MODELISATION DE DONNEES QUALITATIVES

LA REGRESSION DE POISSON

Pierre-Louis GONZALEZ

LA RÉGRESSION DE POISSON

La régression de Poisson permet de modéliser des comptages distribués selon une loi de Poisson en fonction de variables explicatives quantitatives ou qualitatives.

1 - Les données

Y = comptage

X_1 ... X_k Variables explicatives

Exemple

Âge	Région	Y=mélanome	Population
< 35	N	61	2 880 262
35 - 44	N	76	564 535
45 - 54	N	98	592 983
55 - 64	N	104	450 740
65 - 74	N	63	270 908
> 74	N	80	161 850
< 35	S	64	1 074 246
35 - 44	S	75	220 407
45 - 54	S	68	198 119
55 - 64	S	63	134 084
65 - 74	S	45	70 708
> 74	S	27	34 233

2 - Le modèle

Y suit une loi de Poisson de moyenne

$$\mu = \mathbf{N} \exp\left[\beta_0 + \begin{matrix}<35\\35-44\\45-54\\55-64\\65-74\\>74\end{matrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\\\beta_5\\-\beta_1\ldots-\beta_5\end{bmatrix} + \begin{matrix}\mathbf{N}\\\mathbf{S}\end{matrix}\begin{bmatrix}\beta_6\\-\beta_6\end{bmatrix}\right]$$

Effectif population soumise au risque — **Âge** — **Région**

$$+ \begin{matrix}<35\\35-44\\45-54\\55-64\\65-74\\>74\end{matrix}\left|\begin{matrix}\beta_7\\\beta_8\\\beta_9\\\beta_{10}\\\beta_{11}\\-\beta_7\ldots-\beta_{11}\end{matrix}\right|\left|\begin{matrix}-\beta_7\\-\beta_8\\-\beta_9\\-\beta_{10}\\-\beta_{11}\\\beta_7+\ldots+\beta_{11}\end{matrix}\right|$$

N **S**

Âge * Région

$$P(Y=y) = \frac{e^{-\mu}\,\mu^{y}}{y!}$$

Vraisemblance

$$\varphi\left(y_1,\dots y_s \,/\, \beta_0,\beta_1\dots\beta_{11}\right) = \prod_{i=1}^{12} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

où **i** est l'indice de la **i**ème population.

Estimation

On estime les β_j en maximisant la vraisemblance.

Test

$$H_0 : L\beta = 0$$

$$H_1 : L\beta \neq 0$$

Statistique de Wald

$$Q = \left(L\hat{\beta}\right)' \left[\hat{V}ar\left(L\hat{\beta}\right)\right]^{-1} L\hat{\beta}$$

Règle de décision

On rejette H_0 si :

$$Q \geq \chi_{0,95}\left(\text{rang } L\right)$$

II. Exemples

Exemple 1 : Mélanomes

Nous reprenons l'exemple sur le risque de Mélanome présenté dans Tenenhaus (1993) : La régression de Poisson, *Modélisation et Analyse des Données*, n° 4, pp.41-48. Les données proviennent de Koch, Atkinson & Stokes (1986) : Poisson Regression. In Kotz, Johnson & Read (Eds) : *Encyclopedia of Statistical Sciences*, vol. 7, Wiley. Elles concernent des personnes de race blanche atteintes de mélanome dans les années 1969-1971 et sont présentées dans le tableau ci-dessous. L'indice *i* varie de 1 à 6 pour la région Nord et de 7 à 12 pour la région Sud.

Tranche d'âge	Nombre de cas de mélanomes, n_i		Nombre estimé de personnes soumises au risque, N_i	
	Région Nord, n_i	Région Sud, n_i	Région Nord, N_i	Région Sud, N_i
< 35	61	64	2 880 262	1 074 246
35-44	76	75	564 535	220 407
45-54	98	68	592 983	198 119
55-64	104	63	450 740	134 084
65-74	63	45	270 908	70 708
≥75	80	27	161 850	34 233

Le but de l'étude est ici de déterminer si le rapport entre le nombre d'atteints et le nombre d'exposés, n_i / N_i, est à peu près constant ou non selon la région et la tranche d'âge.

