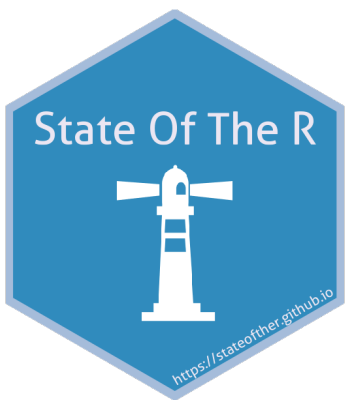
Statistical modelling for biological data with R

Day 2-3 : Linear model with R

Marie-Pierre Etienne

https://marieetienne.github.io

Novembre 2019



Plan

Plan

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 :

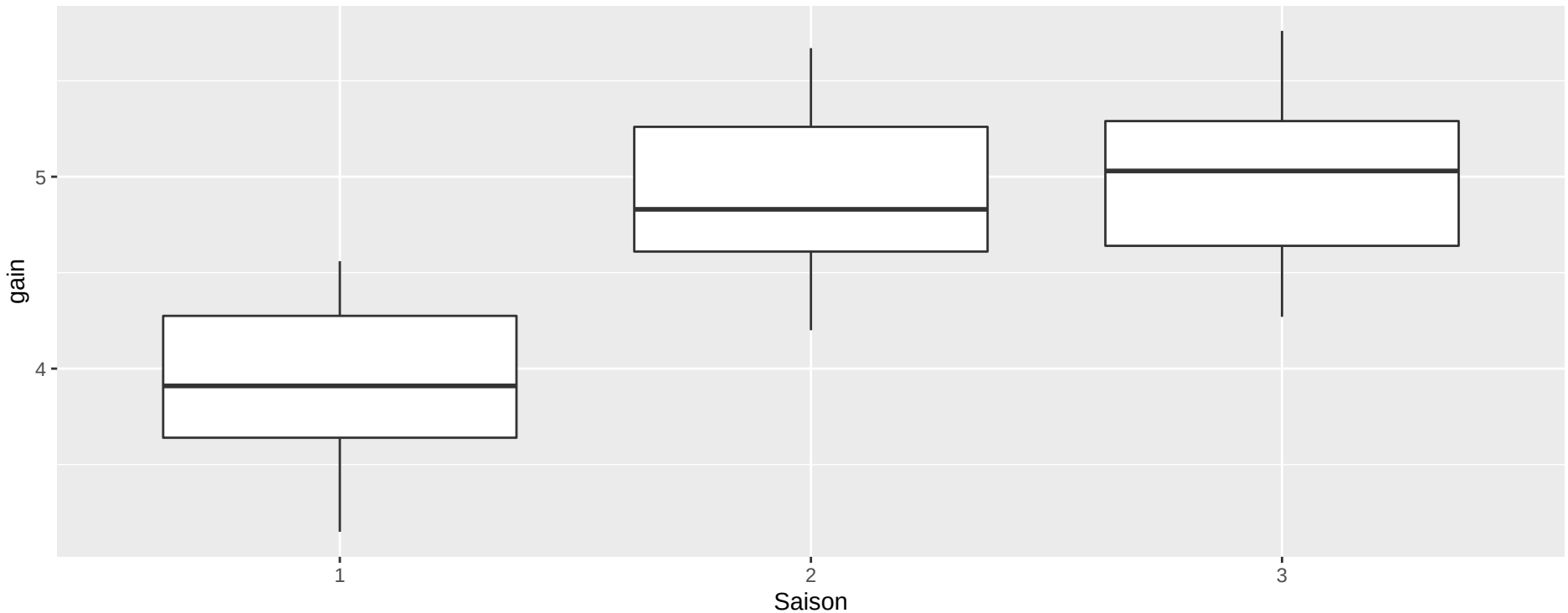
- poids initial,
- poids au retour
- GPS (oui ou non)
- Période de suivi (3 périodes).

