S5_3_GLM : Modèles linéaires généralisés

Table of contents

1 Introduction

1.A Retour sur le modèle linéaire simple

BAS : score quantitatif de dévoleppement cognitif

Dans ce tableau :

Durée de l'allaitement maternel	Moyennes (n)	Coefficients bruts (IC 95%) (n=10 944)	Coefficients partiellement ajustés (IC 95%) (n=10 929)	Coefficients complètement ajustés (IC 95%) (n=10 416)
Jamais	106.5 (3825)	Reference	Reference	Reference
< 2,0 mois	110.2 (2901)	3.7 (2.8–4.6)	0.7 (–0.1–1.5)	0.7 (–0.1–1.5)
2,0–3,9 mois	111.8 (1047)	5.4 (4.1–6.6)	1.2 (0–2.4)	1.2 (0–2.4)
4,0–5,9 mois	113.0 (889)	6.5 (5.1–7.9)	1.2 (0–2.5)	1.0 (–0.3–2.3)
6,0–11,9 mois	114.1 (1392)	7.7 (6.5–8.8)	2.2 (1.2–3.2)	2.0 (1.0–3.0)
12,0 mois	114.2 (890)	7.7 (6.6–8.9)	2.4 (1.3–3.5)	2.4 (1.3–3.6)

- Durée d'allaitement : variable explicative **catégorielle ordinale** en 6 niveaux : Jamais ; < 2 mois ; 2-3,9 mois ; 4-5,9 mois ; 6-11,9 mois ; 12 mois
- Estimation de l'effet de l'allaitement sur le score BAS
- Modèle avec coefficients "bruts" + partiellement ajustés + ajustés
- "Brut" : analyse univariable = régression linéaire simple
- "Ajustement" : analyse multivariable = régression linéaire multiple

Questions :

- Pourquoi le modèle brut est sur 10 944 et le modèle ajusté complet sur 10 416 ?
  - Probablement en raison des données manquantes pour les variables de confusion incluses dans le modèle ajusté complet.
- Pourquoi l'intervalle de confiance est diminué dans le modèle ajusté complet par rapport au modèle brut ?
  - L'ajustement pour les variables de confusion peut réduire la variabilité non expliquée, ce qui peut conduire à des intervalles de confiance plus étroits.
  - Même s'il ne s'agit pas d'une variable de confusion, l'ajout de variables explicatives pertinentes peut améliorer la précision des estimations.
- Pourquoi pas de p-value ?
  - Dans un modèle linéaire, les coefficients s'interprètent comme des différences de moyennes entre les groupes par rapport à la catégorie de référence.
  - C'est différent d'un modèle de régression logistique où les coefficients sont des log-odds ratios, et où les p-values sont souvent utilisées pour tester l'association.
  - 3,7 : différence de moyenne de score de BAS entre le groupe "< 2 mois" et le groupe "Jamais".
- Ce ne sont pas des odds ratios ?

  - Non, dans un modèle linéaire, les coefficients représentent des différences de moyennes, pas des rapports de cotes.
  - Dans un modèle linéaire, la valeur neutre = 0 (pas de différence de moyenne).
  - **Mesures d'associations (OR, RR, HR) sont des rapports dont la valeur neutre = 1.**
- Pourquoi des catégories d'âge maternelle et pas l'âge en continu ?
  - Utilisation de catégories peut faciliter l'interprétation des résultats.
  - Utilisation de l'âge en continue reviendrait à faire l'hypothèse d'une relation linéaire entre l'âge maternel et le score BAS, ce qui peut ne pas être approprié.
- Si on devait avoir une variable quantitative dont il est difficile d'avoir un seuil
  - Binarisation en optimisant un seuil (Youden)
  - Binarisation par la médiane
  - Catégories d'âge

1.B Modèles linéaires généralisés (GLM)

- Extension des modèles linéaires pour permettre de modéliser des variables dépendantes qui ne suivent pas une distribution normale.
- Permettent de modéliser des relations entre une variable dépendante et une ou plusieurs variables indépendantes, tout en tenant compte de la nature spécifique de la variable dépendante.
- Composants principaux des GLM :
  - **Fonction de lien** : établit une relation entre la moyenne de la variable dépendante et les variables indépendantes.
  - **Distribution de la famille exponentielle** : spécifie la distribution de la variable dépendante (ex. binomiale, Poisson, etc.).
- Exemples courants de GLM :
  - Régression logistique : pour les variables dépendantes binaires (oui/non).
  - Régression de Poisson : pour les variables dépendantes de comptage.
  - Régression gamma : pour les variables dépendantes continues et positives.