Nous allons utiliser sur cet exemple la régression de Poisson. On suppose que le comptage $Y_i = n_i$ suit une loi de Poisson de moyenne :

$$\mu_i = N_i \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})$$

Le modèle étudié s'écrit donc :

- $Y_i \sim \text{Poisson}(\mu_i)$
- $\text{Log}(\mu_i) = \text{Log}(N_i) + \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$

Plus précisément, notons A_h la variable indicatrice de la tranche d'âge h (de <35 à ≥75) et R_k la variable indicatrice de la région k (1 = Nord et 2 = Sud). Abandonnant les dernières modalités, le modèle avec interaction s'écrit :

$$\text{Log}(\mu_i) = \text{Log}(N_i) + \beta_0 + \beta_1 A_1 + \ldots + \beta_5 A_5 + \beta_6 R_1 + \beta_7 A_1 * R_1 + \ldots + \beta_{11} A_5 * R_1$$

C'est un modèle saturé puisqu'il y a 12 paramètres pour 12 observations.

Utilisons maintenant la Proc Genmod pour étudier ce modèle.

Les données

OBS	AGE	REGION	MELANOME	EFFECTIF
1	<35	n	1	61
2	<35	s	1	64
3	35-44	n	1	76
4	35-44	s	1	75
5	45-54	n	1	98
6	45-54	s	1	68
7	55-64	n	1	104
8	55-64	s	1	63
9	65-74	n	1	63
10	65-74	s	1	45
11	>74	n	1	80
12	>74	s	1	27
13	<35	n	2	2880201
14	<35	s	2	1074182
15	35-44	n	2	564459
16	35-44	s	2	220332
17	45-54	n	2	592885
18	45-54	s	2	198051
19	55-64	n	2	450636
20	55-64	s	2	134021
21	65-74	n	2	270845
22	65-74	s	2	70663
23	>74	n	2	161770
24	>74	s	2	34206

Le programme (modèle avec interaction)

```
options nocenter nodate nolabel pageno=1;
data melanome;
input id $ age $ region $ cas pop;
logcsp=log(cas/pop);
logpop=log(pop);
cards;
n,<35   <35   n   61 2880262
s,<35   <35   s   64 1074246
n,35-44 35-44 n   76  564535
s,35-44 35-44 s   75  220407
n,45-54 45-54 n   98  592983
s,45-54 45-54 s   68  198119
n,55-64 55-64 n  104  450740
s,55-64 55-64 s   63  134084
n,65-74 65-74 n   63  270908
s,65-74 65-74 s   45   70708
n,>74   >74   n   80  161850
s,>74   >74   s   27   34233
;

proc print data=melanome;
run;

proc genmod data=melanome order=data;
class age region;
model cas=age region age*region/dist=poisson
                                link=log
                                offset=logpop
                                type3 ;
run;
```

Les Résultats

OBS	ID	AGE	REGION	CAS	POP	LOGCSP	LOGPOP
1	n,<35	<35	n	61	2880262	-10.7625	14.8734
2	s,<35	<35	s	64	1074246	-9.7282	13.8871
3	n,35-44	35-44	n	76	564535	-8.9130	13.2438
4	s,35-44	35-44	s	75	220407	-7.9857	12.3032
5	n,45-54	45-54	n	98	592983	-8.7080	13.2929
6	s,45-54	45-54	s	68	198119	-7.9771	12.1966
7	n,55-64	55-64	n	104	450740	-8.3743	13.0186
8	s,55-64	55-64	s	63	134084	-7.6631	11.8062
9	n,65-74	65-74	n	63	270908	-8.3664	12.5095
10	s,65-74	65-74	s	45	70708	-7.3597	11.1663
11	n,>74	>74	n	80	161850	-7.6124	11.9944
12	s,>74	>74	s	27	34233	-7.1451	10.4409

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.MELANOME
Distribution	POISSON
Link Function	LOG
Dependent Variable	CAS
Offset Variable	LOGPOP
Observations Used	12

Class Level Information

Class	Levels	Values
AGE	6	<35 35-44 45-54 55-64 65-74 >74
REGION	2	n s

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood	.	2698.0337	.

