Complément formateur : L'algorithme IRLS

IRLS signifie « Iteratively Reweighted Least Squares », en français les moindres carrés itérativement repondérés. C'est la procédure classiquement utilisée dans le cadre du modèle linéaire généralisé (glm).

I- Rappel du principe général de minimisation d'une fonction :

Soit une fonction, par exemple un critère $C(y_i, m(X,\beta))$, qui dépend des valeurs observées y_i, et d'un modèle m associant variables explicatives X et vecteur de paramètres β. Observées et variables explicatives étant connues et constantes, C est une fonction de β, et nous simplifions la notation en $C(\beta)$. Sous certaines conditions, minimiser C en jouant sur β revient à annuler le gradient (les dérivées premières) de C par rapport à β. Si C n'est pas linéaire par rapport à β, il n'est en général pas possible de trouver une solution explicite. On procède alors de manière itérative, en modifiant à chaque pas un vecteur β^κ pour obtenir un vecteur $\beta^{\kappa+1}$ qui diminue C. Soit donc à trouver une correction δ^κ telle que :

$$\beta^{\kappa+1} = \beta^\kappa + \delta^\kappa \qquad \text{vérifiant} \qquad C(\beta^{\kappa+1}) < C(\beta^\kappa)$$

Un minimum est atteint, en espérant que ce soit le bon, pour une valeur de β qui annule le gradient G de C. Partant d'une valeur non nulle $G(\beta^\kappa)$ calculée avec une valeur initiale β^κ du vecteur des paramètres, on cherche un vecteur de correction des paramètres δ^κ tel que $G(\beta^\kappa+\delta^\kappa) = 0$. En développant à l'ordre 1 au voisinage de β^κ, et en notant classiquement ' la dérivée quand il n'y a pas d'ambiguïté il vient :

$$G(\beta^\kappa+\delta^\kappa) = G(\beta^\kappa) + G'(\beta^\kappa)\delta^\kappa + \varepsilon(\delta^\kappa)$$

L'annulation de cette expression en négligeant ε nous donne la correction recherchée :

$$\delta^\kappa = -G'(\beta^\kappa)^{-1}G(\beta^\kappa)$$

avec $G(\beta^\kappa) = \dfrac{\partial C(\beta^k)}{\partial \beta^k}$ et $G'(\beta^\kappa) = H(\beta^\kappa) = \dfrac{\partial G(\beta^k)}{\partial \beta^k} = \dfrac{\partial^2 C(\beta^k)}{\partial \beta^{k\,2}}$

G est un vecteur de dimension p, nombre de paramètres du modèle. H est le Hessien de C, c'est à dire la matrice des dérivées secondes du critère par rapport aux paramètres. C'est donc une matrice pxp.

Pour trouver δ^κ, il faut connaître G et H. Si on calcule effectivement H, on applique la méthode de Newton-Raphson. Une alternative qui converge sans biais vers la solution est de remplacer H par son espérance. Elle est connue sous le nom d'algorithme de Fisher (Fisher scoring).

Remarques:

- Minimiser un critère revient à maximiser son opposé et vice-versa.
- Pour un fonctionnement optimal de l'algorithme et une bonne estimation de la matrice de variance covariance, l'utilisateur doit en principe fournir G et H. Dans la pratique, H ou son espérance est presque toujours estimé par l'algorithme, et G lui-même est estimé numériquement par la plupart des algorithmes.

II- Application à la vraisemblance des modèles de la famille exponentielle.

Si C est le critère du maximum de vraisemblance, c'est à dire le produit des densités de probabilité $f(y_i)$ des réalisations y_i il vient :

$C = \prod_{i=1}^{n} f(y_i)$ n, le nombre d'observations

avec f de la famille exponentielle, c'est à dire pouvant s'écrire sous la forme :

$f(y_i) = \exp\{\frac{\omega_i}{\varphi}(y_i\theta_i - b(\theta_i)) + c(y_i,\varphi)\}$ avec :

φ constante d'échelle, appelé aussi facteur de dispersion (1 dans le cas de la loi binomiale)
ω_i constantes de poids (n_i, nombre de réalisations binaires de y_i dans le cas binomial)
θ_i paramètre canonique

et : $E(Y_i) = \mu_i = \frac{db(\theta_i)}{d\theta_i} = b'(\theta_i)$ $Var(Y_i) = V_i = \frac{d^2b(\theta_i)}{d\theta_i^2}\frac{\varphi}{\omega_i} = b''(\theta_i)\frac{\varphi}{\omega_i}$

Partant du principe qu'il est plus facile de manipuler des sommes que des produits, on prend le logarithme de C. La fonction log étant monotone, maximiser log(C) (ou minimiser -log(C)) revient à maximiser C.

$$\log(C) = \sum_{i=1}^{n}\frac{\omega_i}{\varphi}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\varphi)$$

Il nous faut donc les dérivées premières (G) et secondes (H) de log(C) par rapport à β.

