MODELES LINEAIRES GENERALISES

Famille de modèles:

- Présentés pour la première fois sous ce nom par Nelder et Wedderburn (1972)

- Exposés de façon complète par Mc Cullagh et Nelder (1989)

Les modèles linéaires généralisés permettent d'étudier la liaison entre une variable dépendante ou réponse Y et un ensemble de variables explicatives ou prédicteurs X_1 …. X_K

Ils englobent:- **le modèle linéaire général** (régression multiple, analyse de la variance et analyse de la covariance)

- le modèle log-linéaire

- la régression logistique

- la régression de Poisson.

I. Présentation des modèles linéaires généralisés

Les modèles linéaires généralisés sont formés de trois composantes:

la variable de réponse Y, composante aléatoire à laquelle est associée une loi de probabilité.

- les variables explicatives X_1 …. X_K utilisées comme prédicteurs dans le modèle définissent sous forme d'une combinaison linéaire la **composante déterministe.**

- **le lien** décrit la relation fonctionnelle entre la combinaison linéaire des variables X_1 ….X_K et l'espérance mathématique de la variable de réponse Y.

I.1.1 Composante aléatoire

La loi de probabilité de la composante aléatoire appartient à la famille **exponentielle**

Notons (Y_1, …., Y_n) un échantillon aléatoire de taille n de la variable de réponse Y, les variables aléatoires Y_1…. Y_n étant supposées indépendantes.

Y_i peut être binaire (succès-échecs, présence-absence)
Loi de Bernoulli, loi binomiale,

Y_i peut être distribuée selon une **loi de Poisson**

Y_i peut être distribuée selon une **loi normale**

I.1.2 Composante déterministe

La composante déterministe, exprimée sous forme d'une combinaison linéaire $\beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$ (appelée aussi prédicteur linéaire) précise quels sont les prédicteurs.

Certaines des variables X_j peuvent se déduire de variables initiales utilisées dans le modèle, par exemple:

- $\mathbf{X_3 = X_{1*}X_2}$ de façon à étudier l'interaction entre X_1et X_2

- ou encore $\mathbf{X_4 = X_1^2}$ de façon à prendre en compte un effet non linéaire de la variable X_1

I.1.3 Lien

La troisième composante d'un modèle linéaire généralisé est le **lien entre la composante aléatoire et la composante déterministe.**

Il spécifie comment l'espérance mathématique de Y notée μ est liée au prédicteur linéaire construit à partir des variables explicatives.

On peut modéliser l'espérance μ directement (régression linéaire usuelle) ou modéliser une fonction monotone $g(\mu)$ de l'espérance:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K$$

La fonction de lien g(μ) = log(μ) permet par exemple de modéliser le logarithme de l'espérance. Les modèles utilisant cette fonction de lien sont des **modèles log-linéaires.**

La fonction de lien g(μ) = $\log(\frac{\mu}{1-\mu})$ **modélise le logarithme du rapport des chances.** Elle est appelée **logit** et est adaptée au cas où μ est comprise entre 0 et1 (par exemple la probabilité de succès dans une loi binomiale).

A toute loi de probabilité de la composante aléatoire est associée une fonction spécifique de l'espérance appelée paramètre canonique.

Pour la distribution normale il s'agit de l'espérance elle-même.

Pour la distribution de Poisson le paramètre canonique est le logarithme de l'espérance.

Pour la distribution binomiale le paramètre canonique est le logit de la probabilité de succès.

La fonction de lien qui utilise le paramètre canonique dans la famille des modèles linéaires généralisés, est appelée la fonction de lien canonique.

En pratique, dans de nombreux cas les modèles linéaires généralisés sont construits en utilisant la fonction de lien canonique.

I.2 Loi de probabilité de la réponse Y

La loi de probabilité de la réponse Y_i doit appartenir à la **famille exponentielle**.

$$f(y_i, \theta_i, \varphi, \omega_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\varphi)} \omega_i + c(y_i, \varphi, \omega_i) \right) \quad (1)$$

Les fonctions a, b et c sont spécifiées en fonction du type de loi exponentielle.

