

Analyse de la variance (anova)

quantitatif ~ f(qualitatif)

Benjamin Louis

16/10/2019 (MàJ: 18/10/2019)

Définition et objectif

Modèle général :

$$y_{ij...nr} = \mu + \alpha_i + \beta_j + \dots + \gamma_n + \dots + \alpha\beta_{ij} + \dots + \beta\gamma_{jn} + \epsilon_{ij...nr}$$

$$\epsilon_{ij...nr} \rightarrow N(0, \sigma^2) \quad \text{cov}(\epsilon_{ij...nr}, \epsilon_{i'j'...n'r'}) = 0$$

- μ : intercept
- α, β, γ : variables qualitatives
- α_i : valeur pour la modalité i de la variable associée au paramètre α
- $\alpha\beta$: interaction entre α et β
- ϵ : erreur résiduelle

Objectif principal :

Tester et étudier l'influence de variables qualitatives sur une variable numérique.

Hypothèses :

- Sur les résidus
- Sur les coefficients : $\alpha_{ref} = 0$ ou $\sum_i \alpha_i = 0$

Exemple

- Expérimentation en plein champ avec I = 3 variétés différentes de blé et J = 2 types de fongicides
- 54 parcelles : 9 répétitions pour chaque pair de variété et fongicide
- Question : existe t-il des différences de rendement entre les variétés de blé et selon le fongicide utilisé ?
- Model :

$$rendement_{ijr} = \mu + \alpha_{variete=i} + \beta_{fongicide=j} + \alpha\beta_{variete=i,fongicide=j} + \epsilon_{ijr}$$

Estimation par moindres carrés (modèle à 1 facteur !)

```
# Par défaut dans R  
#  
#
```

```
options(contrasts=c("contr.sum","contr.sum"))  
# ou  
FactoMineR::AovSum()
```

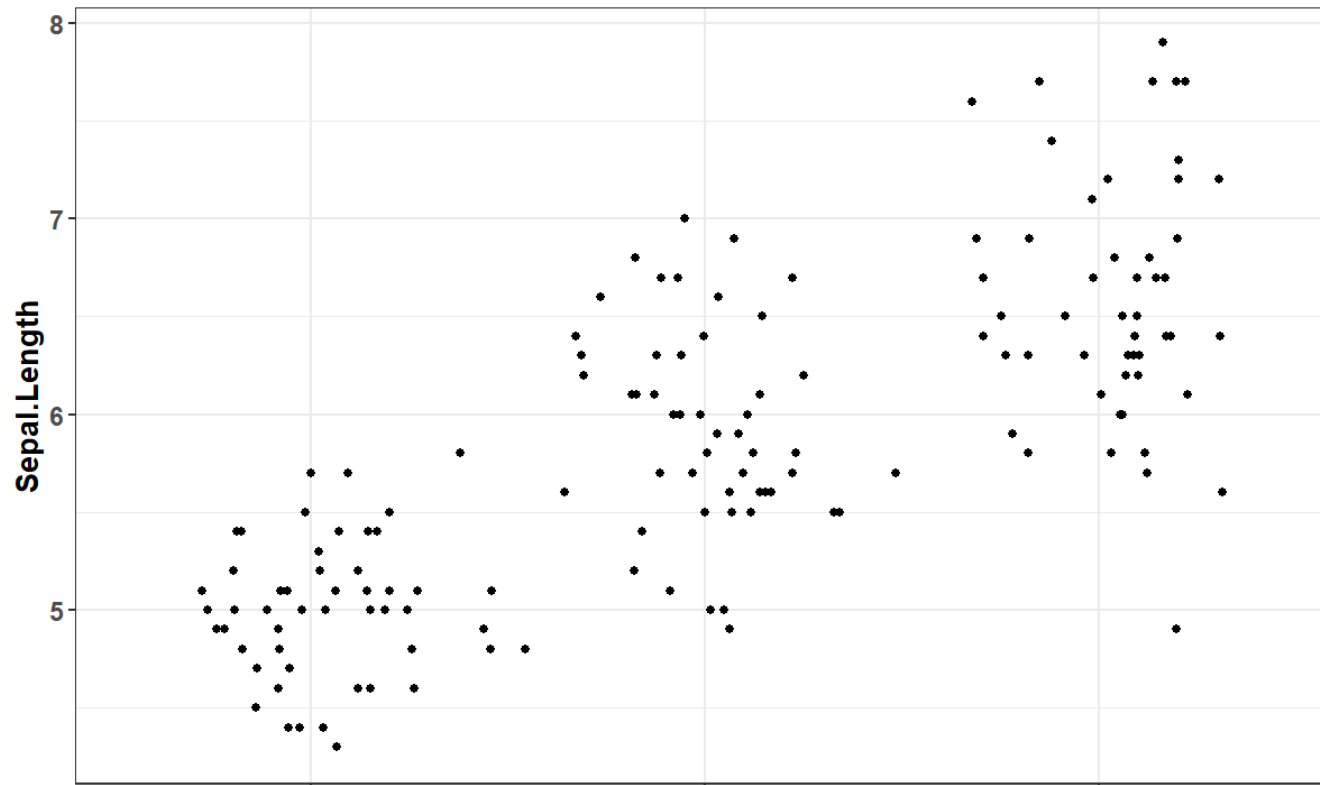
- $\hat{\mu} = \bar{y}_{i=ref}$
- $\hat{\alpha}_{i=ref} = 0$
 $\hat{\alpha}_{i \neq ref} = \bar{y}_{i \neq ref} - \mu$

- $\hat{\mu} = \bar{y}$
- $\hat{\alpha}_i = \bar{y}_i - \mu$

-
- $\hat{y}_{ik} = \hat{\mu} + \hat{\alpha}_i$
 - $e_{ik} = y_{ik} - \hat{y}_{ik}$
 - $\hat{\sigma}^2 = \frac{\sum_{ik} \epsilon_{ik}^2}{n-I}$
 - $\hat{\sigma}_{\hat{\alpha}_i}^2 = \frac{I-1}{n} \hat{\sigma}^2$

