concluding in relation to the MON863 GMO maize: *Considering that no significant difference has been observed between the results obtained for MON863 maize and for the other varieties of maize, one might, therefore, conclude that the new plant is nutritionally equivalent* (AFSSA, Saisine 2003-0215, p. 6).

The absence of statistically significant difference does not allow to conclude on equivalence. The European Food Safety Authority (EFSA) published a "Scientific Opinion" on this topic.

In particular, EFSA considers that *statistical methodology should not be focussed exclusively on either differences or equivalences, but should provide a richer framework within which the conclusions of both types of assessment are allowed. Both approaches are complementary: statistically significant differences may point at biological changes caused by the genetic modification, but may not be relevant from the viewpoint of food safety. On the other hand, equivalence assessments may identify differences that are potentially larger than normal natural variation, but such cases may or may not be cases where there is an indication for true biological change caused by the genetic modification. A procedure combining both approaches can only aid the subsequent toxicological assessment following risk characterization of the statistical results.*

EFSA also propose the following classification of the possible outcomes:

After adjustment of the equivalence limits, a single confidence limit (for the difference) serves visually for assessing the outcome of both tests (difference and equivalence). Here, only the upper adjusted equivalence limit is considered. Shown are: the mean of the GM crop on an appropriate scale (square), the confidence limits (whiskers) for the difference between the GM crop and its conventional couterpart (bar shows confidence interval), a vertical line indicating zero difference (for proof of difference), and vertical lines indicating adjusted equivalence limits (for proof of equivalence).

For outcome types 1, 3 and 5 the null hypothesis of no difference cannot be rejected: for outcomes 2, 4, 6 and 7 the GM crop is different from its conventional counterpart. Regarding interpretation of equivalence, four categories (i) - (iv) are identified: in category (i) the null hypothesis of non-equivalence is rejected in favour of equivalence; in categories (ii) equivalence is more likely than not (further evaluation may be required), (iii) non-equivalence is more likely than not (further evaluation required) and (iv) non-equivalence is established (further evaluation required).

5.3 One sample test

Even if equivalence testing is mainly used for testing the equivalence between two populations, a one sample equivalence test is also possible.

Assume that, for some $\delta > 0$, we want to test

$$H_0 : \text{``}|\mu_x - \mu_0| \geq \delta \text{''} \quad \text{versus} \quad H_1 : \text{``}|\mu_x - \mu_0| < \delta \text{''}$$

Let $z_i = x_i - \mu_0$, $i = 1, 2, \ldots, n_x$, and let $\mu_z = \mu_x - \mu_0$. Then, the test reduces to use the (z_i) for testing

$$H_0 : ``|\mu_z| \geq \delta" \quad \text{versus} \quad H_1 : ``|\mu_z| < \delta"$$

```
mu0 <- 500
delta <- 10
alpha <- 0.05
z <- x - mu0
tost(z, alpha=alpha, epsilon=delta)
```

```
##
##  One Sample TOST
##
## data:  z
## df = 38
## sample estimates:
## mean of x
##  13.70769
##
## Epsilon: 10
## 95 percent two one-sided confidence interval (TOST interval):
##   2.035311 25.380074
## Null hypothesis of statistical difference is: not rejected
## TOST p-value: 0.7023006
```

Here, the two-sided null hypothesis is the union of the one sided hypotheses $``\mu_z \geq \delta"$ and $``\mu_z \leq -\delta"$. Thus,

$$\begin{aligned} p_{\text{value}} &= \mathrm{P}(\bar{z} < \bar{z}^{\text{obs}} \mid \mu_z = \delta) + \mathrm{P}(\bar{z} > \bar{z}^{\text{obs}} \mid \mu_z = -\delta) \\ &= \mathrm{P}(t_{n_z-1} < \frac{\bar{z}^{\text{obs}} - \delta}{s_{\bar{z}}}) + \mathrm{P}(t_{n_z-1} > \frac{\bar{z}^{\text{obs}} + \delta}{s_{\bar{z}}}) \\ &= Ft_{n_x-1}\left(\frac{\bar{x}^{\text{obs}} - \mu_0 - \delta}{s_{\bar{x}}}\right) + 1 - Ft_{n_x-1}\left(\frac{\bar{x}^{\text{obs}} - \mu_0 + \delta}{s_{\bar{x}}}\right) \end{aligned}$$

```
z.mean <- mean(z)
zbar.sd <- sd(z)/sqrt(nx)
p.value <- pt((z.mean-delta)/zbar.sd, nx-1) + 1 - pt((z.mean+delta)/zbar.sd, nx-1)
p.value
```

```
## [1] 0.6592176
```

Marc Lavielle
January 17th, 2018

1 Introduction

When we perform a large number of statistical tests, some will have p-values less than 0.05 purely by chance, even if all the null hypotheses are really true.

More precisely, if we do a large number m of statistical tests, and for $m_{0.}$ of them the null hypothesis is actually true, we would expect about 5% of the $m_{0.}$ tests to be significant at the 0.05 level, just due to chance: these significant results are false discoveries (or false positives). On the other hand, if some alternative hypothesis are true, we can miss some of them: these non significant results are false negatives.

	Null hypothesis true	Alternative hypothesis true	
Test non significant	m_{00}	m_{10}	$m_{.0}$
Test significant	m_{01}	m_{11}	$m_{.1}$
	$m_{0.}$	$m_{1.}$	m

If important conclusions and decisions are based on these false positives, it is then important to control the *family-wise error rate* (FWER):

- the family-wise error rate is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests

$$FWER = \mathrm{P}(m_{01} \geq 1)$$

When true positives are expected, it is possible to miss some of them. We then necessarily need to accept false positives if we want to limit the number of these false negatives. It is important in such situation to control the false discovery rate (FDR)

- The false discovery rate (FDR) is the expected proportion of false discoveries among the discoveries

$$FDR = \mathrm{E}\left(\frac{m_{01}}{m_{01} + m_{11}}\right) = \mathrm{E}\left(\frac{m_{01}}{m_{.1}}\right)$$

Several procedures exist for controlling either the FWER or the FDR.

2 Distribution of the p-values

2.1 Introduction

The health effects of a Roundup-tolerant genetically modified maize, cultivated with or without Roundup, and Roundup alone, were studied during a 2 years study in rats.

For each sex, one control group had access to plain water and standard diet from the closest isogenic non-transgenic maize control; six groups were fed with 11, 22 and 33% of GM NK603 maize either treated or not with Roundup. The final three groups were fed with the control diet and had access to water supplemented with different concentrations of Roundup.

A sample of 200 rats including 100 males and 100 females was randomized into 20 groups of 10 rats of the same sex. Within each group, rats received the same diet. For each sex, there are therefore nine experimental groups and one control group.

The file ratSurvival.csv reports the lifespan (in days) for each animal. Here, the experiment stopped after 720 days. Then, the reported survival time is 720 for those animals who were still alive at the end of the experiment.

See the opinion of the Haut Conseil des Biotechnologies for more information about this study.

Here is a summary of the data,

```
data <- read.csv("ratSurvival.csv")
summary(data)
```

```
##       time              regimen      gender
##  Min.   :100.0   control    :20   female:100
##  1st Qu.:590.0   NK603-11%  :20   male  :100
##  Median :660.0   NK603-11%+R:20
##  Mean   :631.6   NK603-22%  :20
##  3rd Qu.:720.0   NK603-22%+R:20
##  Max.   :720.0   NK603-33%  :20
##                  (Other)    :80
```

```
levels(data$regimen)
```

```
##  [1] "control"     "NK603-11%"   "NK603-11%+R" "NK603-22%"   "NK603-22%+R"
##  [6] "NK603-33%"   "NK603-33%+R" "RoundUp A"   "RoundUp B"   "RoundUp C"
```

2.2 Single comparison between 2 groups

One objective of this study is the comparison of the survival between the control group and the experimental groups.

Consider for instance the control group of females

```
data.control <- subset(data, regimen=="control" & gender=="female")
data.control$time
```

```
##  [1] 540 645 720 720 720 720 720 720 720 720
```

Only 2 rats of this group died before the end of the experiment. On the other hand, 7 females of the group fed with 22% of maize NK693 died during the experiment.

```
data.test <- subset(data, regimen=="NK603-22%" & gender=="female")
data.test$time
```

```
##  [1] 290 475 480 510 550 555 650 720 720 720
```

A negative effect of the diet on the survival means that the rats of the experimental group tend to die before those of the control group. Then, we would like to test H_0 : *the NK603 11% diet has no effect on the survival of female rats''* against H_1: *the NK603 11% diet leads to decreased survival time for female rats''*.

In terms of survival functions, that means that, under H_1, the probability to be alive at a given time t is lower for a rat of the experimental group than for a rat of the control group. We then would like to test

$$H_0 : \text{``P}(T_{\text{test}} > t) = \text{P}(T_{\text{control}} > t), \text{ for any } t > 0 \text{''} \quad \text{versus} \quad H_1 : \text{``P}(T_{\text{test}} > t) < \text{P}(T_{\text{control}} > t), \text{ for any } t > 0 \text{''}$$

Because of the (right) censoring process, we cannot just compare the mean survival times using a t-test. On the other hand, we can use the Wilcoxon-Mann-Whitney test which precisely aims to compare the ranks of the survival times in both groups.

```
wilcox.test(data.test$time, data.control$time, alternative="less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data.test$time and data.control$time
## W = 22, p-value = 0.01144
## alternative hypothesis: true location shift is less than 0
```

Here, the p-value should lead us to reject the null hypothesis and conclude that 22% of the GM maize in the diet has a negative effect on the survival.

