

# **MATH 60604**

## **Modélisation statistique**

### **§ 4b - Régression logistique**

Léo Belzile

HEC Montréal  
Département de sciences de la décision

# Modèles linéaires généralisés pour données binaires

- Si notre variable réponse est binaire, on peut supposer que  $Y_i$  suit une loi Bernoulli de paramètre  $\pi_i$ ,  $Y_i \sim \text{Bin}(\pi_i)$ , où

$$\pi_i = P(Y_i = 1 \mid \mathbf{X}_i) = E(Y_i \mid \mathbf{X}_i).$$

- La fonction de liaison la plus courante pour les réponses binaires est la fonction **logit**

$$g(z) := \text{logit}(z) = \ln \left( \frac{z}{1-z} \right).$$

soit la **fonction quantile de la loi logistique** qui ici relie  $E(Y_i \mid \mathbf{X}_i) = \pi_i(\mathbf{X}_i)$  et  $\eta_i$ .

# Fonction de liaison logit

- Le modèle logistique spécifie

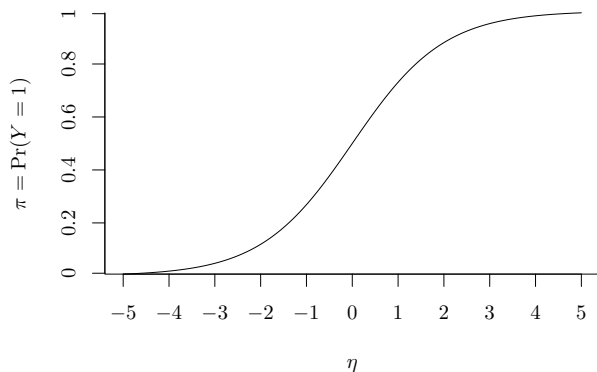
$$\eta_i = \ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

- Ce modèle peut être exprimé sur l'échelle de la moyenne à l'aide de la fonction inverse de **logit** (expit),

$$E(Y_i | \mathbf{X}_i) = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})}.$$

- Cela donne une expression pour la moyenne  $\pi_i = E(Y_i | \mathbf{X}_i)$  en fonction des variables explicatives  $\mathbf{X}_i$ , mais...
  - à quoi ressemble cette fonction?
  - que nous apprend-elle sur la relation entre  $\pi_i$  et  $\eta_i$  (et donc  $\mathbf{X}_i$ )?

# Fonction de répartition logistique



- On voit que  $\pi$  est une **fonction croissante** de  $\eta = \beta_0 + \sum_{j=1}^p \beta_j X_j$ .
  - Si  $\beta_j$  est positif et  $X_j$  augmente,  $P(Y=1)$  augmente aussi.
  - Si  $\beta_j$  est négatif et  $X_j$  augmente,  $P(Y=1)$  diminue.
- La relation entre  $P(Y=1)$  et  $\eta$  (et donc aussi  $X_j$ ) est **nonlinéaire**.

# Interprétation des paramètres de la régression logistique

- Quantifier les effets des paramètres  $\beta$  de la régression logistique est compliqué à cause de la nonlinéarité.
- L'effet des coefficients est en terme de **cote** et de **rapports de cote**.
- Soit  $\pi = P(Y = 1 | X_1, \dots, X_p)$  et le modèle de régression logistique

$$\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

- Si on prend l'exponentielle de chaque côté de l'équation précédente, on obtient

$$\text{cote}(Y | \mathbf{X}) = \frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})} = \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p),$$

où  $\pi(\mathbf{X})/\{1 - \pi(\mathbf{X})\}$  est la cote de  $P(Y = 1 | \mathbf{X})$  par rapport à  $P(Y = 0 | \mathbf{X})$ .

- L'utilisation de la fonction de liaison logit donne un modèle pour le **log de la cote**.
- La cote pour une variable binaire  $Y$  est le rapport

$$\text{cote}(\pi) = \frac{\pi}{1 - \pi} = \frac{P(Y = 1)}{P(Y = 0)}.$$

- Par exemple, une cote de 4 signifie que la probabilité de  $Y = 1$  est quatre fois plus élevée que la probabilité de  $Y = 0$ .
- Une cote de 0,25 à l'inverse veut dire que la probabilité de  $Y = 1$  est seulement le quart de celle pour  $Y = 0$ , ou encore que la probabilité de  $Y = 0$  est quatre fois plus élevée que celle pour  $Y = 1$ .

$P(Y = 1)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
cote	0,11	0,25	0,43	0,67	1	1,5	2,33	4	9
cote (frac.)	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9

# Interprétation de l'ordonnée à l'origine en termes de cote

- Si  $X_1 = \dots = X_p = 0$ , il est clair que

$$\text{cote}(Y \mid \mathbf{X} = \mathbf{0}_p) = \exp(\beta_0)$$

and

$$P(Y = 1 \mid \mathbf{X} = \mathbf{0}_p) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

représente la probabilité que  $Y = 1$  quand  $\mathbf{X} = \mathbf{0}_p$ .

- Comme pour la régression linéaire,  $X_1 = \dots = X_p = 0$  peut être impossible; dans ce cas,  $\beta_0$  ne s'interprète pas.

# Interprétation des paramètres en termes de rapports de cotes

On considère pour faire simple un modèle logistique de la forme

$$\text{logit}(\pi) = \beta_0 + \beta_1 x.$$

Le facteur  $\exp(\beta_1)$  est le changement dans la cote quand  $X$  augmente d'une unité,

$$\text{cote}(Y \mid X = x + 1) = \exp(\beta_1) \times \text{cote}(Y \mid X = x).$$

- Si  $\beta_1 = 0$ , le rapport des cotes vaut un:  $X$  n'impacte pas la cote de  $Y$
- Si  $\beta_1$  est positif, le rapport des cotes  $\exp(\beta_1)$  excède un: si  $X$  croît, la cote de  $Y$  augmente.
- Si  $\beta_1$  est négatif, le rapport des cotes  $\exp(\beta_1)$  est inférieur à un: si  $X$  croît, la cote de  $Y$  décroît.

Quand il y a plusieurs variables explicatives, l'interprétation de  $\beta_1$  est identique, mais elle n'est valide que quand **toutes les autres variables explicatives sont égales par ailleurs.**



# Interprétation de $\beta_k$ en terme de rapport de cote

Dans le modèle logistique, le **rapport de cotes** quand  $X_k = x_k + 1$  versus  $X_k = x_k$  quand  $X_j = x_j$  ( $j = 1, \dots, p, j \neq k$ ) est

$$\frac{\text{cote}(Y \mid X_k = x_k + 1, X_j = x_j, j \neq k)}{\text{cote}(Y \mid X_k = x_k, X_j = x_j, j \neq k)} = \frac{\exp\left(\beta_0 + \sum_{j=1}^p \beta_j x_j + \beta_k\right)}{\exp\left(\beta_0 + \sum_{j=1}^p \beta_j x_j\right)} \\ = \exp(\beta_k).$$

Quand  $X_k$  augmente d'une unité **et que la valeur de toutes les autres covariables est constante**, la cote de  $Y$  est multipliée par un  $\exp(\beta_k)$ .

- La cote augmente si  $\exp(\beta_k) > 1$ , c'est-à-dire si  $\beta_k > 0$ .
- La cote diminue si  $\exp(\beta_k) < 1$ , c'est-à-dire si  $\beta_k < 0$ .