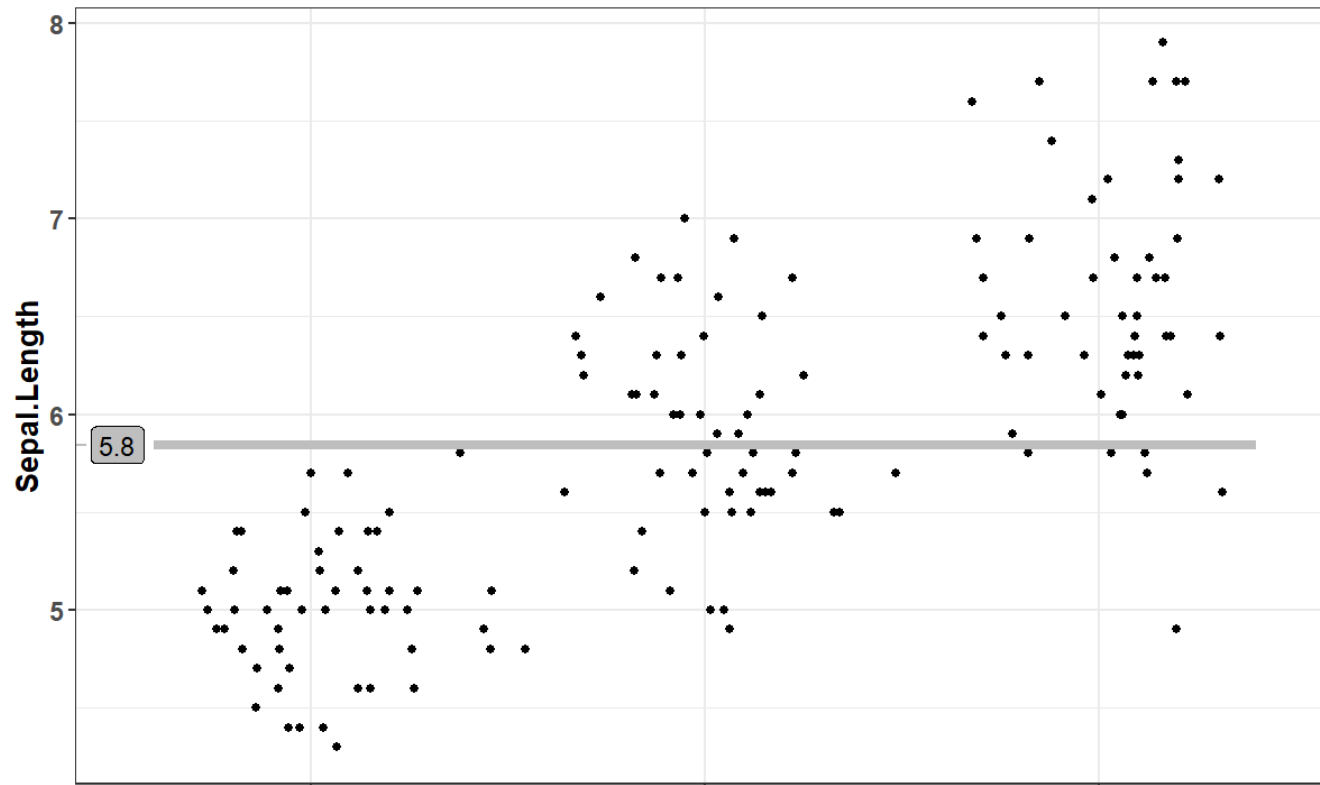
Illustration

Sepal.Length_{ik}



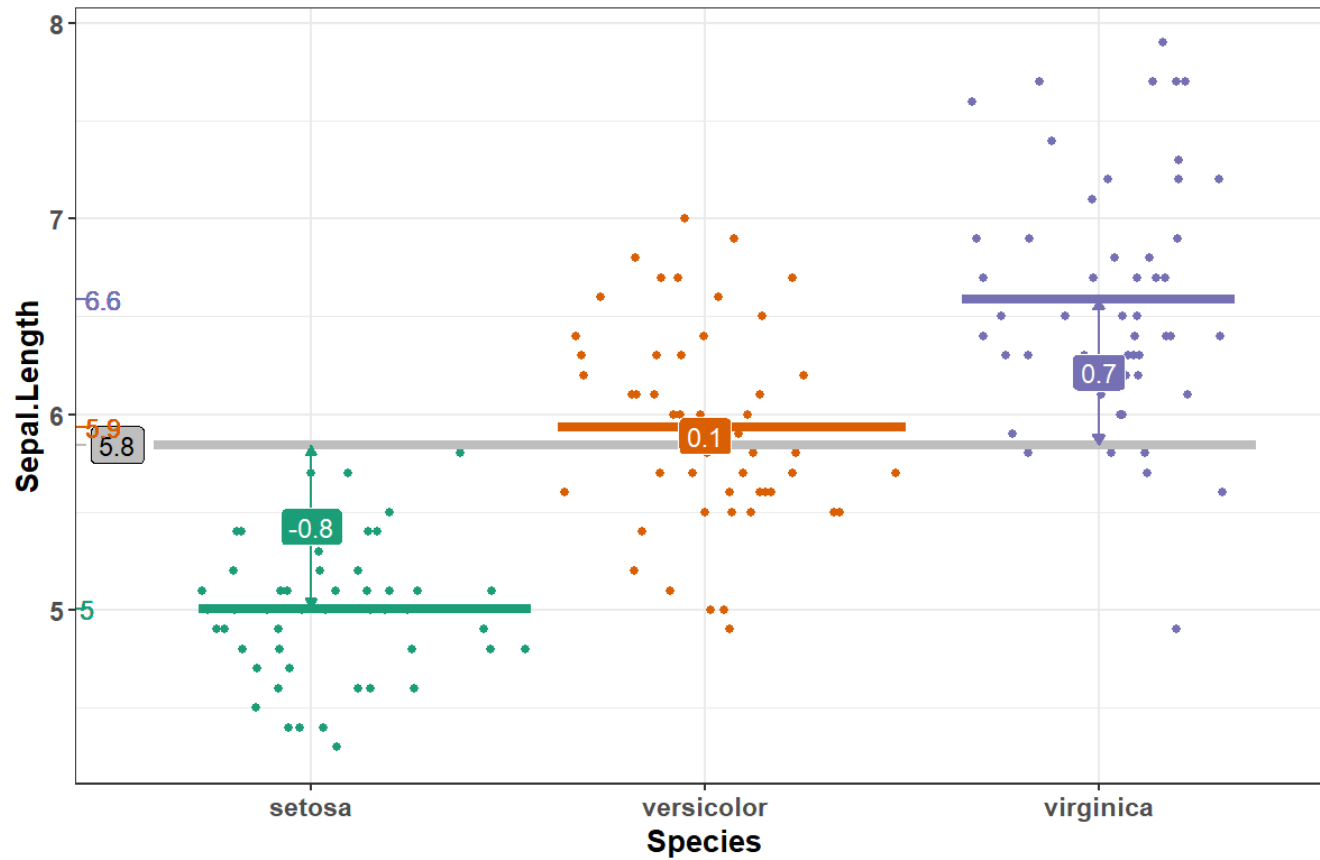
Illustration

$$\text{Sepal.Length}_{ik} = \mu$$



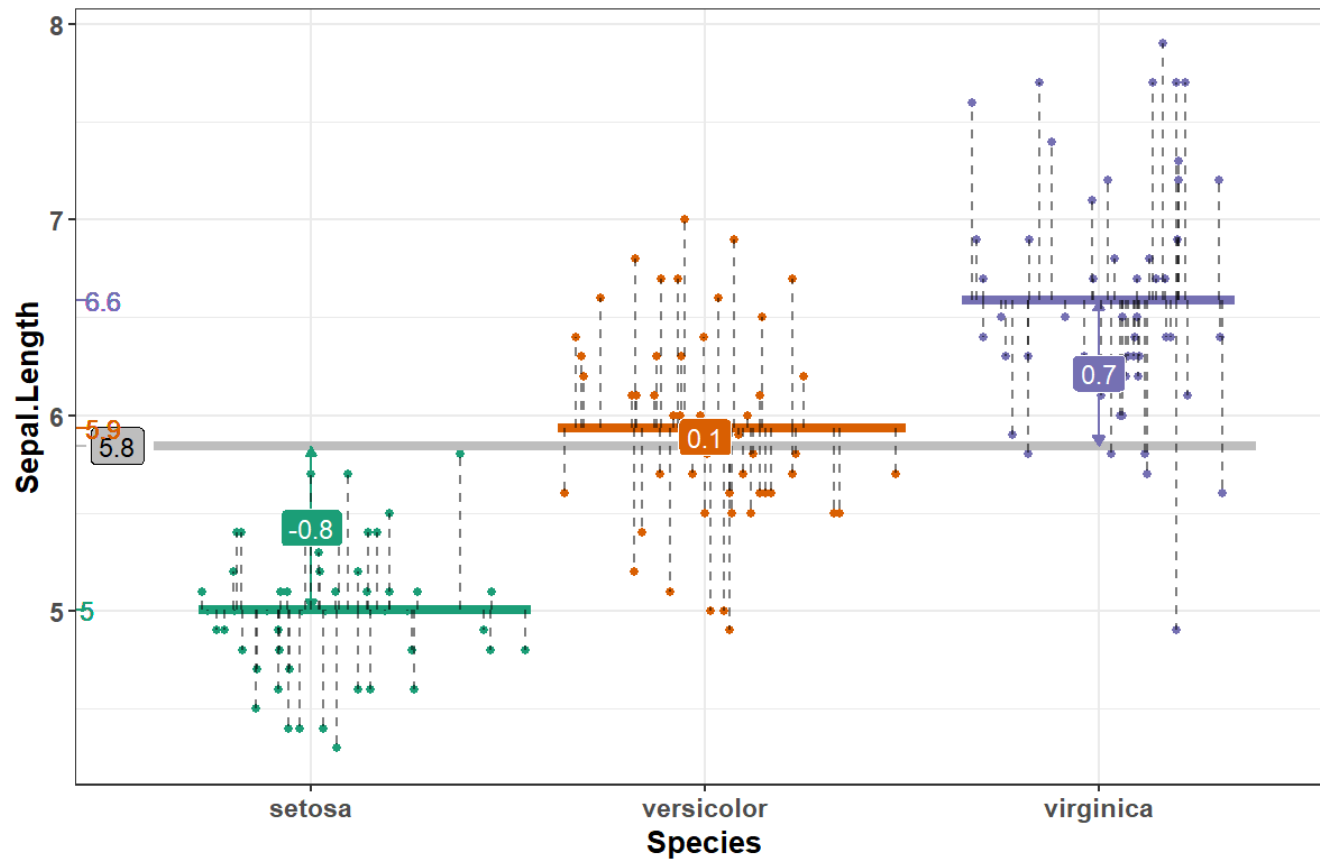
Illustration

$$\text{Sepal.Length}_{ik} = \mu + \alpha_i, \quad i = \{\text{setosa} \mid \text{versicolor} \mid \text{virginica}\}$$



Illustration

$$\text{Sepal.Length}_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \quad i = \{\text{setosa} \mid \text{versicolor} \mid \text{virginica}\}$$



Test de l'effet d'une variable

➡ La variabilité de y entre les modalités de la variable est elle *significativement* plus grande que la variabilité résiduelle ?

⚠ Dans le cas d'un plan équilibré ⚠ :

$$\sum_{ij} (y_{ij} - \bar{y})^2 = \frac{n}{I} \sum_i (\hat{\alpha}_i)^2 + \sum_{ij} (e_{ij})^2$$

| | | | | |
|--------------------|---|---|---|---|
| Variabilité totale | = | Variabilité interclasse (<i>between group</i>) | + | Variabilité interclasse (<i>whithin group</i>) |
| SS_T | = | SS_F | + | SS_R |

Test : à quel point SS_F est plus grande que SS_R ? (Plus il y a de groupes/observations, plus les valeurs risquent d'être élevées !)

