

Modèles linéaires généralisés

22 mai 2018

M1 Actuariat, année 2017-2018

Durée : 2h

Une feuille, seulement recto, manuscrite est autorisée.

Toutes les réponses doivent être soigneusement justifiées.

Pour les questions de l'exercice 1, on ne demande pas uniquement la formule du cours qui donne le résultat : il faut présenter une preuve des résultats.

Exercice 1 Nous considérons le modèle linéaire généralisé

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

avec $\mu = E(Y)$ et Y appartient à la famille exponentielle.

- i) Quelles sont les raisons qui poussent à utiliser le modèle linéaire généralisé par rapport au modèle linéaire classique en tarification ?
- ii) Définir la fonction variance et en expliquer l'importance en mettant en avant la différence avec le modèle linéaire classique.
- iii) Décrire la procédure de construction du test entre modèles emboîtés après en avoir spécifié les hypothèses nulle et alternative. Et dans un modèle de régression linéaire classique comment procède-t-on ?
- iv) Présenter une mesure de la qualité d'ajustement du modèle en en expliquant l'utilisation. Et dans le modèle de régression linéaire classique comment juge-t-on de la quantité d'ajustement ?
- v) On suppose maintenant que $Y \sim BN(\mu, k)$ de densité :

$$f(y) = \frac{\Gamma(y + \frac{1}{k})}{y! \Gamma(\frac{1}{k})} \left(\frac{1}{1 + k\mu} \right)^{1/k} \left(\frac{k\mu}{1 + k\mu} \right)^y \quad y = 0, 1, 2, \dots$$

- a) Montrer que la loi binomiale négative appartient à la famille exponentielle tout en spécifiant le paramètre de la moyenne θ , le paramètre de dispersion, les fonctions b et c .
- b) Trouver la fonction lien canonique ainsi que la fonction variance $V(\mu)$.
- c) Avec les éléments trouvés aux points précédents, calculer $E(Y)$ et $Var(Y)$.
- d) Montrer que la déviance est donnée par $D = 2 \sum_{i=1}^n \{ y_i \ln(\frac{y_i}{\mu_i}) - (y_i + \frac{1}{k}) \ln(\frac{y_i + 1/k}{\mu_i + 1/k}) \}$.
- e) Écrire les résidus de déviance.
- vi) On cherche maintenant à expliquer le taux de rachat en assurance par certaines variables explicatives. Quel type de modèle proposeriez-vous ? L'écrire et commenter. Une variable offset serait-elle pertinente ? Oui ? Non ? Pourquoi ?

Exercice 2 Nous considérons un portefeuille d'assurance Automobile, dont les polices ont générée au moins un sinistre. On dispose des variables suivantes :

Variable	Modalités
clm_amt	montant de sinistres
bluebk	valeur de la voiture
npolicy1	nombre de polices souscrites par l'assuré
mvrcat	points cumulés
DENSITY	densité de la population de la zone de résidence de l'assuré

Grâce à la procédure PROC GENMOD de SAS, on a pu obtenir les résultats présentés dans le Tableau ci-dessous :

Distribution		Gamma		log gamma	
Link Function		Log			
Dependent Variable		clm_amt			
Number of Observations Read	2746				
Number of Observations Used	2746				
Class Level Information					
Class	Value	Design Variables			
npolicy1	0	1			
	1	0			
mvrcat	1	1 0 0 0 0			
	2	0 1 0 0 0			
	3	0 0 1 0 0			
	4	0 0 0 1 0			
	5	0 0 0 0 1			
	6+	0 0 0 0 0			
DENSITY	Highly Rural	1 0 0			
	Highly Urban	0 1 0			
	Rural	0 0 1			
	Urban	0 0 0			
Critère		DF	Valeur	Valeur/DF	
Deviance		2734	1906.1893	0.6972	
Scaled Deviance		2734	3024.5955	-1.1063	
Pearson Chi-Square		2734	4105.1339	1.5015	
Scaled Pearson X2		2734	6513.7126	2.3825	
Log Likelihood			-26269.9604		

