MAT7381 Exercices
Chapitre 8 – Modèle logistique

8.1 Les données suivantes[1] portent sur la fréquence de certaines malformations du système nerveux chez les nouveau-nés de quelques localités en Australie. Le but de l'étude est de déterminer si la dureté de l'eau y est pour quelque chose. On considère comme variable concomitante le type de travail (manuel ou pas).

Localité	Dureté de l'eau (parties par million)	Travailleurs non manuels		Travailleurs manuels	
		Avec malformation	Sans malformation	Avec malformation	Sans malformation
Cardiff	110	19	4091	78	9424
Newport	100	8	1515	24	4610
Swansea	95	14	2394	53	5526
Glamorgan E.	42	26	3163	145	13217
Glamorgan O.	39	16	1979	84	8195
Glamorgan C.	161	25	4838	65	7803
Monmouth V.	83	18	2362	86	9962
Montmouth (autre)	122	9	1604	24	3172

a) Ajuster un modèle logistique liant la dureté de l'eau à la fréquence des malformations :

i) Assurez-vous que la relation existe réellement et estimer les coefficients.

ii) Estimer la fréquence des malformations dans une localité dans laquelle la dureté de l'eau est de 150 parties par million.

iii) Déterminer un intervalle de confiance pour le paramètre estimé en ii).

iv) Estimer la dureté de l'eau telle que le taux de malformation moyen soit de 1/1000.

v) Estimer approximativement l'écart-type de l'estimateur en iv)

b) Le coefficient de la variable « eau » est négatif, ce qui veut dire que le taux de malformation diminue lorsque l'eau durcit. Est-ce vraiment le cas ? Faites intervenir le type de travail dans votre analyse, et réduisez s'il y a lieu le nombre de paramètres. Le coefficient de la variable « eau » change-t-il de signe ?

8.2 On administre une certaine dose d'un poison à des groupes d'une cinquantaine d'insectes, et on observe le nombre (y) d'insectes morts ou moribonds. Voici les données :

dose	n (nombre d'insectes exposés)	y (nombre morts ou moribonds)
0,2	50	44
7,7	50	49
5,1	49	44
3,8	46	42
2,6	48	24
0,0	49	0

a) Estimez les paramètres d'une régression logistique, $\pi(x) = \dfrac{e^{\beta_o+\beta_1 x}}{1+e^{\beta_o+\beta_1 x}}$, où x est la dose, testez le modèle et testez l'hypothèse que la pente est nulle.

b) Supposez que la fonction logistique est la fonction de répartition de la variable X = dose mortelle (la dose minimale à laquelle l'insecte meurt). On veut déterminer un intervalle de confiance pour la médiane de cette distribution.

(i) Montrez que la médiane est $-\beta_o/\beta_1$.

[1] Loew, C. R., Roberts, C. J., et Lloyd, S (1971) Malformations of central nervous system and softness of local water supply. *British Medical Journal* **15** , 357-61. Rapporté dans McCullagh, P. et Nelder, J. A. Generalized Linear Models[1], 2nd Edition, Chapman & Hill 1989.

(ii) Déterminer la variance de l'estimateur $-\hat{\beta}_o/\hat{\beta}_1$ (utiliser le fait, pas fournis par le logiciel) que le coefficient de corrélation estimé entre $\hat{\beta}_o$ et $\hat{\beta}_1$ est -0.67638606. Pour déterminer une formule approximative, utiliser la technique qui consiste à remplacer une fonction $f(X;Y)$ de deux variables aléatoires X et Y de moyennes μ_X et μ_Y par la fonction linéaire $f(X ; Y) = f(\mu_X ; \mu_Y) + \frac{\partial f}{\partial X}(X-\mu_X)+\frac{\partial f}{\partial Y}(Y-\mu_Y)$, les dérivées partielles étant évaluées à $(\mu_X ; \mu_Y)$.

