

Modèles linéaires généralisés

14 mai 2019

M1 Actuariat, année 2018-2019

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

- 1) On considère la variable aléatoire Y de loi gaussienne inverse, notée $Y \sim IG(\mu, \sigma^2)$, $\mu > 0$, $\sigma^2 > 0$ de densité :

$$f(y) = \frac{1}{\sqrt{2\pi y^3} \sigma} \exp \left\{ -\frac{1}{2y} \left(\frac{y-\mu}{\mu\sigma} \right)^2 \right\} \quad y > 0$$

- a) Montrer que la densité d'une loi gaussienne inverse peut se mettre sous la forme 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) Montrer que la déviance est donnée par $D = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}$.
 - d) Écrire les résidus de Pearson. Que faire avec ces résidus ?
 - e) En quel cas est-il raisonnable de choisir la loi IG pour Y ?
- 2) Maintenant supposons que Y pourrait être ajustée par une loi lognormale, i.e. $Y \sim LN(\mu, \sigma)$ de densité
- $$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(\ln y - \mu)^2}{2\sigma^2} \right\} \quad y > 0$$
- i) La loi lognormale appartient-elle à la famille exponentielle ? Oui ? Non ? Le montrer.
 - ii) Quel modèle estimer pour prendre en compte un coût de sinistre lognormale ?
- 3) On cherche maintenant à expliquer le choix par un assuré du niveau de garantie en assurance santé au moyen d'un ensemble de variables explicatives. Quel type de modèle proposeriez-vous ? L'écrire et commenter.

Exercice 2 Nous considerons un portefeuille d'assurance RC Automobile en Australie. On dispose du nombre de sinistres (**claims**) observées durant une période de 12 mois entre 1984 et 1986 dans les 176 zones géographiques de la Nouvelle-Galles du Sud, Australie. Les zones géographiques sont groupées en 13 départements (**sd**). On dispose aussi du nombre d'accidents (**accidents**) ainsi que de la taille de la population (**population**) pour chaque zone géographique. On note **logpop** le logarithme de la variable **population** et **logacc** le logarithme de la variable **accidents**.

Grâce à la procédure PROC GENMOD de SAS nous avons estimé un modèle avec loi Binomiale Négative (BN) et un modèle Quasi Poisson dont les résultats sont présentés dans les tableaux ci-dessous.

Modèle 1 :

Distribution	Negative Binomial				
Link Function	Log				
Dependent Variable	claims				
Offset Variable	logpop				
Critere	DF	Valeur	Valeur/DF		
Deviance	162	193.1413	1.1922		
Scaled Deviance	162	193.1413	1.1922		
Pearson Chi-Square	162	219.7335	1.3564		
Scaled Pearson X2	162	219.7335	1.3564		
Log Likelihood		651894.2334			
BIC		2082.8954			
Parametre	DF	Estimation	Erreur		
			standard		
Intercept	1	-6.2331	0.3165		
sd	1	-0.5863	0.2539		
sd	2	-0.7193	0.2938		
sd	3	-0.5731	0.2540		
sd	4	-0.7031	0.2909		
sd	5	-0.7517	0.2712		
sd	6	-0.5282	0.2569		
sd	7	-0.9872	0.2454		
sd	8	-0.5123	0.3570		
sd	9	-0.5986	0.2403		
sd	10	-0.4272	0.2484		
sd	11	-0.5368	0.2512		
sd	12	-0.7862	0.2525		
sd	13	0	0.0000		
logacc	1	0.2369	0.0397		
Dispersion	1	0.1424	0.0167		
			Wald 95 Limites		
			de confiance %		
			Khi 2		
			Pr > Khi 2		
Intercept	1	-6.2331	-6.8536 -5.6127	387.74	<.0001
sd	1	-0.5863	-1.0840 -0.0886	5.33	0.0209
sd	2	-0.7193	-1.2952 -0.1434		0.0144
sd	3	-0.5731	-1.0709 -0.0752	5.09	0.0241
sd	4	-0.7031	-1.2732 -0.1331	5.84	0.0156
sd	5	-0.7517	-1.2832 -0.2201	7.68	0.0056
sd	6	-0.5282	-1.0318 -0.0247	4.23	0.0398
sd	7	-0.9872	-1.4680 -0.5063	16.19	<.0001
sd	8	-0.5123	-1.2120 0.1875	2.06	0.1513
sd	9	-0.5986	-1.0696 -0.1276	6.20	0.0127
sd	10	-0.4272	-0.9140 0.0596	2.96	0.0854
sd	11	-0.5368	-1.0292 -0.0444	4.57	0.0326
sd	12	-0.7862	-1.2810 0.2913	9.70	0.0018
sd	13	0	0.0000 0.0000	.	.
logacc	1	0.2369	0.0397 0.1590	35.56	<.0001
Dispersion	1	0.1424	0.0167 0.1131	0.1792	<i>Ici $\phi=1$ car</i>

