S8_1_GLM_Poisson

Table of contents

1 Plan du cours

1. Introduction : données de comptage, données longtiduinales, données en cluster
2. Modèles de Poisson pour les données de comptage

- Loi de Poisson
- Interprétation des coefficients
- Cas des données sur-dispersées (variance > moyenne)

2 Introduction : données de comptage, données longtiduinales, données en cluster

2.A Données de comptage

Les données de comptage sont des données qui représentent le nombre d'occurrences d'un événement dans un intervalle de temps ou d'espace donné.

- **Variable aléatoire discrète**
  - Valeur dans les entiers naturels positifs (0, 1, 2, …)
  - Exemples : nombre de visites à l'hôpital, nombre d'accidents de la route, nombre de naissances, etc.
- **Ordre de grandeur des comptages**
  - Souvent, les comptages sont de taille modérée (par exemple, de 0 à quelques dizaines).
  - Cependant, ils peuvent aussi être très élevés dans certains contextes (par exemple, le nombre de visiteurs sur un site web).
- **Caractéristiques des données de comptage**
  - Beaucoup de zéros (événements rares)
  - **Pas d'unité !** : les comptages sont des nombres absolus
  - Distribution asymétrique : la distribution est souvent biaisée à droite (peu de grandes valeurs)
  - Non normale : les données de comptage ne suivent pas une distribution normale
  - Relation moyenne-variance : dans une distribution de Poisson, **la variance est souvent égale voire supérieure à la moyenne (sur-dispersion)**
  - On peut comparer les deux car pas d'unité
- **Lien avec la distribution de Poisson**
  - La distribution de Poisson est souvent utilisée pour modéliser les données de comptage.
  - Elle est caractérisée par un paramètre λ qui représente à la fois la moyenne et la variance des comptages.
  - λ correspond au taux moyen d'occurrence de l'événement par unité de temps ou d'espace.
- **Notion d'exposition**
  - Dans certains cas, les données de comptage sont associées à une notion d'exposition, c'est-à-dire la durée ou la surface pendant laquelle les événements peuvent se produire.

- Par exemple, le nombre d'accidents de la route peut être rapporté au nombre de kilomètres parcourus.
- L'exposition est souvent utilisée comme un offset dans les modèles de régression pour ajuster les comptages en fonction de la durée ou de la surface d'observation.

Loi de Poisson

- Paramètre $\lambda > 0$ (ne peut pas être négatif)
- Espérance : $\mathbb{E}(X) = \lambda$
- Variance : Var$(X) = \lambda$

= distribution de probabilité discrète

- décrivant le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe,
- lorsque ces événements se produisent avec une moyenne constante et indépendamment du temps écoulé depuis le dernier événement.

$P(X = k) = \frac{\lambda^k}{k!} \times e^{-\lambda}$, pour $k = 0, 1, 2, ...$

2.A.1 Exemple de données de comptage

10 000 tirages d'une variable aléatoire

- avec loi de Poisson de paramètre $\lambda = 1$ (donc moyenne et variance égales à 1)
- avec loi de Poisson de paramètre $\lambda = 5$ (donc moyenne et variance égales à 5)
- avec loi de Poisson de paramètre $\lambda = 20$ (donc moyenne et variance égales à 20)

	Pandemic year 1		Pandemic year 2	
	Cohen's d effect size (95% CI)	p value	Cohen's d effect size (95% CI)	p value
Whole cohort				
Executive function				
Verbal Reasoning	0·15 (0·12 to 0·17)	<0·0001	0·02 (0·00 to 0·04)	0·12
Working memory				
Paired Associate Learning	0·77 (0·30 to 0·37)	<0·0001	0·74 (0·72 to 0·76)	<0·0001
Self-Ordered Search	0·15 (0·13 to 0·18)	<0·0001	0·15 (0·13 to 0·18)	<0·0001
Digit Span	0·19 (0·17 to 0·21)	<0·0001	0·14 (0·12 to 0·16)	<0·0001
Composite	0·51 (0·49 to 0·53)	<0·0001	0·47 (0·44 to 0·49)	<0·0001

Plus le paramètre λ est grand, plus la distribution ressemble à une distribution normale.

Pour $\lambda = 1$ ou 5, la distribution est asymétrique et concentrée sur les petites valeurs.

Donc cette loi a moins d'intérêt pour des comptages élevés.

2.A.2 Approximation binomiale / Poisson

- Loi binomiale $\mathfrak{B}(n, p)$ = nombre de succès dans n essais indépendants, avec probabilité p de succès à chaque essai.
  - Exemple : nombre d'appels au SAMU dans une journée parmi 10 000 habitants, avec une probabilité p d'appel faible.

Démarche de modélisation :

- La ville contient un nombre n d'habitants (très grand)
- Chaque habitant a une probabilité p (très petite) d'appeler le SAMU dans la journée
- Les appels surviennent de manière indépendante entre les habitants
- Le nombre total d'appels X suit une loi binomiale $B(n, p)$, d'espérance $\lambda = n \times p$

Cognitive decline in older adults in the UK during and after the COVID-19 pandemic: a longitudinal analysis of PROTECT study data

Anne Corbett, Gareth Williams, Byron Creese, Adam Hampshire, Vincent Hayman, Abbie Palmer, Akos Filakovzsky, Kathryn Mills, Jeffrey Cummings, Dag Aarsland, Zunera Khan, Clive Ballard

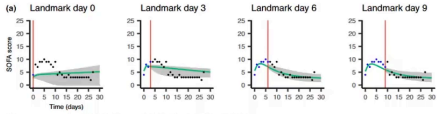
Summary

Background Although the long-term health effects of COVID-19 are increasingly recognised, the societal restrictions during the COVID-19 pandemic hold the potential for considerable detriment to cognitive and mental health, particularly because major dementia risk factors—such as those related to exercise and dietary habits—were affected during this period. We used longitudinal data from the PROTECT study to evaluate the effect of the pandemic on cognition in older adults in the UK.

En augmentant n et en diminuant p de manière à ce que le produit $n \times p$ reste constant, la loi binomiale $\mathfrak{B}(n, p)$ se rapproche de la loi de Poisson (de paramètre $\lambda = n \times p$).

2.A.3 Surdispersion

Modèles non linéaires mixtes (NLMM)
Ex : Evolution d'un biomarqueur



Lavalley-Morelle A et al. CPT Pharmacometrics Syst Pharmacol. 2022 Nov;11(11):1472-1484.
doi: 10.1002/psp4.12856

Exemple : données de transcriptomique Bowtie

- Chaque gène a un niveau d'expression (entier positif) sur chaque échantillon
- On représente, pour chaque gène, la variance en fonction de la moyenne
- Sous l'hypothèse d'une loi de Poisson, les points devraient tomber proche de la diagonale $y = x$

Or : ce n'est pas la cas !

