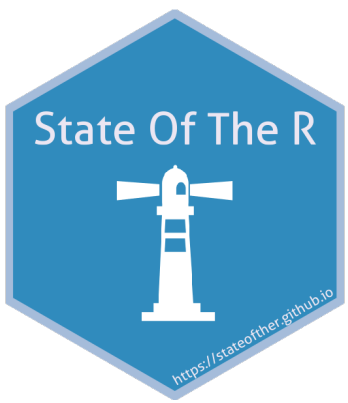
Statistical modelling for biological data with R

Day 2-3 : Linear model with R

Marie-Pierre Etienne

https://marieetienne.github.io

Novembre 2019



Plan

Plan

Les notions clés de la statistique paramétrique

- paramètres
- estimateurs
- estimations

Si besoin Rappel

Plan

Plan

Plan

Contexte

- Les manchots élèvent leurs petits en couple et s'alimentent alternativement en haute mer (voyage d' une dizaine de jours)
- Une étude pour identifier les facteurs de variation dans l'efficacité de leur alimentation.
- Débute après la naissance des petits et se poursuit jusqu'au départ des petits.

Dispositif :

Pour identifier les déplacements des manchots, on équipe certains individus de transmetteurs GPS, pesant 450g et ayant une surface frontale de 14 cm^2, ce qui représente 2.4% de la section d'un oiseau de 24kg. *Est ce un désavantage competitif ?*

Variables mesurées :

- poids initial,
- poids au retour
- GPS (oui ou non)
- Période de suivi (3 périodes).

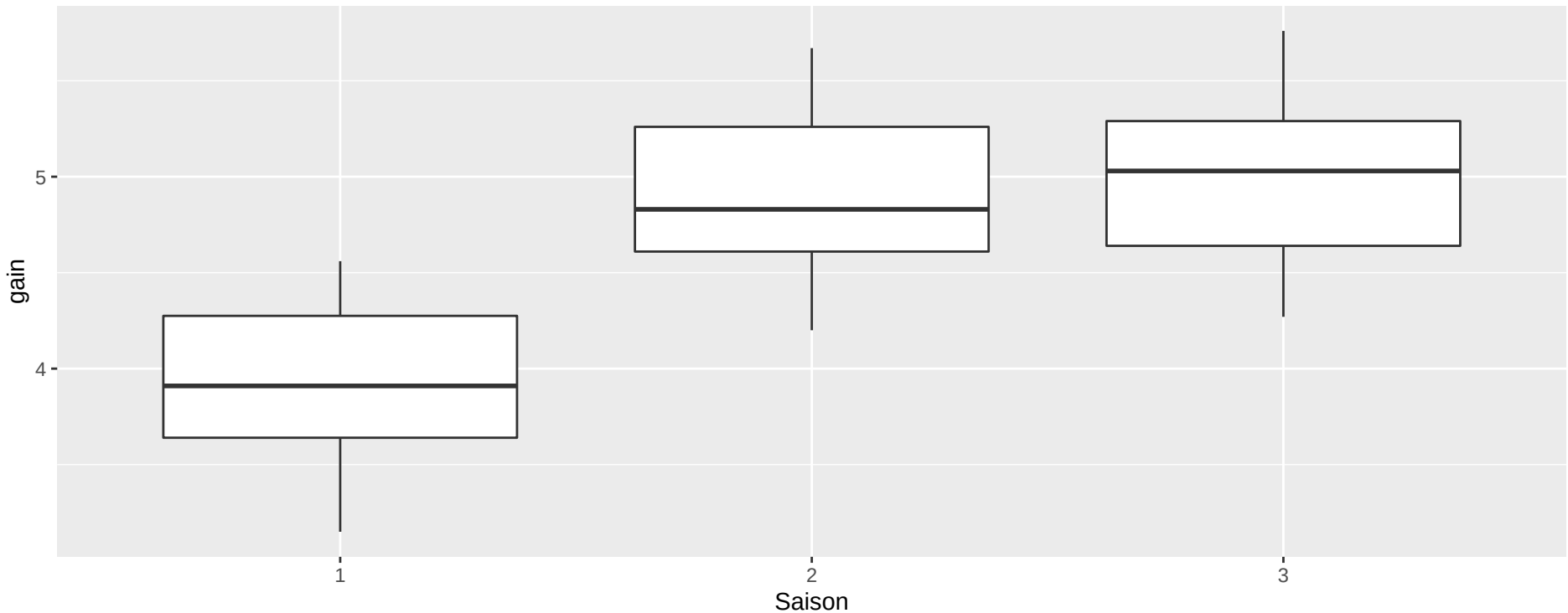
Présentation des données

```
manchots <- read.table('../../Datasets/Manchots.csv',
                       header=T, sep = ";")

manchots %>% mutate(GPS = as.factor(GPS),
                    Saison = as.factor(Saison),
                    gain = PoidsFinal - PoidsInit) -> manchots

manchots %>% group_by(GPS, Saison) %>%
  summarize( n = n())
```

```
## # A tibble: 6 x 3
## # Groups:   GPS [2]
##   GPS   Saison     n
##   <fct> <fct>  <int>
## 1 0     1         27
## 2 0     2         13
## 3 0     3         19
## 4 1     1          3
## 5 1     2          8
## 6 1     3         10
```

Des représentation graphiques

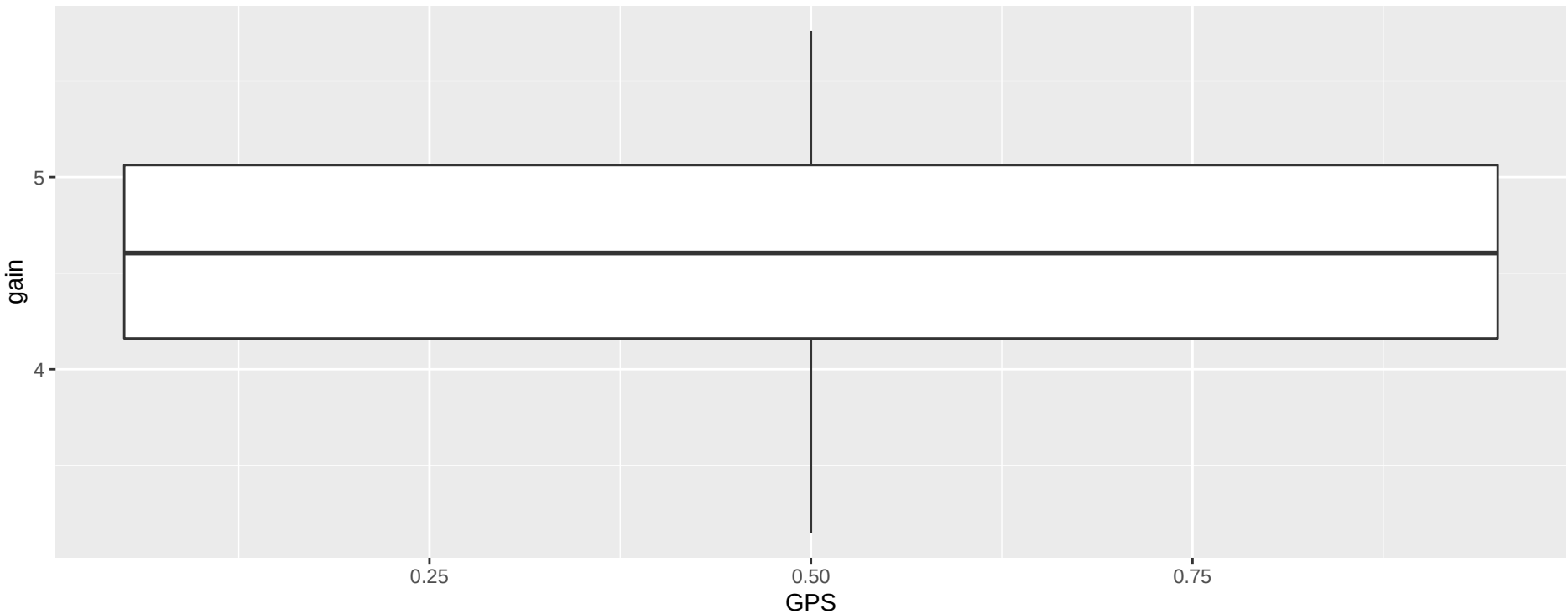
```
manchots %>% ggplot() + geom_boxplot(aes(x= Saison, y=gain))
```



Des représentation graphiques

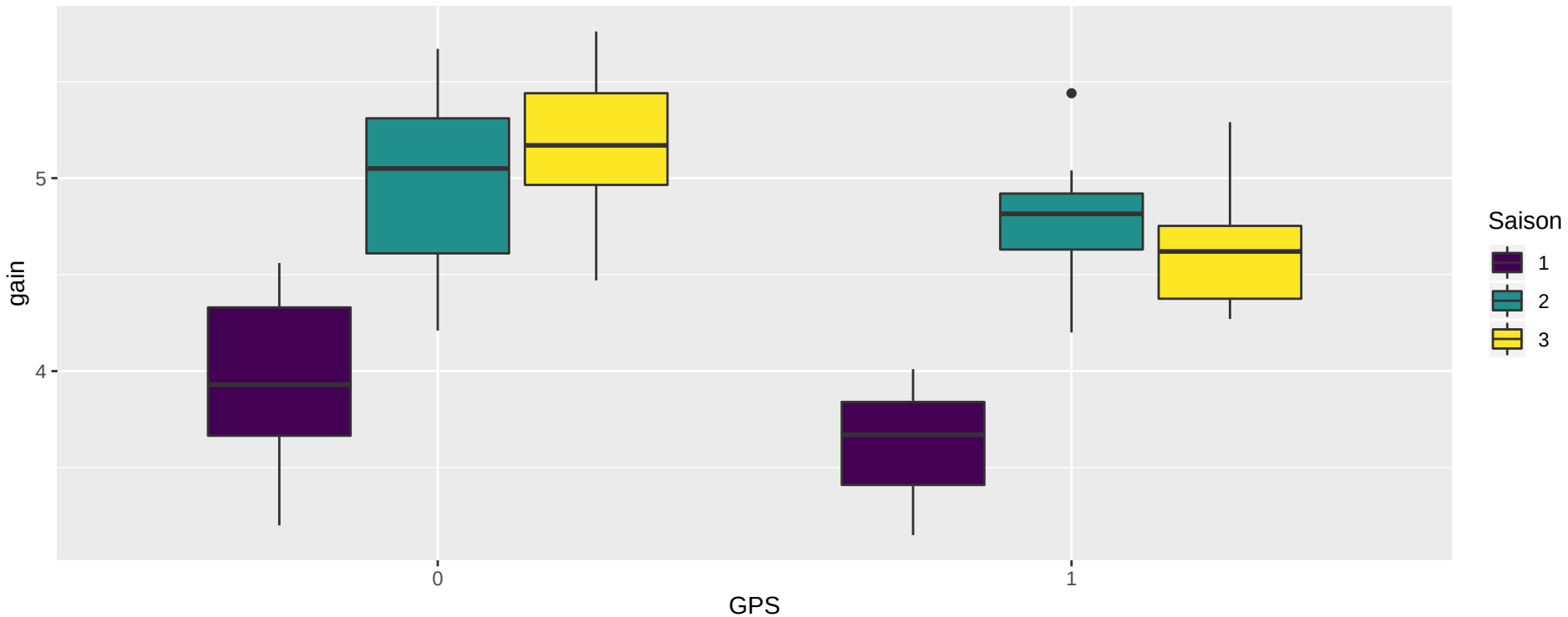
```
manchots %>% ggplot() + geom_boxplot(aes(x= GPS, y=gain))
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)
```



Des représentation graphiques

```
manchots %>% ggplot() +
  geom_boxplot(aes(x= GPS, y=gain, fill = Saison)) +
  scale_fill_viridis_d()
```



Plan

Contexte

Les prairies exploitées de manière intensive constituent des écosystèmes très fréquents

Kirwan Laura et al. (2014) examine l'effet d'une diversification expérimentale des cultures sur le rendement des prairies.

