Projet de biostatistiques - Ankylostome

Amandine LIAGRE - Florian BUCQUET - Rachid ABDELJABBAR

03-02-2024

Contents

Introduction

A travers ce projet, nous allons utiliser les données provenant d'une enquête réalisée sur un échantillon d'individus en Egypte. Plus particulièrement, nous avons des informations concernant l'infection des individus par l'ankylostome. En marchant pieds nus, les individus sont contaminés via les larves des ankylostomes vivant en terre. L'infection peut aussi se produire via une ingestion d'aliments contaminés par des larves. Les différents symptômes posibles sont des éruptions et lésions cutannées aux endroits où les larves ont pénétré la peau, de la fièvre, des douleurs épigastriques, des diarrhées, de la toux, inflammation de l'intestin ... . Dans les cas les plus graves, le malade peut être victime d'une perte de sang (les larves dans l'intestin se nourrissent de sang en étant accroché à sa paroi et il en résulte une potentielle anémie pour le malade), d'insuffisance cardiaque. Il existe des médicaments antiparasitaires pour traiter cette infection (albendazole, mébendazole).

L'ankylostome vit particulièrement bien dans la terre et une température aux alentours des 18°C afin que les oeufs puissent éclore. Les oeufs d'ankylostomes ont l'allure suivante:

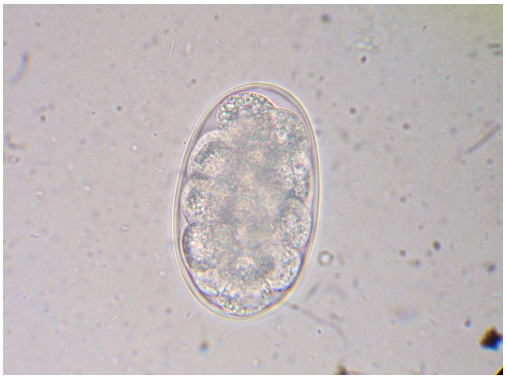


Figure 1: Oeufs d'Ankylostosme, par Joel Mills - CC BY-SA 3.0

Et par la suite, deviennent les des vers se propageant vers l'intestin:

Figure 2: Vers d'Ankylostosme, par CDC's Public Health Image Library

Voici le cycle parasitaire de l'ankylostome:

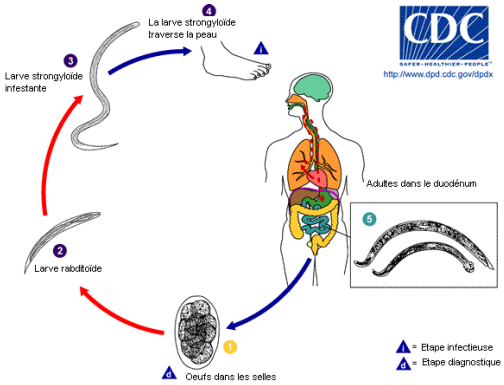


Figure 3: Cycle parasitaire de l'ankylostome, par CDC - Department of Parasitic Diseases - Domaine public

I. Lecture des données et vérification

A travers le tableau suivant, voici un récapitulatif de nos données:

