Statistique et aide à la décision
Session 2

David Causeur
Agrocampus Ouest
IRMAR CNRS UMR 6625

Plan

Modèle de régression pour un risque

Soit Y une variable à K groupes $\{y_1, \ldots, y_K\}$.

Soit X une variable explicative, quantitative ou catégorielle.

Il y a **un effet de X sur Y** si, pour deux valeurs $x \neq x'$,

$$\{\pi_1(x), \ldots, \pi_K(x)\} \quad \neq \quad \{\pi_1(x'), \ldots, \pi_K(x')\},$$

où $\pi_k(x) = \mathbb{P}(Y = y_k \mid X = x),\ k = 1, \ldots, K.$

Modèle de régression pour un risque

Soit Y une variable à $K = 2$ groupes $\{y_1 = -1, y_2 = +1\}$.

Soit X une variable explicative, quantitative ou catégorielle.

Il y a **un effet de X sur Y** si, pour deux valeurs $x \neq x'$,

$$\pi(x) \quad \neq \quad \pi(x'),$$

où $\pi(x) = \mathbb{P}(Y = +1 \mid X = x),\ k = 1, \ldots, K.$

▶ RStudio

Modèle pour une variable réponse à deux groupes

On suppose que

- Si $Y = +1$ alors $X \sim \mathcal{N}(\mu_1, \sigma)$
- Si $Y = -1$ alors $X \sim \mathcal{N}(\mu_0, \sigma)$

... comme en analyse de la variance à un facteur.

Modèle pour une variable réponse à deux groupes

Si on observe que $x \leq X \leq x+h$, que vaut

$$\pi(x;h) = \mathbb{P}(Y=+1 \mid x \leq X \leq x+h) \text{ ?}$$

D'après le théorème de Bayes,

$$\begin{aligned}\pi(x;h) &= \frac{\mathbb{P}(x \leq X \leq x+h \mid Y=+1)}{\mathbb{P}(x \leq X \leq x+h)}\mathbb{P}(Y=+1),\\ &= \frac{\mathbb{P}(x \leq X \leq x+h \mid Y=+1)}{\mathbb{P}(x \leq X \leq x+h)}p,\end{aligned}$$

où $p = \mathbb{P}(Y=+1)$ est la probabilité a priori de $Y=+1$.

Modèle pour une variable réponse à deux groupes

On en déduit l'expression de l'odds :

$$\begin{aligned}
\text{odds}(x;h) &= \frac{\pi(x;h)}{1-\pi(x;h)}, \\
&= \frac{\mathbb{P}(x \le X \le x+h \mid Y=+1)}{\mathbb{P}(x \le X \le x+h \mid Y=-1)} \frac{p}{1-p}, \\
&= \frac{F_{\mu_1,\sigma}(x+h) - F_{\mu_1,\sigma}(x)}{F_{\mu_0,\sigma}(x+h) - F_{\mu_0,\sigma}(x)} \frac{p}{1-p},
\end{aligned}$$

où $F_{\mu,\sigma}(x) = \mathbb{P}(U \le x)$, avec $U \sim \mathcal{N}(\mu,\sigma)$.

Lorsque h tend vers 0 :

$$\text{odds}(x) = \frac{\pi(x)}{1-\pi(x)} = \lim_{h\to 0} \frac{\pi(x;h)}{1-\pi(x;h)},$$

Modèle pour une variable réponse à deux groupes

Lorsque h tend vers 0 :

$$\begin{aligned} \text{odds}(x) &= \frac{p}{1-p} \lim_{h \to 0} \frac{\frac{F_{\mu_1,\sigma}(x+h)-F_{\mu_1,\sigma}(x)}{h}}{\frac{F_{\mu_0,\sigma}(x+h)-F_{\mu_0,\sigma}(x)}{h}}, \\ &= \frac{p}{1-p} \frac{F'_{\mu_1,\sigma}(x)}{F'_{\mu_0,\sigma}(x)}, \end{aligned}$$

où $F'_{\mu,\sigma}(x) = f_{\mu,\sigma}(x)$:

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Modèle pour une variable réponse à deux groupes