2.3 A single comparison… among many others

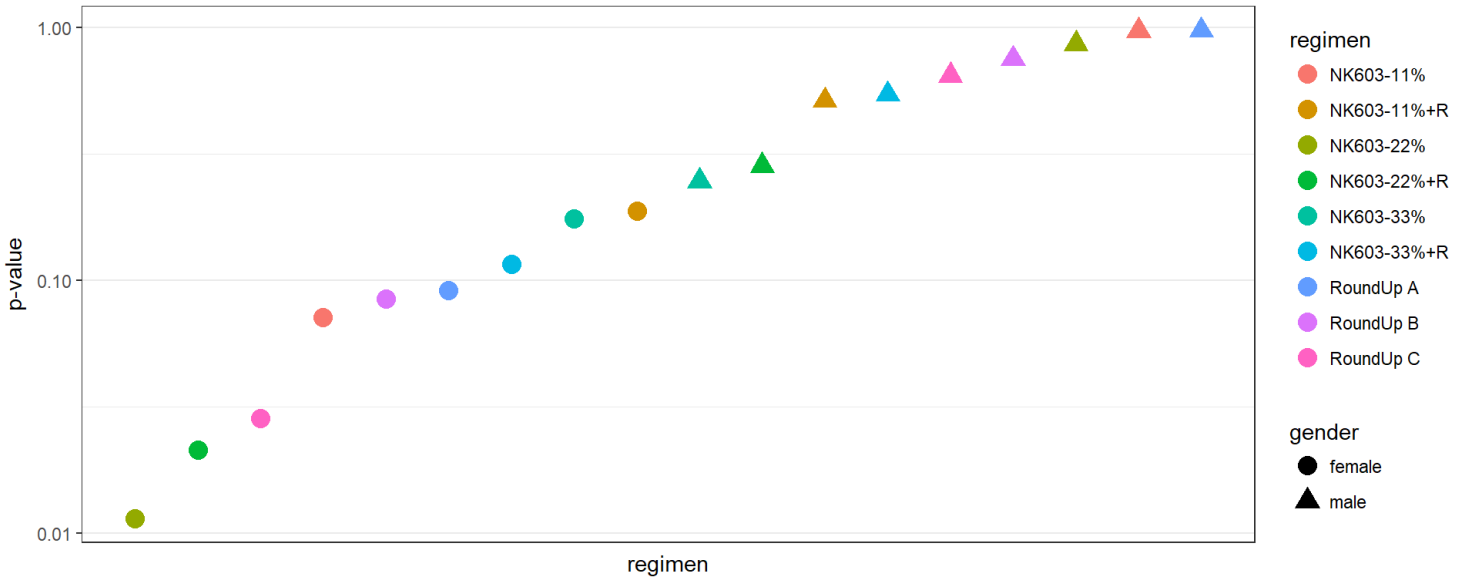
Should we really accept this conclusion as it stands? No, because we don't know the whole storyRemember that there are 9 experimental groups for each sex. Then, 18 comparisons with the control groups are performed.

```
library(ggplot2) ; theme_set(theme_bw())
pval <- stat <- gender <- regimen <- NULL
for (g in levels(data$gender)) {
  data.control <- subset(data, regimen=="control" & gender==g)
  for (r in levels(data$regimen)) {
    if (r != "control") {
      data.test <- subset(data, gender==g & regimen==r)
      wt <- wilcox.test(data.test$time, data.control$time, alternative="less")
      pval <- c(pval,wt$p.value)
      stat <- c(stat,wt$statistic)
      gender <- c(gender, g)
      regimen <- c(regimen, r)
    }
  }
}
R <- data.frame(gender=gender, regimen=regimen,  stat=stat, p.value=pval)
R <-  R[order(R$p.value),]
R
```

```
##    gender     regimen stat    p.value
## 3  female   NK603-22% 22.0 0.01143779
## 4  female NK603-22%+R 25.0 0.02131745
## 9  female   RoundUp C 26.5 0.02845455
## 1  female   NK603-11% 33.0 0.07166095
## 8  female   RoundUp B 34.0 0.08459162
## 7  female   RoundUp A 34.5 0.09156610
## 6  female NK603-33%+R 36.0 0.11556850
## 5  female   NK603-33% 39.0 0.17583926
## 2  female NK603-11%+R 40.0 0.18797769
## 14   male   NK603-33% 40.5 0.24733242
## 13   male NK603-22%+R 42.0 0.28493944
## 11   male NK603-11%+R 50.0 0.51508635
## 15   male NK603-33%+R 51.0 0.54532608
## 18   male   RoundUp C 54.5 0.64764235
## 17   male   RoundUp B 58.5 0.75283120
## 12   male   NK603-22% 64.0 0.86465902
## 10   male   NK603-11% 74.5 0.97160483
## 16   male   RoundUp A 75.5 0.97688776
```

Let us plot the ordered p-values:

```
ggplot(data=R) + geom_point(aes(x=1:18,color=regimen,y=p.value,shape=gender), size=4) +
  scale_y_log10() + xlab("regimen") + ylab("p-value") + scale_x_continuous(breaks=NULL)
```



If we then decide to only report the largest observed differences, associated to the smallest p-values, how can we conclude that these differences are statistically significant?

2.4 The Bonferroni correction

Imagine that we perform m comparisons and that all the m null hypotheses are true, i.e. $m = m_0$. .. If we use the same significance level α_m for the m tests, how should we choose α_m in order to control the family-wise error rate (FWER)?

$$\begin{aligned} \text{FWER} &= \text{P}(m_{01} \geq 1) \\ &= 1 - \text{P}(m_{01} = 0) \\ &= 1 - (1 - \alpha_m)^m \end{aligned}$$

Then, if we set FWER= α , the significance level for each individual test should be

$$\begin{aligned} \alpha_m &= 1-(1-\alpha)^{\frac{1}{m}} \qquad \text{(Sidak correction)} \\ &\simeq \frac{\alpha}{m} \qquad \text{(Bonferroni correction)} \end{aligned}$$

Let p_k be the p-value of the k-th test. Using the Bonferroni correction, the k-th test is significant if

$$p_k \leq \alpha_m \quad \Longleftrightarrow \quad m\,p_k \leq \alpha$$

We can then either compare the original p-value p_k to the corrected significance level α/m , or compare the *adjusted* p-value $p_k^{(\text{bonferroni})} = \min(1, m\,p_k)$ to the critical value α.

```
m <- nrow(R)
R$pv.bonf <- pmin(1, R$p.value*m)
R
```

```
##    gender     regimen stat    p.value   pv.bonf
## 3  female   NK603-22% 22.0 0.01143779 0.2058803
## 4  female NK603-22%+R 25.0 0.02131745 0.3837141
## 9  female   RoundUp C 26.5 0.02845455 0.5121818
## 1  female   NK603-11% 33.0 0.07166095 1.0000000
## 8  female   RoundUp B 34.0 0.08459162 1.0000000
## 7  female   RoundUp A 34.5 0.09156610 1.0000000
## 6  female NK603-33%+R 36.0 0.11556850 1.0000000
## 5  female   NK603-33% 39.0 0.17583926 1.0000000
## 2  female NK603-11%+R 40.0 0.18797769 1.0000000
## 14   male   NK603-33% 40.5 0.24733242 1.0000000
## 13   male NK603-22%+R 42.0 0.28493944 1.0000000
## 11   male NK603-11%+R 50.0 0.51508635 1.0000000
## 15   male NK603-33%+R 51.0 0.54532608 1.0000000
## 18   male   RoundUp C 54.5 0.64764235 1.0000000
## 17   male   RoundUp B 58.5 0.75283120 1.0000000
## 12   male   NK603-22% 64.0 0.86465902 1.0000000
## 10   male   NK603-11% 74.5 0.97160483 1.0000000
## 16   male   RoundUp A 75.5 0.97688776 1.0000000
```

Using the Bonferroni correction, none of the 18 comparisons is significant.

Remark: the function `p.adjust` proposes several adjustements of the p-values for multiple comparisons, including the Bonferroni adjustment:

```
p.adjust(pval, method = "bonferroni")
```

```
##  [1] 1.0000000 1.0000000 0.2058803 0.3837141 1.0000000 1.0000000 1.0000000
##  [8] 1.0000000 0.5121818 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [15] 1.0000000 1.0000000 1.0000000 1.0000000
```

The Bonferroni correction is appropriate when a single false positive in a set of tests would be a problem. It is mainly useful when there are a fairly small number of multiple comparisons and very few of them might be significant. The main drawback of the Bonferroni correction is its lack of power: it may lead to a very high rate of false negatives.