Paramètre	DF	Estimation	Erreur standard	Wald 95% Limites de confiance %	Khi 2	Pr > Khi 2
Intercept	1	8.6621	0.0372	8.5893 8.7350	54358.1	<.0001 ***
mvrcat	1	-0.0635	0.0501	-0.1616 0.0346	1.61	0.2048
mvrcat	2	0.1087	0.0516	0.0076 0.2097	4.44	0.0350 **
mvrcat	3	-0.0134	0.0524	-0.1161 0.0892	0.07	0.7974
mvrcat	4	0.1070	0.0422	0.0243 0.1898	6.42	0.0113 *
mvrcat	5	0.1138	0.0563	0.0035 0.2241	4.09	0.0432 *
bluebk	1	0.0251	0.0022	0.0208 0.0295	128.53	<.0001 ***
bluebk*bluebk	1	-0.0006	0.0002	-0.0009 -0.0003	17.98	<.0001 ***
npolicy1	0	0.0859	0.0308	0.0256 0.1463	7.78	0.0053 **
DENSITY	Highly Rural	1 0.1089	0.1436	-0.1726 0.3904	0.58	0.4483
DENSITY	Highly Urban	1 -0.1158	0.0325	-0.1796 -0.0520	12.66	0.0004 ***
DENSITY	Rural	1 -0.1238	0.0832	-0.2869 0.0394	2.21	0.1371
Scale	1	1.5867	0.0391	1.5118 1.6653		

NOTE: The scale parameter was estimated by maximum likelihood.

Statistiques LR pour Analyse de Type 3

Source	DF	Khi 2	Pr > Khi 2
mvrcat	5	18.27	0.0026
bluebk	1	122.52	<.0001
bluebk*bluebk	1	16.30	<.0001
mpolicy1	1	7.79	0.0053
DENSITY	3	14.94	0.0019

Nous avons également représenté (Figure 1) les résidus de déviance contre les \hat{y}_i .

