Devoir de statistiques

Master 1 Intelligence Artificielle
2024/2025
Prof: M. Mbaye FAYE

Sujet

Implémentation du modèle de régression poisson pour expliquer et prédire les crises épileptiques

Fait par: Adji Fatou Mahmoud Ibrahima MBAYE
Outils: R, R Studio

Plan

INTRODUCTION

L’épilepsie est une pathologie neurologique marquée par des crises récurrentes, dont la gestion repose souvent sur des traitements médicamenteux. Après l’arrêt de ces traitements, il est essentiel de comprendre les facteurs qui influencent la survenue des crises.

Ce rapport explore l'impact de trois variables — le type de traitement, la consommation d'alcool, et l’estime de soi — sur le nombre de crises survenant durant l'année suivant l’arrêt du traitement. L’objectif est d’appliquer un modèle statistique pour identifier les facteurs significatifs et mieux comprendre leur influence sur les événements épileptiques.

MÉTHODOLOGIES

Description du dataset

Le dataset utilisé pour cette analyse comprend 75 observations de quatre variables :

- **Esteem** : L'estime de soi des patients, mesurée sur une échelle numérique (plus la valeur est élevée, plus l’estime de soi est grande).
- **Alcohol** : Indicateur binaire de consommation d'alcool (1 : oui, 0 : non).
- **Treatment** : Type de traitement suivi par les patients (1 : haute dose, 2 : faible dose, 3 : placebo).
- **Events** : Nombre de crises d’épilepsie observées pendant l'année suivant la fin du traitement. Cette variable est de type comptage.

Méthode statistique

L’objectif principal de cette analyse est d’étudier la relation entre le nombre de crises d’épilepsie (variable dépendante) et les variables explicatives. En raison de la nature de la variable dépendante (comptage d'événements), nous avons choisi d'utiliser la régression de Poisson. Ce modèle permet de modéliser des événements rares ou comptés sur une période donnée en fonction de variables explicatives.

Qu’est ce que la régression de Poisson?

La régression de Poisson est un modèle linéaire généralisé utilisé pour les données de comptage. Ce modèle est basé sur la distribution de Poisson, qui décrit la probabilité de l'occurrence d'un certain nombre d'événements dans un intervalle de temps ou une zone donnée

.Elle suppose que le logarithme de son espérance est une combinaison linéaire des variables explicatives.

Le modèle est défini comme suit :

$$\log(\lambda i)=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp$$

Où :

λi est le nombre attendu d'événements pour l'observation i,

$\beta 0$ est l'intercept,

$\beta 1,\beta 2,\ldots,\beta p$ sont les coefficients estimés pour les variables explicatives x1,x2,...,xp.

Dans le cadre de cette étude, la régression de Poisson est utilisée pour analyser l'impact des variables telles que le type de traitement, la consommation d'alcool, et l'estime de soi sur le nombre de crises d'épilepsie observées chez les patients. Ce modèle est particulièrement adapté aux données de comptage, comme celles concernant les événements épileptiques.

PRETRAITEMENT DES DONNEES

Importation et chargement des données

```
library(foreign)
medData <- read.spss("med.poissonregression.equaltimes.sav",to.data.frame=TRUE)
```

Nous importons nos données depuis un fichier SPSS (.sav) à l'aide du package foreign, que nous stockons dans notre variable **medData.**

Exploration des données

- **Structure du dataset**

```
> str(medData)
'data.frame':   75 obs. of  4 variables:
 $ esteem    : num  13 15 16 15 21 10 18 17 17 16 ...
 $ alcohol   : num  0 0 0 0 0 1 0 0 1 0 ...
 $ treatment: num  1 1 1 1 1 1 1 1 1 1 ...
 $ events    : int  5 4 4 5 6 6 4 7 4 3 ...
 - attr(*, "variable.labels")= Named chr(0)
  ..- attr(*, "names")= chr(0)
 - attr(*, "codepage")= int 1252
```