➡ Il faut normaliser avec les *degrés de liberté* (*df*)!

$$CM_F = \frac{SS_F}{I - 1}$$

$$CM_R = \frac{SS_R}{n - (I - 1) - 1}$$

Dans R

```
options(contrasts=c("contr.sum","contr.sum"))
mod = lm(Sepal.Length ~ Species, data = iris)
anova(mod)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: Sepal.Length
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## Species      2  63.212   31.606   119.26 < 2.2e-16 ***
```

```
## Residuals 147  38.956    0.265
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test de l'effet d'une variable

Hypothèse :

$H_0 : \alpha_i = 0, \forall i$, tous les coefficients sont égaux, donc nuls

$H_1 : \exists \alpha_i \neq 0$, un coefficient au moins n'est pas nul

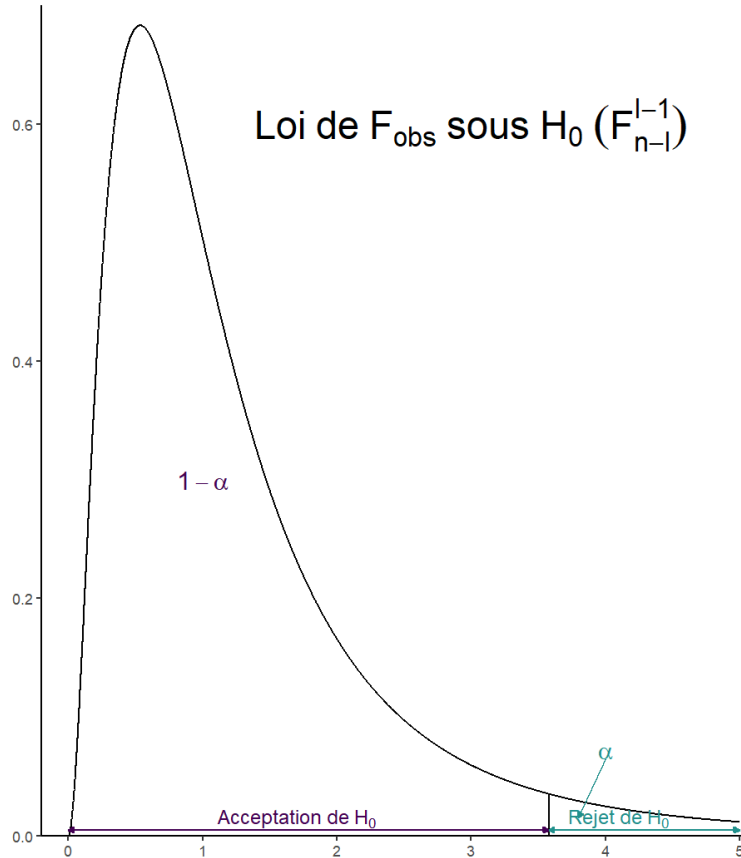
Statistique de test :

$$F_{obs} = \frac{CM_F}{CM_R}$$

Sous H_0 :

$$F_{obs} \rightarrow F_{n-I}^{I-1}$$

Test de l'effet d'une variable



Règle de décision

$$F_{\text{obs}} < F_{n-I}^{I-1}(1 - \alpha) \Rightarrow \text{acceptation de } H_0$$

$$F_{\text{obs}} \geq F_{n-I}^{I-1}(1 - \alpha) \Rightarrow \text{rejet de } H_0$$

Dans R

```
options(contrasts=c("contr.sum","contr.sum"))  
mod = lm(Sepal.Length ~ Species, data = iris)  
anova(mod)
```

```
## Analysis of Variance Table  
##  
## Response: Sepal.Length  
##           Df Sum Sq Mean Sq F value    Pr(>F)  
## Species      2  63.212   31.606  119.26 < 2.2e-16 ***  
## Residuals 147  38.956    0.265  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pf(119.26, 2, 147, lower.tail = FALSE)
```

Exemples avec plus de facteurs

```
library(agricolae)
data(greenhouse)
data <- greenhouse$greenhouse1
mod <- lm(weight ~ variety*method, data = data)
anova(mod)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: weight
```

```
##           Df Sum Sq Mean Sq  F value Pr(>F)
## variety      2   2297    1148    0.7279 0.4835
## method       3 483322   161107 102.1204 <2e-16 ***
## variety:method 6 150619    25103  15.9120 <2e-16 ***
## Residuals   468 738326    1578
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

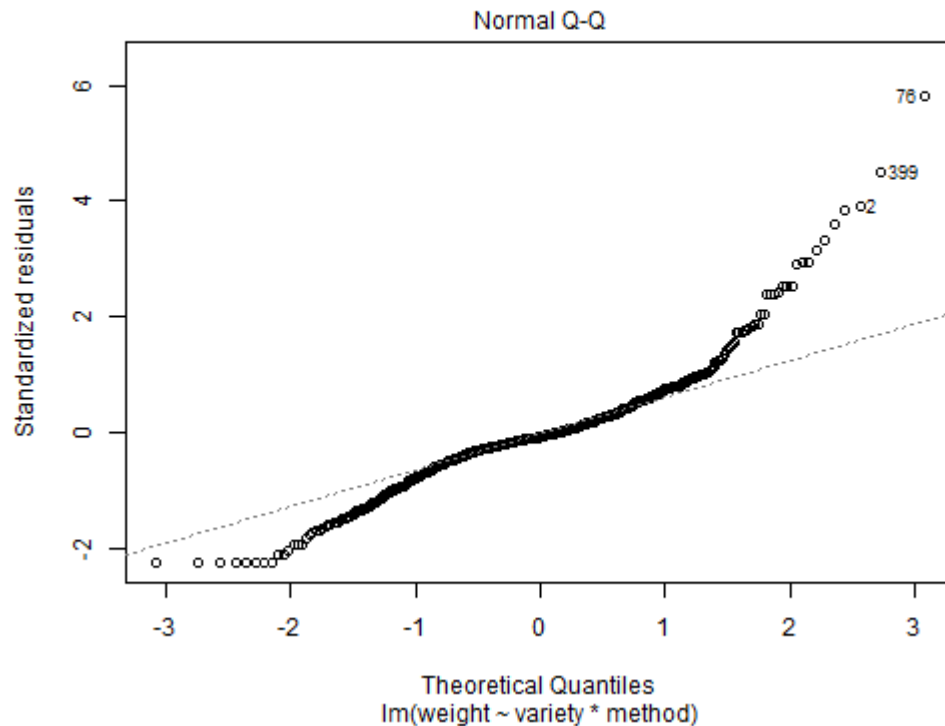
Énormément d'information donné par le `summary()` !!!

```
summary(mod)

##
## Call:
## lm(formula = weight ~ variety * method, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -88.547 -17.871  -3.492  15.626 227.944
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      72.664      1.813   40.081 < 2e-16 ***
## variety1         1.056      2.564    0.412  0.68067
## variety2         1.990      2.564    0.776  0.43799
## method1          2.022      3.140    0.644  0.51996
## method2         11.814      3.140    3.762  0.00019 ***
## method3        -50.440      3.140  -16.063 < 2e-16 ***
## variety1:method1 -35.589      4.441   -8.014 8.92e-15 ***
## variety2:method1  18.681      4.441    4.207 3.11e-05 ***
## variety1:method2   2.014      4.441    0.454  0.65035
## variety2:method2  10.335      4.441    2.327  0.02038 *
## variety1:method3   2.855      4.441    0.643  0.52054
## variety2:method3  -9.963      4.441   -2.243  0.02533 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.72 on 468 degrees of freedom
## Multiple R-squared:  0.4629,    Adjusted R-squared:  0.4502
## F-statistic: 36.66 on 11 and 468 DF,  p-value: < 2.2e-16
```