1.B.1 Modèle logistique

Reprend presque toutes les caractéristiques de la régression linéaire, mais est adapté pour les variables dépendantes binaires.

- possibilité d'obtenir des coefficients bruts ou ajustés
- codage des variables explicatives catégorielles
- interprétation des résultats

SAUF

- La variable expliquée (critère de jugement) est **BINAIRE** (oui/non, succès/échec, malade/-pas malade)
- La force de l'association est mesurée par des **odds ratios (OR)** au lieu de différences de moyennes

> **!** Important
>
> **PLAN DU COURS**
>
> 1. Mesures d'association : RR et OR
> 2. Modèle logistique univariable
>   - Interprétation des coefficients
>   - Test statistiques sur les coefficients
> 3. Modèle logistique multivariable
>   - Notion d'interaction
>   - Construction de modèles multivariables

2 Définition des risques

2.A Exemple

- 256 patients avec transplantation de moelle osseuse
- Variable à expliquer Y : rejet de la greffe (oui/non) = variable binaire dichotomique
- Covariable X : dose faible (X = 1) vs dose élevée (X = 0) d'un médicament immunosuppresseur

Étude du lien entre dose de médicament et risque de rejet de la greffe.

Tableau de contingence :

	Rejet (Y=1)	Pas de rejet (Y=0)	Total
Dose faible (X=1)	37	71	108
Dose élevée (X=0)	31	117	148
Total	68	188	256

- Pourcentage de rejet dans chaque groupe :
  - Dose faible : 37/108 34%
  - Dose élevée : 31/148 21%
- Comparaison de pourcentages = test du chi2

```r
# Tableau de contingence
tableau <- matrix(c(37, 71, 31, 117), ncol = 2)
```

```r
colnames(tableau) <- c("Rejet", "Pas de rejet")
rownames(tableau) <- c("Dose faible", "Dose élevée")
tableau
```

```
            Rejet Pas de rejet
Dose faible    37           31
Dose élevée    71          117
```

```r
# Test du chi2
chisq.test(tableau, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  tableau
X-squared = 5.6732, df = 1, p-value = 0.01723
```

Syntaxe :

- `matrix(c(a, b, c, d), ncol = 2)` : création du tableau de contingence
- `chisq.test(tableau, correct = FALSE)` : test du chi2
  - correct = FALSE : pas de correction de Yates pour les petits effectifs

Interprétation des résultats :

- LIEN significatif
- mais pas de mesure de la force
- pour mesurer la force : on pourrait faire 34/21 = 1,62 = RR

Questions :

- Est-ce une mesure de la force de l'association entre dose et rejet ?
  - Non, le test du chi2 indique s'il y a une association statistiquement significative, mais ne quantifie pas la force de cette association.
- Est-ce une mesure de la signification statistique ?
  - Oui, le test du chi2 fournit une p-value qui indique si l'association observée est statistiquement significative.
- Est-ce que ça dépend de l'effectif ?
  - Oui, la puissance du test du chi2 dépend de la taille de l'échantillon. Des échantillons plus grands peuvent détecter des associations plus faibles.

2.B Mesures d'associations

Risque : soit un facteur de risque X (Oui = 1, Non = 0) et un événement Y

- R1 = P(Y=1 | X=1) = probabilité que l'événement se produise chez les sujets exposés au facteur de risque X=1

- Dans l'exemple, R1 = 37/108 = 0,3426 (risque que le rejet se produise dans le groupe dose faible)
- R0 = P(Y=1 | X=0) = probabilité que l'événement se produise chez les sujets non exposés au facteur de risque X=0
- Dans l'exemple, R0 = 31/148 = 0,2095 (risque que le rejet se produise dans le groupe dose élevée)

Ratio de risque = Risque Relatif (RR)

