

# Modèles linéaires généralisés

13 mai 2022

M1 Actuariat, année 2021 - 2022

*Durée : 2h*

*Une feuille, seulement recto, manuscrite est autorisée.*

**Toutes les réponses doivent être soigneusement justifiées.**

**Exercice 1** Soit  $Y$  une variable aléatoire Binomiale Négative,  $Y \sim BN(\mu, k)$ , 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 peut se mettre sous la forme exponentielle tout en spécifiant le paramètre de la moyenne  $\theta$ , le paramètre de dispersion  $\phi$ , les fonctions  $a$ ,  $b$  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  $Var(Y)$  (on utilisera les formules de calcul d'espérance et variance de la famille exponentielle).
- 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.
- f) Expliquer dans quel contexte nous préférerons utiliser la loi binomiale négative plutôt que la loi de Poisson et quelles seraient les conséquences si, dans le contexte évoqué, on utilisait, à tort, une Poisson.
- g) Et quelle est l'utilité d'un modèle quasi-Poisson ?
- h) Nous avons estimé un modèle log-Binomiale Négative pour le nombre de sinistres. Quelles sont les étapes restantes en vu du calcul de la prime pure en assurance auto ? Les expliciter et donner la formule qui aurait permis de calculer la prime pure. Pourquoi ne pas utiliser un simple modèle de régression linéaire ?
- i) 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 des données issues d'une complémentaire santé individuelle proposant 5 formules de garanties aux assurés. Nous disposons des variables suivantes :

Variable	Descriptif	Modalités
conso	consommation en santé d'un assuré	
age	tranches d'âge de l'assuré	[0,5[, [5,10[, ..., 65 ans et +
formule	niveau de garantie choisi	A, B, C, D, E
zone	zone géographique	0, 1, 2, 3, 4
lien	lien entre assuré et bénéficiaire	A(Autres), C(Conjoint), E(Enfant), P(Assuré principal)

Nous cherchons à expliquer la variable *conso*. Grâce à la procédure PROC GENMOD de SAS, nous avons estimé un modèle et nous avons pu obtenir les résultats présentés dans le Tableau ci-dessous :

Distribution	Gaussienne inverse					
Link Function	Log					
Dependent Variable	conso					
Critere			DF	Valeur	Valeur/DF	
Deviance			37E4	5894.8172	0.0157	
Scaled Deviance			37E4	374317.0002	1.0001	
Pearson Chi-Square			37E4	9557.2328	0.0255	
Scaled Pearson X2			37E4	606877.9693	1.6214	
Log Likelihood				-2616728.458		
Parametre	DF	Estimation	Erreurs standard	Wald 95% Limites de confiance %	Khi 2	Pr > Khi 2
Intercept	1	7.0961	0.0290	7.0393 7.1259	59999.1	<.0001
age [0,5[	1	-0.4750	0.0326	-0.5388 -0.4112	212.90	<.0001
age [5,10[	1	-0.9237	0.0320	-0.9864 -0.8609	832.31	<.0001
age [10,15[	1	-0.7447	0.0327	-0.8088 -0.6806	518.31	<.0001
age [15,20[	1	-0.7506	0.0321	-0.8136 -0.6877	546.59	<.0001
age [20,25[	1	-0.8462	0.0251	-0.8955 -0.7970	1134.27	<.0001
age [25,30[	1	-0.7654	0.0243	-0.8134 -0.7175	978.48	<.0001
age [30,35[	1	-0.6801	0.0247	-0.7285 -0.6316	757.18	<.0001
age [35,40[	1	-0.6461	0.0245	-0.6942 -0.5980	693.51	<.0001
age [40,45[	1	-0.5448	0.0250	-0.5938 -0.4959	475.45	<.0001
age [45,50[	1	-0.4330	0.0257	-0.4834 -0.3827	284.04	<.0001
age [50,55[	1	-0.3128	0.0264	-0.3645 -0.2610	140.37	<.0001
age [55,60[	1	-0.2042	0.0261	-0.2553 -0.1530	61.24	<.0001
age [60,65[	1	-0.1346	0.0259	-0.1853 -0.0839	27.04	<.0001
age [65 et +	0	0.0000	0.0000	0.0000 0.0000	.	.
formule A	1	-1.1180	0.0195	-1.1563 -1.0798	3282.71	<.0001
formule B	1	-0.9824	0.0194	-1.0205 -0.9443	2554.44	<.0001
formule C	1	-0.7697	0.0198	-0.8086 -0.7309	1508.02	<.0001
formule D	1	-0.5184	0.0203	-0.5582 -0.4786	651.75	<.0001
formule E	0	0.0000	0.0000	0.0000 0.0000	.	.
zone 0	1	0.2041	0.0212	0.1625 0.2457	93.28	<.0001
zone 1	1	0.0424	0.0107	0.0214 0.0633	15.71	<.0001
zone 2	1	-0.0132	0.0126	-0.0378 0.0115	1.10	0.2952
zone 3	1	-0.0733	0.0093	-0.0916 -0.0551	61.96	<.0001
zone 4	0	0.0000	0.0000	0.0000 0.0000	.	.
lien A	1	-0.2161	0.0541	-0.3221 -0.1101	15.96	<.0001
lien C	1	-0.0994	0.0117	-0.1223 -0.0766	72.63	<.0001
lien E	1	-0.2782	0.0215	-0.3203 -0.2360	167.10	<.0001
lien P	0	0.0000	0.0000	0.0000 0.0000	.	.
Scale	1	0.1255	0.0001	0.1252 0.1258		

