Le modèle linéaire généralisé (GLM)

Les observations à expliquer peuvent être :

Discrètes :

- de type présence/absence, 0/1, succés/échec
  Modélisation du risque de survenue d'une défaillance, d'une pathologie. Les données à expliquer sont du type porteur / non porteur, contrôlées par des facteurs de risque.
- des effectifs
  Prédiction du taux d'exhautivité d'un registre d'handicap, de la taille d'une population porteuse d'un virus, de la taille d'une population animalière ...
  Nombres de défaillances

Continues ... mais où les hypothèses de loi gaussienne sont rejetées !

- observations disymétriques
- observations où la variance est une fonction de la moyenne.
  Exemple : Variance = Moyenne^2

Fiabilité du logiciel : modélisation du temps inter défaillance.

Évaluation d'algorithmes évolutionnaires : modélisation de la valeur atteinte avec un nombre fixé d'itération.

Trois hypothèses permettent de caractériser un **GLM** :

- la distribution de la variable à expliquer

$$f(y_i, \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right),$$

θ_i est un paramètre *canonique*
ϕ un paramètre de *dispersion* > 0.
Les fonctions a, b et c sont spécifiques à chaque distribution.

$$\begin{aligned} E(Y_i) &= \mu_i = b'(\theta_i) \\ V(Y_i) &= a(\phi)b''(\theta_i) \end{aligned}$$

exemple : la loi Gamma

$$\forall y \in R_+ , \quad G(\alpha, \lambda)(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} , \quad \lambda > 0, \alpha > 0 .$$

$$\theta = \frac{\lambda}{\alpha} , \quad b(\theta) = \log(\theta) , \quad \phi = \frac{1}{\alpha} , \quad a(\phi) = -\phi .$$

- **le prédicteur linéaire** l'expression de la linéarité mettant en jeu les variables explicatives

$$\eta = X\beta,$$

- **la fonction de lien** la fonction qui relie le prédicteur à l'espérance mathématique de la variable à expliquer (*fonction de lien* g).

$$\begin{aligned} \eta &= g(\mu) \\ X\beta &= g(b'(\theta)) \end{aligned}$$

g est la fonction de lien canonique si $g = b'^{-1}$

La log-vraisemblance $\mathcal{L}$

$$L(\beta; y) = \sum_{i=1}^{N} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right] = \sum_{i=1}^{N} L_i(\beta_i; y_i) \ .$$

où $\theta_i = (g \circ b')^{-1}(X\beta)_i$

Les dérivées de vraisemblance pour $i \in \{1, ..., N\}$, $j \in \{1, ..., d\}$:

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial \eta_i}{\partial \beta_j} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial L_i}{\partial \theta_i} = X_{ij} \frac{1}{g'(\mu_i)} \frac{1}{b''(\theta_i)} \frac{y_i - \mu_i}{a(\phi)} \ ,$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{N} X_{ij} \frac{1}{g'(\mu_i)^2 Var(Y_i)} g'(\mu_i) \ (y_i - \mu_i) \ .$$

Estimateur maximum de vraisemblance

et son équation normale :

$$X' W_\beta^{-1} \frac{\partial \eta}{\partial \mu} (y - \mu) = 0$$

avec

$$W_\beta = (Var(Y_i) g'(\mu_i)^2 \delta_{ij})_{i,j=1,\ldots,N}$$

L'estimation du paramètre inconnu β par maximum de vraisemblance se fait à l'aide d'un algorithme itératif.

$$\begin{aligned} \beta^{[t+1]} &= \beta^{[t]} - \left(E\left[\left\{ \frac{\partial^2 L}{\partial \beta \partial \beta'} \right\} \right]^{[t]} \right)^{-1} \frac{\partial L}{\partial \beta}^{[t]} \\ &= \beta^{[t]} + \left(X' W_{\beta^{[t]}}^{-1} X \right)^{-1} X' W_{\beta^{[t]}}^{-1} \frac{d\eta}{d\mu}^{[t]} \left(Y - \mu^{[t]} \right) \end{aligned}$$

$$\beta^{[t+1]} = \left(X' W_{\beta^{[t]}}^{-1} X\right)^{-1} X' W_{\beta^{[t]}}^{-1} z^{[t]}$$

avec $z^{[t]} = X\beta^{[t]} + \dfrac{d\eta}{d\mu}^{[t]} (y - \mu^{[t]})$

Modèle linéarisé $\mathcal{M}^{[t]}$:

$$Z = X\beta + \varepsilon$$

où $E(\varepsilon) = 0$
$Var(\varepsilon) = W_{\beta^{[t]}}$