- $RR = R_1/R_0$.
- Dans le cas de l'exemple au dessus :
  - R_1 = 37/108 = 0,3426 (risque de rejet dans le groupe dose faible)
  - R_0 = 31/148 = 0,2095 (risque de rejet dans le groupe dose élevée)
  - $RR = R_1$ / R_0 = 0,3426 / 0,2095 = 1,61 / R0 = (37/108) / (31/148) = 1,62

Cote ou Odds :

- $Odds = R/(1-R)$ = probabilité que l'événement se produise / probabilité que l'événement ne se produise pas
- Par exemple si R vaut 0,75 (risque = 75%)
- O = 0,75 / (1-0,75) = 3 donc la cote est de 3 (3 fois plus de chances que l'événement se produise que de ne pas se produire)

Rapport de cotes = Odds Ratio (OR)

- $OR = \frac{R_1/1-R_1}{R_0/1-R_0}$

RR et OR = 2 manières différentes de mesurer l'association entre variables binaires

2.C Interprétation des risques

RR	OR
• Facilement interprétable – Facteur par lequel le risque de Y est multiplié si le facteur X est présent • Pas estimable pour tous les types d'études épidémiologiques (notamment pas pour cas témoins) • Moins bonnes propriétés statistiques et mathématiques que l'OR	• Interpétation moins immédiate (rapport des cotes souvent confondus avec RR) • Obtenus dans analyses statistiques multivariées • Estimable dans les enquêtes cas-témoin (maladie à prévalence non estimable)

2.C.1 Exemple

- 100 hommes : 90 ont bu du vin la semaine précédente
  - Odds d'un homme d'avoir bu du vin : 9:1

- 100 femmes : 20 ont bu
  - Odds : 20/80 : 1:4
- OR = 9/0,25 = 36 donc peu interprétable
- RR = 0.9/0.2 = 4,5

NB : l'OR est symétrique mais pas le RR !

- OR de ne pas boire de vin : 0.11/4 = 1/36 = 1/OR d'avoir bu
- RR de ne pas avoir bu : 0.1/0.8 = 1/8 = 0.125 de 1/RR !!

2.D Calcul des OR et RR dans un tableau de contingence

	Rejet (Y=1)	Pas de rejet (Y=0)	Total
Dose faible (X=1)	a	b	a+b
Dose élevée (X=0)	c	d	c+d
Total	a+c	b+d	n

Risques :

- Risque dans le groupe exposé (X=1) : $R_1 = \frac{a}{a+b}$
- Risque dans le groupe non exposé (X=0) : $R_0 = \frac{c}{c+d}$
- Risque relatif (RR) : $RR = \frac{R_1}{R_0} = \frac{a/(a+b)}{c/(c+d)}$

Côtes :

- Cote dans le groupe exposé (X=1) : $O_1 = \frac{a}{b}$
- Cote dans le groupe non exposé (X=0) : $O_0 = \frac{c}{d}$
- Odds ratio (OR) : $OR = \frac{R_1/(1-R_1)}{R_0/(1-R_0)} = \frac{a/b}{c/d} = \frac{ad}{bc}$

> **Tip**
>
> Imaginons qu'on multiplie les effectifs des réponse = oui
>
> - OR : $\frac{(2a)(d)}{(b)(2c)} = \frac{2a*d}{b*2c}$ donc OR ne change pas (un 2 de chaque côté se simplifie)
>   - Donc OR beaucoup plus utilisé ! (peut être calculé dans bcp + de cas)
> - RR : $\frac{(2a)/(2a+b)}{(2c)/(2c+d)}$ donc RR change (le 2 ne se simplifie pas)
>   - Le RR dépend de la proportion rejet / non rejet !!

2.D.1 Exemple numérique avec R

```r
# Effectifs
a <- 37  # Rejet, Dose faible
b <- 31  # Pas de rejet, Dose faible
c <- 71  # Rejet, Dose élevée
d <- 117 # Pas de rejet, Dose élevée

#création et affichage tableau de contingence
tableau <- matrix(c(a, b, c, d), ncol = 2)
colnames(tableau) <- c("Rejet", "Pas de rejet")
rownames(tableau) <- c("Dose faible", "Dose élevée")
tableau
```

```
            Rejet Pas de rejet
Dose faible    37           71
Dose élevée    31          117
```

```r
# Calcul des OR et RR
RR <- a/(a+c)/(b/(b+d))
OR <- (a*d)/(b*c)

# Affichage des résultats
OR
```

```
[1] 1.966833
```

```r
RR
```

```
[1] 1.635603
```

OR = 1.966833

RR = 1.635603

> **Tip**
>
> - A-t-on toujours OR > RR ?
>   - Oui, sauf lorsque l'événement est rare (risque faible), où OR et RR peuvent être proches.
>   - Dans tous les cas **plus éloigné de 1** que le RR.

