Sciences & Technologie **A** – **N°35**, Juin 2012, 13-21.

ESTIMATION DU MODELE LINEAIRE GENERALISE ET APPLICATION

Malika CHIKHI[1], Michel CHAVANCE[2]

[1]Faculté des Sciences, Département de Mathématiques, Université Mentouri Constantine (Algérie).
[2]INSERM, unité 472, Avenue P. Vaillant Couturier, Villejuif (France).

Reçu le 05/06/2010 – Accepté le 21/02/2012

Résumé

Cet article présente le modèle linéaire généralisé englobant les techniques de modélisation telles que la régression linéaire, la régression logistique, la régression log linéaire et la régression de Poisson . On Commence par la présentation des modèles des lois exponentielles pour ensuite estimer les paramètres du modèle par la méthode du maximum de vraisemblance. Par la suite on teste les coefficients du modèle pour voir leurs significations et leurs intervalles de confiances, en utilisant le test de Wald qui porte sur la signification de la vraie valeur du paramètre basé sur l'estimation de l'échantillon.
Mots clés *:* modèles linéaires généralisés, composantes aléatoires, le lien binomial, familles exponentielles, méthode du maximum de vraisemblance, régression logistique.

Abstract

This paper presents the generalized linear model including modeling techniques such as linear regression, logistic regression, log-linear regression, Poisson regression, is starting with the presentation of exponential models. Then we estimate the model parameters by the method of maximum likelihood. Subsequently we test the model coefficients to see their meanings and their confidence intervals, using the Wald test which focuses on the significance of the true value of the parameter based on the sample estimate.

Keywords*: generalized linear models, random components , link binomial, exponential families, method of maximum likelihood logistic regression, log linear model*

ملخص.

تعرض هذه الوثيقة نموذج معمم بما في ذلك تقنيات النموذجة الخطية مثل التحليل الخطى، والانحدار اللوجستي، والانحدار الخطي اللوغاريتم، والانحدار بواشون، تبدأ في عرض نماذج الأسي. بعد ذلك تقدير معالم النموذج باستخدام طريقة من احتمال الحد الأقصى. اختبار بعد ذلك لمعرفة معاملات نموذج معانيها وفواصل الثقة، وذلك باستخدام اختبار والد الذي يركز على أهمية القيمة الحقيقية للمعلمة على أساس تقدير العينة

الكلمات المفتاحية: **نماذج خطية معممة، مكونات عشوائية،وتوقع، الرابط ذي الحدين، متعدد الحدود، والأسر الأسية، وطريقة الحد الأقصى احتمال الانحدار اللوجستي والأسر الأسية، وطريقة الحد الأقصى احتمال الانحدار اللوجستي**

Introduction :

En statistiques, le modèle linéaire généralisé (**GLM**) est une généralisation de la régression linéaire. Il permet d'étudier la liaison entre une variable réponse Y et un ensemble de variables explicatives ou prédicteurs$(X_1, X_2, \dots, X_n)$. (1,2)

Les modèles linéaires généralisés ont été formulés par (3,4) comme un moyen d'unifier les modèles statistiques y compris la régression linéaire, la régression logistique , la Régression log linéaire , la régression de Poisson. Ils proposent une méthode itérative dénommée méthode des moindres carrés ré-pondérés itérativement pour l'estimation du maximum de vraisemblance des paramètres du modèle. L'estimation du maximum de vraisemblance reste populaire et est la méthode par défaut dans de nombreux logiciels de calculs statistiques.

2-PRESENTATION DU MODELE LINEAIRE GENERALISE :

2-1-Les composantes du modèle linéaire généralisé :

Les modèles linéaires généralisés sont caractérisées par trois composantes : la composante aléatoire, le prédicteurs linéaire ou composante déterministe et la fonction lien.

- **La composante aléatoire** est définie par la distribution de probabilité de La variable réponse y. Elle peut être choisie dans la famille exponentielle à laquelle appartiennent les lois normales, binomiales, de Poisson, gamma, etc... Une propriété de ces lois est que pour chacune d'elle, il existe une relation spécifique entre l'espérance $E(Y) = \mu$ et la variance $Var(Y) = a(\phi)V(\mu)$, souvent $a(\phi) = \phi$

- **La composante déterministe** est définie par la fonction linéaire des variables explicatives, utilisées comme prédicteurs dans le modèle. Dans un modèle linéaire généralisé, l'espérance mathématique de Y, notée μ, varie en fonction des valeurs des variables explicatives. Le prédicteur linéaire est exprimé sous forme d'une combinaison linéaire. $\eta = \beta_0 + \beta_1 X_1 + \cdots \dots \dots \dots + \beta_n X_n$.
La fonction lien exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire. Elle spécifie comment l'espérance mathématique de Y , notée μ, est liée au prédicteur linéaire construit à partir des variables explicatives.

2-2-definition du modèle linéaire généralisé :

1- La composante aléatoire

Elle est définie par la distribution de probabilité de La variable réponse.