- Le **paramètre canonique θ_i** est inconnu. Le paramètre canonique θ_i est une fonction de l'espérance $\theta_i = g_c(\mu_i)$.
- Le **paramètre de dispersion φ** est supposé connu.
- Si ce n'est pas le cas, ce paramètre (dit paramètre de nuisance car pour certaines valeurs la densité (1) peut ne pas appartenir à la famille exponentielle) est estimé préalablement et considéré ensuite comme connu. On le note φ_0 ; le plus souvent : $a(\varphi_0) = \varphi_0$.
- ω_i est un poids.

Les lois de probabilités telles que **la loi normale, la loi binomiale, la loi de Poisson, la loi Gamma et la loi de Gauss inverse** appartiennent à la famille exponentielle décrite précédemment.

Par ailleurs les propriétés de la **fonction score**, permettent d'établir que:

$$\mu_i = E(Y_i) = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

$$Var(Y_i) = a(\varphi_0) b''(\theta_i) \quad \text{où} \quad b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$$

TABLEAU 1 : Composantes de la famille exponentielle

Distribution	$\theta(\mu)$	$b(\theta)$	$a(\varphi_0)$
Normale $N(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2
Bernouilli $B(1,\mu)$	$\log\frac{\mu}{1-\mu}$	$\log(1+e^{\theta})$	1
Poisson $P(\mu)$	$\log\mu$	e^{θ}	1
Gamma $G(\mu,\nu)$	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\frac{1}{\nu}$
Gauss Inverse $I.G(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	σ^2

TABLEAU 2 : Espérance et variance

Distribution	$E(Y) = b'(\theta)$	$V(Y) = b''(\theta)a(l_0)$
Normale	$\mu = \theta$	σ^2
Bernouilli	$\mu = \frac{e^{\theta}}{1+e^{\theta}}$	$\mu(1-\mu)$
Poisson	$\mu = e^{\theta}$	μ
Gamma	$\mu = -\frac{1}{\theta}$	$\frac{\mu^2}{\nu}$
Gauss Inverse	$\mu = (-2\theta)^{-1/2}$	$\mu^3\sigma^2$

II Principes d'estimation d'un modèle linéaire généralisé

Pour la plupart des modèles linéaires généralisés, les équations qui déterminent les paramètres au sens du maximum de vraisemblance sont non linéaires et les estimateurs n'ont pas d'autres expressions formulables que comme solutions de ces équations.

Les logiciels calculent les estimations en utilisant un algorithme itératif pour la résolution d'équations non linéaires.

II Principes d'estimation d'un modèle linéaire généralisé

Un algorithme populaire pour atteindre cet objectif est appelé **« Fisher scoring »** et a été proposé initialement pour ajuster des modèles probit.

Pour la régression logistique binomiale et pour les modèles log-linéaires de Poisson cet algorithme se simplifie et n'est alors qu'une version du très connu **algorithme de Newton-Raphson.**

L’algorithme de Newton-Raphson

L’algorithme de Newton-Raphson approxime le log de la fonction de vraisemblance dans un voisinage du paramètre initial par une fonction polynomiale qui a la forme d’une parabole concave.

Elle a la même pente et la même courbure dans les conditions initiales que la log-fonction de vraisemblance. Il est facile de déterminer le maximum de ce polynôme d’approximation.
Ce maximum fournit la seconde étape du processus d’estimation et l’on reprend la procédure décrite précédemment.

L’algorithme de Newton-Raphson

Les approximations successives convergent rapidement vers les estimations au sens du maximum de vraisemblance.

Les logiciels usuels ne demandent pas à l’utilisateur de préciser les conditions initiales.

L’algorithme de Newton-Raphson

Chaque étape de l’algorithme de Newton Raphson constitue un ajustement de type moindres carrés pondérés.

Ceci est une généralisation des moindres carré ordinaires qui prend en compte la non-constance de la variance de Y dans les modèles.

