Introduction au logiciel R

Partie 2

Magalie HOUÉE-BIGOT et Mathieu EMILY

Institut Agro - Agrocampus Ouest

Planning prévisionnel

Les principaux objectifs de la formation sont :

- Jour 1 matin : Introduction au logiciel R/RStudio
  - Gestion de données
  - Gestion du système de pacakges
  - Travail en mode projet collaboratif
  - Rapport automatisé
- Jour 1 après-midi : Manipulation et résumé de données statistiques
  - Importation de données
  - Manipulation de données
  - Résumés quantitatifs et visualisation simple
- Jour 2 matin : Exemple du modèle linéaire
  - Analyse de la variance
  - Régression simple et multiple
- Jour 2 après-midi : Prise d'autonomie

Exemple de vers de terre

Young rice seedlings (*Oryza sativa*, cv. Moroberekan) were grown for three months under a 600 μmol photons $m^{-}2$ $s^{-}1$ artificial light source, at 28°C -1 and 24 °C night-1 temperatures and at 75%±5% air moisture. Pots (10 cm in diameter) were filled with 1 kg of a sandy ultisol from Lamto savannah (Ivory Coast).

- 3 factors (= 3 traitements)
  - `millsonia` (TRUE/FALSE). Introduction ou non de vers de terre l'espèce *millsonia anomala*
  - `chuniodrilus` (TRUE/FALSE). Introduction ou non de vers de terre l'espèce *Chuniodrilus zielae*
  - `azote`. Cinq niveaux de concentrations en NH4+ (0,25,100,400,1600 μmol^{-1})
- 2 variables réponses : `bmaer` la biomasse aérienne et `bmrac` la biomasse racinaire
- 3 répétitions par condition expérimentale ($5x2x2x3 = 60$ observations au total)

Contexte

Certaines questions statistiques sont induites par les résumés préliminaires

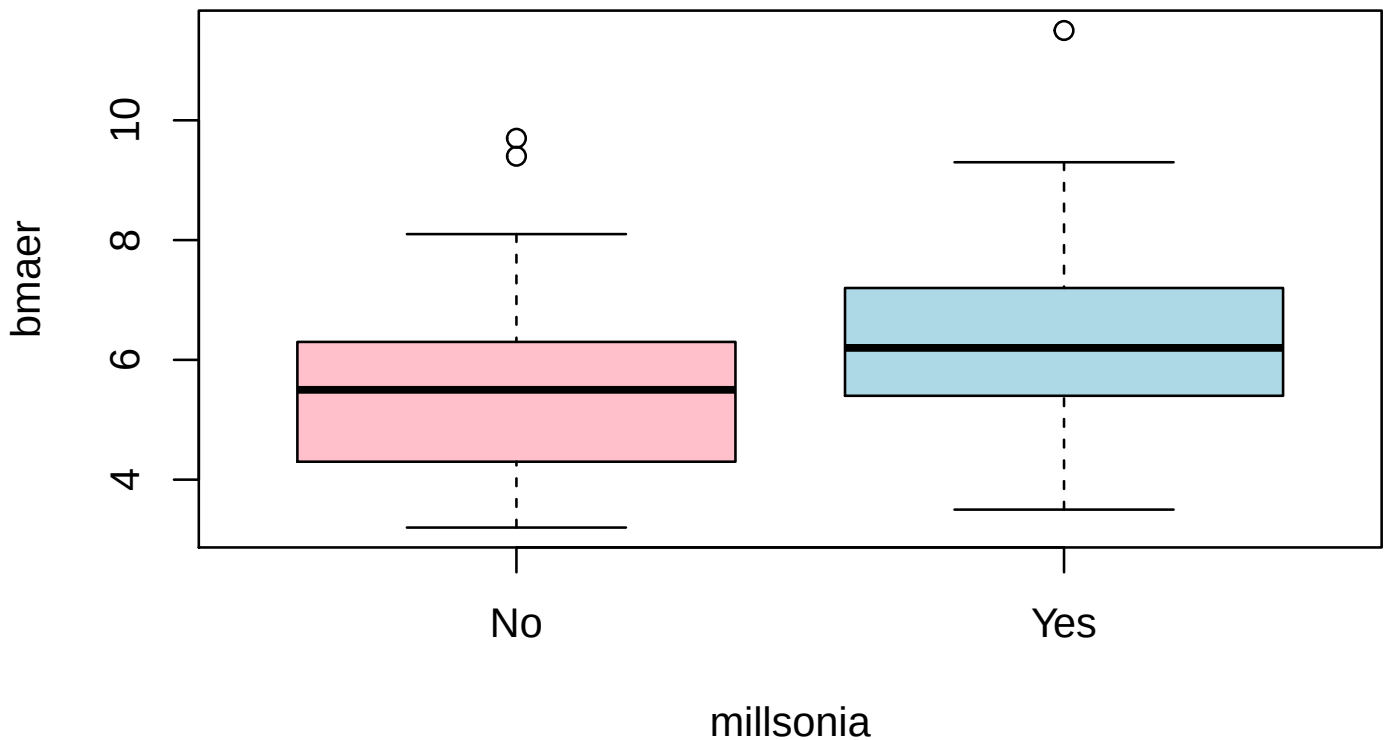
- La biomasse aérienne est-elle influencée par la présence de vers de terre?
- La biomasse aérienne peut-elle être utilisée comme un bon proxy pour la biomasse racinaire?
- La biomasse augmente-t-elle avec le niveau d'azote?
  - Comment?
  - La présence de vers de terre influence-t-elle la relation biomasse/azote?