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT			1	-7.1451	0.1925	1378.4195	0.0001
AGE	<35		1	-2.5831	0.2295	126.7061	0.0001
AGE	35-44		1	-0.8406	0.2244	14.0294	0.0002
AGE	45-54		1	-0.8320	0.2275	13.3784	0.0003
AGE	55-64		1	-0.5180	0.2300	5.0709	0.0243
AGE	65-74		1	-0.2145	0.2434	0.7767	0.3781
AGE	>74		0	0.0000	0.0000	.	.
REGION	n		1	-0.4673	0.2226	4.4080	0.0358
REGION	s		0	0.0000	0.0000	.	.
AGE*REGION	<35	n	1	-0.5670	0.2856	3.9417	0.0471
AGE*REGION	<35	s	0	0.0000	0.0000	.	.
AGE*REGION	35-44	n	1	-0.4600	0.2757	2.7831	0.0953
AGE*REGION	35-44	s	0	0.0000	0.0000	.	.
AGE*REGION	45-54	n	1	-0.2635	0.2728	0.9330	0.3341
AGE*REGION	45-54	s	0	0.0000	0.0000	.	.
AGE*REGION	55-64	n	1	-0.2439	0.2739	0.7928	0.3733
AGE*REGION	55-64	s	0	0.0000	0.0000	.	.
AGE*REGION	65-74	n	1	-0.5395	0.2960	3.3209	0.0684
AGE*REGION	65-74	s	0	0.0000	0.0000	.	.
AGE*REGION	>74	n	0	0.0000	0.0000	.	.
AGE*REGION	>74	s	0	0.0000	0.0000	.	.
SCALE			0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
AGE	5	715.9897	0.0001
REGION	1	108.1919	0.0001
AGE*REGION	5	6.2149	0.2859

Commentaires

1) On vérifie que la déviance et le khi-deux de Pearson sont nuls puisque le modèle est saturé.

2) La vraisemblance des données s'écrit

$$L = \prod_{i=1}^{N} e^{-\mu_i} \mu_i^{y_i} / y_i!$$

et son logarithme

$$l = \sum_{i=1}^{N} \left[-\mu_i + y_i \log(\mu_i) - \log(y_i!) \right]$$

Le *Log Likelihood* l^* fournit par la Proc Genmod correspond en fait à la partie de l qui dépend des μ_i, c'est à dire

$$l^* = \sum_{i=1}^{N} \left[-\mu_i + y_i \log(\mu_i) \right]$$

Il est donc possible d'obtenir des *Log Likelihood* l^* positifs (!), ce qui est le cas sur cet exemple. Les tests LRT peuvent être construits à partir des *Log Likelihood* l^*.

3) L'interaction étant non significative, on passe maintenant au modèle sans interaction.

Le programme (modèle additif)

```
proc genmod data=melanome order=data;
class age region;
model cas=age region/dist=poisson
                    link=log
                    offset=logpop
                    type3 obstats residuals;
contrast '<35 vs 35-44'    age -1  1  0  0  0  0;
contrast '35-44 vs 45-54'  age  0 -1  1  0  0  0;
contrast '45-54 vs 55-64'  age  0  0 -1  1  0  0;
contrast '55-64 vs 65-74'  age  0  0  0 -1  1  0;
contrast '65-74 vs >74'    age  0  0  0  0 -1  1;
contrast '<35 vs 35-44'    age -1  1  0  0  0  0 / wald;
contrast '35-44 vs 45-54'  age  0 -1  1  0  0  0 / wald;
contrast '45-54 vs 55-64'  age  0  0 -1  1  0  0 / wald;
contrast '55-64 vs 65-74'  age  0  0  0 -1  1  0 / wald;
contrast '65-74 vs >74'    age  0  0  0  0 -1  1 / wald;
make 'obstats'  out=a ;
run;

data aa;
merge melanome a;

proc plot data=aa;
       plot logcsp*xbeta='*' $ id;

run;
```