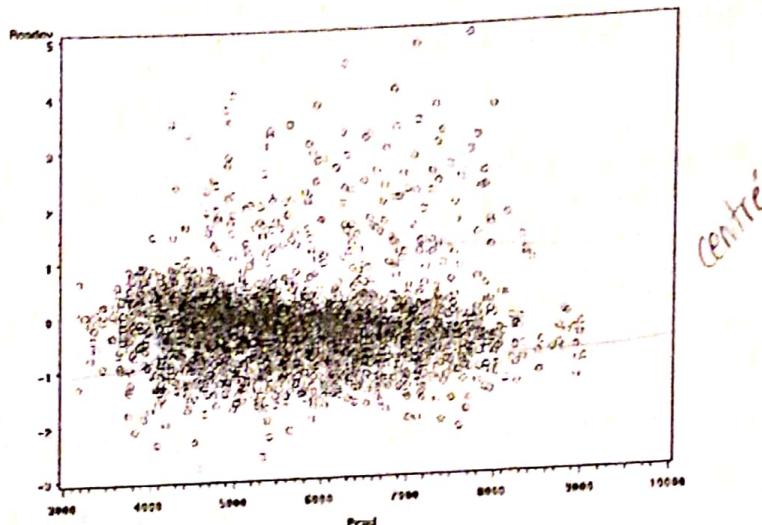


Fig. 1. Résidus de déviance

- 1) Écrire le modèle et présenter les caractéristiques de l'assuré de référence.
- 2) Combien vaut le paramètre de dispersion ϕ ? Quelle est la différence entre Deviance et Scaled Deviance dans le tableau ci-dessus?
- 3) Commenter l'ajustement du modèle ainsi que la significativité des variables explicatives. Les signes des β correspondent à vos intuitions?
- 4) Quelles informations peut-on tirer de l'analyse de type III effectuée sur ce modèle?
- 5) Que suggérez-vous afin d'améliorer le modèle estimé?
- 6) Nous effectuons un test dont l'hypothèse nulle est $H_0 : \beta_{HighlyRural} = \beta_{Rural}$. Pourquoi effectuer ce test? Quel type d'approche utiliser pour construire ce test?
- 7) Quel est le montant moyen de sinistres de l'assuré de référence? Quel est le montant moyen de sinistres d'un assuré qui a cumulé 5 points?
- 8) Définir les résidus de déviance. Les résidus de déviance de ce modèle vous semblent-ils satisfaisants? Oui? Non? Pourquoi?

Exercice 1:

i) En tarification, on s'intéresse à $E[S] = E[N]E[X]$ montant simétrique
pas gaussien

ii) Soit Y une variable aléatoire suivant une loi appartenant à la famille exponentielle

$$\text{Var}(Y) = a(\phi) b''(\theta)$$

$\frac{\text{Var}}{V(Y)} \rightarrow$ ça définit la relati° entre moyenne de Y et variance de Y .

↳ Cela met en avant que dans les MLG, la variance n'est pas constante \neq modèle gaussien où

$$\text{Var}(Y_i) = \text{Var}(E_i) = \sigma^2 \quad \forall i$$

(homoschématie)

iii) En GLM. Test entre les modèles emboîtés

$$H_0: \beta = \beta_0 = (\beta_0, \dots, \beta_q) \text{ vs } H_1: \beta = \beta_1 = (\beta_0, \dots, \beta_p) \text{ avec } q < p$$

Il y a ensuite 3 approches qui vont consister à réaliser un test du χ^2 .

* Test du rapport de vraisemblance: $\lambda = \frac{\bar{L}}{L} \leftarrow$ vraisemblance sans contrainte
"avec"

$$\rightarrow 2 \ln(\lambda) = 2(\bar{L} - L) \sim \chi_q^2$$

Test du rapport de vraisemblance
Le rapport de vraisemblance est défini comme
 $\lambda = \frac{\bar{L}}{L}$
avec \bar{L} et L les vraisemblances des modèles sans et avec contraintes.
La statistique du test du rapport de vraisemblance est:
 $2 \ln(\lambda) = 2(\bar{L} - L) \xrightarrow{\text{sous } H_0} \chi_q^2$
avec q le nombre de lignes de la matrice C .

* Test de Wald: $\hat{\beta}$ EMV de β tq $\hat{\beta} \sim N(\beta, \phi(X'WX)^{-1})$

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

Test de Wald
 $\hat{\beta}$ est l'EMV de β .
 $\beta \sim N(\beta, \phi(X'WX)^{-1})$
avec W matrice diagonale de pondération et
 $[W]_{i,i} = \frac{1}{Var(Y_i)} (\frac{\partial \mu_i}{\partial \beta})^2$
Sous H_0 : $C\hat{\beta} - r \sim N(0, \phi(C(X'WX)^{-1}C))$
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 Score: $(l'(\hat{\beta}))' [Var(l'(\beta))]^{-1} l'(\hat{\beta}) \sim \chi_q^2$

$$\text{avec } l'(\hat{\beta}) = \phi^{-1} X'WG(y - \mu)$$

G la matrice diagonale d'éléments $g'(\mu_i)$

Test du Score
Ce test est basé sur la dérivée de l en β , appelée score. Or,
 $l'(\beta) = \phi^{-1} X'WG(y - \mu)$
avec G matrice diagonale d'éléments $g'(\mu_i)$ et W matrice diagonale d'éléments $[g'(\mu_i)^2 V(\mu_i)]$. On peut montrer que
 $E[l'(\beta)] = 0$
et $Var[l'(\beta)] = E[l'(\beta)[l'(\beta)]^T] = \phi^{-1} X'WX$
La statistique du score est donnée par :
 $(l'(\hat{\beta}))' [Var(l'(\beta))]^{-1} l'(\hat{\beta}) \xrightarrow{\text{sous } H_0} \chi_q^2$