Plan

Plan

Visualisation des données

```
boxplot(bmaer~millsonia,data=data.EarthWorms,col=c("pink","lightblue"))
```



Comparaison de 2 moyennes: test de la normalité

A-t-on bien la normalité de la biomasse en présence et en absence de millsonia?

```
by(data.EarthWorms$bmaer, data.EarthWorms$millsonia, shapiro.test)
```

```
## data.EarthWorms$millsonia: No
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.93372, p-value = 0.06168
##
## ------------------------------------------------------------
## data.EarthWorms$millsonia: Yes
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94336, p-value = 0.112
```

On accepte l'hypothèse de normalité dans les 2 conditions expérimentales

Comparaison de 2 moyennes : test d'égalite des variances

Quel test utiliser ? Celui avec variances égales ou inégales ?

```
var.test(bmaer ~ millsonia, conf.level=.95, data=data.EarthWorms)
```

```
##
##  F test to compare two variances
##
## data:  bmaer by millsonia
## F = 1.0132, num df = 29, denom df = 29, p-value = 0.9722
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4822296 2.1286499
## sample estimates:
## ratio of variances
##           1.013162
```

On accepte l'hypothèse d'égalité des variances $\Longrightarrow$ on considère que les variances sont égales

Test de comparaison de 2 moyennes (suite et fin. . . enfin presque)

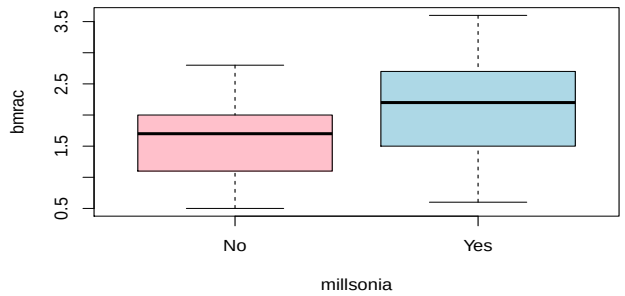
```
res <- t.test(bmaer ~ millsonia, alternative="two.sided", conf.level=.95,
              var.equal=TRUE, data=data.EarthWorms)
res
```

```
##
##  Two Sample t-test
##
## data:  bmaer by millsonia
## t = -1.924, df = 58, p-value = 0.05927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7819469  0.0352802
## sample estimates:
##  mean in group No mean in group Yes
##          5.543333          6.416667
```

On considère que la présence ou l'absence de vers de terre millsonia n'influence pas la biomasse aérienne

Quid de la biomasse racinaire? (1)

```
boxplot(bmrac~millsonia,data=data.EarthWorms,col=c("pink","lightblue"))
```




```
var.test(bmrac ~ millsonia, conf.level=.95, data=data.EarthWorms)
```

```
##
##  F test to compare two variances
##
## data:  bmrac by millsonia
## F = 0.57779, num df = 29, denom df = 29, p-value = 0.1456
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2750096 1.2139428
## sample estimates:
## ratio of variances
##           0.577794
```

Nous considérons l'égalité des variances

Quid de la biomasse racinaire? (2)

```
res <- t.test(bmrac ~ millsonia, alternative="two.sided", conf.level=.95,
              var.equal=TRUE, data=data.EarthWorms)
res
```

```
##
##  Two Sample t-test
##
## data:  bmrac by millsonia
## t = -3.0594, df = 58, p-value = 0.003355
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9319094 -0.1947572
## sample estimates:
##  mean in group No mean in group Yes
##          1.583333           2.146667
```

On peut conclure que la présence de vers de terre Millsonia influence la biomasse racinaire.