Assume, for instance that all the female rats of all the experimental groups die before the end of the experiment. Each of the

$m = 18$ original tests is significant, but the Bonferroni correction would lead to considering none of these tests as significant:

```
x <- subset(data, regimen=="control" & gender=="female")$time
y <- seq(610,700,length=10)
c(wilcox.test(x,y,"greater")$p.value, wilcox.test(x,y,"greater")$p.value*18)
```

```
## [1] 0.004444026 0.079992464
```

2.5 Permutation test

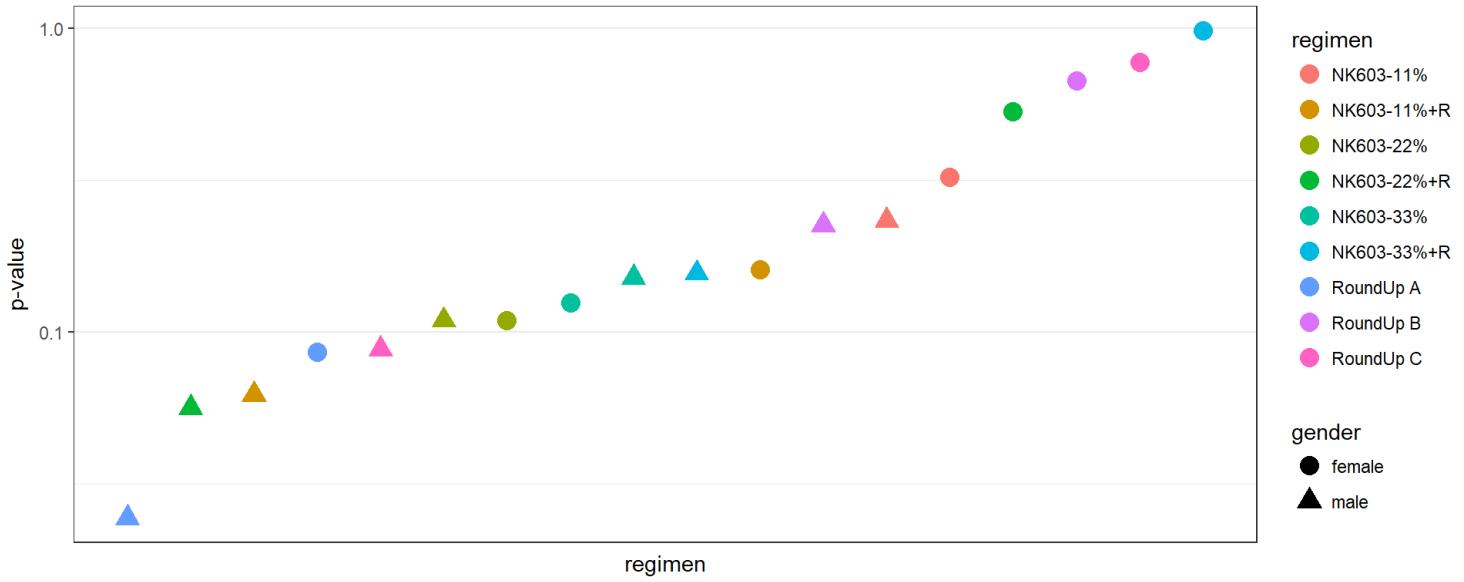
A permutation test (also called a randomization test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. If the labels are exchangeable under the null hypothesis, then the resulting tests yield exact significance levels. Prediction intervals can also be derived.

In our example, imagine that the null hypothesis is true. We can then randomly exchange the labels (i.e. the regimen) and perform the 18 comparisons between the experimental groups and the control groups.

```
set.seed(100)
dperm.m <- subset(data,  gender=="male")
n.m <- dim(dperm.m)[1]
dperm.m$regimen <- dperm.m$regimen[sample(n.m)]
dperm.f <- subset(data,  gender=="female")
n.f <- dim(dperm.f)[1]
dperm.f$regimen <- dperm.f$regimen[sample(n.f)]
dperm <- rbind(dperm.m,dperm.f)

pval <- gender <- regimen <- NULL
for (g in levels(dperm$gender)) {
  dperm.control <- subset(dperm, regimen=="control" & gender==g)
  for (r in levels(dperm$regimen)) {
    if (r != "control") {
      dperm.test <- subset(dperm, gender==g & regimen==r)
      wt <- wilcox.test(dperm.test$time, dperm.control$time, alternative="less")
      pval <- c(pval,wt$p.value)
      gender <- c(gender, g)
      regimen <- c(regimen, r)
    }
  }
}
R.p <- data.frame(gender=gender, regimen=regimen, p.value=pval)
Ro.p <-  R.p[order(R.p$p.value),]
```

The test statistics and the p-values now really behave how they are supposed to behave under the null hypothesis

```
ggplot(data=Ro.p) + geom_point(aes(x=1:18,color=regimen,y=p.value,shape=gender), size=4) +
  scale_y_log10() + xlab("regimen") + ylab("p-value") + scale_x_continuous(breaks=NULL)
```



If we now repeat the same experiment using many different permutations, we will be able to estimate the m distributions of the m test statistics as well as the m distributions of the m p-values under the null hypothesis.

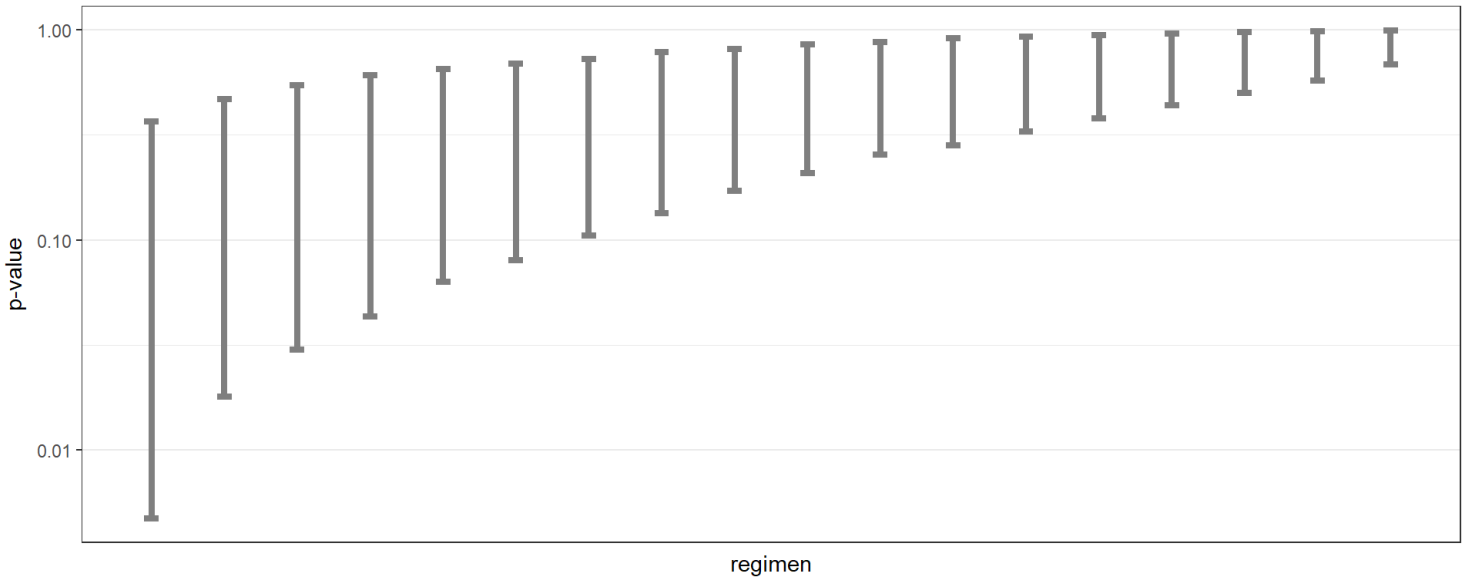
```
L <- 1000
dperm.m <- subset(data,  gender=="male")
n.m <- dim(dperm.m)[1]
dperm.f <- subset(data,  gender=="female")
n.f <- dim(dperm.f)[1]
PV <- ST <- NULL

for (l in (1:L)) {
  dperm.m$regimen <- dperm.m$regimen[sample(n.m)]
  dperm.f$regimen <- dperm.f$regimen[sample(n.f)]
  dperm <- rbind(dperm.m,dperm.f)
  pval <- stat <- NULL
  for (g in levels(dperm$gender)) {
    dperm.control <- subset(dperm, regimen=="control" & gender==g)
    for (r in levels(dperm$regimen)) {
      if (r != "control") {
        dperm.test <- subset(dperm, gender==g & regimen==r)
        wt <- wilcox.test(dperm.test$time, dperm.control$time, alternative="less")
        pval <- c(pval,wt$p.value)
        stat <- c(stat,wt$statistic)
      }
    }
  }
  PV <- rbind(PV,sort(pval))
  ST <- rbind(ST,sort(stat))
}
```

We can estimate, for instance, prediction intervals of level 90% for the $m = 18$ ordered p-values

```
q <- apply(PV, MARGIN = 2, quantile, probs = c(0.05, 0.5, 0.95))
q <- as.data.frame(t(q))
names(q) <- c("low","median","up")
q$rank <- 1:dim(q)[1]
```