Les observations recueillies en des points où la variabilité est plus faible sont affectées d’un poids plus important dans la détermination des paramètres.

A chaque itération les poids sont remis à jour d’où le terme parfois utilisé de **« moindres carrés itérativement repondérés ».**

L’algorithme de Newton-Raphson

L’algorithme de Newton-Raphson utilise la matrice d’information de Fisher.

Cette matrice contient l’information concernant la courbure de la fonction de log-vraisemblance au point d’estimation.

Plus grande est la courbure, plus l’information apportée au sujet des paramètres du modèle est importante.

En effet, les écarts-types des estimateurs sont les racines carrées des éléments diagonaux de l’inverse de la matrice d’information de Fisher.

Plus la courbure de la fonction de log-vraisemblance est importante, plus les écarts-types sont petits.

L'algorithme de Newton-Raphson

Plus la courbure de la fonction de log-vraisemblance est importante, plus les écarts-types sont petits.

Ceci est raisonnable dans la mesure où **une grande courbure implique que le log de la vraisemblance diminue rapidement quand on s'éloigne de l'estimation** $\hat{\beta}$

En conséquence les données observées ont plus de chances d'apparaître si β prend la valeur $\hat{\beta}$ qu'une valeur éloignée de $\hat{\beta}$

III Construction pratique d'un modèle linéaire généralisé

Après avoir indiqué brièvement comment choisir le type de modèle à construire nous présenterons :

- **les statistiques permettant d'apprécier l'adéquation du modèle aux données**
- **les tests d'hypothèse concernant les coefficients du modèle**
- **la construction d'intervalles de confiance pour les coefficients du modèle**
- **l'estimation de la moyenne**
- **l'analyse des résidus**

III.1 Le choix du modèle

Le plus souvent le choix de la loi de probabilité de la fonction de réponse découle naturellement de la nature du problème étudié.

On peut alors choisir comme fonction de lien la fonction de lien canonique associée à la loi de probabilité de la fonction de réponse étudiée.

III.1 Le choix du modèle

Il est toujours possible d'utiliser d'autres fonctions de lien. La procédure GENMOD de SAS propose par exemple les fonctions de liens suivantes :

- **Identité**
- **Logit**
- **Probit**
- **Puissance**
- **Logarithme**
- **Gompit (complémentaire log log)**

III.1 Le choix du modèle

On peut parfois essayer différentes fonction de réponses et retenir celle qui minimise la **déviance D** (définition dans le paragraphe III.2).

Notons que le statisticien peut enfin utiliser d'autres fonctions de réponse et d'autres fonctions de lien que celles proposées en standard par les logiciels.

III.2 Adéquation du modèle

Deux statistiques sont utiles pour juger de l'adéquation du modèle aux données :

- la déviance normalisée (scaled deviance)

- la statistique du khi-deux de Pearson

Pour mesurer l'adéquation du modèle étudié aux données, on construit tout d'abord un **modèle saturé.**

D* :déviance normalisée :

$$D^* = 2\log\lambda = 2\log\left(\frac{L(b_{max};y)}{L(b;y)}\right)$$
$$= 2\left[\ell(b_{max};y) - \ell(b;y)\right]$$

Lorsque le modèle étudié est exact, la déviance normalisée D* suit approximativement une loi du khi-deux à n-K degrés de liberté.

Loi de probabilité	Valeur de φ	Déviance
Normale	σ^2	$\sum_i (y_i - \mu_i)^2$
Binomiale(*)	1	$2\sum_i m_i \left\{ y_i \log(y/\hat{\mu}) + (1-y_i)\log((1-y_i)/(1-\hat{\mu}_i)) \right\}$
Poisson	1	$2\sum_i \left\{ y_i \log(y/\hat{\mu}) - (y_i - \hat{\mu}_i) \right\}$
Gamma	$\frac{1}{\nu}$	$2\sum_i \left\{ -\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i \right\}$
Gauss Inverse	σ^2	$2\sum_i \left\{ (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i^2 y_i \right\}$

La statistique du khi-deux de Pearson est définie par :

$$\chi^2 = \sum (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i)$$

Le khi-deux de Pearson normalisé est égal à: X^2/φ.