Plan

Modèle statistique

Sous l'hypothèse d'égalité des variances, le test de comparaison de moyenne peut se réécrire comme un modèle linéaire

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$ et $Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$

```
mod.one.way <- lm(bmrac ~ millsonia,data=data.EarthWorms)
anova(mod.one.way)
```

```
## Analysis of Variance Table
##
## Response: bmrac
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## millsonia  1  4.7602  4.7602  9.3601 0.003355 **
## Residuals 58 29.4963  0.5086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Test global:

$$\mathcal{H}_0 : \forall i \ \alpha_i = 0 \ \ vs. \ \ \mathcal{H}_1 : \exists i \ \alpha_i \neq 0$$

Modèle avec un facteur explicatif > 2 modalités

```
mod.one.way.2 <- lm(bmrac ~ azote.fac,data=data.EarthWorms)
anova(mod.one.way.2)
```

```
## Analysis of Variance Table
##
## Response: bmrac
##           Df Sum Sq Mean Sq F value Pr(>F)
## azote.fac  4  1.599 0.39975  0.6732 0.6134
## Residuals 55 32.658 0.59377
```

- Interpretion des coefficients

```
print(mod.one.way.2)
```

```
##
## Call:
## lm(formula = bmrac ~ azote.fac, data = data.EarthWorms)
##
## Coefficients:
##       (Intercept)       azote.facLow    azote.facMedium      azote.facHigh
##           1.87500           -0.15000           -0.15833            0.29167
## azote.facVeryHigh
##           -0.03333
```

Coefficient interpretation

Attention à la contrainte utilisée

```
contrasts(data.EarthWorms$azote.fac) <- contr.sum(n=5)
mod.one.way.2.2 <- lm(bmrac ~ azote.fac,data=data.EarthWorms)
print(mod.one.way.2.2)
```

```
##
## Call:
## lm(formula = bmrac ~ azote.fac, data = data.EarthWorms)
##
## Coefficients:
## (Intercept)   azote.fac1   azote.fac2   azote.fac3   azote.fac4
##      1.8650       0.0100      -0.1400      -0.1483       0.3017
```

- Pour `mod.one.way.2`, "VeryLow" est la modalité de référence ($\alpha_1 = 0$)
  - μ correspond à la moyenne espérée pour la condition azote=VeryLow
  - Chaque α_i est la différence de moyenne entre azote=i et azote=VeryLow
- Pour `mod.one.way.2.2`: $\alpha_5 = -\alpha_1 - \alpha_2 - \alpha_3 - \alpha_4$
  - μ correspond à la moyenne globale attendue
  - Chaque α_i est la différence entre la moyenne pour azote=i et la moyenne globale

Analyse Post-Hoc - comparaison entre les modalités d'un facteur

```
library(multcomp)
summary(glht(mod.one.way.2,linfct = mcp(azote.fac = "Tukey")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = bmrac ~ azote.fac, data = data.EarthWorms)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## Low - VeryLow == 0       -0.150000   0.314582  -0.477    0.989
## Medium - VeryLow == 0    -0.158333   0.314582  -0.503    0.987
## High - VeryLow == 0       0.291667   0.314582   0.927    0.885
## VeryHigh - VeryLow == 0 -0.033333   0.314582  -0.106    1.000
## Medium - Low == 0        -0.008333   0.314582  -0.026    1.000
## High - Low == 0           0.441667   0.314582   1.404    0.628
## VeryHigh - Low == 0       0.116667   0.314582   0.371    0.996
## High - Medium == 0        0.450000   0.314582   1.430    0.611
## VeryHigh - Medium == 0    0.125000   0.314582   0.397    0.995
## VeryHigh - High == 0     -0.325000   0.314582  -1.033    0.839
## (Adjusted p values reported -- single-step method)
```

Plan

Modèle statistique

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

```
mod.two.way.inter <- lm(bmaer ~ millsonia+azote.fac+millsonia:azote.fac,data=data.EarthWorms)
anova(mod.two.way.inter)
```