Les résultats

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	6.2149	1.2430
Scaled Deviance	5	6.2149	1.2430
Pearson Chi-Square	5	6.1151	1.2230
Scaled Pearson X2	5	6.1151	1.2230
Log Likelihood	.	2694.9262	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-6.8941	0.1079	4080.0957	0.0001
AGE	<35	1	-2.9447	0.1320	497.2981	0.0001
AGE	35-44	1	-1.1473	0.1268	81.8851	0.0001
AGE	45-54	1	-1.0316	0.1242	68.9792	0.0001
AGE	55-64	1	-0.7029	0.1240	32.1532	0.0001
AGE	65-74	1	-0.5790	0.1364	18.0049	0.0001
AGE	>74	0	0.0000	0.0000	.	.
REGION	n	1	-0.8195	0.0710	133.1138	0.0001
REGION	s	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
AGE	5	796.7436	0.0001
REGION	1	124.2203	0.0001

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
<35 vs 35-44	1	206.3042	0.0001	LR
35-44 vs 45-54	1	1.0595	0.3033	LR
45-54 vs 55-64	1	8.9551	0.0028	LR
55-64 vs 65-74	1	0.9974	0.3179	LR
65-74 vs >74	1	17.7438	0.0001	LR
<35 vs 35-44	1	220.9204	0.0001	Wald
35-44 vs 45-54	1	1.0581	0.3036	Wald
45-54 vs 55-64	1	8.9924	0.0027	Wald
55-64 vs 65-74	1	1.0068	0.3157	Wald
65-74 vs >74	1	18.0049	0.0001	Wald

Observation Statistics

CAS	Pred	Xbeta	Std	HessWgt	Lower	Upper	Resraw
61	67.6998	-10.6583	0.0952	67.6998	56.1779	81.5846	-6.6998
64	57.3002	-9.8388	0.0974	57.3002	47.3456	69.3479	6.6998
76	80.0638	-8.8609	0.0880	80.0638	67.3860	95.1267	-4.0638
75	70.9362	-8.0414	0.0897	70.9362	59.5032	84.5660	4.0638
98	94.4150	-8.7452	0.0834	94.4150	80.1708	111.1899	3.5850
68	71.5850	-7.9257	0.0875	71.5850	60.3035	84.9771	-3.5850
104	99.6974	-8.4165	0.0825	99.6974	84.8113	117.1963	4.3026
63	67.3026	-7.5970	0.0882	67.3026	56.6139	80.0094	-4.3026
63	67.8263	-8.2926	0.0998	67.8263	55.7776	82.4776	-4.8263
45	40.1737	-7.4731	0.1061	40.1737	32.6334	49.4563	4.8263
80	72.2979	-7.7136	0.0994	72.2979	59.5022	87.8453	7.7021
27	34.7021	-6.8941	0.1079	34.7021	28.0857	42.8772	-7.7021

Observation Statistics

Reschi	Resdev	StResdev	StReschi	Reslik
-0.8143	-0.8283	-1.3321	-1.3095	-1.3183
0.8851	0.8686	1.2852	1.3095	1.2985
-0.4542	-0.4581	-0.7425	-0.7361	-0.7386
0.4825	0.4780	0.7293	0.7361	0.7332
0.3690	0.3667	0.6264	0.6303	0.6290
-0.4237	-0.4273	-0.6357	-0.6303	-0.6327
0.4309	0.4279	0.7548	0.7602	0.7585
-0.5245	-0.5302	-0.7685	-0.7602	-0.7642
-0.5860	-0.5932	-1.0411	-1.0285	-1.0326
0.7614	0.7469	1.0089	1.0285	1.0178
0.9058	0.8904	1.6651	1.6939	1.6857
-1.3075	-1.3609	-1.7632	-1.6939	-1.7355