Données

- six sites différents `COUNTRY`
- Sur chaque site 48 parcelles (`PLOT`) on été ensemencées avec un mélange de graines.
- Proportion en herbes (`G`) et en légumineuses (`L`).
- Sur chaque parcelle,un indice de diversité (indice de Shannon), `H` variant entre 1 et 4
- Expérience menée entre 2003 et 2006
- Mesure de rendement `HARV_YIELD`

Question

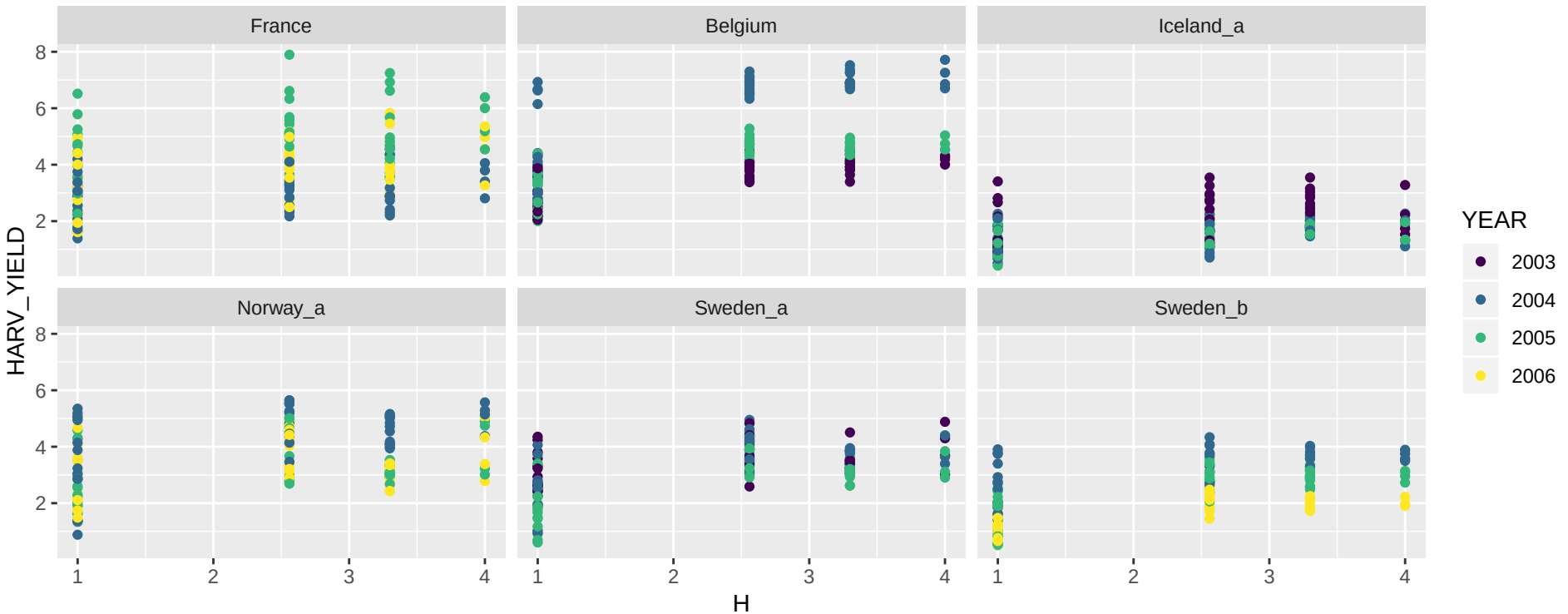
Impact de la biodiversité sur le rendement

Présentation des données

```
biomass <- read.table(file = 'https://marieetienne.github.io/datase
                      sep = ',',
                      header = TRUE)

biomass %>% mutate(Hfact = as.factor(biomass$H)) -> biomass

biomass %>% as_tibble %>% print(n = 4)

## # A tibble: 864 x 8
##   COUNTRY    YEAR  PLOT     G     L HARV_YIELD     H Hfact
##   <fct>     <int> <int> <dbl> <dbl>       <dbl> <dbl> <fct>
## 1 Belgium    2003    12     1     0        2.69     1 1
## 2 Iceland_a  2003    12     1     0        2.82     1 1
## 3 Sweden_a   2003    12     1     0        2.61     1 1
## 4 Belgium    2004    12     1     0        3.44     1 1
## # ... with 860 more rows
```

Représentation des données

```
biomass %>%  ggplot() + facet_wrap(~COUNTRY) + geom_point(aes(x=H,
```



Plan

Plan

Version mathématique

Le modèle d'analyse de variance à 1 facteur s'écrit :

$$Y_{ik} = \mu + \alpha_i + E_{ik}$$

avec $E_{ik} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ et $cov(E_{ik}, E_{i'k'}) = 0 \quad \forall\ (i,k) \neq (i',k')$

- μ effet de référence
- α_i effet différentiel du niveau i du premier facteur

Objectif de l'analyse de variance : étudier si Y varie selon les modalités du fateur

Version matricielle

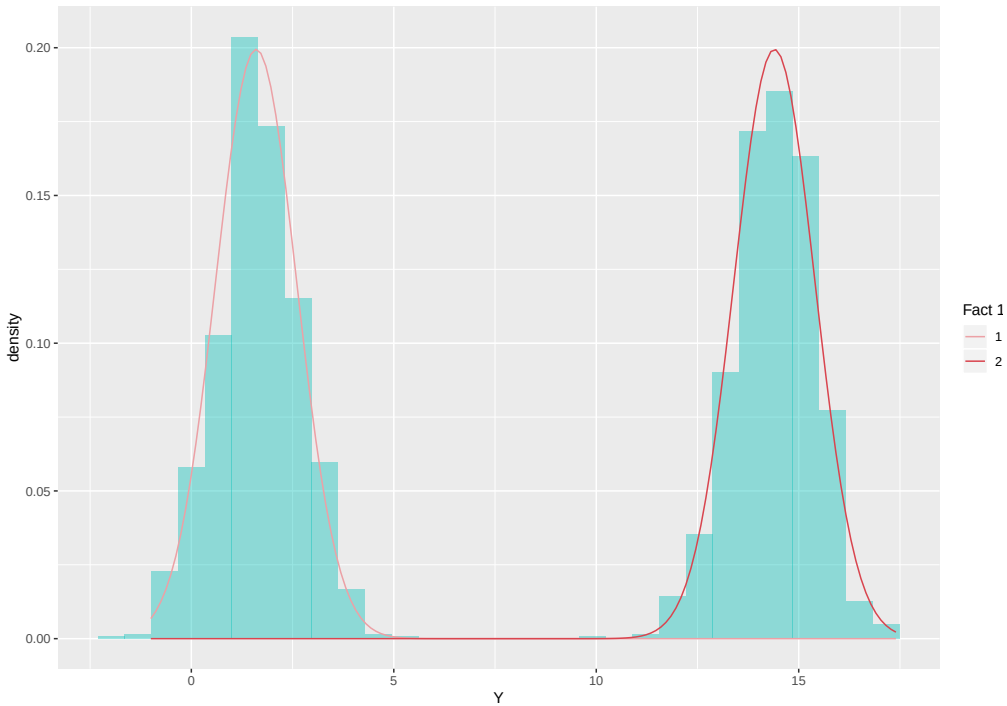
Le modèle d'analyse de variance à 2 facteurs s'écrit :

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Version graphique

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```



Plan

Version mathématique

Le modèle d'analyse de variance à 2 facteurs s'écrit :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$$

avec $E_{ijk} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ et $cov(E_{ijk}, E_{i'j'k'}) = 0 \quad \forall\ (i,j,k) \neq (i',j',k')$

- μ effet de référence
- α_i effet différentiel du niveau i du premier facteur
- β_j effet différentiel du niveau j du second facteur
- $(\alpha\beta)_{ij}$ effet différentiel de l'interaction des niveaux i et j

Objectif de l'analyse de variance : étudier parmi ces effets ceux qui influent sur Y

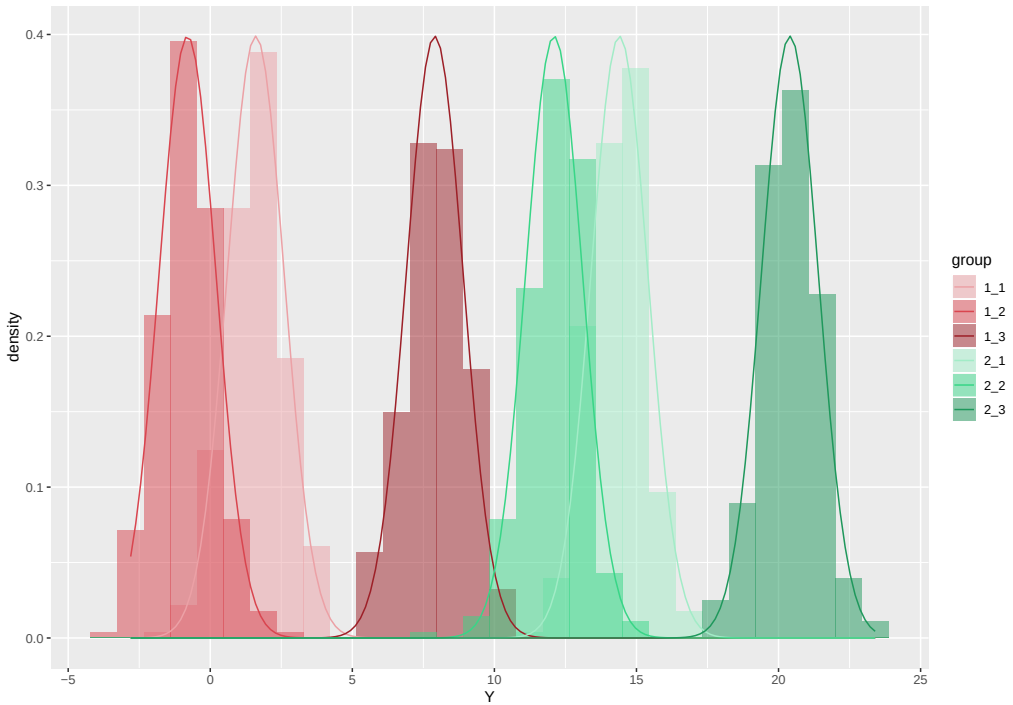
Version matricielle

Le modèle d'analyse de variance à 2 facteurs s'écrit :