```
## Analysis of Variance Table
##
## Response: bmaer
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## millsonia            1  11.441 11.4407  8.4226  0.005499 **
## azote.fac            4 106.008 26.5019 19.5106 1.002e-09 ***
## millsonia:azote.fac  4   5.331  1.3328  0.9812  0.426322
## Residuals           50  67.917  1.3583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- L'interaction n'est pas significative : il est préférable de l'enlever du modèle pour améliorer l'estimation des autres facteurs

Sélection de facteurs (1)

- Test global

```
data.EarthWorms$millsonia <- as.factor(data.EarthWorms$millsonia)
levels(data.EarthWorms$millsonia) <- c("No","Yes")
mod.two.way.nointer <- lm(bmaer ~ millsonia+azote.fac,data=data.EarthWorms)
anova(mod.two.way.nointer)
```

```
## Analysis of Variance Table
##
## Response: bmaer
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## millsonia   1  11.441 11.4407  8.4343  0.005325 **
## azote.fac   4 106.008 26.5019 19.5379 5.436e-10 ***
## Residuals 54  73.248  1.3564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sélection de facteurs (2)

- Analyse post-hoc

```
summary(glht(mod.two.way.nointer,linfct = mcp(millsonia = "Tukey")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = bmaer ~ millsonia + azote.fac, data = data.EarthWorms)
##
## Linear Hypotheses:
##                Estimate Std. Error t value Pr(>|t|)
## Yes - No == 0   0.8733     0.3007   2.904  0.00532 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Plan

Plan

Modèle statistique

La biomasse aérienne peut-elle être utilisée comme un bon proxy pour la biomasse racinaire?

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ et $Cov(\varepsilon_i, \varepsilon_{i'}) = 0$

```
mod.reg.simple <- lm(bmrac ~ bmaer,data=data.EarthWorms)
```

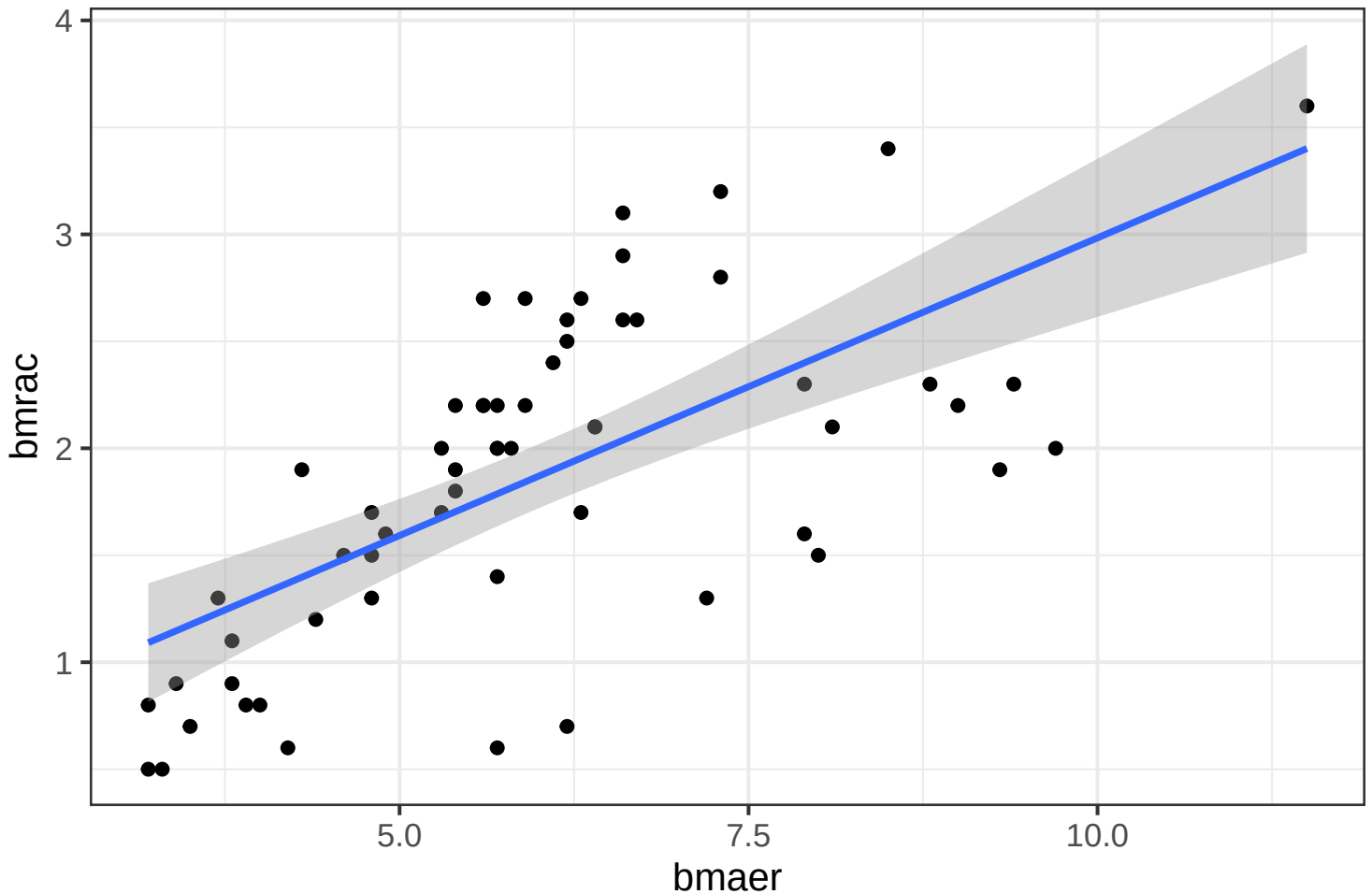
Interprétation

```
summary(mod.reg.simple)
```

```
##
## Call:
## lm(formula = bmrac ~ bmaer, data = data.EarthWorms)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22622 -0.40902  0.05252  0.45466  1.06246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20085    0.26194   0.767    0.446
## bmaer        0.27829    0.04198   6.630 1.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5797 on 58 degrees of freedom
## Multiple R-squared:  0.4311, Adjusted R-squared:  0.4213
## F-statistic: 43.95 on 1 and 58 DF,  p-value: 1.224e-08
```