Soit $Y_1, Y_2, \dots, Y_n$ des variables d'un n échantillon aléatoire de la variable réponse$Y$, ces variables étant supposées indépendantes admettant des distributions issues d'une famille exponentielle.
Chaque observation Y_i admet une fonction de densité de la forme :
$f(y_i, \theta_i, \phi) = exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$ [1]
Le paramètre θ_i est appelé **paramètre naturel** de la famille exponentielle. La fonction $a(\phi)$ a la forme de $a(\phi) = \frac{\phi}{\omega_i}$ ou les poids ω_i sont connus, ϕ est appelé **paramètre de dispersion.**

Remarques

L'expression précédente représente la forme la plus générale des modèles linéaires généralisés. Elle englobe l'ensemble des lois usuelles utilisant un ou deux paramètres, tels que la loi normale, l'inverse de la loi normale, la loi gamma, la loi Poisson et la loi binomiale.
Dans lequel ϕ est un paramètre de nuisance. Si ϕ est une constante connue (en générale elle est égale à 1) l'expression [1] se met sous la forme canonique suivante :

$$f(y_i; \theta) = a(\theta_i)b(y_i)exp\ [y_i\, Q(\theta_i)] \quad [2]$$

En posant :

$$Q(\theta_i) = \left[\theta_i / a(\phi)\right],\ a(\theta_i) = exp\left[-b(\theta_i) / a(\phi)\right],$$

$$b(y_i) = exp\,[c(y_i, \phi)].$$

- Le terme $Q(\theta)$ est appelé le paramètre naturel de la distribution. Tout autre paramètre de la distribution est considéré comme un paramètre de nuisance(1,2).
- Cette famille [2] comprend de nombreuses distributions importantes telles-que la loi de Poisson et la loi binomiale.
La valeur du paramètre θ_i dépend des valeurs des variables explicatives.

2- La composante déterministe :

La composante déterministe du modèle se rapporte à un vecteur d'un ensemble de variables explicatives $\eta = \eta_1, \dots \eta_n$ par un modèle linéaire

$$\eta = X\beta$$

La matrice X se compose de n valeurs des variables explicatives, β est le vecteur des paramètres du modèle, le vecteur η est appelé prédicteur linéaire. La fonction lien spécifie comment l'espérance mathématique de Y , notée μ, est liée au prédicteur linéaire construit à partir des variables explicatives.

3- la fonction lien :

La fonction lien exprime une relation fonctionnelle entre la composante déterministe et la composante aléatoire.

Soit $\mu = E(y)$ on pose $\eta = g(\mu)$.

Ou g est appelée fonction lien, c'est une fonction différenciable et monotone.

Donc On peut modéliser l'espérance μ directement comme dans la régression linéaire, ou modéliser une fonction monotone $g(\mu)$ de l'espérance. On a alors :

$$g(\mu) = X'\beta$$

La fonction lien qui associe la moyenne μ au paramètre naturel est appelée fonction de lien canonique. 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 Poisson le paramètre canonique est le logarithme de l'espérance. Pour la distribution binomiale le paramètre canonique est la probabilité de succès. [3,4].Le tableau (1) résume les différents Types de modèles couverts par le modèle linéaire généralisé.

Composante aléatoire	Lien	Nature des variables de la composante déterministe	Modèle
Normale	Identité	Quantitatives	Régression
Normale	Identité	Qualitatives	Analyse du var
Normale	Identité	Mixtes	Analyse du cov
Binomiale	Logit	Mixtes	Régression logistique
Poisson	log	Mixtes	Modèles log − linéaires
Multinomiale	Logit généralisé	Mixtes	Modèles à réponses multinomiales

Tableau 1. Récapitulatif des principaux modèles

Les lois de probabilités telles que la loi normale, la loi de poisson, la loi binomiale, la loi gamma et la loi de Gauss inverse appartiennent à la famille des modèles linéaires généralisés ,elles sont définies ci-dessous.

2-2-1- La distribution de Poisson :

Soit y_i désigne l'effectif de l'iem cellule distribué selon une loi de Poisson de paramètres$E(y_i) = \mu_i$ et $VAR(y_i) = \mu_i$. Sa distribution de probabilité est définie par :

$$f(y_i;\mu_i) = e^{-\mu_i}\frac{{\mu_i}^{y_i}}{y_i!}$$

$$\log f(y_i;\mu_i) = (y_i log\mu_i - \mu_i - logy_i!)$$

La forme exponentielle[2] de La distribution de Poisson est définie par:

$f(y_i;\mu_i) = exp(y_i log\mu_i - \mu_i - logy_i!)$
$f(y_i;\theta) = a(\theta_i)b(y_i)exp\ [y_i\,Q(\theta_i)]$ [2]
$f(y_i;\theta) = exp(-\mu_i)\ \frac{1}{y_i!}exp\ [y_i\,log\mu_i]$ [2]

Les composantes du modèle sont donnés par :
$\mathbf{Q}(\boldsymbol{\theta_i}) = log\mu_i$, $\boldsymbol{a}(\boldsymbol{\theta_i}) = exp(-\mu_i)$, $\boldsymbol{b}(\boldsymbol{y_i}) = \frac{1}{y_i!}$.
La fonction de lien canonique est $\boldsymbol{g}(\mu_i) = log\mu_i$, elle permet de modéliser le logarithme de l'espérance, le modèle utilisant ce lien est le modèle $\boldsymbol{log - lin\acute{e}aire}$.