On simplifie :

$$
\begin{aligned}
\frac{\pi(x)}{1-\pi(x)} &= \frac{p}{1-p}\exp\Big[-\frac{1}{2\sigma^2}\Big((x-\mu_1)^2-(x-\mu_0)^2\Big)\Big], \\
&= \frac{p}{1-p}\exp\Big[-\frac{1}{2\sigma^2}\Big(-2(\mu_1-\mu_0)x+(\mu_1^2-\mu_0^2)\Big)\Big], \\
&= \frac{p}{1-p}\exp\Big[\frac{\mu_1-\mu_0}{\sigma^2}\Big(x-\frac{\mu_1+\mu_0}{2}\Big)\Big].
\end{aligned}
$$

Par conséquent,

- Si $\mu_0 = \mu_1$, alors $\pi(x) \equiv p$.
- L'effet de x sur $\pi(x)$ dépend de $(\mu_1-\mu_0)/\sigma^2$.

Modèle pour une variable réponse à deux groupes

Finalement :

$$\begin{aligned}\log\frac{\pi(x)}{1-\pi(x)} &= \log\frac{p}{1-p}+\frac{\mu_1-\mu_0}{\sigma^2}\left(x-\frac{\mu_1+\mu_0}{2}\right),\\ \operatorname{logit}\pi(x) &= \operatorname{logit}p+\left[\frac{\mu_1-\mu_0}{\sigma^2}\left(x-\frac{\mu_1+\mu_0}{2}\right)\right].\end{aligned}$$

où logit : $\pi \mapsto \log \pi/(1-\pi)$.

Modèle pour une variable réponse à deux groupes

Le **modèle de régression logistique de** $Y = \pm 1$ sur X est :

$$\operatorname{logit} \mathbb{P}(Y = +1 \mid X = x) = \beta_0 + \beta_1 x.$$

où β_0 et β_1 sont les paramètres du modèle.

- `mod = glm(y~x,family=binomial(link=logit)...)`
- RStudio

Modèle pour une variable réponse à deux groupes

odds-ratio lorsque X est quantitative

$$
\begin{aligned}
\text{odds-ratio} &= \frac{\text{odds}(x+1)}{\text{odds}(x)}, \\
&= \frac{\exp\left(\beta_0 + \beta_1(x+1)\right)}{\exp\left(\beta_0 + \beta_1 x\right)}, \\
&= \exp\left(\beta_1\right).
\end{aligned}
$$

- `questionr::odds.ratio(mod)`
-

Plan

Modèle pour une variable réponse à K groupes

Le **modèle de régression logistique multinomiale de** $Y \in \{y_1, \dots, y_K\}$ sur X est :

$$\left\{\begin{array}{lcl} \log \frac{\mathbb{P}(Y=y_2 \mid X=x)}{\mathbb{P}(Y=y_1 \mid X=x)} & = & \beta_0^{(2)} + \beta_1^{(2)} x \\ \log \frac{\mathbb{P}(Y=y_3 \mid X=x)}{\mathbb{P}(Y=y_1 \mid X=x)} & = & \beta_0^{(3)} + \beta_1^{(3)} x \\ & \vdots & \\ \log \frac{\mathbb{P}(Y=y_K \mid X=x)}{\mathbb{P}(Y=y_1 \mid X=x)} & = & \beta_0^{(K)} + \beta_1^{(K)} x \end{array}\right.$$

où $\beta_0^{(k)}$ et $\beta_1^{(k)}$ sont les $2(K-1)$ paramètres du modèle.

- `mod = multinom(y~x,data=...)`
- RStudio

Modèles avec plusieurs effets

- Deux variables explicatives quantitatives

$$\text{logit}\, \mathbb{P}(Y = +1) \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

où $\pi(x) = \mathbb{P}(Y = +1 \mid X_1 = x_1, X_2 = x_2)$.

- Deux facteurs

$$\text{logit}\, \pi_{ij} \quad = \quad \mu + \alpha_i + \beta_j$$

où $\pi_{ij} = \mathbb{P}(Y = +1)$ pour un individu de modalité i pour le 1er facteur et du groupe j pour le 2ème facteur.

Modèles avec plusieurs effets

- Deux facteurs et leur interaction

$$\text{logit}\, \pi_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

- Une variable explicative quantitative et un facteur

$$\text{logit}\, \pi_i(x) = \mu + \alpha_i + (\beta + \gamma_i)x$$

où $\pi_i(x) = \mathbb{P}(Y = +1 \mid X = x)$ pour un individu de modalité i.

► RStudio