Compléments en régression logistique

① Interprétation des coefficients en regression logistique

En regression logistique, les parametres $\beta_1, \dots, \beta_p$ que l'on estime s'interprètent comme des **logarithmes d'odds ratio** (rapports de côte)

(1.1) Odd = côte = chance = risque

Soit Y une variable binaire $\begin{cases} 1 = \text{succès, presence} \dots \\ 0 = \text{échec, absence} \dots \end{cases}$

$Y \sim \text{Bernoulli}(p)$: $\begin{cases} p = \mathbb{P}(Y=1) \\ 1-p = \mathbb{P}(Y=0) \end{cases}$

On définit alors :

$\underline{\text{côte}} = \dfrac{p}{1-p}$: Si un événement (ici $Y=1$) a une forte probabilité p, il a de grandes chances (chance, risque) d'intervenir :

$\begin{cases} p = 0 \iff \text{côte} = 0 \\ p \approx 1 \iff \text{côte} \approx +\infty \end{cases}$

(1.2) Odds ratio = rapport de côtes

Mesure la liaison entre deux variables qualitatives binaires X et Y. On note alors :

$p_1 = \mathbb{P}(Y=1 \,/\, X=1)$

$p_0 = \mathbb{P}(Y=1 \,/\, X=0)$

On définit alors l'odds ratio (OR) de Y associé à X par :

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

Par exemple : Y = presence / absence d'une maladie

X = variable d'exposition à la maladie par exemple le sexe : $\begin{cases} 1 = \text{homme} \\ 0 = \text{femme} \end{cases}$

$\Rightarrow \frac{p_1}{1-p_1}$: risque de présence de la maladie chez les hommes

$\frac{p_0}{1-p_0}$ = risque de présence de la maladie chez les femmes

$\Rightarrow$ Si :

OR ≈ 1, la probabilité d'être malade est identique chez les hommes et chez les femmes donc le risque de maladie n'est pas associé au sexe

$\begin{cases} OR > 1, \text{ le risque d'être malade est plus grand chez les hommes que chez les femmes} \\ OR < 1, \text{ c'est le contraire} \end{cases}$

↳ Dans les deux cas, il existe une association entre le risque de maladie et le sexe.

Autre exemple : Y = présence / absence d'une maladie
X = âge du patient (quantitative)

$\Rightarrow$ On définit la variable binaire $\begin{cases} X = x \\ X = x+1 \end{cases}$ où x = âge donné.

et on note

$p_1 = \mathbb{P}(Y = 1 \ / X = x+1)$

$p_0 = \mathbb{P}(Y = 1 \ / X = x)$

Dans ce cas :

$OR = \frac{p_1/1-p_1}{}$ mesure le risque que $Y = 1$ (maladie

$p_0/1-p_0$ par exemple) lorsque la variable X (l'âge par exemple) augmente d'une unité.

⇒ Si :

OR ≈ 0, le risque de maladie n'est pas associé à l'âge.

OR > 1, le risque de maladie ↗ lorsque l'âge augmente (d'un an).

OR < 1, c'est l'inverse.

(1.3) Les coefficients en regression logistique = log(odds ratio)

On peut montrer qu'en regression logistique les coefficients $\beta_1, \ldots, \beta_p$ associés aux p variables explicatives $X^1, \ldots, X^p$ s'interprètent comme des logarithmes d'odds ratios. On a vu que les variables explicatives X^j sont soit quantitatives, soit binaires (indicatrices des modalités des variables qualitatives).

→ Si X^j binaire : $\begin{cases} p_1 = P(Y=1 / X^j=1) \text{ et} \\ p_0 = P(Y=1 / X^j=0) \end{cases}$

$e^{\beta_j} = \frac{p_1/1-p_1}{p_0/1-p_0}$ est l'odds ratio du risque de survenue de l'évènement (Y=1) chez les personnes exposées au facteur X^j par rapport aux personnes non exposées, ajusté sur les autres variables explicatives du modèle (toute chose égale par ailleurs - TCEPA)

→ Si X^j quantitative : $p_x = P(Y=1 / X^j = x+1)$

$p_0 = \mathbb{P}(Y=1 / X=x)$

$e^{\beta_j} = \frac{p_1/1-p_1}{p_0/1-p_0}$ est l'odds ratio du risque de survenue de l'événement $Y=1$ pour une augmentation d'une unité du facteur X^j, toute chose égale par ailleurs.

② Log-vraisemblance en regression logistique

Echantillon $(X_1, Y_1), \dots, (X_i, Y_i), \dots, (X_n, Y_n)$

et d'après le modèle :

$$p_i = \mathbb{P}(Y=y_i / X=x_i) = \frac{\exp(\beta_0 + x_i^T \beta)}{1 + \exp(\beta_0 + x_i^T \beta)}$$

$$\Rightarrow l(\beta_0, \beta) = \log \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i / X_i = x_i)$$

$$= \log \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= \sum_{i=1}^{n} [y_i \log p_i + (1-y_i) \log(1-p_i)]$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right]$$

Or $1 - p_i = 1 - \frac{\exp(\beta_0 + x^T\beta)}{1+\exp(\beta_0 + x^T\beta)} = \frac{1}{1+\exp(\beta_0 + x^T\beta)}$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p) = \beta_0 + x^T\beta$$

$$\Rightarrow \underline{l(\beta_0, \beta) = \sum_{i=1}^{n} y_i (\beta_0 + x^T\beta) - \log(1 + \exp(\beta_0 + x^T\beta))}$$