2-2-2- La distribution binomiale :

Notons $Y_1, Y_2, \dots, Y_n$ des variables d'un échantillon aléatoire de taille n de la variable de réponse$Y$, ces variables étant supposées indépendantes. Chaque variable Y_i est binaire (succès-échec), ou de façon plus générale Y_i peut être le nombre de succès au cours d'un certain nombre d'essais. On supposera alors que la composante aléatoire est distribuée selon une loi binomiale de paramètres $E(y_i) = \pi_i$ et $VAR(y_i) = \pi_i(1-\pi_i)$. Sa distribution de probabilité est définie par :

$f(y_i,\pi_i) = C_n^{y_i}{\pi_i}^{y_i}(1-\pi_i)^{n-y_i}$
$logf(y_i,\pi_i) = logC_n^{y_i} + y_i log\pi_i + (n-y_i)\log(1-\pi_i)$
$logf(y_i,\pi_i) = \{y_i log\frac{\pi_i}{1-\pi_i} + n\log(1-\pi_i) + logC_n^{y_i}\}$

La forme exponentielle [2],de La distribution binomiale est donnée par:

$f(y_i,\pi_i) = (1-\pi_i)^n C_n^{y_i}\ exp\left[y_i log\frac{\pi_i}{1-\pi_i}\right]$
$f(y_i;\theta) = a(\theta_i)b(y_i)exp\ [y_i\,Q(\theta_i)]$ [2]

Les paramètres sont : $\mathbf{Q}(\boldsymbol{\theta_i}) = log\frac{\pi_i}{1-\pi_i}$
, $\boldsymbol{a}(\boldsymbol{\theta_i}) = (1-\pi_i)^{\boldsymbol{n}}$, $\boldsymbol{b}(\boldsymbol{y_i}) = C_n^{y_i}$

La fonction lien canonique est $\boldsymbol{g}(\pi_i) = \boldsymbol{log}\frac{\pi_i}{1-\pi_i}$, elle est appelée fonction $logit$de π.

2-2-3- La distribution normale :

Soit $Y_1 \dots Y_n$ des variables d'un n échantillon aléatoire gaussien de la variable$Y$, ces variables sont supposées indépendantes. Alors la composante aléatoire est distribuée selon une distribution normale de paramètres(μ, σ^2), les fonctions de densités de ces variables s'écrivent.

$$f(y_i, \mu_i, {\sigma_i}^2) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-1}{2{\sigma_i}^2}(y_i-\mu_i)^2}$$

$$\log f(y_i, \mu_i, {\sigma_i}^2) = log\,\frac{1}{\sigma_i\sqrt{2\pi}} - \frac{1}{2{\sigma_i}^2}(y_i-\mu_i)^2$$

$$\log f(y_i, \mu_i, {\sigma_i}^2) = exp[-\frac{1}{2{\sigma_i}^2}{y_i}^2 + \frac{2}{2{\sigma_i}^2}y_i\mu_i - \frac{1}{2{\sigma_i}^2}{\mu_i}^2 - log\sigma_i\sqrt{2\pi}]$$

La forme exponentielle de la loi Normale selon la forme du GLM[**2**] est:

$$\log f(y_i, \mu_i, {\sigma_i}^2) = exp\left\{\left[\frac{y_i\mu_i}{{\sigma_i}^2}\right] + \left[-\frac{{\mu_i}^2}{2{\sigma_i}^2}\right] + \left[-\frac{{y_i}^2}{2{\sigma_i}^2} - \frac{1}{2}log[2\pi{\sigma_i}^2]\right]\right\}$$

Les composantes du modèle sont:

$\mathbf{Q}(\boldsymbol{\theta_i}) = \mu_i$, $\boldsymbol{a}(\boldsymbol{\theta_i}) = exp\left[-\frac{\mu^2}{2}\right]$, $\boldsymbol{b}(\boldsymbol{y_i}) = \exp\{-\frac{{y_i}^2}{2} - \frac{1}{2}log[2\pi{\sigma_i}^2]\}$, $\boldsymbol{\phi} = {\sigma_i}^2$.

La famille gaussienne se met sous la forme canonique [2] c'est une famille exponentielle de paramètre de dispersion $\phi = {\sigma_i}^2$ et de paramètre naturel $\boldsymbol{\theta_i} = E(y_i) = \mu_i$ Donc la fonction de lien canonique est La fonction identité $g(\mu_i) = \mu_i$.
Le tableau suivant indique les composantes de la famille exponentielle pour des lois usuelles.

Distribution	$\theta(\mu)$	$b(\theta)$	$a(\phi)$
Normale$N(\mu, \sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2
Binomiale$B(1, \mu)$	$log\,\frac{\mu}{1-\mu}$	$\log(1 + e^{\theta})$	1
Poisson$P(\mu)$	$log\mu$	e^{θ}	1
Gamma$G(\mu, v)$	$\frac{-1}{\mu}$	$-\log(-\theta)$	1
Gauss inverse $IG(\mu, \sigma^2)$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2

Tableau 2. Tableau récapitulatif des composantes des lois de la famille exponentielle

3-Principe d'estimation d'un modèle linéaire généralisé :

3-1- La méthode des moindres carrées :

Les paramètres du modèle linéaire sont classiquement estimés par la méthode des moindres carrées qui vise à minimiser la somme des carrés des écarts entre les valeurs observées Y_i et les valeurs prédites $\mu_i = \sum_{j=0}^{p}\beta_j X_{ij}$ [3].

