Projet

Vincent Mouillot & Inès Rouached

20/12/2021

Contents

1 Introduction

Notre jeu de données contient des données recensées sur les participants d'une grande manifestation en Inde en 1990.
Pour cela, les habitudes alimentaires sur ces individus ont été répertoriées. Suivant la problématique posée, différentes approches (méthodes) apprises en cours et en TP de biostatistiques seront mises en avant.
L'objectif principal sera de construire de potentiels modèle qui permettent d'expliquer et de constater les liens statistiques qui peuvent exister. Un grand échantillon des manifestants a donc été interrogé afin d'étudier l'impact de certains plats ou boissons vendus sur place et, a terme, de pouvoir définir le ou les responsables de ces intoxications.

Notre table est composée de onze variables dont une variable identifiant (id), deux caractérisant l'individu (le sexe et l'âge), quatre concernent les aliments consommés lors de la manifestation (boeuf au curry, oeufs frits, eau et quantité d'éclair) et quatre concernent les symptômes ressentis par l'individu (nausée, vomissements, douleur(abdominale) et diarrhée).
1094 personnes font partie de notre échantillon.

Ces variables sont codés de la manière suivante :
Variables quantitatives :

id : identifiant
Age : l'âge en années

Variables qualitatives :

sex : Sexe, 0 : femme / 1 : Homme
Boeuf : le sujet a-t-il mangé du boeuf au curry lors de l'événement ? 1 : Oui / 2 : Non
Oeuf : le sujet a-t-il mangé des oeufs frits lors de l'événement ? 1 : Oui / 2 : Non
Eau : le sujet a-t-il bu de l'eau distribuée ce jour là ? 1 : Oui / 2 : Non
Eclair : combien d'éclairs l'individu a mangé ce jour là. Les valeurs 80 et 90 correspondent respectivement à "a mangé des éclair sans se souvenir combien" et "donnée manquante".

Viennent ensuite les variables de symptôme : nausée, vomissement, douleur (abdominale) et diarrhée, constatés chez l'individu. Les symptômes constatés ont été codés de la façon suivante :
0 : pas de symtôme / 1 : symptôme / 99 : non renseigné.

Afin de mieux comprendre ces données, nous allons commencer par en faire une première lecture descriptive.

2 Statistiques descriptives

Regardons si notre jeu de donnée contient des données manquantes :

```
Intoxications <- read_excel("Intoxications.xls")
colnames(Intoxications) <- c("id", "Sexe", "Age", "Boeuf", "Oeuf", "Eclair", "Eau", "Nausee", "Vomi", "
Intoxications$Sexe=factor(Intoxications$Sexe)

Intoxications$Eclair[Intoxications$Eclair==90]<-NA
#Intoxications$éclair[Intoxications$éclair==80]<-NA
Intoxications$Nausee[Intoxications$Nausee == 99] <- NA
Intoxications$Vomi[Intoxications$Vomi == 99] <- NA
Intoxications$Douleur[Intoxications$Douleur == 99] <- NA
Intoxications$Diarrhee[Intoxications$Diarrhee == 99] <- NA
Intoxications$Boeuf[Intoxications$Boeuf==9]<-NA
Intoxications$Oeuf[Intoxications$Oeuf==9]<-NA
Intoxications$Eau[Intoxications$Eau == 9] <- NA
sum(is.na(Intoxications))
```

```
## [1] 133
```

Après recodage des données manquantes, on observe qu'il y a dans la table 133 données manquantes. Parmis ces 133 valeurs manquantes, il y en a 117 qui appartiennent à la question sur les éclairs.

Ainsi, comme on a peu d'individus avec une ou des informations manquantes ($\frac{133}{1094} \approx 12.2\%$), on a décidé de les retirer du jeu de données.