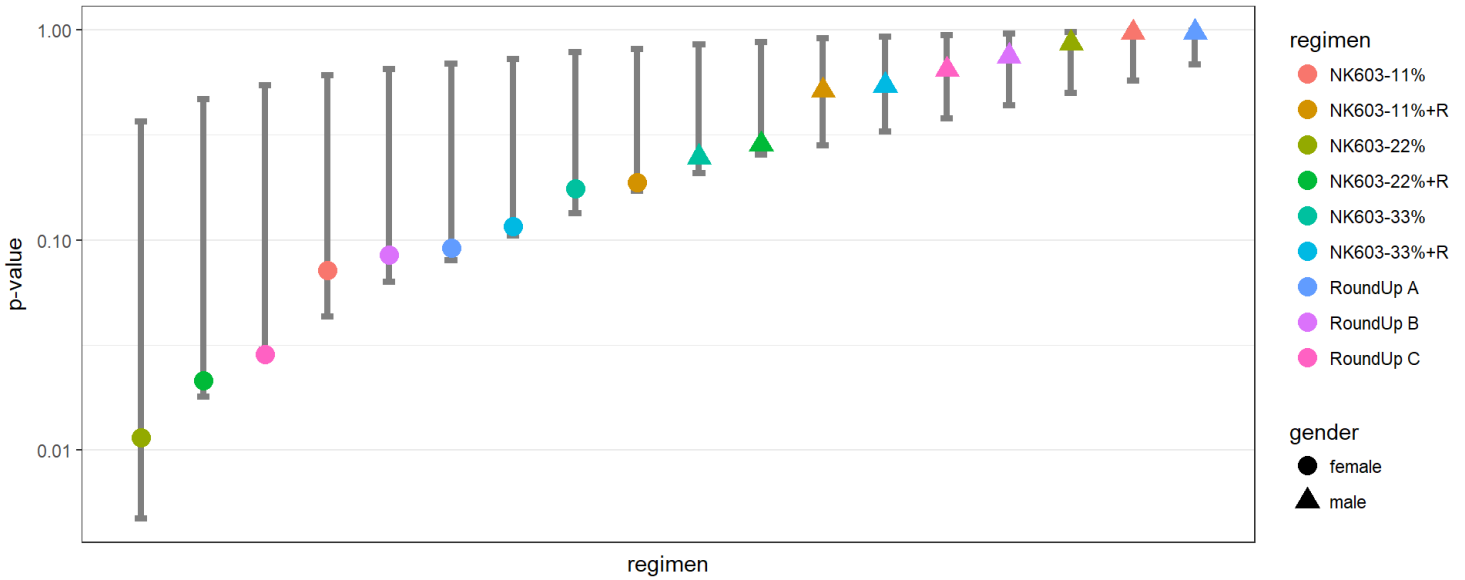
and plot them...

```
pl <- ggplot(data=q) +
  geom_errorbar(aes(x=rank, ymin=low, ymax=up), width=0.2,size=1.5,colour="grey50") +
  scale_y_log10() + xlab("regimen") + ylab("p-value") + scale_x_continuous(breaks=NULL)
pl
```



... with the original p-values

```
pl +  geom_point(data=R, aes(x=1:18,color=regimen,y=p.value,shape=gender), size=4)
```



Here, all the p-values, including the smallest ones, belong to the 90% prediction intervals: all the observed p-values behave individually how they are expected to behave under the null hypothesis.

In particular, when 18 comparisons are performed, it's not unlikely under the null hypothesis to obtain a smallest p-value less than or equal to the observed one (0.011).

The probability of such event can easily be estimated by Monte Carlo simulation. Let $p_{(1),\ell}$ be the smallest p-value obtained from the ℓ-th replicate of the Monte Carlo. Then,

$$\mathrm{P}(p_{(1)} \leq p_{(1)}^{\mathrm{obs}}) \simeq \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{1}\{p_{(1),\ell} \leq p_{(1)}^{\mathrm{obs}}\}$$

```
mean(ST[,1]<R$stat[1])
```

```
## [1] 0.139
```

3 Controlling the False Discovery Rate

3.1 Detecting associations

(Example from Handbook of Biological Statistics)

García-Arenzana et al. (2014) tested associations of 25 dietary variables with mammographic density, an important risk factor for breast cancer, in Spanish women. They found the following results

```
data <- read.csv("dietary.csv")
data
```

```
##              dietary p.value
## 1     Total calories   0.001
## 2          Olive oil   0.008
## 3         Whole milk   0.039
## 4         White meat   0.041
## 5           Proteins   0.042
## 6               Nuts   0.061
## 7  Cereals and pasta   0.074
## 8         White fish   0.205
## 9             Butter   0.212
## 10        Vegetables   0.216
## 11      Skimmed milk   0.222
## 12          Red meat   0.251
## 13             Fruit   0.269
## 14              Eggs   0.275
## 15         Blue fish   0.340
## 16           Legumes   0.341
## 17     Carbohydrates   0.384
## 18          Potatoes   0.569
## 19             Bread   0.594
## 20              Fats   0.696
## 21            Sweets   0.762
## 22    Dairy products   0.940
## 23 Semi-skimmed milk   0.942
## 24        Total meat   0.975
## 25    Processed meat   0.986
```

We can see that five of the variables show a significant p-value (<0.05). However, because García-Arenzana et al. (2014) tested 25 dietary variables, we would expect one or two variables to show a significant result purely by chance, even if diet had no real effect on mammographic density.

Applying the Bonferroni correction, we divide $\alpha = 0.05$ by the number of tests ($m = 25$) to get the Bonferroni critical value, so a test would have to have $p < 0.002$ to be significant. Under that criterion, only the test for total calories is significant.

An alternative approach is to control the false discovery rate, i.e the expected proportion of ``discoveries" (significant results) that are actually false positives. FDR control offers a way to increase power while maintaining some principled bound on error.

Imagine for instance that we compare expression levels for 20,000 genes between liver tumors and normal liver cells. We are

going to do additional experiments on any genes that show a significant difference between the normal and tumor cells. Then, because we don't want to miss genes of interest, we are willing to accept up to 25% of the genes with significant results being false positives. We'll find out they're false positives when we do the followup experiments. In this case, we would set the false discovery rate to 25%.

The Benjamini-Hochberg (BH) procedure controls the FDR... and it is simple to use!

Indeed, for a given α and a given sequence of **ordered** p-values $P_{(1)}, P_{(2)}, , P_{(m)}$, it consists in computing the m *adjusted* p-values defined as

$$P_{(i)}^{\mathrm{BH}} = \min(P_{(i)} \frac{m}{i} , P_{(i+1)}^{\mathrm{BH}})$$

```
p.bh <- data$p.value
m <- length(p.bh)
for (i in ((m-1):1))
  p.bh[i] <- min(data$p.value[i]*m/i , p.bh[i+1])
data$p.bh <- p.bh
data
```

```
##              dietary p.value      p.bh
## 1     Total calories   0.001 0.0250000
## 2          Olive oil   0.008 0.1000000
## 3         Whole milk   0.039 0.2100000
## 4         White meat   0.041 0.2100000
## 5           Proteins   0.042 0.2100000
## 6               Nuts   0.061 0.2541667
## 7  Cereals and pasta   0.074 0.2642857
## 8         White fish   0.205 0.4910714
## 9             Butter   0.212 0.4910714
## 10        Vegetables   0.216 0.4910714
## 11      Skimmed milk   0.222 0.4910714
## 12          Red meat   0.251 0.4910714
## 13             Fruit   0.269 0.4910714
## 14              Eggs   0.275 0.4910714
## 15         Blue fish   0.340 0.5328125
## 16           Legumes   0.341 0.5328125
## 17     Carbohydrates   0.384 0.5647059
## 18          Potatoes   0.569 0.7815789
## 19             Bread   0.594 0.7815789
## 20              Fats   0.696 0.8700000
## 21            Sweets   0.762 0.9071429
## 22    Dairy products   0.940 0.9860000
## 23 Semi-skimmed milk   0.942 0.9860000
## 24        Total meat   0.975 0.9860000
## 25    Processed meat   0.986 0.9860000
```

Notice that we could equivalently use the function `p.adjust`:

```
p.adjust(data$p.value, method = "BH")
```

```
##  [1] 0.0250000 0.1000000 0.2100000 0.2100000 0.2100000 0.2541667 0.2642857
##  [8] 0.4910714 0.4910714 0.4910714 0.4910714 0.4910714 0.4910714 0.4910714
## [15] 0.5328125 0.5328125 0.5647059 0.7815789 0.7815789 0.8700000 0.9071429
```

```
## [22] 0.9860000 0.9860000 0.9860000 0.9860000
```

Then, the discoveries, i.e. the significant tests, are those with an ajusted p-value less than α.

It can be shown that this procedure guarantees that for *independent tests*, and for *any alternative hypothesis*,

$$\begin{aligned} \text{FDR} &= \text{E}\left(\frac{m_{01}}{m_{01}+m_{11}}\right) \\ &\leq \frac{m_{0\cdot}}{m}\alpha \\ &\leq \alpha \end{aligned}$$

where $m_{0\cdot}$ is the (unknown) total number of true null hypotheses, and where the first inequality is an equality with continuous p-value distributions.

In our example, the first five tests would be significant with $\alpha = 0.25$, which means that we expect no more than 25% of these 5 tests to be false discoveries.