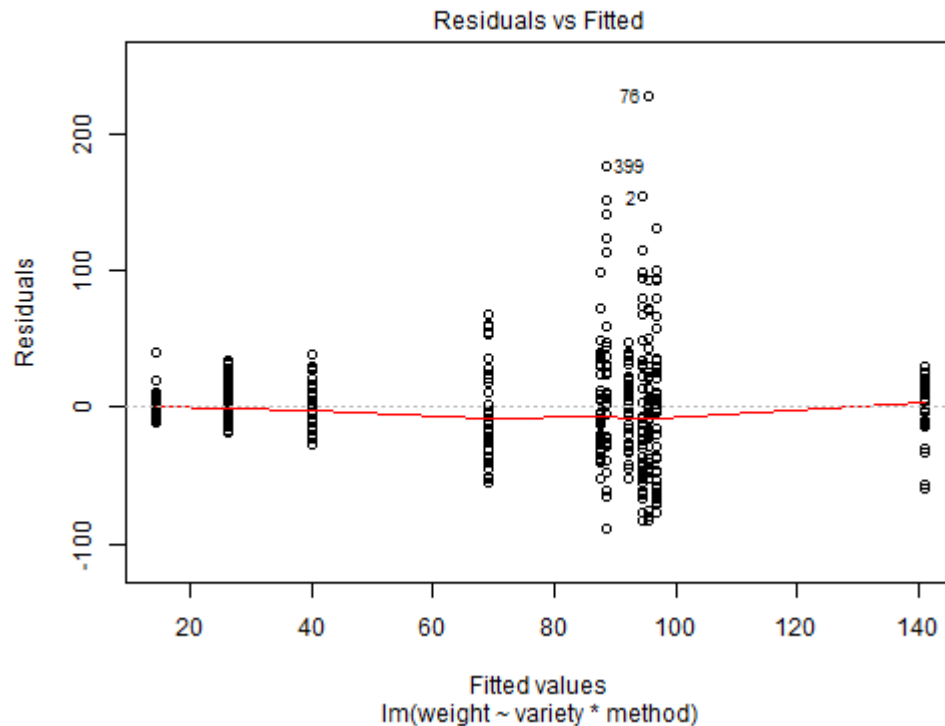
Vérification hypothèses

```
plot(mod, which = 2)
```



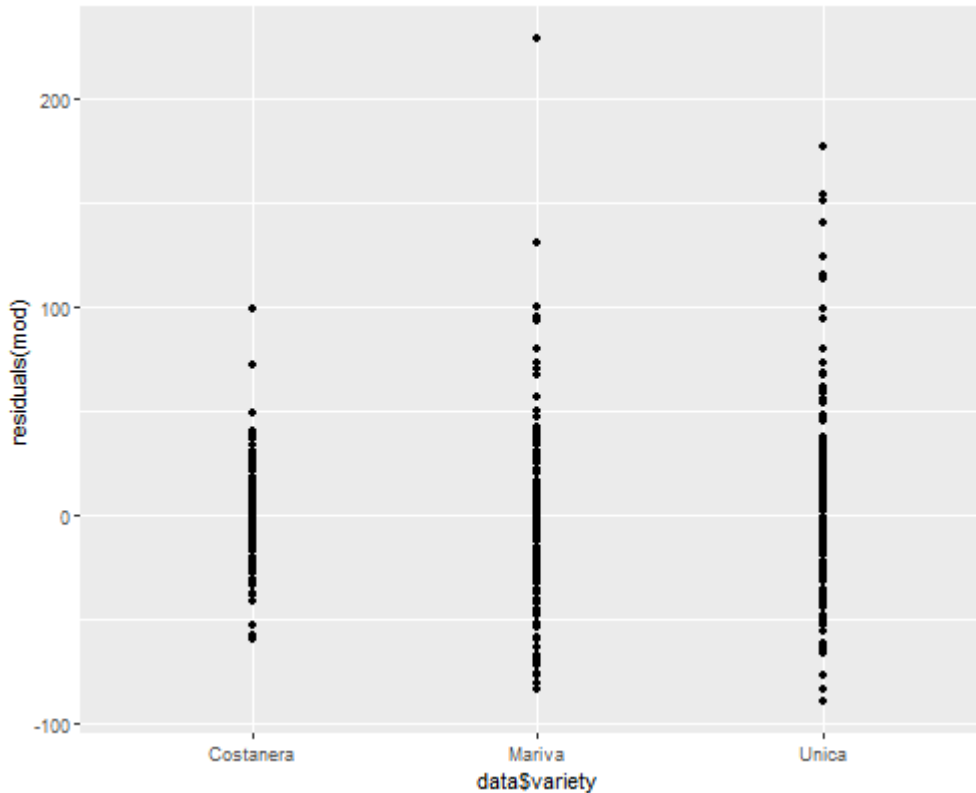
Vérification hypothèses

```
plot(mod, which = 1)
```



