

Modèles linéaires généralisés

Esterina Masiello

M1 Actuarial 2022 – 2023

N.B. Ces transparents ne sont pas exhaustifs, ils doivent être complétés avec le cours fait au tableau !

Esterina Masiello

Modèles linéaires généralisés

1 / 56

Espérance et variance de Y

On peut montrer que si Y appartient à la famille exponentielle, alors

$$\mu = E(Y) = b'(\theta)$$

$$Var(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$$

où $V(\mu)$ est appelé fonction variance.

*var fonction de la moyenne donc pas forcément constante
 $V(\mu) \neq$ variance.*

Esterina Masiello

Modèles linéaires généralisés

3 / 56

Le modèle linéaire généralisé (MLG)

Le MLG présente trois composantes :

- 1) composante aléatoire du modèle : les v.a. à expliquer Y_1, \dots, Y_n dont les densités appartiennent à la famille exponentielle

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

Souvenir
 $a(\phi) = \frac{\phi}{w_i}$

$y \in S$, avec S un sous-ensemble de \mathbb{N} ou de \mathbb{R} ;

- 2) composante déterministe : les vecteurs explicatifs du modèle $X_1 = (X_{11}, \dots, X_{1n})'$, ..., $X_p = (X_{p1}, \dots, X_{pn})'$;

- 3) une fonction lien g (fonction réelle, déterministe et strictement monotone) telle que

$$g_n(E(Y)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

où $g_n : \mathbb{R}^n \rightarrow \mathbb{R}^n$ définie par $g_n(x_1, \dots, x_n) = (g(x_1), \dots, g(x_n))$.

Esterina Masiello

Modèles linéaires généralisés

2 / 56

Fonction de lien canonique

La fonction de lien canonique est la fonction lien qui associe la moyenne μ_i au paramètre canonique θ_i . Elle est telle que

$$g(\mu_i) = \theta_i.$$

Or, $\mu_i = b'(\theta_i)$ d'où $g^{-1} = b'$.

Loi de probabilité	Fonction de lien canonique
Normale	$\eta = \mu$ (identité)
Poisson	$\eta = \ln(\mu)$ (log)
Gamma	$\eta = \frac{1}{\mu}$ (inverse)
Gaussienne inverse	$\eta = \frac{1}{\mu^2}$
Binomiale	$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$ (logistique)

Esterina Masiello

Modèles linéaires généralisés

4 / 56

Estimation des paramètres du MLG - Méthode MV

Soit $Y = (Y_1, \dots, Y_n)$ le vecteur à expliquer dont la densité s'écrit sous la forme :

$$f_Y(y) = \exp \left\{ \sum_{i=1}^n \frac{\omega_i (y_i \theta_i - b(\theta_i))}{\phi} + \sum_{i=1}^n c_i(y_i; \phi) \right\}$$

Et β ? Il est caché en θ !

Soient X_1, \dots, X_p les p vecteurs explicatifs et g la fonction lien. Alors, $\forall i \in \{1, \dots, n\}$,

$$g(E[Y_i]) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$$

Or, $E[Y_i] = b'(\theta_i)$, b' inversible et g bijective. Il est donc possible d'exprimer, $\forall i \in \{1, \dots, n\}$, θ_i en fonction de $\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$.

Équations de vraisemblance

La fonction de log-vraisemblance s'écrit

$$\begin{aligned} l(\theta(\beta), y, \phi) &= \sum_{i=1}^n \ln(f(y_i; \phi, \theta_i)) \quad \text{avec } f(y_i; \phi, \theta_i) \\ &= \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{\phi} + \sum_{i=1}^n c_i(y_i, \phi) \end{aligned}$$

Pour $j = 0, \dots, p$, nous considérons

$$\frac{\partial l(\theta(\beta); y, \phi)}{\partial \beta_j} = 0$$

Or,

$$\frac{\partial \ln(f(y_i; \phi, \theta))}{\partial \beta_j} = \frac{\partial \ln(f(y_i; \phi, \theta_i))}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j}$$

Estimation des paramètres du MLG - Méthode MV (ctd)

En effet, $\forall i \in \{1, \dots, n\}$

$$\theta_i = (b')^{-1}(E[Y_i]) = T(E[Y_i])$$

$$E[Y_i] = g^{-1}(\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi})$$

et donc $\forall i \in \{1, \dots, n\}$

$$\theta_i = T(E[Y_i]) = (T \circ g^{-1})(\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi})$$

