Introduction au modèle linéaire général

Résumé

Introductions au modèle linéaire général.

Retour au plan du cours. Travaux pratiques .

1 Introduction

L'objet de ce chapitre est d'introduire le cadre théorique global permettant de regrouper tous les modèles (linéaire gaussien, logit, log-linéaire) de ce cours et qui cherchent à exprimer l'espérance d'une variable réponse Y en fonction d'une combinaison linéaire des variables explicatives. Le *modèle linéaire généralisé* développé initialement en 1972 par Nelder et Wedderburn et dont on trouvera des exposés détaillés dans Nelder et Mc Cullagh (1983), Agresti (1990) ou Antoniadis et al. (1992), n'est ici qu'esquissé afin de définir les concepts communs à ces modèles : famille exponentielle, estimation par maximum de vraisemblance, tests, diagnostics, résidus. Il est mis en œuvre dans plusieurs logiciels dont GLIM, `glm` de Splus, `genmod` et `insight` de SAS.

2 Composantes des modèles

Les modèles catalogués dans la classe des modèles linéaires généralisés sont caractérisés par trois composantes.

2.1 Distribution

La *composante aléatoire* identifie la distribution de probabilités de la variable à expliquer. On suppose que l'échantillon statistique est constitué de n variables aléatoires $\{Y_i; i = 1, \ldots, n\}$ indépendantes admettant des distributions issues d'une *structure exponentielle*. Cela signifie que les lois de ces variables sont dominées par une même mesure dite de référence et que la famille de leurs densités par rapport à cette mesure se met sous la forme :

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right\}. \qquad (1)$$

Cette formulation inclut la plupart des lois usuelles comportant un ou deux paramètres : gaussienne, gaussienne inverse, gamma, Poisson, binomiale.... Le paramètre θ_i est appelé *paramètre naturel* de la famille exponentielle.

Attention, la mesure de référence change d'une structure exponentielle à l'autre, la mesure de Lebesgues pour une loi continue, une mesure discrète combinaison de masses de Dirac pour une loi discrète. Consulter Antoniadis et al. (1992) pour une présentation générale des structures exponentielles et des propriétés asymptotiques des estimateurs de leurs paramètres.

Pour certaines lois, la fonction u est de la forme :

$$u(\phi) = \frac{\phi}{\omega_i}$$

où les poids ω_i sont les poids connus des observations, fixés ici à 1 pour simplifier ; ϕ est appelé alors *paramètre de dispersion*, c'est un paramètre de nuisance intervenant, par exemple lorsque les variances des lois gaussiennes sont inconnues, mais égal à 1 pour les lois à un paramètre (Poisson, binomiale). L'expression de la structure exponentielle (1) se met alors sous la *forme canonique* en posant :

$$\begin{aligned} Q(\theta) &= \frac{\theta}{\phi}, \\ a(\theta) &= \exp\left\{-\frac{v(\theta)}{\phi}\right\}, \\ b(y) &= \exp\{w(y, \phi)\}, \end{aligned}$$

on obtient

$$f(y_i, \theta_i) = a(\theta_i)b(y_i)\exp\{y_iQ(\theta_i)\}. \qquad (2)$$

2.2 Prédicteur linéaire

Les observations planifiées des variables explicatives sont organisées dans la matrice $\mathbf{X}$ de planification d'expérience (design matrix). Soit $\boldsymbol{\beta}$ un vecteur de p paramètres, le prédicteur linéaire, *composante déterministe* du modèle, est le vecteur à n composantes :

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

2.3 Lien

La troisième composante exprime une *relation fonctionnelle* entre la composante aléatoire et le prédicteur linéaire. Soit $\{\mu_i = E(Y_i); i = 1, \ldots, n\}$, on pose

$$\eta_i = g(\mu_i) \quad i = 1, \ldots, n$$

où g, appelée *fonction lien*, est supposée monotone et différentiable. Ceci revient donc à écrire un modèle dans lequel une *fonction de la moyenne* appartient au sous-espace engendré par les variables explicatives :