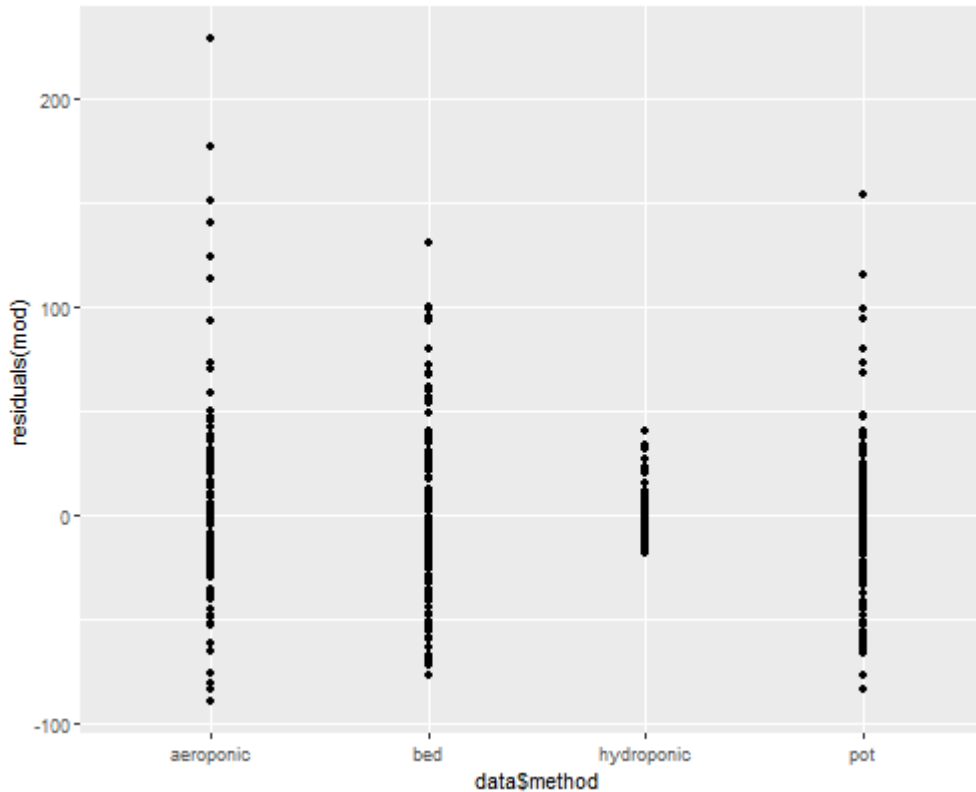
Vérification hypothèses

```
qplot(x = data$variety, y = residuals(mod), geom = "point")
```



Vérification hypothèses

```
qplot(x = data$method, y = residuals(mod), geom = "point")
```



Tests post-hoc (comparaisons multiples)

```
library(emmeans)
pairs(emmeans(mod, ~method))
```

NOTE: Results may be misleading due to involvement in interactions

| ## | contrast | estimate | SE | df | t.ratio | p.value |
|----|--|----------|------|-----|---------|---------|
| ## | aeroponic - bed | -9.79 | 5.13 | 468 | -1.910 | 0.2253 |
| ## | aeroponic - hydroponic | 52.46 | 5.13 | 468 | 10.231 | <.0001 |
| ## | aeroponic - pot | -34.58 | 5.13 | 468 | -6.744 | <.0001 |
| ## | bed - hydroponic | 62.25 | 5.13 | 468 | 12.141 | <.0001 |
| ## | bed - pot | -24.79 | 5.13 | 468 | -4.834 | <.0001 |
| ## | hydroponic - pot | -87.04 | 5.13 | 468 | -16.975 | <.0001 |
| ## | | | | | | |
| ## | Results are averaged over the levels of: variety | | | | | |
| ## | P value adjustment: tukey method for comparing a family of 4 estimates | | | | | |

Tests post-hoc (comparaisons multiples)

```
pairs(emmeans(mod, ~variety|method))
```

```
## method = aeroponic:
## contrast      estimate    SE  df t.ratio p.value
## Costanera - Mariva   -55.20  8.88 468  -6.216  <.0001
## Costanera - Unica    -48.40  8.88 468  -5.449  <.0001
## Mariva - Unica        6.81  8.88 468   0.767  0.7237
##
## method = bed:
## contrast      estimate    SE  df t.ratio p.value
## Costanera - Mariva   -9.26  8.88 468  -1.042  0.5507
## Costanera - Unica    18.46  8.88 468   2.079  0.0953
## Mariva - Unica       27.72  8.88 468   3.121  0.0054
##
## method = hydroponic:
## contrast      estimate    SE  df t.ratio p.value
## Costanera - Mariva    11.88  8.88 468   1.338  0.3748
## Costanera - Unica     -0.15  8.88 468  -0.017  0.9998
## Mariva - Unica       -12.03  8.88 468  -1.355  0.3656
##
## method = pot:
## contrast      estimate    SE  df t.ratio p.value
## Costanera - Mariva    48.84  8.88 468   5.499  <.0001
## Costanera - Unica     46.49  8.88 468   5.234  <.0001
## Mariva - Unica       -2.35  8.88 468  -0.265  0.9621
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

p-value !!!

Tableau d'Anova

| | fongicide | variete | interaction |
|---------|-----------|-----------|-------------|
| p-value | 0.0500205 | 0.0006926 | 0.231347 |

Comparaisons fongicides

| | Estimate | p-value |
|-------|----------|-----------|
| F1-F2 | 9.81 | 0.0500205 |

Comparaisons variétés

| | Estimate | p-value |
|-------|----------|-----------|
| V1-V2 | 2.58 | 0.0284387 |
| V1-V3 | 3.36 | 0.0173694 |
| V2-V3 | 0.78 | 0.1120579 |

Quelles conclusions peut-on faire de ces résultats ?

p-value !!!

- **Les *p-value* dépendent de la qualité ET de la quantité des données !**
- **Le seuil de 0.05 est arbitraire !**
- **Les résultats des tests statistiques ne sont pas la fin de l'analyse. Il est important de prendre du recul et d'évaluer la qualité des résultats !**
- **Les résultats des tests ne constituent pas des conclusions/décisions ! Ce sont des outils qui aident les experts à conclure/décider !**

Exercice 1

Retrouver les valeurs du tableau d'analyse du jeu de données iris

Exercice 2

Analyse du fichier *cow.csv* : mesure de rendement de production laitière par différentes vaches. Les vaches ont reçu soit la ration R1, soit la ration R2 de nourriture et sont à des âges différents (1ère, 2ème, 3ème ou 4ème lactation).

Existe t-il une influence de la ration et/ou de la lactation sur la production de lait ?

Correction exercice 1