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Version graphique



Plan

Version mathématique

$$Y_k = \mu + \beta x_k + E_k$$

avec $E_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

- μ effet de référence
- β effet de la variable x_k

Objectif de la régression simple : Quantifier l'effet de x sur Y, prédire Y

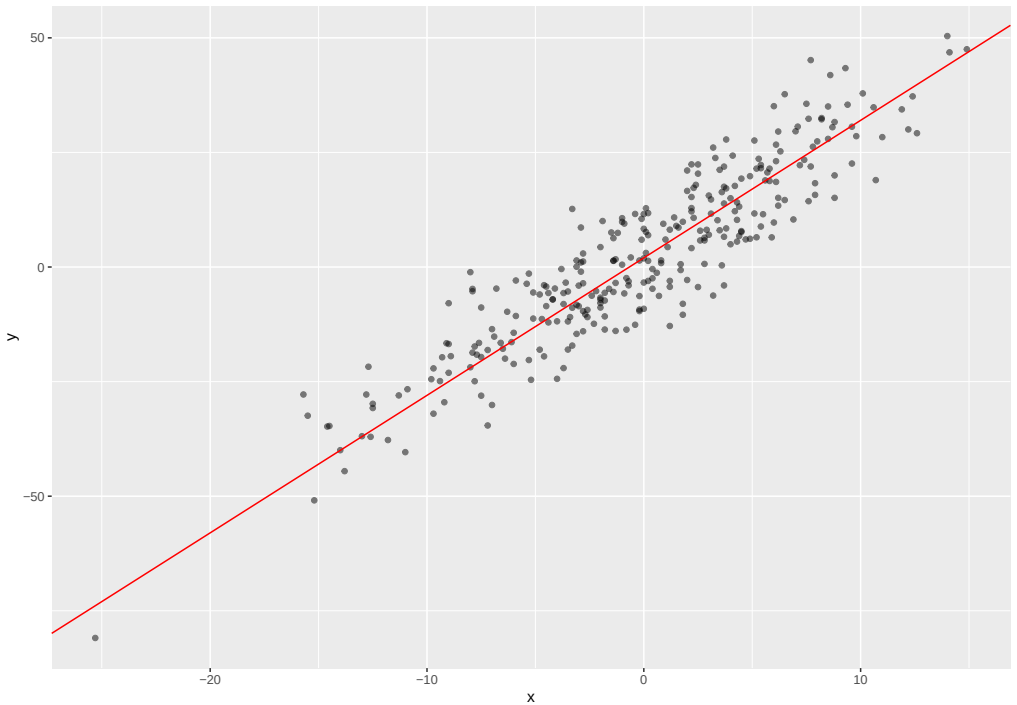
Version matricielle

Le modèle de régression simple s'écrit:

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Modèle vision graphique



Plan

Version mathématique

$$Y_k = \mu + \beta_1 x_k^{(1)} + \beta_2 x_k^{(2)} + \ldots + \beta_p x_k^{(p)} + E_k$$

avec $E_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

- μ effet de référence
- β_1 effet de la variable $x_k^{(1)}$
- ...
- β_p effet de la variable $x_k^{(p)}$

Objectif de la régression multiple : identifier les variables x liées à Y, prédire$Y$

Version matricielle

Le modèle de régression multiple s'écrit:

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Plan

Version mathématique

$$Y_{ik} = \mu + \alpha_i + \beta x_k + \gamma_i x_k + E_k$$

avec $E_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

- μ effet de référence
- α_i effet différentiel du niveau i du facteur
- β effet de la variable x_k
- γ_i effet différentiel du niveau i sur le lien entre x et Y.

Objectif de l'analyse de la covariance : comparer des droites de régression

Version matricielle

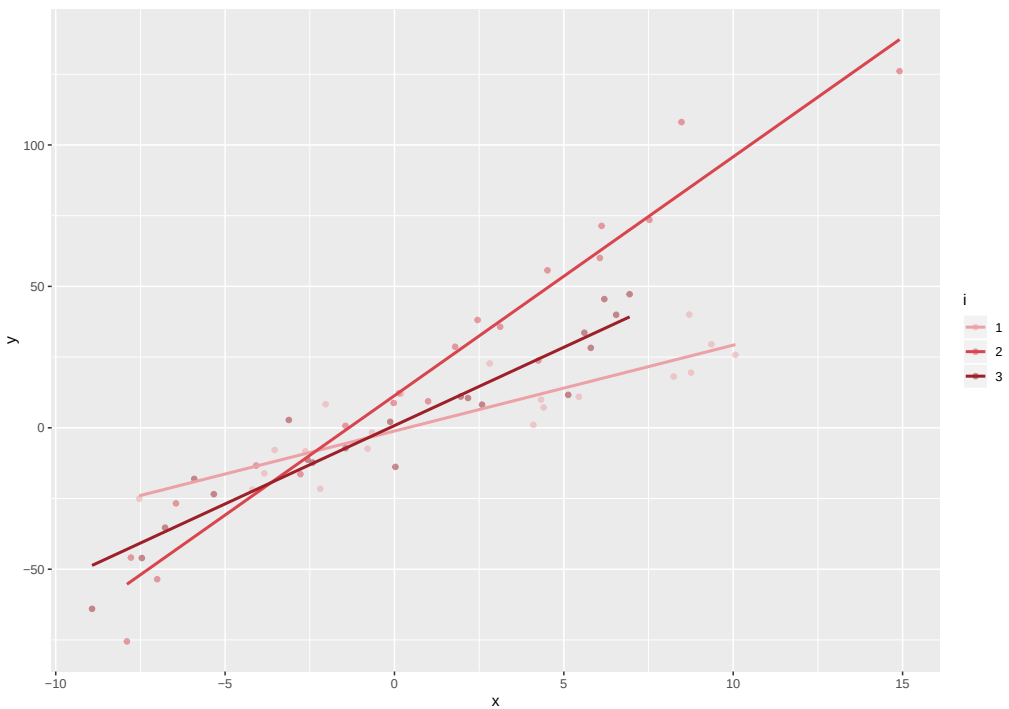
Le modèle d'analyse de la covariance s'écrit:

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Vision graphique

```
ggplot(data = dta, aes(x=x, y=y, col = i)) +
  geom_point( alpha = 0.5) +
  geom_smooth(data = dta, method = 'lm', formula = y ~ x, se = FALS
  scale_fill_manual(values = anova_colour)  +
  scale_colour_manual(values = anova_colour)
```



Plan

Estimation par maximum de vraisemblance

Si $(\boldsymbol{X}'\boldsymbol{X})$ est inversible

$$\hat{\theta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- Attention dans le cas de l'analyse de la variance
- Attention dans le cas de variable colinéaires

Plan

Test sur les paramètres

Vers la décomposition de la variance

Test de modèles emboîtés

Table de décomposition de la variance

Sélection automatique de variables

Diagnostics

Plan

Plan

Test de comparaison de 2 moyennes

Question : Les poids des poulpes mâles et femelles sont-ils égaux ?

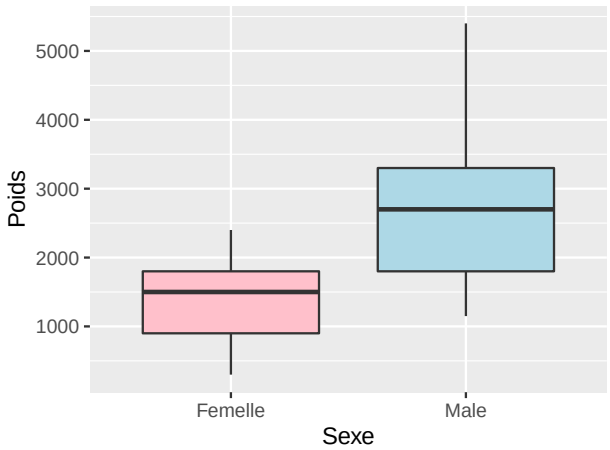
Importons et visualisons les données:

```
poulpe <- read.table("https://r-stat-sc-donnees.github.io/poulpe.cs
summary(poulpe)
```

```
##      Poids           Sexe
##  Min.   : 300   Femelle:13
##  1st Qu.:1480   Male   :15
##  Median :1800
##  Mean   :2099
##  3rd Qu.:2750
##  Max.   :5400
```

Visualisation des données

```
library(ggplot2)
poulpe %>% ggplot() + aes(x=Sexe,y=Poids) + geom_boxplot(fill=c("pi
```



Pour un graphe interactif en html:

```
library(plotly)
poulpe %>% ggplot() + aes(x=Sexe,y=Poids) + geom_boxplot(fill=c("pi
ggplotly()
```

Avec les lignes de code R:

```
boxplot(Poids ~ Sexe, col=c("pink","lightblue"), data=poulpe)
```

Comparaison de 2 moyennes: test de la normalité

A-t-on bien la normalité des poids pour les mâles et femelles ?

```
by(poulpe$Poids, poulpe$Sexe, shapiro.test)
```

```
## poulpe$Sexe: Femelle
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.97109, p-value = 0.9069
##
## ------------------------------------------------------------
## poulpe$Sexe: Male
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.93501, p-value = 0.3238
```

On accepte l'hypothèse de normalité des poids pour les femelles, et pour les mâles

Comparaison de 2 moyennes : test d'égalite des variances

Quel test utiliser ? Celui avec variances égales ou inégales ?

```
var.test(Poids ~ Sexe, conf.level=.95, data=poulpe)
```

```
##
##  F test to compare two variances
##
## data:  Poids by Sexe
## F = 0.28833, num df = 12, denom df = 14, p-value = 0.03713
## alternative hypothesis: true ratio of variances is not equal to
## 95 percent confidence interval:
##  0.09452959 0.92444666
## sample estimates:
## ratio of variances
##          0.2883299
```

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

Test de comparaison de 2 moyennes (suite et fin)

```
res <- t.test(Poids~Sexe, alternative="two.sided", conf.level=.95,
              var.equal=FALSE, data=poulpe)
res

##
##  Welch Two Sample t-test
##
## data:  Poids by Sexe
## t = -3.7496, df = 22.021, p-value = 0.001107
## alternative hypothesis: true difference in means is not equal to
## 95 percent confidence interval:
##  -2010.624  -578.607
## sample estimates:
## mean in group Femelle    mean in group Male
##              1405.385              2700.000
```

On considère que les poids moyennes des mâles et femelles sont différents

Les mâles sont plus lourds (2700) que les femelles (1405.4)

Plan

Problématique et données

Question : Peut-on prévoir le maximum d'ozone en fonction de données climatiques (température, nébulosité, vitesse du vent, max d'ozone de la veille) ?

Importons et visualisons les données:

```
ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt"
library(tidyverse)
ozone.m <- ozone %>% select(1:11)
ozone.m %>% select(1:4) %>% summary()
```

```
##      maxO3             T9              T12             T15
##  Min.   : 42.00   Min.   :11.30   Min.   :14.00   Min.   :14.90
##  1st Qu.: 70.75   1st Qu.:16.20   1st Qu.:18.60   1st Qu.:19.27
##  Median : 81.50   Median :17.80   Median :20.55   Median :22.05
##  Mean   : 90.30   Mean   :18.36   Mean   :21.53   Mean   :22.63
##  3rd Qu.:106.00   3rd Qu.:19.93   3rd Qu.:23.55   3rd Qu.:25.40
##  Max.   :166.00   Max.   :27.00   Max.   :33.50   Max.   :35.50
```

Avec les lignes de code R:

```
ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt"
```

Visualisation des liaisons par paires de variables

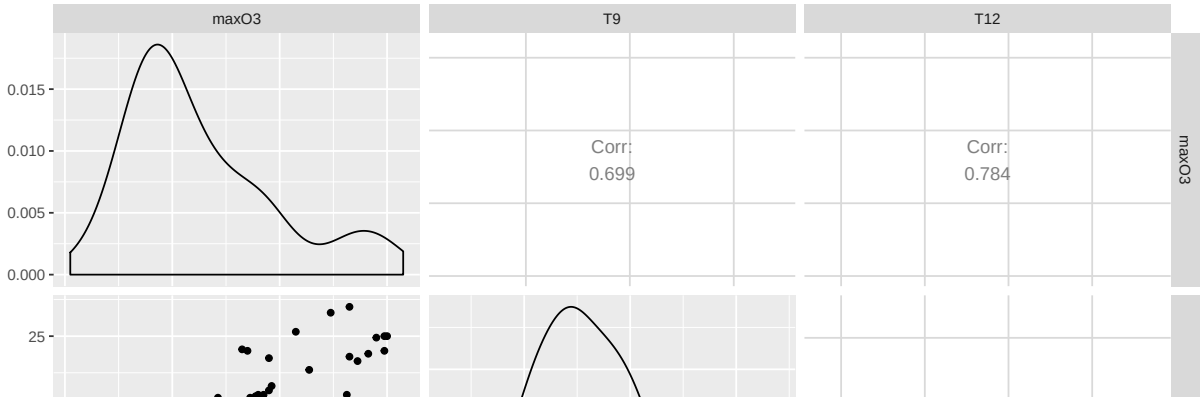
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ozone.m %>% select(1:3) %>% ggpairs()
```



Construction du modèle complet