Donc la somme des carrées des écarts entre les réponses observées et prédites est une fonction quadratique des paramètres inconnus :

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=0}^{p}\beta_j X_{ij}\right]^2.$$

3-2- La méthode du maximum de vraisemblance

L'estimation des paramètres β_j est calculée par la maximisation du log de vraisemblance du modèle linéaire généralisé. Cette estimation s'applique à toutes les lois de distributions appartenant à la famille exponentielle de la forme[2]

3-2-1-Estimation des coefficients par la méthode du maximum de vraisemblance :

Soit Y une variable qui obéit à une loi de distribution : $f(y_i; \theta_{i,} \Phi)$ .A partir d'un certain nombre d'observations de Y,$(Y_1, Y_2, \dots\dots, Y_n)$, on détermine les valeurs inconnues des paramètresθ_i . La méthode du maximum de vraisemblance postule que cette valeur de θ_i devrait être celle qui maximise la probabilité et d'obtenir les valeurs observées de Y.La procédure d'estimation par la méthode du maximum de vraisemblance (1,2) de la première expression du modèle linéaire généralisé est définie par :

La Fonction de vraisemblance du MLG est exprimée par :

$$f(y_i; \theta_i, \phi) = exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$$

En général Il est plus pratique d'opérer sur la transformation logarithmique de $f(y_i; \theta_i, \phi)$ qui s'exprime comme une somme de fonctions de θ, plutôt qu'un produit de fonctions comme c'est le cas pour $f(y_i; \theta_i, \phi)$). Que l'on désigne par $logf(y_i; \theta_i, \phi)$ et on la note $l(y_i; \theta_{i,} \Phi)$.

Le log de vraisemblance de l'ième observation est exprimée par:

$$l(y_i; \theta_{i,} \Phi) = \{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$$

La Maximisation du log de vraisemblance (1,2) est en fonction des premières et deuxièmes dérivées. alors on l'applique aux résultats de la vraisemblance :

$$\frac{\partial l}{\partial \theta_i} = \{[y_i - b'(\theta_i)]/a(\phi)\}$$

$$\frac{\partial^2 l}{\partial {\theta_i}^2} = -\left. b''(\theta_i) \right/ a(\phi)$$

Les conditions de régularités des lois appartenant aux familles exponentielles (1,2) sont vérifiées et permettent d'écrire :

$$E\left(\frac{\partial l}{\partial \theta_i}\right) = 0 \;\; et \;\; -E\left(\frac{\partial^2 l}{\partial {\theta_i}^2}\right) = E\left(\frac{\partial l}{\partial \theta_i}\right)^2$$

Alors

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

Et comme

$$\boldsymbol{E\left\{\left. b''(\theta_i) \right/ a(\phi)\right\} = E\{[\;\; Y_i - b'(\theta_i)]/a(\phi)^2 \;\;]\} =}$$

$$\boldsymbol{Var(Y_i)/a^2\,(\phi)}$$

Donc

$$\boldsymbol{Var(Y_i) = \; b''(\theta_i) a(\phi)}$$

Ainsi ϕ est appelé paramètre de dispersion lorsque, μ est la fonction d'identité.

Équations de vraisemblance

Soit X La matrice qui se compose de p observations des variables explicatives, β un vecteur de p paramètres du modèle et η le prédicteur linéaire à n composantes définie par .

$$\eta = X\,\beta$$

La fonction lien est supposée être monotone et différenciable telle que :

$$\eta_i = g(\mu_i)$$

La fonction lien canonique est donnée par :

$$g(\mu_i) = \boldsymbol{\theta_i}$$

Soit n observations indépendantes et $\boldsymbol{\theta}$ dépend de β ,le log de vraisemblance est définie par :

$$L(\beta) = \sum_{i=1}^{n} log\, f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} l(y_i; \theta_i, \phi\;)$$

Pour obtenir les équations de vraisemblance, on Calcule :

$$\frac{\partial\; l_i}{\partial \boldsymbol{\beta_j}} = \frac{\partial l_i}{\partial \boldsymbol{\theta_i}} \frac{\partial \boldsymbol{\theta_i}}{\partial \boldsymbol{\mu_i}} \frac{\partial \mu_i}{\partial \boldsymbol{\eta_i}} \frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{\beta_j}}$$

Comme :

$$\frac{\partial l_i}{\partial \boldsymbol{\theta_i}} = \{[\boldsymbol{y_i - b'(\theta_i)}]/\boldsymbol{a(\phi)}\} = \{[\boldsymbol{y_i - \mu_i}]/\boldsymbol{a(\phi)}\}$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{\theta_i}} = \frac{\boldsymbol{Var(Y_i)}}{\boldsymbol{a(\phi)}} = \; \boldsymbol{b''(\theta_i)}$$

$\frac{\partial \eta_i}{\partial \boldsymbol{\beta_j}} = x_{ij}$, car $\eta_i = \sum_j \beta_j x_{ij}$

Puisque, $\frac{\partial \mu_i}{\partial \eta_i}$ dépends de la fonction lien $\eta_i = g(\mu_i)$ du modèle alors :

$$\frac{\partial\; l_i}{\partial \beta_j} = \frac{(y_i - \mu_i\;)}{a(\phi)} \times \frac{a(\phi)}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij}$$

Les équations de vraisemblance sont données par :

$\sum_{i=1}^{n} \frac{(y_i - \mu_i\;) x_{ij}}{Var(Y_i)} \times \frac{\partial \mu_i}{\partial \eta_i} = 0$ .Avec j=1,…, p

Ces équations sont non linéaires en β.pour les résoudre on utilise des méthodes itératives telles que la méthode de newton Raphson (ou l'on utilise le **Hessien**) ou la méthode des scores de fisher (on utilise la matrice d'information).

