

Modèles linéaires généralisés

12 mai 2021

M1 Actuariat, année 2020 - 2021

Durée : 1h30

*Une feuille, seulement recto, manuscrite est autorisée.
Toutes les réponses doivent être soigneusement justifiées.*

Exercice 1 Soit Y une variable aléatoire binomiale, $Y \sim B(n, \pi)$.

- Montrer que la loi Binomiale peut se mettre sous la forme exponentielle en précisant le paramètre de la moyenne θ , le paramètre de dispersion ϕ , les fonctions a , b et c .
 - Trouver la fonction lien canonique ainsi que la fonction variance $V(\mu)$.
 - Vérifier que l'on trouve bien $E(Y) = n\pi$ et $Var(Y) = n\pi(1 - \pi)$ en utilisant les formules de calcul d'espérance et variance de la famille exponentielle.
 - Calculer la déviance et rappeler l'idée qui se cache derrière la construction de cet indicateur.
- diff avec modèle salué Diff partie 1*

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		
Critère	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	Erreur 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	2*Log-		
		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.
- 2) Calculer la consommation moyenne en santé d'un assuré dans la tranche d'âge [45, 50[ayant choisi le niveau de garantie A et résidant dans la zone 2. Combien vaut n , taille de l'échantillon ?
- 3) Comment expliquer la différence de valeur entre Deviance et Scaled Deviance dans le tableau ci-dessus ?

- 4) Que reste-il à faire pour améliorer le modèle ?
- 5) Y a-t-il un ou des effets d'interaction qu'il aurait été pertinent de tester ?
- 6) Nous avons choisi une loi Gaussienne inverse. Quelles auraient été les alternatives ?
En quel cas aurait-il été pertinent d'utiliser la loi Tweedie au lieu de la loi Gaussienne inverse ?
continue - non catégorielle
- 7) 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* ?
- 8) Dans le graphique suivant nous représentons les résidus de déviance contre les $\hat{\mu}_i$.

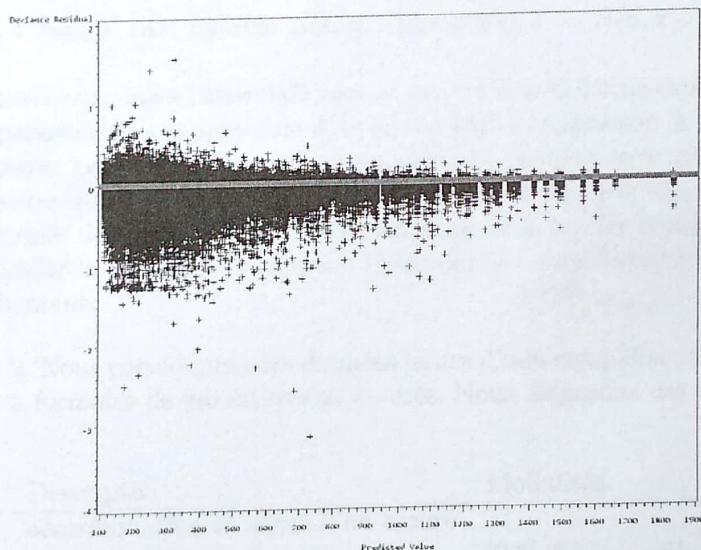


Fig. 1. Résidus de déviance

Non il faut
faire cylindrigat
autour de
ici entenir

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

- 9) Cette étude a été effectuée afin de constituer un tarif. Quelles auraient été les étapes supplémentaires dans la constitution du tarif ? Les expliciter. Donner la formule qui aurait permis de calculer la prime pure. $E[\mathbf{x}] = E[\mathbf{X}] E[\mathbf{S}]$
- 10) 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.

$$f(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$

$$\ln(f(y)) = \ln\left(\binom{m}{y}\right) + y \ln(\pi) + (m-y) \ln(1-\pi)$$

$$= y \ln\left(\frac{\pi}{1-\pi}\right) + m \ln(1-\pi) + \ln\left(\binom{m}{y}\right)$$