NOTE: The scale parameter was estimated by maximum likelihood.

## Statistiques LR pour Analyse de Type 1

Source	vraisemblance	DF	Khi 2	Pr > Khi 2
Intercept	-5251164.9			
age	-5242795.5	13	7638.47	<.0001
formule	-5233946.7	4	8848.77	<.0001
zone	-5233651.6	4	295.13	<.0001
lien	-5233430.0	3	221.59	<.0001

## Statistiques LR pour Analyse de Type 3

Source	DF	Khi 2	Pr > Khi 2
age	13	3931.99	<.0001
formule	4	6794.23	<.0001
zone	4	294.79	<.0001
lien	3	221.59	<.0001

- 1) Commentez globalement les résultats du modèle à la fois en terme d'ajustement du modèle aux données (à noter que  $\chi^2_{370000,0.95} = 371416$ ) et en terme de significativité des variables explicatives du modèle. Que proposez-vous afin d'améliorer le modèle ?
- 2) Comment expliquer la différence de valeur entre Deviance et Scaled Deviance dans le tableau ci-dessus ?
- 3) Calculer la consommation moyenne en santé d'un assuré dans la tranche d'âge [20, 25[ ayant choisi le niveau de garantie  $E$  et résidant dans la zone 1. Combien vaut  $n$ , taille de l'échantillon ?
- 4) Nous avons choisi une loi Gaussienne inverse pour modéliser la variable  $conso$ . Quelles auraient été les alternatives ? En quel cas aurait-il été pertinent d'utiliser la loi Tweedie au lieu de la loi Gaussienne inverse ?
- 5) Préciser quelle approche (parmi les trois vues en cours) est utilisée par SAS pour les analyses de type 1 et 3. Décrire ensuite la procédure de construction de la statistique du test. Pourquoi nous n'obtenons pas la même valeur de la statistique du test pour la variable  $âge$  quand on effectue une analyse de type 1 et une analyse de type 3 alors que nous obtenons la même valeur pour la variable  $lien$  ?
- 6) Dans le graphique suivant nous représentons les résidus de déviance contre les  $\hat{\mu}_t$ .