Les éléments de la matrice d'information sont données par $\{ E\left(\frac{\partial^2 L(\beta)}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_k}}\right) \}$, elle est notée par F et égale à :

$$\mathbf{F} = \boldsymbol{X'WX}$$

De terme général

$$F \;= E\left(\frac{\partial^2 L(\beta)}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_k}}\right) = -\sum_{i=1}^{n} \frac{x_{ik} x_{ij}}{Var(Y_i)} \times \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Ou W est la matrice diagonale, ces éléments sont définies par :

$$w_i = \frac{1}{Var(Y_i)} \times \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

La procédure d'estimation par la méthode du maximum de vraisemblance de la deuxième expression du modèle linéaire généralisé est, (plusieurs simplifications interviennent) :

$$\eta_i = \frac{\partial \mu_i}{\partial \boldsymbol{\theta_i}} = \sum_j \beta_j x_{ij}$$

$$\frac{\partial \mu_i}{\partial\; \eta_i} = \frac{\partial \mu_i}{\partial \boldsymbol{\theta_i}} = \frac{\partial b'(\boldsymbol{\theta_i})}{\partial \boldsymbol{\theta_i}} = b''(\boldsymbol{\theta_i})$$

Ainsi

$$\frac{\partial\; l_i}{\partial \beta_j} \;= \frac{(y_i - \mu_i\;)}{Var(Y_i)} \times b''(\boldsymbol{\theta_i}) \times x_{ij} = \frac{(y_i - \mu_i\;)}{a(\phi)} \times x_{ij}$$

Les termes $\left(\frac{\partial^2 L(\beta)}{\partial \boldsymbol{\beta_j} \partial \boldsymbol{\beta_k}}\right)$ ne dépendent plus de y_i alors on montre que le Hessien est égale à la matrice d'information et donc les méthodes de newton Raphson coïncident avec les méthodes de résolutions des scores de Fisher.

Si, en plus $\boldsymbol{a(\phi)}$ est constante pour les observations, alors les équations de vraisemblance s'écrivent :

$$\boldsymbol{X'y = X'\mu}$$

La valeur de la fonction en son maximum joue un rôle central dans la définition du test statistique du rapport de vraisemblance.

3-2-2-Exemples d'estimation des paramètres des lois usuelles par La méthode du maximum de vraisemblance

1-loi binomiale

Notons $Y_1, Y_2, \ldots, Y_n$ des variables d'un n échantillon aléatoire de la variable réponse$Y$, ces variables sont supposées indépendantes. Chaque variable Y_i est binaire (succès-échec), ou de façon plus générale Y_i peut être le nombre de succès au cours d'un certain nombre d'essais. On supposera alors que la composante aléatoire est distribuée selon une loi binomiale de paramètres $E(y_i) = \pi_i$

et $VAR(y_i) = \pi_i(1-\pi_i)$. Sa distribution de probabilité est définie par :

$$f(y_i,\pi_i) = C_n^{y_i}\pi_i{}^{y_i}(1-\pi_i)^{n-y_i}$$
$$logf(y_i,\pi_i) = \{y_i log\frac{\pi_i}{1-\pi_i} + n\log(1-\pi_i) + logC_n^{y_i}\}$$
$$f(y_i;\theta_i,\phi) = exp\{y_i log\frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)n + logC_n^{y_i}\}$$
$$f(y_i;\theta_i,\phi) = exp\{[y_i\theta_i + exp(\theta_i)]/1 - logC_n^{y_i}\}$$

Les composantes du modèle GLM sont donnés par :

$\theta_i = log\frac{\pi_i}{1-\pi_i}$, $b(\theta_i) = n\log(1-\pi_i)$, $c(y_i,\phi) = logC_n^{y_i}$,$a(\phi) = 1$.

Le paramètre naturel est $\theta_i = log\frac{\pi_i}{1-\pi_i}$, *la moyenne est la variance sont :*

$$E(Y_i) = b'(\theta_i) = \mu_i = \pi_i = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$Var(Y_i) = b''(\theta_i) = n\pi_i(1-\pi_i) = n\mu_i(1-\mu_i)$$

La fonction de lien canonique est $\boldsymbol{g(\pi_i) = log\frac{\pi_i}{1-\pi_i}}$, elle modélise le logarithme du rapport des cotes. Elle est appelée fonction logit, le modèle utilisant ce lien est appelé modèle logistique

2- Loi de Poisson

Considérons le cas général, ou l'on observe y_i décès pour n personnes-temps cumulées dans une population ouverte. On suppose que le nombre de décès obéit à une loi de poisson de paramètre μ_i . On veut déterminer la valeur deμ_i qui maximise la fonction $L(\mu_i)$.

$$f(y_i,\mu_i) = e^{-\mu_i}\frac{\mu_i{}^{y_i,}}{y_i!}$$
$$\log f(y_i,\mu_i) = \{y_i log\mu_i - \mu_i - logy_i!\}$$

La forme exponentielle de La distribution de Poisson est définie ainsi :

$$f(y_i,\mu_i) = exp\{y_i log\mu_i - \mu_i - logy_i!\}$$
$$f(y_i;\theta_i,\phi) = exp\{[y_i\theta_i - exp(\theta_i)]/1 - logy_i!\} \quad [1]$$

Les composantes du modèle sont donnés par :

$\theta_i = \log(\mu_i)$, $b(\theta_i) = exp(-\theta_i)$, $c(y_i,\phi) = -logy_i!$,$a(\phi) = 1$.