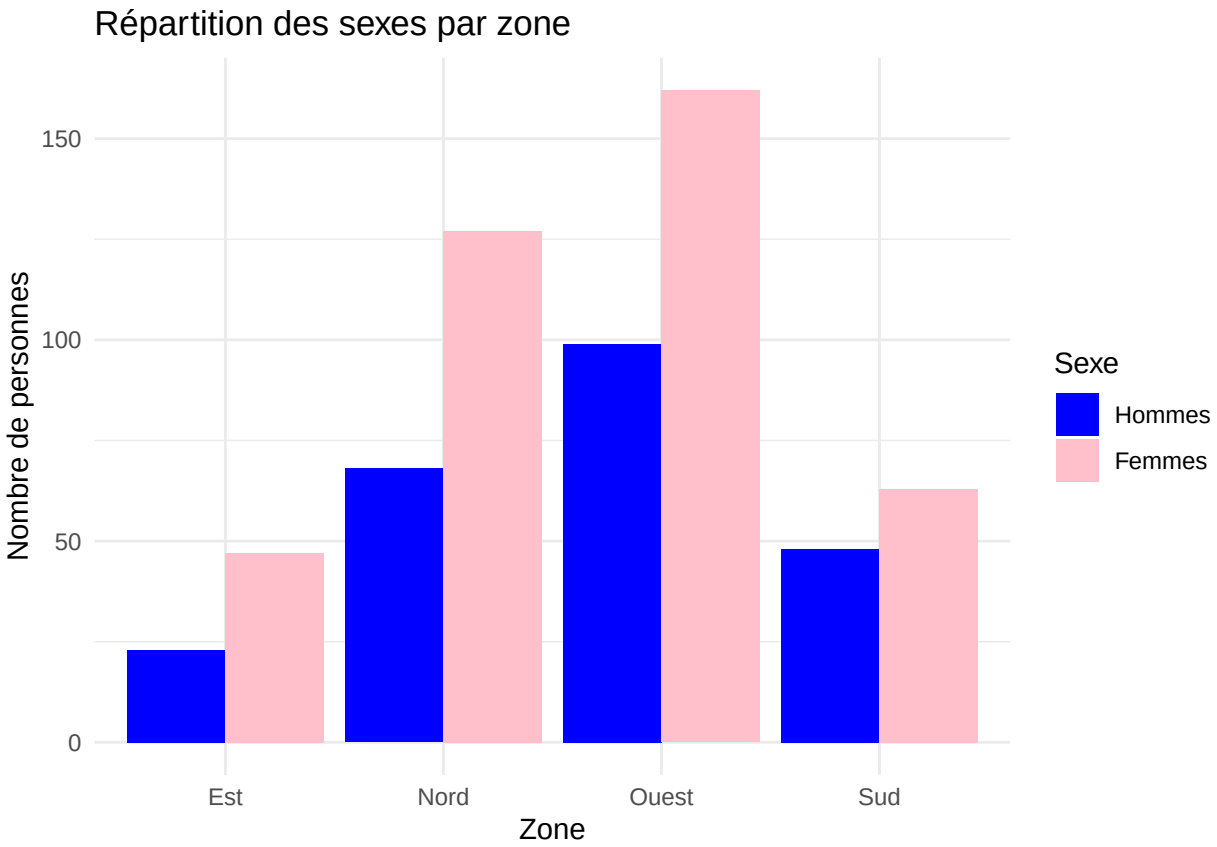
Nom de la variable	Type	Modalités ou exemples de modalités
id	int	440, 336, 60, ...
age	int	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, ...
agegr	chr	<15 yrs, 15-59 yrs, 60+ yrs
zone	chr	Nord, Ouest, Sud, Est
sexe	int	0 (masculin ?), 1 (féminin ?)
chaussures	chr	"no", "yes"
nb.oeufs	int	0, 46, 184, 989, 1150, 690, ...
intensite	chr	0, "[1;1.999]", "[2;+]"
ageclasses	chr	<16 ans, 16-49, 49 et plus

Voici l'allure générale de nos données:

```
##    id age   agegr  zone sexe chaussures nb.oeufs   intensite ageclasses
## 1 440   2 <15 yrs  Nord    0         no        0           0    <16 ans
## 2 336   2 <15 yrs Ouest    1         no       46 "[1;1.999]"    <16 ans
## 3  60   2 <15 yrs  Nord    1         no      184 "[1;1.999]"    <16 ans
## 4 100   2 <15 yrs Ouest    1         no        0           0    <16 ans
## 5 281   2 <15 yrs   Sud    1         no        0           0    <16 ans
## 6  90   2 <15 yrs Ouest    0         no        0           0    <16 ans
```

I.1 Informations concernant les individus

La table de données contient 238 hommes et 399 femmes et ils sont répartis de la manière suivante selon la zone géographique:



Regardons les catégories d'âges. Trois variables sont à notre disposition: age, agegr et ageclasses.

Concernant la variable **age**:

Statistique	Valeur
Min.	2.00
1st Qu.	9.00
Median	23.00
Mean	25.94
3rd Qu.	40.00
Max.	78.00

Concernant la variable **agegr**:

Catégorie	Valeur
<15 yrs	259
15-59 yrs	331
60+ yrs	47

Concernant la variable **ageclasses**:

Catégorie	Valeur
<16 ans	259

Catégorie	Valeur
16-49 ans	331
49 et plus	47

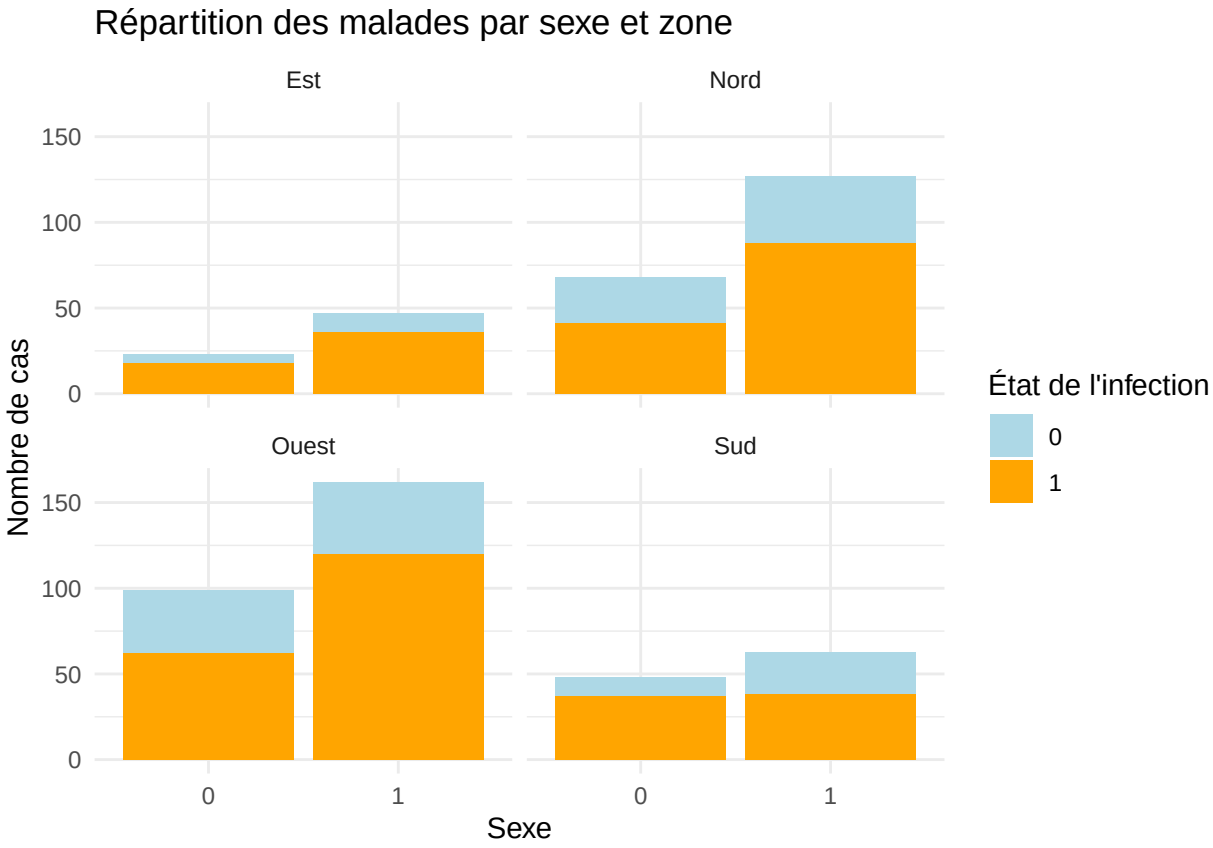
I.2 Création de la variable “malade” et observations

Afin de réaliser notre étude, nous créons la variable **malade**. Nous considérons qu'un individu est infecté si la variable **nb.oeufs** est supérieur à 0. La variable malade vaudra 0 si la variable **nb.oeufs** est égal à 0, sinon elle vaudra 1.