Il faut donc une loi plus générale (ex. loi binomiale négative) pour modéliser ces données.

2.A.4 Excès de 0

Modèles linéaires généralisés mixtes (GLMM)
Ex : logistique mixte

Bouzid D et al. BMC Med Educ. 2022 Dec 13;22(1):861. d
doi: 10.1186/s12909-022-03919-1

Table 1 Summary of the factors influencing students' scores variability

	Station 1				Sta
	Score median- IQR (Q1-Q3) or % success	Agreement N (%)	Students' ICC (%)	Raters'ICC (%)	Sco me or
score (/100)	60 (50-70)	21 (25)	60,2	23,0	60 (
Item 1	65	76 (89)	97	3	61
Item 2	93	75 (88)	82	18	64
Item 3	94	74 (87)	*	*	99
Item 4	95	76 (89)	*	*	39
Item 5	33	67 (79)	31	48	38
Item 6	48	80 (94)	92	0	79

Exemple : nombre d'épisode de migraine durant la dernière année

- Les nons malades n'ont pas d'épisode (0) = zéros "structurels" ;
- Les malades peuvent aussi ne pas avoir d'épisode = zéros "aléatoires", mais aussi 1, 2, …
- Il en résulte un excès de 0 par rapport à une distribution "standard"

Solution : un modèle de mélange (on modélise à part les "faux" 0 : "zero-inflated model")

- 1ère partie du modèle estime la probabilité d'être toujours à 0 (par exemple "non malade") ;
- 2nde partie du modèle estime le comptage conditionnel (0, 1, 2, …) chez les individus susceptibles d'avoir des événements (par exemple les malades), souvent avec une Poisson ou une binomiale négative.
- Ce type de modèle permet de mieux tenir compte de la structure des données et d'obtenir des estimations plus réalistes des effets des covariables.

3 Modèle de poisson

3.A Modèle

On observe

- une réponse Y de type comptage (0, 1, 2, …)
- des covariables explicatives X1, X2, …, Xp, dont on souhaite examiner l'association avec la moyenne de Y.

On modélise le paramètre λ de la loi de Poisson (moyenne et variance de Y) via une fonction de lien exponentielle :

$$Y \sim Poisson(\lambda)$$

$$\lambda = \mathbb{E}(Y|X) = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Donc λ est l'exponentielle d'un prédicteur linéaire = fonction log

Le lien assure que λ est toujours positif.

C'est différent d'une régression logistique où la moyenne est un "expit" du prédicteur linéaire (donc fonction logit inverse)

> **Note**
>
> **Modèle linéaire** : suppose que chaque variable explicative a un effet additif sur la moyenne de la variable réponse, avec une relation linéaire directe.
>
> - Linéaire = proportionnel et additif : chaque variable explicative influence la moyenne de la variable réponse de manière proportionnelle à son coefficient.
> - Additif car les effets des différentes variables explicatives s'additionnent pour déterminer la moyenne de la variable réponse.

Liens en régressions = fonction de lien = façon de transformer la variable qu'on veut expliquer en quelque chose qu'on peut modéliser linéairement.

La fonction de lien doit être adaptée à la variable

Exprimer une probabilité p (entre 0 et 1) : on la transforme entre -infini et +infini avec la fonction logit ou probit

- logit : fonctionne en faisant le logarithme du rapport entre la probabilité d'un événement et la probabilité de son complément (1 - p)
- probit : utilise la fonction normale cumulative inverse pour transformer la probabilité en une valeur sur l'axe des réels.
  - on prend une probabilité p
  - on trouve la valeur x telle que la probabilité d'obtenir une valeur ≤ à x dans une distribution normale standard soit égale à p.
  - Cela permet de modéliser des données binaires en utilisant une approche basée sur la distribution normale.
- Exprimer une moyenne μ positive (comptage) : on la transforme entre 0 et +infini avec la fonction log

> **Important**
>
> **Régression linéaire** : moyenne est égale au prédicteur linéaire
>
> $$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$
>
> **Régression logistique** : moyenne est une fonction "expit" du prédicteur linéaire
>
> $$\mathbb{E}(Y|X) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$
>
> **Régression de Poisson** : moyenne est une fonction exponentielle du prédicteur linéaire
>
> $$\mathbb{E}(Y|X) = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Bonus : tableau comparatif liens GLM

3.B Interprétation des coefficients

1. Si la variable X est binaire

$\lambda_0 = \mathbb{E}(Y \| X = 0) = exp(\beta_0)$
$\lambda_1 = \mathbb{E}(Y \| X = 1) = exp(\beta_0 + \beta_1) = exp(\beta_0) \times exp(\beta_1)$
Donc le ratio des moyennes est
$\frac{\lambda_1}{\lambda_0} = \frac{\mathbb{E}(Y \| X=1)}{\mathbb{E}(Y \| X=0)} = \frac{exp(\beta_0+\beta_1)}{exp(\beta_0)} = exp(\beta_1)$
Le coefficient s'inteprète comme un **log d'un risque relatif**, c'est à dire le logarithme du ratio des moyennes entre les deux groupes.

2. Si la variable X est catégorielle
    - On choisit une catégorie de référence (par exemple la première catégorie)
    - Comme das les autres modèles linéaires ou GLM
    - Chaque coefficient β_k correspond au log du ratio des moyennes entre la catégorie k et la catégorie de référence.
3. Si la variable X est continue
    - exp(β_1) correspond au RR associé à une augmentation de 1 unité de X.

Dans un modèle multivariable, l'hypothèse est un effet additif** des variables explicatives sur le log(λ), donc un effet **multiplicatif** sur λ.**

Tip

Exemple
Imaginons que
$\lambda =$ moyenne du nombre de visites aux urgences par mois.
On ajuste un modèle de Poisson :

$$\log(\lambda) = \beta_0 + 0.4\,(\text{fumeur}) + 0.7\,(\text{diabétique})$$

Cas 1 — Personne non fumeuse et non diabétique
Les deux indicateurs valent 0.

$$\lambda = e^{\beta_0}$$

Cas 2 — Personne fumeuse uniquement

$$\lambda = e^{\beta_0} \times e^{0.4}$$

Comme $e^{0.4} \approx 1.49$, cela correspond à **+49%** sur la moyenne du comptage.

Cas 3 — Personne fumeuse + diabétique

$$\lambda = e^{\beta_0} \times e^{0.4} \times e^{0.7}$$

Comme $e^{0.7} \approx 2.01$, cela correspond à **+101%** supplémentaires.
Ainsi, l'augmentation est :

- 1.49 (effet fumeur)
- $\times 2.01$ (effet diabète)
- = **multiplication des deux effets**

Les effets sont donc **additifs sur** $\log(\lambda)$,
mais ils deviennent **multiplicatifs sur** λ après exponentiation.
C'est pourquoi, dans une régression de Poisson,
les coefficients exponentiés e^{β} s'interprètent comme des **risques relatifs**.