La vue de la structure de notre base de données (**str( )**) montre que notre jeu de données est constitué de 75 observations réparties sur 4 variables: 1 variable catégorique, 3 variables numériques.

```
> summary(medData)
     esteem         alcohol         treatment         events
 Min.   : 7.00   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
 1st Qu.:13.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 4.000
 Median :15.00   Median :0.0000   Median :2.000   Median : 6.000
 Mean   :15.11   Mean   :0.3067   Mean   :2.067   Mean   : 6.093
 3rd Qu.:18.00   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.: 7.000
 Max.   :23.00   Max.   :1.0000   Max.   :3.000   Max.   :14.000
```

Summary( ) nous montre les statistiques générales sur les différentes variables.

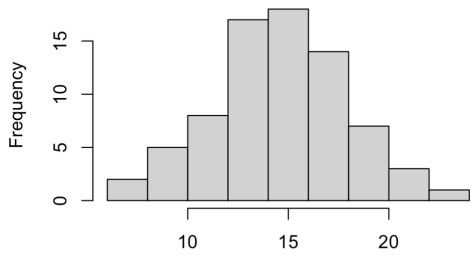
- **Constats initiaux**

```
> # Vérifier s'il y a des lignes dupliquées
> print("Lignes dupliquées: ", sum(duplicated(medData)))
[1] "Lignes dupliquées: "
> # Nombre total de valeurs manquantes
> print("Valeurs manquantes: ",sum(is.na(medData)))
Error in print.default("Valeurs manquantes: ", sum(is.na(medData))) :
  invalid printing digits 0
```

Absence de doublons : Aucune ligne dupliquée n'a été détectée
Valeurs manquantes: Aucune valeur manquante

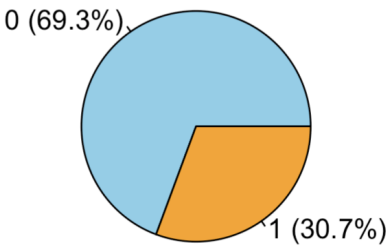
- **Analyses univariées**

Esteem:



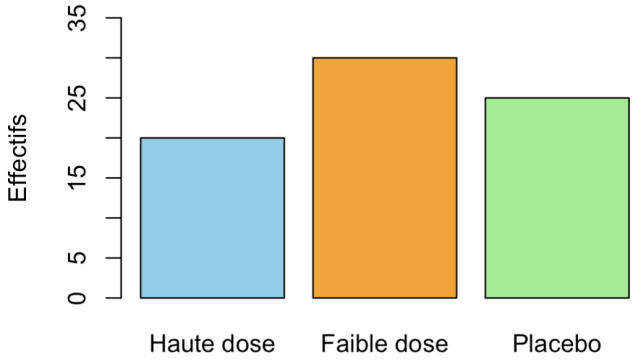
- Esteem minimum: 7
- Esteem maximum: 23
- Esteem moyen: 15

Alcohol:



- 31% de nos patients boit de l'alcool
- 69% de nos patients ne boit pas d'alcool

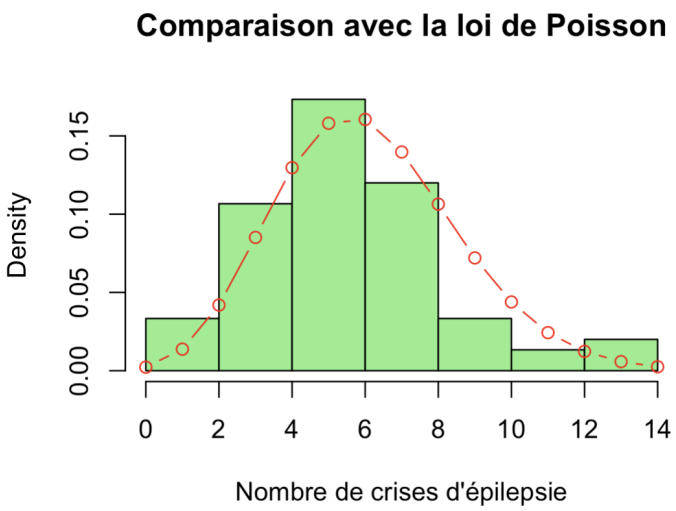
Treatment:



- Haute dose(1): 20
- Faible dose(2): 30
- Placebo (3): 25

Ils sont relativement équilibrés.