On remplace donc, dans la fonction de vraisemblance, $(\theta_1, \dots, \theta_n)$ par l'expression donné ci-dessus. On notera les estimateurs MV des paramètres $(\beta_0, \dots, \beta_p)$ et ϕ par $(\hat{\beta}_0, \dots, \hat{\beta}_p)$ et $\hat{\phi}$ respectivement. La valeur ajustée sera donnée par

$$\hat{y}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_p X_{pi})$$

Équations de vraisemblance (ctd)

$$\frac{\partial \ln(f(y_i; \phi, \theta_i))}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}$$

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

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = x_{ji} \frac{\partial \mu_i}{\partial \eta_i}$$

On obtient alors

$$\frac{\partial \ln(f(y_i; \theta, \phi))}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ji} \frac{\partial \mu_i}{\partial \eta_i}}{\frac{\phi}{\omega_i} b''(\theta_i)}$$

Car si $Y \in$ Famille expo
alors $E[Y_i] = b'(\theta_i)$

$$\begin{aligned} g(\mu_i) &= x_i^\top \underline{\beta} \\ \underline{\beta} &\text{ SCORE} \\ x_i^\top &= (1 x_{i1} - x_{i2} - x_{ip}) \end{aligned}$$

Equations de vraisemblance (ctd)

Finalement les équations de vraisemblance s'écrivent :

$$\sum_{i=1}^n \omega_i (y_i - \mu_i) \frac{x_{ji}}{b''(\theta_i) g'(\mu_i)} = 0 \quad j \in \{0, \dots, p\}$$

En général pas de solutions explicites!

Méthodes numériques :

- Newton-Raphson (qui utilise le Hessian)
 - du score de Fisher (qui utilise la matrice d'information)
- * Le Hessian est la matrice des dérivées secondes : $[H]_{ij} = \frac{\partial^2 \mathcal{L}(y_i; \beta)}{\partial \beta_i \partial \beta_j}$

* La matrice d'information est la matrice $I = X'WX$ de terme général :

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

et où W est la matrice diagonale de pondération $[W]_{ii} = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

Qualité d'ajustement

Nous considérons deux critères :

- 1) Déviance
- 2) Statistique de Pearson.

Equations de vraisemblance (ctd)

Si on utilise la fonction de lien canonique associée à la structure exponentielle, alors plusieurs simplifications interviennent :

$$\begin{aligned} \eta_i &= \theta_i = x_i' \beta \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) \end{aligned}$$

Ainsi on obtient :

$$\sum_{i=1}^n \omega_i (y_i - \hat{\mu}_i) x_{ji} = 0 \quad j = 0, \dots, p$$

$= g'(x_i' \hat{\beta})$ VALEUR AJUSTÉE PAR LE MODÈLE \hat{y}_i
↳ c'est un résidu!

1) Déviance

Le modèle saturé est le modèle qui possède autant de paramètres que d'observations. Il est caractérisé par $\hat{\mu}_i = y_i$, $i = 1, \dots, n$.

Nous considérons un modèle avec $p + 1 < n$. Idée : le modèle décrit bien les données lorsque $L \simeq L_{SAT}$ et mal lorsque $L << L_{SAT}$.

Comme mesure de la qualité de l'ajustement, nous pouvons considérer la statistique du rapport de vraisemblance

$$\lambda = \frac{L_{SAT}}{L}$$

1) Déviance (ctd)

Ou encore

$$\ln(\lambda) = \ln(L_{SAT}) - \ln(L)$$

La statistique

$$D = 2\ln(\lambda) = 2(\ln(L_{SAT}) - \ln(L))$$

est appelée déviance réduite ou normalisée (Scaled Deviance en SAS) et

$$D^* = \phi D$$

déviance non réduite (déviance en SAS). Il est possible de montrer que

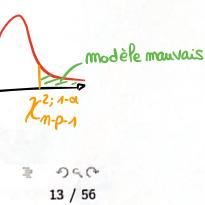
$$D \sim \chi^2_{n-p-1}$$

Le modèle est de mauvaise qualité si $D_{obs} > \chi^2_{n-p-1; 1-\alpha}$

La règle empirique considère que l'ajustement est convenable si $\frac{D}{n-p-1}$ n'est pas beaucoup plus grand que 1.