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis				
Source	2*Log-vraisemblance	DDL	Khi-2	Pr > Khi-2
Intercept	1303674.84			
sd	1303754.87	12	80.02	<.0001
logacc	1303788.47	1	33.60	<.0001

LR Statistics For Type 3 Analysis			
Source	DDL	Khi-2	Pr > Khi-2
sd	12	29.92	0.0029
logacc	1		<.0001

Modèle 2 :

Distribution	Poisson		
Link Function	Log		
Dependent Variable	claims		
Offset Variable	logpop		

Critere	DF	Valeur	Valeur/DF
Deviance	162	14274.3008	88.1130
Scaled Deviance	162	162.0000	1.0000
Pearson Chi-Square	162	15534.8815	95.8943
Scaled Pearson X2	162	176.3064	1.0883
Log Likelihood		7321.8155	
BIC		15572.3957	

Parametre	DF	Estimation	standard	Erreur		Wald	95 Limites de confiance %	Khi 2	Pr > Khi 2
Intercept	1	-6.2328	0.4284	-7.0724	-5.3931	211.69	<.0001		
sd	1	1	-1.0174	0.3195	-1.6437	-0.3911	10.14	0.0015	
sd	2	1	-1.1400	0.3345	-1.7956	-0.4843	11.61	0.0007	
sd	3	1	-1.1319	0.3277	-1.7742	-0.4896	11.93	0.0006	
sd	4	1	-1.2479	0.3402	-1.9147	-0.5811	13.45	0.0002	
sd	5	1	-1.2250	0.3775	-1.9649	-0.4851	10.53	0.0012	
sd	6	1	-0.9049	0.3431	-1.5773	-0.2324	6.96	0.0084	
sd	7	1	-1.2411	0.3779	-1.9818	-0.5004	10.78	0.0010	
sd	8	1	-0.7279	0.6963	-2.0926	0.6367	1.09	0.2958	
sd	9	1	-0.8366	0.3358	-1.4948	-0.1783	6.21	0.0127	
sd	10	1	-0.8950	0.3625	-1.6055	-0.1846	6.10	0.0135	
sd	11	1	-0.9350	0.3673	-1.6549	-0.2151	6.48	0.0109	
sd	12	1	-1.1593	0.4191	-1.9807	-0.3380	7.65	0.0057	
sd	13	0	0.0000	0.0000	0.0000	0.0000	.	.	.
logacc	1	0.2811	0.0501	0.1830	0.3793	31.50	<.0001		
Scale	0	9.3869	0.0000	9.3869	9.3869				

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

- 1) Écrire le modèle 1 et expliquer la présence de la variable offset. N'aurons-nous pas pu simplement inclure la variable population dans le modèle en tant que variable explicative ? Pourquoi l'inclure en tant que variable offset ?
- 2) Commentez globalement les résultats du modèle 1 à la fois en terme d'ajustement du modèle aux données et en terme de significativité des variables explicatives du modèle.
- 3) Pourquoi observons-nous une ligne de zéros en correspondance de la modalité 13 de la variable sd ?
- 4) Pour la modalité 2 de la variable sd calculer la valeur de la statistique du test du χ^2 . Quelle approche parmi les trois vues en cours a été utilisée pour construire la statistique du test ? La présenter.
- 5) On considère maintenant l'analyse de type 3. Sans faire le calcul mais en expliquant la raison de ce résultat, donner la valeur de la statistique du test du χ^2 pour la variable logacc.
- 6) Le modèle 2 est un modèle Quasi Poisson (on maximise une quasi-vraisemblance avec fonction variance de Poisson). Pourquoi a-t-on choisi d'estimer ce genre de modèle ?
- 7) Pourquoi le scale parameter n'a pas été estimé par maximum de vraisemblance comme dans le cas de la BN ?