Lecture des données :

Voici les 5 premières lignes de notre jeu de données :

```
## # A tibble: 5 x 13
##      id Sexe    Age Boeuf Oeuf  Eclair Eau   Nausee Vomi  Douleur Diarrhee
##   <dbl> <fct> <dbl> <fct> <fct>  <dbl> <fct> <fct>  <fct> <fct>   <fct>
## 1     1 1        13 1     1        1   1     1      1     1       1
## 2     2 1        14 1     1        0   1     0      0     0       0
## 3     3 1        13 1     1        0   1     0      0     0       0
## 4     4 1        15 1     1        0.5 1     0      0     0       0
## 5     5 1        14 1     1        0   1     0      0     0       0
## # ... with 2 more variables: Symptomes <dbl>, Malade <fct>
```

La fonction summary() permet d'avoir la description statistique de notre table de donnée.

Pour une variable donnée, la fonction renvoie 5 valeurs : le minimum (Min.), le premier quartile (1st Qu.), la médiane (Median), la moyenne (Mean), le troisième quartile (3rd Qu.) et le maximum (Max).

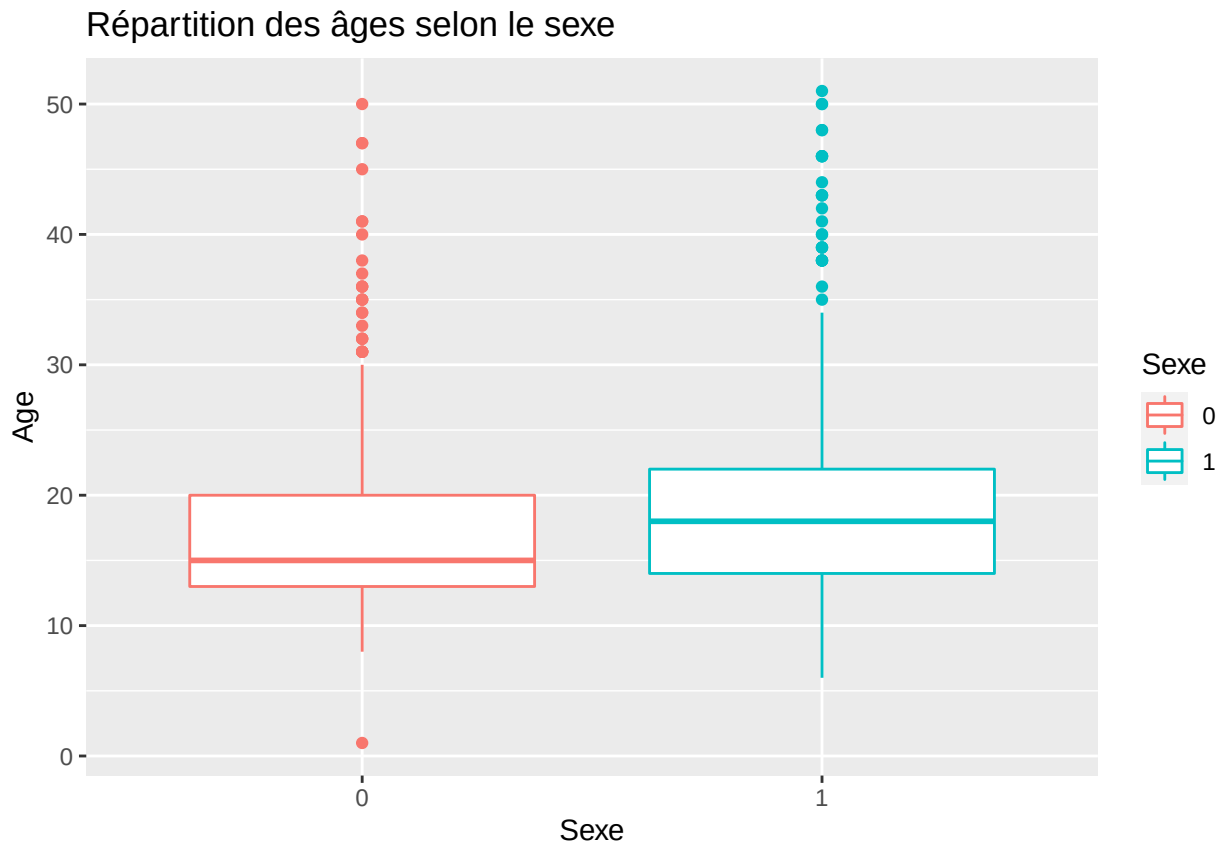
```
Intoxications %>% summary()
```

```
##        id         Sexe         Age        Boeuf   Oeuf        Eclair
##  Min.   :   1.0   0:318   Min.   : 1.00   0: 61   0: 61   Min.   : 0.000
##  1st Qu.: 234.5   1:593   1st Qu.:14.00   1:850   1:850   1st Qu.: 0.000
##  Median : 483.0           Median :17.00                   Median : 2.000
##  Mean   : 505.5           Mean   :18.77                   Mean   : 1.702
##  3rd Qu.: 781.5           3rd Qu.:21.00                   3rd Qu.: 2.000
```

```
##  Max.   :1094.0           Max.   :51.00                    Max.   :20.000
##  Eau     Nausee  Vomi    Douleur Diarrhee   Symptomes     Malade
##  0: 23   0:557   0:572   0:600   0:712    Min.   :0.000   0:526
##  1:888   1:354   1:339   1:311   1:199    1st Qu.:0.000   1:385
##                                           Median :0.000
##                                           Mean   :1.321
##                                           3rd Qu.:3.000
##                                           Max.   :4.000
```