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad i = 1, \ldots, n.$$

La fonction lien qui associe la moyenne μ_i au paramètre naturel est appelée *fonction lien canonique*. Dans ce cas,

$$g(\mu_i) = \theta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

2.4 Exemples

2.4.1 Loi gaussienne

Dans le cas d'un échantillon gaussien, les densités d'une famille de lois $\mathcal{N}(\mu_i, \sigma^2)$ s'écrit :

$$\begin{aligned} f(y_i, \mu_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\} \exp\left\{y_i \frac{\mu_i}{\sigma^2}\right\} \end{aligned}$$

En posant

$$\begin{aligned} Q(\theta_i) &= \frac{\theta_i}{\phi} = \frac{\mu_i}{\sigma^2} \\ a(\theta_i) &= \exp\left\{-\frac{1}{2}\frac{\mu_i^2}{\sigma^2}\right\} \\ b(y_i) &= \exp\left\{-\frac{1}{2}\frac{y_i^2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}. \end{aligned}$$

la famille gaussienne se met sous la forme canonique (2) qui en fait une famille exponentielle de paramètre de dispersion $\phi = \sigma^2$ et de paramètre naturel

$$\theta_i = E(Y_i) = \mu_i$$

et donc de fonction lien canonique, la fonction *identité*.

2.4.2 Loi de Bernouilli

Considérons n variables aléatoires binaires indépendantes Z_i de probabilité de succès π_i et donc d'espérance $E(Z_i) = \pi_i$. Les fonctions de densité de ces variables sont éléments de la famille :

$$f(z_i, \pi_i) = \pi_i^{z_i}(1 - \pi_i)^{1-z_i} = (1 - \pi_i) \exp\left\{z_i \ln \frac{\pi_i}{1 - \pi_i}\right\},$$

qui est la forme canonique d'une structure exponentielle de paramètre naturel

$$\theta_i = \ln \frac{\pi_i}{1 - \pi_i}.$$

Cette relation définit la fonction *logit* pour fonction lien canonique associée à ce modèle. La loi binomiale conduit à des résultats identiques en considérant les sommes de n_i (n_i connus) variables de Bernouilli.

2.4.3 Loi de Poisson

On considère n variables indépendantes Y_i de loi de Poisson de paramètre $\mu_i = E(Y_i)$. Les Y_i sont par exemple les effectifs d'une table de contingence. Ces variables admettent pour densités :

$$f(y_i, \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} = \exp\{-\mu_i\} \frac{1}{y_i!} \exp\{y_i \ln \mu_i\}$$

qui sont issues d'une structure exponentielle et, mises sous la forme canonique, de paramètre naturel

$$\theta_i = \ln \mu_i$$

définissant comme fonction lien canonique le *logarithme* pour ce modèle.

3 Estimation

L'estimation des paramètres β_j est calculée en maximisant la log-vraisemblance du modèle linéaire généralisé. Celle-ci s'exprime pour toute famille de distributions mise sous la forme (1) d'une structure exponentielle.

3.1 Expression des moments

Notons $\ell(\theta_i, \phi; y_i) = \ln f(y_i; \theta_i, \phi)$ la contribution de la $i$ème observation à la log-vraisemblance.

$$\ell(\theta_i, \phi; y_i) = [y_i\theta_i - v(\theta_i)]/u(\phi) + w(y_i, \phi).$$

L'étude du maximum de la log-vraisemblance nécessite la connaissance des dérivées :

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_i} &= [y_i - v'(\theta_i)]/u(\phi) \\ \frac{\partial^2 \ell}{\partial \theta_i^2} &= -v''(\theta_i)/u(\phi). \end{aligned}$$

Pour des lois issues de structures exponentielles, les conditions de régularité vérifiées permettent d'écrire :

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{et} \quad -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = E\left(\frac{\partial \ell}{\partial \theta}\right)^2.$$

Alors,

$$E(Y_i) = \mu_i = v'(\theta_i)$$

et comme

$$E\{v''(\theta_i)/u(\phi)\} = E\{[Y_i - v'(\theta_i)]/u(\phi)\}^2 = \text{Var}(Y_i)/u^2(\phi)$$

il vient donc :

$$\text{Var}(Y_i) = v''(\theta_i)u(\phi)\ ;$$

justifiant ainsi l'appellation de *paramètre de dispersion* pour ϕ lorsque u est la fonction identité.