Remark 1: The BH procedure is equivalent to consider as significant the non adjusted p-values smaller than a threshold P_{BH} defined as

$$P_{\text{BH}} = \max_i \{P_{(i)} : \ P_{(i)} \leq \alpha\frac{i}{m}\}$$

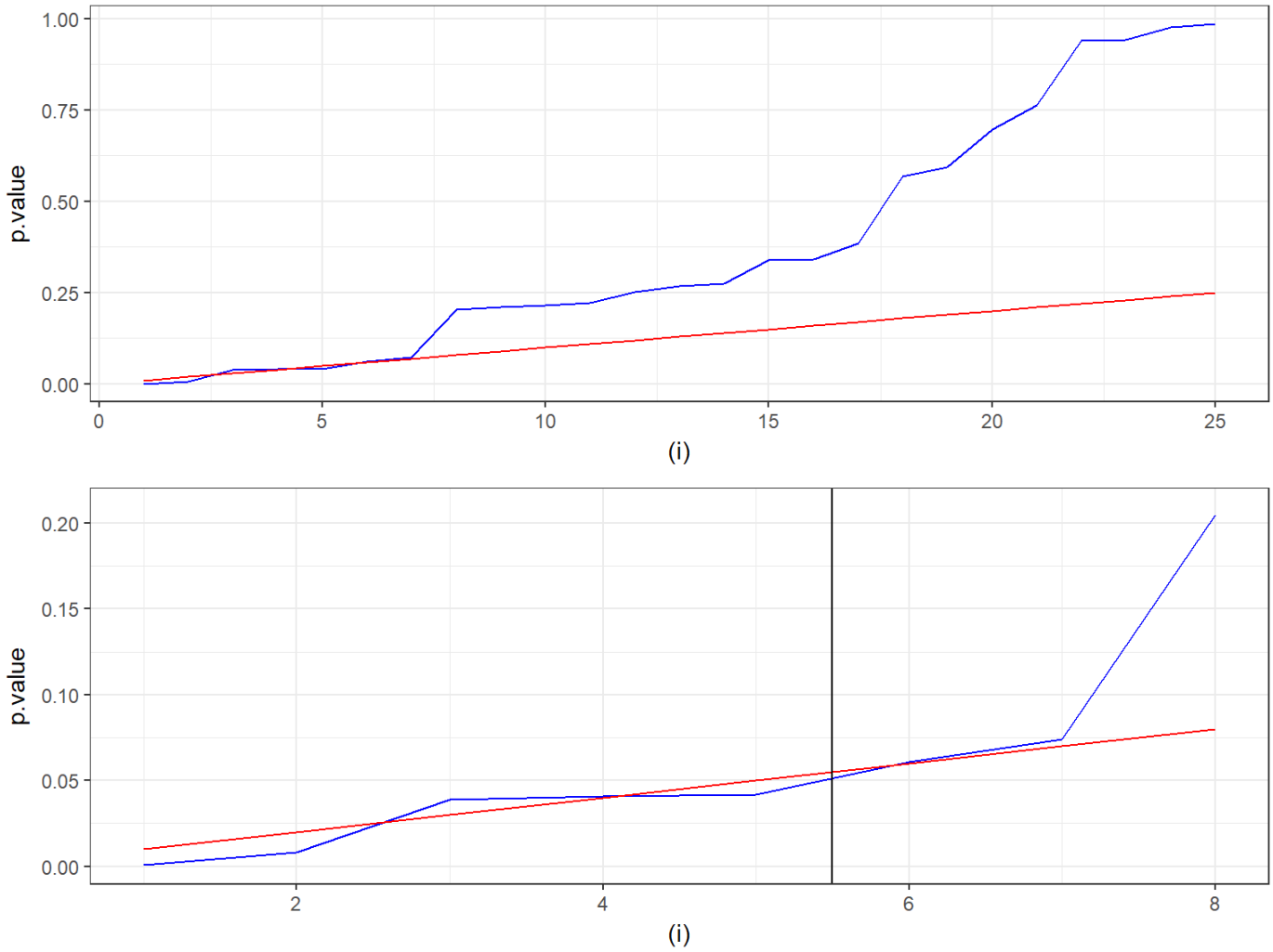
In other words, the largest p-value that has $P_{(i)} < (i/m)\alpha$ is significant, and all of the P-values smaller than it are also significant, even the ones that aren't less than their Benjamini-Hochberg critical value $\alpha \times i/m$

```
alpha <- 0.25
m <- dim(data)[1]
data$critical.value <- (1:m)/m*alpha
data
```

```
##              dietary p.value      p.bh critical.value
## 1     Total calories   0.001 0.0250000           0.01
## 2          Olive oil   0.008 0.1000000           0.02
## 3         Whole milk   0.039 0.2100000           0.03
## 4         White meat   0.041 0.2100000           0.04
## 5           Proteins   0.042 0.2100000           0.05
## 6               Nuts   0.061 0.2541667           0.06
## 7  Cereals and pasta   0.074 0.2642857           0.07
## 8         White fish   0.205 0.4910714           0.08
## 9             Butter   0.212 0.4910714           0.09
## 10        Vegetables   0.216 0.4910714           0.10
## 11      Skimmed milk   0.222 0.4910714           0.11
## 12          Red meat   0.251 0.4910714           0.12
## 13             Fruit   0.269 0.4910714           0.13
## 14              Eggs   0.275 0.4910714           0.14
## 15         Blue fish   0.340 0.5328125           0.15
## 16           Legumes   0.341 0.5328125           0.16
## 17     Carbohydrates   0.384 0.5647059           0.17
## 18          Potatoes   0.569 0.7815789           0.18
## 19             Bread   0.594 0.7815789           0.19
## 20              Fats   0.696 0.8700000           0.20
## 21            Sweets   0.762 0.9071429           0.21
## 22    Dairy products   0.940 0.9860000           0.22
## 23 Semi-skimmed milk   0.942 0.9860000           0.23
## 24        Total meat   0.975 0.9860000           0.24
```

```
## 25    Processed meat   0.986 0.9860000           0.25
```

```
library(gridExtra)
pl1 <- ggplot(data) + geom_line(aes(x=1:m,y=p.value), colour="blue") +
geom_line(aes(x=1:m,y=critical.value), colour="red") +xlab("(i)")
grid.arrange(pl1,pl1 + xlim(c(1,8)) + ylim(c(0,0.21)) + geom_vline(xintercept=5.5))
```



The largest p-value with $P_{(i)} < (i/m)\alpha$ is *proteins*, where the individual p-value (0.042) is less than the $(i/m)\alpha$ value of 0.050. Thus the first five tests would be significant.

Remark 2: The FDR is not bounded by α, but by $(m_{0.}/m)\alpha$. We could increase the global power of the tests and get a FDR equal to the desired level α, either by defining the critical values as $(i/m_{0.})\alpha$, or by multiplying the adjusted p-values by $m_{0.}/m$.

Unfortunately, $m_{0.}$ is unknown... but it can be estimated, as the number of non significant tests for instance.

```
m0.est <- sum(data$p.bh>alpha)
data$crit.valc <- round(data$critical.value*m/m0.est,4)
data$p.bhc <- round(data$p.bh*m0.est/m,4)
head(data,10)
```

```
##             dietary p.value      p.bh critical.value crit.valc  p.bhc
## 1    Total calories   0.001 0.0250000           0.01    0.0125 0.0200
## 2         Olive oil   0.008 0.1000000           0.02    0.0250 0.0800
## 3        Whole milk   0.039 0.2100000           0.03    0.0375 0.1680
```

```
## 4         White meat   0.041 0.2100000          0.04    0.0500 0.1680
## 5           Proteins   0.042 0.2100000          0.05    0.0625 0.1680
## 6               Nuts   0.061 0.2541667          0.06    0.0750 0.2033
## 7  Cereals and pasta   0.074 0.2642857          0.07    0.0875 0.2114
## 8         White fish   0.205 0.4910714          0.08    0.1000 0.3929
## 9             Butter   0.212 0.4910714          0.09    0.1125 0.3929
## 10        Vegetables   0.216 0.4910714          0.10    0.1250 0.3929
```

We would consider the 7 first p-values as significant using this new correction.

3.2 A Monte Carlo simulation

Let us perform a Monte Carlo simulation to better understand the impact of these corrections.

Let us assume that we observe $(x_{ij}, 1 \leq i \leq n_x, 1 \leq j \leq m)$ and $(y_{ij}, 1 \leq i \leq n_y, 1 \leq j \leq m)$ where ,

$$x_{ij} \underset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_{x,j}\ ,\ \sigma_x^2)$$
$$y_{ij} \underset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_{y,j}\ ,\ \sigma_y^2)$$

For $j = 1, 2, \ldots, m$, we want to test $H_{0,j} : \mu_{x,j} = \mu_{y,j}$ versus $H_{1,j} : \mu_{x,j} \neq \mu_{y,j}$.

For the simulation, we will use $m = 140$, $n_x = n_y = 50$, $\sigma_x^2 = \sigma_y^2 = 1$ and $\mu_{x,j} = 0$ for $1 \leq j \leq m$.

Furthermore, the null hypothesis is true for the first $m_{0,\cdot} = 120$ variables, assuming that $\mu_{y,j} = 0$ for $1 \leq j \leq m_{0,\cdot}$. Then, for $m_{0,\cdot} + 1 \leq j \leq m$, $\mu_{y,j}$ varies from 0.3 to 0.6, which means that the alternative hypothesis is true for these $m_{1,\cdot} = 20$ variables.

```
nx <- 50
ny <- 50
m0 <- 120
m1 <- 20
mu.x <- 0
mu.y <- c(rep(0,m0),seq(0.3,0.6,length=m1))
```

For each of the $L = 1000$ simulated replicate of the same experiment, we will randomly sample observation x and y from the model and perform a t-test for each of the m variables. We therefore get m p-values for each of these L replicates.

```
L <- 1000
m <- m0+m1
set.seed <- 12345
pval <- matrix(ncol=L,nrow=m)
for (l in (1:L))
{
  x.sim <- matrix(rnorm(nx*m, mu.x), ncol=nx)
  y.sim <- matrix(rnorm(ny*m, mu.y), ncol=ny)
  dat <- cbind(x.sim, y.sim)
  pval[,l] <- apply(dat, 1, function(dat) {
    t.test(x = dat[1:nx], y = dat[(nx + 1):(nx + ny)])$p.value})
}
```

Setting the significance level α to 0.2, we can compute for each replicate the numbers of true and false discoveries m_{11} and m_{01}, as well as the numbers of true and false nondiscoveries m_{00} and m_{10}.