Nous allons d'abord observer la répartition des âges des hommes et des femmes au sein de l'échantillon.

```
ggplot(Intoxications, aes(group = Sexe, x = Sexe, y = Age, color=Sexe)) +
 geom_boxplot() +
 ggtitle("Répartition des âges selon le sexe")
```



Les femmes sont codées par un 0 et les hommes par un 1.
L'échantillon est composé de 721 hommes et 373 femmes.
On peut supposer que l'échantillon est représentatif du public de la manifestation puisqu'on a un échantillon de taille importante.
Alors on observe une population plutôt jeune. Les femmes ont en moyenne 15 ans et aux alentours de 18 ans pour les hommes.
Nous allons voir à présent les différences de consommations des plats mise en cause dans les intoxications alimentaires constatées.

```
c(sum(Intoxications$Boeuf, na.rm=T), sum(Intoxications$Oeuf, na.rm=T), sum(Intoxications$Eau, na.rm=T))
```

```
## [1] 1043 1047 1117
```

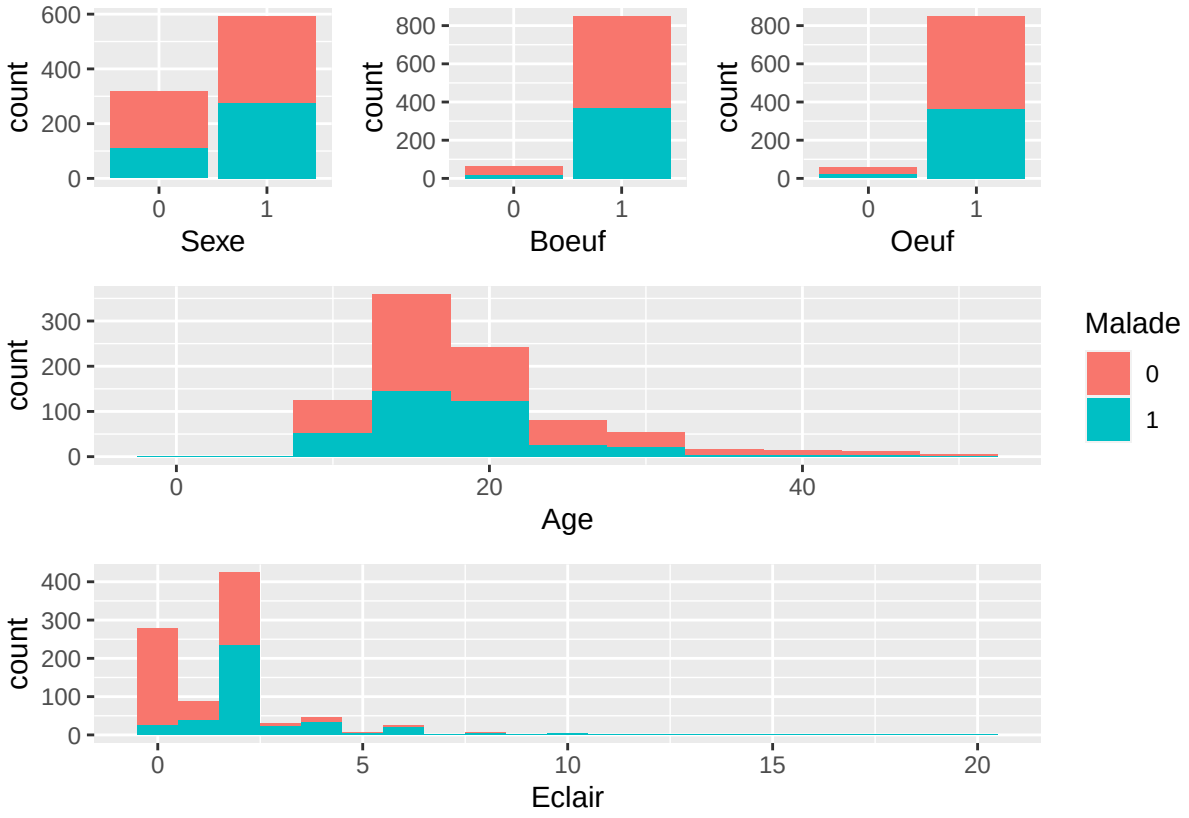
```
table(Intoxications$Eclair)
```

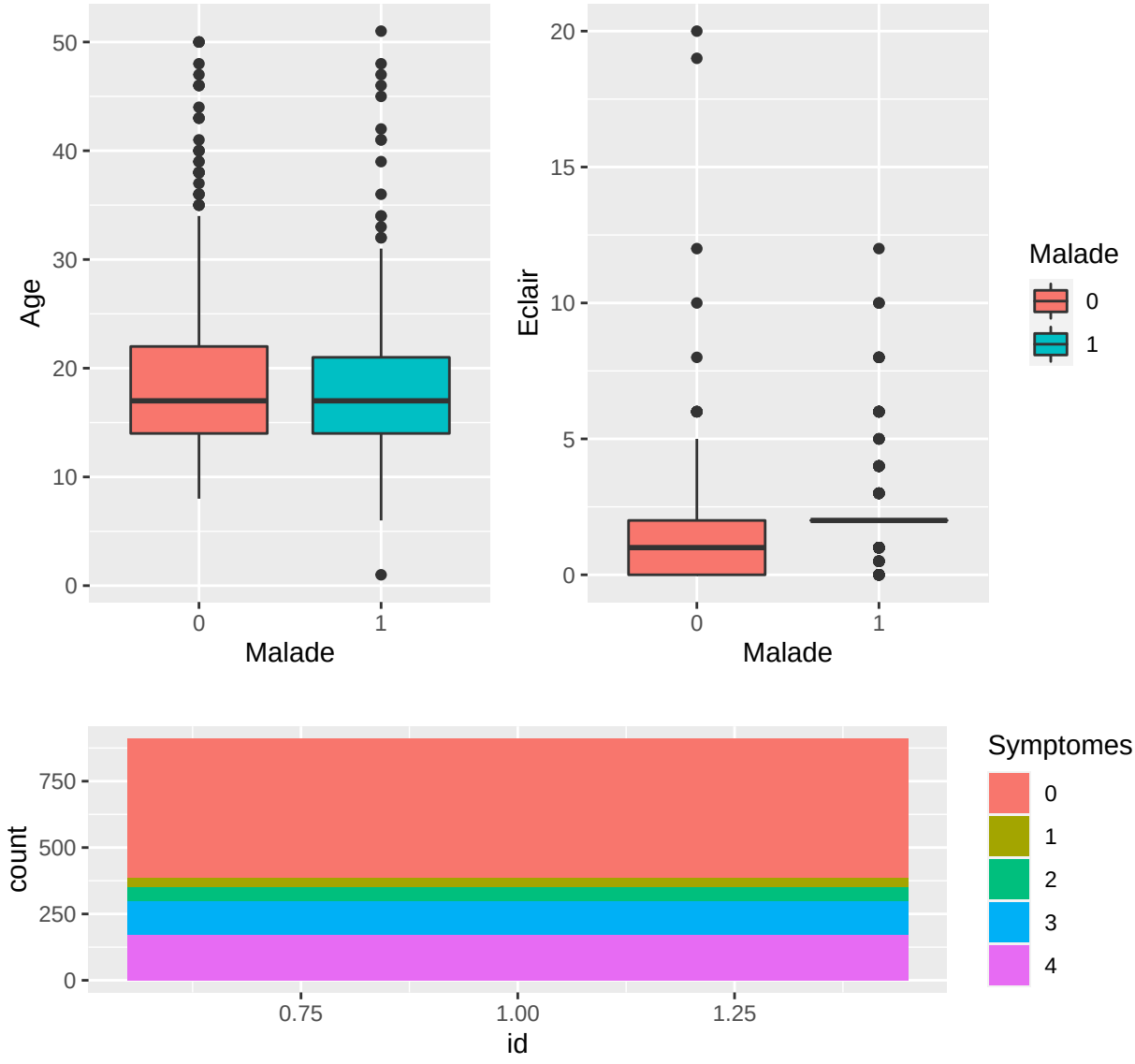
```
##
##   0 0.5   1  10  12  19   2  20   3   4   5   6   8  80  90
## 294  15  90   4   2   1 446   1  31  47   7  28   6   5 117
```

```
sum(Intoxications$Boeuf & Intoxications$Oeuf & Intoxications$Eau, na.rm=T)
```

```
## [1] 950
```