Events:




```
> cat("Moyenne des événements:", mean_events, "\nVariance des événements:", var_events)
Moyenne des événements: 5.65 6.133333 6.4
Variance des événements: 6.555676
```

- Moyenne= ~ 6
- Variance = 6,56

On peut en déduire que ces données suivent une loi de Poisson. Notre hypothèse est donc vérifiée. La régression Poisson peut être utilisée.

Préparation des données

La variable TREATMENT est qualitative. Nous devons la recoder. Nous optons pour un codage disjonctif simple, en prenant (TREATMENT = 3, placebo) comme modalité de référence. De ce fait, nous modélisons l'écart par rapport au placebo. Nous affichons ensuite les sommes pour vérifier les cohérences:

```r
#Recodage treatment - high
T1 <- rep(0,nrow(D))
T1[medData$treatment==1] <- 1
print(sum(T1))

#Recodage treatment - low
T2 <- rep(0,nrow(D))
T2[medData$treatment==2] <- 1
print(sum(T2))
```

```
> print(sum(T1))
[1] 20
> print(sum(T2))
[1] 30
```

LE MODÈLE: entrainement, ajustement, prediction

Définition du modèle

Dans la foulée, nous étudions events en fonction de treatment (T1, T2 encodé précédemment), alcohol et esteem:

```r
poisson_model <- glm(events ~ T1 + T2 + alcohol + esteem,
                     family = poisson(link = "log"),
                     data = medData)
summary(poisson_model)
```

L'output de summary:

```
> summary(poisson_model)

Call:
glm(formula = events ~ T1 + T2 + alcohol + esteem, family = poisson(link = "log"),
    data = medData)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.74912    0.22836  12.038  < 2e-16 ***
T1          -1.15316    0.14570  -7.914 2.48e-15 ***
T2          -0.45741    0.09939  -4.602 4.19e-06 ***
alcohol      0.00426    0.10315   0.041   0.9671
esteem      -0.03385    0.01387  -2.441   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 122.839  on 74  degrees of freedom
Residual deviance:  39.214  on 70  degrees of freedom
AIC: 316.74

Number of Fisher Scoring iterations: 4
```