3.2 Équations de vraisemblance

Considérons p variables explicatives dont les observations sont rangées dans la matrice de plan d'expérience $\mathbf{X}$, $\boldsymbol{\beta}$ un vecteur de p paramètres et le prédicteur linéaire à n composantes

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

La fonction lien g est supposée monotone différentiable telle que : $\eta_i = g(\mu_i)$;

c'est la fonction lien canonique si : $g(\mu_i) = \theta_i$.

Pour n observations supposées indépendantes et en tenant compte que $\boldsymbol{\theta}$ dépend de $\boldsymbol{\beta}$, la log-vraisemblance s'écrit :

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \ell(\theta_i, \phi; y_i).$$

Calculons

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Comme

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_i} &= [y_i - v'(\theta_i)]/u(\phi) = (y_i - \mu_i)/u(\phi), \\ \frac{\partial \mu_i}{\partial \theta_i} &= v''(\theta_i) = \text{Var}(Y_i)/u(\phi), \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \quad \text{car} \quad \eta_i = \mathbf{x}_i'\boldsymbol{\beta}, \\ \frac{\partial \mu_i}{\partial \eta_i} & \quad \text{dépend de la fonction lien} \quad \eta_i = g(\mu_i), \end{aligned}$$

Les équations de la vraisemblance sont :

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0 \quad j = 1, \ldots, p.$$

Ce sont des équations non-linéaires en $\boldsymbol{\beta}$ dont la résolution requiert des méthodes itératives dans lesquelles interviennent le Hessien (pour Newton-Raphson) ou la *matrice d'information* (pour les Scores de Fisher). La matrice

d'information est la matrice

$$\mathfrak{I} = \mathbf{X}'\mathbf{W}\mathbf{X}$$

de terme général

$$[\mathfrak{I}]_{jk} = E\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k} = -\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

et où $\mathbf{W}$ est la matrice diagonale de "pondération" :

$$[\mathbf{W}]_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2.$$

3.3 Fonction lien canonique

Dans le cas particulier où la fonction lien du modèle linéaire généralisé utilisée est la fonction lien canonique associée à la structure exponentielle alors plusieurs simplifications interviennent :

$$\begin{aligned} \eta_i &= \theta_i = \mathbf{x}_i'\boldsymbol{\beta}, \\ \frac{\partial\mu_i}{\partial\eta_i} &= \frac{\partial\mu_i}{\partial\theta_i} = \frac{\partial v'(\theta_i)}{\partial\theta_i} = v''(\theta_i). \end{aligned}$$

Ainsi,

$$\frac{\partial\ell_i}{\partial\beta_j} = \frac{(y_i-\mu_i)}{\text{Var}(Y_i)} v''(\theta_i)x_{ij} = \frac{(y_i-\mu_i)}{u(\phi)}x_{ij}.$$

De plus, comme les termes $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}$ ne dépendent plus de y_i, on montre que le Hessien est égal à la matrice d'information et donc les méthodes de résolution du score de Fisher et de Newton-Raphson coïncident.

Si, de plus, $u(\phi)$ est constante pour les observations, les équations de vraisemblance deviennent :

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

Ainsi, dans le cas gaussien, le modèle s'écrivant $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ avec la fonction de lien canonique identité, on retrouve la solution :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

qui coïncide avec celle obtenue par minimisation des moindres carrés.

4 Qualité d'ajustement

Il s'agit d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Plusieurs critères sont proposés.

4.1 Déviance

Le modèle estimé est comparé avec le modèle dit *saturé*, c'est-à-dire le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la *déviance* D des log-vraisemblances $\mathcal{L}$ et $\mathcal{L}_{\text{sat}}$:

$$D = -2(\mathcal{L} - \mathcal{L}_{\text{sat}})$$

qui est le logarithme du carré du rapport des vraisemblances. Ce rapport remplace ou "généralise" l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation par moindres carrés.

On montre qu'asymptotiquement, D suit une loi du χ^2 à $n-p$ degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

Attention, l'approximation de la loi du χ^2 peut être douteuse. De plus, dans le cas de données non groupées (modèle binomial), le cadre asymptotique n'est plus adapté car le nombre de paramètres estimés tend également vers l'infini avec n et il ne faut plus se fier à ce test.