Modèle emboité (test des q derniers coeff significatifs). En régr. lin. classique

$$H_0: \beta_{p+1} = \dots = \beta_q = 0$$

vs

$$H_1: \beta_j \in [\beta_{p+1}, \dots, \beta_q] \text{ tq } \beta_j \neq 0$$

\rightarrow Test de Fisher.

$$F^* = \frac{SCR(H_0) - SCR(H_1)}{SCR(H_1)} \times \frac{m-p-1}{q} \sim F_{q, m-p-1}$$

iv) En GLM: La mesure de la qualité d'ajustement est la Déviance (ou stat de Pearson).

© Théo Jalabert

→ On considère la stat du rapport de vraisemblance

$$\lambda = \frac{L_{\text{SAT}}}{L} \quad L_{\text{SAT}}: \text{vraisemblance du modèle saturé.}$$

Idee: à $L \approx L_{\text{SAT}}$ qualité d'ajustement bonne
Si $L_{\text{SAT}} \gg L \Rightarrow$ pas bon!

On réécrit λ pour obtenir la déviance

$$\rightarrow D = 2 \ln(\lambda) = 2 [L_{\text{SAT}} - L]$$

$$\text{avec } D \sim \chi^2_{m-p_1}$$

Le modèle est mauvais si $D_{\text{obs}} > \chi^2_{m-p_1, 1-\alpha}$ ou ajustement convenable si $\frac{D}{m-p_1} \sim 1$

Dans le modèle linéaire classique: mesure avec R^2 (coeff de détermination).

$$R^2 = \frac{\text{Variance expliquée}}{\text{Variance totale}} = \frac{SCE}{SCT} \quad \text{ou } R^2 = 1 - \frac{m-1}{m-p_1} \times \frac{SCE}{SCT} \quad \text{on veut } R^2 \sim 1.$$

y) $Y \sim BN(\mu, k) \rightarrow f(y) = \frac{\Gamma(y+1/k)}{y! \Gamma(1/k)} \left(\frac{1}{1+k\mu}\right)^{1/k} \left(\frac{k\mu}{1+k\mu}\right)^y \quad y \in \mathbb{N}.$

a) $\ln(f(y)) = y \ln\left(\frac{k\mu}{1+k\mu}\right) + \frac{1}{k} \ln\left(\frac{1}{1+k\mu}\right) + \ln\left(\frac{\Gamma(y+1/k)}{y! \Gamma(1/k)}\right)$

$$\Rightarrow f(y) = \exp\left[y \ln\left(\frac{k\mu}{1+k\mu}\right) + \frac{1}{k} \ln\left(\frac{1}{1+k\mu}\right) + \ln\left(\frac{\Gamma(y+1/k)}{y! \Gamma(1/k)}\right)\right]$$

$$\Rightarrow \theta = \ln\left(\frac{k\mu}{1+k\mu}\right) \Rightarrow 1-e^\theta = 1 - \frac{k\mu}{1+k\mu} = \frac{1}{1+k\mu}$$

$$\Rightarrow b(\theta) = -\frac{1}{k} \ln\left(\frac{1}{1+k\mu}\right) = -\frac{1}{k} \ln(1-e^\theta)$$

$$\phi = 1, a(\phi) = 1 \\ c(y, \phi) = \ln\left(\frac{\Gamma(y+1/k)}{y! \Gamma(1/k)}\right)$$

Donc la loi binomiale négative appartient bien à la famille exponentielle.