2.D.2 Intervalles de confiance des OR et RR

- Calcul des intervalles de confiance (IC) pour les OR et RR

$$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

3 Régression logistique univariable

3.A Introduction

Exemple.1 : on veut étudier le relation entre pathologie et plusieurs FDR

Exemple 2 : on veut étudier le lien entre survenue d'un évènement et plusieurs évènements intercurrents

- Variable dépendante Y :
  - binaire dichotomique (0/1)
  - Loi de Bernoulli = success/failure
- Variables explicatives x1, x2, …, xp :
  - quantitatives et/ou qualitatives
  - peuvent être continues ou discrètes
- On a des données chez n patients numérotés i = 1, 2, …, n
  - Y_i : réponse binaire pour le patient i
  - $x_{i1}, x_{i2}, ..., x_{ip}$: covariables explicatives pour le patient i

3.B Transformation logistique

- Au lieu de modéliser le lien entre Y_i et les x_i directement, on veut un modèle pour $p_i = E(Y = 1/X = x_i)$
  - C'est à dire la probabilité que $Y_i = 1$ sachant les valeurs des covariables explicatives pour le patient i
  - Exemple : dose plus élevée associée à une plus grande probabilité de réponse
- p_i n'est pas directement observée
  - Peut être estimée chez tous les patients avec $X = x_i$ si variable catégorielle ou groupements
- Problème : $p_i = P(Y_i = 1 \| X = x_i)$ est une probabilité
  - Doit être comprise entre 0 et 1 (donc les modèles linéaires ne sont pas adaptés car supposent la normalité des résidus)
  - Relation linéaire entre p_i et les x_i peut donner des valeurs en dehors de [0, 1]
- Solution : transformation logistique (fonction logit)
  - $q = logit(p) = \ln\left(\frac{p}{1-p}\right)$
  - même transformation que R/(1-R) = Odds
  - prendre le log permet d'avoir quelque chose entre -∞ et +∞
- Fonction expit (inverse du logit) :
  - $p = expit(q) = \frac{e^q}{1+e^q} = \frac{1}{1+e^{-q}}$
  - permet de revenir à une probabilité entre 0 et 1

3.C Modèle logistique

On modélise logit de $p = P(Y = 1/X = x)$ comme une combinaison linéaire des variables explicatives x1, x2, ..., xp

$logit(p) = \alpha + \beta x$

Transformation inverse (expit) :

$p = \frac{1}{1+exp(-\alpha+\beta x)} = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$

3.C.1 Si X est une variable binaire (0/1) :

- quand X = 0, $logit(p_0) =$ $alpha = logit(R_O$
- quand X = 1, $logit(p_1) =$ $alpha +$ $beta = logit(R_1)$
- OR =
  - $= \frac{R_1/(1-R_1)}{R_0/(1-R_0)}$ ou $ln(OR) = logit(R_1) - logit(R_0) =$ $beta)$
  - $= \frac{exp(alpha+beta)}{exp(\quad alpha)}$
  - $= exp($ $beta)$

Valeur neutre pour l'OR = 1 donc valeur neutre pour $\quad = 0$

3.C.2 Si X est une variable catégorielle à 3 valeurs (0, 1, 2) :

- Équivalent à 2 variables binaires indicatrices :
  - $x_1 = 1$ si X = 1, 0 sinon
  - $x_2 = 1$ si X = 2, 0 sinon
- Modèle à 2 variables :
  - $logit(p) =$ $alpha +$ $beta_1 x_1 +$ $beta_2 x_2$
  - quand X = 0, $logit(p_0) =$ $alpha = logit(R_0)$
  - quand X = 1, $logit(p_1) =$ $alpha +$ $beta_1 = logit(R_1)$
  - quand X = 2, $logit(p_2) =$ $alpha +$ $beta_2 = logit(R_2)$

- $OR_1 = exp($ $beta_1)vs0$
- $OR_2 = exp($ $beta_2)vs0$
- OR de 2 vs 1 = $exp($ $beta_2)/exp($ $beta_1) = OR2/OR1$

3.C.3 Si X est une variable catégorielle à K valeurs

- Définir une classe de référence (X=0)
- Il y aura K-1 OR par rapport à la classe de référence

Attention au choix de la classe de référence !