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

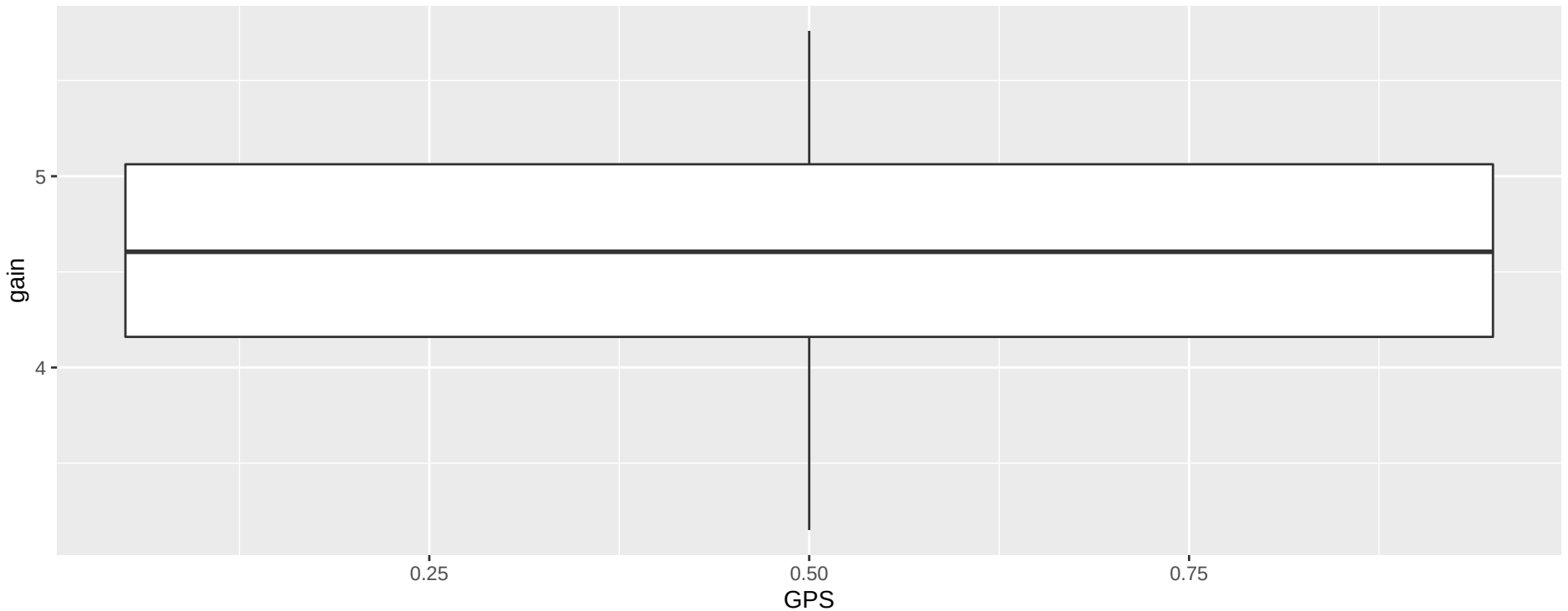
```
manchots %>% ggplot() + geom_boxplot(aes(x= Saison, y=gain))
```