Au sein de l'échantillon, chaque plat a été consommé par une large proportion d'individus.
Le boeuf au curry, l'oeuf frit et l'eau ont été pris par environ 1000 individus chacun, et 950 personnes ont consommé les trois. En plus petite proportion, 683 personnes (Echantillon totale - personne n'ayant pas mangé d'éclair - données manquantes = 1094-294-117 = 683) ont pris des éclairs, entre une moitié et 8 éclairs. On observe également que 5 personnes ayant déclaré avoir mangé des éclairs ne se souviennent plus de combien.




Age
50
40
30
20
10
0
0
1
Malade
Eclair
20
15
10
5
0
0
1
Malade
Malade
0
1
count
750
500
250
0
0.75
1.00
1.25
id
Symptomes
0
1
2
3
4

3 Regression logistique

```
f <- stepAIC(object = glm(formula = Malade ~ Sexe * Eau * Boeuf * Oeuf * Age * Eclair,
    family = "binomial",
    data = Intoxications),
    direction = "backward")
```

```
f$converged
```

```
## [1] FALSE
```

Méthode backaward pas adapté car trop de cas particulier et pas d'individus respectant ceux-ci.

```
g <- glm(formula = Malade ~ 1,
    family = "binomial",
    data = Intoxications)

fitforw <- stepAIC(object = g,
    direction = "forward",
    scope = list(lower = g,
                 upper = ~ Sexe * Age * Boeuf * Oeuf * Eau * Eclair))
```

```
summary(fitforw)
```