Atention aux regroupements de catégories (si faibles effectifs)

3.C.4 Si X est une variable continue / quantitative

Hypothèse : relation linéaire de logit(p) avec x (sur le logOR)

$f(x) = P(Y = 1/X = x = \frac{1}{1+exp(-(\alpha+\beta x))}$

- Risque que Y = 1 sachant que X prend la valeur x

Entre deux valeurs x1 et x2 :

- $ln(OR) = \beta(x_1 - x_0)$; OR = $exp(\beta(x_1 - x_0))$

NB : si pour des variables QUANTITATIVES, l'OR a une UNITÉ

- Pour 1 an d'âge
- pour 10mg de morphine.

3.D Fonction de vraisemblance

- Estimation des paramètres du modèle (, 1, 2, ..., p) par la méthode du maximum de vraisemblance
- Vraisemblance = probabilité d'observer les données telles qu'elles ont été observées, en fonction des paramètres du modèle
- On cherche les valeurs des paramètres qui maximisent cette probabilité
- Ordinateur cherche les meilleur coefficient α et β qui maximisent la vraisemblance

3.E Test de Wald

Test de Wald = Basé sur le logOR

- logciels fournissent β et l'erreur d'estimation SE(β) (erreur standard)
- H0 = $\beta = 0$ (OR = 1)

- $\beta/\text{se}(\beta)$ suit une loi normale centrée réduite N(0,1)
- comparaison de $\beta/\text{se}(\beta)$ à 1,96 pour un test bilatéral au seuil 5%
- calcul du IC à partir de β et $\text{se}(\beta) = \beta \pm 1{,}96*\text{se}(\beta)$

Intervalle de confiance sur OR

- on prend l'exponentielle des bornes de l'IC pour b : [exp(b1) ; exp(b2)]
- intervalle non symétrique autour de exp(b)
- Si IC pour β contient 0, IC pour OR contient 1 alors rejet de H0

3.F Exemple avec R

Exemple du rejet de greffe

```
# Données
beta <- log(OR)
beta
```

```
[1] 0.6764248
```

```
# Estimation de l'erreur standard
se_beta <- sqrt(1/37 + 1/71 + 1/31 + 1/117)
se_beta
```

```
[1] 0.2862108
```

```
# Test de Wald
beta/se_beta
```

```
[1] 2.36338
```

```
# ici > 1,96 donc on rejette H0

# pvalue
2 * (1 - pnorm(abs(beta/se_beta)))
```

```
[1] 0.0181091
```

- OR = 1.96
- $\beta = \log(\text{OR}) = 0{,}68$ (b = coefficient)
- $\text{se}(\beta) = 0{,}29$ (écart-type de l'estimateur)
- Rapport coeffcient/écart-type : $\text{b/se(b)} = 2{,}36 > 1{,}96$ donc on rejette H0

Il y a donc un lien significatif entre infection et décès

Calcul de l'IC à 95% pour OR

```r
# Calcul des bornes de l'IC pour b
beta1 <- b - 1.96 * se_beta
beta2 <- b + 1.96 * se_beta
beta1
```

```
[1] 30.43903
```

```r
beta2
```

```
[1] 31.56097
```

```r
#exponentielle des bornes pour obtenir l'IC pour OR
ic_or <- c(exp(beta1), exp(beta2))
ic_or
```

```
[1] 1.657683e+13 5.090453e+13
```

Donc IC à 95%

- pour β : 0,68 $\pm$ 1,96 * 0.29 = [0,11 ; 1,24]
- pour OR : exponentielle de l'IC de β [1,12 ; 3,45].