Des représentation graphiques

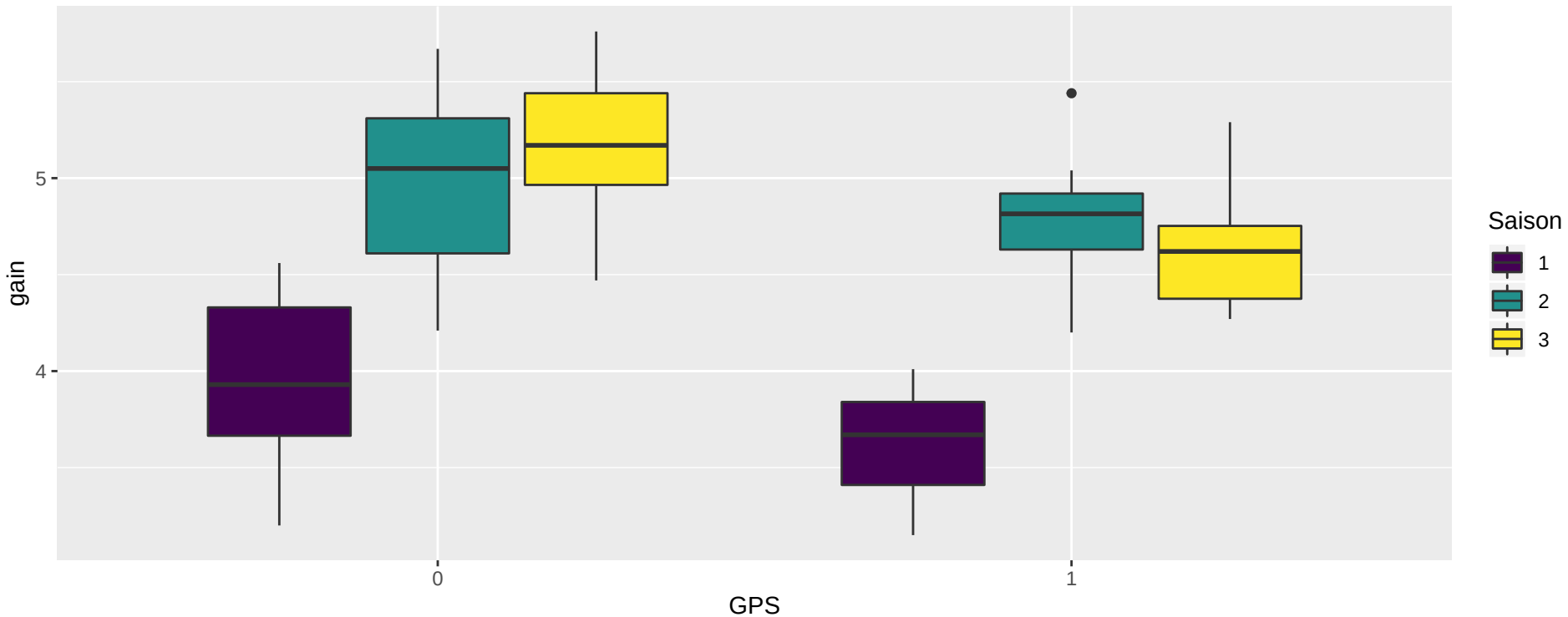
```
manchots %>% ggplot() + geom_boxplot(aes(x= GPS, y=gain))
```

```
## 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

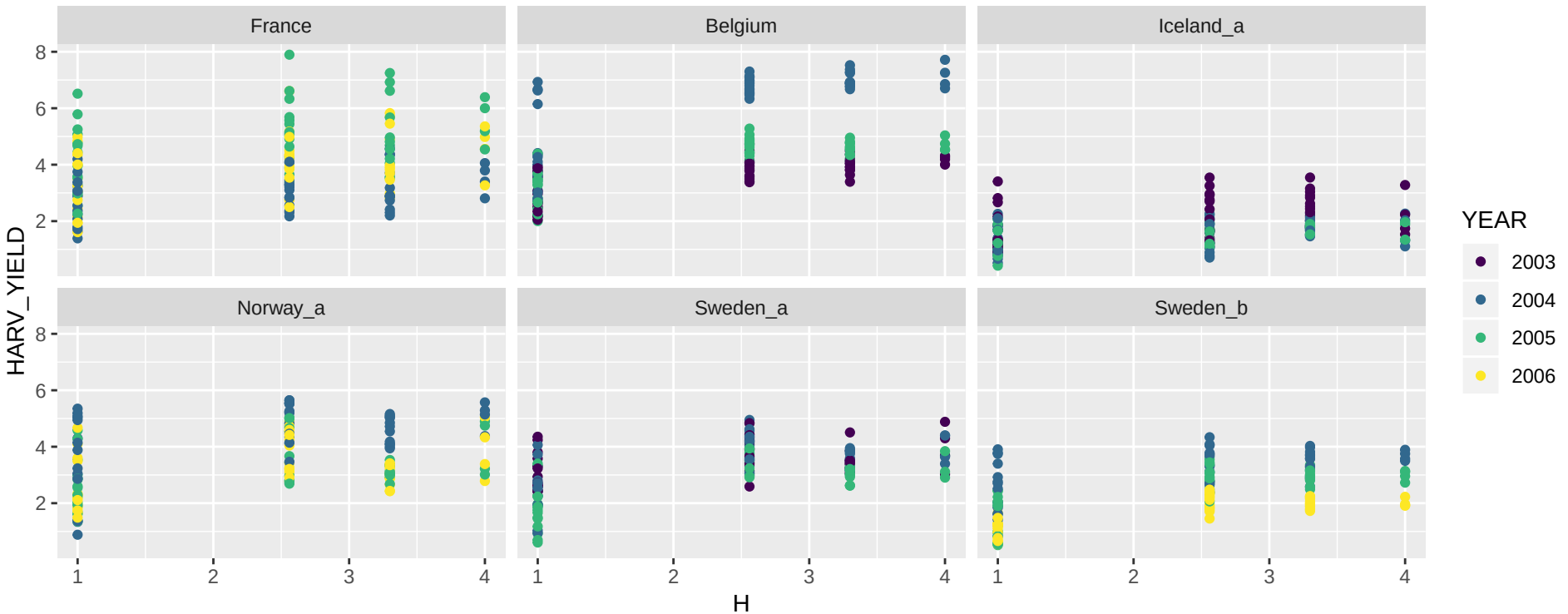
Impact de la biodiversité sur le rendement

Présentation des données