3.C Tests statistiques

Comme pour les autres GLM, l'ordinateur estime, par la méthode du maximum de vraisemblance :

- les coefficients β.
- les erreurs standards associées.

On dispose donc :

- du test de Wald pour chaque coefficient (H0 : $\beta_k = 0$)
- du test du rapport de vraisemblance pour le modèle global (H0 : tous les $\beta_k = 0$) si l'on compare 2 modèles emboités.

! Important

Attention à la surdispersion !

- p-value du test de Wald peut être trop petite (risque de faux positifs augmenté).
- 95% IC peut être trop étroit = diminution de la couverture réelle.

3.D Exemple avec R

3.D.1 1

On crée une variable X simulant 10 000 tirages et suivant une loi de Poisson de paramètre 3.

On crée ensuite une variable $Y = 4 \times X$.

Lorsqu'on multiplie une variable par une constante :

- Sa moyenne est multipliée par cette constante
- Sa variance est multipliée par le carré de cette constante

Donc la variance de Y est 16 fois celle de X = sur-dispersion (variance = 4* la moyenne)

```
X <- rpois(10000,3) #simule 10000 tirages d'une loi de poisson de paramètre 3
Y <- 4*X
fit <- glm(Y ~ 1, family = poisson()) #Y ~ 1 veut dire qu'il n'y a pas de covariable explicativ
summary(fit)
```

```
Call:
glm(formula = Y ~ 1, family = poisson())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.48260    0.00289     859   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42953  on 9999  degrees of freedom
Residual deviance: 42953  on 9999  degrees of freedom
AIC: 83341

Number of Fisher Scoring iterations: 5
```

- Coefficient estimé : 2.48
- Erreur standard : 0.00298
- z value = rapport coefficient estimé / erreur standard = 858
- p-value < 2e-16 (p value petite car grand échantillon)

Dispersion parameter for poisson family taken to be 1 : indique que l'on a supposé une variance égale à la moyenne (pas de sur-dispersion prise en compte)

Pour trouver le paramètre λ de la loi de Poisson, on exponentie le coefficient estimé :

```
exp(coef(fit))
```

```
(Intercept)
    11.9724
```

On trouve une moyenne de 11.96 (≈ 12), ce qui est cohérent avec la moyenne de Y (4*3 = 12).

Donc 2 messages :

1. Loi de Poisson : moyenne = variance et données peuvent être sur-dispersées
2. Même en présence de sur-dispersion, l'estimation de la moyenne par le modèle de Poisson reste correcte.

3.D.2 2

2 vecteurs :

- $Y1$ contient 100 tirages d'une loi de Poisson de paramètre 3 (moyenne théorique = 3)
- $Y2$ contient 100 tirages d'une loi de Poisson de paramètre 4 (moyenne théorique = 4)

Donc RR théorique = $\text{RR}_{\text{theorique}} = \frac{\lambda_1}{\lambda_0} = \frac{4}{3} \approx 1{,}33.$

Objectif en ajustant un modèle de Poisson sur la variable Y est d'estimer ce RR à partir des données simulées.

Variable de comptage Y définie par :

$$Y = 4 \times (Y1 + Y2)$$

Vecteur explicatif X :

- Indicateur binaire
- prenant la valeur 0 pour les 100 premières observations et la valeur 1 pour les 100 suivantes.

Puis ajuster un modèle de Poisson :

$$\log(\lambda) = \beta_0 + \beta_1 X$$

et afficher le résumé du modèle.

```r
set.seed(123)

# Deux vecteurs de Poisson
Y1 <- rpois(100,3)
Y2 <- rpois(100,4)

# Variable de comptage finale
Y <- 4 * c(Y1, Y2)

# Variable explicative binaire
X <- c(rep(0, 100), rep(1, 100))
```

Estimation du modèle de Poisson :

```r
# Modèle de Poisson
fit <- glm(Y ~ X, family = poisson())
summary(fit)
```

```
Call:
glm(formula = Y ~ X, family = poisson())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.46470    0.02916  84.522   <2e-16 ***
X            0.33258    0.03821   8.704   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 813.15  on 199  degrees of freedom
Residual deviance: 736.34  on 198  degrees of freedom
AIC: 1594.4

Number of Fisher Scoring iterations: 5
```

Intercept : 2,46 (correspond à la strate X=0)

Variable X : 0,33 (correspond à la différence entre les strates X=1 et X=0)

On exponentie le coefficient de X pour l'interpréter comme un risque relatif :

```
exp(coef(fit))
```

```
(Intercept)           X
  11.760000    1.394558
```

Donc finalement, malgré la sur-dispersion, on trouve un RR de 1,39, ce qui est cohérent avec le ratio des moyennes théoriques (4/3 = 1,33).

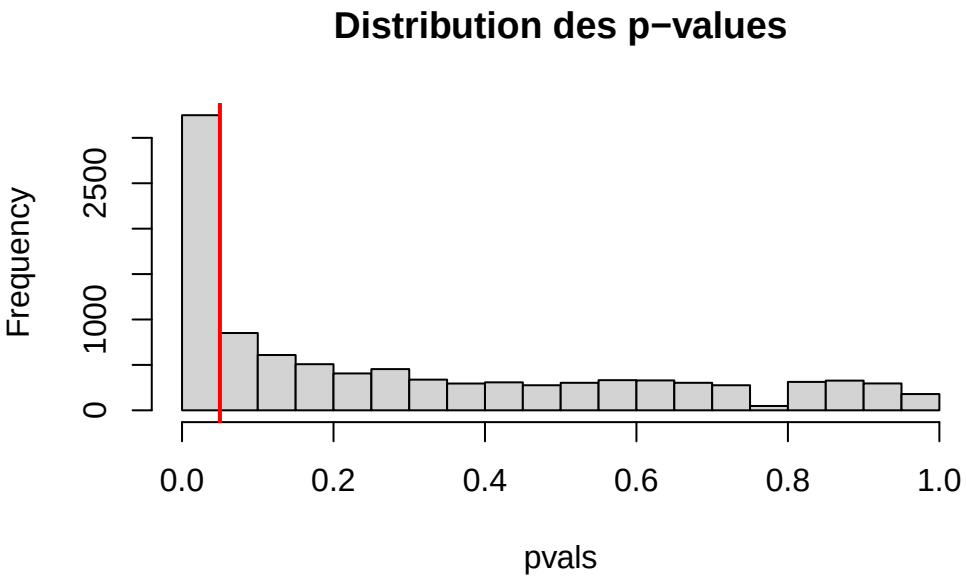
3.D.3 3

On souhaite étudier la distribution des p-values obtenues lorsqu'on répète 10 000 fois la simulation suivante :

1. On simule deux vecteurs :
   - $Y1$: 100 tirages d'une loi de Poisson de paramètre 3
   - $Y2$: 100 tirages d'une loi de Poisson de paramètre 3
2. On construit ensuite la variable de comptage :
   - $Y = 4 \times (Y1 + Y2)$
3. On définit une variable explicative binaire :
   - $X = \{0, 0, \dots, 0, 1, 1, \dots, 1\}$
   - avec 100 zéros suivis de 100 1.