4 Régression logistique multivariable

4.A Principe

4.A.1 Principe général

Variable dichotomique Y dont on veut étudier le lien avec p variables continues ou discrètes : x_1, x_2, ..., x_p
On dit que Y est

- la variable dépendante
- la variable à expliquer

On dit que les X_k sont

- les variables indépendantes
- les variables explicatives
- les covariables

Extension de la régression logistique univariable :
$logit(P(Y = 1 \| x_1, ..., x_p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$

4.A.2 Hypothèse de multiplicativité des effets

Le modèle logistique multivarié suppose la **multiplicativité** des effets des facteurs de risque

Cas particulier : deux facteurs de risque X_1 et X_2 prenant deux valeurs (0/1) (ex tabac et alcool et cancer du poumon)

- OR pour $X_1 = 1$ vs 0 : $OR_1 = exp(\beta_1)$
- OR pour $X_2 = 1$ vs 0 : $OR_2 = exp(\beta_2)$
- OR pour $X_1 = 1$ et $X_2 = 1 = OR_{1,2} = exp(\beta_1 + \beta_2) = exp(\beta_1) * exp(\beta_2) = OR_1 * OR_2$

Si OR de cancer de poumon de 3 pour tabac et de 2 pour alcool ; OR pour tabac + alcool = 3 * 2 = 6

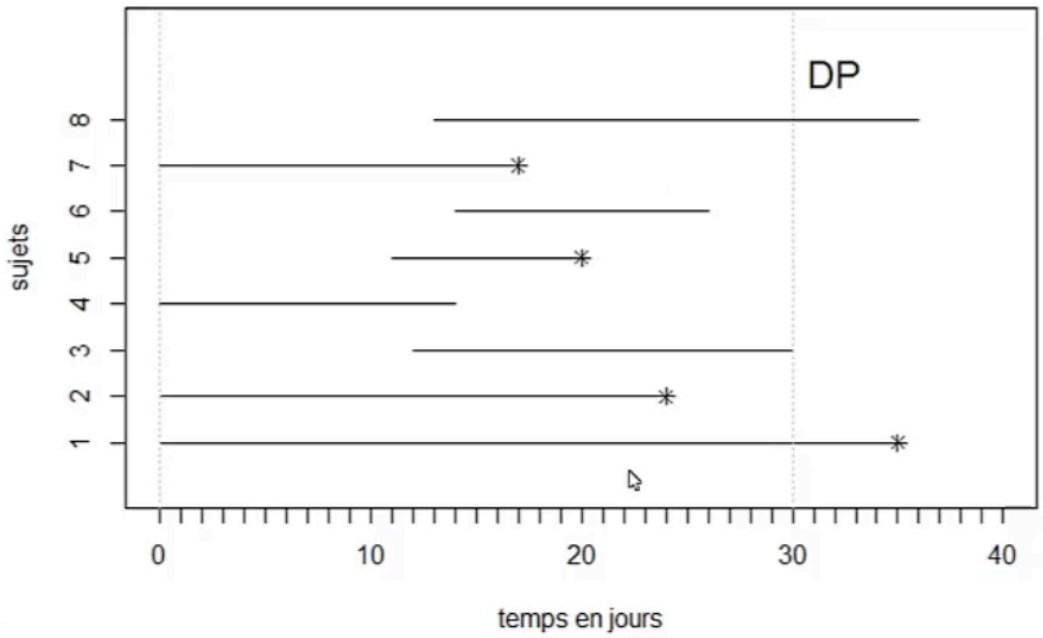
Mais il est possible que l'effet combiné soit différent de la multiplication des effets individuels (interaction) !!!

Dans ce cas, il faut ajouter un terme d'interaction dans le modèle

On peut très bien faire un tableau avec les β et les $se(\beta)$!

- Dans ce cas : les **coefficients** sont comparés à 0
- Pour les variables continues exprimer les coefficients et les OR **DANS L'UNITÉ** d'origine !
- **IC de l'OR :** $exp(\beta \pm 1,96 * se(\beta))$

A partir du graphique déterminer le temps de participation T_k pour chaque patient et l'état I_k en T_k.



16/11/20

13

4.A.3 Notion d'interaction

Interaction entre deux variables explicatives X_1 et X_2 :
$logit(P(Y = 1 \| x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$
Rajoute un coefficient et lève l'hypothèse de multiplicativité des OR

Tip

Définitions essentielles : , estimateur, estimation
Dans un modèle logistique, on étudie l'association entre une variable explicative X (par exemple, dose d'un traitement) et la probabilité d'un événement Y (par exemple, rejet de greffe).
Le modèle s'écrit :
$\text{logit}(p) = \alpha + \beta X$
$p = P(Y = 1 \mid X)$
Dans cette équation :

- α est une constante appelée intercept ;
- β est un paramètre du modèle, qui mesure l'effet de X sur la probabilité de l'événement.