Nous souhaitons maintenant comparer les résultats obtenus avec un modèle Quasi Poisson aux résultats que l'on aurait obtenu en estimant un modèle log-Poisson, partiellement présentés dans le tableau ci-dessous :

Modèle 3 :

Distribution	Poisson
Link Function	Log
Dependent Variable	claims
Offset Variable	logpop

Critere	DF	Valeur	Valeur/DF
Deviance	162	14274.3008	88.1130
Scaled Deviance	162	14274.3008	88.1130

*Valeurs de la stat de
rest du X² de signif.*

Parametre	DF	Estimation	Erreur standard	Wald 95% Limites de confiance %		Khi 2	Pr > Khi 2
				-	-		
Intercept	1	-6.2328	0.0456	-6.3232	-6.1433	18652.8	<.0001
sd 1	1	-1.0174	0.0340	-1.0841	-0.9507	893.18	<.0001
sd 2	1	-1.1400	0.0356	-1.2098	-1.0701	1023.18	<.0001
sd 3	1	-1.1319	0.0349	-1.2003	-1.0635	1051.24	<.0001
sd 4	1	-1.2479	0.0362	-1.3189	-1.1768	1185.53	<.0001
sd 5	1	-1.2250	0.0402	-1.3038	-1.1462	927.81	<.0001
sd 6	1	-0.9049	0.0366	-0.9765	-0.8332	612.89	<.0001
sd 7	1	-1.2411	0.0403	-1.3200	-1.1622	950.28	<.0001
sd 8	1	-0.7279	0.0742	-0.8733	-0.5825	96.31	<.0001
sd 9	1	-0.8366	0.0358	-0.9067	-0.7664	546.76	<.0001
sd 10	1	-0.8950	0.0386	-0.9707	-0.8193	537.17	<.0001
sd 11	1	-0.9350	0.0391	-1.0117	-0.8583	570.98	<.0001
sd 12	1	-1.1593	0.0446	-1.2469	-1.0718	674.31	<.0001
sd 13	0	0.0000	0.0000	0.0000	0.0000	.	.
logacc	1	0.2811	0.0053	0.2707	0.2916	2775.40	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

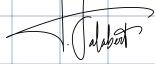
- 8) Comment expliqueriez-vous le mauvais ajustement de ce modèle ?
- 9) En quoi les résultats du modèle de Poisson et du modèle Quasi Poisson sont différents ? En quoi sont-ils similaires ? Êtes-vous étonnés par la significativité de toutes les variables explicatives du modèle de Poisson ? Oui ? Non ? Pourquoi ?

Finalement nous décidons de retenir le modèle 1.

- 10) Que faire pour les modalités 8 et 10 de la variable sd ?
- 11) Supposons maintenant que la compagnie ait adopté un mécanisme de bonus-malus et que cela ait incité un certain nombre d'assurés à ne pas déclarer certains sinistres. Nous observons donc des "zéros structurels". Quel type de modèle proposeriez-vous afin d'améliorer les résultats du modèle BN ? L'écrire et en expliquer le principe.

Exercice 1:

© Théo Jalabert



1) a) Classique

b) Passer par def de fonction Liem / Variance. $V(\mu) = b''(0)$

c) Partir de la def $D = 2(p_{SAT} - p)$

d) Residus de Pearson (Shade 26) Que faire? esp=0 et homoscedastique
Centre

On les représente et on s'attend à observer des résidus centrés sur l'axe 0 des abscisses et répartis de part et d'autre de l'axe

e) Gaussienne inverse: quand la var à expliquer est quantitative continue
prononcée
→ Schima.

et asymétrique sym assez

2) i) Non car on n'arrive pas à la mettre sous forme exp

ii) Lognormale ~~GCM~~

⇒ Lognormale → On prend le log ⇒ Va gaussienne.

des coûts de
Simplicité

On estime un modèle gaussien sur le log des coûts.

3) Va quantitative avec modalités A, B, C, D, E sans ordre entre modalités
⇒ Modèle de régress° nominal.