4. On ajuste un modèle de Poisson :
   - $\log(\lambda) = \beta_0 + \beta_1 X$
5. On extrait la p-value associée à β_1.
6. On répète cette simulation 10 000 fois et on visualise la distribution empirique des p-values.

```r
set.seed(123)

# Variable explicative (fixe)
X <- c(rep(0, 100), rep(1, 100))

# Fonction qui renvoie la p-value du coefficient de X
test <- function() {

    # Simulations Poisson indépendantes
    Y1 <- rpois(100, lambda = 3) #Y1 suit une loi de Poisson de paramètre 3
    Y2 <- rpois(100, lambda = 3) #Y2 suit une loi de Poisson de paramètre 3 aussi !

    # Variable de réponse : concaténer Y1 et Y2, puis multiplier par 4 pour générer de la sur-d
    Y <- 4 * c(Y1, Y2)

    # Modèle de Poisson
    fit <- glm(Y ~ X, family = poisson)

    # p-value du coefficient de X (ligne 2, colonne 4) = pvalue du test de Wald
    pval <- summary(fit)$coefficients[2, 4]

    return(pval)
}

# 10 000 répétitions
pvals <- replicate(10000, test())

# Histogramme des p-values
hist(pvals, breaks = 30, main = "Distribution des p-values")
abline(v = 0.05, col = "red", lwd = 2)
```



Combien de p-values sont inférieures à 0,05 en valeur absolue ?

```
sum_05 <- sum(pvals < 0.05)
prop_05 <- sum_05 / length(pvals) * 100
cat("Il y a", sum_05, "p-values < 0.05 soit", round(prop_05, 1), "%\n")
```

```
Il y a 3249 p-values < 0.05 soit 32.5 %
```

On constate une augmentation du nombre de p-values faibles (inférieures à 0,05) par rapport à une distribution uniforme.

Du fait de la sur-dispersion, le test de Wald produit trop souvent des p-values faibles, ce qui augmente le risque de faux positifs.

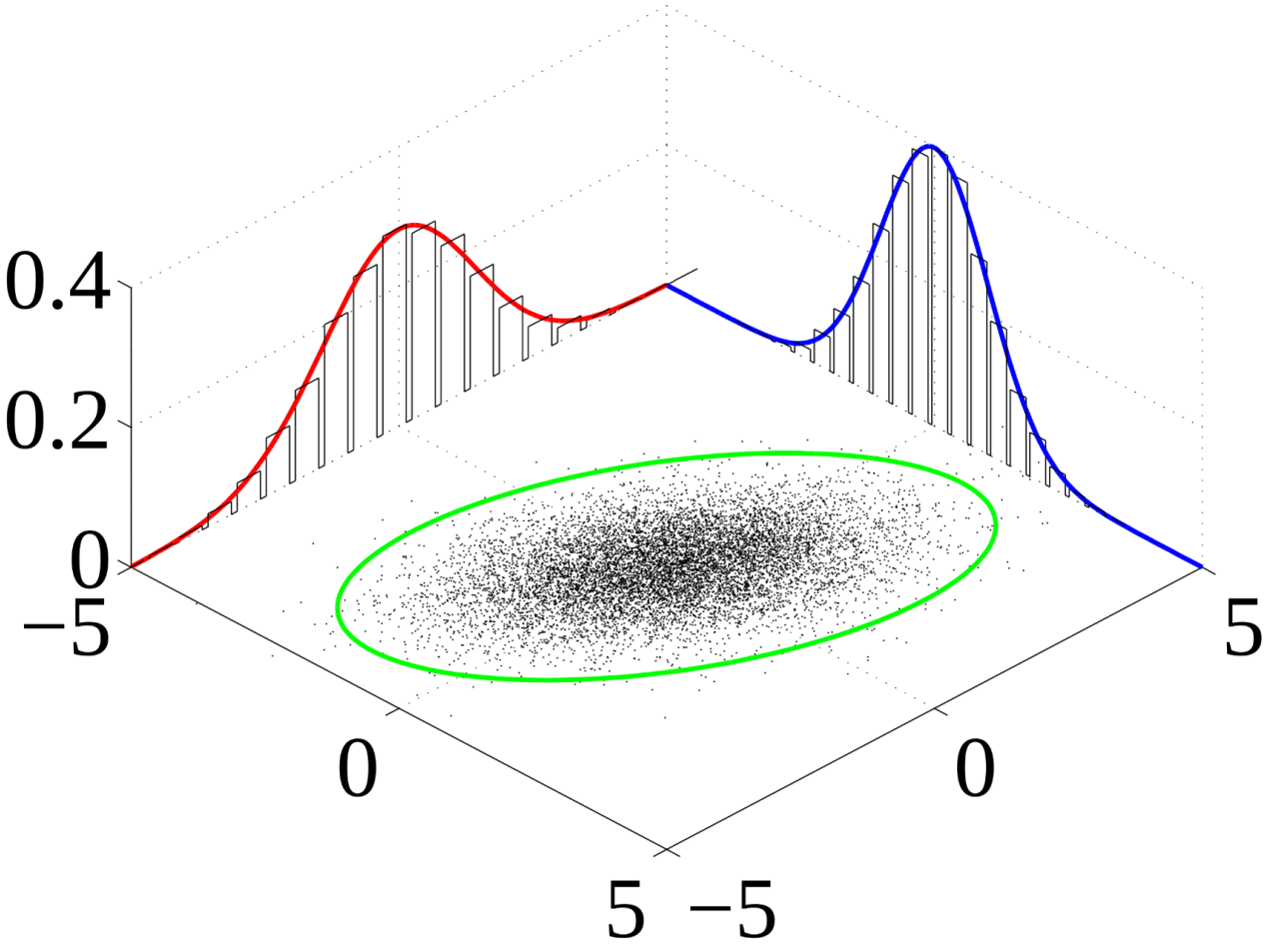
Le problème, c'est pas l'estimation du RR (de la moyenne), mais l'inférence statistique (tests, IC) qui est biaisée par la sur-dispersion (= la variance importante de la variable).