Le paramètre naturel est $\theta_i = \log(\mu_i)$ *, la moyenne est la variance sont :*

$$E(Y_i) = b'(\theta_i) = exp(\theta_i) = \mu_i$$
$$Var(Y_i) = b''(\theta_i) = exp(\theta_i) = \mu_i$$

Ainsi $\boldsymbol{g(\mu_i)}$ est égale au paramètre naturel θ_i alors g est le log de la fonction, La fonction de lien canonique est $\eta_i = log\mu_i$, elle permet de modéliser le logarithme de l'espérance, le modèle utilisant ce lien est le modèle $\boldsymbol{log - linéaire}$.

3- La distribution normale :

Soit $Y_1 \ldots Y_n$ des variables d'un n échantillon aléatoire gaussien de la variable$Y$, ces variables sont supposées indépendantes. Alors la composante aléatoire est distribuée selon une distribution normale de paramètres(μ,σ^2). On veut déterminer les valeurs de μ *et* σ^2 qui maximise la fonction de vraisemblance. Cette fonction est construite à partir de la fonction de densité appliquée à chacune des valeurs observées.

$$f(y_i,\mu_i,{\sigma_i}^2) = \frac{1}{\sigma_i\sqrt{2\pi}}e^{\frac{-1}{2{\sigma_i}^2}(y_i-\mu_i)^2}$$

La forme exponentielle de La distribution la loi Normale est définie par :

$$\log f(y_i,\mu_i,{\sigma_i}^2) = \exp\left\{\left[\frac{2y_i\mu_i}{2{\sigma_i}^2}\right] + \left[-\frac{{\mu_i}^2}{2{\sigma_i}^2}\right] + \left[-\frac{1{y_i}^2}{2{\sigma_i}^2} - \frac{1}{2}log(2\pi{\sigma_i}^2)\right]\right\}$$
$$f(y_i;\theta_i,\phi) = exp\left\{[y_i\theta_i - exp(\theta_i)]/2{\sigma_i}^2 + (-\frac{1{y_i}^2}{2{\sigma_i}^2} - \frac{1}{2}log[2\pi{\sigma_i}^2])\right\}$$

Les composantes du modèle sont:

$\theta_i = \mu_i$, $b(\theta_i) = \frac{{\mu_i}^2}{2}$, $c(y_i,\phi) = (-\frac{1{y_i}^2}{2{\sigma_i}^2} - \frac{1}{2}log[2\pi{\sigma_i}^2])$,$a(\phi) = {\sigma_i}^2$.

Le paramètre naturel est $\theta_i = \mu_i$ *la moyenne est la variance sont :*

$$E(Y_i) = b'(\theta_i) = \theta_i = \mu_i = \frac{\sum_i y_i}{n}$$
$$Var(Y_i) = b''(\theta_i) = \sigma^2 = \frac{1}{n}\sum_i (y_i-\mu)^2$$

La famille gaussienne se met sous la forme canonique $\eta_i = \mu_i$ c'est une famille exponentielle de paramètre de dispersion $\phi = {\sigma_i}^2$ et de paramètre naturel $\boldsymbol{\theta_i} = E(y_i) = \mu_i$ Donc la fonction de lien canonique est La fonction identité, $\boldsymbol{g(\mu_i)} = \mu_i$ le modèle utilisant ce lien est appelé modèle linéaire.

Le tableau suivant indique l'espérance et la variance des probabilités usuelles appartenant à la famille exponentielle en utilisant les fonctions de scores.

Distribution	$\mu = E(Y) = b'(\theta)$	$V(Y) = b''(\theta)a(\phi)$
Normale$N(\mu,\sigma^2)$	θ	σ^2
Binomiale$B(1,\mu)$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$
Poisson$P(\mu)$	e^θ	μ
Gamma$G(\mu,v)$	$\frac{-1}{\theta}$	$\frac{\mu^2}{2}$
Gauss inverse $IG(\mu,\sigma^2)$	$\frac{1}{\sqrt{-2\theta}}$	$\mu^3\sigma^2$

Tableau-3- l'espérance et la variance des lois usuelles.

4-Construction pratique d'un modèle linéaire généralisé

La Construction pratique d'un modèle linéaire généralisé (7)et l'interprétation des résultats est basée sur :

- Le choix du modèle.
- 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'analyse des déviances et des résidus.

4-1- 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.

Il est toujours possible d'utiliser d'autre fonction de lien par exemple : l'identité, le Logit, Probit, Puissance et Logarithme.

4-2- Adéquation du modèle :

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

- La déviance normalisée.
- La statistique du Khi-deux de Pearson.