Commentaires

1) On vérifie que la statistique LRT de l'interaction Age*Région vaut :

$$2[\log L(\text{Age, Région, Age*Région}) - \log L(\text{Age, Région})] = 2(2698.0337 - 2694.9262) = 6.2150$$

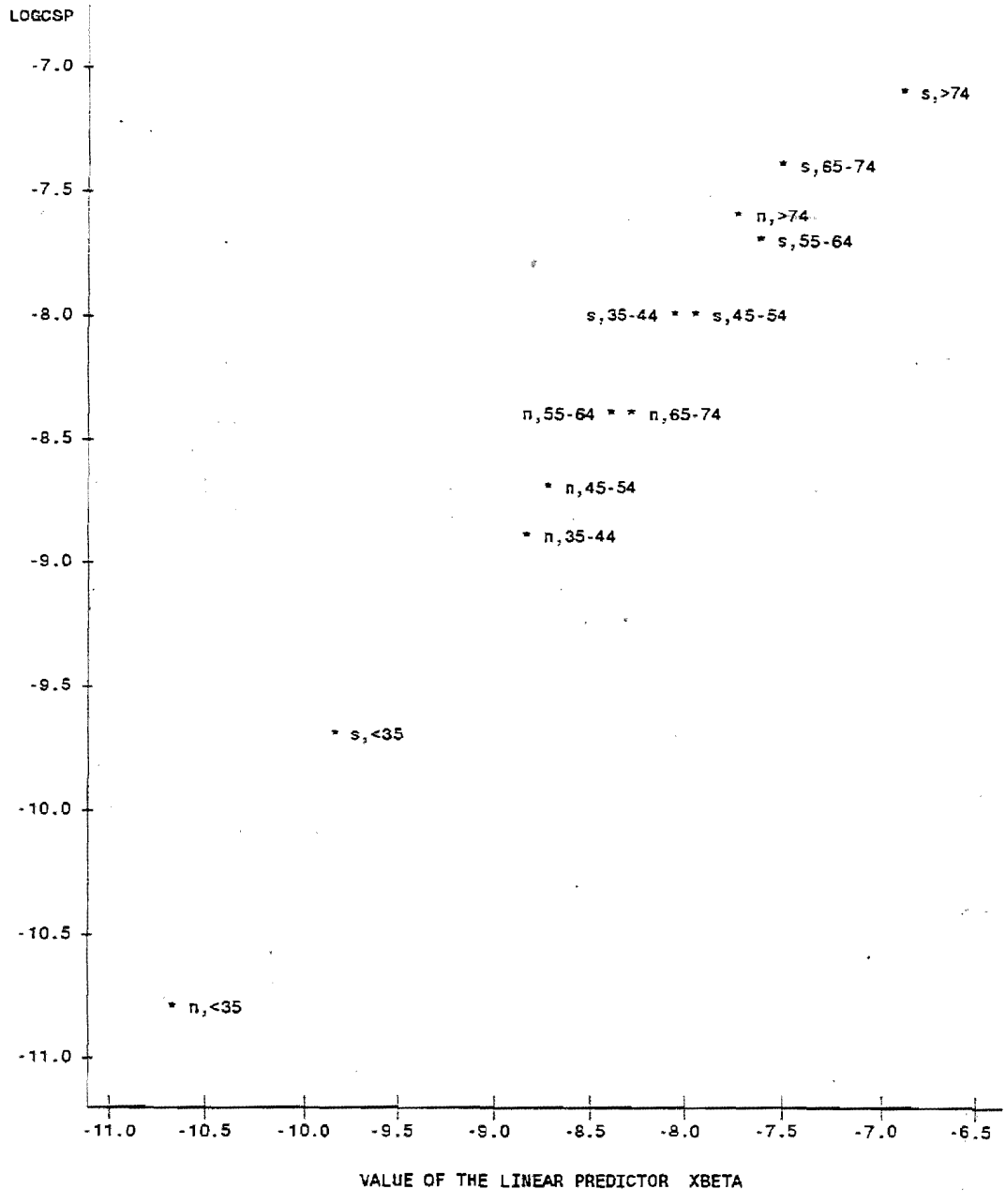
2) Les facteurs Age et Région sont significatifs. Le modèle additif étudié est accepté. La déviance normalisée et le khi-deux de Pearson normalisé divisés par leur degrés de liberté N-p = 12-7 = 5 sont proches de 1 (respectivement 1.2430 et 1.2230).