```
# Groupe
i <- length(unique(iris$Species))
# Observation
n <- nrow(iris)
# Mu
mu <- pull(iris, Sepal.Length) %>% mean()
# Coef
dtalpha <- iris %>%
  group_by(Species) %>%
  summarize(alphas = mean(Sepal.Length) - mu)
# SS du facteur
SSF <- (n/i)*sum(pull(dtalpha, alphas)^2)
# Residu
dt <- left_join(iris, dtalpha, by = "Species") %>%
  mutate(preds = mu + alphas) %>%
  mutate(e = preds - Sepal.Length)
# SS residuel
SSR <- sum(pull(dt, e)^2)
# Carrée moyens
CMF <- SSF/(i - 1)
CMR <- SSR/(n - i)
# Stat de test
Fobs <- CMF/CMR
# p-value
pval <- pf(Fobs, i - 1, n - i, lower.tail = FALSE)
```

Correction exercice 1

```
tibble(  
  Df = c(i - 1, n - i),  
  SS = c(SSF, SSR),  
  CM = c(CMF, CMR),  
  F = c(Fobs, NA),  
  p = c(pval, NA)  
)
```

```
## # A tibble: 2 x 5  
##       Df    SS    CM    F      p  
##   <dbl> <dbl> <dbl> <dbl>   <dbl>  
## 1     2  63.2  31.6  119. 1.67e-31  
## 2    147  39.0  0.265   NA   NA
```

```
mod <- lm(Sepal.Length~Species, data = iris)  
anova(mod)
```

```
## Analysis of Variance Table  
##  
## Response: Sepal.Length  
##           Df Sum Sq Mean Sq F value    Pr(>F)        
## Species      2  63.212   31.606   119.26 < 2.2e-16 ***  
## Residuals  147  38.956    0.265                  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

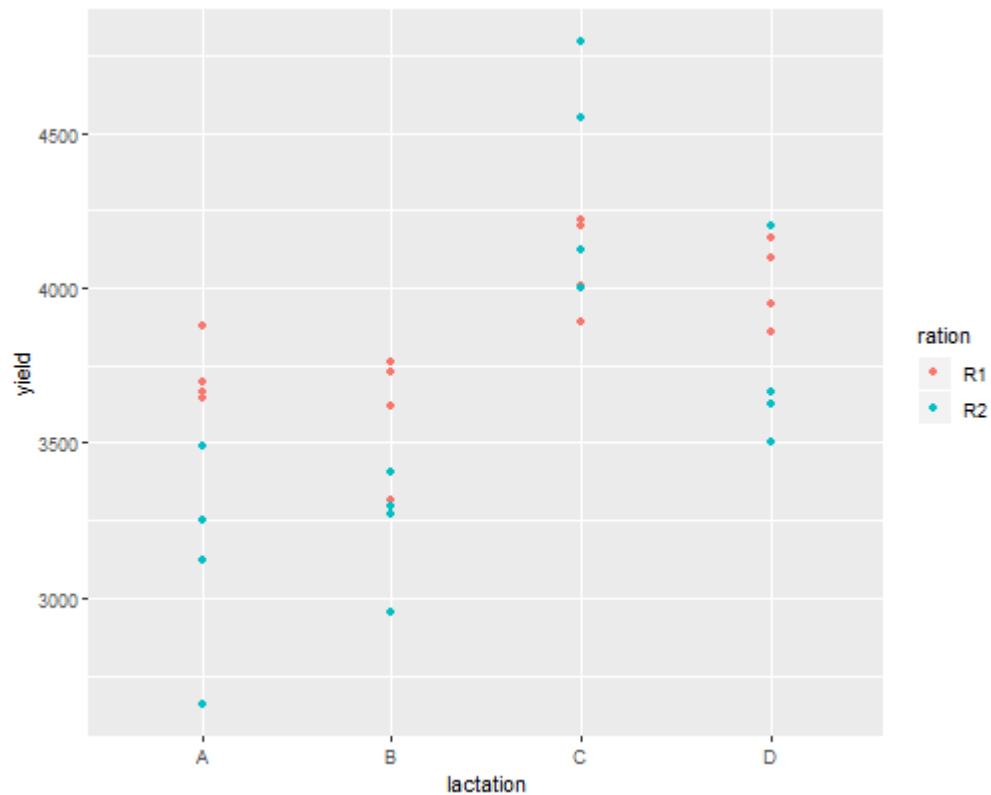
Correction exercice 2

```
library(readr)
library(ggplot2)
library(emmeans)
library(dplyr)
data <- read_delim(here::here("static/data/cows.csv"), delim = ";")
summary(data)
```

| ## | ration | lactation | yield |
|----|------------------|------------------|--------------|
| ## | Length:32 | Length:32 | Min. :2660 |
| ## | Class :character | Class :character | 1st Qu.:3470 |
| ## | Mode :character | Mode :character | Median :3716 |
| ## | | | Mean :3738 |
| ## | | | 3rd Qu.:4030 |
| ## | | | Max. :4792 |