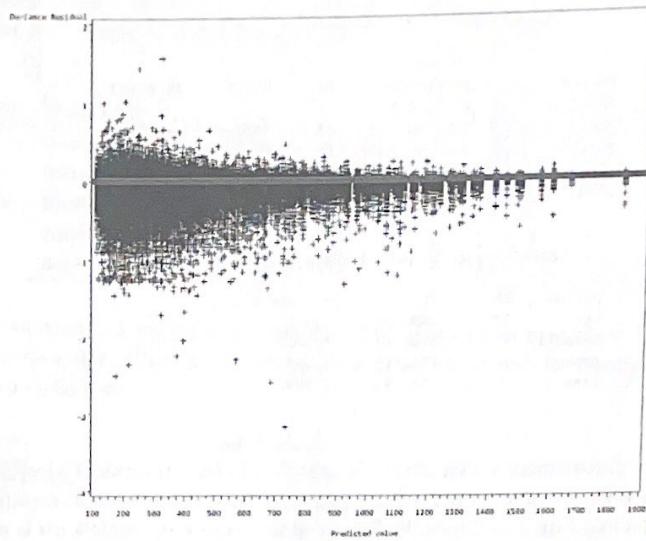


Fig. 1. Résidus de déviance

Les résidus de ce modèle sont-ils satisfaisants ? Oui ? Non ? Pourquoi ?

- 7) Nous souhaitons maintenant expliquer la variable *formule* (niveau de garantie choisi par l'assuré). Quel type de modèle proposez-vous ? Le décrire et en expliquer le fonctionnement.

### Exercice 1: $Y \sim \text{BN}(\mu, k)$

© Théo Jalabert

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

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

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

$$b(\theta) = -\frac{1}{k} \ln(1-e^\theta)$$

$$a(\phi) = 1, \phi = 1$$

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

Donc la loi binomiale négative fait bien partie de la famille exponentielle.

$$b) \theta = \ln\left(\frac{k\mu}{1+k\mu}\right) \rightarrow \text{fonction logitique}$$

$$\begin{aligned} \sqrt{(\mu)} &= b'(\theta) \\ &= \frac{1}{k} \frac{e^\theta}{(1-e^\theta)^2} \\ &= \frac{1}{k} \frac{\frac{k\mu}{1+k\mu}}{\left(\frac{1-k\mu}{1+k\mu}\right)^2} = \mu(1+k\mu) \end{aligned}$$

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

$$\text{et } \text{Var}(Y) = \frac{1}{a(\phi)} \text{Var}(\mu) \\ = \text{Var}(\mu) = \mu(1+k\mu)$$

$$d) L_{\text{SAT}} = \prod_{i=1}^m f(y_i) = \exp\left[\sum_{i=1}^m \left(y_i \ln\left(\frac{k\mu_i}{1+k\mu_i}\right) + \frac{1}{k} \ln\left(\frac{1}{1+k\mu_i}\right) + c(y_i, \phi)\right)\right] \text{ modèle saturé} \Rightarrow \mu_i = y_i$$

$$L = \prod_{i=1}^m f(y_i) = \exp\left[\sum_{i=1}^m \left(y_i \ln\left(\frac{k\hat{\mu}_i}{1+k\hat{\mu}_i}\right) + \frac{1}{k} \ln\left(\frac{1}{1+k\hat{\mu}_i}\right) + c(y_i, \phi)\right)\right] \text{ modèle estimé} \Rightarrow \mu_i = \hat{\mu}_i$$

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

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

$$= 2 \sum_{i=1}^m \left[ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{k}\right) \ln\left(\frac{y_i + \frac{1}{k}}{\hat{\mu}_i + \frac{1}{k}}\right) \right].$$

e)  $\hat{\sigma}_i^D = \text{Signe}(y_i - \hat{\mu}_i) \sqrt{d_i}$

$$= \text{Signe}(y_i - \hat{\mu}_i) \sqrt{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{k}\right) \ln\left(\frac{y_i + \frac{1}{k}}{\hat{\mu}_i + \frac{1}{k}}\right)}$$

f) Le problème avec la loi de Poisson est que l'espérance est égale à la variance alors que souvent, empirique on observe une variance empirique supérieure à la moyenne empirique (on parle alors de phénomène de surdispersion).