```
reg.mul <- lm(maxO3~., data=ozone.m)
summary(reg.mul)

## Call:
## lm(formula = maxO3 ~ ., data = ozone.m)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.24442   13.47190   0.909   0.3656
## T9          -0.01901    1.12515  -0.017   0.9866
## T12          2.22115    1.43294   1.550   0.1243
## T15          0.55853    1.14464   0.488   0.6266
## Ne9         -2.18909    0.93824  -2.333   0.0216 *
## Ne12        -0.42102    1.36766  -0.308   0.7588
## Ne15         0.18373    1.00279   0.183   0.8550
## Vx9          0.94791    0.91228   1.039   0.3013
## Vx12         0.03120    1.05523   0.030   0.9765
## Vx15         0.41859    0.91568   0.457   0.6486
## maxO3v       0.35198    0.06289   5.597 1.88e-07 ***
##
```

Sélection de variables

```
library(FactoMineR)
select <- RegBest(ozone.m$maxO3, ozone.m[,2:11])
select$summary ; select$best
```

```
##                                    R2       Pvalue
## Model with 1 variable   0.6150674 1.512025e-24
## Model with 2 variables  0.7012408 2.541031e-29
## Model with 3 variables  0.7519764 1.457692e-32
## Model with 4 variables  0.7622198 1.763434e-32
## Model with 5 variables  0.7630603 1.449905e-31
## Model with 6 variables  0.7635768 1.130263e-30
## Model with 7 variables  0.7637610 8.556709e-30
## Model with 8 variables  0.7638390 6.076804e-29
## Model with 9 variables  0.7638407 4.066941e-28
## Model with 10 variables 0.7638413 2.545665e-27
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.76225   11.10038   0.879    0.381
```

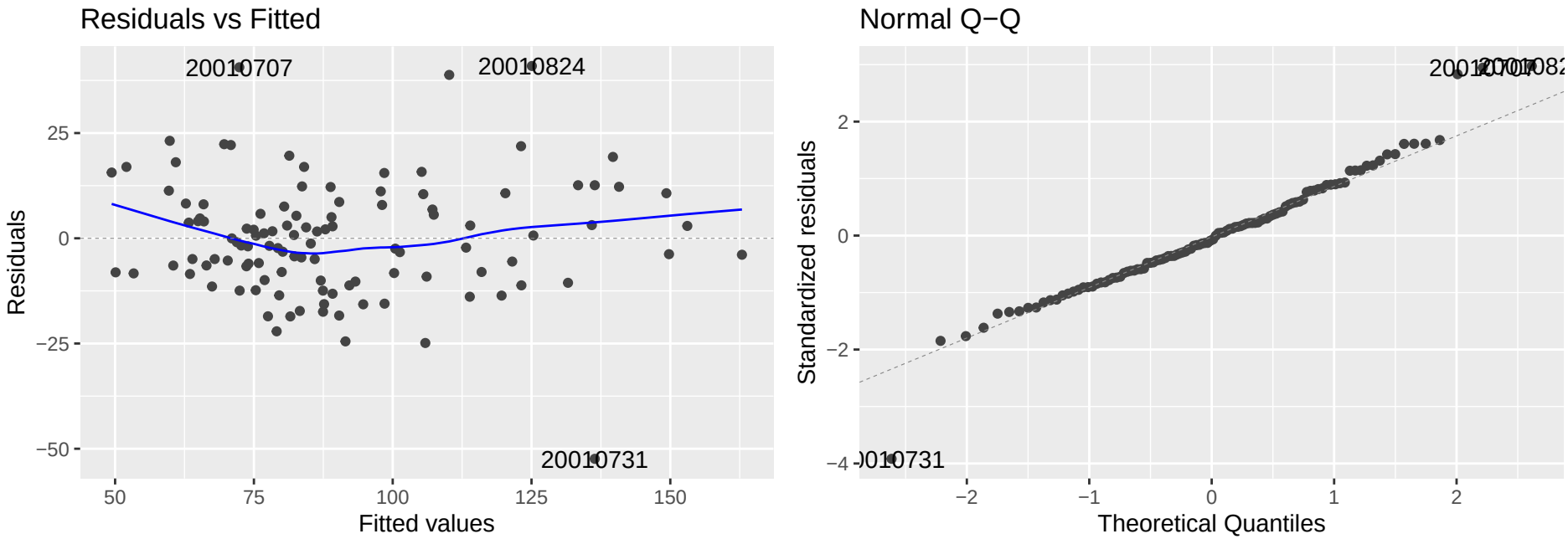
Construction du modèle final

```
reg.fin <- lm(maxO3~T12+Ne9+Vx9+maxO3v, data=ozone.m)
summary(reg.fin)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12 + Ne9 + Vx9 + maxO3v, data = ozone.m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.396  -8.377  -1.086   7.951  40.933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.63131   11.00088   1.148 0.253443
## T12          2.76409    0.47450   5.825 6.07e-08 ***
## Ne9         -2.51540    0.67585  -3.722 0.000317 ***
## Vx9          1.29286    0.60218   2.147 0.034055 *
## maxO3v       0.35483    0.05789   6.130 1.50e-08 ***
## ---
```

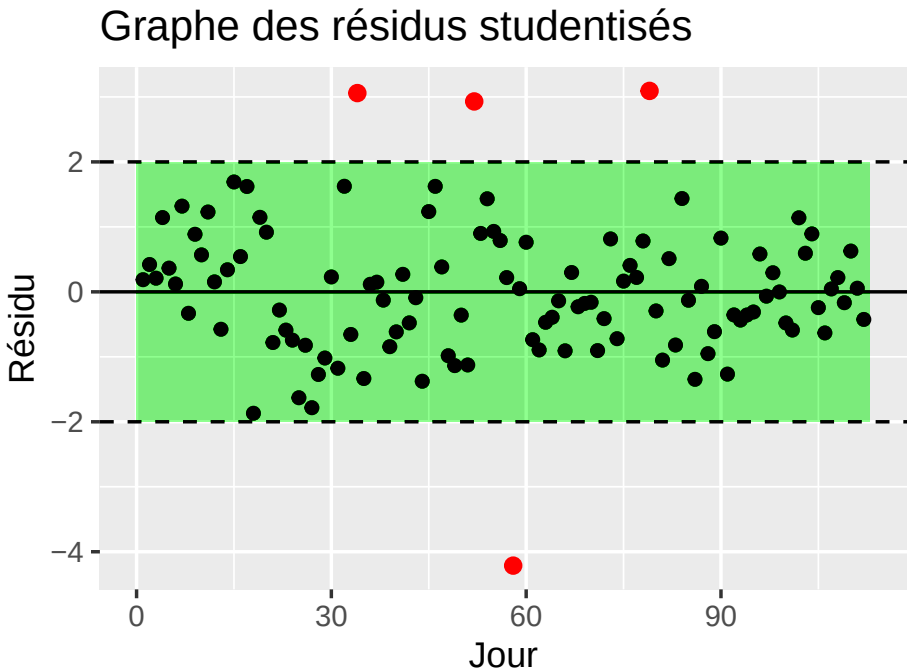
Analyser les résidus

```
library(ggfortify)
autoplot(reg.fin)
```



Analyser les résidus (suite)

```
residutib <- tibble(jour = 1:112, residu = rstudent(reg.fin))
residutib %>% ggplot() + aes(x=jour, y=residu) + geom_point() +
  labs(x="Jour", y="Résidu", title = "Graphe des résidus studentisé
  geom_abline(slope=0, intercept=c(-2,0,2), linetype=c(2,1,2)) +
  geom_rect(aes(xmin=0, xmax=113, ymin=-2, ymax=2), alpha=0.002,fil
  geom_point(data = residutib %>% filter(abs(residu)>2), cex=2, col
```



Avec les lignes de code R :

Prévoir une nouvelle valeur

Et comment prédire le maximum d'ozone pour de nouvelles valeurs ?

```
xnew <- matrix(c(19,8,2.05,70),nrow=1)
colnames(xnew) <- c("T12","Ne9","Vx9","maxO3v")
xnew <- as.data.frame(xnew)
predict(reg.fin,xnew,interval="pred")
```

```
##        fit      lwr      upr
## 1 72.51437 43.80638 101.2224
```

Plan

Problématique et données

Question : Y a-t-il un effet de la pluie et du vent sur le maximum d'ozone ?
Y a-t-il un effet de l'interaction de ces deux facteurs ?

Importation des données:

```
ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt"
summary(ozone[,c("maxO3","vent","pluie")])
```

```
##       maxO3             vent       pluie
##  Min.   : 42.00   Est  :10   Pluie:43
##  1st Qu.: 70.75   Nord :31   Sec  :69
##  Median : 81.50   Ouest:50
##  Mean   : 90.30   Sud  :21
##  3rd Qu.:106.00
##  Max.   :166.00
```

Le modèle d'analyse de variance

```
library('tidyverse')
```

Le modèle d'analyse de variance à 2 facteurs s'écrit :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

avec $\varepsilon_{ijk} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ et $cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \quad \forall\ (i,j,k) \neq (i',j',k')$

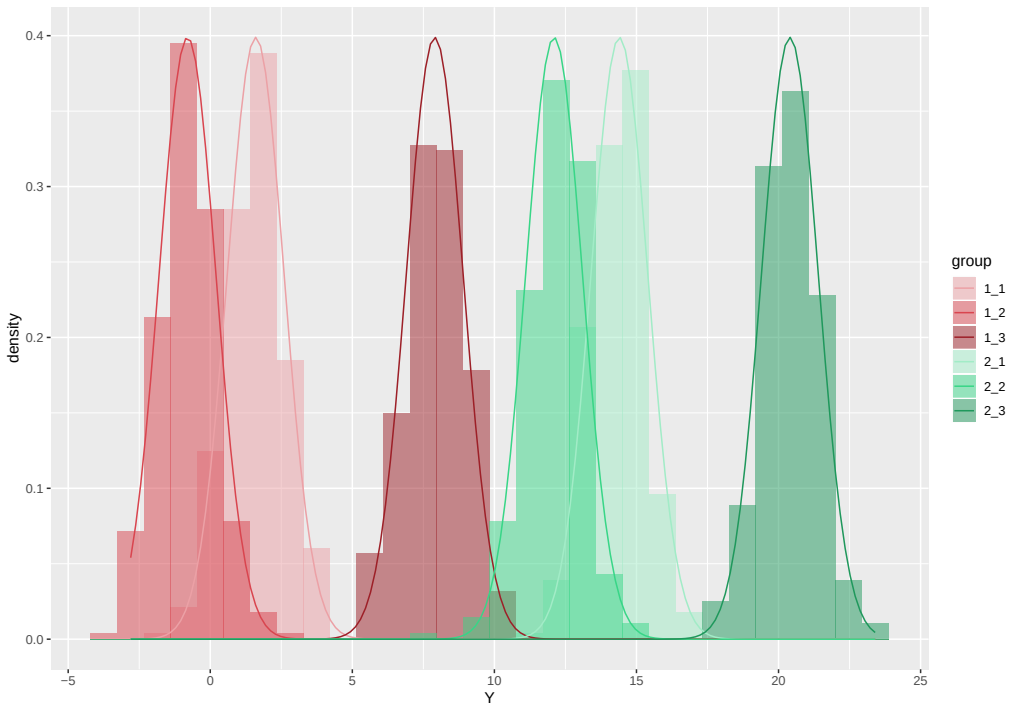
- μ effet de référence
- α_i effet différentiel du niveau i du premier facteur
- β_j effet différentiel du niveau j du second facteur
- $(\alpha\beta)_{ij}$ effet différentiel de l'interaction des niveaux i et j

Objectif de l'analyse de variance : étudier parmi ces effets ceux qui influent sur Y

Test des effets

Décomposition de la variabilité

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```