3.E Variance "sandwich"

Dans un modèle de Poisson, on cherche à "maximiser" la log-vraisemblance :

- **vraisemblance** : se fait dans tous les modèles statistiques
  - objectif : trouver les paramètres qui rendent les données observées les plus probables
  - fonction de vraisemblance : fonction qui mesure à quel point un certain choix de paramètres rend les données observées probable. C'est une mesure de la plausibilité des données en fonction des paramètres du modèle.
- **log-vraisemblance** : logarithme de la fonction de vraisemblance. permet de rendre plus simple à manipuler (produit devient somme)
- **maximum de vraisemblance** : méthode d'estimation des paramètres du modèle en choisissant les valeurs qui maximisent la log-vraisemblance.
- **courbure de la log-vraisemblance** :

- plus la courbe est pointue autour de son maximum, plus l'estimation est précise (erreur standard faible)
- plus la courbe est plate autour de son maximum, moins l'estimation est précise (erreur standard élevée)
- la matrice de variance/covariance des paramètres est liée à la courbure de la log-vraisemblance.



Donc la matrice de variance/covariance des paramètres peut être estimé :

1. Directement par l'inverse de la courbure de la log-vraisemblance au voisinage du maximum de vraisemblance (estimation classique)
    - Mais biaisé si sur-dispersion ++
2. Avec l'estimateur "sandwich" (robuste aux violations des hypothèses du modèle, notamment la sur-dispersion)
    - Dans lequel la matrice de la courbure est multipliée gauche et à droite par une autre matrice
    - Obtient la variance "sandwich"
    - Méthode très robuste à une mauvaise "spécification" du modèle

NB : "spécification" = fait de choisir un modèle statistique particulier pour représenter les données,

en supposant que ce modèle capture correctement la structure sous-jacente des données.

3.E.1 Exemple avec R

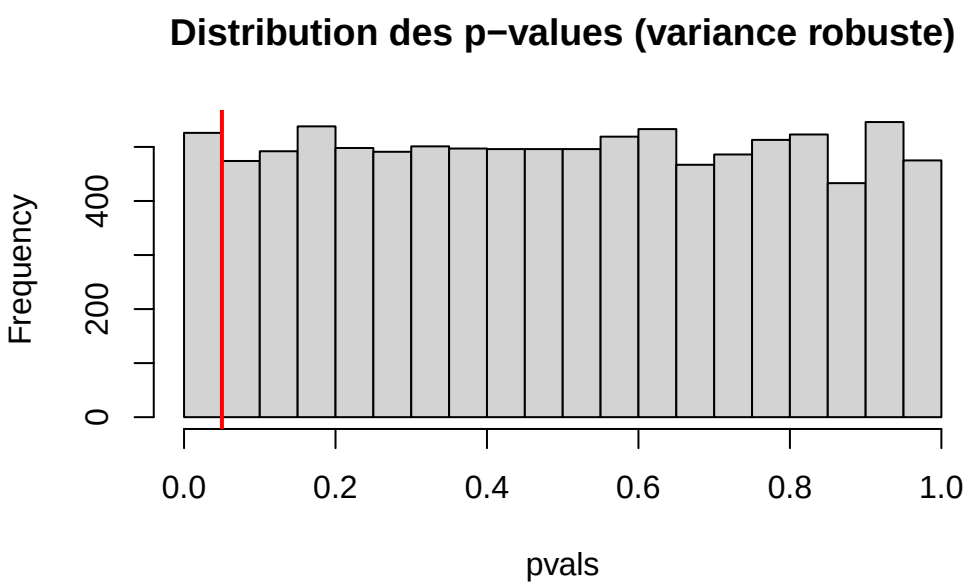
1. On génère
   - $Y1 \sim \text{Pois}(3)$
   - $Y2 \sim \text{Pois}(3)$
2. On définit :
   - $Y = 4 \times (Y1 + Y2)$
3. On ajuste un modèle de Poisson :
   - $\log(\lambda) = \beta_0 + \beta_1 X$
4. Mais cette fois, on utilise un estimateur de variance robuste (sandwich)
   - donc la p-value provient de : `coeftest(fit, vcov. = sandwich)`
5. On extrait la p-value du coefficient de X.
6. On répète 10 000 fois et on trace l'histogramme des p-values.

```r
set.seed(123)

library(sandwich)
library(lmtest)
```

```
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
# Variable explicative fixe
X <- c(rep(0, 100), rep(1, 100))

# Fonction de simulation
test <- function() {

    Y1 <- rpois(100, 3)
    Y2 <- rpois(100, 3)
    Y  <- 4 * c(Y1, Y2)

    fit <- glm(Y ~ X, family = poisson)

    # p-value robuste (sandwich)
```

```r
    pv <- coeftest(fit, vcov. = sandwich)[2, 4] #P-value du test de Wald avec variance robuste

    return(pv)
}

# 10 000 répétitions
pvals <- replicate(10000, test())

# Histogramme des p-values
hist(pvals, breaks = 30, main = "Distribution des p-values (variance robuste)")
abline(v = 0.05, col = "red", lwd = 2)
```



Combien de p-values sont inférieures à 0,05 en valeur absolue ?

```r
sum_05 <- sum(pvals < 0.05)
prop_05 <- sum_05 / length(pvals) * 100
cat("Il y a", sum_05, "p-values < 0.05 soit", round(prop_05, 1), "%\n")
```

```
Il y a 526 p-values < 0.05 soit 5.3 %
```

L'estimation robuste de la variance permet de corriger le problème de sur-dispersion

3.F Traitement de la surdispersion

1. Correction des variances des estimateurs (estimateur robuste de la variance, bootstrap)
    1. **Estimateur robuste de la variance (sandwich)**
        - Simple à mettre en œuvre
        - Permet d'obtenir des tests et IC corrects
        - Ne modifie pas l'estimation des coefficients R
    2. **Bootstrap**

- Permet d'obtenir des intervalles de confiance et des tests basés sur la distribution empirique des estimateurs
- Plus coûteux en calcul
- Ne modifie pas l'estimation des coefficients R

2. Utilisation d'une autre loi pour modéliser les données
    1. **Quasi-Poisson** : (`family = quasipoisson()` dans R) - Modifie l'estimation de la variance en ajoutant un paramètre de dispersion estimé à partir des données - Permet d'obtenir des tests et IC corrects
    2. **Binomiale négative avec θ fixé** : (`family = negative.binomial(theta = 1)` dans R, avec `theta` paramètre de dispersion fixé)
        - Modifie à la fois l'estimation des coefficients et de la variance
        - Permet d'obtenir des tests et IC corrects
    3. **Binomiale négative avec θ estimé** : (`family = negative.binomial()` dans R, avec `theta` paramètre de dispersion estimé à partir des données) ou fonction `glm.nb()` du package `MASS`

    - Modifie à la fois l'estimation des coefficients et de la variance
    - Permet d'obtenir des tests et IC corrects