Si on ne prend pas en compte la surdispersion, on risque de sous-estimer la variance des estimateurs ce qui entraîne une surestimation de la valeur de la statistique du test du  $\chi^2$  de significativité.

→ On se déplace donc vers la région de rejet du test. Cela veut dire qu'on juge significative une variable qui ne l'est probablement pas.

Donc, s'il y a surdispersion, on utilise Y-BN et pas Poisson.

g) Avec une quasi-Poisson,  $\hat{\beta}$  est le même que l'on aurait obtenu si on avait estimé une régression de Poisson (et qu'on n'avait pas pris en compte la surdispersion) mais les écarts-types sont "gonflés" ce qui permet de résoudre les problèmes énoncés ci-dessus.

h) Principe Pur :  $E[S] = E[Y]E[N]$   
 Coût des sinistres       $\xrightarrow{nb}$  nb de sinistres

Prochaines étapes : Estimer un modèle pour le coût des sinistres → MODELE POUR VARIABLE RÉPONSE CONTINUE  
 (Régression Gamma/Gaussianne Inverse/Tweaked)

Car ici : Y et N ne sont pas Gaussien  $\Rightarrow$  impossible d'appliquer un modèle de régression classique.

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

$$\Rightarrow h(\mu) = h(m) + x'\beta$$

variable offset

## Exercice 2:

© Théo Jalabert

$$1) D_{\text{obs}} = 374317,0002 \text{ et } \chi^2_{370000, 0,95} = 371416$$

$$\Rightarrow D_{\text{obs}} > \chi^2_{370000, 0,95} \Rightarrow \text{on rejette le modèle.}$$

Toutes les variables sont signif sauf Zone<sub>2</sub> p.value = 0,2952 > 0,05

Or le modèle est faux, il surestime donc il rend "artificiellement" toutes les variables signif.

→ Variable offset pour tenir compte de la taille des <sup>les</sup> zones.

2) En SAS, Deviance correspond à la déviance non réduite, ici  $D^* = 5894,8172$   
et Scaled Deviance correspond à la déviance, ici  $D = 374317,0002$

$$D^* = \phi D \quad \Rightarrow \phi = \frac{D^*}{D} = 0,015748$$

↑  
Coeff de dispersion

$$3) \text{Comme moyenne} = \exp(7,0961 - 0,8/62 + 0,0/24) \\ = 540,395$$

$$\begin{aligned} m &= 370000 + \rho + 1 \\ &= 370000 + 24 + 1 \\ &= 370025 \end{aligned}$$

4) Alternatives : Poisson / Gamma / Tweedie.

Entre la Gaussienne Inverse et la Gamma, on préfère la Gaussienne Inverse en cas d'asymétrie plus marquée.

La Tweedie est en général utilisée lorsque l'on souhaite modéliser directement la charge totale (et donc ne pas estimer 2 modèles, un pour le nb de sinistres et un pour le montant des sinistres) car par exemple on ne dispose pas des coûts individuels mais seulement de la charge globale sur l'année.

5) Le Test du rapport de vraisemblance est utilisé par SAS pour les analyses de type 1 et 3

Le rapport de vraisemblance est défini comme:  $\lambda = \frac{\tilde{L}}{L}$   
avec  $\tilde{L}$  et  $L$  les vraisemblances des modèles sans et avec contraintes.

La stat du test du rapport est:  $2 \ln(\lambda) = 2(\tilde{L} - L) \stackrel{\text{sous}}{\sim} \chi^2_q$  avec  $q$  le nombres de lignes de la matrice C.  
 $H_0: CB = r$ .

Entre analyse type 1 et type 3, stat du test  $\neq$  car:

© Théo Jalabert



Analyse type 1: Test rapport de vraisemblance, analyse séquentielle

Analyse type 3: Test de Significativité avec toutes les autres variables.

6) Non il me sont pas satisfaisants car pour l'être leur répartition autour de l'axe des absisses devrait former un cylindre.  
Et ici, entourer.

7) modèle à variable réponse ordinaire