Correction exercice 2

```
ggplot(data) +  
  aes(x = lactation, y = yield, color = ration) +  
  geom_point()
```



Correction exercice 2

```
mod <- lm(yield ~ ration + lactation + ration:lactation, data = data)
anova(mod)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: yield
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---------------------|----|---------|---------|---------|-----------|-----|
| ## ration | 1 | 448404 | 448404 | 7.3174 | 0.01236 | * |
| ## lactation | 3 | 3631550 | 1210517 | 19.7540 | 1.136e-06 | *** |
| ## ration:lactation | 3 | 838933 | 279644 | 4.5634 | 0.01148 | * |
| ## Residuals | 24 | 1470708 | 61280 | | | |

```
## ---
```

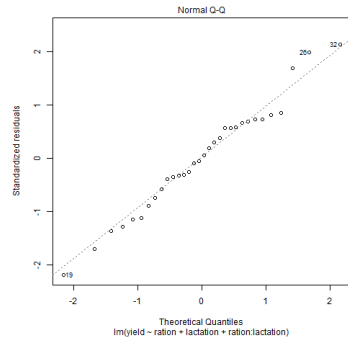
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correction exercice 2

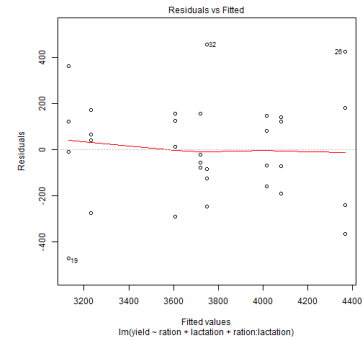
```
summary(mod)
```

```
##
## Call:
## lm(formula = yield ~ ration + lactation + ration:lactation, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -471.0 -133.0   0.5  142.8  456.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3737.63     43.76  85.411  < 2e-16 ***
## ration1         118.37     43.76   2.705  0.012362 *
## lactation1     -311.63     75.80  -4.111  0.000397 ***
## lactation2     -317.62     75.80  -4.191  0.000325 ***
## lactation3      484.87     75.80   6.397  1.29e-06 ***
## ration1:lactation1  176.62     75.80   2.330  0.028525 *
## ration1:lactation2   69.63     75.80   0.919  0.367450
## ration1:lactation3 -261.88     75.80  -3.455  0.002059 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 247.5 on 24 degrees of freedom
## Multiple R-squared:  0.7698,    Adjusted R-squared:  0.7027
## F-statistic: 11.47 on 7 and 24 DF,  p-value: 2.562e-06
```

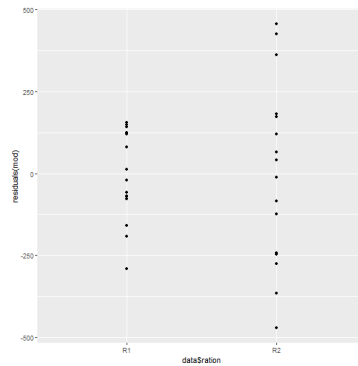
```
plot(mod, which = 2)
```



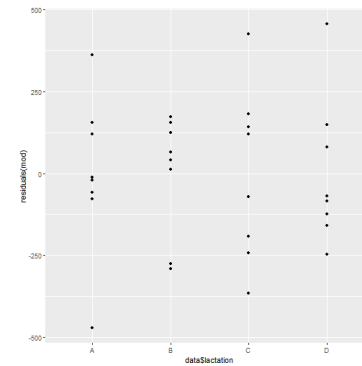
```
plot(mod, which = 1)
```



```
qplot(x = data$ration,
      y = residuals(mod),
      geom = "point")
```



```
qplot(x = data$lactation,
      y = residuals(mod),
      geom = "point")
```



Correction exercice 2

```
library(emmeans)
pairs(emmeans(mod,~ration))
```

NOTE: Results may be misleading due to involvement in interactions

```
## contrast estimate SE df t.ratio p.value
## R1 - R2          237 87.5 24 2.705 0.0124
##
## Results are averaged over the levels of: lactation
```

```
pairs(emmeans(mod,~lactation))
```

NOTE: Results may be misleading due to involvement in interactions

```
## contrast estimate SE df t.ratio p.value
## A - B             6 124 24 0.048 1.0000
## A - C            -796 124 24 -6.435 <.0001
## A - D            -456 124 24 -3.684 0.0060
## B - C            -802 124 24 -6.484 <.0001
## B - D            -462 124 24 -3.733 0.0053
## C - D             340 124 24 2.751 0.0508
##
## Results are averaged over the levels of: ration
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Correction exercice 2

```
pairs(emmeans(mod,~ration|lactation))
```

```
## lactation = A:  
## contrast estimate SE df t.ratio p.value  
## R1 - R2          590 175 24  3.371  0.0025  
##  
## lactation = B:  
## contrast estimate SE df t.ratio p.value  
## R1 - R2          376 175 24  2.148  0.0420  
##  
## lactation = C:  
## contrast estimate SE df t.ratio p.value  
## R1 - R2         -287 175 24 -1.640  0.1141  
##  
## lactation = D:  
## contrast estimate SE df t.ratio p.value  
## R1 - R2          268 175 24  1.531  0.1388
```

Correction exercice 2

```
pairs(emmeans(mod,~lactation|ration))
```

```
## ration = R1:
## contrast estimate SE df t.ratio p.value
## A - B          113 175 24  0.646  0.9161
## A - C         -358 175 24 -2.045  0.1999
## A - D         -295 175 24 -1.685  0.3530
## B - C         -471 175 24 -2.691  0.0577
## B - D         -408 175 24 -2.331  0.1190
## C - D           63 175 24  0.360  0.9836
##
## ration = R2:
## contrast estimate SE df t.ratio p.value
## A - B          -101 175 24 -0.577  0.9380
## A - C         -1235 175 24 -7.055 <.0001
## A - D          -617 175 24 -3.525  0.0088
## B - C         -1134 175 24 -6.478 <.0001
## B - D          -516 175 24 -2.948  0.0331
## C - D           618 175 24  3.531  0.0087
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```