We can then compute the proportion of wrongly rejected null hypotheses

```
alpha=0.2
m01 <- colSums(pval[1:m0,] < alpha)
mean(m01/m0)
```

```
## [1] 0.2009333
```

As expected, this proportion is close to α which is precisely the probability to wrongly reject the null hypothesis.

The proportion of correctly rejected null hypotheses is an estimate of the power of the test

```
m11 <- colSums(pval[(m0+1):m,] < alpha)
mean(m11/m1)
```

```
## [1] 0.80705
```

The False Discovery Rate is estimated as the proportion of false discoveries

```
mean(m01/(m01+m11))
```

```
## [1] 0.5952523
```

This means that among the significant results, about 60% of them are false discoveries.

Let us now apply the Bonferroni correction, and compute the proportion of wrongly and correctly rejected null hypotheses and the proportion of false discoveries

```
pval.b <- apply(pval,2, function(pval) {p.adjust(pval,method="bonferroni")})
m01.b <- colSums(pval.b[1:m0,] < alpha)
m11.b <- colSums(pval.b[(m0+1):m,] < alpha)
c(mean(m01.b/m0), mean(m11.b/m1),mean(m01.b/(m01.b+m11.b),na.rm=TRUE))
```

```
## [1] 0.001283333 0.180950000 0.037476158
```

Very few of the true null hypotheses are rejected, which may be a good point, but the price to pay is a very low power (less that 20%).

The familywise error rate (FWER) is the probability $\mathrm{P}(m_{01} \geq 1)$ to reject *at least* one of the true null hypothesis. This probability remains quite small when the Bonferroni correction is used.

```
mean(m01.b>=1)
```

```
## [1] 0.148
```

The Benjamini-Hochberg correction increases the power and controls the FDR as expected since the proportion of false discoveries remains below the level α.

```
pval.bh <- apply(pval,2, function(pval) {p.adjust(pval,method="BH")})
```

```
m01.bh <- colSums(pval.bh[1:m0,] < alpha)
m11.bh <- colSums(pval.bh[(m0+1):m,] < alpha)
c(mean(m01.bh/m0), mean(m11.bh/m1),mean(m01.bh/(m01.bh+m11.bh),na.rm=TRUE))
```

```
## [1] 0.01786667 0.43115000 0.17250494
```

Lastly, we can slightly improve the BH procedure by multiplying the p-values by the ratio $\hat{m}_0/m$

```
pval.bhc <- pval.bh
for (l in (1:L)) {
  m0e <- sum(pval.bh[,l]>alpha)
  pval.bhc[,l] <- pval.bh[,l]*m0e/m
}

m01.bhc <- colSums(pval.bhc[1:m0,] < alpha)
m11.bhc <- colSums(pval.bhc[(m0+1):m,] < alpha)
c(mean(m01.bhc/m0), mean(m11.bhc/m1),mean(m01.bhc/(m01.bhc+m11.bhc),na.rm=TRUE))
```

```
## [1] 0.02068333 0.44910000 0.18481846
```

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57: 289-300.

García-Arenzana, N., E.M. Navarrete-Muñoz, V. Lope, P. Moreo, S. Laso-Pablos, N. Ascunce, F. Casanova-Gómez, C. Sánchez-Contador, C. Santamariña, N. Aragonés, B.P. Gómez, J. Vioque, and M. Pollán. 2014. Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925.

Statistical tests: exercices

1 Gene expression

The dataset geHT.csv consists of gene expression measurements for ten genes under control and treatment conditions, with four replicates each.

1. Test the hypothesis that the mean of the control expression values is 2000.
2. Test that there is no difference overall between the treatments and controls for any of the genes (test that the whole experiment didn't work or there are no differentially expressed genes)
3. Test if the variances for the gene expression are the same under treatment or control conditions

2 Smoking, no smoking

1. There are 88 smokers among a group of 300 people of a same population. Test that the proportion of smokers in this population is less than or equal to 0.25, greater than or equal to 0.25, equal to 0.25. Show that we can use an exact test, or a test relying on an approximation.
2. There are 90 smokers in another group of 400 people, coming from another population. Can we conclude that the proportion of smokers are different in these two populations?

3 Alzheimer's disease

Dementia is the result of various cerebral disorders, leading to an acquired loss of memory and impaired cognitive ability. The most common forms are Alzheimer's disease and vascular dementia.

In a study, patients were treated either with Cerebrolysin or Donepezil. The datafile `scoreAD.csv` reports the difference of a score obtained by these patients before and after treatment (a negative score indicates an improvment).

1. Test if Cerebrolysin and Donepezil have a beneficial effect on patients.
2. A doctor claims that the score decrease is greater than 2 in average for patients who take Cerebrolysin. What do you think of this hypothesis? What should be the null hypothesis?
3. Test if the two drugs can be considered a equivalent, considering that the two drugs are equivalent if the

difference between the effects is *i)* less than 2 in average, *ii)* less than 4.

4 Epileptic activity

It is frequently assumed that the daily numbers of epilepy seizures are independent Poisson random variables.

1. The daily numbers of epilepy seizures of a given patient are reported in the datafile `epilepsy1.csv`.Use the Poisson dispersion test and the Chi-Square goodness-of-fit test to test if this data follows a Poisson distribution.
2. Compare the Type I error rate and the power of these two tests via Monte Carlo simulation.

5 Type 2 diabetes

An investigator is exploring whether the expression levels of genes significantly differ between a sample of healthy individuals and a sample of individuals with Type 2 diabetes. He performs a separate t-test comparing the two samples for 5,000 different genes, and uses $\alpha = 0.05$ as his cutoff. His analysis identifes 411 genes as having different expression levels between the two samples.

1. The investigator reasons that because he carried out his t-tests using a type I error rate of 5%, he should expect about 5% of the 411 genes that he discovered to be type I errors. Is this reasoning correct or incorrect? If it is incorrect, what's wrong with it?
2. What is the investigator's false discovery rate?

6 Identification of genes

Breast cancer is the most common malignant disease in Western women. In these patients, it is not the primary tumour, but its metastases at distant sites that are the main cause of death.

Prognostic markers are needed to identify patients who are at the highest risk for developing metastases, which might enable oncologists to begin tailoring treatment strategies to individual patients. Gene-expression signatures of primary breast tumours might be one way to identify the patients who are most likely to develop metastatic cancer.

The datafile `geneMFS.csv` contains the expression level of 11 genes and the metastasis-free survival (the period until metastasis is detected) for 527 patients.

The objective of this study is to identify which genes may be good or poor prognosis for the development of matastasis.

1. Graphically compare the distribution of the gene expressions in the groups of patients with early metastasis (MFS <1000) and late metastasis (MFS>1000).
2. Compare the gene expression levels in these two groups using a parametric test.
3. Compare these results with those obtained using a non parametric test.

Regression models: exercices

1 Polynomial regression

The file `ratWeight.csv` consists of rat weights measured over 14 weeks during a subchronic toxicity study related to the question of genetically modified (GM) corn.

We will only consider the weight of rat `B38625`.

Based on this data, our objective is to build a regression model of the form

$$y_j = f(x_j) + e_j \quad ; \quad 1 \leq j \leq n$$

We will restrict ourselves to polynomial regression, by considering functions of the form

$$\begin{aligned} f(x) &= f(x; c_0, c_1, c_2, \ldots, c_d) \\ &= c_0 + c_1 x + c_2 x^2 + \ldots + c_d x^d \end{aligned}$$

Fit the ``best'' polynomial to this data

2 Nonlinear regression

1. Load the `ratWeight.csv` datafile and plot the weight of the females of the control group
2. Select the ID `B38837` and fit a polynomial model to the growth curve of this female rat.
3. Fit a Gompertz model $f_1(t) = Ae^{-be^{-kt}}$ to this data.
4. Fit the two following growth models:

Asymptotic regression model:

$$f_2(t) = A(1 - be^{-kt})$$

Logistic curve:

$$f_3(t) = \frac{A}{1 + e^{-\gamma(t-\tau)}}$$

5. Propose two other parametrizations of the asymptotic regression model which involves

i. the weight at birth w_0 (when $t = 0$), the limit weight w_∞ (when $t \to \infty$) and k
ii. the weight at birth, the weight at the end of the study w_{14} and the ratio $r = (w_{14} - W_7)/(w_7 - w_0)$

Can we compare these models?

6. We will now use model f_{2a}. Check that the estimate of $\beta = (w_0, w_\infty, k)$ obtained with the `nls` function is the least squares estimate.
7. Check that this estimate is also the least squares estimate of the linearized model. Then, how are computed the standard errors of $\hat{\beta}$?
8. Compute 90% confidence intervals for the model parameters using several approaches (linearization, parametric bootstrap, profile likelihood)
9. Compute a 90% confidence interval for the predicted weight and a 90% prediction interval for the measured weight using the delta method.
10. Compare the predicted weight of this rat with the predicted weight for ID `B38837`.

Mixed effects models: exercices

1 Linear mixed effects models

1.1 The Pastes data

1. Check the documentation, the structure and a summary of the Pastes data from the lme4 package.
2. Build a model for this data
3. What is the main cause of the variability of the paste strength?