Comme la déviance normalisée, cette statistique est distribuée approximativement selon une loi du khi-deux à n-K degrés de liberté si le modèle étudié est exact.

Remarque 1: Utilisation de ces statistiques pour estimer le paramètre de dispersion φ inconnu.

On peut utilise la déviance D et la statistique khi-deux de Pearson pour estimer le paramètre de dispersion φ inconnu en identifiant ces statistiques normalisées à leur moyenne égale à n-K.

L'utilisation de la déviance donne : $\hat{\varphi} = \frac{D}{n-k}$

Celle du khi-deux de Pearson conduit à : $\hat{\varphi} = \frac{\chi^2}{n-k}$

Remarque 2: La sur-dispersion

La sur-dispersion est un phénomène qui concerne la modélisation de données selon une loi binomiale ou selon une loi de Poisson. Il y a sur-dispersion lorsque la déviance normalisée ou le khi-deux de Pearson normalisés sont nettement supérieurs à 1.

Pour remédier à ce problème on modifie la fonction $V(\mu)$ de ces lois en la multipliant par un paramètre de sur-dispersion φ :

loi binomiale : $V(\mu) = \varphi\ \mu\ (\ 1-\mu\ )$

loi de Poisson : $V(\mu) = \varphi\ \mu$

Remarque 2: La sur-dispersion

La sur-dispersion n'intervient pas au niveau de l'estimation de β.

Par contre la sur-dispersion a pour effet de rendre les résultats des tests de Wald et des tests LRT trop significatifs.

En effet quand on corrige le problème de la sur dispersion, et que l'on améliore donc le modèle, du fait que la matrice de variance-covariance de b (soit J^{-1})est multipliée par φ : les log vraisemblances et les valeurs des khi-deux de Wald sont-elles divisées par φ.

Des coefficients significativement différents de zéro avant correction de la sur-dispersion ne le sont plus forcément après correction.

Remarque 2: La sur-dispersion

L'approche proposée qui consiste à remplacer la fonction $a(\varphi) = 1$ par $a(\varphi) = \varphi$ semble conduire à des résultats tout à fait acceptables sur le plan pratique.

Néanmoins avec ces formules on n'est plus en présence de lois de probabilité. On appelle alors quasi-vraisemblance la pseudo fonction de vraisemblance associée.

III.3 Tests d'hypothèses concernant les coefficients du modèle

III.3.1 Le test de Wald

Sous l'hypothèse $\mathbf{H_0 : L'\beta = 0}$ la statistique :

$S = (L'b)'(L'J^{-1} L)^{-1}(Lb)$

suit une loi du khi-deux à r degrés de liberté, où r est le rang de L.

III.3.1 Le test de Wald

En notant $z = \hat{\beta} / ASE$ où ASE est l'erreur standard asymptotique de $\hat{\beta}$, il découle de ce résultat général que le test d'un coefficient du modèle se base sur la statistique z^2 distribuée sous l'hypothèse de nullité du coefficient selon une loi du **khi-deux à un degré de liberté** .

III.3.2 Le test LRT

Notons b* l'estimation du maximum de vraisemblance de β sous l'hypothèse H0 : L'β =0

$$l(b^*, y) = \max_{H_0: L'\beta = 0} l(\beta, y)$$

Sous l'hypothèse H0 la statistique :

S = 2 [l(b,y) – l(b*,y)]

suit approximativement une loi du Khi-deux à r degrés de liberté, où r est le rang de L

III.4 Intervalles de confiance pour les coefficients du modèle

III.4.1 L'intervalle de confiance de Wald

Du fait que b est approximativement distribuée selon une loi normale $N(\beta, J^{-1})$ on déduit l'intervalle de confiance de Wald de βj :

$$[\, b_j - z_{1-\alpha/2} s_j \,;\, b_j + z_{1-\alpha/2} s_j \,]$$

où $z_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ d'une loi normale réduite et s_j le j-ième terme de la diagonale de J^{-1}