Calcul de G. Le deuxième terme (la fonction c) dans log(C) ne dépendant pas de β il vient :

$G(\beta) = \sum_{i=1}^{n}\frac{\omega_i}{\varphi}(y_i - \frac{db(\theta_i)}{d\theta_i})\frac{\partial\theta_i}{\partial\beta}$ Or on peut écrire : $\frac{\partial\theta_i}{\partial\beta} = \frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta}$ avec :

μ_i espérance de Y_i
$\eta_i = X_i^T \beta$ prédicteur linéaire, dépendant des variables explicatives au point i et des paramètres β (X est la classique matrice d'incidence nxp construite d'après les valeurs des variables explicatives, (X_i^T est la transposée du vecteur fourni par la $i^{ème}$ ligne).
$\eta_i = g(\mu_i)$ g fonction de lien, choisie a priori et sans paramètre

On écrit bien que l'espérance $E(Y_i) = \mu_i = g^{-1}(X_i^T \beta) = m(X_i, \beta)$ est modélisée comme une fonction de X et de β. Cette formulation fait apparaître en outre un espace, l'espace des η_i, où le modèle est linéaire par rapport à ses paramètres.

Or $E(Y_i) = \mu_i = \frac{db(\theta_i)}{d\theta_i} = b'(\theta_i)$ $Var(Y_i) = V_i = \frac{d^2b(\theta_i)}{d\theta_i^2}\frac{\varphi}{\omega_i} = b''(\theta_i)\frac{\varphi}{\omega_i}$ donc :

$\frac{d\mu_i}{d\theta_i} = b''(\theta_i)$ $\frac{\partial\eta_i}{\partial\mu_i} = g'(\mu_i)$ $\frac{\partial\eta_i}{\partial\beta} = X_i$ (soit pour chaque paramètre $\frac{\partial\eta_i}{\partial\beta_j} = x_{ij}$)

On peut alors écrire : $\frac{\partial\theta_i}{\partial\beta} = \frac{1}{b''(\theta_i)}\frac{1}{g'(\mu_i)}X_i = \varphi\frac{X_i}{V_i\omega_i g'(\mu_i)}$

et il vient : $G(\beta) = \sum_{i=1}^{n}\frac{\omega_i}{\varphi}(y_i - \mu_i)\varphi\frac{X_i}{V_i\omega_i g'(\mu_i)} = \sum_{i=1}^{n}\frac{(y_i - \mu_i)}{V_i g'(\mu_i)}X_i$

Calcul de E(H). H est une matrice pxp dont nous noterons le terme général $h_{j,k}$.

On peut montrer que $E\left(h_{j,k}\right) = E\left(\frac{\partial^2 \log(L(\beta))}{\partial\beta_j \partial\beta_k}\right) = -E\left(\frac{\partial \log(L(\beta))}{\partial\beta_j}\frac{\partial \log(L(\beta))}{\partial\beta_k}\right)$

donc $-E(h_{j,k}) = E(\sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V_i g'(\mu_i)} x_{i,j} \sum_{i'=1}^{n} \frac{(y_{i'}-\mu_{i'})}{V_{i'} g'(\mu_{i'})} x_{i',k})$

$$= E(\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{(V_i g'(\mu_i))^2} x_{i,j} x_{i,k}) + E(\sum_{i\neq i'}^{n} \frac{(y_i-\mu_i)(y_{i'}-\mu_{i'})}{V_i V_{i'} g'(\mu_i) g'(\mu_{i'})} x_{i,j} x_{i',k})$$

Si les observations sont indépendantes et le modèle valide, alors la covariance de $(Y_i, Y_{i'})$ est nulle, donc l'espérance des termes $(y_i - \mu_i)(y_{i'} - \mu_{i'})$ est nulle. Et comme l'espérance des termes en $(y_i - \mu_i)^2$ est égale à la variance de Y_i notée V_i il vient :

$$-E(h_{j,k}) = \sum_{i=1}^{n} \frac{1}{V_i g'(\mu_i)^2} x_{ij} x_{ik}$$

Remarque : l'inverse de -E(H) est dans ce cas la matrice asymptotique de variance covariance des paramètres. Si le nombre d'observations est "suffisamment grand", sa diagonale donne les variances non biaisées des paramètres. -E(H) est alors appelée matrice d'information et parfois notée I.

