Compléments régression logistique

1) Interprétation des coefficients.

1.1) Odd = côte = chance = risque

Y = variable binaire $\begin{cases} 1 = \text{succès, présence...} \\ 0 = \text{échec, absence...} \end{cases}$

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

$\text{côte} = \frac{p}{1-p}$: si un évènement (ici $Y=1$) a une faible probabilité, alors il a de grandes chances d'intervenir :

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

1.2) (rapport de côtes) Odd-ratio = mesure de liaison entre une variable binaire Y et une variable X.

$\boxed{Y=1}$ est l'évènement d'intérêt. $\boxed{X = \text{variable binaire}}$

Exemple : Y = présence / absence d'une maladie $\Rightarrow \begin{cases} Y=1 : \text{présence} \\ Y=0 : \text{absence} \end{cases}$

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

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

$$\boxed{OR = \frac{p_1/1-p_1}{p_0/1-p_0}}$$ où $\begin{cases} p_1 = \mathbb{P}(Y=1 / X=1) \\ p_0 = \mathbb{P}(Y=1 / X=0) \end{cases}$

Interprétation de l'odds-ratio :

* Si OR proche de 1, la proba d'être malade est identique chez les hommes et chez les femmes
:> le risque de maladie n'est pas associé au sexe.

* Si OR ≠ 1, il y a une association entre la maladie et le sexe

→ Si OR > 1, le numérateur est plus grand que le dénominateur donc le risque de maladie est plus grand chez les hommes que chez les femmes

→ Si OR < 1, c'est le contraire

⇒ l'OR mesure le risque d'1 variable relativement à une autre.

Si X est une variable quantitative

On définit la variable binaire

$$\begin{cases} X = x \\ X = x+1 \end{cases}$$

et on note
$$\begin{cases} p_1 = \mathbb{P}(Y=1 \,/\, X=x) \\ p_0 = \mathbb{P}(Y=1 \,/\, X=x+1) \end{cases}$$

On définit alors l'odds-ratio de Y associé à X par

$$OR = \frac{p_1/1-p_1}{p_0/1-p_0}$$ = mesure du risque que $Y=1$ (maladie par exemple) lorsque la variable X augmente d'une unité

Exemple X = âge du patient

* Si OR proche de 1, le risque de maladie n'est pas associé à l'âge
* Si $OR > 1$, le risque de maladie est plus grand lorsque l'âge augmente de 1 an
* Si $OR < 1$, le risque de maladie est plus petit lorsque l'âge augmente de 1 an

1.3) Interprétation des coefficients en régression logistique

Les paramètres $\beta_0, \beta_1, \dots, \beta_p$ du modèle de régression logistique s'interprètent comme des logarithmes d'odds ratio.

1.3.1) Si X^j est binaire

$$e^{\beta_j} = OR = \frac{p_1/1-p_1}{p_0/1-p_0} \quad \text{où} \left\{ \begin{array}{l} p_1 = \mathbb{P}(Y=1 / X^j = 1) \\ p_0 = \mathbb{P}(Y=1 / X^j = 0) \end{array} \right. ,$$

toutes choses égales par ailleurs (pour les autres variable fixés)
(TCEPA)

$\Rightarrow$ e^{β_j} = odds ratio du risque de survenue de l'événement (Y=1) chez les sujets exposés à la variable (facteur) X^j, par rapport aux sujets non exposés, ajusté sur les autres variables explicatives du modèle (TCEPA)

1.3.2) Si X^j est quantitative

$$e^{\beta_j} = OR = \frac{p_1/1-p_1}{p_0/1-p_0} \quad \text{où} \left\{ \begin{array}{l} p_1 = \mathbb{P}(Y=1 / X^j = x) \\ p_0 = \mathbb{P}(Y=1 / X^j = x+1) \end{array} \right.$$

toutes choses égales par ailleurs (TCEPA)

e^{β_j} = odds ratio du risque de survenue de l'événement (Y=1) pour une augmentation d'1 unité de la variable (facteur) X^j, toutes choses égales par ailleurs

2) $\boxed{\text{Log. vraisemblance}}$ en regression logistique

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$$

$$x_i = (1, x_{i1}, \dots, x_{ip})^T$$

⚠ Loi conditionelle Y/X

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

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

Notation :

$$y_i = 1 \Rightarrow \boxed{\mathbb{P}(Y_i = 1 / X = x_i) = p_i}$$

$$y_i = 0 \Rightarrow \mathbb{P}(Y_i = 0 / X = x_i) = 1 - p_i$$

D'où

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

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

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

Donc

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + \exp(-x_i^T \beta)} \right) + (1 - y_i) \log \Bigg(\underbrace{1 - \frac{1}{1 + \exp(-x_i^T \beta)}}_{\frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)}} \Bigg) \right]$$

$$= \sum_{i=1}^{n} \left[\log \left(\frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)} \times \frac{\exp(x_i^T \beta)}{\exp(x_i^T \beta)} \right) + y_i \left(\log \left(\frac{1}{1 + \exp(-x_i^T \beta)} \right) - \log \left(\frac{\exp(-x_i^T \beta)}{1 + \exp(-x_i^T \beta)} \right) \right) \right]$$

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

$$= \sum_{i=1}^{n} \left[\overbrace{\log(1)}^{=0} - \log\left(1+\exp(x_i^T \beta)\right) + y_i \overbrace{\log(1)}^{=0} - y_i \log\left(\exp(-x_i^T \beta)\right) \right]$$

$$= \sum_{i=1}^{n} \left[-\log\left(1+\exp(x_i^T \beta)\right) + y_i x_i^T \beta \right]$$