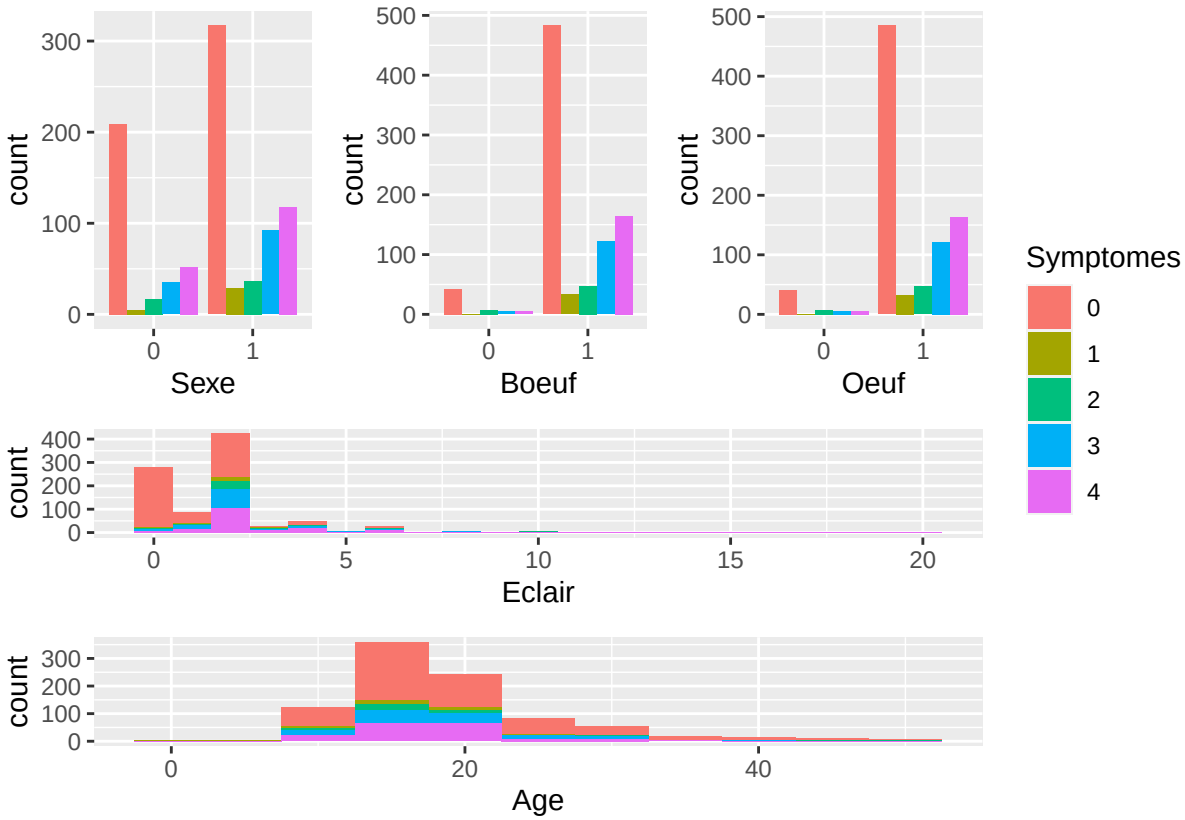
```
##
## Call:
## glm(formula = Malade ~ Eclair + Sexe + Age + Eclair:Age + Sexe:Age,
##     family = "binomial", data = Intoxications)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9523  -0.9377  -0.6372   1.0968   2.0518
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.882057   0.407034  -7.081 1.43e-12 ***
## Eclair       1.182897   0.136469   8.668  < 2e-16 ***
## Sexe1        1.763691   0.424342   4.156 3.23e-05 ***
## Age          0.060464   0.018349   3.295 0.000983 ***
## Eclair:Age  -0.026794   0.004676  -5.730 1.01e-08 ***
## Sexe1:Age   -0.065478   0.020923  -3.129 0.001752 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1241.0  on 910  degrees of freedom
## Residual deviance: 1064.2  on 905  degrees of freedom
## AIC: 1076.2
##
## Number of Fisher Scoring iterations: 5
```