4.2 Test de Pearson

Un test du χ^2 est également utilisé pour comparer les valeurs observées y_i à leur prévision par le modèle. La statistique du test est définie par

$$X^2 = \sum_{i=1}^{I} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\text{Var}}(\hat{\mu}_i)}$$

(μ_i est remplacé par $n_i\pi_i$ dans le cas binomial) et on montre qu'elle admet asymptotiquement la même loi que la déviance.

En pratique ces deux approches conduisent à des résultats peu différents et, dans le cas contraire, c'est une indication de mauvaise approximation de la loi asymptotique. Sachant que l'espérance d'une loi du χ^2 est son nombre de

degrés de liberté et, connaissant les aspects approximatifs des tests construits, l'usage est souvent de comparer les statistiques avec le nombre de degrés de liberté. le modèle peut être jugé satisfaisant pour un rapport D/ddl plus petit que 1.

5 Tests

Deux critères sont habituellement proposés pour aider au choix de modèle.

5.1 Rapport de vraisemblance

Comme dans le cas de la régression multiple où un test permet de comparer un modèle avec un modèle réduit, le rapport de vraisemblance ou la différence de déviance est une évaluation de l'apport des variables explicatives supplémentaires dans l'ajustement du modèle. La différence des déviances entre deux modèles *emboîtés* respectivement à q_1 et q_2 $(q_2 > q_1)$ variables explicatives

$$\begin{aligned} D_2 - D_1 &= 2(\mathcal{L}_1 - \mathcal{L}_{\text{sat}}) - 2(\mathcal{L}_2 - \mathcal{L}_{\text{sat}}) \\ &= 2(\mathcal{L}_1 - \mathcal{L}_2) \end{aligned}$$

suit approximativement une loi du χ^2 à $(q_2 - q_1)$ degrés de liberté pour les lois à 1 paramètre (binomial, Poisson) et une loi de Fisher pour les lois à deux paramètres (gaussienne). Ceci permet donc de tester la significativité de la diminution de la déviance par l'ajout de variables explicatives ou la prise en compte d'interactions.

5.2 Test de Wald

Ce test est basé sur la forme quadratique faisant intervenir la matrice de covariance des paramètres, l'inverse de la matrice d'information observée $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$. Cette matrice est calculée à partir du Hessien approché par l'algorithme de maximisation. Elle généralise la matrice $(\mathbf{X}'\mathbf{X})^{-1}$ utilisée dans le cas du modèle linéaire gaussien en faisant intervenir une matrice $\mathbf{W}$ de pondération. Ainsi, test de Wald et test de Fisher sont équivalents dans le cas particulier du modèle gaussien.

Si la matrice $\mathbf{K}$, dite *contraste*, définit l'ensemble H_0 des hypothèses à tester sur les paramètres :

$$\mathbf{K}'\beta = 0,$$

on montre que la statistique

$$(\mathbf{K}'\mathbf{b})'(\mathbf{K}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\mathbf{b}$$

suit asymptotiquement une loi du χ^2.

Attention, le test de Wald, approximatif, peut ne pas être précis si le nombre d'observations est faible.

6 Diagnostics

De nombreux indicateurs, comme dans le cas de la régression linéaire multiple, sont proposés afin d'évaluer la qualité ou la robustesse des modèles estimés. Ils concernent la détection des valeurs influentes et l'étude graphique des résidus. La définition de ces derniers pose quelques difficultés.

6.1 Effet levier

On construit la matrice de projection (hat matrix)

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}')\mathbf{W}^{1/2},$$

relative au produit scalaire de matrice $\mathbf{W}$, sur le sous-espace engendré par les variables explicatives. Les termes diagonaux de cette matrice supérieurs à $(3p/n)$ indiquent des valeurs potentiellement influentes. Le graphe représentant les points d'ordonnées h_{ii} et d'abscisses le numéro de l'observation les visualise.

6.2 Résidus

Avec des erreurs centrées, additives, c'est-à-dire dans le cas du modèle gaussien utilisant la fonction lien identité, il est naturel de définir des résidus par :

$$\varepsilon_i = y_i - E(y_i) = y_i - \mu_i.$$

comme dans le cas du modèle linéaire. Ce cadre est ici inadapté au cas général et différents substituts sont proposés. Chacun possède par ailleurs une version *standardisée* et une version *studentisée*.