b) $\delta_i = \ln\left(\frac{k\mu_i}{1+k\mu_i}\right) = g(\mu_i) \quad \text{où } g: x \mapsto \ln\left(\frac{kx}{1+kx}\right) \text{ fonction logistique}$

$$\sqrt{V(\mu)} = b''(\theta) \Rightarrow \sqrt{V(\mu)} = \frac{1}{k} \frac{e^\theta}{(1-e^\theta)^2} = \frac{1}{k} \frac{\frac{k\mu}{1+k\mu}}{\left(\frac{1}{1+k\mu}\right)^2} = \mu(1+k\mu)$$

c) $E(Y) = b'(\theta) = \frac{1}{k} \frac{e^\theta}{1-e^\theta} = \frac{1}{k} (1+k\mu) \times \frac{k\mu}{1+k\mu} = \mu$.

et $\text{Var}(Y) = a(\phi) b''(\theta) = 1 \times b''(\theta) = \mu(1+k\mu)$

$$d) D = 2[L_{SAT} - L]$$

$$L_{SAT} = \prod_{i=1}^m f(y_i) = \exp \left[y_i \theta_i - b(\theta_i) + c(y_i; \phi) \right] \quad \text{modèle saturé} \Rightarrow y_i = \mu_i$$

$$\Rightarrow L_{SAT} = \exp \left[y_i \ln \left(\frac{\mu_i}{1+\mu_i} \right) + \frac{1}{k} \ln \left(\frac{1}{1+\mu_i} \right) + c(y_i; \phi) \right]$$

$$L = \prod_{i=1}^m f(y_i) = \exp \left[\sum_{i=1}^m (y_i \theta_i - b(\theta_i) + c(y_i; \phi)) \right] \quad \text{modèle estimé} \Rightarrow \mu_i = \hat{\mu}_i$$

$$\Rightarrow L = \exp \left[\sum_{i=1}^m \left(y_i \ln \left(\frac{\hat{\mu}_i}{1+\hat{\mu}_i} \right) + \frac{1}{k} \ln \left(\frac{1}{1+\hat{\mu}_i} \right) + c(y_i; \phi) \right) \right]$$

$$\begin{aligned} \Rightarrow 2[L_{SAT} - L] &= 2 \left[\sum_{i=1}^m \left(y_i \ln \left(\frac{\mu_i}{1+\mu_i} \right) - \frac{1}{k} \ln \left(\frac{1}{1+\mu_i} \right) - y_i \ln \left(\frac{\hat{\mu}_i}{1+\hat{\mu}_i} \right) - \frac{1}{k} \ln \left(\frac{1}{1+\hat{\mu}_i} \right) \right) \right] \\ &= 2 \sum_{i=1}^m \left(y_i \left(\ln \left(\frac{\mu_i}{1+\mu_i} \right) - \ln \left(\frac{\hat{\mu}_i}{1+\hat{\mu}_i} \right) \right) + \frac{1}{k} \left(\ln \left(\frac{1}{1+\mu_i} \right) - \ln \left(\frac{1}{1+\hat{\mu}_i} \right) \right) \right) \\ &= 2 \sum_{i=1}^m \left(y_i \ln \left(\frac{\mu_i}{\hat{\mu}_i} \right) - y_i \ln \left(\frac{1+k\mu_i}{1+k\hat{\mu}_i} \right) - \frac{1}{k} \ln \left(\frac{1+k\mu_i}{1+k\hat{\mu}_i} \right) \right) \\ &= 2 \sum_{i=1}^m \left(y_i \ln \left(\frac{\mu_i}{\hat{\mu}_i} \right) - (y_i + \frac{1}{k}) \ln \left(\frac{\mu_i + \frac{1}{k}}{\hat{\mu}_i + \frac{1}{k}} \right) \right) \end{aligned}$$

$$e) \hat{\pi}_i^D = \text{Signe}(y_i - \hat{\mu}_i) \times \sqrt{d_i} = \text{Signe}(y_i - \hat{\mu}_i) \times \sqrt{y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + \frac{1}{k}) \ln \left(\frac{y_i + \frac{1}{k}}{\hat{\mu}_i + \frac{1}{k}} \right)}$$

v.) Modèle de type: $g(\frac{\mu}{m}) = x' \beta$ avec la fonction lien ln.
et Y suit une loi de Poisson ou loi binomiale négative.