3) Le modèle estimé s'écrit

$$\log(\mu_i) = \log(N_i) - 6.89 - 2.94\text{Age}_{<35} - 1.15\text{Age}_{35\text{-}44} - 1.03\text{Age}_{45\text{-}54} - 0.70\text{Age}_{55\text{-}64} - 0.58\text{Age}_{65\text{-}74} - 0.82\text{Nord}$$

L'ajustement entre les données et le modèle est visualisé dans la figure ci-dessous où on a représenté en ordonné le logarithme de n_i/N_i , noté LOGCSP, et en abscisse

$$\text{XBETA} = -6.89 - 2.94\text{Age}_{<35} - 1.15\text{Age}_{35\text{-}44} - 1.03\text{Age}_{45\text{-}54} - 0.70\text{Age}_{55\text{-}64} - 0.58\text{Age}_{65\text{-}74} - 0.82\text{Nord}$$



4) L'examen des coefficients des variables indicatrices de l'âge suggère de regrouper les âges 35-44 et 45-54 et également les âges 55-64 et 65-74. Nous avons donc construit les contrastes permettant de comparer les tranches d'âge adjacentes. Nous avons utilisé les statistiques de Wald et LRT. Les résultats donnés par ces deux statistiques sont très voisins et confirment la possibilité de regroupement. Nous avons donc étudié un troisième modèle en réalisant ces regroupement.

Le programme (modèle additif simplifié)

```
data b;
set melanome;
age1=(age = "<35");
age2=(age = "35-44") or (age="45-54");
age3=(age = "55-64") or (age="65-74");

proc genmod data=b order=data;
class region;
model cas=age1 age2 age3 region/dist=poisson
                    link=log
                    offset=logpop
                    type3 residuals waldci lrci;
contrast 'age'   age1 1,
                 age2 1,
                 age3 1 /e;
contrast 'age'   age1 1,
                 age2 1,
                 age3 1 / wald;
run;
```

Les résultats

Parameter Information

Parameter	Effect	REGION
PRM1	INTERCEPT	
PRM2	AGE1	
PRM3	AGE2	
PRM4	AGE3	
PRM5	REGION	n
PRM6	REGION	s

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	8.2709	1.1816
Scaled Deviance	7	8.2709	1.1816
Pearson Chi-Square	7	8.2329	1.1761
Scaled Pearson X2	7	8.2329	1.1761
Log Likelihood	.	2693.8982	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-6.8962	0.1079	4081.6288	0.0001
AGE1		1	-2.9443	0.1320	497.1665	0.0001
AGE2		1	-1.0880	0.1122	94.0913	0.0001
AGE3		1	-0.6558	0.1140	33.0717	0.0001
REGION	n	1	-0.8165	0.0710	132.2439	0.0001
REGION	s	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Normal Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9500

Parameter	Confidence Limits	
PRM1	Lower	-7.1078
PRM1	Upper	-6.6846
PRM2	Lower	-3.2031
PRM2	Upper	-2.6855
PRM3	Lower	-1.3078
PRM3	Upper	-0.8682
PRM4	Lower	-0.8793
PRM4	Upper	-0.4323
PRM5	Lower	-0.9556
PRM5	Upper	-0.6773

