Modèles linéaires généralisés

3 mai 2017

M1 Actuariat, année 2016-2017

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) Obtenir les équations de vraisemblance. Comment trouver $\hat{\beta}$? Montrer ensuite comment on peut re-écrire et interpréter les équations de vraisemblance en cas de fonction lien canonique.
ii) Décrire la procédure de construction du test de significativité de Wald après en avoir specifié les hypothèses nulle et alternative. Et dans un modèle de régression linéaire classique comment procède-t-on ?
iii) Présenter les résidus de déviance et en expliquer l'utilisation/interprétation. Ces résidus sont-ils normaux ?
iv) Comment détecter une valeur influente dans ce modèle ? Et dans un modèle de régression linéaire classique ?
v) On suppose maintenant que $Y \sim G(\mu, \nu)$ de densité :

$$f(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu}\right)^{\nu} e^{\frac{-y\nu}{\mu}} \qquad\qquad y > 0$$

   a) Montrer que la loi gamma 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 $Var(Y)$ ($E(Y) = \mu$ et $Var(Y) = \frac{\mu^2}{\nu}$).
   d) Montrer que la déviance est donnée par $D = 2\nu \sum_{i=1}^{n}\{-\ln(\frac{y_i}{\hat{\mu}_i}) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\}$.
   e) Écrire les résidus de déviance.
vi) Présenter un exemple pratique où on estimerais un modèle de régression nominale. Écrire ensuite le modèle.

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`), du nombre de personnes tuées ou blessées (`ki`), de la densité de population (`pop_density`) 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`.

La procédure PROC GENMOD de SAS nous a permis d'estimer plusieurs modèles linéaires généralisés dont les résultats sont présentés dans les tableaux ci-dessous.

<u>**Modèle 1**</u> :

Distribution	Poisson
Link Function	Log
Dependent Variable	claims
Offset Variable	logpop

Number of Observations Read	176
Number of Observations Used	176

Critere	DF	Valeur	Valeur/DF
Deviance	160	11782.7760	73.6424
Scaled Deviance	160	11782.7760	73.6424
Pearson Chi-Square	160	12905.0062	80.6563
Scaled Pearson X2	160	12905.0062	80.6563
Log Likelihood		646392.6605	
BIC		13091.2118	

Parametre		DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept		1	-4.7836	0.0650	-4.9110	-4.6563	5420.05	<.0001
sd	1	1	-1.0345	0.0351	-1.1033	-0.9656	867.01	<.0001
sd	2	1	-0.9349	0.0363	-1.0060	-0.8638	664.58	<.0001
sd	3	1	-1.0332	0.0352	-1.1023	-0.9641	859.16	<.0001
sd	4	1	-1.1917	0.0365	-1.2632	-1.1203	1068.55	<.0001
sd	5	1	-1.0869	0.0404	-1.1660	-1.0078	724.90	<.0001
sd	6	1	-0.7752	0.0367	-0.8471	-0.7032	446.19	<.0001
sd	7	1	-1.2232	0.0403	-1.3022	-1.1443	922.60	<.0001
sd	8	1	-0.7848	0.0742	-0.9303	-0.6393	111.78	<.0001
sd	9	1	-0.7953	0.0358	-0.8654	-0.7251	493.85	<.0001
sd	10	1	-0.8318	0.0386	-0.9075	-0.7561	463.34	<.0001
sd	11	1	-0.8740	0.0391	-0.9507	-0.7973	498.97	<.0001
sd	12	1	-1.1875	0.0447	-1.2751	-1.0999	705.49	<.0001
sd	13	0	0.0000	0.0000	0.0000	0.0000	.	.
logacc		1	0.0223	0.0097	0.0033	0.0413	5.32	0.0211
ki		1	0.0002	0.0000	0.0002	0.0002	1291.41	<.0001
pop_density		1	1.0847	0.0249	1.0360	1.1335	1901.98	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DDL	Khi-2	Pr > Khi-2
sd	12	1365.77	<.0001
logacc	1	5.33	0.0209
ki	1	1251.76	<.0001
pop_density	1	1848.16	<.0001

Modèle 2 :

Distribution	Negative Binomial
Link Function	Log
Dependent Variable	claims
Offset Variable	logpop

Critere	DF	Valeur	Valeur/DF
Deviance	160	193.5675	1.2098
Scaled Deviance	160	193.5675	1.2098
Pearson Chi-Square	160	221.2694	1.3829
Scaled Pearson X2	160	221.2694	1.3829
Log Likelihood		651895.1392	
BIC		2091.4249	

