Compléments Régression Logistique

Le modèle linéaire gaussien permet une modélisation linéaire de l'espérance d'un prédictand gaussien Y, conditionnée par les q prédicteurs :

$$E[Y|X_1,...,X_q] = \beta_0 + \sum_{j=1}^{q} \beta_j X_j$$

D'après la théorie du modèle linéaire généralisé, pour une variable aléatoire non gaussienne Y distribuée selon une loi appartenant à la famille exponentielle, il est également possible de modéliser linéairement non pas directement son espérance conditionnelle mais une fonction de cette espérance :

$$g(E[Y|X_1,...,X_q]) = \beta_0 + \sum_{j=1}^{q} \beta_j X_j$$

La fonction g est appelée fonction de lien canonique, à chaque loi de la famille exponentielle est associée une fonction de lien canonique. Un tel modèle intègrera la relation entre espérance et variance d'une variable non gaussienne (relation qui induit un comportement hétéroscédastique de l'erreur dans le modèle gaussien). La loi normale appartient à la famille exponentielle, le modèle linéaire gaussien constitue donc un modèle linéaire généralisé particulier, exploitant la fonction de lien identité.
Après estimation des paramètres par maximisation de la vraisemblance, les estimations de l'espérance conditionnelle du prédictand Y sont obtenues par exploitation de la fonction réciproque de g.

$$\hat{E}[Y|X_1,...,X_q] = g^{-1}\left(\hat{\beta}_0 + \sum_{j=1}^{q} \hat{\beta}_j X_j\right)$$

(comme pour le modèle linéaire gaussien, des prédicteurs qualitatifs peuvent également être exploités)

Famille (ou classe) exponentielle de lois :

La loi suivie par la variable aléatoire Y appartient à la famille exponentielle si la densité de probabilité f(y) (ou la probabilité P(Y=y) pour les lois discrètes) peut s'exprimer sous la forme suivante :

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

avec a, b, c trois fonctions arbitraires,

θ paramètre naturel (ou canonique) de la loi, fonction de E[Y] :

$\theta = g(E[Y])$ avec g = fonction de lien canonique,

ϕ paramètre d'échelle (absent si la loi ne comporte qu'un paramètre)

On montre alors que : $E[Y]=b'(\theta)$ et $V[Y]=a(\phi)b''(\theta)$ (cf. démo en fin de document)

- Cas de la loi normale de paramètre μ et σ^2 ($E[Y]=\mu$ et $V[Y]=\sigma^2$ ne sont pas liées) **:**

La densité normale f(Y) peut s'exprimer sous la forme exponentielle :

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(y-\mu)^2\right] = \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \ln\left(\sqrt{2\pi\sigma^2}\right)\right]$$

On peut alors définir les quantités suivantes :

θ = paramètre naturel = μ = E[Y] → fonction de lien g = identité

$\phi = \sigma^2 = a(\phi)$ → fonction a = identité

$b(\theta)=\theta^2/2$ et $c(y,\phi) = -\frac{y^2}{2\phi} - \ln\left(\sqrt{2\pi\phi}\right)$

La loi normale appartient donc à la famille exponentielle, la fonction de lien canonique étant l'identité.

- Cas de la loi de Bernoulli de paramètre p :

P(Y=1) = p = probabilité de succès
P(Y=0) = 1 – p

E[Y] = p
V[Y] = p(1-p) = E[Y] (1- E[Y])

On peut donc exprimer la loi de probabilités ainsi, avec y = 0 ou 1 :

$$P(Y=y) = p^y(1-p)^{1-y} = \exp\left[y\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right]$$

On peut alors définir les quantités suivantes :

θ = paramètre naturel = ln(p/(1-p)) = g(E[Y]) → fonction de lien g = fonction logit

La loi de Bernoulli ne comportant qu'un seul paramètre, le paramètre d'échelle ϕ n'intervient pas.

$a(\phi) = 1$, $c(y, \phi) = 0$ et $b(\theta) = -\ln(1-p) = \ln(1+e^{\theta})$

La loi de Bernoulli appartient donc à la famille exponentielle, la fonction de lien canonique étant la fonction logit. Un modèle de régression logistique exploitant q prédicteurs sera donc défini ainsi :

$$g(E[Y|X_1,...,X_q]) = \ln\left(\frac{P(Y=1|X_1,...,X_q)}{1-P(Y=1|X_1,...,X_q)}\right) = \beta_0 + \sum_{j=1}^{q}\beta_j X_j$$

Après estimation des paramètres du modèle théorique, les estimations des probabilités de succès conditionnelles seront obtenues via la fonction logistique :

$$\hat{P}(Y=1|X_1,...,X_q) = g^{-1}(\hat{\beta}_0 + \sum_{j=1}^{q}\hat{\beta}_j X_j) = \frac{1}{1+\exp\left[-(\hat{\beta}_0 + \sum_{j=1}^{q}\hat{\beta}_j X_j)\right]}$$

Remarque : il existe la régression logistique multinomiale pour un prédictand à plus de 2 modalités.

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

La log-vraisemblance l(θ) s'exprime ainsi (en considérant une seule mesure de Y) :

$$l(\theta) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

$$l'(\theta) = \frac{y - b'(\theta)}{a(\phi)} \quad \rightarrow \quad E[l'(\theta)] = \frac{E[Y] - b'(\theta)}{a(\phi)}$$

D'après la théorie de la vraisemblance, cette espérance est nulle pour la vraie valeur de θ, on a donc :

$$E[Y] = b'(\theta)$$

$$l''(\theta) = \frac{-b''(\theta)}{a(\phi)} = E[l''(\theta)]$$

D'après la théorie de la vraisemblance, on a : $E[(l'(\theta))^2] = -E[l''(\theta)] = I(\theta) = Inf.Fisher$

$$E[(l'(\theta))^2] = E\left[\frac{(Y - b'(\theta))^2}{a^2(\phi)}\right] = \frac{E[(Y - E[Y])^2]}{a^2(\phi)} = \frac{V[Y]}{a^2(\phi)}$$

On obtient donc :

$$V[Y] = b''(\theta)a(\phi)$$

(cf. Rappel Information de Fisher sur moodle pour les propriétés de la vraisemblance utilisées ici)