Visualisation

```
ggplot(data = data.EarthWorms) +
  aes(x = bmaer, y = bmrac) +
  geom_point() +
  geom_smooth(method = 'lm', se = TRUE)
```



Plan

Modèle statistique

$$Y_{ik} = \mu + \alpha_i + \beta x + \gamma_i x_{ik} + \varepsilon_{ik}$$

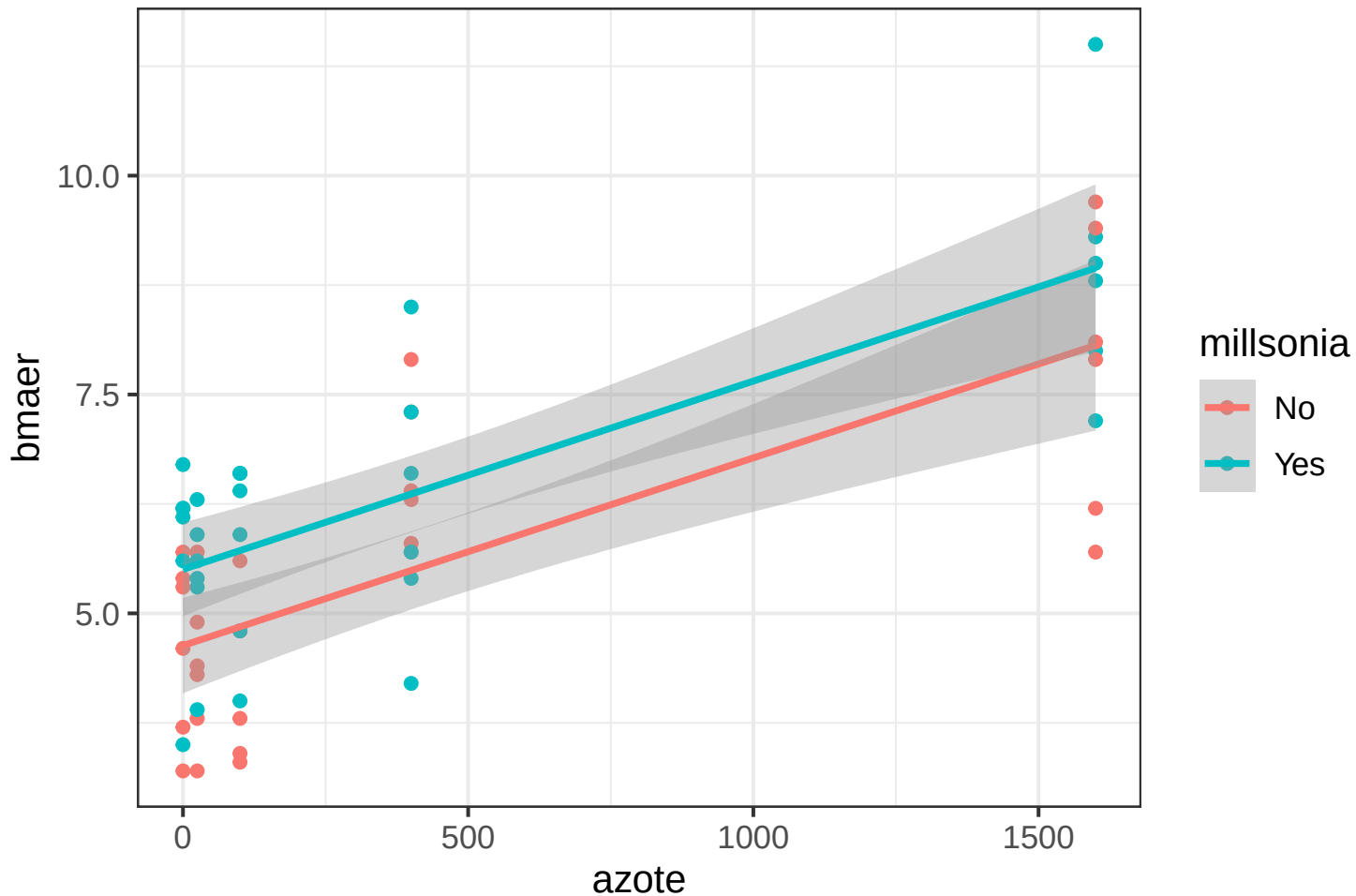
- $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma)$ et $Cov(\varepsilon_{ik}, \varepsilon_{i'k'}) = 0$