```
load('../../Datasets/Biomass.Rdata')

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>     <fct> <fct> <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 :

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

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

- 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

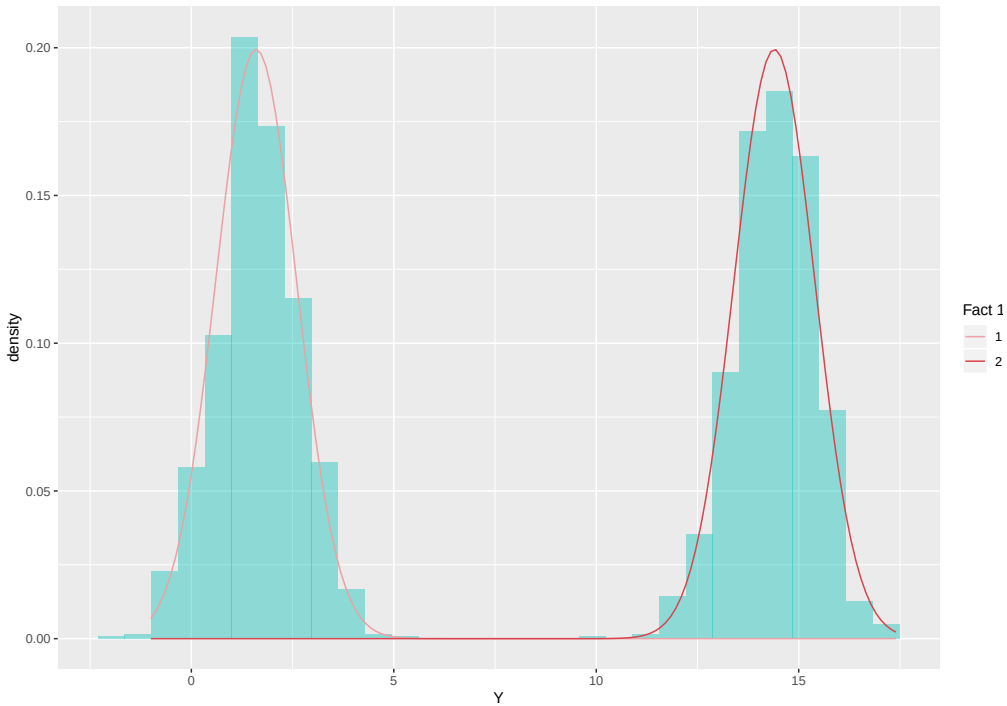
Le modèle d'analyse de variance à 2 facteurs s'écrit :