```
## Saving 10 x 7 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```

Test des effets

Différents calculs de variances

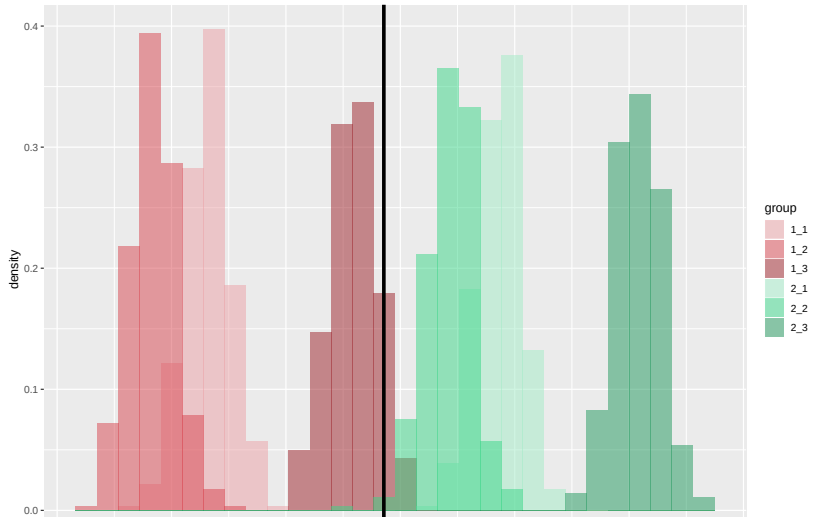
La **variabilité totale** est définie par

$$RSS_{M0} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(y_{ijk} - y_{\bullet\bullet\bullet})^2.$$

Correspond à la variabilité résiduelle dans le modèle

$$M_0 : \quad Y_{ijk} = \mu + \varepsilon_{ijk}$$

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```



Test des effets

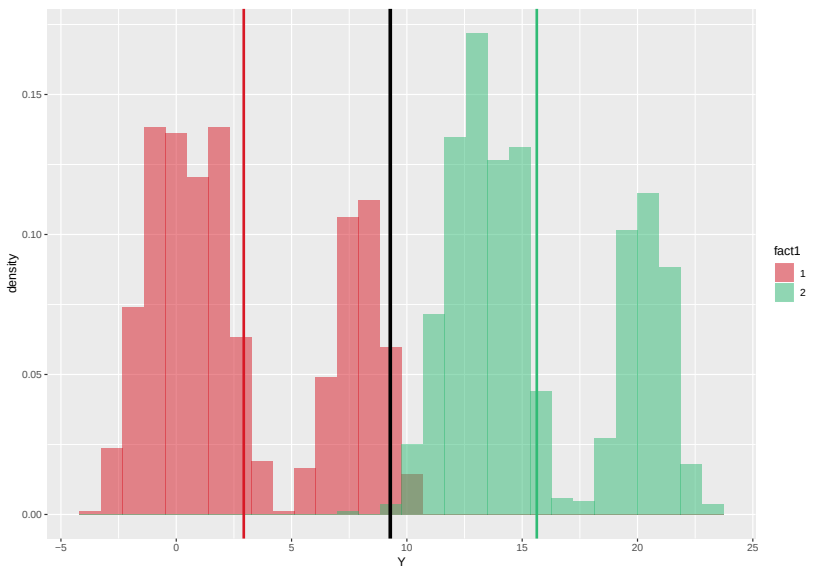
La **variabilité attribuée au facteur 1** *peut être* définie par

$$RSS_{M0} - RSS_{M1},$$

avec

$$RSS_{M1} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (y_{ijk} - y_{i\bullet\bullet})^2.$$

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```



Test des effets

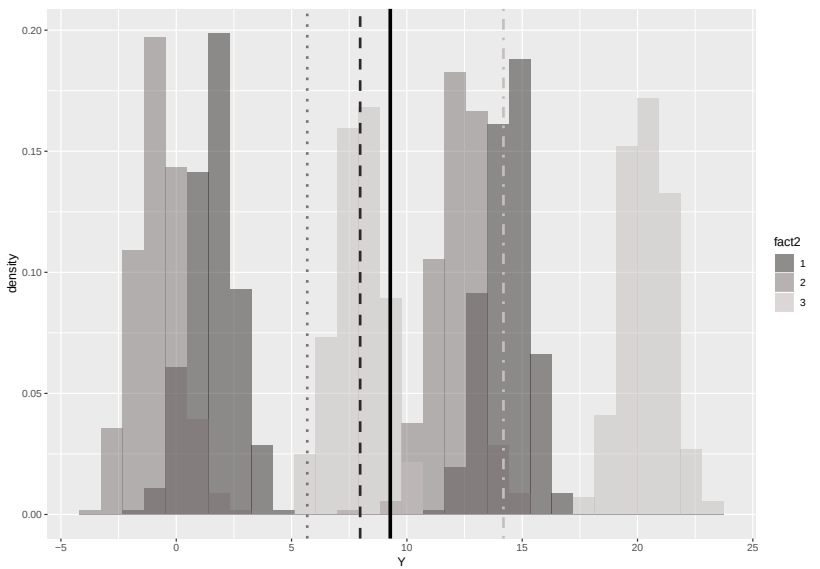
La **variabilité attribuée au facteur 2** *peut être* définie par

$$RSS_{M0} - RSS_{M2},$$

avec

$$RSS_{M2} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (y_{ijk} - y_{\bullet j \bullet})^2.$$

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth
```



Test des effets

Les effets sont testés grâce à l'analyse de la variabilité

```
anova_comp <- lm(Y ~ fact1 + fact2 + fact1:fact2, data = dta )

anova(anova_comp)
```

```
## Analysis of Variance Table
##
## Response: Y
##               Df Sum Sq Mean Sq   F value    Pr(>F)
## fact1          1  72641   72641 69888.232 < 2.2e-16 ***
## fact2          2  23243   11621 11181.110 < 2.2e-16 ***
## fact1:fact2    2     26      13    12.316 4.868e-06 ***
## Residuals   1794   1865       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test des effets

Table d'analyse de la variance (`anova`)

Source	Mesure	H_0
Fact1	$RSS_{M_0} - RSS_{M_1}$	M_0 et M_1 sont équivalents
Fact2	$RSS_{M_1} - RSS_{M_{12}}$	M_1 et M_{12} sont équivalents
Interaction	$RSS_{M_12} - RSS_{M_{comp}}$	M_{12} et M_{comp} sont équivalents

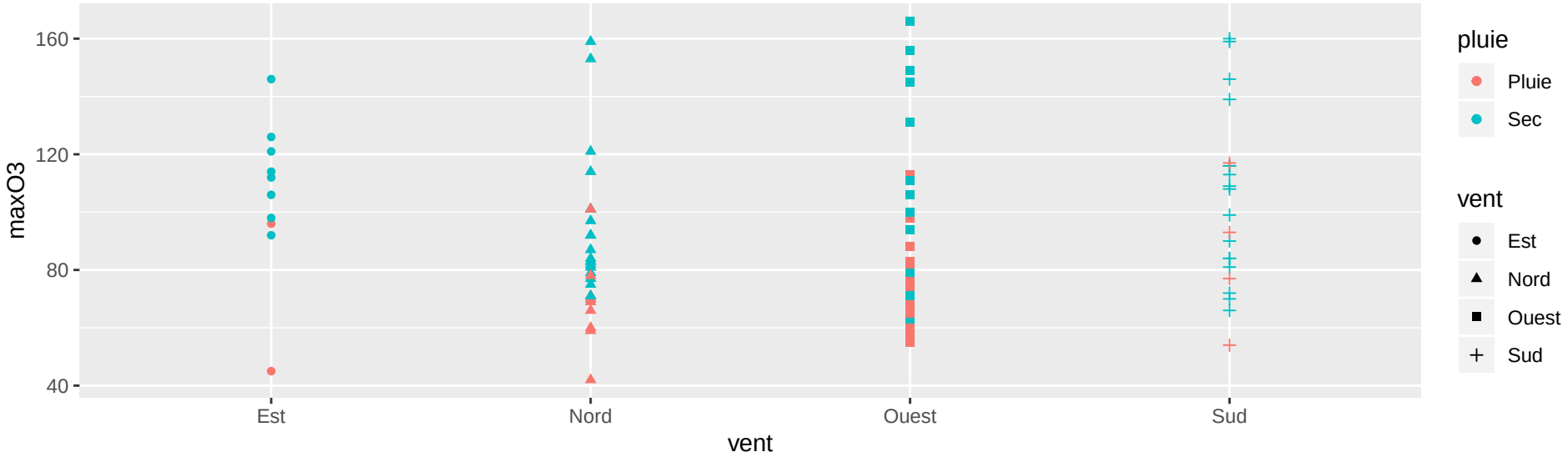
Test des effets

Table d'analyse de la variance (`Anova`)

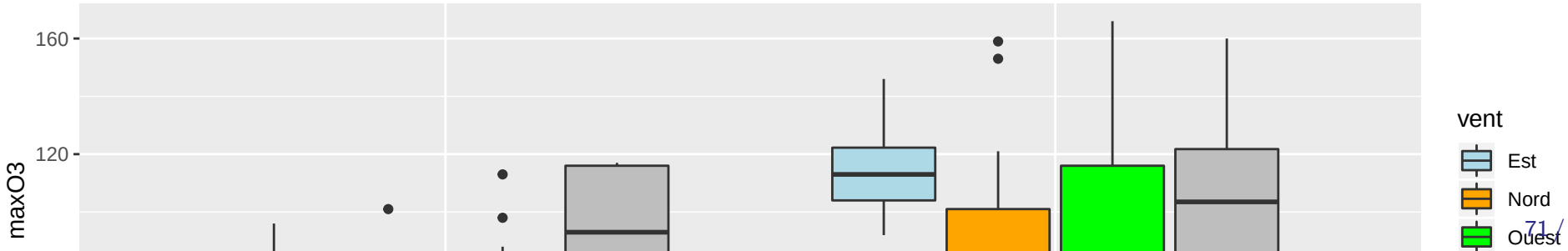
Source	Mesure	H_0
Fact1	$RSS_{M_2} - -RSS_{M_{12}}$	M_2 et M_{12} sont équivalents
Fact2	$RSS_{M_1} - RSS_{M_{12}}$	M_1 et M_{12} sont équivalents
Interaction	$RSS_{M_12} - RSS_{M_{comp}}$	M_{12} et M_{comp} sont équivalents

Visualisation des données avec ggplot2

```
library(ggplot2)
ozone %>% ggplot() + aes(y=maxO3, x=vent) + geom_point(aes(col=plui
```