Etant donné que $\sum_i \alpha_i = 0$, on a

- μ est la moyenne générale attendue quand $x = 0$
- α_i correspond à la différence de moyenne quand $x = 0$
- β est la pente de la droite de régression
- γ_i correspond à la différence des pentes par rapport à i

Visualisation avec ggplot2

```
ggplot(data = data.EarthWorms) +
  aes(x = azote, y = bmaer,col=millsonia) +
  geom_point() +
  geom_smooth(method = 'lm', se = TRUE)
```



Modélisation et tests avec R

```
mod.ancova.inter <- lm(bmaer ~ millsonia+azote+millsonia:azote,data=data.EarthWorms)
anova(mod.ancova.inter)
```

```
## Analysis of Variance Table
##
## Response: bmaer
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## millsonia       1  11.441  11.441  8.2025  0.005878 **
## azote           1 101.148 101.148 72.5195 1.103e-11 ***
## millsonia:azote  1   0.000   0.000  0.0002  0.988869
## Residuals      56  78.107   1.395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod.ancova.nointer <- lm(bmaer ~ millsonia+azote,data=data.EarthWorms)
anova(mod.ancova.nointer)
```

```
## Analysis of Variance Table
##
## Response: bmaer
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## millsonia  1  11.441  11.441   8.349  0.005449 **
## azote      1 101.148 101.148  73.814 7.221e-12 ***
## Residuals 57  78.107   1.370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plan

La chenille processionnaire du pin

- Une variable réponse : `NbNests` le nombre de nids
- 10 variables explicatives
  - `Altitude` (m), `Pente` (en degrés), `NbPines`, `Height` (de l'arbre), `Diameter` (of the tree), `Density` (densité de population), `Orientation` (à partir du sud), `MaxHeight`, `NbStrat` (nombre de couches de végétation), `Mix` (une score de mixité de population)
  - 33 observations

```
data.Caterpillar <- read.table("./Data/Caterpillar.csv",header=TRUE)
head(data.Caterpillar,n=3)
```

```
##   Altitude Slope NbPines Height Diameter Density Orientation MaxHeight NbStrat
## 1     1200    22       1    4.0     14.8     1.0         1.1       5.9     1.4
## 2     1342    28       8    4.4     18.0     1.5         1.5       6.4     1.7
## 3     1231    28       5    2.4      7.8     1.3         1.6       4.3     1.5
```

Modèle statistique

$$Y_k = \beta_0 + \sum_i \beta_i x_k + \varepsilon_k$$

- $\varepsilon_k \sim \mathcal{N}(0, \sigma)$
- Hypothèses: $Cov(\varepsilon_k, \varepsilon_{k'}) = 0$