III.4.2 L'intervalle de confiance basé sur le rapport des vraisemblances

La « fonction de vraisemblance profil » (profile likelihood function) de β_j est définie par :

$$l^*(\beta_j, y) = \max_{\tilde{\beta}} l(\beta, y)$$

où $\tilde{\beta}$ est le vecteur β avec le $j^{ème}$ élément fixé à β_j

III.4.2 L'intervalle de confiance basé sur le rapport des vraisemblances

Si β_j est la vraie valeur du paramètre, alors

$$2\left[l(b,y) - l^*(\beta_j, y)\right]$$

suit approximativement une loi de khi-deux à un degré de liberté.

On obtient un intervalle de confiance de β_j d'ordre 1-α en considérant l'ensemble des β_j tels que :

$$2\left[\ell(b,y) - \ell^*(\beta_j, y)\right] \leq \chi^2_{1-\alpha}(1) :$$

$$\left\{\beta_j : \ell^*(\beta_j, y) \geq \ell(b,y) - 0.5\chi^2_{1-\alpha}(1)\right\}$$

III.5 Estimation de la moyenne

La moyenne μ_i est estimée par $\mu_i = g^{-1}(x_i'b)$, x_i étant le vecteur contenant les valeurs des prédicteurs.

Pour obtenir l'intervalle de confiance de μ_i on utilise l'intervalle de confiance de $x_i'b$. La variance de $x_i'b$ étant estimée par $x_i'J^{-1}x_i$, l'intervalle de confiance de μ_i d'ordre $1-\alpha$ est donné par la formule :

$$g^{-1}\left(x_i'b \pm z_{1-\alpha/2}\sqrt{x_i'J^{-1}x_i}\right)$$

Dans le cas où l'argument de l'inverse de la fonction de lien est extérieur au domaine valide on ne peut construire d'intervalle de confiance.

III.6 Les résidus

Pour l'observation i le résidu observé $r_i = y_i - \hat{\mu}_i$ n'a qu'un intérêt restreint.

En effet si l'on prend l'exemple d'un modèle de Poisson, l'écart-type d'un effectif est $\sqrt{\hat{\mu}_i}$, de grosses différences tendent à apparaître si μ_i prend une valeur élevée.

III.6 Les résidus

Résidu de Pearson est un résidu « standardisé » de cette différence défini par :

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Résidu déviance :

où d_i représente la contribution de l'observation i à la déviance D et fournit un diagnostic individuel concernant la linéarité du modèle.

$$r_{D_i} = \sqrt{d_i}\,\text{signe}(y_i - \hat{\mu}_i)$$

le résidu de Pearson normalisé $r_{P_i}^{*} = \dfrac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1-h_i)}}$

le résidu déviance normalisé $r_{D_i}^{*} = \dfrac{\text{signe}(y_i - \hat{\mu}_i)d_i}{\sqrt{\varphi(1-h_i)}}$

le résidu vraisemblance normalisé

$$r_{G_i}^{*} = \text{signe}(y_i - \hat{\mu}_i)\sqrt{(1-h_i)r_{D_i}^{*2} + h_i r_{P_i}^{*2}}$$

h_i est le levier associé à l'observation i

On doit prêter attention aux résidus normalisés dépassant 2 en valeur absolue.

Composante aléatoire	Lien	Variables explicatives	Modèle
Normale	**Identité**	**Quantitatives**	Régression
Normale	**Identité**	**Qualitatives**	ANOVA
Normale	**Identité**	**Mixtes**	Analyse de la covariance
Binomiale	**Logit**	**Mixtes**	Régression logistique
Poisson	**Log**	**Mixtes**	Modèles log-linéaire
Multinomiale	**Logit généralisé**	**Mixtes**	Modèles à réponses multinomiales