1.2 The orange data

1. Check the documentation, the structure and a summary of the Orange data.
2. Plot this data
3. Fit a linear model to this data
4. Fit different linear mixed effect models to this data and compare them
5. Compute confidence intervals on the parameters, compute the individual parameters and plot the random effects for the ``best?????? model
6. Compare the ML and the REML estimates for this model

1.3 The Oats data

1. Check the documentation, the structure and a summary of the Oats data from the nlme package.
2. Plot the data in such a way as to visualize the effect of the fertilizer-concentration on the yield as well as possible differences between blocks and varieties.
3. Fit the Oats data with linear mixed effects model, assuming a linear effect of the fertilizer concentration on the yield.

2 Nonlinear mixed effects models

S??ralini *et al.* published in 2007 the paper ???New Analysis of a Rat Feeding Study with a Genetically Modified Maize Reveals Signs of Hepatorenal Toxicity???. The authors of the paper pretend that, after the consumption of MON863, rats showed slight but dose-related significant variations in growth.

The objective of this exercise is to highlight the flaws in the methodology used to achieve this result, and show how to properly analyse the data.

We will restrict our study to the male rats of the study fed with 11% of maize (control and GMO)

1. Load the `ratWeight.csv` data, select the male rats fed with 11% of maize and plot the growth curves of the control and GMO groups.
2. Fit a Gompertz growth model $f_1(t) = A\exp(-\exp(-b(t-c)))$ to the complete data (males fed with 11% of maize) using a least square approach, with the same parameters for the control and GMO groups.
3. Fit a Gompertz growth model to the complete data (11% male) using a least square approach, with different parameters for the control and GMO groups.

Hint: write the model as

$$y_{ij} = A_0 e^{-e^{-b_0(t_{ij}-c_0)}} \mathbb{1}_{\text{regime}_i=\text{Control}} + A_1 e^{-e^{-b_1(t_{ij}-c_1)}} \mathbb{1}_{\text{regime}_i=\text{GMO}} + e_{ij}$$

4. Check out the results of the paper displayed Table 1, for the 11% males.
5. Plot the residuals and explain why the results of the paper are wrong.
6. We popose to use instead a mixed effects model for testing the effect of the regime on the growth of the 11% male rats. The codes below show how to fit a Gompertz model to the data
   - assuming the same population parameters for the two regime groups,
   - using lognormal distributions for the 3 parameters (setting transform.par=c(1,1,1))
   - assuming a diagonal covariance matrix Ω (default)

Create first the saemixData object

```
library(saemix)

data <- read.csv("ratWeight.csv")
data.male11 <- subset(data, gender=="Male" & dosage=="11%")

saemix.data<-saemixData(name.data=data.male11,
                        name.group=c("id"),
                        name.predictors=c("week"),
                        name.response=c("weight"))
```

Implement then the structural model and create the saemixModel object. Initial values for the population parameters should be provided.

```
gompertz.model <- function(psi,id,x) {
  t <- x[,1]
  A<-psi[id,1]
  b<-psi[id,2]
  c<-psi[id,3]
  ypred<- A*exp(-exp(-b*(t-c)))
```

```
  return(ypred)
}

saemix.gompertz.model0<-saemixModel(model=gompertz.model,
                          psi0=c(A=500,b=0.2,c=0.2),
                          transform.par=c(1,1,0))
```

Run saemix for estimating the population parameters, computing the individual estimates, computing the FIM and the log-likelihood (linearization)

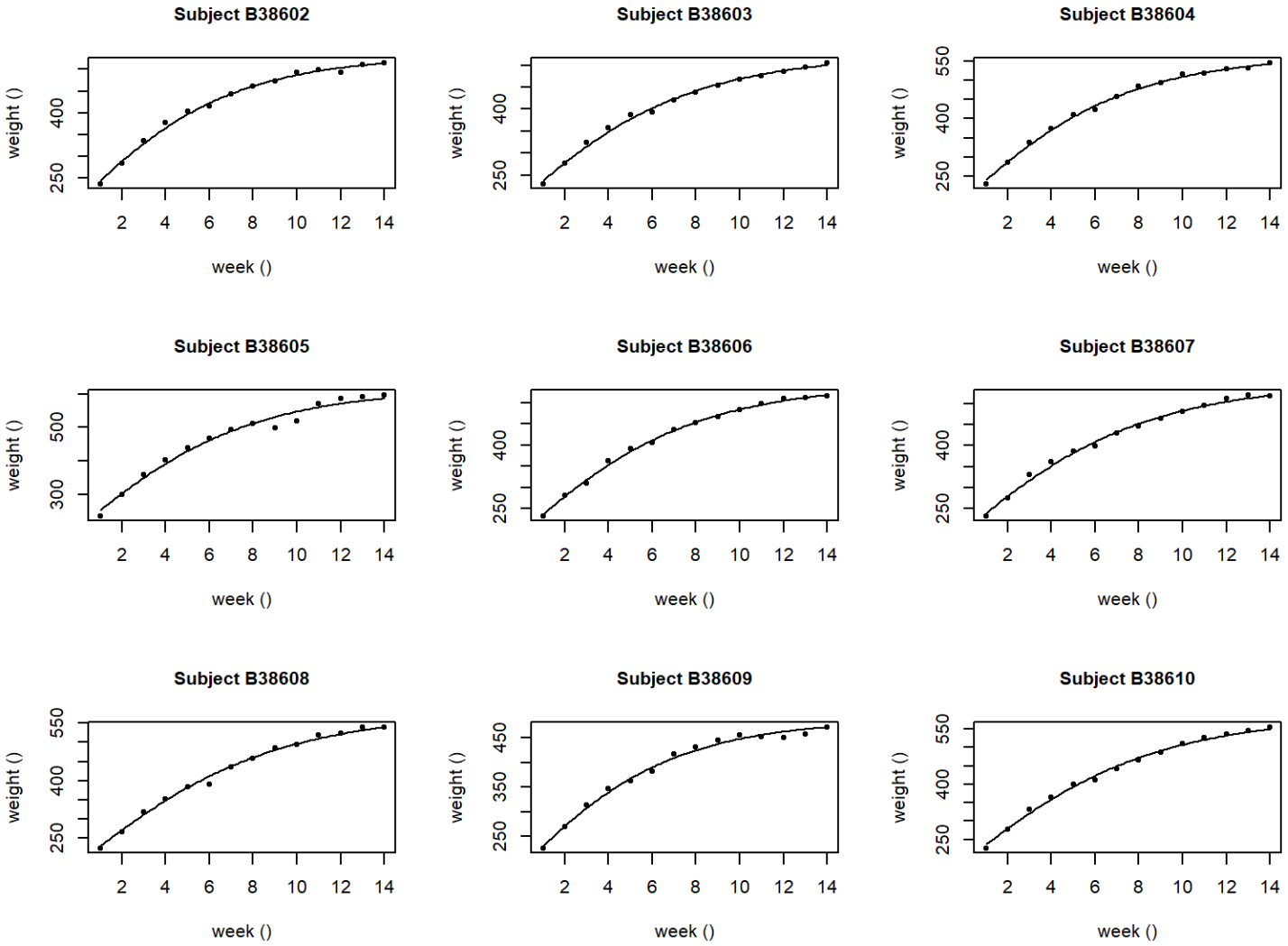
```
saemix.options<-list(seed=632545, displayProgress=FALSE, algorithms=c(1,1,0))
saemix.gompertz.fit0 <- saemix(saemix.gompertz.model0,saemix.data,saemix.options)
```

```
summary(saemix.gompertz.fit0)
```