```
data.Caterpillar.f <- data.Caterpillar[,-12]
mod.init <- lm(NbNests~.,data=data.Caterpillar.f)
```

Résultats (1)

```
summary(mod.init)
```

```
##
## Call:
## lm(formula = NbNests ~ ., data = data.Caterpillar.f)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03941 -0.26272 -0.02351  0.21953  1.35140
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.561849   2.096950   4.083 0.000493 ***
## Altitude    -0.002956   0.001038  -2.847 0.009374 **
## Slope       -0.034821   0.014510  -2.400 0.025311 *
## NbPines      0.035385   0.066586   0.531 0.600454
## Height      -0.501564   0.378701  -1.324 0.198955
## Diameter     0.108739   0.069495   1.565 0.131925
## Density     -0.032715   1.044915  -0.031 0.975305
## Orientation -0.203959   0.669598  -0.305 0.763535
## MaxHeight    0.028180   0.157007   0.179 0.859201
## NbStrat     -0.862409   0.572133  -1.507 0.145945
## Mix         -0.448124   0.513764  -0.872 0.392499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5493 on 22 degrees of freedom
## Multiple R-squared:  0.6809, Adjusted R-squared:  0.5359
## F-statistic: 4.695 on 10 and 22 DF,  p-value: 0.001203
```

Résultats (2)

- Comment interpréter que `Density` ne soit pas significatif alors que

```
summary(lm(NbNests~Density,data=data.Caterpillar.f))
```

```
##
## Call:
## lm(formula = NbNests ~ Density, data = data.Caterpillar.f)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11809 -0.49721 -0.04085  0.36506  1.68191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9590     0.3192   6.136 8.33e-07 ***
## Density      -0.6409     0.1658  -3.865 0.000531 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6729 on 31 degrees of freedom
## Multiple R-squared:  0.3252, Adjusted R-squared:  0.3034
## F-statistic: 14.94 on 1 and 31 DF,  p-value: 0.0005309
```

Tests par comparaison de modèles

```
mod.2 <- lm(NbNests~Altitude+Slope+NbPines+
              Height+Diameter+Orientation+MaxHeight+NbStrat+Mix,data=data.Caterpillar.f)
1-pchisq(-2*logLik(mod.2)+2*logLik(mod.init),df=1)

## 'log Lik.' 0.9694124 (df=11)
anova(mod.init,mod.2)

## Analysis of Variance Table
##
## Model 1: NbNests ~ Altitude + Slope + NbPines + Height + Diameter + Density +
##     Orientation + MaxHeight + NbStrat + Mix
## Model 2: NbNests ~ Altitude + Slope + NbPines + Height + Diameter + Orientation +
##     MaxHeight + NbStrat + Mix
##   Res.Df    RSS Df   Sum of Sq     F Pr(>F)
## 1     22 6.6369
## 2     23 6.6372 -1 -0.00029572 0.001 0.9753
```

Résultats (3)

- Lien avec la corrélation?

```
cor(data.Caterpillar.f[c("Altitude","Slope","Density")])
```

```
##             Altitude     Slope   Density
## Altitude  1.0000000 0.1205209 0.5146683
## Slope     0.1205209 1.0000000 0.3006666
## Density   0.5146683 0.3006666 1.0000000
```

- Un besoin de régularisation

$$Critere = Mesured'adequation + Penalisation$$

- Par exemple, nous pouvons utiliser le AIC (Akaike Information Criteria)

$$AIC = 2p - 2\log(L) = 2p + \mathcal{D}$$

Sélection de variables avec step

```
mod.selected <- step(mod.init,direction="both",trace=FALSE)
```

- Première étape

```
Start:  AIC=-30.93
NbNests ~ Altitude + Slope + NbPines + Height + Diameter + Density +
    Orientation + MaxHeight + NbStrat + Mix