Interprétations

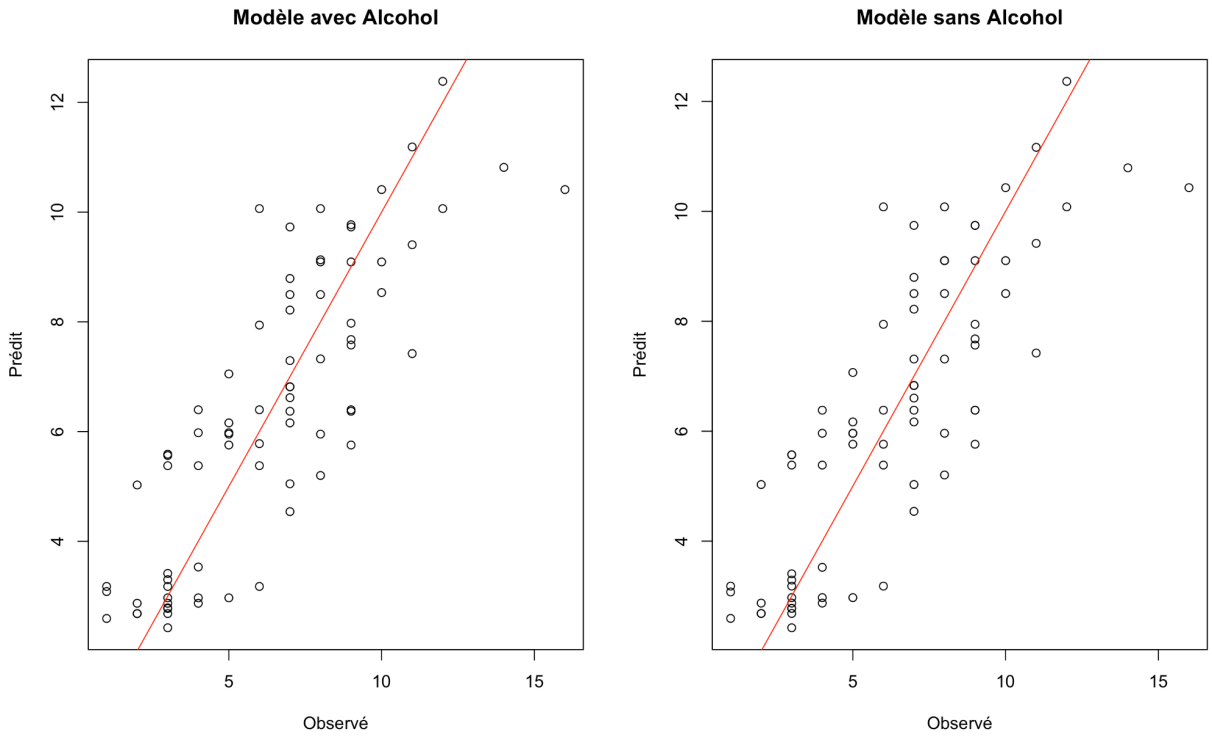
- Signification des coefficients
    - **Intercept** (2.74912) : C'est le logarithme du nombre attendu d'événements pour le groupe placebo (T1=0, T2=0), sans consommation d'alcool (alcohol=0) et avec une estime de soi nulle. Très significatif (p < 2e-16).
    - **T1** (-1.15316) : Par rapport au placebo, le traitement à haute dose réduit significativement le logarithme du nombre d'événements (p = 2.48e-15). Pour interpréter en termes de taux d'incidence : exp(-1.15316) = 0.32, donc le traitement à haute dose réduit le nombre d'événements d'environ **68%** par rapport au placebo.
    - **T2** (-0.45741) : Le traitement à faible dose réduit également significativement le logarithme du nombre d'événements (p = 4.19e-06). En termes de taux : exp(-0.45741) = 0.63, soit une réduction d'environ **37%** par rapport au placebo.
    - **alcohol** (0.00426) : L'effet de l'alcool est très faible et non significatif (p = 0.9671). Ce prédicteur pourrait être retiré du modèle.
    - **esteem** (-0.03385) : L'estime de soi a un effet négatif significatif (p = 0.0147). Pour chaque augmentation d'une unité d'estime de soi, le nombre attendu

d'événements est multiplié par exp(-0.03385) = 0.967, soit une réduction d'environ 3.3%.

- Qualité du modèle:
    - La déviance nulle (122.839) comparée à la déviance résiduelle (39.214) montre que le modèle explique une part importante de la variabilité.
    - L'AIC est de 316.74

Ajustements

Compte tenu de la non-significativité de la variable 'alcohol', nous avons trouvé judicieux de faire sans. Ci-dessous, la comparaison entre les deux modèles:



Le retrait de la variable 'alcohol' n'a pas impacté la qualité prédictive du modèle. Ceci confirme que la variable 'alcohol' n'apporte pas d'information significative. Le modèle sans “alcohol” est donc préférable

- ## Performances du modèle

```
> rmse <- sqrt(mean((medData$events - predict(new_poisson_model, type="response"))^2))
> rmse
[1] 1.775962
```

Pour mesurer les performances du modèle, nous avons calculé la Root Mean Square Error (RMSE) et avons obtenu une valeur de 1,78. Ce qui signifie que les prédictions du modèle s'écartent de 1.78 événements par rapport aux valeurs réelles, représentant une erreur de prédiction d'environ 2 événements en moyenne.