Si ya ordre → Modèle de variable réponse ordinaire

Exo 2:

1) Va à expliquer claims fact° Lem: log

→ Modèle: $\log(E[Claims]) = \log pop + \beta_0 + \beta_1 sd_1 + \dots + \beta_{12} sd_{12} + \beta_3 \log acc$

↑
-6,2328

taille échantillon: 176 zones geo

Y a offset pour prendre en compte que les + zones ont des tailles de pop + ⇒ nb moyes de similitude proportionnelles à la taille de la pop.

Logpop pas en va explicative car on enlève l'estimat° d'un β supplémentaire.

© Théo Jalabert

TJ

2) tout est significatif sauf $\beta_{el10} \Rightarrow$ globalement OK

modélisation de la va sd $\geq 5\%$

Qualité d'ajustement \rightarrow regarder deviance

Rappel en SAS Scale Deviance - Deviance

$$162 \text{ ddl} = n - p - 1$$

$I_{ci} = 1,13$ pas trop éloigné de 1 \Rightarrow ajustement convaincant

mais on pourrait comparer 193,1613 avec
la quantile à 162 ddl.

$$\text{Dobs} = 193,1613 > \chi^2_{162, 0,95} = 192 \\ \Rightarrow \text{ajustement mauvais.}$$

3) Car la modéliser β_0 de la va β est la référence

4) $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$ test de significat du χ^2 de Wald.

Stat du test du χ^2 : $\frac{\hat{\beta}_2^2}{\text{Var}(\hat{\beta}_2)} = \left(-\frac{0,7193}{0,2938} \right)^2$

L'approche utilisée est celle de Wald parmi (rapport de vraisemblance, Wald, Score)

$$H_0: C\beta = r \\ r=0 \quad C = (001 \quad 0 \quad 0)$$

5) 33,60 là où que l'analyse de type 1 car c'est la dernière va qui est testée.

Analyse de type 3: logacc: $H_0: g(\mu) = \text{Logpop} + \beta_0 + \beta_1 \text{sd}$

$$H_1: g(\mu) = \text{Logpop} + \beta_0 + \beta_1 \text{sd} + \beta_2 \text{logacc.}$$

6) Il y a un pb de sous-dispersio des mes données (binomial négative) \rightarrow Quasi-Poisson

Var empirique $>$ Var théorique \Rightarrow Modèle binomial nég \rightarrow ici Mauvais résultats.

Quasi-vraisemblance

7) Il n'a pas de vraisemblance car on a pas fait d'hyp sur les de y.

8) Slide 35

9) Similaires en β mais écarts types + élevés qu'en log-poisson chp. value très petit car

on sous estime \Rightarrow faux.

© Théo Jalabert



Non car signif. faux.

le) 2^{INB}

Exercice 1.

1) a) $Y \sim IG(\mu, \sigma^2)$ $\mu > 0, \sigma^2 > 0$. Y appartient à la famille exponentielle si sa densité peut s'écrire:

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

$$\text{Ici } f(y) = \frac{1}{\sqrt{2\pi y^3 \sigma^2}} \exp \left\{ -\frac{1}{2y} \left(\frac{y-\mu}{\mu \sigma} \right)^2 \right\} \quad y > 0.$$

$$\Rightarrow \ln(f(y)) = -\frac{1}{2y} \left(\frac{y-\mu}{\mu \sigma} \right)^2 - \ln(\sqrt{2\pi y^3 \sigma^2})$$

$$\Rightarrow f(y) = \exp \left[-\frac{y}{2\mu^2 \sigma^2} + \frac{1}{\mu \sigma^2} - \frac{1}{2y \sigma^2} - \ln(\sqrt{2\pi y^3 \sigma^2}) \right]$$

Par identification:

$$\theta = -\frac{1}{2\mu^2} \text{ et } \alpha(\phi) = \sigma^2 \Rightarrow \phi = \sigma^2$$

$$b(\theta) = -\frac{1}{\mu} = -\sqrt{-2\theta} \Rightarrow b'(\theta) = -(-2 \times \frac{1}{2\sqrt{-2\theta}}) = \frac{1}{\sqrt{-2\theta}} = \mu$$

$$b''(\theta) = \frac{1}{(-2\theta)^{3/2}} = \frac{1}{(\sqrt{-2\theta})^3} = \mu^3$$