Un paramètre est une quantité théorique, supposée fixe (c'est à dire constante), qui décrit la relation vraie dans la population étudiée. Dans le modèle logistique, on montre que :
$\beta = \log(\text{OR})$,
$\text{OR} = e^{\beta}$.
Ainsi :

- si $\beta = 0$, alors OR $= 1$: il n'existe pas d'association entre X et Y ;
- si $\beta > 0$, alors OR > 1 : présence d'un effet augmentant le risque ;
- si $\beta < 0$, alors OR < 1 : présence d'un effet protecteur.

En pratique, on ne connaît jamais la vraie valeur de β.
On cherche donc à l'estimer à partir d'un échantillon.
Un estimateur est une méthode de calcul permettant d'approcher un paramètre à partir des données.
La valeur obtenue par cet estimateur dans un échantillon donné est appelée estimation.
Dans la régression logistique, l'estimateur utilisé pour β est l'estimateur du maximum de vraisemblance, noté $\widehat{\beta}$ ("bêta chapeau").
La valeur numérique affichée par le logiciel (par exemple, 0.68) est l'estimation de β dans l'échantillon étudié.

Principe du maximum de vraisemblance : comment obtient-on $\widehat{\beta}$?
Pour chaque individu i, le modèle logistique calcule une probabilité prédite de l'événement :
$p_i = P(Y_i = 1 \mid X_i) = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}}$.
Si $Y_i = 1$, la probabilité d'observer cette valeur est p_i. Si $Y_i = 0$, cette probabilité est $1 - p_i$.
On définit alors la vraisemblance totale des paramètres α et β comme la probabilité d'observer l'ensemble des données :
$L(\alpha, \beta) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1-Y_i}$.

Le principe du maximum de vraisemblance consiste à choisir les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui maximisent cette probabilité.
Autrement dit, on sélectionne les paramètres qui rendent les données observées les plus plausibles selon le modèle.
Pour des raisons numériques, on maximise généralement la log-vraisemblance :
$\ell(\alpha, \beta) = \sum_{i=1}^{n} [Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)]$,
ce qui est équivalent à maximiser la vraisemblance elle-même.
Le logiciel recherche alors les valeurs de α et β qui maximisent $\ell(\alpha, \beta)$. La valeur obtenue pour β est l'estimation $\hat{\beta}$.

Erreur standard : quantifier l'incertitude sur l'estimation $\hat{\beta}$
Si l'on refaisait l'étude plusieurs fois sur des échantillons différents, on n'obtiendrait pas toujours la même estimation $\hat{\beta}$. L'estimateur $\hat{\beta}$ varie d'un échantillon à l'autre : il s'agit d'une variable aléatoire.
L'erreur standard $\mathrm{SE}(\hat{\beta})$ mesure l'ampleur typique de cette variabilité. Elle correspond à la racine carrée de la variance théorique de l'estimateur. Plus $\mathrm{SE}(\hat{\beta})$ est petit, plus l'estimateur est précis.
Cette erreur standard est calculée (ou approximée) automatiquement par le logiciel à partir de la dérivée seconde de la log-vraisemblance (matrice d'information de Fisher).

Standardisation de l'estimateur et loi normale centrée réduite
Dans les modèles logistiques, et plus largement dans les modèles estimés par maximum de vraisemblance, un résultat fondamental est que, sous certaines conditions (échantillon suffisamment grand, modèle correctement spécifié), l'estimateur $\hat{\beta}$ suit approximativement une loi normale de moyenne β et de variance $\mathrm{SE}(\hat{\beta})^2$.
Sous l'hypothèse nulle $H_0 : \beta = 0$, on a donc :
$\hat{\beta} \approx \mathcal{N}(0, \mathrm{SE}(\hat{\beta})^2)$.
Pour exploiter ce résultat, on standardise l'estimateur, c'est-à-dire qu'on le transforme en une variable sans unité, mesurant la distance à 0 en nombre d'erreurs standard :
$Z = \frac{\hat{\beta}-0}{\mathrm{SE}(\hat{\beta})} = \frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})}$. Cette variable Z suit alors approximativement une loi normale centrée réduite $\mathcal{N}(0, 1)$, c'est-à-dire une loi normale de moyenne 0 et de variance 1.
Cette propriété est à la base du test de Wald.