Parametre		DF	Estimation	Erreur standard	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept		1	-6.1827	0.3803	-6.928	-5.4373	264.25	<.0001
sd	1	1	-0.7091	0.2676	-1.2337	-0.1846	7.02	0.0081
sd	2	1	-0.7070	0.2926	-1.2805	-0.1334	5.84	0.0157
sd	3	1	-0.5656	0.2526	-1.0608	-0.0705	5.01	0.0252
sd	4	1	-0.6960	0.2892	-1.2628	-0.1292	5.79	0.0161
sd	5	1	-0.7392	0.2704	-1.2692	-0.2092	7.47	0.0063
sd	6	1	-0.5155	0.2561	-1.0174	-0.0136	4.05	0.0441
sd	7	1	-0.9804	0.2440	-1.4586	-0.5022	16.15	<.0001
sd	8	1	-0.5006	0.3550	-1.1964	0.1952	1.99	0.1585
sd	9	1	-0.5885	0.2391	-1.0570	-0.1199	6.06	0.0138
sd	10	1	-0.4165	0.2475	-0.9017	0.0687	2.83	0.0925
sd	11	1	-0.5278	0.2498	-1.0174	-0.0382	4.46	0.0346
sd	12	1	-0.7833	0.2515	-1.2763	-0.2903	9.70	0.0018
sd	13	0	0.0000	0.0000	0.0000	0.0000	.	.
logacc		1	0.2251	0.0582	0.1110	0.3392	14.95	0.0001
ki		1	0.0000	0.0001	-0.0001	0.0001	0.07	0.7945
pop_density		1	0.5420	0.4037	-0.2492	1.3333	1.80	0.1794
Dispersion		1	0.1404	0.0166	0.1114	0.1770		

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

LR Statistics For Type 3 Analysis

Source	DDL	Khi-2	Pr > Khi-2
sd	12	30.80	0.0021
logacc	1	14.70	0.0001
ki	1	0.07	0.7942
pop_density	1	1.81	0.1791

Modèle 3 :

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	Wald 95Limites de confiance %		Khi 2	Pr > Khi 2
Intercept		1	-6.2331	0.3165	-6.8536	-5.6127	387.74	<.0001
sd	1	1	-0.5863	0.2539	-1.0840	-0.0886	5.33	0.0209
sd	2	1	-0.7193	0.2938	-1.2952	-0.1434	5.99	0.0144
sd	3	1	-0.5731	0.2540	-1.0709	-0.0752	5.09	0.0241
sd	4	1	-0.7031	0.2909	-1.2732	-0.1331	5.84	0.0156
sd	5	1	-0.7517	0.2712	-1.2832	-0.2201	7.68	0.0056
sd	6	1	-0.5282	0.2569	-1.0318	-0.0247	4.23	0.0398
sd	7	1	-0.9872	0.2454	-1.4680	-0.5063	16.19	<.0001
sd	8	1	-0.5123	0.3570	-1.2120	0.1875	2.06	0.1513
sd	9	1	-0.5986	0.2403	-1.0696	-0.1276	6.20	0.0127

sd	10	1	-0.4272	0.2484	-0.9140	0.0596	2.96	0.0854
sd	11	1	-0.5368	0.2512	-1.0292	-0.0444	4.57	0.0326
sd	12	1	-0.7862	0.2525	-1.2810	0.2913	9.70	0.0018
sd	13	0	0.0000	0.0000	0.0000	0.0000	.	.
logacc		1	0.2369	0.0397	0.1590	0.3148	35.56	<.0001
Dispersion		1	0.1424	0.0167	0.1131	0.1792		

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	33.60	<.0001

Modèle 4 :

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	Erreur standard	Wald 95Limites 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) Entre les modèles 1 et 2 lequel n'est pas adapté et pourquoi ? Et si on avait gardé le modèle qui n'est pas adapté quelles auraient été les conséquences ?
2) Commentez globalement les résultats du modèle 2 tout en expliquant le passage au modèle 3.
3) Écrire le modèle 3 et presenter les caractéristiques de l'assuré de référence. Pourquoi avons-nous inclus une variable offset dans ce modèle ?
4) Expliquer ce que chaque élément de la ligne qui correspond à $sd8$ représente en commentant en même temps.
5) Nous avons effectué un test statistique dont l'hypothèse nulle est $H_0 : \beta_7 = \beta_8$ et avons obtenu une p-value de 0.1018. Expliquer la raison pour laquelle ce test a été effectué. Que faire au vu des résultats du test ?