$$\Rightarrow h(\mu) = \underbrace{h(m)}_{\text{variable offset}} + x' \beta$$

Exercice 2:

1) Analyse de type III: p-value très petite \Rightarrow on rejette H₀ (sans mvrcat) donc mvrcat significative.

$$\text{Modèle: } \log(\mu) = \beta_0 + \beta_1 \mathbb{1}_{\text{mvrcat}=1} + \dots + \beta_5 \mathbb{1}_{\text{mvrcat}=5} + \beta_6 \text{bluebk} + \beta_7 \text{bluebk}^2 + \beta_8 \text{mpolicy} + \beta_9 \text{Density}_{high_rural} + \beta_{10} \text{Density}_{high_urban}$$

$$+ \beta_{11} \text{Density}_{rural}$$

Ici, Y suit une loi Gamma.

L₀ Assur de référence: c'est celle qui présente les modalités de réf (celles qui sont dans le tableau).

2) Le paramètre de dispersion estimé est: $\hat{\phi} = \frac{1}{\text{scale}} = 0,63024$ (vraie pour la Gamma)

Dans le tableau, Scaled Deviance = D = $2 \ln(\lambda) = 2(L_{SAT} - \bar{L})$ \leftarrow Deviance

et Deviance = $D^* = \phi D$ \leftarrow Deviance non réduite.

$$\hookrightarrow 1906,1893 = \frac{1}{1,5857} \times 304,5855$$

3) Ici la $\frac{\text{déviance}}{m-p-1}$ vaut $1/1063 \rightarrow$ proche de 1 donc ajustement bon a priori.

\hookrightarrow Si on a le quantile $\chi^2_{m-p-1, \alpha}$, il faut comparer au quantile Si: $D_{obs} > \chi^2_{m-p-1, \alpha} \Rightarrow$ pas bon.

Les Va sont signif sauf:

* mvrca1

* mvrca2

* Density_{Highrural}

* Density_{Rural}

On peut essayer de regrouper certaines de ces variables entre elles pour déterminer si la nouvelle va est signif.

TCEPA

Les signes des β correspondent à l'intuition "moins on a de pts plus on est à risque \Rightarrow sinistre" car l'assurance de ref à 6^e points (attention aux coeffs des variables non signif).

Pour les autres aussi.

4) L'analyse de type III, p/r aux variables qualitatives, nous dit lesquelles sont significatives, suppose des quest° p/r aux groupes de modalités.

5) \rightarrow On peut retirer la variable la moins significative et au vue de l'analyse de type 3 il faut essayer de grouper des modalités par ex: highrural et rural.

6) $H_0: C\beta = 0$ avec $C = (0, -1, 0, 1, \dots)$ \oplus Test de Wald

Si on accepte $H_0 \rightarrow$ alors on peut les grouper.

7) Le montant moyen est: $\exp(\text{intercept})$ Δ à ne pas oublier la fonction Com et l'offset.

$$\mu = m e^{x'\beta}$$

⊕ les qualitatives

* Pour assurer de ref c'est $\exp(8,6621 + 0,0251 \text{bluebk} - 0,0006 \text{bluebk}^2)$

* Pour assurer à 5 pts c'est $\exp(8,6621 + 0,0251 \text{bluebk} - 0,0006 \text{bluebk}^2 + 0,1138)$

$$8) \hat{\sigma}_i D = \text{sigme}(y_i - \hat{\mu}_i) \sqrt{d_i} \quad \text{où } D = \sum_{i=1}^m d_i$$

Les $\hat{\sigma}_i D$ doivent être, en valeur absolue, bcp + grand que 1 sinon le modèle fait à la déviance alors que l'on souhaite une faible déviance.

Ici ils sont centré sur l'axe des abscisses \Rightarrow globalement c'est plutôt bon.