Test de Wald : définition, interprétation, limites
Principe du test
On souhaite tester :
$H_0 : \beta = 0$
vs
$H_1 : \beta \neq 0$.
Le test de Wald utilise la statistique :
$W = \frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})}$.
Sous $H_0 : \beta = 0$, W suit approximativement une loi normale centrée réduite. Pour un test bilatéral au seuil de 5 %, on rejette H_0 si :
$|W| > 1.96$.
La p-value se déduit de la loi normale. L'intervalle de confiance de β est :

$\hat{\beta} \pm 1.96 \times \text{SE}(\hat{\beta})$.
L'intervalle de confiance de l'odds ratio s'obtient en exponentiant les bornes.
Limites et précautions sur le test de Wald*
Le test de Wald repose sur des approximations asymptotiques. Il peut être peu fiable dans plusieurs situations : 1. Échantillons de petite taille

- `L'approximation normale est mauvaise et le test peut être trop libéral ou trop conservate`

2. Événements rares
    - Si l'événement étudié est très rare, $\hat{\beta}$ peut avoir une distribution très asymétrique. Les intervalles de confiance peuvent être trop étroits ou décalés.
3. Effets très importants (odds ratios très élevés)
    - Dans ces cas, la distribution de $\hat{\beta}$ est fortement asymétrique. Les intervalles basés sur Wald peuvent être inadaptés.
4. Séparation complète ou quasi-complète
    - Si une covariable prédit parfaitement l'événement (par exemple, tous les événements surviennent dans un seul groupe), $\hat{\beta}$ "diverge" et le test de Wald n'est plus utilisable.

Dans ces situations, il est souvent préférable d'utiliser des méthodes alternatives comme la régression logistique pénalisée (méthode de Firth), des tests exacts, ou le test du rapport de vraisemblance.

Test du rapport de vraisemblance (Likelihood Ratio Test)
Le test du rapport de vraisemblance (LRT) compare deux modèles emboîtés : un modèle réduit (sans la covariable testée) et un modèle complet (avec la covariable).
Pour tester $\beta = 0$, on compare :

- modèle réduit : $\text{logit}(p) = \alpha$;
- modèle complet : $\text{logit}(p) = \alpha + \beta X$.

Le logiciel calcule la log-vraisemblance maximale des deux modèles :

- $\ell_{\text{réduit}}$: $modlesansX$;
- ℓ_{complet} : $modleavecX$.

La statistique de test est :
$\Lambda = 2(\ell_{\text{complet}} - \ell_{\text{réduit}})$.
Sous $H_0 : \beta = 0$ et sous des conditions générales, Λ suit une loi du χ^2 avec un degré de liberté égal au nombre de paramètres testés (ici, 1). On en déduit une p-value.
Le LRT présente plusieurs avantages :

- il repose sur la comparaison globale des modèles et non sur l'approximation locale autour de $\hat{\beta}$;

- il est plus robuste que le test de Wald lorsque l'échantillon est petit, lorsque les événements sont rares, ou lorsque l'effet étudié est important ;
- il permet de tester simultanément plusieurs coefficients (par exemple, pour des variables qualitatives à plusieurs modalités).

Résumé

- β est le paramètre du modèle logistique correspondant au log(odds ratio).
- $\widehat{\beta}$ est son estimation obtenue par maximum de vraisemblance, en choisissant les valeurs qui rendent les données observées les plus probables.
- L'erreur standard de $\widehat{\beta}$ mesure la variabilité de l'estimation d'un échantillon à l'autre.
- Le test de Wald repose sur la standardisation de $\widehat{\beta}$, qui suit approximativement une loi normale centrée réduite sous l'hypothèse nulle.
- Ce test peut être peu fiable en cas de petits échantillons, d'événements rares ou d'effets extrêmes.
- Le test du rapport de vraisemblance compare un modèle avec et sans la covariable, en s'appuyant sur la log-vraisemblance. Il est souvent plus robuste et peut être utilisé pour tester plusieurs coefficients simultanément.