6) Quel est le nombre moyen de sinistres de l'assuré de référence ? Quel est le nombre moyen de sinistres d'un assuré qui réside dans le départment 10 ?
7) Comment expliqueriez-vous la différence de la valeur de la statistique du test du χ^2 effectué pour la variable `sd` dans le tableau d'analyse de type 1 et de type 3 ? Et pourquoi obtenons-nous la même valeur dans les deux analyses pour la variable `logacc` ?
8) Le modèle 4 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 ? Et si on avait estimé un modèle de Poisson, en quoi les résultats des deux modèles (Poisson et quasi-Poisson) auraient été différents et/ou similaires ?
9) Quel modèle choisir entre le modèle 3 et le modèle 4 ? Justifier votre choix. Y a-t-il moyen d'améliorer le modèle choisi ?

$$Y \sim \Gamma(\mu, \nu) \quad f(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu}\right)^{\nu} e^{-\frac{y\nu}{\mu}}$$

$$\Rightarrow \ln(f(y)) = -\ln(y) - \ln(\Gamma(\nu)) + \nu \ln(y\nu) + \nu \ln(1/\mu) - \frac{y\nu}{\mu}$$

$$\Rightarrow f(y) = \exp\left(\frac{-y/\mu + \ln(1/\mu)}{1/\nu} - \ln(y) + \nu \ln(y\nu) - \ln\Gamma(\nu) \right)$$

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

$$\theta = -1/\mu \longrightarrow b(\theta) = -\ln(-\theta)$$

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

$$\theta_i = \frac{-1}{\mu_i} \qquad V(\mu) = b''(\theta_i) = \frac{1}{\theta^2} = \mu^2$$

$$\mathbb{E}[Y] = b'(\theta) = \frac{1}{-\theta} = \mu$$

$$Var(Y) = a(\phi)V(\mu) = \frac{1}{\nu} \times \mu^2 = \frac{\mu^2}{\nu}$$

Exercice 1:

$$g(\mu) = x'\beta \qquad \text{avec } \mu = \mathbb{E}[Y] \text{ et } Y \in \mathcal{F}_{exp}.$$

i) Notons f_Y la densité de Y.

$$\mathcal{L}(y;\theta;\phi) = \prod_{i=1}^{m} f_{Y_i}(y_i;\theta_i;\phi)$$

On sait que $Y \in \mathcal{F}_{exp}$, on a donc:

$$f_{Y_i}(y_i;\theta;\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}$$

$$\Rightarrow \mathcal{L}(y;\theta;\phi) = \exp\left(\sum_{i=1}^{m}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right]\right)$$

Or $a(\phi) = \frac{\phi}{\omega_i}$

$$\Rightarrow \ln(\mathcal{L}(y;\theta;\phi)) = \frac{1}{\phi}\sum_{i=1}^{m}\frac{y_i\theta_i - b(\theta_i)}{\omega_i} + \sum_{i=1}^{m} c(y_i,\phi)$$

On sait que $\forall j \in [\![0,p]\!]$, $\frac{\partial \ln(\mathcal{L})}{\partial \beta_j} = 0$

Or $$\frac{\partial \ln(f(y_i;\phi;\theta_i))}{\partial \beta_j} = \frac{\partial \ln(f(y_i;\phi;\theta_i))}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \beta_j}$$

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

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \quad \text{et} \quad \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = x_{ji}\frac{\partial \mu_i}{\partial \eta_i} \qquad \text{avec } \eta_i = x_i'\beta \leftarrow \text{Score}.$$

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

Et $\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i)$

$$\Rightarrow \frac{\partial \ln(f(y_i,\phi,\theta_i))}{\partial \beta_j} = \frac{\omega_i (y_i - \mu_i) x_{ji}}{\phi\, b''(\theta_i)\, g'(\mu_i)} = 0$$

D'où $$\frac{\partial \ln(\mathcal{L})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{m} \omega_i (y_i - \mu_i)\frac{x_{ji}}{b''(\theta_i)g'(\mu_i)} = 0 \qquad j \in [\![0,p]\!].$$

On trouve $\hat{\beta}$ par MV (Max de vraisemblance) ou par MCO (Moindres Carrés Ordinaires).

Si on utilise la fonct° lien canonique, il vient:

$$\eta_i = \theta_i = x_i'\beta \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i)$$

$$\Rightarrow \frac{\partial \ln(\mathcal{L})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{m} \omega_i (y_i - \mu_i) x_{ji} = 0$$

ii) Test de Significativité de Wald:

$\hat{\beta}$ est EMV tq $\hat{\beta} \sim \mathcal{N}(\beta, \phi(X'WX)^{-1})$ avec W matrice diagonale.