$$C(y, \phi) = -\frac{1}{2y\phi} - \ln(\sqrt{2\pi y^3 \phi})$$

D'où la gaussienne inverse appartient à la famille exponentielle

b) Fonction liée canonique : $\theta = g(\mu) = -\frac{1}{2\mu^2}$

Fonction Variance : $V(\mu) = b''(\theta) = \frac{1}{(-2\theta)^{3/2}} = \frac{1}{(\sqrt{-2\theta})^3} = \mu^3$

c) $D = 2(\ln(L_{\text{SAT}}) - \ln(L))$ $\hat{\mu}_i$ car modèle estimé

$$\begin{aligned} L &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi y_i^3 \sigma^2}} \exp \left\{ -\frac{1}{2y_i} \left(\frac{y_i - \hat{\mu}_i}{\mu \sigma} \right)^2 \right\} \\ &= (\sqrt{2\pi \sigma^2})^{-m} \left(\prod_{i=1}^m y_i^{-3/2} \right) \exp \left\{ \sum_{i=1}^m -\frac{1}{2y_i} \left(\frac{y_i - \hat{\mu}_i}{\mu \sigma} \right)^2 \right\} \end{aligned}$$

$$\Rightarrow \ln(L) = -m \ln(\sqrt{2\pi \sigma^2}) - \frac{3}{2} \sum_{i=1}^m \ln(y_i) - \frac{1}{2} \sum_{i=1}^m \frac{1}{y_i} \left(\frac{y_i - \hat{\mu}_i}{\mu \sigma} \right)^2$$

$$\text{De même, on a } h(L_{SAT}) = -m \ln(\sqrt{2\pi}\sigma) - \frac{3}{2} \sum_{i=1}^m h(y_i) - \frac{1}{2} \sum_{i=1}^m \frac{1}{y_i} (y_i - \mu_i)^2$$

© Théo Jalabert

Or ici il s'agit du modèle saturé $\Rightarrow y_i = \mu_i$

$$\Rightarrow h(L_{SAT}) = -m \ln(\sqrt{2\pi}\sigma) - \frac{3}{2} \sum_{i=1}^m h(y_i)$$

$$\Rightarrow D = 2(h(L_{SAT}) - h(L))$$

$$= 2(-m \ln(\sqrt{2\pi}\sigma) - \frac{3}{2} \sum_{i=1}^m h(y_i) + m \ln(\sqrt{2\pi}\sigma) + \frac{3}{2} \sum_{i=1}^m h(y_i) + \frac{1}{2} \sum_{i=1}^m \frac{1}{y_i} (y_i - \hat{\mu}_i)^2)$$

$$= \sum_{i=1}^m \frac{1}{y_i} \frac{1}{\mu_i^2 \sigma^2} (y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2} \sum_{i=1}^m \frac{(y_i - \hat{\mu}_i)^2}{\mu_i^2 y_i}$$

d) Les résidus de Pearson sont définis par :

$$r_i^P = \frac{\sqrt{\mu_i} (y_i - \hat{\mu}_i)}{\sqrt{V(\mu_i)}}$$

On remarque que ces résidus sont centrés et homoskedastiques.

On va donc les représenter et on s'attend à observer des résidus centrés sur l'axe des abscisses et répartis de part et d'autre de l'axe.

e) Il est raisonnable de choisir la loi LG pour Y lorsque la variable à expliquer est quantitative continue et asymétrique mais également si nos données présentent une espérance (μ) et variance (σ^2) telles que : $\sigma^2 = \lambda \mu^3$ où λ est un paramètre de dispersion.

$$2) Y \sim LG(\mu, \sigma) \text{ tq } f(y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{1}{2\sigma^2} (\ln(y) - \mu)^2} \quad y > 0$$

$$i) h(f(y)) = -h(\sigma) - h(y) - \frac{1}{2} h(2\sigma) - \frac{1}{2\sigma^2} (\ln(y))^2 - 2\mu \ln(y) + \mu^2$$

$$\Rightarrow f(y) = \exp\left(-\frac{\ln(y)^2}{2\sigma^2} + \frac{\mu \ln(y)}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - h(\sigma) - h(y) - \frac{1}{2} h(2\sigma)\right)$$

\Rightarrow La loi log normale n'appartient pas à la famille exponentielle puisqu'il n'est pas possible d'écrire sa densité sous la forme :

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

ii) Pour prendre en compte un coût de sinistre lognormal, on prend le log des coûts de sinistres ce qui nous fournit une variable gaussienne.