Un modèle linéaire généralisé est défini par la loi de probabilité $f(y;\theta)$ de la réponse Y et la nature de la fonction de lien g reliant l'espérance μ et Y aux variables explicatives

$$X_1, X_2, \dots, X_k \ , \ g(\mu_i) = x'_i\beta.$$

On note b l'estimation du maximum de vraisemblance de β. Pour mesurer l'adéquation du modèle étudié aux données, on construit tout d'abord un modèle saturé, modèle basé sur la même loi de probabilité et la même fonction de lien, mais contenant autant de variables explicatives indépendantes que de données : ce modèle permet de reconstruire parfaitement les données. On note b_{max} l'estimation du vecteur des paramètresβ pour ce modèle.[8]

4-2-1- La déviance :

Dans le modèle linéaire, la décomposition de la somme des carrées en somme des carrées expliquée par le modèle, SCL, et somme des carrées résiduelle, SCR, fournit une mesure d'adéquation classique, le coefficient de détermination $R^2 = \frac{SCL}{SCL+SCR}$

Il fournit une mesure pratique d'adéquation parce qu'il est compris entre 0 et 1 et que sa borne supérieure peut être atteinte à condition d'utiliser un nombre suffisamment élevé de coefficients (modèle dit saturé). (5,6)

Ce coefficient, dont le calcul ne fait intervenir que les valeurs observées et prédites, pourrait être calculé pour n'importe quel modèle linéaire généralisé, cependant il ne bénéficierait pas des mêmes avantages. Il suffit d'imaginer le cas de données binaires comme dans une enquête épidémiologique ou tous les sujets de l'échantillon ont des profils de risque différents, par exemple parce que certains risques sont mesurés sur une échelle continue. Les observations ne peuvent prendre n'importe quelle valeur de l'intervalle [0,1].

On préfère utiliser un coefficient appelé déviance $D = 2(L_{max} - L_0)$, qui représente le double du logarithme d'un rapport de vraisemblance. Il est toujours possible de construire un modèle comprend autant de paramètres que d'observations. Ce modèle comprend autant de paramètres que d'observations distinctes (modèle dit saturé) et puisqu'il ne permet pas de les résumer. Cependant on peut le considérer comme une référence : aucune modèle ne s'ajute mieux et sa log-vraisemblance L_{max} est la plus élevée parmi tous ceux de la famille considérée. La log-vraisemblance l_0 du modèle dont on veut mesurer l'adéquation est inférieure, mais si la différence n'est pas trop élevée on pourra affirmer qu'il s'ajuste bien aux données. Sous l'hypothèse que le modèle considéré est correct, on montre que$2(l_{max} - l_0)$ suit asymptotiquement une distribution de chi-2 dont le nombre de degrés de liberté est égal à la différence entre le nombre de paramètres des 2 modèles. On peut remarquer qu'il est strictement équivalent de comparer deux modèles emboités en calculant la différence de leurs déviances ou le double de la différence de leur log-vraisemblances(7).

Dans le modèle linéaire qui suppose des observations normales de variance constante, la déviance $\frac{\sum(yi-\widehat{yi})^2}{s^2}$ est apparentée a la somme des carrés résiduelles SCR= $\sum_{(yi-\widehat{yi})}2$

Dans un modèle logistique, qui suppose des observations $\boldsymbol{y_i}$ distribuées selon des lois binomiales B (n_i-π_i), la vraisemblance du modèle saturé qui prédit les probabilités π_i est v_0=$\prod \pi_i^{y_i}(1-\pi)^{n_i - y_i}$, celle du modèle saturé qui prédit les fréquences observées f_i=$y_{i\,/n_i}$ est V_{max}=$\prod f_i^{y_i}(1-f_i)^{n_i - y_i}$ et la déviance vaut donc :

$$D= 2\sum[y_i \ log\frac{f_i}{n_i} +(n_i\text{-}y_i)\log\frac{1-f_i}{1-\pi_i}]$$

$$D=2\sum O_{ij}\log\frac{O_{ij}}{E_{ij}}$$

Ou j=0,1 indice les deux modalités de réponse possibles. De façon générale, le tableau suivant donne la forme de la déviance pour quatre des principales distributions de la famille exponentielle.

Distribution	Déviance
Normale	$\sum(y-\widehat{\mu})^2$
Poisson	$\sum[ylog\frac{y}{\hat{\mu}}-(y-\widehat{\mu})]$
Binomiale	$2\left[\left[\sum ylog\frac{y}{\hat{\mu}}+(n-y)log\left(\frac{n-y}{n-\hat{\mu}}\right)\right]\right]$
Gamma	$2\sum(-log\frac{y}{\hat{\mu}}+\frac{y-\hat{\mu}}{\hat{\mu}}]$

Tableau-4- Déviance des différentes distributions de la famille exponentielle

Notons que les seconds termes dans la déviance des lois de poisson ou gamma sont en général nuls, et donc omis : c'est une conséquence des équations du score quand l'ordonnées à l'origine est estimée.

Exemples de modèle emboitée :

Pour décrire le risque de la maladie $\boldsymbol{M}$ en fonction des $\boldsymbol{p}$ variables $\boldsymbol{X\{X_1, X_2, \dots, X_p\}}$ observées sur $\boldsymbol{n}$ individus, considérons les modèles suivants :

$$\boldsymbol{g_0 = \alpha}$$

$$\boldsymbol{g_1 = \alpha + \beta_1 X_1}$$

$$\boldsymbol{g_2 = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$$

$$\boldsymbol{g_3 = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2}$$

$$\boldsymbol{gs = modèle\ saturé}$$

Remarquons que tous ces modèles sont emboîtés, C'est-à-dire qu'en termes de coefficients, $\boldsymbol{g_s}$ contient$\boldsymbol{g_3}$, qui contient$\boldsymbol{g_2}$, qui contient$\boldsymbol{g_1}$, qui contient $\boldsymbol{g_0}$

Pour la comparaison des modèles $\boldsymbol{g_1}$ $\boldsymbol{et}$ $\boldsymbol{g_3}$ par exemple, on a :

$$\begin{aligned} D_1 - D_3 &= [-2L_1 - (-2L_s)] - [-2L_3 - (-2L_s)] \\ &= -2[L1_1 - L_3] \\ &= -2log\,(V_1/V_3) \end{aligned}$$

Cette statistique obéit à un test du khi – carré à 3 degré de liberté.