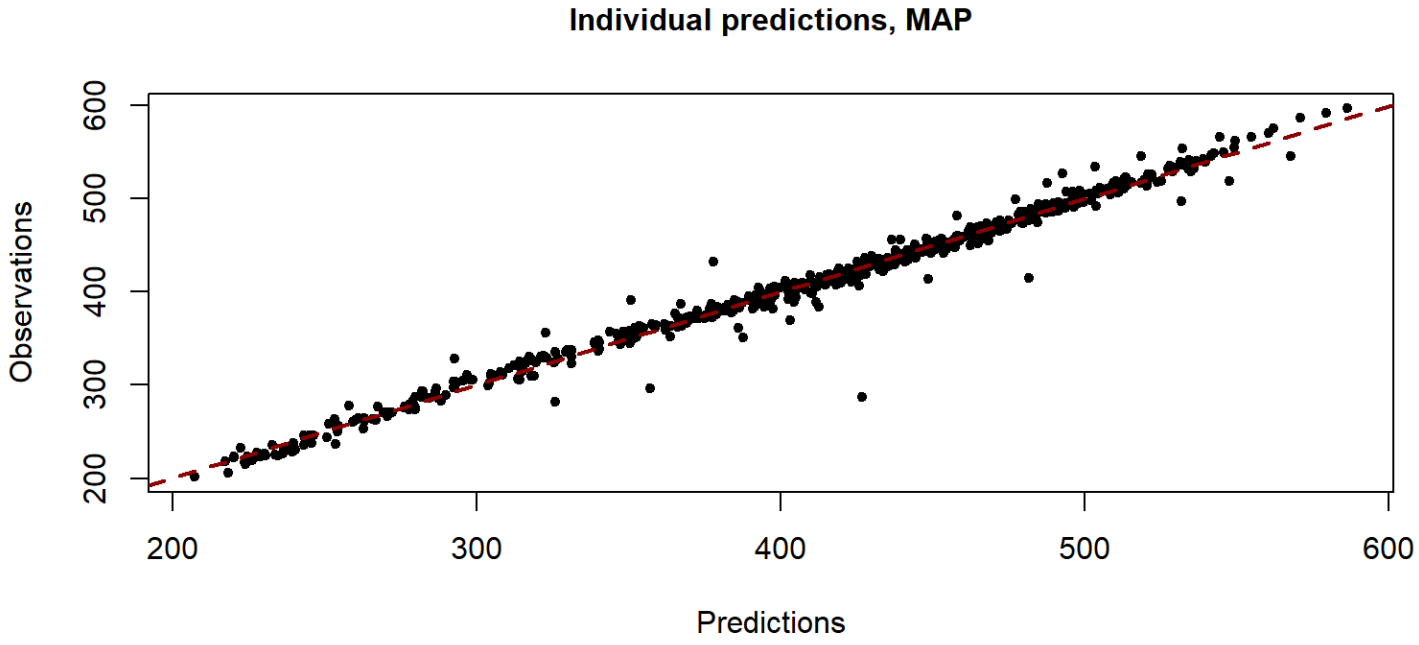
```
## ----------------------------------------------------
## -----------------  Fixed effects  ------------------
## ----------------------------------------------------
##   Parameter Estimate     SE  CV(%)
## 1         A  529.069 7.9284   1.50
## 2         b    0.214 0.0065   3.06
## 3         c    0.041 0.0820 199.02
## 4         a   12.297 0.4101   3.33
## ----------------------------------------------------
## -----------  Variance of random effects  -----------
## ----------------------------------------------------
##   Parameter Estimate     SE CV(%)
## A  omega2.A   0.0081 0.0019 24.04
## b  omega2.b   0.0221 0.0080 36.01
## c  omega2.c   0.2068 0.0588 28.45
## ----------------------------------------------------
## ------  Correlation matrix of random effects  ------
## ----------------------------------------------------
##          omega2.A omega2.b omega2.c
## omega2.A 1.00     0.00     0.00
## omega2.b 0.00     1.00     0.00
## omega2.c 0.00     0.00     1.00
## ----------------------------------------------------
## ---------------  Statistical criteria  -------------
## ----------------------------------------------------
## Likelihood computed by linearisation
##       -2LL= 4703.692
##       AIC = 4717.692
##       BIC = 4729.514
## ----------------------------------------------------
```

Display some diagnostic plots:

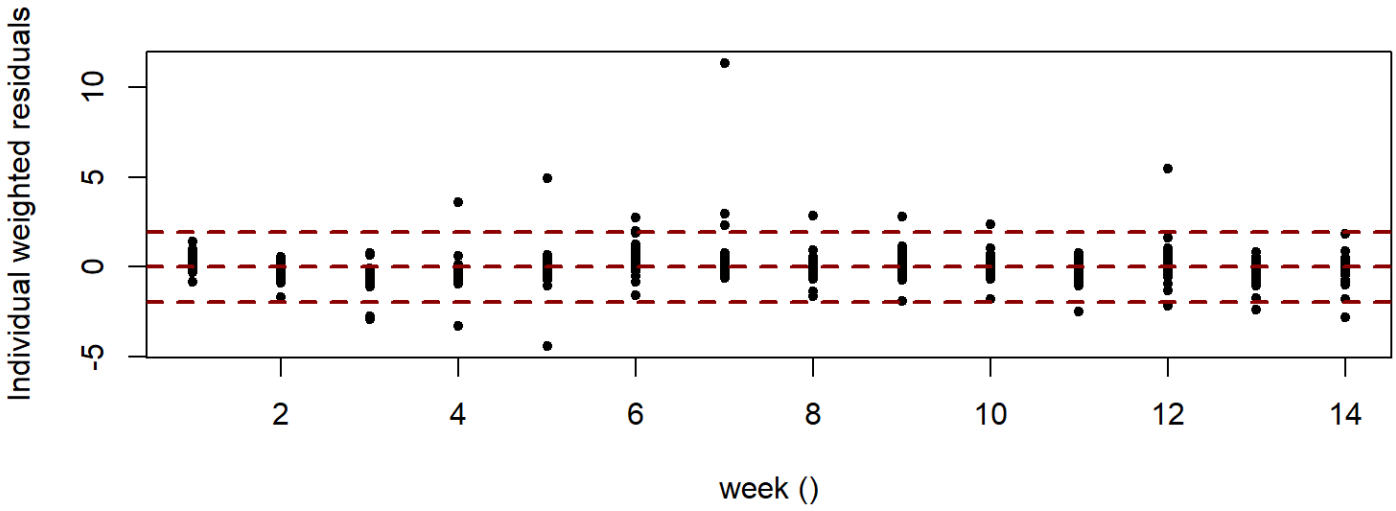
```
#Individual predictions
saemix.gompertz.fit0 <- predict(saemix.gompertz.fit0)

# Individual plot for subject 1 to 9,
saemix.plot.fits(saemix.gompertz.fit0,ilist=c(1:9),smooth=TRUE)
```

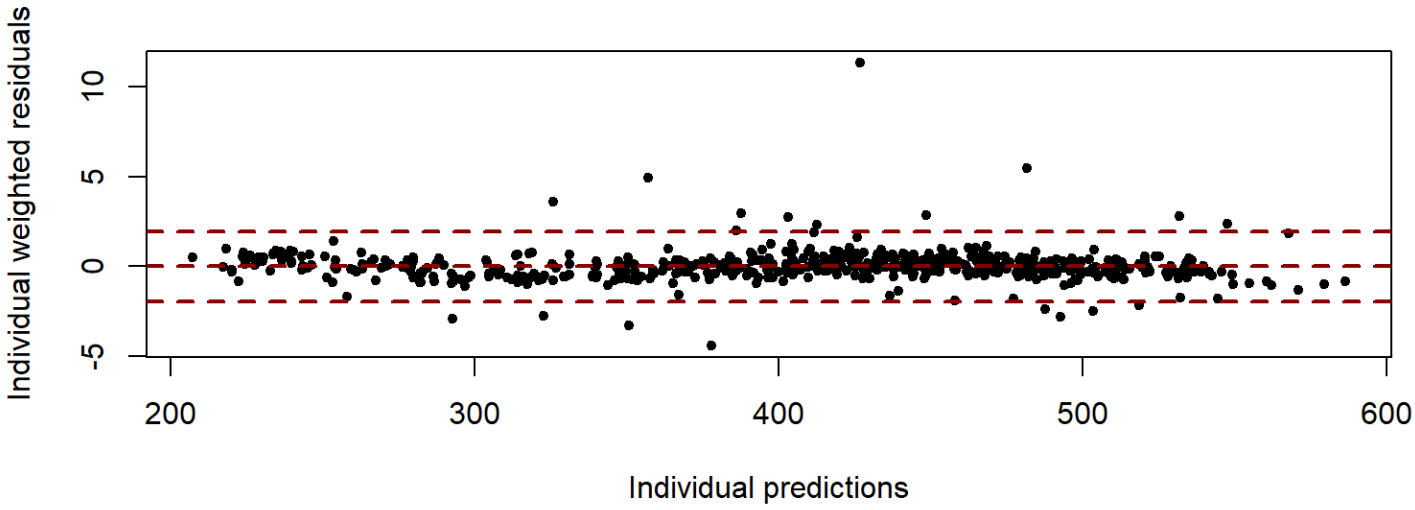



```
# Diagnostic plot: observations versus population predictions
saemix.plot.obsvspred(saemix.gompertz.fit0,level=1)
```




```
# Scatter plot of residuals
saemix.plot.scatterresiduals(saemix.gompertz.fit0, level=1)
```





Correlation matrix of the estimates:

```
fim <- -saemix.gompertz.fit0@results@fim #Fisher information matrix
cov.est <- solve(fim) # covariance matrix of the estimates
d <- sqrt(diag(cov.est)) # s.e. of the estimates
cov.est/(d%*%t(d)) # correlation matrix of the estimates
```

```
##            [,1]       [,2]       [,3]         [,4]        [,5]
## [1,]  1.0000000 -0.1846297 -0.0207690  0.000000000  0.00000000
## [2,] -0.1846297  1.0000000  0.1447597  0.000000000  0.00000000
## [3,] -0.0207690  0.1447597  1.0000000  0.000000000  0.00000000
## [4,]  0.0000000  0.0000000  0.0000000  1.000000000 -0.03431771
## [5,]  0.0000000  0.0000000  0.0000000 -0.034317707  1.00000000
## [6,]  0.0000000  0.0000000  0.0000000  0.000503899 -0.02306556
## [7,]  0.0000000  0.0000000  0.0000000 -0.010303060 -0.09232785
##              [,6]        [,7]
## [1,]  0.000000000  0.00000000
```

```
## [2,]  0.000000000  0.00000000
## [3,]  0.000000000  0.00000000
## [4,]  0.000503899 -0.01030306
## [5,] -0.023065562 -0.09232785
## [6,]  1.000000000 -0.05432513
## [7,] -0.054325132  1.00000000
```

Fit the same model to the same data, assuming different population parameters for the control and GMO groups. Can we conclude that the regime has an effect on the growth of the 11% male rats?

7. Use an asymptotic regression model $f(t) = w_\infty + (w_0 - w_\infty)e^{-kt}$ to test the effect of the regime on the growth of the 11% male rats.
8. Should we accept the hypothesis that the random effects are uncorrelated?
9. In conclusion, what is the ``best?????? model to fit this data?