Puis on estime un modèle gaussien sur le log des coûts.

3) Supposons que les variables soient qualitatives avec des modalités (ex: modalités A, B, C, D et E) et qu'il n'y ait pas d'ordre entre les modalités.

Nous devrions alors utiliser un modèle de régression multinomiale.
S'il y avait un ordre, on utiliserait un modèle à variable réponse ordinaire.

→ Slide 43

Exercice 2:

1) Modèle 1: $\log(\text{IF[claims]}) = \log pop + \beta_0 + \beta_1 \text{sd}_1 + \dots + \beta_2 \text{sd}_{12} + \beta_3 \log acc$

La variable population n'est pas inclus en variable explicative car cela nous évite l'estimation d'un coeff supplémentaire.

On l'inclus en tant que variable offset car on doit prendre en compte que les 176 zones géographiques ont des tailles de populations différentes.

Cela nous permet donc d'estimer le nb moyens de sinistres proportionnellement à la taille de la population.

2) En regardant les valeurs de p.value, toutes les variables sont significatives à l'exception des modalités 8 et 10 de la variable sd (p.value égale à 15,13% > 5% (resp 8,54% > 5%).

En ce qui concerne la qualité d'ajustement, on s'intéresse à la déviance.

Ici, Valeur/DF ~ 1,19 → ajustement qui semble convenable

Pour être précis, on regarde $\chi^2_{162, 0.95} = 192$
 $\Rightarrow D_{obs} > \chi^2_{162, 0.95}$ \Rightarrow ajustement mauvais.

3) Car la modalité 13 de la variable sd est celle de référence.

4) $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$ test de significativité de Wald.

$$\text{Stat du test du } \chi^2: \frac{\hat{\beta}_2^2}{\text{Var}(\hat{\beta}_2)} = \left(\frac{-0,793}{0,2938} \right)^2 = 5,9940$$

L'approche utilisée parmi (rapport de Vraisemblance, Wald, Score) est celle de Wald.

$$H_0: C\beta = 0$$

$$z=0 \quad C = (0 \ 0 \ 1 \ 0 \ - \ 0)$$

5) Valeur de la stat du test = 33,60

© Théo Jalabert



Cette valeur est identique à celle calculée dans l'analyse de type 1 car c'est la dernière variable qui est testée.

Analyse de type 3: logacc: $H_0: g(\mu) = \log \text{pop} + \beta_0 + \beta_1 \text{sd}$

$$H_1: g(\mu) = \log \text{pop} + \beta_0 + \beta_1 \text{sd} + \beta_2 \log \text{acc}.$$

6) Il y a un problème de surdispersion de mes données dans le cas du modèle avec loi Binomiale Négative. Cela nous amène donc à estimer le modèle Quasi-Poisson.

Var empirique > moyenne empirique \Rightarrow {modèle binom. nég. \rightarrow mauvais résultats
Quasi-vraisemblance.}

7) Car il n'a pas de vraisemblance puisque l'on n'a pas fait d'hypothèses sur la loi de y .

8) Le mauvais ajustement de ce modèle s'explique par la surdispersion des observations qui n'est pas prise en compte dans ce modèle.

9) Les paramètres estimés sont les mêmes et la déviance également.

Les différences se situent dans l'estimation des écarts-types des paramètres. En effet, les erreurs du modèle de Poisson sont plus élevées compte tenu de la surdispersion des paramètres. La significativité de toutes les variables explicatives du modèle de Poisson s'explique justement par des valeurs élevées de la variance des paramètres qui poussent la stat de Wald vers la région de rejet de l'hypothèse de non-significativité des variables explicatives.

10) Ces modalités n'étant pas significatives par rapport à la modalité de référence, il conviendrait de les regrouper avec cette dernière.

11) Le modèle à zéro inflation est plus adapté.

Il s'écrit:

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

où π_i est la probabilité d'avoir un zéro structurel
et ici: $p_i(k) = \frac{e^{-\lambda_i} \lambda_i^k}{k!}$

Le principe est que le résultat observé est celui de deux processus : un premier qui génère des zéros structurels et un second qui génère des sinistres aléatoires.