III- Notions de variable de travail et de poids

Si on introduit un terme w_i (que l'on qualifiera de pondération) et un terme y_i^* tels que:

$w_i = 1/ (V_i\, g'(\mu_i)^2)$ et $y_i^* = (y_i - \mu_i)\, g'(\mu_i)$

alors les termes généraux de I et G s'écrivent :

$$-E(h_{j,k}) = \sum_{i=1}^{n} w_i x_{ij} x_{ik} \qquad G_j = \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V_i g'(\mu_i)} x_{i,j} = \sum_{i=1}^{n} w_i (y_i-\mu_i) g'(\mu_i) x_{i,j} = \sum_{i=1}^{n} w_i y_i^* x_{i,j}$$

Un calcul très simple montre en outre que $var(Y_i^*) = 1/w_i$

Sous forme matricielle, et en notant y* le vecteur de terme y_i^*, on a :

$G = X^T W y^*$ et $-E = X^T W X$

Et si on indice par k les termes obtenus à l'étape k du processus itératif on peut écrire :

$$\beta^{\kappa+1} = \beta^{\kappa} + (X^T W^k X)^{-1} X^T W^k y^{k*}$$

en multipliant β^{κ} par $(X^T W^k X)^{-1} (X^T W^k X)$ il vient :

$$\beta^{\kappa+1} = (X^T W^k X)^{-1} [X^T W^k X \beta^{\kappa} + X^T W^k y^{k*}] = (X^T W^k X)^{-1} X^T W^k (X\beta^{\kappa} + y^{k*})$$

Si on note maintenant $z^k = X\beta^{\kappa} + y^{k*}$ on obtient :

$$\beta^{\kappa+1} = (X^T W^k X)^{-1} X^T W^k z^k$$

Où l'on reconnaît la solution d'une régression linéaire pondérée de z^k sur X.

Conclusion, dans le cas où les observées suivent une loi de distribution de la famille exponentielle, *chaque étape du processus itératif de maximisation de la vraisemblance est équivalente à une régression linéaire d'une variable de travail bien choisie pondérée par l'inverse de sa variance*. C'est le principe de l'algorithme IRLS.

IV- Maximum de vraisemblance et statistique du χ^2 de Pearson

A chaque étape k du processus itératif, la phase de régression linéaire conduit donc à minimiser la somme des carrés des écarts pondérés :

$$SCEP = \sum_{i=1}^{n} w_i^k (z_i^k - X_i^T \beta^{k+1})^2$$

avec $z_i^k = X_i^t \beta^k + y_i^{k*} = X_i^t \beta^k + (y_i - \mu_i^k)\, g'(\mu_i^k)$ et $(w_i^k)^{-1} = V_i\, g'(\mu_i^k)^2$

donc $$SCEP = \sum_{i=1}^{n} \frac{(X_i^T \beta^k + (y_i - \mu_i^k) g'(\mu_i^k) - X_i^T \beta^{k+1})^2}{V_i g'(\mu_i^k)_i^2}$$

A la convergence $\beta^{k+1} = \beta^{\kappa}$ et on obtient comme somme des écarts pondérés :

$$SCEP = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V_i} = \chi^2$$

A la convergence, ce critère est la somme des carrés des écarts entre observées et estimées divisés par la variance des estimées, c'est à dire la statistique du χ^2 de Pearson. Un simple développement limité montre que ce χ^2 est une approximation quadratique de la déviance au minimum. Tous les résultats asymptotiques sur les estimateurs du maximum de vraisemblance, et en particulier tous les tests, sont basés sur cette propriété.

Considérons maintenant ce critère. Toujours sous certaines conditions, le minimiser revient à annuler l'expression :

$$\frac{\partial SCEP}{\partial \beta} = -2 \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V_i} \frac{\partial \mu_i}{\partial \beta} - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V_i^2} \frac{\partial V_i}{\partial \beta} \qquad \text{or :} \quad \frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{X_i}{g'(\mu_i)}$$

donc $$\frac{\partial SCEP}{\partial \beta} = -2 \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V_i g'(\mu_i)} X_i - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V_i^2} \frac{\partial V_i}{\partial \beta} = -2G(\beta) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V_i^2} \frac{\partial V_i}{\partial \beta}$$

Si V_i est connu, le deuxième terme est nul et minimiser ce critère revient à annuler notre gradient initial G. Conclusion, si les observées suivent une loi appartenant à la famille exponentielle, et si leur variance est connue a priori, faire du maximum de vraisemblance ou des moindres carrés pondérés par l'inverse de la variance revient au même, que l'on soit dans l'échelle initiale des observées, ou dans une échelle où le modèle est linéaire. Mais si V_i dépend de β, minimiser la distance du χ^2 n'est plus équivalent à maximiser la vraisemblance.