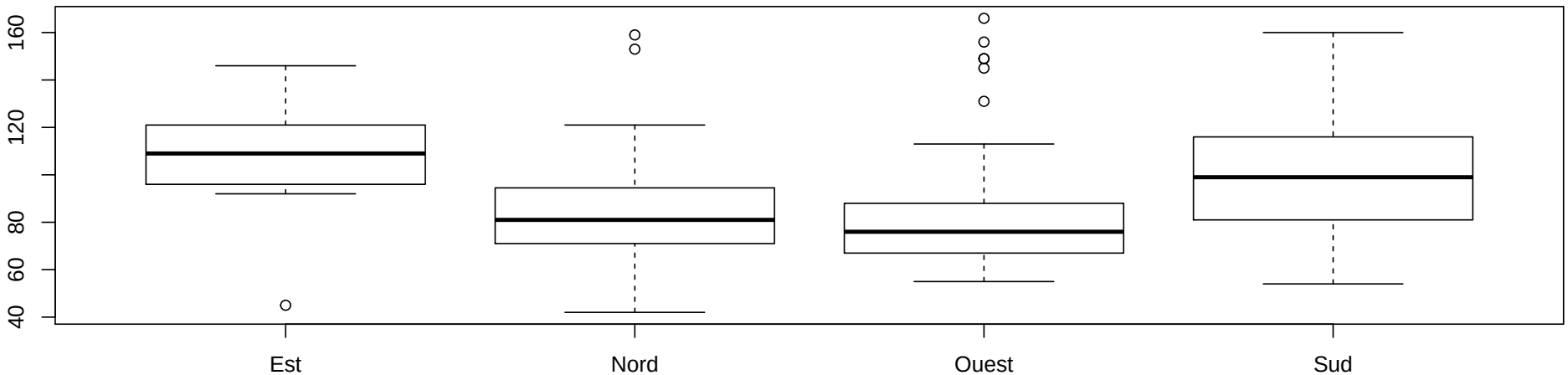



```
ozone %>% ggplot() + aes(pluie, maxO3) + geom_boxplot(aes(fill=vent
  scale_fill_manual(values=c("lightblue","orange","green","grey"))
```

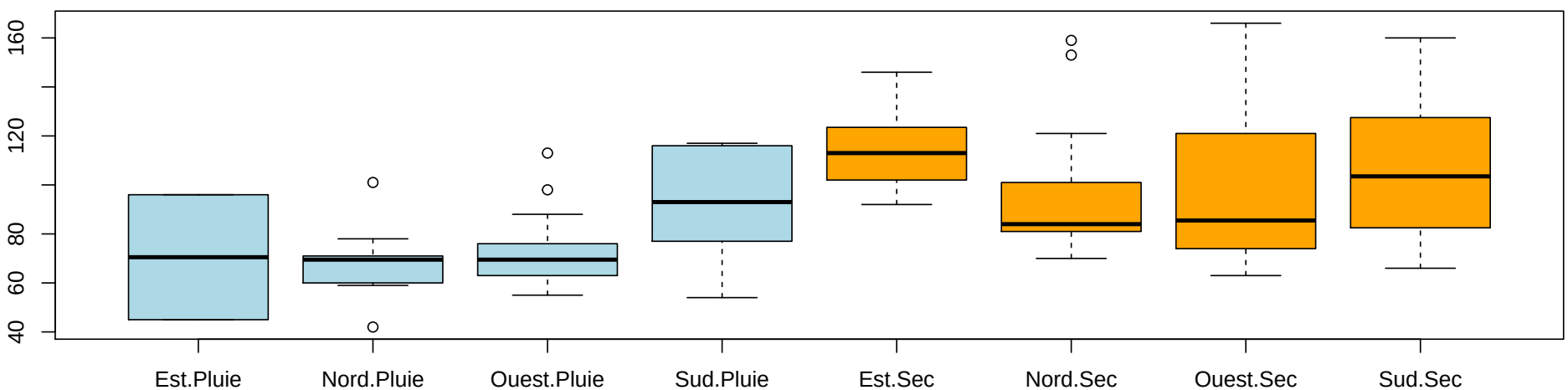


Visualisation des données en R

```
boxplot(maxO3~vent, data = ozone)
```

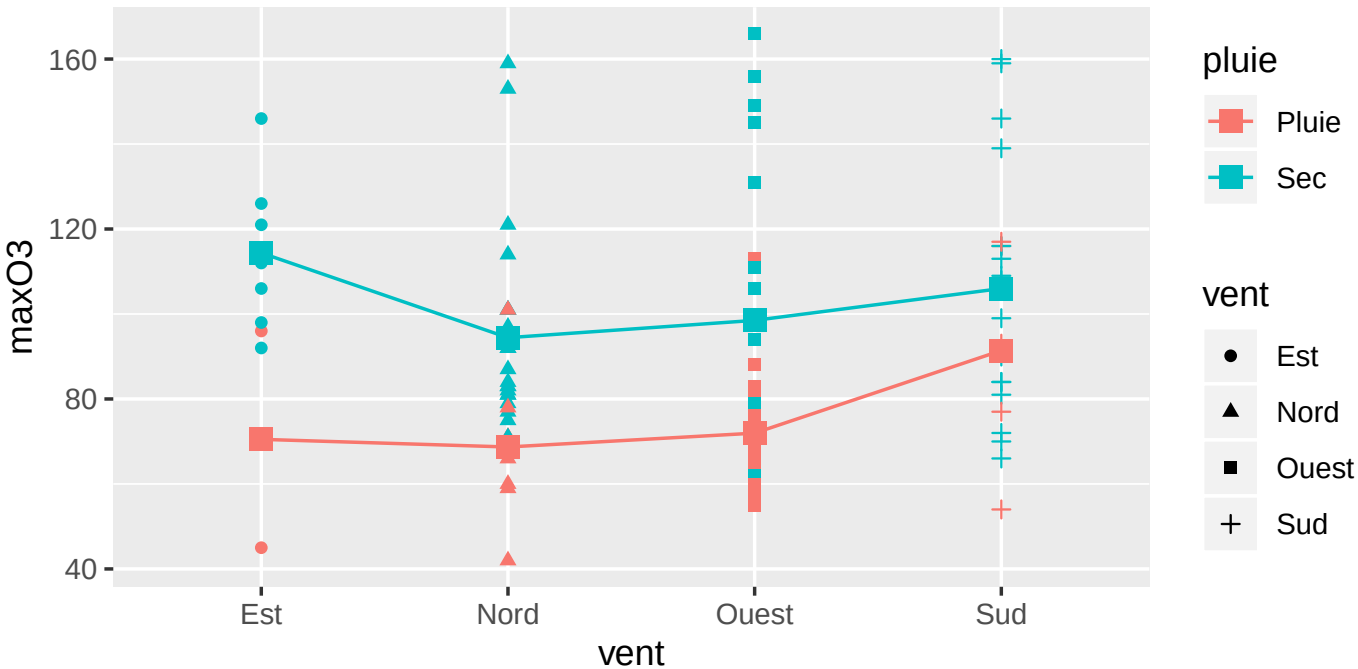



```
boxplot(maxO3~vent*pluie, data = ozone, col=c(rep("Lightblue",4),re
```



Visualisation de l'interaction

```
ozone %>% ggplot() + aes(x = vent, y = maxO3, group = pluie) +
  geom_point(aes(color = pluie, shape=vent)) +
  stat_summary(fun.y = mean, geom = "point", size=3, shape=15,aes(c
  stat_summary(fun.y = mean, geom = "line", aes(color = pluie))
```

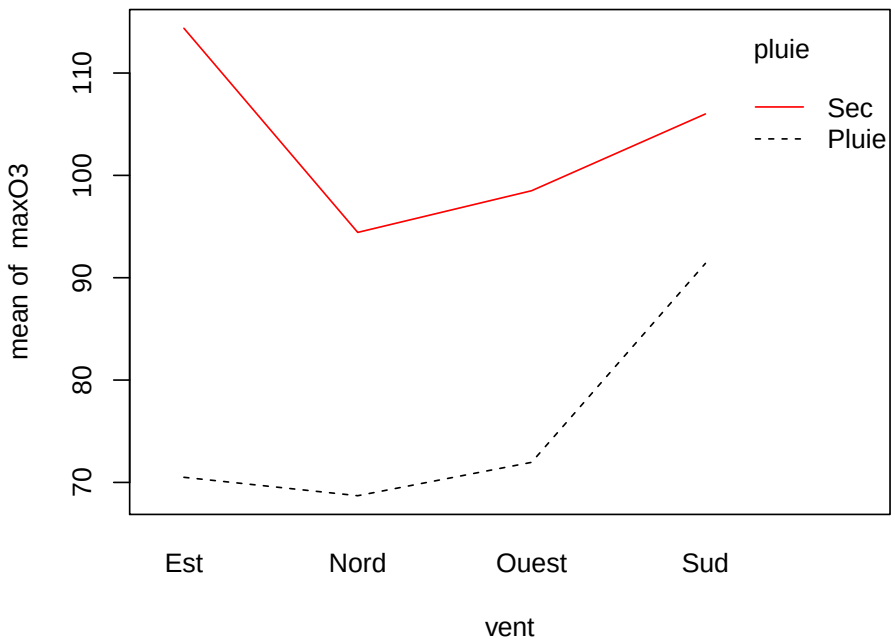


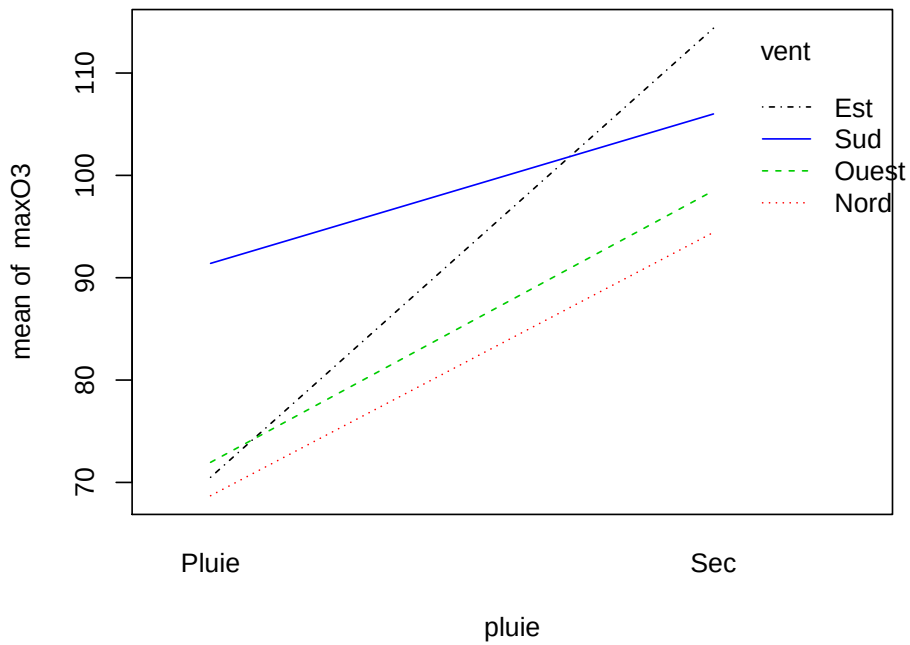
Visualiser l'autre graphe d'interaction (une ligne brisée par direction du vent) et conserver le graphe le plus explicite

```
ozone %>%  ggplot() + aes(x = pluie, y = maxO3, group = vent, color
```