Le tableau suivant donne l'analyse de la déviance

Modèle	déviation	DDL	test du RV
Modèle 1	L_0-L_1	1	-2[L_0-L_1]
Modèle 2	L_1-L_2	2	-2[L_1-L_2]
Modèle 3	L_2-L_3	1	-2[L_2-L_3]
Résidu	L_3-L_s	n-1-4	-2[L_3-L_s]
Total	L_0- L_s	n-1	-2[L_0- L_s]

Où $\boldsymbol{n}$ représente le nombre de modalités. Nous rappelons que s'il y a $\boldsymbol{p}$ variables dichotomiques indépendantes, alors il y a 2 modalités différentes.

4-2-2- Les résidus :

L'observation des résidus permet d'identifier les observations qui s'ajustent mal au modèle. Les résidus bruts représentent les différences $Y_i - \widehat{Y}_i$ entre réponse observée et prédite. Leur variance vaut $\sigma^2(1-h_{ii})$. On peut leur préférer les résidus de Pearson $\frac{Y_i-\widehat{Y}_i}{s}$, de déviance $(1-h_{ii})$, dont les carrées représente la contribution de l'observation au X^2 de Pearson, ou les résidus studentisés $\frac{Y_i-\widehat{Y}_i}{s\sqrt{1-h_{ii}}}$ de variance unité.[5,6]

Ces trois types de résidus existent sous une forme dite de prédiction qui utilise la régression effectuée sur l'ensemble de l'échantillon moins le point considéré. En appelant X_i la matrice des variables explicatives sans la ligne x_i , β_i et s_i les estimations correspondantes des coefficients β et l'écart type résiduel σ, $\widehat{Y}_i$ la valeur prédite par x_i et β_i ,on obtient les résidus de prédiction $Y_i - \widehat{Y}_i$, $\frac{Y_i-\widehat{Y}_i}{s_i\sqrt{1-h_{ii}}}$ et $\frac{Y_i-\widehat{Y}_i}{s_i\sqrt{1-h_{ii}}}$,noter que dans cette formule, le bras de levier h_{ii}, qui ne dépend que des covariables X, et calculé sur la matrice complète. Il est plus pertinent de comparer à une distribution de Student (ou pratiquement à une distribution normale) la valeur des résidus studentisés de prédiction que celle des résidus de Student simples car une observation suffisamment excentrée à la fois du point de vue des Co variables(bras de levier élevé) et du point de vue de la variable réponse Y,a un résidus nul.[3]

CONCLUSION

Le développement de la théorie des modèles linéaires généralisés au cours des vingt dernières années , a permis d'unifier différents types de modèles dédiés aussi bien aux données quantitatives qu' aux données qualitatives . La composante aléatoire entrant dans la définition d'un modèle linéaire généralisé doit appartenir à la famille des lois exponentielles, ce qui ne constitue pas une restriction importante. En effet, comme nous l'avons vu, la plupart des lois de probabilités usuelles Poisson, Binomiale, Normale, Gamma en font partie et on peut donc en déduire de nombreuses applications en modélisation statistique. Notons que l'on peut également définir des extensions multivariées des modèles linéaires généralisés dans le cas où la variable de réponse est un vecteur à plusieurs composantes. Un trait intéressant de la famille des modèles linéaire généralises réside dans le fait que l'algorithme est le même pour tous les modèles, quelles que soient le choix de la loi de probabilité de la réponse et de la fonction de lien. Ainsi les procédures offertes dans les logiciels permettent de construire de nombreuses options de modélisation d'une réponse binaires et polytomique.

BIBLIOGRAPHIE

[1]. Alan Agresti, « categorical data analysis », John Wiley ans Sons, Inc, 1991.

[2]. Alan. Agresti, « Analysis of Ordinal Categorial Data », John Wiley and Sons Inc., 2010.

[3]. M.C. Cullagh and J. Nelder, « Generalised Linear Models », London Chapman and Hall, 1983.

[4]. D. Mc Faden, « Regression Based Specification Tests for the Multinomial Logit Models », *Journal of Econometrics*, 34, 1987, pp.63-82.

[5]. Michel Chavance « Le modèle linéaire Généralise ». polycopié, Décembre, 2008.

[6]. Dobson, A.J.; Barnett, A.G. (2008) . Introduction aux modèles linéaire généralisés, troisième édition. Londres : Colporteur et Hall/CRC.

[7]. P.L. Gonzeles, « Modèles linéaire généralisé» , Modèles statistique pour données qualitatives , Droesbeke, Lejeune et Saporta Editeurs, Chapitre 5, pages 84-98, Technip, 2005..

[8]. Raymond H. Myers, Douglas C. Montgomery, G. Geoffrey Vining, Timothy J. Robinson, «Generalized Linear Models», Amazon France, 2012.