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

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

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 :

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

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

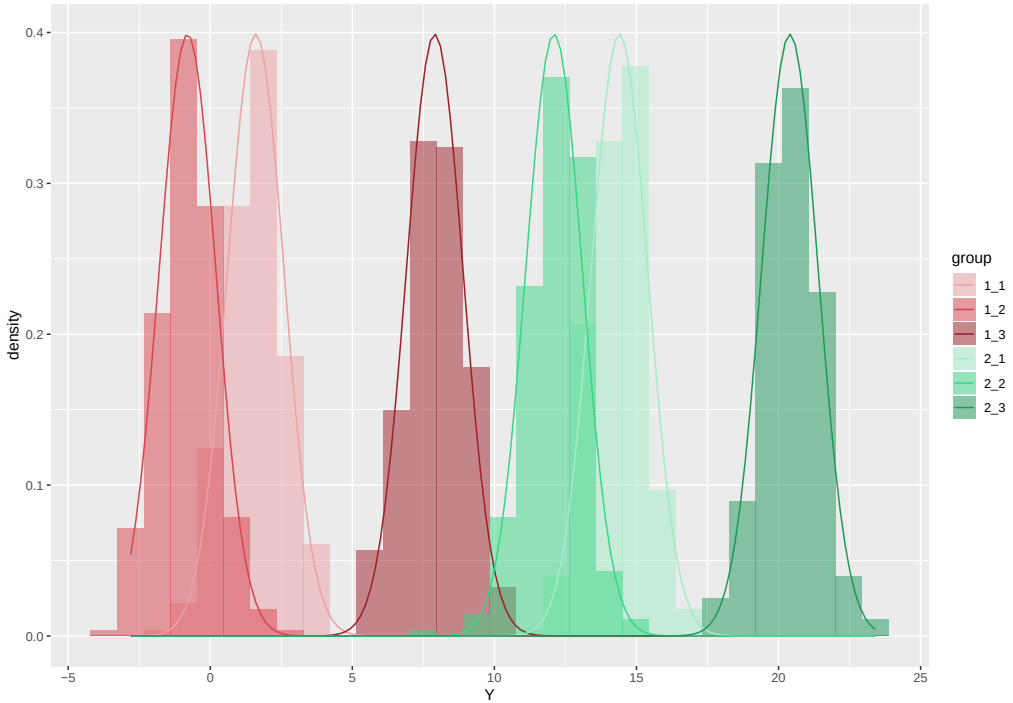