Esterina Masiello

Modèles linéaires généralisés



13 / 56

Estimation du paramètre de dispersion ϕ

Différentes possibilités :

a) $\tilde{\phi} = \frac{D^*}{n-p-1}$ peu utilisé en pratique.

b) $\hat{\phi} = \frac{1}{n-p-1} (y - \hat{\mu})^t I_n(\hat{\mu})(y - \hat{\mu}) = \frac{X^2}{n-p-1}$

estimateur de Pearson

N.B. L'estimateur par default de ϕ en SAS est l'EMV.

Esterina Masiello

Modèles linéaires généralisés

15 / 56

2) Statistique de Pearson

La statistique du test est définie par

$$X^2 = \sum_{i=1}^n \frac{(\gamma_i - \hat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi} \sum_{i=1}^n \omega_i \frac{(y_i - \hat{\mu}_i)^2}{V(\mu_i)}$$

$$= \frac{\phi}{\omega_i} \frac{V(\mu_i)}{V(Y_i)}$$

Aussi,

$$X^2 \text{ normalisé} = \frac{X^2}{\phi}$$

Comme pour la déviance, nous avons $X^2 \sim \chi^2_{n-p-1}$

Esterina Masiello

Modèles linéaires généralisés

14 / 56

Tests d'hypothèses sur les paramètres – Test entre les modèles emboités

$$H_0 : \beta = \beta_0 = (\beta_0, \beta_1, \dots, \beta_q)^t$$

$$H_1 : \beta = \beta_1 = (\beta_0, \beta_1, \dots, \beta_p)^t \text{ avec } q < p < n.$$

La statistique du test (du rapport de vraisemblance) est

$$\Delta = D_0 - D_1 = 2 \left(\ln \left(L_{\hat{\beta}_0}(y) \right) - \ln \left(L_{\hat{\beta}_1}(y) \right) \right)$$

Il est possible de montrer que

$$\Delta \underset{\text{approx}}{\sim} \chi^2_{p-q}$$

On n'accepte pas H_0 lorsque $\Delta_{obs} > \chi^2_{p-q, 1-\alpha}$.

p : nb de variables explicatives du modèle sous H_0
q : nb de variables explicatives du modèle sous H_1

Esterina Masiello

Modèles linéaires généralisés

16 / 56

Tests d'hypothèses sur les paramètres

Plus en général, nous pouvons écrire :

$$H_0 : C\beta = r$$

avec C matrice connue et r un ensemble de valeurs données.

Soient $\hat{\beta}$ l'EMV de β sans contraintes et $\tilde{\beta}$ l'EMV sous la contrainte $C\beta = r$.

Nous considérons trois approches:

- a) Test du rapport de vraisemblance
- b) Test de Wald
- c) Test du Score

Test de Wald

$\hat{\beta}$ est l'EMV de β .

$$\hat{\beta} \sim \mathcal{N}(\beta, \phi(X'WX)^{-1})$$

avec W matrice diagonale de pondération et

$$[W]_{i,i} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \mu_i} \right)^2$$

Sous H_0 ,

$$C\hat{\beta} - r \sim \mathcal{N}(0, \phi C(X'WX)^{-1}C')$$

La statistique du test de Wald est donc définie comme :

$$(C\hat{\beta} - r)' \{ \phi C(X'WX)^{-1}C' \}^{-1} (C\hat{\beta} - r) \sim \chi_q^2$$

Test du rapport de vraisemblance

Le rapport de vraisemblance est défini comme

$$\lambda = \frac{\hat{L}}{\tilde{L}}$$

avec \hat{L} et \tilde{L} les vraisemblance des modèles sans et avec contraintes.

La statistique du test du rapport de vraisemblance est:

$$2 \ln(\lambda) = 2(\hat{l} - \tilde{l}) \underset{\text{sous } H_0}{\sim} \chi_q^2$$

avec q le nombre de lignes de la matrice C .

Test de Wald (ctd)

Nous voulons effectuer un test d'un seul paramètre:

$$H_0 : \beta_j = r_j$$

En ce cas, $C = (0, \dots, 0, \underbrace{1}_{j^{\text{eme}}}, 0, \dots, 0)$ et $Var(\hat{\beta}_j) = \phi \psi_j$. La

statistique du test de Wald devient alors $\frac{(\hat{\beta}_j - r_j)^2}{\phi \psi_j} \sim \chi_1^2$.