3.G Traitement des excès de 0

- Modèles « hurdle » et « zero-inflated » : mélange entre
    - Un modèle logistique pour les zéros
    - Et un modèle de compte pour le reste
- Différences (ténues) dans la formulation mathématique
    - Fonction `zeroinfl()` ou `hurdle()` du package `pscl`
    - Les covariables peuvent être différentes pour les deux modèles (problème de sélection)

Modèle `hurdle` :

- Modélise la probabilité d'avoir un comptage > 0 avec un modèle logistique
- Puis modélise la distribution des comptages conditionnels à être > 0 avec un modèle de Poisson ou binomiale négative tronquée

Modèle `zero-inflated` :

- Modélise la probabilité d'être dans le groupe des zéros structurels avec un modèle logistique
- Puis modélise la distribution des comptages avec un modèle de Poisson ou binomiale négative pour les individus susceptibles d'avoir des événements (y compris les zéros aléatoires)

4 Résumé

4.A Données de comptage et histoire binomiale → Poisson

Les données de comptage correspondent au **nombre d'occurrences** d'un événement dans un intervalle (temps, espace, patients, etc.).
Exemples : nombre d'appels au SAMU par jour, nombre de visites aux urgences par mois, nombre d'épisodes de migraine sur un an.

On peut raconter l'histoire suivante :

- Il y a un nombre très élevé d'unités (habitants, patients, jours à risque, etc.), noté n.
- Chaque unité a une **très petite probabilité** p de présenter l'événement sur la période.
- Les événements entre unités sont supposés indépendants.

Le nombre total d'événements X suit alors **une loi binomiale** $\mathcal{B}(n, p)$, d'espérance

$$\mathbb{E}(X) = np.$$

Quand : - n est **très grand**, - p est **très petit**, - et le produit np reste **constant** (noté λ),

alors la binomiale se **rapproche d'une loi de Poisson** de paramètre λ :

- $X \sim \text{Poisson}(\lambda)$,
- $\mathbb{E}(X) = \lambda$,
- $\text{Var}(X) = \lambda$.

C'est ce qui justifie l'usage de la loi de Poisson pour de nombreux phénomènes de comptage.

4.B Modèle de Poisson et interprétation des coefficients

Dans un **GLM de Poisson**, on modélise la moyenne conditionnelle

$$\lambda_i = \mathbb{E}(Y_i \mid X_i)$$

d'un comptage Y_i par un lien logarithme :

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

Ainsi :

- $\lambda_i = \exp(\beta_0 + \cdots + \beta_p X_{ip}) > 0$ (moyenne toujours positive),
- le modèle est **linéaire** sur $\log(\lambda)$.

Interprétation des coefficients :

- Si X est **binaire** (0 / 1), alors
  - $\mathbb{E}(Y \mid X = 0) = \lambda_0$,
  - $\mathbb{E}(Y \mid X = 1) = \lambda_1$,

  - et
    * $\frac{\lambda_1}{\lambda_0} = \exp(\beta_1)$.
    * $\exp(\beta_1)$ est donc un **rapport de moyennes** (souvent interprété comme un **risque relatif**).

- Si X est **catégorielle**, chaque coefficient correspond au **log-rapport de moyenne** par rapport à une catégorie de référence (comme en régression linéaire ou logistique).
- Si X est **continue**, $\exp(\beta_1)$ est le **facteur multiplicatif** sur la moyenne associé à une augmentation d'une unité de X.

En multivariable :

- les effets sont **additifs** sur $\log(\lambda)$,
- donc **multiplicatifs** sur λ après exponentiation,
- les $\exp(\beta)$ s'interprètent comme des **risques relatifs / rapports de moyennes ajustés**.

4.C Surdispersion : définition, conséquences et illustration R

Le modèle de Poisson impose

$$\text{Var}(Y \mid X) = \mathbb{E}(Y \mid X) = \lambda.$$

En pratique, on observe fréquemment :

- une **variance empirique** nettement **supérieure** à la moyenne,
- c'est la **surdispersion** :
  - $\text{Var}(Y) > \mathbb{E}(Y)$.

Dans le code, plusieurs exemples montrent ce phénomène :

- Création d'une variable $X \sim \text{Poisson}(3)$,
- puis définition de $Y = 4X$.

Alors :

- $\mathbb{E}(Y) = 4\,\mathbb{E}(X) = 4 \times 3 = 12$,
- $\text{Var}(Y) = 4^2\,\text{Var}(X) = 16 \times 3 = 48$,

ce qui donne une variance **4 fois plus grande** que la moyenne.

Même constat dans la simulation plus complexe :

- $Y_1 \sim \text{Poisson}(3)$, $Y_2 \sim \text{Poisson}(3)$,
- $Y = 4(Y_1 + Y_2)$.

Là encore, la variance est **très supérieure** à la moyenne, donc il y a surdispersion.

Effets sur l'inférence :

- Les **coefficients** $\hat{\beta}$ (et donc les rapports de moyennes) restent souvent **correctement estimés**.

- En revanche, les **erreurs standards** calculées sous le modèle de Poisson supposent à tort $\text{Var}(Y) = \mathbb{E}(Y)$.
- En présence de surdispersion :
- les erreurs standards sont **trop petites**,
- les tests de Wald donnent **trop de p-values petites**,
- les intervalles de confiance sont **trop étroits**,
- l'**erreur de type I** (faux positifs) est **inflée**.

C'est ce que montrent les simulations R :

- on répète 10 000 fois la génération de données surdispersées,
- on ajuste un modèle de Poisson standard,
- on récupère la p-value du coefficient de X à chaque fois,
- l'histogramme des p-values montre **un excès de valeurs < 0,05**,
- la proportion de p-values < 0,05 dépasse largement les 5 % attendus sous H_0.

Conclusion :
le problème n'est pas l'estimation des moyennes / RR, mais la validité des tests et des IC si l'on ne corrige pas la surdispersion.

4.D Variance robuste (« variance sandwich »)

L'idée générale est la suivante :

- On garde le **modèle de Poisson** pour estimer les $\hat{\beta}$.
- On change **la façon d'estimer la variance de** $\hat{\beta}$.
- On remplace la variance « classique » (basée sur la courbure de la log-vraisemblance sous Poisson) par une **variance robuste**, appelée **variance sandwich**.

Cela a pour effet :

- de **ne pas modifier** les estimations $\hat{\beta}$,
- mais d'**augmenter** les erreurs standards lorsque la surdispersion est présente,
- donc de **corriger** les p-values et les intervalles de confiance.