Likelihood Ratio Based Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9500

Parameter	Confidence Limits		Parameter Values				
			PRM1	PRM2	PRM3	PRM4	PRM5
PRM1	Lower	-7.1132	-7.1132	-2.7604	-0.9046	-0.4759	-0.7562
PRM1	Upper	-6.6898	-6.6898	-3.1168	-1.2601	-0.8241	-0.8798
PRM2	Lower	-3.2026	-6.7592	-3.2026	-1.2195	-0.7867	-0.8265
PRM2	Upper	-2.6843	-7.0505	-2.6843	-0.9392	-0.5076	-0.8065
PRM3	Lower	-1.3041	-6.7324	-3.1018	-1.3041	-0.8126	-0.8280
PRM3	Upper	-0.8640	-7.0744	-2.7719	-0.8640	-0.4841	-0.8058
PRM4	Lower	-0.8759	-6.7390	-3.0982	-1.2419	-0.8759	-0.8225
PRM4	Upper	-0.4284	-7.0685	-2.7751	-0.9188	-0.4284	-0.8108
PRM5	Lower	-0.9551	-6.8045	-2.9631	-1.1058	-0.6653	-0.9551
PRM5	Upper	-0.6767	-6.9928	-2.9259	-1.0707	-0.6466	-0.6767

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
AGE1	1	362.7350	0.0001
AGE2	1	77.2131	0.0001
AGE3	1	29.7461	0.0001
REGION	1	123.4360	0.0001

Coefficients For age

Parameter	ROW1	ROW2	ROW3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
5	0	0	0
6	0	0	0

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
age	3	794.6877	0.0001	LR
age	3	605.9982	0.0001	Wald

Commentaires

1) On peut comparer le modèle additif complet et le modèle simplifié à l'aide d'un test LRT. La statistique LRT vaut 2(2694.9262 - 2693.8982) = 2.056 à comparer au fractile $\chi^2_{0.95}(2) = 5.99$. On accepte donc le modèle simplifié. On peut retrouver ce résultat à l'aide d'un test sur le modèle additif complet. On utilise le programme ci-dessous :

```
proc genmod data=melanome order=data;
class age region;
model cas=age region/ dist=poisson
                      link=log
                      offset=logpop
                      type3;
       contrast 'modèle complet vs simplifié'
       age  0 -1  1  0  0  0,
       age  0  0  0 -1  1  0/e;
run;
```

Et on obtient le résultat :

Coefficients For modèle complet vs si

Parameter	ROW1	ROW2
1	0	0
2	0	0
3	-1	0
4	1	0
5	0	-1
6	0	1
7	0	0
8	0	0
9	0	0

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
modèle complet vs si	2	2.0560	0.3577	LR

2) On peut aussi remarquer que la déviance divisé par ses degrés de liberté a diminué en passant du modèle additif complet au modèle simplifié (respectivement 1.243 et 1.1816)

3) Le modèle simplifié estimé s'écrit

$$\log(\mu_i) = \log(N_i) - 6.90 - 2.94\text{Age}_{<35} - 1.09\text{Age}_{35\text{-}54} - 0.66\text{Age}_{55\text{-}74} - 0.82\text{Nord}$$

4) Dans cet exemple, les tailles des populations soumises au risque sont élevées par rapport au nombre de cas observés. En fait on peut aussi considérer que Y_i suit une loi binomiale bin(N_i ;p_i) où p_i est la probabilité qu'un individu tiré au hasard dans la population soumise au risque présente un mélanome. Cette loi binomiale est approchée par une loi de Poisson de moyenne $\mu_i = N_i p_i$. Nous avons donc essayé de modéliser ces données en utilisant la loi binomiale et en conservant la fonction de lien logarithme :

$$\log(p_i) = \beta_0 + \beta_1\text{Age}_{<35} + \beta_2\text{Age}_{35\text{-}54} + \beta_3\text{Age}_{55\text{-}74} + \beta_4\text{Nord}$$

Voici le nouveau programme.