Pour un test de tous les paramètres sauf β_0 ,

$$C_{p \times (p+1)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Test du Score

Ce test est basé sur la dérivée de l en $\tilde{\beta}$, appelée score. Or,

$$l'(\beta) = \phi^{-1} X' W G(y - \mu)$$

avec G matrice diagonale d'éléments $g'(\mu_i)$ et W matrice diagonale d'éléments $[(g'(\mu_i))^2 V(\mu_i)]^{-1}$. On peut montrer que

$$E[l'(\beta)] = 0$$

et

$$\text{Var}[l'(\beta)] = E[l'(\beta)[l'(\beta)]^t] = \phi^{-1} X' W X$$

La statistique du score est donnée par :

$$(l'(\tilde{\beta}))^t [Var(l'(\beta))]^{-1} l'(\tilde{\beta}) \stackrel{\text{sous } H_0}{\sim} \chi_q^2$$

avec $l'(\tilde{\beta}) = \phi^{-1} X' W G(y - \tilde{\mu})$.

Diagnostiques

a) Effet levier

Pour mesurer un tel effet, on se sert de la matrice de projection H telle que $\hat{Y} = HY$.

Dans le cas MLG, on a

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

$$\text{avec } W = \text{diag} \left\{ (V(\mu_i))^2 \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \frac{\phi}{\omega_i} \right\}. \text{ Comme}$$

$Tr(H) = p + 1$, la moyenne vaut $(p + 1)/n$ et les valeurs qui correspondent à $h_{ii} > 2 \frac{p+1}{n}$ doivent faire l'objet d'un examen approfondi.

b) Distance de Cook

$$C_i = \frac{1}{p+1} (\hat{\beta} - \hat{\beta}_{(i)})^t X^t W X (\hat{\beta} - \hat{\beta}_{(i)})$$

Analyse des résidus

Nous définissons différents types de résidus :

- a) Résidus lignes
- b) Résidus de Pearson
- c) Résidus de déviance
- d) Résidus de Anscombe

Résidus lignes et résidus de Pearson

a) Les résidus lignes sont définis comme

$$r_i = y_i - \mu_i$$

Le résidu empirique est

$$\hat{r}_i = y_i - \hat{\mu}_i$$

b) Les résidus de Pearson sont définis par

$$r_i^p = \frac{\sqrt{\omega_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

On remarque que $X^2 = \sum_{i=1}^n (r_i^p)^2$. De plus, $E[r_i^p] = 0$ et

$$\text{Var}(r_i^p) = \frac{\text{Var}(Y_i)}{\frac{V(\mu_i)}{\omega_i}} = \phi. \text{ Le résidu de Pearson empirique est}$$

$$\hat{r}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{V(\hat{\mu}_i)}{\omega_i}}}$$

Résidus de déviance et résidus de Anscombe

c) On peut considérer que chaque observation y_i contribue à hauteur d'une quantité d_i à la déviance, i.e. $D = \sum_{i=1}^n d_i$. Le résidu de déviance est défini comme

$$r_i^D = \text{signe}(y_i - \mu_i) \times \sqrt{d_i}$$

d) Les résidus de Anscombe sont basés sur la différence $h(y_i) - h(\hat{\mu}_i)$ où h est choisie de façon à ce que $h(Y)$ soit approximativement normale. h est telle que

$$h'(y) = (V(y))^{-\frac{1}{3}}$$

Le résidu de Anscombe est défini par

$$\frac{h(y_i) - h(\hat{\mu}_i)}{h'(\hat{y}_i)\sqrt{V(\hat{\mu}_i)}} \underset{\text{approx}}{\sim} \mathcal{N}$$

Comme exemple, nous prenons la loi Gaussienne inverse et calculons les résidus de Anscombe. Nous avons

$$\begin{aligned} V(\mu) &= \mu^3 \\ h'(y) &= (y^3)^{-\frac{1}{3}} = y^{-1} \\ h(y) &= \ln(y) \end{aligned}$$

D'où les résidus de Anscombe :

$$\frac{\ln(y_i) - \ln(\hat{y}_i)}{\sqrt{\hat{y}_i}}$$

Lois de la famille exponentielle et leur paramètres

Loi Y	θ	$b(\theta)$	ϕ	$E[Y]$	$V(\mu) = \frac{Var(Y)}{\phi}$
$\mathcal{B}(n, p)$	$\ln\left(\frac{p}{1-p}\right)$	$n \ln(1 + e^\theta)$	1	np	$np(1-p)$
$\mathcal{P}(\lambda)$	$\ln(\lambda)$	e^θ	1	λ	λ
$\mathcal{N}(m, \sigma^2)$	m	$\frac{\theta^2}{2}$	σ^2	m	1
$\Gamma(m, \nu)$	$-\frac{1}{m}$	$-ln(-\theta)$	ϕ	m	m^2
$IG(m, \sigma^2)$	$-\frac{1}{2m^2}$	$-\sqrt{-2\theta}$	$\frac{\sigma^2}{m^2}\omega$	m	m^3
$BN(\mu, k)$	$\ln\left(\frac{k\mu}{1+k\mu}\right)$	$-\frac{1}{k} \ln(1 - ke^\theta)$	1	μ	$\mu(1 + k\mu)$