Dans le code :

- 1 ajustement d'abord d'un `glm(..., family = poisson)`,
- puis

```r
library(sandwich)
library(lmtest)
coeftest(fit, vcov. = sandwich)
```

```
z test of coefficients:
```

```
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 2.464704   0.054807 44.9704 < 2.2e-16 ***
X           0.332577   0.070939  4.6882 2.756e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `sandwich` calcule la matrice de variance robuste,
- `coeftest()` recalcule les tests en utilisant cette variance.

En répétant la simulation 10 000 fois avec la variance robuste :

- l'histogramme des p-values devient beaucoup plus proche de l'uniforme,
- la proportion de p-values < 0,05 redevient ≈ 5 %,
- le contrôle de l'erreur de type I est rétabli.

Message clé :

- En présence de surdispersion, on doit obligatoirement l'adresser,
- La variance sandwich est une solution simple et robuste pour obtenir des tests valides sans changer les coefficients.

□

4.E Excès de zéros : modèles Hurdle et Zero-Inflated

En plus de la surdispersion, les données de questionnaire ou de pratique clinique présentent souvent un excès de zéros :

- certains individus sont structurellement à 0 (non concernés) : zéros « structurels »,
- d'autres pourraient avoir des événements mais en ont 0 sur la période : zéros « aléatoires ».

Une Poisson simple ne peut pas produire autant de zéros, même en étant surdispersée. On utilise alors des modèles à deux composantes.

4.E.1 Modèle Hurdle

Principe : 1. Partie 1 : franchir la marche (hurdle)

```
-   On modélise la probabilité d'avoir un comptage strictement positif :

-   $\mathbb{P}(Y > 0 \mid X)$.