V- Interprétation de la variable de travail en terme de transformation :

Une transformation intéressante est celle qui fait passer y_i dans une échelle où sa liaison avec X est linéaire. Mais dans le cas général, et si l'on note g cette transformation par analogie avec la fonction de lien, $z_i = g(y_i)$ n'est pas toujours calculable, ni surtout $var(Z_i)$. Si par contre on écrit y_i comme la somme de son espérance et d'une erreur :

$$y_i = \mu_i + (y_i - \mu_i)$$

$(y_i - \mu_i)$ étant supposé petit, on applique un développement limité de la transformation g au point μ_i, ce qui nous donne une valeur approchée de $g(y_i)$ calculable pour tout y_i, où l'on reconnaît notre variable de travail à la convergence :

$$g(y_i) = g(\mu_i + (y_i - \mu_i)) \approx g(\mu_i) + (y_i - \mu_i)\ g'(\mu_i) = \eta_i + (y_i - \mu_i)\ g'(\mu_i) = z_i$$

A la convergence, z_i s'interprète donc non comme une transformation de y_i, mais comme un développement limité de cette transformation au voisinage de la solution recherchée μ_i.

Nous avons également : $\text{var}(Z_i) = \text{var}(Y_i\ g'(\mu_i)) = V_i\ g'(\mu_i)^2 = 1/w_i$

La pondération est bien l'inverse de la variance de Z_i. Comme on recherche une liaison linéaire entre z et X, une simple régression pondérée suffirait pour trouver le résultat. Mais z_i et $\text{var}(Z_i)$ dépendent de la vrai valeur inconnue de μ_i et il faut procéder de manière itérative.

Application au lien logit sur des données binomiales. Nos notations deviennent :

y_i	nombre de succès de n_i processus binaires de probabilité π_i
g	fonction logit
$p_i = y_i/n_i$	probabilité observée pour l'observation i, réalisation d'une variable aléatoire P_i
π_i et V_i	espérance et variance de P_i

Nous avons : $g(\pi_i) = \text{logit}(\pi_i) = \log(\dfrac{\pi_i}{1-\pi_i})$ et $V_i = \text{var}(P_i) = \dfrac{\pi_i(1-\pi_i)}{n_i}$

donc $g'(\pi_i) = \dfrac{\partial g(\pi_i)}{\partial \pi_i} = ... = \dfrac{1}{\pi_i(1-\pi_i)} = \dfrac{1}{\pi_i V_i}$ et il vient : $z_i = \eta_i + \dfrac{(p_i - \pi_i)}{\pi_i(1-\pi_i)}$

Pondération : $w_i^{-1} = V_i g(\pi_i)^2 = \dfrac{\pi_i(1-\pi_i)}{n_i}(\dfrac{1}{\pi_i(1-\pi_i)})^2 = \dfrac{1}{n_i\pi_i(1-\pi_i)}$

donc $w_i = n_i\pi_i(1-\pi_i)$

VI- Mise en oeuvre de l'algorithme IRLS

IRLS procède donc pas à pas. Il utilise des valeurs de départ η_{i0} et μ_{i0} pour effectuer un premier calcul de z_{i0} et w_{i0} et faire la régression linéaire. Ceci conduit à obtenir un vecteur de paramètre β_1 qui permet de calculer $\eta_{i1} = x_i^T \beta_1$ et $\mu_{i1} = g^{-1}(\eta_{i1})$. Et l'on recommence le processus jusqu'à stabilisation.

Comme valeurs de départ, le plus simple est d'estimer les μ_i par les y_i. mais on retombe sur le problème du calcul de $\eta_i = g(\mu_i)$ pour les valeurs 0 ou 1. Une solution est de modifier les probabilités observées en ramenant un peu les probabilités vers 0,5 (dans le cas binaire, toutes les valeurs deviennent égales à 0,5).

$$\hat{p}_{i0} = p_i + \frac{0,5 - p_i}{n_i}$$