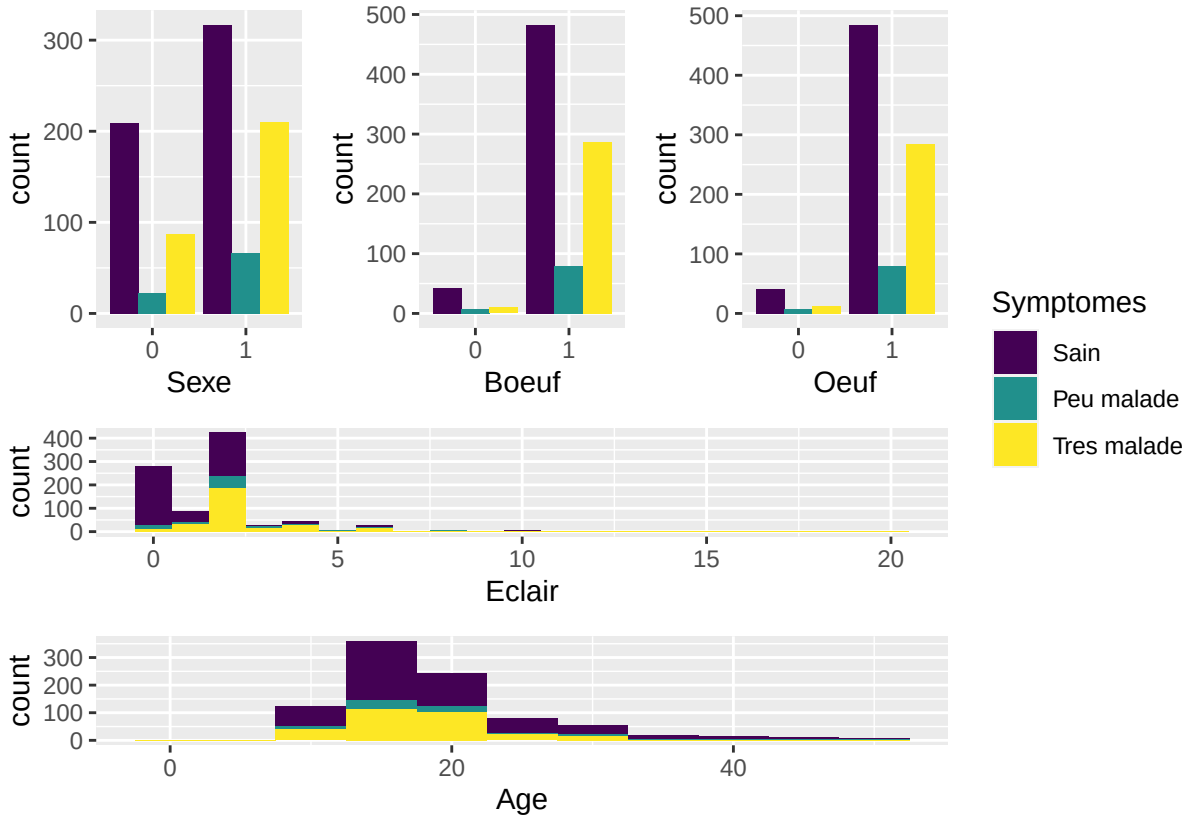
```
confint(fitforw)
```

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

```
##                    2.5 %      97.5 %
## (Intercept) -3.69626702 -2.09577353
## Eclair       0.92349620  1.46284960
## Sexe1        0.93806971  2.60399814
## Age          0.02442529  0.09686471
## Eclair:Age  -0.03698148 -0.01781954
## Sexe1:Age   -0.10689009 -0.02462878
```

4 Regression polytomique ordonnée






```
g <- polr(Symptomes ~ 1, data = Intoxications)

fitforw <- stepAIC(object = g,
    direction = "forward",
    scope = list(lower = g,
                 upper = ~Sexe * Age * Boeuf * Oeuf * Eau * Eclair))
```

```
summary(fitforw)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = Symptomes ~ Eclair + Eau + Sexe + Age + Eclair:Eau +
##     Eclair:Sexe + Sexe:Age + Eclair:Age + Eclair:Sexe:Age, data = Intoxications)
##
## Coefficients:
##                     Value Std. Error t value
## Eclair            1.44825    0.41029  3.5298
## Eau1              0.28675    0.71061  0.4035
## Sexe1             3.34859    0.78106  4.2872
## Age               0.08213    0.02892  2.8397
## Eclair:Eau1       0.27803    0.19510  1.4250
## Eclair:Sexe1     -1.09126    0.35605 -3.0649
## Sexe1:Age        -0.11008    0.03290 -3.3460
```

```
## Eclair:Age       -0.03723    0.01293 -2.8789
## Eclair:Sexe1:Age  0.03016    0.01330  2.2680
##
## Intercepts:
##                          Value   Std. Error t value
## Sain|Peu malade          4.0843  1.0199     4.0048
## Peu malade|Tres malade   4.5835  1.0218     4.4855
##
## Residual Deviance: 1473.189
## AIC: 1495.189

## Waiting for profiling to be done...

##
## Re-fitting to get Hessian
```