Déviance pour les lois de la famille exponentielle

Loi	Déviance
Normale	$\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_{i=1}^n \left(y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right)$
Binomiale	$2 \sum_{i=1}^n n_i \left(y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i) \ln\left(\frac{n_i - y_i}{1 - \hat{\mu}_i}\right) \right)$
Gamma	$2\nu \sum_{i=1}^n \left(-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)$
Gaussienne inverse	$\frac{1}{\sigma^2} \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{\hat{\mu}_i y_i}$
Binomiale Négative	$2 \sum_{i=1}^n \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i + \frac{1}{k}) \ln\left(\frac{y_i + \frac{1}{k}}{\hat{\mu}_i + \frac{1}{k}}\right) \right]$



1) Modèles pour données de comptage

- 1a) Régression de Poisson
- 1b) Régression binomiale négative

1a) Régression de Poisson (ctd)

Nous considérons un modèle simple avec une seule variable explicative quantitative x_1 . Alors $x = (1, x_1)'$, $\beta = (\beta_0, \beta_1)$ et le modèle s'écrit : $g(\mu) = \beta_0 + \beta_1 x_1$.

- lien identité: effet additif

$$(\beta_0 + \beta_1(x_1 + 1)) - (\beta_0 + \beta_1 x_1) = \beta_1$$

- lien log: effet multiplicatif

$$e^{\beta_0 + \beta_1(x_1 + 1)} = e^{\beta_0 + \beta_1 x_1} e^{\beta_1}$$

1a) Régression de Poisson (ou régression log-linéaire)

Le modèle est le suivant :

$$g(\mu) = x'\beta \text{ avec } Y \sim \mathcal{P}(\mu)$$

Pour la fonction lien, le plus souvent nous choisissons :

- fonction lien identité: $\mu = x'\beta$
- fonction lien log: $\ln(\mu) = x'\beta$

N.B. Avec la fonction lien log, la valeur ajustée $\hat{\mu} = \exp\{x'\hat{\beta}\}$ est positive alors que cela n'est pas garanti pour une fonction lien identité.

1a) Régression de Poisson (ctd)

Nous considérons maintenant un modèle avec une seule variable explicative catégorielle à r modalités avec r le niveau de référence. Le modèle s'écrit alors

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1}$$

avec x_1, \dots, x_{r-1} les variables indicatrices des niveaux.

- lien identité: β_j représente le changement de la moyenne de Y dû à la variable correspondante au niveau j par rapport au niveau de référence.
- lien log: $\mu = e^{\beta_0}$ quand la variable prend la modalité correspondante au niveau de référence et $\mu = e^{\beta_0 + \beta_j} = e^{\beta_0} e^{\beta_j}$ quand elle prend la modalité j (effet multiplicatif).

1a) Régression de Poisson (ctd)

Lorsque toutes les variables sont catégorielles, chaque assuré est représenté par un vecteur x dont les composantes valent 0 ou 1. Dans ce cas, $\mu = e^{x'\beta}$ apparaît comme un produit de coefficients de majoration ou de réduction par rapport à l'individu de référence du portefeuille. Plus précisément

$$\begin{aligned}\mu_i &= e^{x'_i \beta} \\ &= e^{\beta_0} \prod_{j=1}^p (e^{\beta_j x_{ij}}) \\ &= e^{\beta_0} \prod_{j|x_{ij}=1} e^{\beta_j}\end{aligned}$$

1a) Régression de Poisson (ctd)

Au final, e^{β_0} est la moyenne de l'individu de référence du portefeuille tandis que chacun des facteurs e^{β_j} traduit l'influence d'un critère de segmentation.

Si $\beta_j > 0$, les individus présentant la caractéristique j subiront une majoration de prime par rapport à la prime de référence e^{β_0} .

Si $\beta_j < 0$ il y aura réduction de prime.