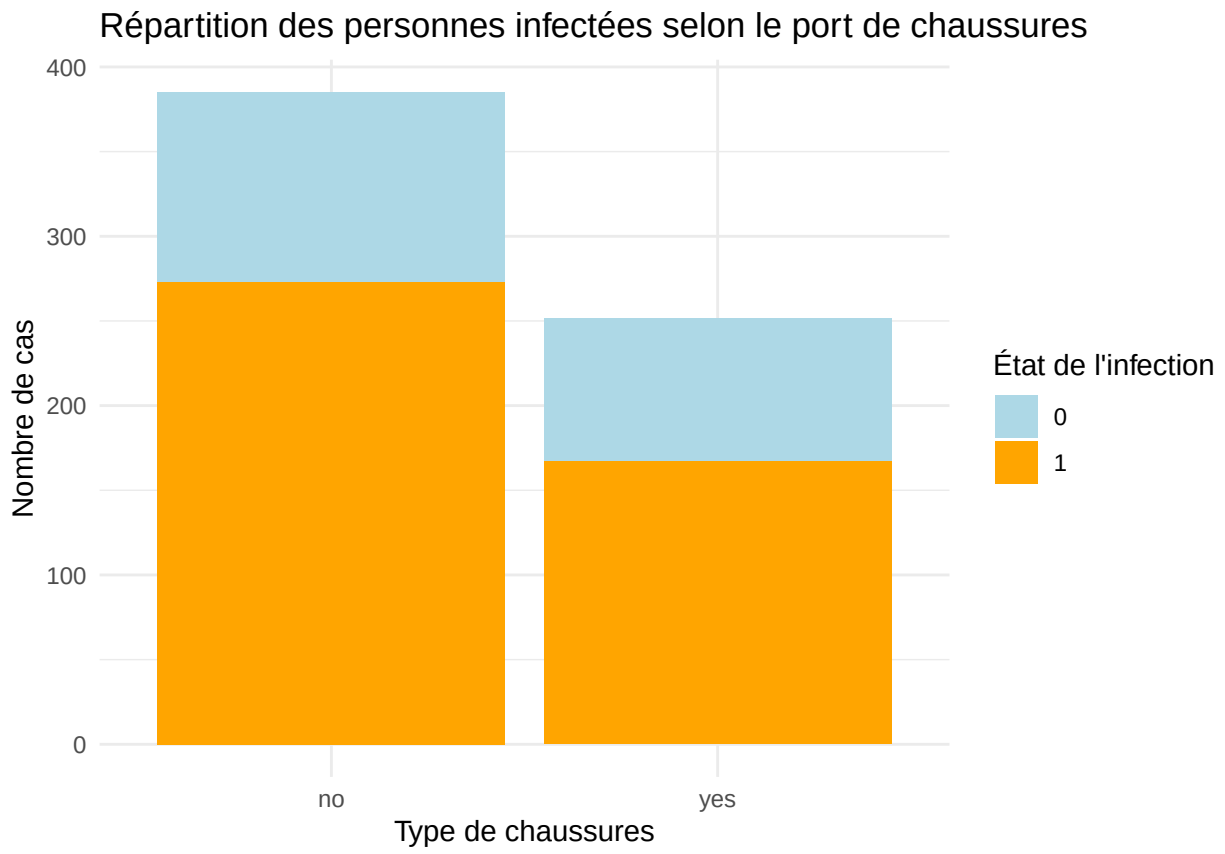
Suite à cette création nous constatons qu'il y a 197 personnes non malades (31% de l'échantillon) et 440 personnes malades (69% de l'échantillon), pour un total de 637 personnes.