Graphe : visualisation de l'interaction

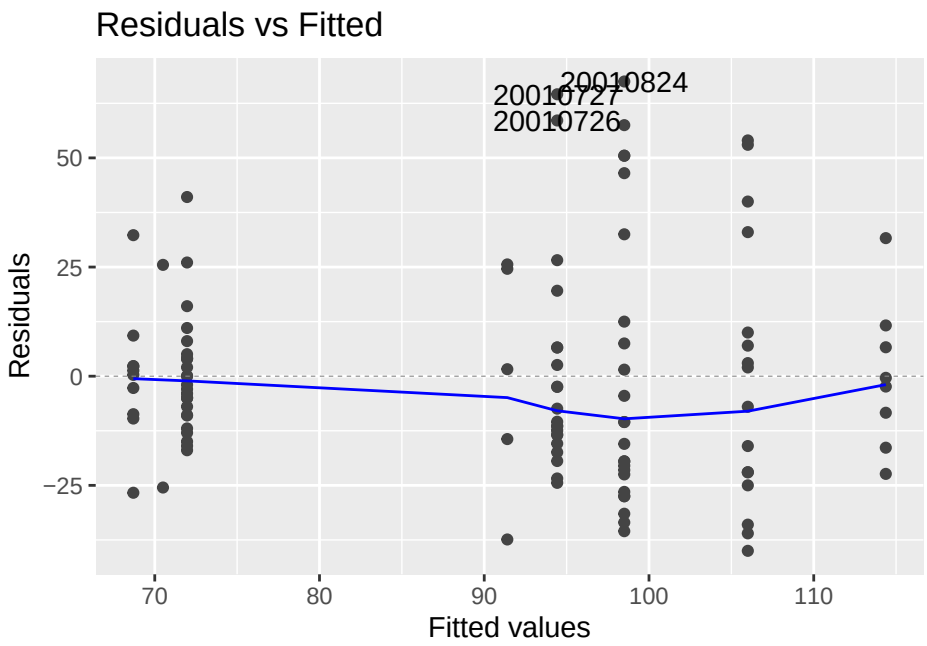
```
with(ozone,interaction.plot(vent,pluie,maxO3,col=1:nlevels(pluie)))
with(ozone,interaction.plot(pluie,vent,maxO3,col=1:nlevels(vent)))
```

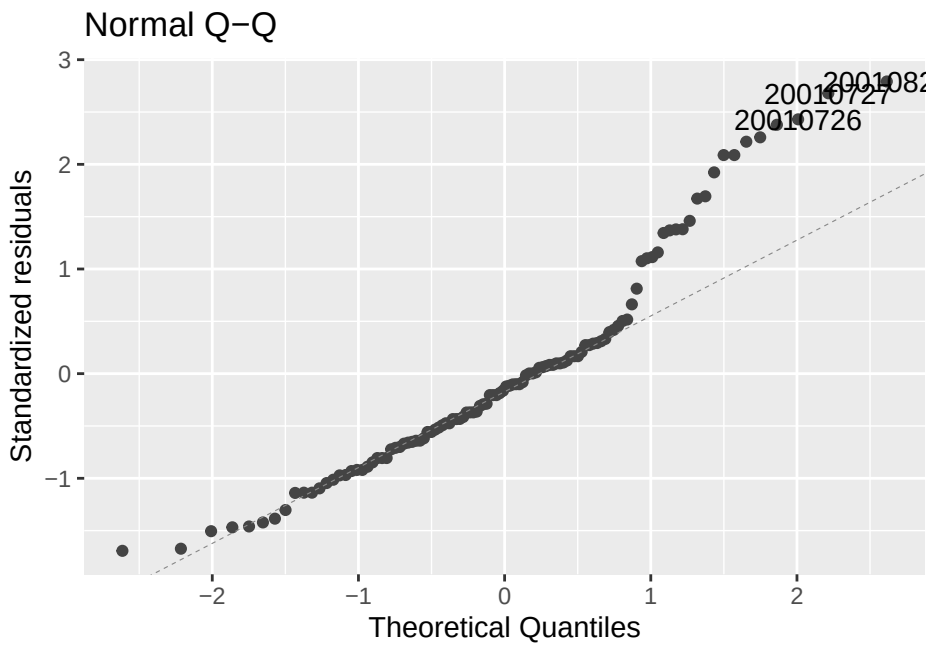


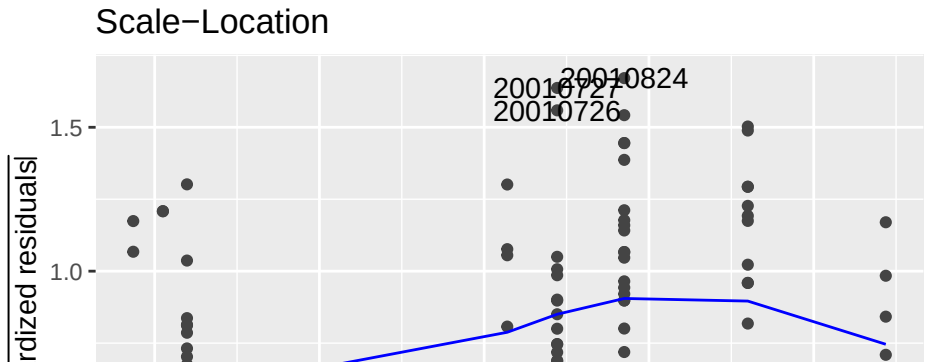


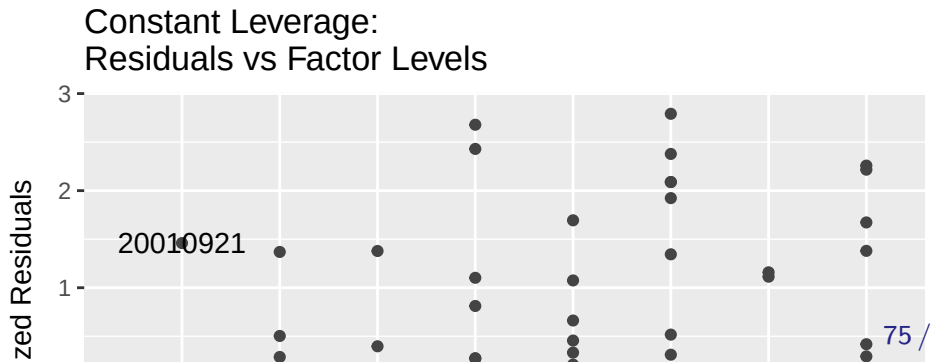
Validité du modèle

```
library(ggfortify)
mod.interaction <- lm(maxO3 ~ vent + pluie + vent:pluie, data=ozone
autoplot(mod.interaction)
```



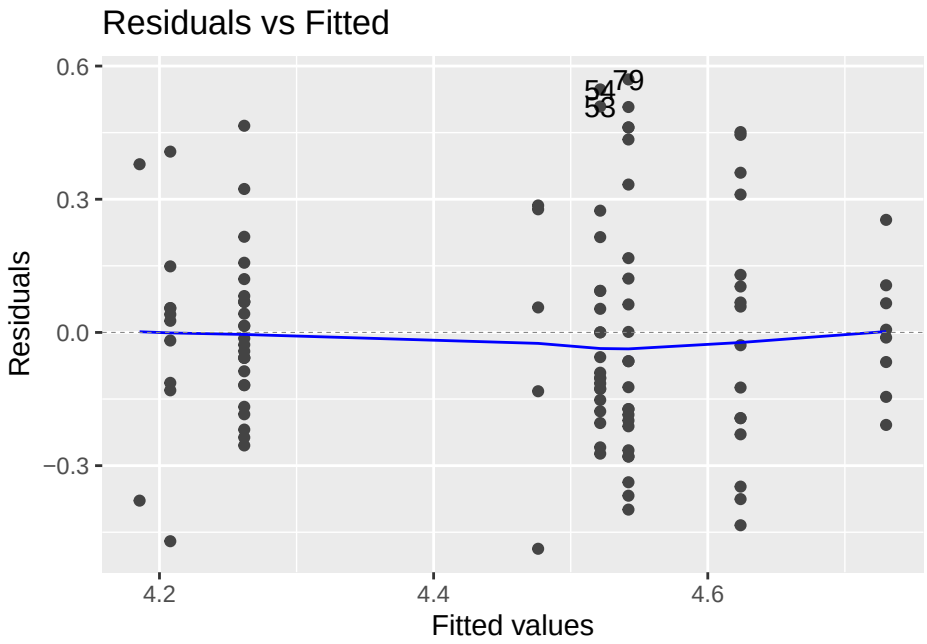


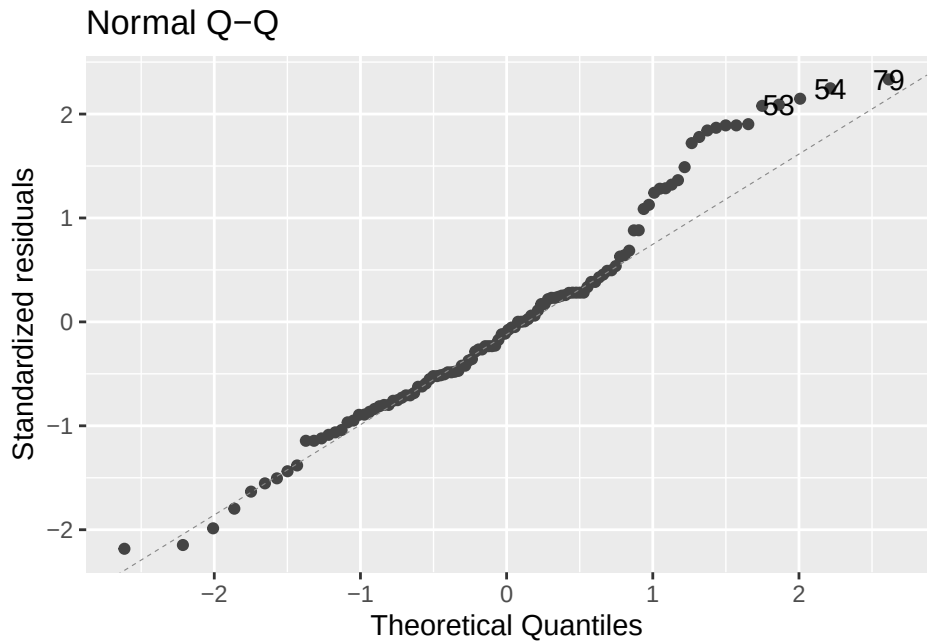


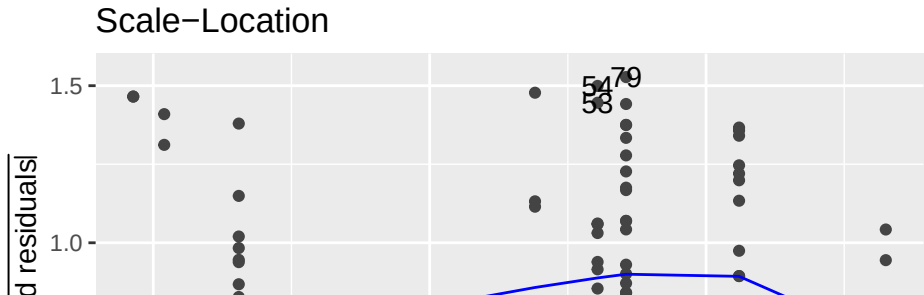


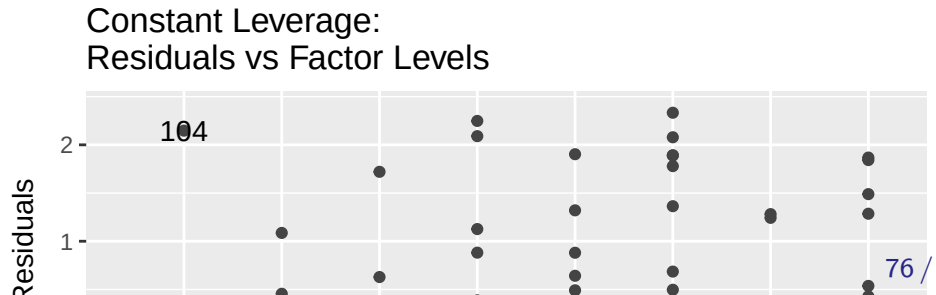
Validité du modèle

```
library(ggfortify)
ozone %>% mutate(log_maxO3 = log(maxO3)) -> ozone
mod.interaction <- lm(log_maxO3 ~ vent + pluie + vent:pluie, data=o
autoplot(mod.interaction)
```