1b) Surdispersion et régression binomiale négative

En pratique, on observe souvent un phénomène de surdispersion qu'il faut prendre en compte puisque la surdispersion entraîne une sous-estimation de la variance des estimateurs et, en conséquence,

- des intervalles de confiance trop étroits
- une surestimation des statistiques du χ^2 du test de significativité de chaque coefficient.

Comment prendre en compte la surdispersion ?

En estimant un modèle de régression binomiale négative avec lien log

$$\ln(\mu) = \ln(n) + x'\beta \text{ avec } Y \sim BN(\mu, k)$$

Quasi vraisemblance

Rappelez-vous que pour la famille exponentielle $Var(Y) = \phi V(\mu)$. Nous aimerais modéliser la surdispersion en considérant

$$Var(Y) = \phi\mu \quad \text{avec } \phi > 1.$$

En ce cas, les équations de vraisemblance

$$\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta_j} \frac{y_i - \mu_i}{\phi V(\mu_i)} \quad (*)$$

ne correspondent pas à la dérivée de la fonction de logvraisemblance d'une famille exponentielle. Le problème est résolu en maximisant la quasi-vraisemblance, $Q(\beta)$, définie comme toute fonction de β dont la dérivée $Q'(\beta)$ est donnée par (*).

Exemple : Quasi Poisson

Supposons que $V(\mu) = \mu$ et $Var(Y) = \phi V(\mu) = \phi\mu$ (alors que pour la loi de Poisson $Var(Y) = E(Y) = \mu$).

Nous pouvons montrer que

$$Q(\beta) = \sum_{i=1}^n (y_i \ln(\mu_i) - \mu_i)$$

En maximisant $Q(\beta)$, nous obtenons le même $\hat{\beta}$ que celui obtenu par une régression de Poisson mais les écarts-types sont multipliés par un facteur $\sqrt{\phi}$.

→ Voir page suivante

2a) Régression logistique

Le modèle de régression logistique s'écrit

$$g(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = x'\beta$$

avec $Y \sim \mathcal{B}(\pi)$ et $\pi = P(Y = 1)$. La fonction lien est la fonction lien logit ($g(\mu) = \frac{\mu}{1-\mu}$) qui est la fonction lien canonique pour Bernoulli.

Le rapport $\frac{\pi}{1-\pi}$ est appelé rapport de côtes (odds ratio) ou simplement côte.

Nous trouvons facilement

$$\pi = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

2) Modèles pour variables réponses catégorielles

2a) Régression logistique

2b) GLM multivarié

2b*) Modèle de régression nominale

2b**) Variable réponse ordinaire

2b**a) Modèle logistique cumulatif

2b**b) Modèle loglog complémentaire cumulatif

2b**c) Modèle probit cumulatif

La fonction lien la plus populaire pour une réponse Bernoulli est la logit mais nous pouvons utiliser d'autres fonctions liens :

- fonction lien probit

$$probit(\pi) = \Phi^{-1}(\pi)$$

avec Φ^{-1} l'inverse de la normale centrée réduite. La valeur ajustée est

$$\hat{\pi} = \Phi(x'\hat{\beta})$$

- fonction lien loglog complémentaire

$$cloglog(\pi) = \ln(-\ln(1-\pi))$$

La valeur ajustée est

$$\hat{\pi} = 1 - \exp(-\exp(x'\hat{\beta}))$$

MODELES A INFLATION DE ZEROS (ZERO INFLATION MODEL)

(ou MODELE ZERO-MODIFIÉ)

ZINB

Il s'agit de modèles alternatifs (à ceux vus auparavant pour les variables de comptage) que l'on peut utiliser lorsqu'on observe trop de personnes non sinistres dans la population totale (par rapport à ce qu'on aurait observé dans un modèle Poissonien par exemple). C'est comme si la sinistralité observée était le résultat de deux sortes de processus : un processus qui génère des zeros structuels et un processus qui génère des sinistres aléatoires.

Ex: mécanisme de Bonus-Malus qui génère un biais puisqu'il peut inciter les assurés à ne pas déclarer un sinistre (de montant raisonnable) pour éviter une majoration de prime l'année suivante.

Un MODELE ZERO-MODIFIÉ est un mélange entre une masse en zéro et un modèle classique de comptage (Poisson ou BN).

On note π_i la probabilité de ne pas déclarer un sinistre (proba d'un zéro). On considère, par exemple, un modèle logistique :

$$\pi_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$$

Pour le modèle de comptage, on note $p_i(k)$ la probabilité que l'ième individu ait k sinistres (modéliser par exemple par une loi de Poisson).