I.3 Quelques analyses préliminaires via des graphiques

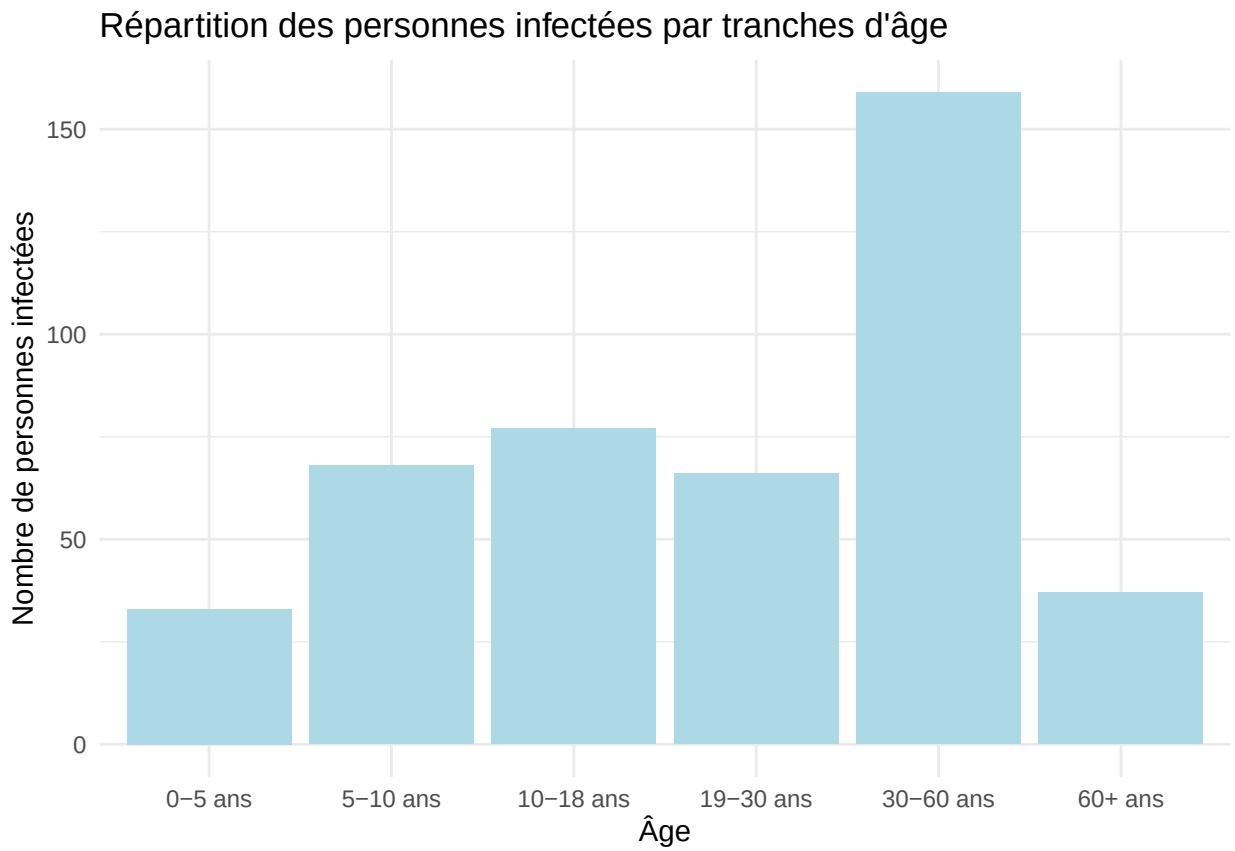
La répartition des sexes par zone géographique selon l'infection:



La répartition des personnes infectées et le port de chaussures:



La répartition des personnes infectées selon leur âge. Pour cela nous allons créer une nouvelle variable age_categ car les deux autres variables qui classent les âges ne nous semblent pas représentatives. Voici les tranches choisies:
- 0 à 5 ans exclus,
- 5 ans à 10 ans exclus,
- 10 ans à 18 ans exclus,
- 19 ans à 30 ans exclus,
- 30 à 60 ans exlus,
- les plus de 60 ans



II. Modèle (à changer)

Le modèle choisi est le suivant : MODELE TROP GROS a priori pour interpréter avec les odds

$$malade = \beta_0 + \beta_1 * age + \beta_2 * sexe + \beta_3 * chaussuresyes + \beta_4 * zoneNord + \beta_5 * zoneOuest + \beta_6 * zoneSud$$

```
### MODELE DE REGRESSION ###
res <- glm(malade ~ age + sexe + chaussures + zone, family="binomial", data=data)
```

Les valeurs des coefficients sont les suivantes:

Variable	Coefficient	Notation
(Intercept)	0.80710299	$\hat{\beta}_0$
age	0.02076273	$\hat{\beta}_1$
sexe	0.11107693	$\hat{\beta}_2$
chaussuresyes	-0.41702146	$\hat{\beta}_3$
zoneNord	-0.55788369	$\hat{\beta}_4$
zoneOuest	-0.38105327	$\hat{\beta}_5$
zoneSud	-0.46207829	$\hat{\beta}_6$

Intervalles de confiance:

```
## Waiting for profiling to be done...

##                       2.5 %       97.5 %
## (Intercept)     0.11681771  1.535147016
## age             0.01091286  0.031091722
## sexe           -0.28877314  0.508351401
## chaussuresyes -0.83286244 -0.004462873
## zoneNord      -1.22407046  0.067038198
## zoneOuest     -1.03287792  0.227246137
## zoneSud       -1.17648515  0.221809630
```

III. Prédictions

```
### PREDICTIONS ###
vect_estimations <- round(res$fitted.values)

#Effectif
tab=table(data$malade, vect_estimations)
tab

#Proportion de personnes pour laquelle la prédiction a été mauvaise: 197 (1 + 196)
#1. => 31%
(tab[1,2] + tab[2,1])
(tab[1,2] + tab[2,1])/sum(tab)

#2. Proportion de personnes infectées pour laquelle la prédiction était non infecté: 31% (faux positifs
tab[1,2]/sum(tab[,2])

#3. Proportion de personnes non infectées pour laquelle la prédiction était infectées => 50% (faux néga
tab[2,1]/sum(tab[,1])
```

Conclusion

A MODIFIER

Il est important de noter que pour prévenir la population de ce type d'infection, il vaut mieux éviter de marcher pieds nus, d'utiliser des eaux usées et de bien utiliser des dispositifs de toilettes, d'hygiène pour éviter la présence de selles au sol. Le diagnostic de l'infection peut-être réalisé via un examen d'un échantillon de selles ou d'analyse de sang.