Programme (Réponse binomiale, fonction de lien log, modèle additif simplifié)

```
proc genmod data=b order=data;
class region;
model cas/pop=age1 age2 age3 region/dist=bin
                                    link=log
                                    type3;
contrast 'age'   age1 1,
                 age2 1,
                 age3 1;
run;
```

Résultats

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.B
Distribution	BINOMIAL
Link Function	LOG
Dependent Variable	CAS
Dependent Variable	POP
Observations Used	12
Number Of Events	824
Number Of Trials	6653075

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	8.2745	1.1821
Scaled Deviance	7	8.2745	1.1821
Pearson Chi-Square	7	8.2368	1.1767
Scaled Pearson X2	7	8.2368	1.1767
Log Likelihood	.	-7793.1578	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-6.8963	0.1079	4084.4651	0.0001
AGE1		1	-2.9442	0.1320	497.3351	0.0001
AGE2		1	-1.0880	0.1121	94.1348	0.0001
AGE3		1	-0.6558	0.1140	33.0858	0.0001
REGION	n	1	-0.8164	0.0710	132.2803	0.0001
REGION	s	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
AGE1	1	362.7851	0.0001
AGE2	1	77.2346	0.0001
AGE3	1	29.7556	0.0001
REGION	1	123.4627	0.0001

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
age	3	794.8074	0.0001	LR

Commentaires

Le modèle estimé s'écrit

$$\log(p_i) = -6.90 - 2.94\text{Age}_{<35} - 1.09\text{Age}_{35\text{-}54} - 0.66\text{Age}_{55\text{-}74} - 0.82\text{Nord}$$

On retrouve exactement (en tout cas avec la précision choisie) la régression de Poisson estimée plus haut. Ce résultat est toujours vrai lorsqu'on est dans les conditions de convergence de la loi binomiale vers la loi de Poisson : N_i grand et p_i petit. Ce qui est le cas ici.

5) Les probabilités p_i étant petites, $\log(p_i)$ est très peu différent de $\log(\frac{p_i}{1-p_i})$. Par conséquent la régression de Poisson est ici équivalente à une régression logistique. Nous allons donc vérifier que le modèle

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 Age_{<35} + \beta_2 Age_{35\text{-}54} + \beta_3 Age_{55\text{-}74} + \beta_4 Nord$$

conduit à la même estimation des coefficients de régression que le modèle précédent.

Le programme (Réponse binomiale, fonction de lien logit)

```
proc genmod data=b order=data;
class region;
model cas/pop=age1 age2 age3 region/dist=bin
                                    link=logit
                                    type3;
contrast 'age'    age1 1,
                  age2 1,
                  age3 1;
run;
```

Résultats

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.B
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	CAS
Dependent Variable	POP
Observations Used	12
Number Of Events	824
Number Of Trials	6653075

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	8.2667	1.1810
Scaled Deviance	7	8.2667	1.1810
Pearson Chi-Square	7	8.2292	1.1756
Scaled Pearson X2	7	8.2292	1.1756
Log Likelihood	.	-7793.1539	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-6.8954	0.1080	4077.8795	0.0001
AGE1		1	-2.9449	0.1321	497.1912	0.0001
AGE2		1	-1.0884	0.1122	94.1134	0.0001
AGE3		1	-0.6561	0.1141	33.0836	0.0001
REGION	n	1	-0.8167	0.0710	132.2760	0.0001

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
REGION	s	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
AGE1	1	362.7964	0.0001
AGE2	1	77.2398	0.0001
AGE3	1	29.7586	0.0001
REGION	1	123.4705	0.0001

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
age	3	794.8152	0.0001	LR

Commentaires

Le modèle estimé s'écrit

$$\log\left(\frac{p_i}{1-p_i}\right) = -6.90 - 2.94\text{Age}_{<35} - 1.09\text{Age}_{35\text{-}54} - 0.66\text{Age}_{55\text{-}74} - 0.82\text{Nord}$$

On retrouve exactement (en tout cas avec la précision choisie) la régression de Poisson estimée plus haut. Ce résultat est toujours vrai lorsqu'on est dans les conditions de convergence de la loi binomiale vers la loi de Poisson : N_i grand et p_i petit. Ce qui est le cas ici.