$H_0 : C\beta = r$ donc sous H_0, on a : $C\hat{\beta} - r \sim \mathcal{N}(0, \phi C(X'WX)^{-1}C')$

La stat du test de Wald est donc : $(C\hat{\beta}-r)'[\phi C(X'WX)^{-1}C']^{-1}(C\hat{\beta}-r) \sim \chi^2_q$

Pour 1 seul paramètre cela revient à tester $H_0 : \beta_j = r_j$

$$\Rightarrow C = (0 - 0 \underbrace{1}_{j^{\text{ème}} \text{ élément}} 0 - 0) \qquad Var(\hat{\beta}_j) = \phi \psi_j$$

La stat du test de Wald s'écrit $\frac{(\hat{\beta}_j - r_j)^2}{Var(\hat{\beta}_j)} \sim \chi^2_1$

iii) Résidus de déviance:

$$D = \sum_{i=1}^{m} d_i \qquad r_i^D = \text{signe}(y_i - \mu_i)\sqrt{d_i}$$

Cela permet de mesurer la contribution de chaque observation à la déviance du modèle par rapport au modèle saturé.

Ces résidus sont normaux car $D \sim \chi^2_{m-p-1}$ d'où $\sqrt{d_i} \sim \mathcal{N}$
$\Rightarrow r_i^D$ normale.

iv) Comment détecter une valeur influente?

Pour un MLG:

* Effet de levier:

On se sert de la matrice de projection H tel que $\hat{Y} = HY$ et:

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

On a $Tr(H) = p+1$, la moyenne est $\frac{p+1}{m}$.
Donc les valeurs tq $h_{ii} > \frac{2(p+1)}{m}$ doivent faire l'objet d'un examen approfondi.

* Distance de Cook:

$$C_i = \frac{1}{p+1}(\hat{\beta} - \hat{\beta}_{(i)})' X'WX(\hat{\beta} - \hat{\beta}_{(i)})$$

Pour un modèle de régress° linéaire classique :

* Matrice de project° :

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

Si $h_{ii} > \frac{2(p+1)}{n}$ ou $\frac{3(p+1)}{n}$ ou 0,5 $\Rightarrow$ Influente

* Distance de Cook

$$C_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X'X (\hat{\beta}_{(i)} - \hat{\beta})}{s^2 (p+1)} \sim \mathcal{F}_{p+1,\, n-p-1}$$

$$= \frac{h_{ii}}{(p+1)(1-h_{ii})} \cdot r_i$$

* DFFITS

Si $\underbrace{|DFFITS_i|}_{= |r_i^*| \sqrt{\frac{h_{ii}}{1-h_{ii}}}} > \sqrt{\frac{p+1}{n-p-1}}\, t_{n-p-2;\, 1-\frac{\alpha}{2}}$

V) $Y \sim \Gamma(\mu, \nu)$ $\quad f(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu}\right)^{\nu} e^{-\frac{y\nu}{\mu}}$

a) $\ln(f(y)) = -\ln(\Gamma(\nu)) + (\nu-1)\ln(y) + \nu \ln\left(\frac{\nu}{\mu}\right) - \frac{y\nu}{\mu}$

$$\Rightarrow f(y) = \exp\left(-\frac{y\nu}{\mu} + \nu \ln\left(\frac{1}{\mu}\right) + \nu \ln(y\nu) - \ln(y) - \ln\Gamma(\nu)\right)$$

$$= \exp\left(\frac{-y\frac{1}{\mu} + \ln\left(\frac{1}{\mu}\right)}{1/\nu} + \nu \ln(y\nu) - \ln(y) - \ln\Gamma(\nu)\right)$$

$$\Rightarrow \theta = -\frac{1}{\mu} \qquad b(\theta) = -\ln(-\theta)$$

$$a(\phi) = 1/\nu \qquad \phi = 1/\nu \qquad c(y,\phi) = \frac{1}{\phi}\ln\left(\frac{y}{\phi}\right) - \ln(y) - \ln(\Gamma(1/\phi))$$

b) $\mu_i = -\frac{1}{\theta_i} \quad \Rightarrow g : x \mapsto \frac{-1}{x}$ fonct° lien canonique