Test du modèle complet

```
mod.interaction <- lm(log_maxO3 ~ vent + pluie + vent:pluie, data=o
mod.0 <- lm(log_maxO3 ~ 1, data=ozone)
anova(mod.0, mod.interaction)

## Analysis of Variance Table
##
## Model 1: log_maxO3 ~ 1
## Model 2: log_maxO3 ~ vent + pluie + vent:pluie
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    111 9.5368
## 2    104 6.4740  7    3.0629 7.029 7.355e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette l'hypothèse qu'il n'existe aucun effet car la probabilité critique (0) est inférieure à 5%

Construction du modèle avec interaction

```
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

anova(mod.interaction)

## Analysis of Variance Table
##
## Response: log_maxO3
##             Df Sum Sq Mean Sq F value    Pr(>F)
## vent         3 0.8588 0.28626  4.5985 0.004603 **
```

Choix d'un sous-modèle

```
modele_12 <- lm(log_maxO3 ~ vent + pluie, data = ozone)
anova(modele_12)
```

```
## Analysis of Variance Table
##
## Response: log_maxO3
##            Df Sum Sq Mean Sq F value    Pr(>F)
## vent        3 0.8588 0.28626  4.5994  0.004555 **
## pluie       1 2.0187 2.01866 32.4346 1.094e-07 ***
## Residuals 107 6.6594 0.06224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

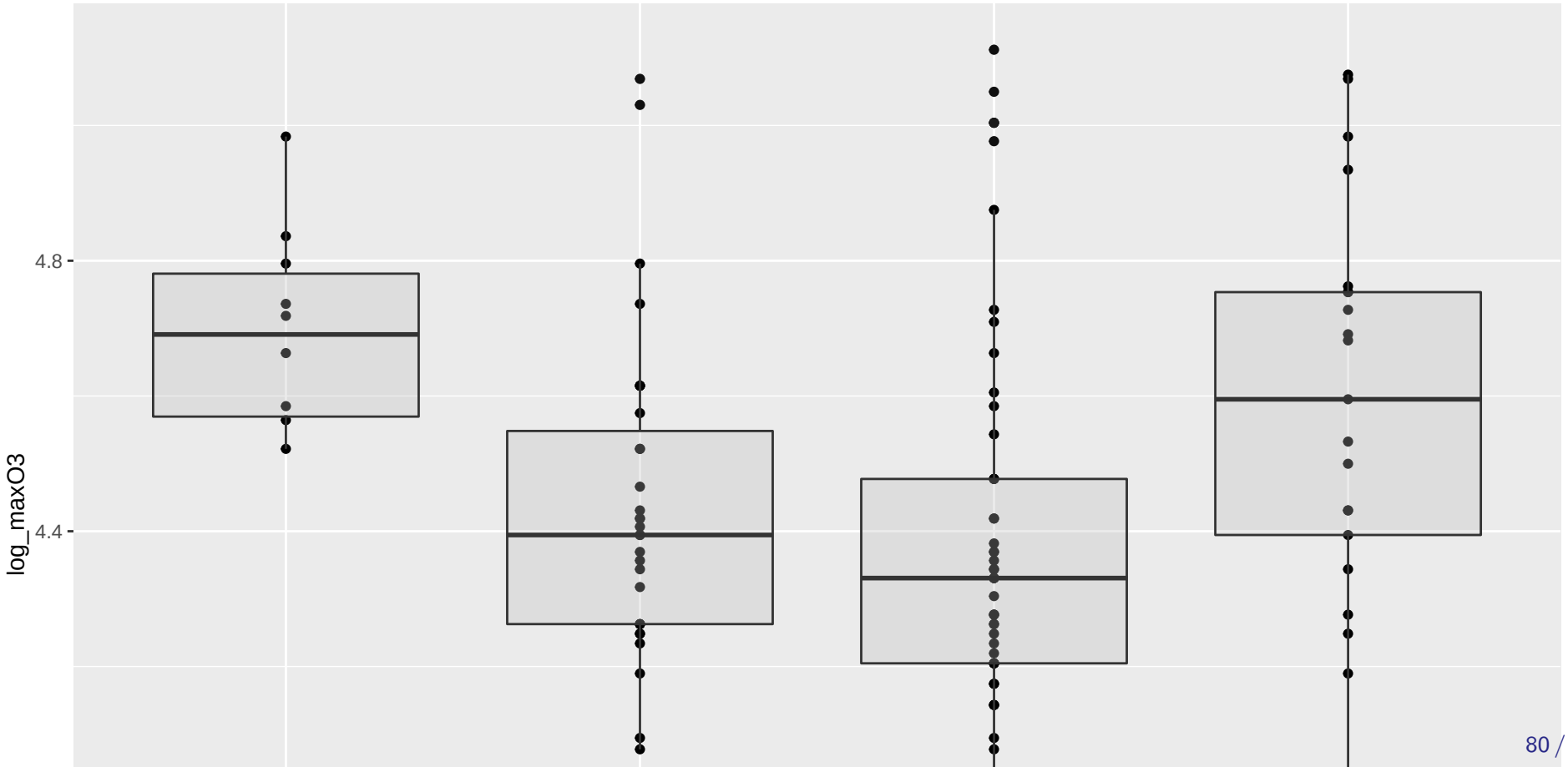
```
Anova(modele_12)
```

```
## Anova Table (Type II tests)
##
## Response: log_maxO3
##           Sum Sq  Df F value    Pr(>F)
## vent      0.3982   3  2.1329    0.1004
```

Qu'est ce que l'effet du vent ??

Visualisation des différences de vent après ajustement à la pluie

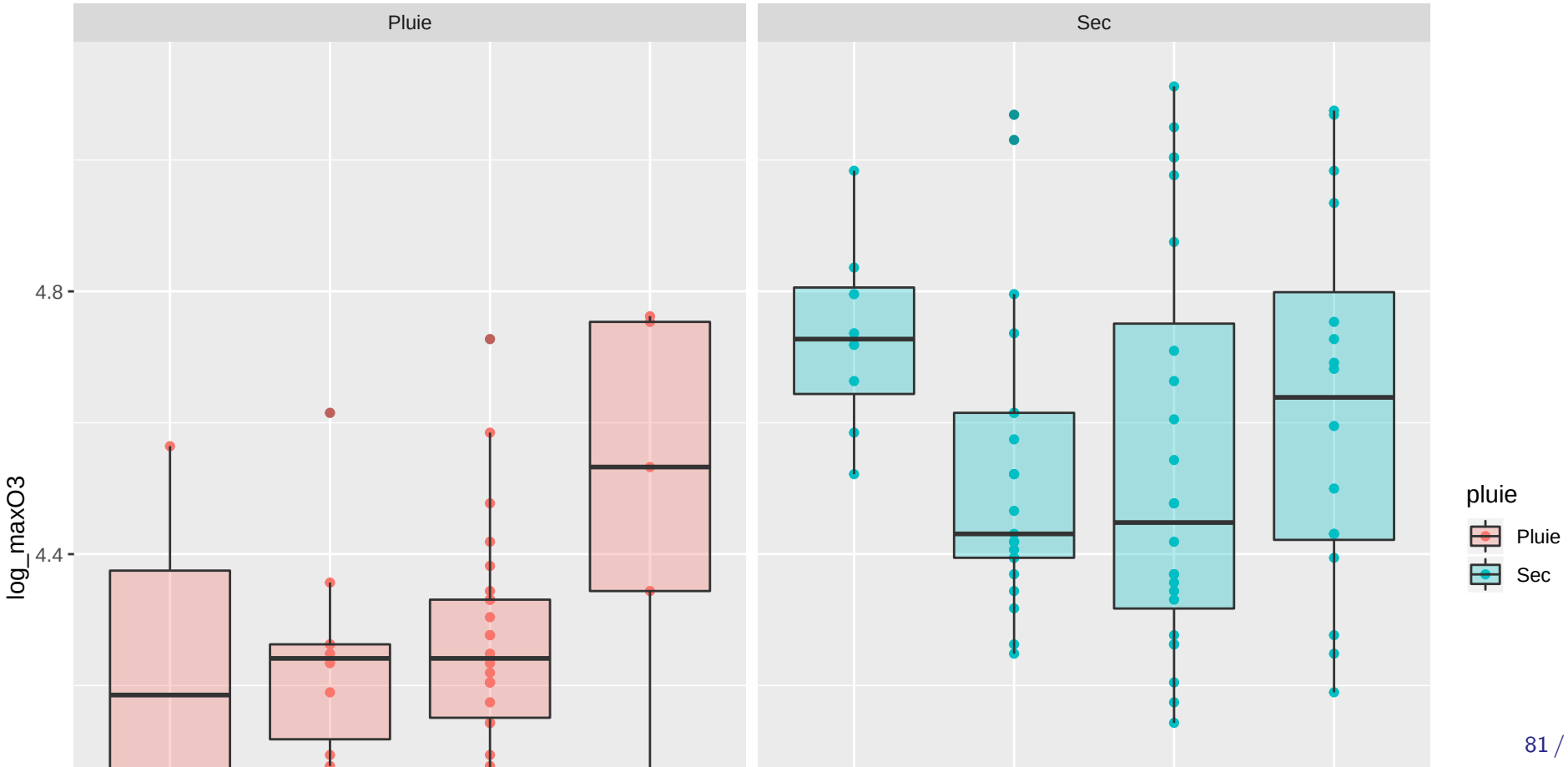
```
ozone %>%   ggplot() +
  geom_point( mapping = aes(x=vent, y=log_maxO3))+
  geom_boxplot( mapping = aes(x=vent, y=log_maxO3), alpha=0.3, fill
```



Qu'est ce que l'effet du vent ??

Visualisation des différences de vent après ajustement à la pluie

```
ozone %>%  ggplot() +  facet_wrap(~pluie)+
  geom_point( mapping = aes(x=vent, y=log_maxO3, col = pluie)) +
  geom_boxplot( mapping = aes(x=vent, y=log_maxO3, fill = pluie), a
```



Estimation des coefficients

Attention à l'interprétation

Dans le modèle complet

```
summary(mod.interaction)
```

```
##
## Call:
## lm(formula = log_maxO3 ~ vent + pluie + vent:pluie, data = ozone
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4873 -0.1727 -0.0230  0.1097  0.5699
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.18551    0.17642  23.724  < 2e-16 ***
## ventNord           0.02239    0.19326   0.116  0.90797
## ventOuest          0.07636    0.18308   0.417  0.67749
## ventSud            0.29073    0.20875   1.393  0.16667
```

Comparaison de moyennes ajustées

```
library('emmeans')
```

```
##
## Attaching package: 'emmeans'
```

```
## The following object is masked from 'package:GGally':
##
##     pigs
```

```
emmeans(modele_12,pairwise~pluie,adjust="hochberg")
```

```
## $emmeans
##  pluie emmean     SE  df lower.CL upper.CL
##  Pluie   4.31 0.0441 107     4.22     4.40
##  Sec     4.60 0.0322 107     4.53     4.66
##
## Results are averaged over the levels of: vent
## Confidence level used: 0.95
##
## $contrasts
```

Références

Daudin, J.-J. (2015). *Le modèle linéaire et ses extensions - Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (Niveau C)* (p. 336 p.). Edition Ellipses.

Faraway, J. J. (2016). *Extending the linear model with r: Generalized linear, mixed effects and nonparametric regression models*. Chapman; Hall/CRC.

Kirwan Laura, Connolly John, Brophy Caroline, Baadshaug Ole, Belanger Gilles, Black Alistair, ... Finn John. (2014). The Agrodiversity Experiment: three years of data from a multisite study in intensively managed grasslands. *Ecology*, *95*, 2680.