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

$$b(\theta) = -m \ln(1-\pi) = m \ln\left(1 - \frac{e^\theta}{1+e^\theta}\right) \\ = m \ln\left(1 + e^{-\theta}\right)$$

$$a(\phi) = 1$$

$$\phi = 1$$

$$c(y, \phi) = \ln\left(\binom{m}{y}\right)$$

Exercice 1: $Y \sim \mathcal{B}(n, \pi) \Rightarrow P(Y=y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \quad y \in \{0, n\}$

© Théo Jalabert

a) $h(P(Y=y)) = y h(\pi) + (n-y) h(1-\pi) + h(\binom{n}{y})$

$$\Rightarrow P(Y=y) = \exp[y h(\frac{\pi}{1-\pi}) + n h(1-\pi) + h(\binom{n}{y})]$$

$$\theta = h(\frac{\pi}{1-\pi}) \Rightarrow 1+e^\theta = \frac{1}{1-\pi}$$

$$\Rightarrow b(\theta) = -n h(1-\pi) = n h(1+e^\theta)$$

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

$$c(y, \phi) = h(\binom{n}{y})$$

b) $\theta_i = h(\frac{\pi_i}{1-\pi_i}) \Rightarrow$ fonction logistique

$$V(\mu) = b''(\theta) = \frac{m e^\theta}{(1+e^\theta)^2} = \frac{m \frac{\pi}{1-\pi}}{(\frac{1}{1-\pi})^2} = m \pi (1-\pi) \Rightarrow$$

$$\text{Loi binomiale } \in \mathcal{F}_{\text{exp}}$$

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

$$\text{Var}(Y) = a(\phi) b''(\theta) = 1 \times V(\mu) = m \pi (1-\pi)$$

d) $D = 2 [h(s_{\hat{\pi}}) - h(L)]$

$$s_{\hat{\pi}} = \prod_{i=1}^n P(Y=y_i) = \exp \left[\sum_{i=1}^n (y_i h(\frac{y_i}{1-y_i}) + m h(1-y_i) + h(\binom{n}{y_i})) \right] \text{ modèle saturé} \Rightarrow y_i = \pi_i$$

$$L = \prod_{i=1}^n P(Y=y_i) = \exp \left[\sum_{i=1}^n (y_i h(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + m h(1-\hat{\pi}_i) + h(\binom{n}{y_i})) \right] \text{ modèle estimé} \Rightarrow \pi_i = \hat{\pi}_i$$

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

Exercice 2:

1) $D_{\text{obs}} = 374317,0002$ 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 Zome2 p.value = 0,2952 > 0,05

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

2) Compte moyenne = $\exp(7,0361 - 0,6330 - 1,1180 - 0,0132)$
= 252,62

La taille de l'échantillon est $n = 370000 + p + 1 = 370000 + 24 + 1 = 370025$

3) En SAS, Deviance correspond à la déviance non réduite, ici $D^* = 5894,8172$
et Scaled Deviance correspond à la déviance, ici $D = 37,317,0002$

© Théo Jalabert

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

↑
Coeff de dispersion.

4) \rightarrow Variable offset pour tenir compte de la taille des \neq^{res} zones...

5) Il aurait pu être pertinent de tester les interactions entre l'âge et le niveau de garantie choisi tout comme l'interaction entre niveau de garantie et zone géo...

6) Alternatives : loi 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.

7) 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: $2 \ln(\lambda) = 2(\tilde{L} - L) \stackrel{\text{sous }}{\sim} \chi_q^2$ avec q le nombres de lignes de la matrice C .
 $H_0: C\beta = r$.

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

Analyse type 1: Test rapport de vraisemblance, analyse séquentielle

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

8) Non il me semblerait pas satisfaisants car pour l'être leur répartition autour de l'axe des absences devrait former un cylindre.
Or ici, entasser.

9) IP faut maintenant estimer un modèle pour le nb de sinistres.

Prisme pur dans un modèle collectif: $E[S] = E[Y]E[N]$

Coût du sinistre nb de sinistres.
/caso

10) modèle à variable réponse ordinaire

© Théo Jalabert