              Df Sum of Sq    RSS     AIC
- Density      1   0.00030 6.6372 -32.926
- MaxHeight    1   0.00972 6.6466 -32.879
- Orientation  1   0.02799 6.6649 -32.788
- NbPines      1   0.08520 6.7221 -32.506
- Mix          1   0.22952 6.8664 -31.805
<none>                     6.6369 -30.927
- Height       1   0.52918 7.1661 -30.396
- NbStrat      1   0.68545 7.3224 -29.684
- Diameter     1   0.73859 7.3755 -29.445
- Slope        1   1.73726 8.3742 -25.255
- Altitude     1   2.44545 9.0824 -22.576
```

- A chaque étape, une et une seule variable est soit ajoutée, soit enlevée du modèle

Modèle final

```
summary(mod.selected)
```

```
##
## Call:
## lm(formula = NbNests ~ Altitude + Slope + Height + Diameter +
##     NbStrat, data = data.Caterpillar.f)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09845 -0.27226  0.00947  0.28545  1.23455
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9981789  1.0041510   5.973 2.27e-06 ***
## Altitude    -0.0022915  0.0007892  -2.904  0.00727 **
## Slope       -0.0338090  0.0135138  -2.502  0.01872 *
## Height      -0.5215956  0.2480264  -2.103  0.04493 *
## Diameter     0.1241452  0.0555627   2.234  0.03394 *
## NbStrat     -0.3849351  0.2187347  -1.760  0.08976 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5249 on 27 degrees of freedom
## Multiple R-squared:  0.6424, Adjusted R-squared:  0.5762
## F-statistic: 9.701 on 5 and 27 DF,  p-value: 2.171e-05
```

Impact sur la prédiction

```
new.d <- data.frame(Altitude=1200,Slope=22,NbPines=1,Height=4,
      Diameter=15,Density=1,Orientation=1,MaxHeight=6,NbStrat=1.5,Mix=1.4)
predict(mod.init,newdata=new.d,interval="confidence")
```

```
##        fit      lwr      upr
## 1 1.919882 1.102417 2.737347
```

```
predict(mod.selected,newdata=new.d,interval="confidence")
```

```
##        fit      lwr      upr
## 1 1.702941 1.368881 2.037002
```

- Nous pouvons constater une réduction de la variabilité de la prédiction

Pour résumer

Modèle linéaire et la fonction `lm`

- Un modèle linéaire permet de modéliser des relations entre une variable quantitative Y et un ensemble de variables explicatives quantitatives et qualitatives
- `lm` permet l'estimation des coefficients du modèle
- La sortie de `lm` fournit
  - Des indicateurs globaux : R^2
  - Des indicateurs indviduels : test de nullité par coefficient
  - Attention : ces tests s'appuient sur des hypothèses mathématiques!

Application

Épaisseur de l'Intima-média

Description

L'athérosclérose est la principale cause de décès chez l'homme après 35 ans et chez la femme après 45 ans dans la plupart des pays développés. C'est un épaississement et une perte d'élasticité des parois internes des artères, dont une des conséquences est l'infarctus du myocarde. La paroi artérielle est constituée de trois couches qui sont respectivement à partir de la lumière artérielle : l'intima, la média et l'adventice. L'épaisseur de l'intima-média est un marqueur reconnu d'athérosclérose. Elle a été mesurée par échographie sur un échantillon de 110 sujets en 1999 dans les CHU de Bordeaux. Des informations sur les principaux facteurs de risques ont aussi été recueillies.

Description	Unité ou Codage	Variable
Sexe	1=Homme; 2=Femme	SEXE
Age le jour de la visite	Années	AGE
Taille	cm	taille
Poids	kg	poids
Statut tabagique	0=Ne fume pas 1=A arrêté de fumer 2=Fume	tabac
Estimation de consommation pour les fumeurs et ex-fumeurs	Nombre de paquets/année	paqan
Activité physique	0=Non; 1=Oui	SPORT
Mesure de l'Intima-Média	mm	mesure
Consommation alcool	0=Ne boit pas 1=Boit occasionnellement 2=Boit régulièrement	alcool

Application

Générer un Rmarkdown

Importer le jeu de données et faire un résumé statistique

Analyser l'effet du facteur `alcool` sur la variable `mesure` :

- Calculer la moyenne par groupe du facteur `alcool`
- Visualiser l'effet de ce facteur sur la variable `mesure`
- Evaluer l'effet du facteur `alcool` sur la variable `mesure`

Evaluer les effets des facteurs qualitatifs sur cette variable?

Evaluer les effets des variables explicatives continues (construire le modèle linéaire)

Construire le modèle complet et analyser ce modèle