$-\ln(-x) \quad -x = \frac{-1}{-x}$

$$V(\mu) = b''(\theta) = \frac{1}{\theta^2} = \mu^2$$

c) $\mathbb{E}[Y] = \mu = b'(\theta) = -1/\theta = \mu$

$$Var(Y) = a(\phi)\, V(\mu) = \frac{\mu^2}{\nu}$$

d) $D = 2[l_{SAT} - \mathcal{L}]$

où $\mathcal{L} = \sum_{i=1}^{n}\left[\dfrac{-y_i \frac{1}{\mu_i} + \ln(1/\mu_i)}{1/\nu} + c(y_i, \phi)\right]$

Dans le modèle saturé, $\mu_i = y_i$

$\Rightarrow D = 2\sum_{i=1}^{n} \nu\left[\ln\left(\frac{1}{y_i}\right) - 1\right] - \nu\left[\ln\left(\frac{1}{\hat{\mu}_i}\right) - \frac{y_i}{\hat{\mu}_i}\right]$

$= 2\nu \sum_{i=1}^{n}\left(-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)$

e) $\hat{r}_i^D = \dfrac{y_i - \hat{\mu}_i}{|y_i - \hat{\mu}_i|}\sqrt{d_i}$

$\Rightarrow \hat{r}_i^D = \dfrac{y_i - \hat{\mu}_i}{|y_i - \hat{\mu}_i|}\sqrt{-\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}}$

vi) Si on s'intéresse à une variable qualitative avec modalités (sans ordre entre les modalités)

$\Rightarrow$ modèle de regression nominale

Les π_j sont liées aux variables explicatives.

On modélise la côte de Y dans la catégorie j p/r au niveau de référence r

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

Exercice 2 :

1) Le modèle 1 n'est pas adapté car sa déviance $= 73{,}64 \gg 1$

$\Rightarrow$ ajustement très mauvais.

La déviance du modèle 2 vaut $1{,}21$ ce qui est mieux !

$\rightarrow$ Si on garde le modèle 1 $\Rightarrow$ ajustement mauvais $\Rightarrow$ résultats mauvais, estimat° fausses ...

2) Le modèle 2 est globalement significatif sauf k_i et pop-density

$\Rightarrow$ on les enlève du modèle.

3) Modèle : $\ln(\text{claims}) = \ln(n) + x'\beta$ $\qquad \log[\text{Claims}] = \log pop + \beta_1 sd_1 + \dots + \beta_{13} sd_{13} + \beta_{14} \log age$

assuré de réf : $\mu = n e^{-6{,}2331}$

$= 176\, e^{-6{,}2331}$

$= 0{,}356$

Var offset : sert à prendre en compte les $\neq$ taille de pop.

4) Ligne sd8 : pour le département 8.

DF = 1 Variable binomiale vaut 1 si indep de sd8 (0 sinon)
Estimat° = −0,5123 estimat° du modèle du Coeff associé à sd8.
Standard error = 0,3570 erreur d'estimat°
IC 95% estimat° par test de signif. de wald.
$Khi\ 2 : \frac{(\beta_{sd8})^2}{Var(\beta_{sd8})} = \frac{(-0,5123)^2}{0,3570^2} = 2,06$

p.value = 0,1513 sert à l'interprétat° de la signif. de la variable.

Ici p.value > 5% => $\beta_8 = 0$ Va non signif.

5) Car dans le modèle 3, la variable sd8 n'est pas signif.
donc on veut tester si on peut la regrouper avec la variable d'un autre départemt

Ici p.value = 0,1018 > 5% => on ne rejette pas H_0 => On conclut que $\beta_7 = \beta_8$.

6) individu de ref : $\mu_i = exp(-6,2331) = 0,001963$

individu du départemt 10 : $\mu_i = exp(-6,2331 - 0,4272)$
$= 0,00128$.

7) Dans analyse de type 1, on teste la signif de la variable sd quand il n'y a que celle-ci dans le modèle (privée des autres).

Analyse de type 3 : test de la va sd quand toutes les autres sont présentes.

Lagacc est la dernière va donc dans les 2 analyses il y a toutes les autres va.

8) Quasi-Poisson : cas de pb de surdispers°
$\mathbb{E}[Y] = \mu$ et $Var(Y) = \phi\mu$

↳ ≠ entre les 2 : surdispers° => surestimer la stat du test de signif de Wald

Du coup, une va jugée pertinente dans le modèle de Poisson, ne l'est pas forcément après avoir pris en compte la surdispers°.

	Modèle 3	Modèle 4
9) Scaled Dev :	1,1922	1,00

⟶ On préf le modèle 4 (qualité d'ajustemt meilleure).
Et il y a un coeff dut qui est signif