Version matricielle

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

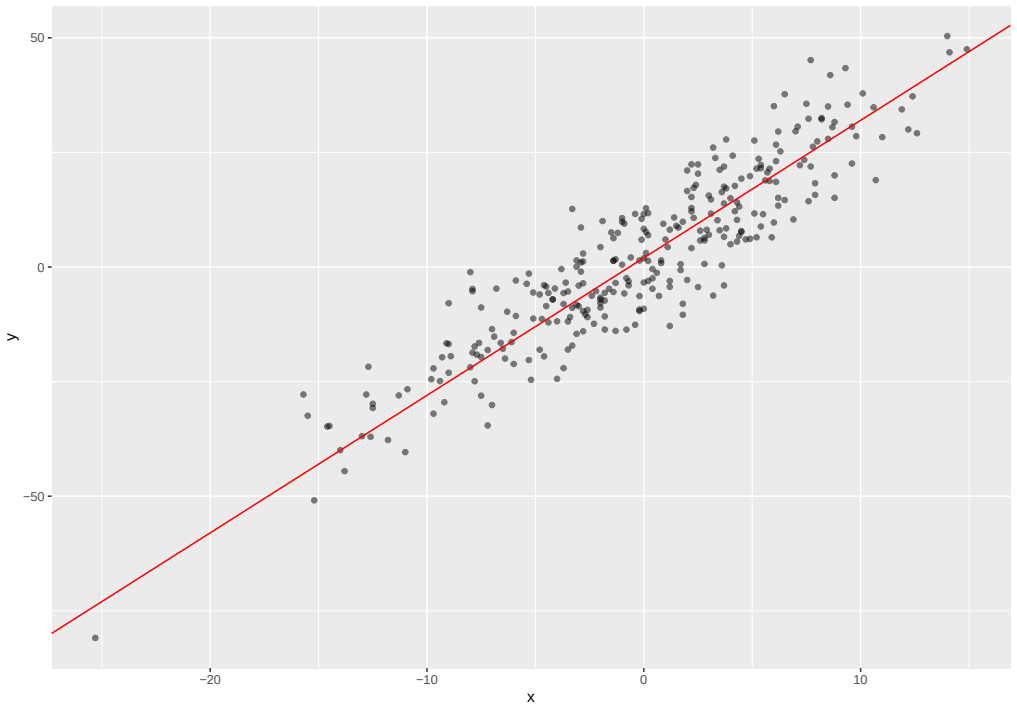
Version matricielle

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

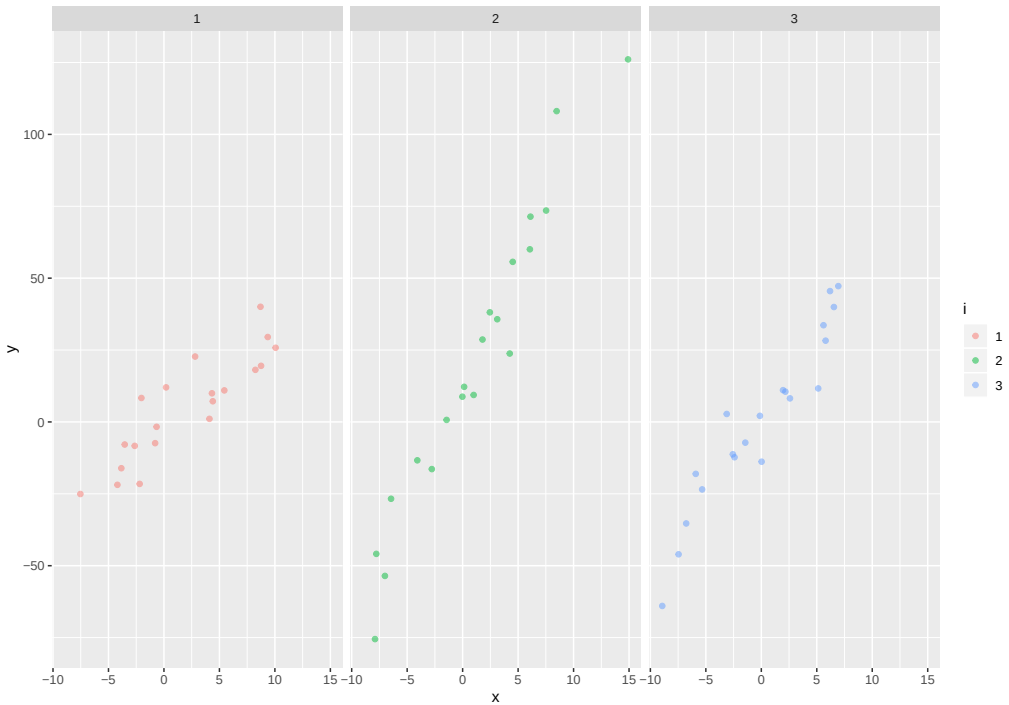
Version matricielle

Le modèle de régression multiple s'écrit:

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

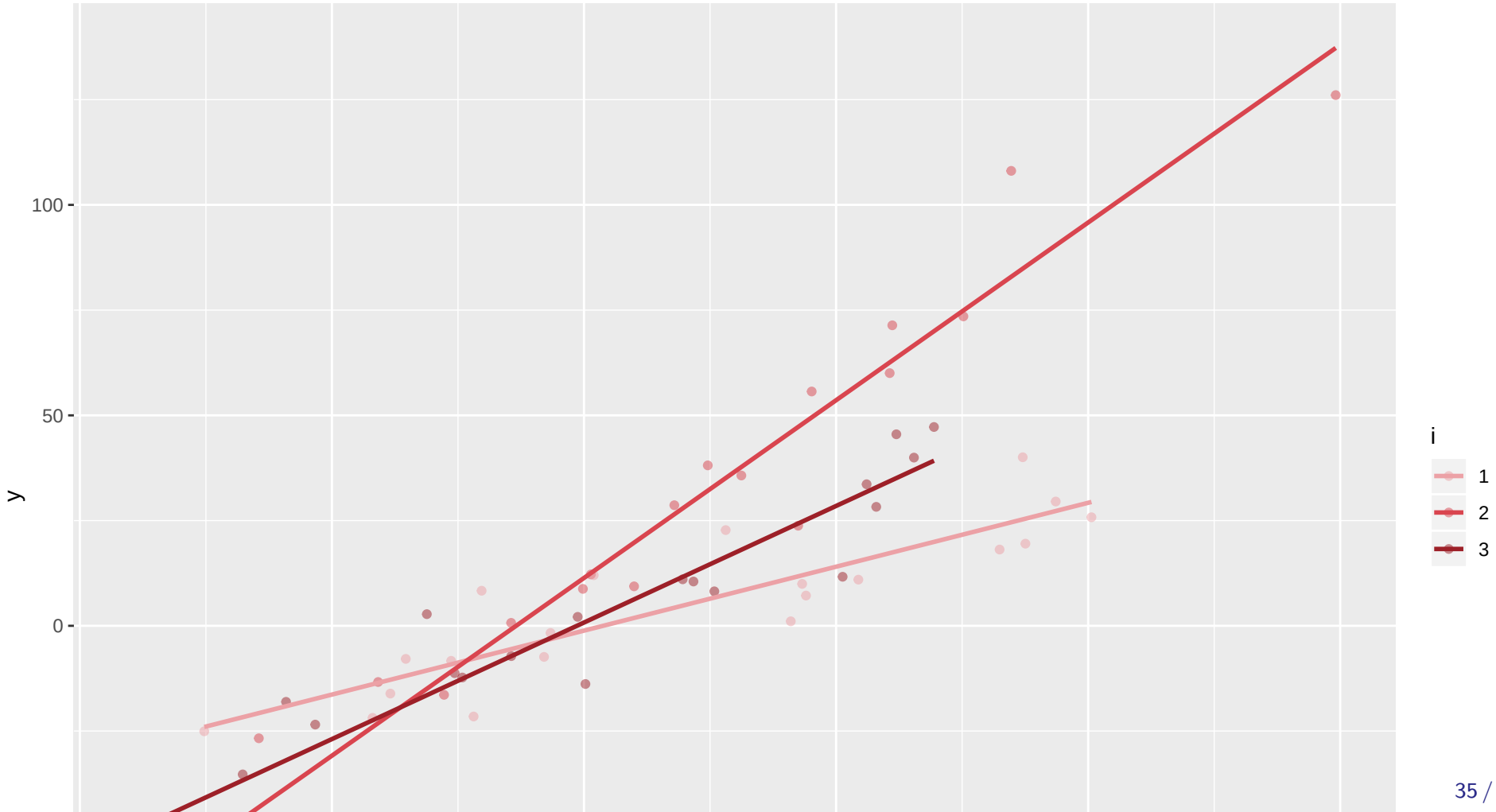
$$\boldsymbol{E} \sim (0, \sigma^2 I_n)$$

Vision graphique



Vision graphique

```
p <- ggplot(data = dta, aes(x=x, y=y, col = i)) +    geom_point( al
p +
  scale_fill_manual(values = anova_colour) + scale_colour_manual(va
```



Plan

Estimation par maximum de vraisemblance

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

Références

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.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2014). *Advanced r*. CRC Press. Retrieved from http://adv-r.had.co.nz/