Pearson

Les résidus obtenus en comparant valeurs observées y_i et valeurs prédites $\hat{y}_i$ sont pondérés par leur précision estimée par l'écart-type : s_i de $\hat{y}_i$. Ceci définit les résidus de Pearson :

$$r_{Pi} = \frac{y_i - \hat{y}_i}{s_i}$$

dont la somme des carrés conduit à la statistique du même nom. Ces résidus mesurent donc la contribution de chaque observation à la significativité du test découlant de cette statistique. Par analogie au modèle linéaire, on vérifie que ce sont également les résidus de la projection par la matrice $\mathbf{H}$.

Ces résidus ne sont pas de variance unité et sont donc difficiles à interpréter. Une estimation de leurs écarts-types conduit à la définition des résidus de Pearson standardisés :

$$r_{Psi} = \frac{y_i - \hat{y}_i}{s_i\sqrt{h_{ii}}}$$

faisant intervenir le terme diagonal de la matrice $\mathbf{H}$.

De plus, prenant en compte que les estimations des écarts-types s_i dépendent de la $i$ème observation et sont donc biaisés, des résidus studentisés sont obtenus en approchant au premier ordre le paramètre de dispersion $s_{(i)}$ calculé sans la $i$ème observation :

$$r_{Pti} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{h_{ii}}}.$$

Déviance

Ces résidus mesurent la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé. Des versions standardisées et studentisées en sont définies comme pour ceux de Pearson.

Anscombe

Les lois des résidus précédents sont inconnues et même dissymétriques. Anscombe a donc proposé de faire opérer une transformation préalable afin de construire des résidus suivant une loi normale :

$$r_{Ai} = \frac{t(y_i) - t(\hat{y}_i)}{t'(y_i)s_i}.$$

L'explicitation de la fonction t dans le cadre du modèle linéaire généralisé est relativement complexe mais le calcul en est fourni par les logiciels. Comme précédemment, des versions standardisées et studentisées sont également calculées.

Un graphe utilisant ces résidus en ordonnées et les numéros d'observation en abscisses permet d'identifier les observations les moins bien ajustées par le modèle.

6.3 Mesure d'influence

De nombreux indicateurs sont proposés afin d'évaluer l'influence d'une observation sur l'estimation d'un paramètre, sur les prédictions ou encore sur la variance des estimateurs. Le plus utilisé, la distance de Cook, mesure globalement l'influence sur l'ensemble des paramètres. C'est la distance, au sens de la métrique définie par l'inverse de la covariance des paramètres, entre le vecteur des paramètres $\mathbf{b}$ estimé avec toutes les observations et celui estimé lorsque la $i$ème observation est supprimée.

$$D_i = \frac{1}{2}(\mathbf{b} - \mathbf{b}_{(i)})'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{b} - \mathbf{b}_{(i)}).$$

Cet indicateur prend simultanément en compte l'effet levier et l'importance du résidu de chaque observation. Le graphe de ces valeurs est donc plus synthétique et interprétable en tenant compte du graphe des résidus et de celui des termes diagonaux de $\mathbf{H}$.

7 Compléments

7.1 Sur-dispersion

Dans certaines situations, par exemple lors d'observations dépendantes, la variance de la variable Y_i supposée binomiale ou de Poisson, qui est théoriquement fixée par le modèle, est plus importante, multipliée par un facteur d'échelle (scale parameter) σ^2. Si ce paramètre est plus grand que 1, on dit qu'il y a sur-dispersion. Une méthode basée sur une maximisation de la formule de *quasi-vraisemblance* est alors utilisée pour estimer à la fois σ et β.

7.2 Variable “offset”

Lorsque la variable à expliquer dans le cas d’un modèle linéaire généralisé dépend également *linéairement* d’une autre variable, cette dernière est déclarée *offset* et sert ainsi à “tarer” le modèle. Exemple : pour modéliser le nombre de sinistres déclarés par catégorie de conducteurs, la variable *nombre de contrats* est déclarée “offset”.