-   Typiquement avec une régression logistique.
```

2. Partie 2 : distribution conditionnelle de Y sachant $Y > 0$
    - On travaille uniquement sur les observations avec $Y > 0$.
    - On ajuste une Poisson tronquée en 0 ou une binomiale négative tronquée.

Les zéros sont donc complètement sortis de la partie « compte » et expliqués à part.

En R, on peut utiliser la fonction `hurdle()` du package `{pscl}`.

4.E.2 Modèle Zero-Inflated

Autre histoire possible :

- une proportion des individus est dans un état « toujours 0 » (non exposés, non malades, etc.),
- le reste suit un modèle de comptage classique (Poisson ou binomiale négative), qui peut lui-même produire des zéros.

On a alors : 1. Un modèle (souvent logistique) pour la probabilité d'être un zéro structurel.

2. Un modèle de comptage pour les individus susceptibles d'avoir des événements (incluant des zéros « aléatoires »).

Ce qu'on appelle :

- des « vrais zéros » de la loi de Poisson,
- et des « faux zéros » venant de personnes non concernées.

En R, cela se fait par exemple avec `zeroinfl()` du package `{pscl}`.

Difficultés méthodologiques

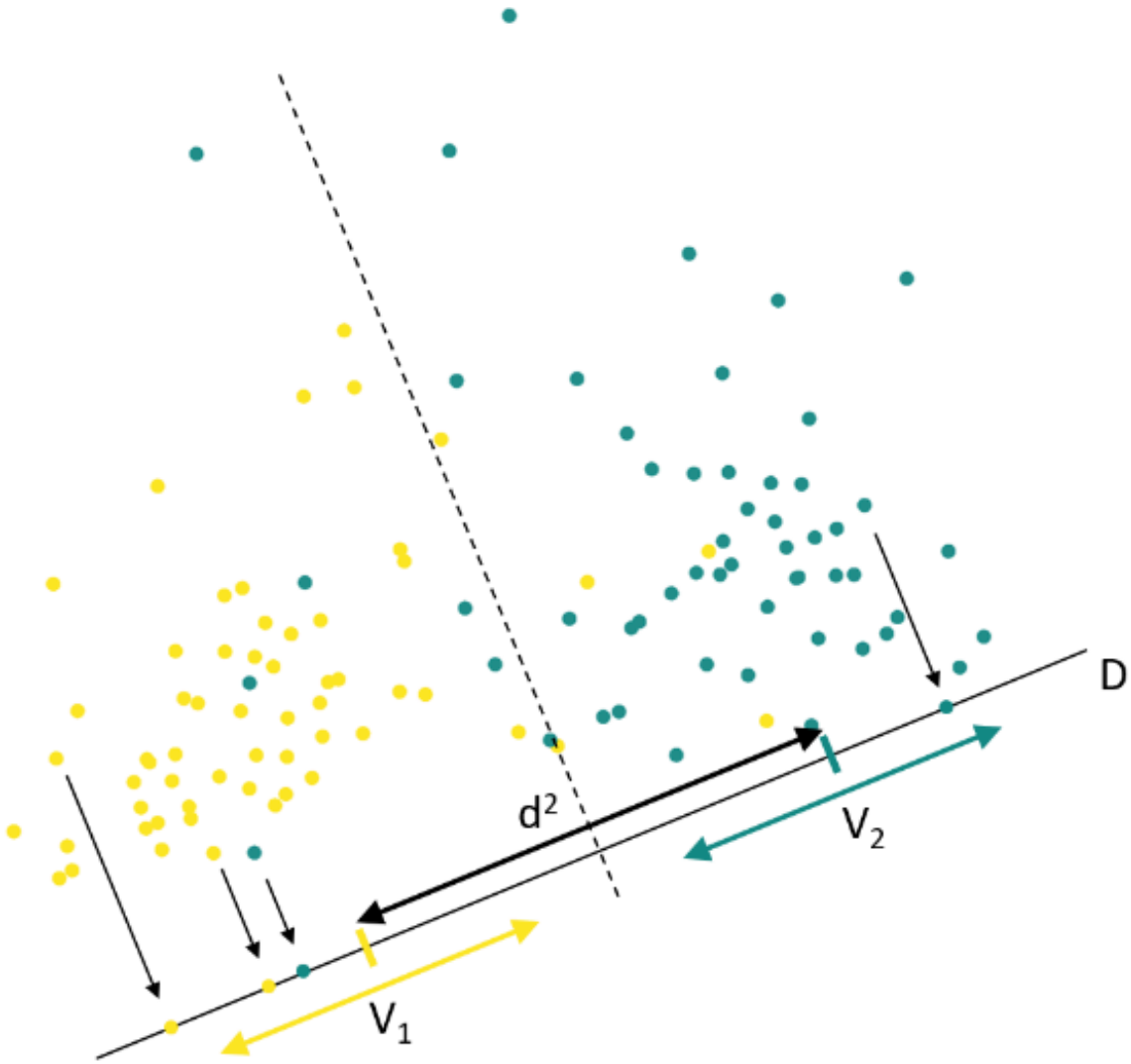
Ces modèles à deux parties posent des questions de spécification :

- Quelle covariable mettre dans la partie « zéro »,
- quelle covariable mettre dans la partie « compte »,
- lesquelles mettre dans les deux,
- avec peu de recommandations standardisées.

4.F En pratique :

- il existe toujours une possibilité de critique : « pourquoi cette covariable est-elle dans la partie zéro et pas dans l'autre ? » ;
- d'où la nécessité de bien réfléchir à la signification des zéros et de justifier le choix des covariables dans chaque sous-modèle.

5 Application



= Association entre l'utilisation d'AOD avec et sans médicaments concomitants et le risque de saignement majeur dans la fibrillation auriculaire non valvulaire

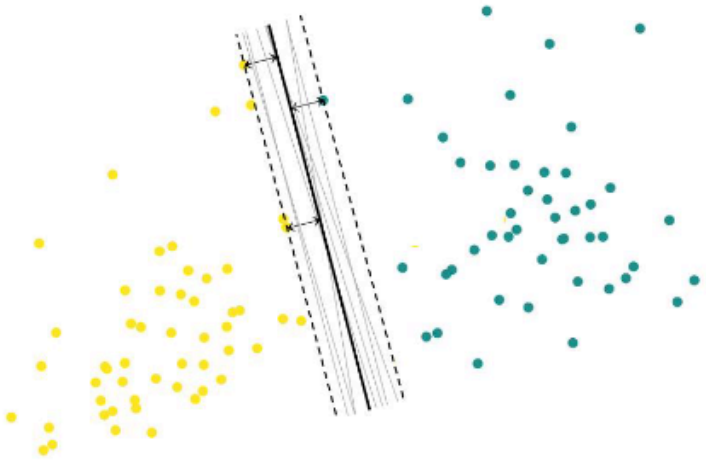


Fig. 3.4 — *Deux groupes de points sont parfaitement séparables par une infinité de droites discriminantes. Celle proposée par le SVM maximise l'écart minimal avec les points des deux groupes. Cette solution semble naturelle, elle est par ailleurs susceptible d'avoir de meilleures performances sur un échantillon de validation.*

Objectif : Estimer le risque relatif (RR) de saignement majeur associé à l'utilisation d'AOD **avec et sans médicaments concomitants**, en ajustant pour les facteurs de confusion potentiels.

Design :

- Étude de cohorte rétrospective
- Patients tous exposés à des AOD, ± médicaments concomitants : atorvastatine, amiodarone, digoxine, inhibiteurs calciques, fluconazole…
- Mesure de l'incidence des évènements (saignements majeurs) pendant la période d'exposition

Analyse :

- On regarde l'incidence d'évènements

Modèles possibles :

- **Modèle logistique et OR** : évènement Y/N à un temps donné
- **Modèle de Cox (survie)** : temps jusqu'à l'évènement : fait de co-prescrie une statine augmente le **risque instantané** de saignement majeur à chaque instant t ?
- **Modèle de Poisson** : nombre d'évènements pendant la période d'exposition : fait de co-prescrie une statine augmente le **taux moyen** de saignement majeur pendant la période d'exposition ? = permet d'obtenir un **taux d'incidence** / **rapport d'incidence**
  - Permet ainsi de calculer un rapport d'incidence ajusté (IRR) pour chaque médicament concomitant
  - le λ du modèle de Poisson correspond au **taux d'incidence** (nombre d'évènements / temps d'exposition)

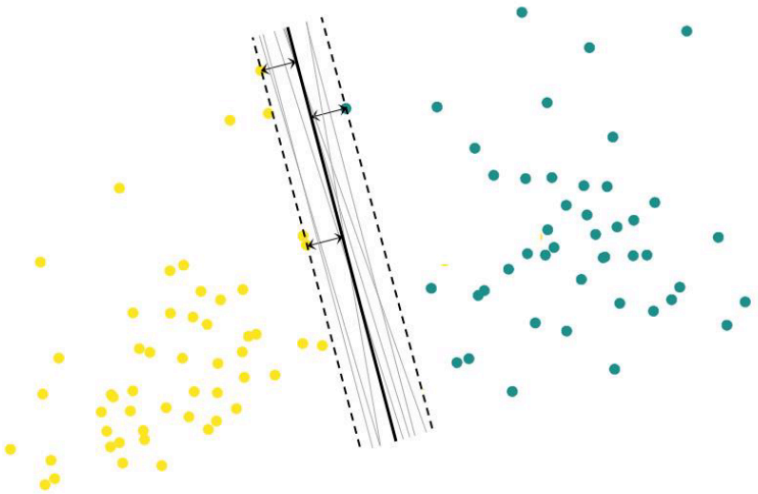


Fig. 3.4 — *Deux groupes de points sont parfaitement séparables par une infinité de droites discriminantes. Celle proposée par le SVM maximise l'écart minimal avec les points des deux groupes. Cette solution semble naturelle, elle est par ailleurs susceptible d'avoir de meilleures performances sur un échantillon de validation.*

Table 3. Major Bleeding Risk Among Patients Taking a NOAC for Nonvalvular Atrial Fibrillation With Concurrent Medications

Concurrent Medication	Person-Quarters With NOAC Use	No. of Bleeding Events	Crude Major Bleeding Incidence Rate (99% CI) per 1000 Person-Years	Adjusted Incidence Rate (99% CI) per 1000 Person-Years	Adjusted Incidence Rate Difference (99% CI) per 1000 Person-Years	Adjusted Rate Ratio (99% CI)
Atorvastatin						
With	123 420	1056	34.22 (31.51 to 36.94)	34.57 (31.87 to 37.50)	–14.38 (–17.76 to –10.99)	0.71 (0.64 to 0.78)
Without	323 617	3459	42.75 (40.88 to 44.63)	48.96 (46.48 to 51.57)	1 [Reference]	1 [Reference]
Digoxin						
With	100 513	1130	44.97 (41.52 to 48.42)	45.69 (42.23 to 49.43)	–4.46 (–8.45 to –0.47)	0.91 (0.83 to 0.99)
Without	346 524	3413	39.40 (37.66 to 41.13)	50.14 (47.34 to 53.11)	1 [Reference]	1 [Reference]

- Person-quarters with NOAC use : nombre de trimestres-personnes d'exposition aux AOD

- No. of bleeding events : nombre de saignements majeurs observés
- Crude major bleeding incidence rate : taux d'incidence brut (non ajusté) de saignement majeur par 1000 personnes-années (se fait en faisant par exemple 1056 / (123420 * 3 mois) * 12 mois * 1000)
- Adjusted incidence rate : taux d'incidence ajusté de saignement majeur par 1000 personnes-années (obtenu par le modèle de Poisson ajusté)
- Pour faire les IC : estimateur Sandiwch (décrit dans les méthodes)

Le problème est qu'ils ont

- à la fois ajusté pour des variables de confusion (âge, sexe, comorbidités, score de risque, etc.)
- et à la fois pondéré par le temps d'exposition (trimestres-personnes)

6 Références

- Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of Statistical Software, 27(8), 1-25.