Annexe Code R

```r
knitr::opts_chunk$set(echo = TRUE)
### LIBRAIRIES UTILISEES ###
library(dplyr)
library(ggplot2)

### LECTURE DES DONNEES ET MODALITES ###
data <- read.csv("Ankylostome.csv")
data <- data %>% select(-c(...1, X))

modalites_uniques <- lapply(data, function(colonne) {
  unique_values <- unique(colonne)
  count_values <- length(unique_values)
  list(Modalites = unique_values, Nombre = count_values)
})

### ALLURE GENERALE DES DONNÉES ###
head(data)

### SEXE DES INDIVIDUS ###
table_sexe <- table(data$sexe)

### REPARTITION SELON LES ZONES ###
ggplot(data, aes(x = zone, fill = as.factor(sexe))) +
  geom_bar(position = "dodge") +
  labs(
    title = "Répartition des sexes par zone",
    x = "Zone",
    y = "Nombre de personnes",
    fill = "Sexe"
  ) +
  scale_fill_manual(
    values = c("0" = "blue", "1" = "pink"),
    labels = c("0" = "Hommes", "1" = "Femmes")
  ) +
  theme_minimal()

### AGE DES INDIVIDUS ###
age <- summary(data["age"])
agegr <- table(data$agegr)
ageclasses <- table(data$ageclasses)

### CREATION DE LA VARIABLE MALADE ET OBSERVATIONS ###
data <- data %>% mutate(malade = ifelse(nb.oeufs == 0, 0, 1))
malades <- table(data$malade)
pourcentages_malades <- prop.table(malades) * 100

### répartition des sexes par zone géographique selon l'infection ###
ggplot(data, aes(x = as.factor(sexe), fill = as.factor(malade))) +
  geom_bar(position = "stack") +
  facet_wrap(~ zone) +
  labs(title = "Répartition des malades par sexe et zone",
```

```
       x = "Sexe",
       y = "Nombre de cas",
       fill = "État de l'infection") +
  scale_fill_manual(values = c("lightblue", "orange")) +
  theme_minimal()

### répartition des personnes infectées et le port de chaussures ###
ggplot(data, aes(x = chaussures, fill = as.factor(malade))) +
  geom_bar(position = "stack") +
  labs(title = "Répartition des personnes infectées selon le port de chaussures",
       x = "Type de chaussures",
       y = "Nombre de cas",
       fill = "État de l'infection") +
  scale_fill_manual(values = c("lightblue", "orange")) +
  theme_minimal()

### CREATION DES CLASSES D'AGE ET GRAPHIQUE ###
data$age_categ <- cut(data$age,
                      breaks = c(0, 5, 10, 18, 30, 60, Inf),
                      labels = c("0-5 ans", "5-10 ans", "10-18 ans", "19-30 ans", "30-60 ans", "60+ ans
                      right = FALSE)

data_malade <- subset(data, malade == 1)

ggplot(data_malade, aes(x = age_categ)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Répartition des personnes infectées par tranches d'âge",
       x = "Âge",
       y = "Nombre de personnes infectées") +
  theme_minimal()

### MODELE DE REGRESSION ###
res <- glm(malade ~ age + sexe + chaussures + zone, family="binomial", data=data)

### VALEURS DES COEFFICIENTS ###
res$coefficients

### INTERVALLES DE CONFIANCE ###
confint(res, level=0.95)

### PREDICTIONS ###
vect_estimations <- round(res$fitted.values)

#Effectif
tab=table(data$malade, vect_estimations)
tab

#Proportion de personnes pour laquelle la prédiction a été mauvaise: 197 (1 + 196)
#1. => 31%
(tab[1,2] + tab[2,1])
(tab[1,2] + tab[2,1])/sum(tab)

#2. Proportion de personnes infectées pour laquelle la prédiction était non infecté: 31% (faux positifs
```

```r
tab[1,2]/sum(tab[,2])

#3. Proportion de personnes non infectées pour laquelle la prédiction était infectées => 50% (faux néga
tab[2,1]/sum(tab[,1])
```