Alors,

$$P(N_i = k) = \begin{cases} \pi_i + (1 - \pi_i)p_i(0) & \text{si } k=0 \\ (1 - \pi_i)p_i(k) & \text{si } k>0 \end{cases}$$

Pour le MODELE ZIP (ZERO INFLATED Poisson), on a $P(N_i = k) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\lambda} & \text{si } k=0 \\ (1 - \pi_i) \frac{\lambda^k e^{-\lambda}}{k!} & \text{si } k>0 \end{cases}$

Pour le MODELE ZINB (ZERO INFLATED BINOMIAL), on a $P(N_i = k) = \begin{cases} \pi_i + (1 - \pi_i)BN(0, n, p) & \text{si } k=0 \\ (1 - \pi_i)BN(k, n, p) & \text{si } k>0 \end{cases}$

$$\text{avec } BN(k, n, p) = \binom{k+n-1}{k} p^k (1-p)^n$$

Slide 40:

Remarque: La fonction liens log-log complémentaire n'est pas symétrique autour de

$T = \frac{1}{2}$, à la différence des fm Logit et probit

utilisée pour l'analyse des données de survie.

Remarque : pour la régression logistique, la déviance D n'est pas considérée comme une mesure de la qualité d'ajustement du modèle utile. Nous avons

$$\hat{\pi}_i = \frac{e^{x'_i \hat{\beta}}}{1 + e^{x'_i \hat{\beta}}}$$

avec $\hat{\beta}$ l'EMV. On peut montrer que

$$D = -2 \sum_{i=1}^n \left[\hat{\pi}_i \ln \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \ln(1 - \hat{\pi}_i) \right]$$

qui dépend seulement des valeurs ajustées. La statistique de Pearson est définie comme

$$X^2 = \sum_{i=1}^n \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)} \stackrel{\text{approx}}{\sim} \chi^2_{n-p-1}$$

2b) GLM multivarié

Nous considérons une variable réponse qualitative avec r catégories ($r > 2$). Nous aurons donc $r - 1$ variables indicatrices Y_j , $j = 1, \dots, r - 1$. La réponse est donc multivariée

$$Y = \{Y_1, \dots, Y_{r-1}\}'$$

(Fahrmeir et Tutz (2001) *Multivariate statistical modeling based on generalized linear models*).

Nous considérons n réalisations indépendantes de Y et notons par $n_j = \sum y_j$ le nombre de fois que la modalité j se présente. La loi jointe de n_1, \dots, n_r est multinomiale

$$f(n_1, \dots, n_r) = \frac{n!}{n_1! \cdots n_r!} \pi_1^{n_1} \cdots \pi_r^{n_r}$$

avec $\pi_j = P(Y = j)$, $\sum_j \pi_j = 1$ et $n = \sum_j n_j$.

2b*) Modèle de régression nominale

Les π_j sont liées aux variables explicatives. On modélise la côte de Y dans la catégorie j par rapport au niveau de référence r

$$\ln \left(\frac{\pi_j}{\pi_r} \right) = \theta_j + x' \beta_j$$

$j = 1, \dots, r - 1$. Nous avons donc

$$\pi_r = \frac{1}{1 + \sum_{k=1}^{r-1} e^{\theta_k + x' \beta_k}}$$

$$\pi_j = \pi_r e^{\theta_j + x' \beta_j} \quad j = 1, \dots, r - 1$$

2b**) Variable réponse ordinaire

Y est une variable réponse avec r ($r > 2$) modalités ordonnées.

Exemple: réponse à la question : "comment évaluez-vous votre santé ?".

Il existe une variable latente non observable, l'état de santé d'un individu, que l'on note Y^* . On considère alors des valeurs seuil $\theta_0, \dots, \theta_r$ telles que

$$y = j \quad \text{si} \quad \theta_{j-1} \leq y^* < \theta_j \quad j = 1, \dots, r$$

2b**) Variable réponse ordinaire (ctd)

Nous considérons les probabilités cumulées

$$\tau_j = P(Y \leq j) = P(Y^* < \theta_j) \quad j = 1, \dots, r$$

que l'on souhaite lier aux variables explicatives. Supposons que Y^* dépend des caractéristiques de l'individu

$$Y^* = -x'\beta + \epsilon$$

avec $E[\epsilon] = 0$. On obtient alors

$$\tau_j = P(\epsilon \leq \theta_j + x'\beta)$$

Selon le choix de loi pour ϵ , on obtient des modèles différents:

2b**) Variable réponse ordinaire (ctd)

2b**a) Modèle logistique cumulatif

2b**b) Modèle loglog complémentaire cumulatif

2b**c) Modèle probit cumulatif

2b**a) Modèle logistique cumulatif

Supposons que ϵ est de loi logistique

$$P(\epsilon \leq x) = \frac{1}{1 + e^{-x}}$$

Alors

$$\tau_j = P(\epsilon \leq \theta_j + x'\beta) = \frac{1}{1 + e^{-(\theta_j + x'\beta)}}$$

et donc le modèle s'écrit

$$\ln \frac{\tau_j}{1 - \tau_j} = \theta_j + x'\beta \quad j = 1, \dots, r - 1$$

Ainsi, on retrouve l'effet de la modalité à travers θ et, de plus, les $r - 1$ équations sont parallèles.

2b**b) Modèle log-log complémentaire cumulatif

Nous considérons la fonction lien log-log complémentaire

$$\ln[-\ln(1 - \tau_j)] = \theta_j + x'\beta \quad j = 1, \dots, r - 1$$

2b**c) Modèle probit cumulatif

Si $\epsilon \sim N$, alors le modèle s'écrit

$$\Phi^{-1}(\tau_j) = \theta_j + x'\beta \quad j = 1, \dots, r-1$$

Test pour l'hypothèse de côtes proportionnelles

Nous voulons tester

$$H_0 : \beta_j = \beta \text{ pour } j = 1, \dots, r-1$$

La statistique du test du score est donnée par

$$\left[l'(\tilde{\beta}) \right]^t \{Var(l'(\beta))\}^{-1} l'(\tilde{\beta}) \sim \chi^2_{p(r-2)} \text{ sous } H_0$$

avec $\tilde{\beta}$ l'estimateur de β dans le modèle sous H_0 et $l'(\tilde{\beta})$ le score.

Si nous n'acceptons pas H_0 , deux possibilités :

- on estime le modèle de régression nominale
- on estime le modèle partiel des côtes proportionnelles.

2b**a) Modèle logistique cumulatif (ctd)

Une hypothèse forte du modèle logistique cumulatif est que les coefficients β sont communs pour tous les niveaux de la variable réponse. En l'absence de cette hypothèse, nous aurions

$$\ln \frac{\tau_j}{1 - \tau_j} = \theta_j + x'\beta_j \quad j = 1, \dots, r-1$$

avec β_j les vecteurs de coefficients de régression spécifiques à chaque groupe j .

Modèle partiel des côtes proportionnelles

Si nous n'acceptons pas l'hypothèse de côtes proportionnelles, nous pouvons considérer le modèle suivant

$$\ln \left(\frac{\tau_j}{1 - \tau_j} \right) = \theta_j + x'\beta + \omega'\alpha_j \quad j = 1, \dots, r-1$$

avec x l'ensemble de toutes les variables explicatives et ω un sous-ensemble de x pour lequel H_0 n'est pas vérifiée.

Les α_j représentent des incrémentations des coefficients β au niveau j .

3) Modèles pour variables réponses continues

Deux solutions possibles:

- modèle linéaire sur la variable réponse transformée
- modèle linéaire généralisé
 - 3a) Régression Gamma
 - 3b) Régression Gaussienne inverse
 - 3c) Régression Tweedie

3a) Régression Gamma

Le modèle s'écrit

$$g(\mu) = x'\beta \text{ avec } Y \sim \Gamma(\mu, \nu)$$

La fonction lien canonique pour la Gamma est la fonction inverse mais souvent nous utilisons la fonction lien *log*.

3b) Régression Gaussienne inverse

Le modèle s'écrit

$$g(\mu) = x'\beta \text{ avec } Y \sim IG(\mu, \sigma^2)$$

La fonction lien canonique de la Gaussienne inverse est $g(\mu) = \mu^{-2}$ mais nous utilisons souvent la fonction lien *log*.

3c) Régression Tweedie

Soit N le nombre de sinistres pour une police et Z_1, \dots, Z_N les coûts individuels. Le coût total est alors donné par

$$Y = \begin{cases} \sum_{j=1}^N Z_j & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

Si $N \sim \mathcal{P}(\lambda)$ et $Z_j \sim \text{Gamma}$ indépendantes et N indépendant de Z , alors Y suit une loi Tweedie.

La loi Tweedie fait partie de la famille exponentielle avec $\text{Var}(Y) = \phi\mu^p$ avec $1 < p < 2$.