8.3 Les données suivantes portent sur des recrues à l'armée américaine.
Les variables sont : :

m nombre de recrues
y nombre de recrues qui ont complété l'école secondaire
age : 0 = <22; 1 = ≥ 22
race : 0 = blanc; 1 = noir
pere : scolarité du père :1 = pas de secondaire; 2 = école secondaire non terminée; 3 = école secondaire terminée

ID	n	y	pere	age	race
1	76	8	1	0	0
2	13	1	2	0	0
3	26	6	3	0	0
4	397	51	1	1	0
5	51	13	2	1	0
6	91	45	3	1	0
7	78	19	1	0	1
8	29	7	2	0	1
9	19	3	3	0	1
10	346	103	1	1	1
11	81	25	2	1	1
12	54	18	3	1	1

On tente d'expliquer la probabilité qu'une recrue complète le secondaire en fonction de son âge, sa race, et la scolarité du père.

a) Montrer qu'on peut se débarrasser des interactions triples.
b) Examiner tour à tour toutes les interactions doubles. Ne retenir que celles qui sont significatives.
c) Vous devriez aboutir à un modèle dans lequel l'âge n'a aucune interaction avec les autres prédicteurs. Dressez un tableau pour chaque groupe d'âge montrant les probabilités de succès estimées dans chaque combinaison de « pere » et « race ». Montrez comment se reflète l'absence d'interactions avec « age » en calculant des « odd-ratio » dans chaque tableau

8.4 Dans des conditions expérimentales bien contrôlées, une respiration profonde peut induire une constriction des vaisseaux sanguins de la peau des doigts. Deux facteurs mesurables peuvent contribuer à cette constriction : le volume V d'air inspiré, et le taux T d'inspiration. La constriction est difficile à mesurer ; donc on se contente de noter sa présence ou absence. Il semblerait que les odds $\pi/(1-\pi)$ dépendent du produit VT, ce qui donnerait pour modèle $\ln[\pi/(1-\pi)] = x_1 + x_2$. Mais on considère un modèle plus général, soit $\log[\pi/(1-\pi)] = \mu + \beta_1 x_1 + \beta_2 x_2$, où x_1 est le logarithme du volume et x_2 est le logarithme du taux d'inspiration. Les données sont présentées en annexe.

a) Déterminez une régression logistique et testez les hypothèses $\beta_1 = 0$ et $\beta_2 = 0$.
b) Testez simultanément les hypothèses $\beta_1 = 1$ et $\beta_2 = 1$
c) Testez l'hypothèse $\beta_1 = \beta_2$

Voici les données :

ID	volume	taux	y	ID	volume	taux	y
1	3.70	0.825	1	21	0.40	2.000	0
2	3.50	1.090	1	22	0.95	1.360	0
3	1.25	2.500	1	23	1.35	1.350	0
4	0.75	1.500	1	24	1.50	1.360	0
5	0.80	3.200	1	25	1.60	1.780	1
6	0.70	3.500	1	26	0.60	1.500	0
7	0.60	0.750	0	27	1.80	1.500	1
8	1.10	1.700	0	28	0.95	1.900	0
9	0.90	0.750	0	29	1.90	0.950	1
10	0.90	0.450	0	30	1.60	0.400	0
11	0.80	0.570	0	31	2.70	0.750	1
12	0.55	2.750	0	32	2.35	0.300	0
13	0.60	3.000	0	33	1.10	1.830	0
14	1.40	2.330	1	34	1.10	2.200	1
15	0.75	3.750	1	35	1.20	2.000	1
16	2.30	1.640	1	36	0.80	3.330	1
17	3.20	1.600	1	37	0.95	1.900	0
18	0.85	1.415	1	38	0.75	1.900	0
19	1.70	1.060	0	39	1.30	1.625	1
20	1.80	1.800	1				

8.5 On prélève les données suivantes sur les types sanguins de 250 personnes tirées d'une certaine population : O : 35, A : 45, B : 130, AB : 40. On suppose que ces observations sont une réalisation d'un vecteur de loi multinomiale de paramètres $n = 250$ et $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$, où $\pi_1 = (1-\alpha-\beta)^2$, $\pi_2 = 2\alpha(1-\beta)-\alpha^2$, $\pi_3 = 2\beta(1-\alpha)-\beta^2$, $\pi_4 = 2\alpha\beta$, avec les conditions $\alpha > 0$, $\beta > 0$ et $\alpha + \beta < 1$.

a) Montrez d'abord que ces π_i sont bien les probabilités des 4 phénotypes lorsqu'on suppose que α est la probabilité d'hériter A d'un parent et β est la probabilité d'hériter B.

b) Estimez α et β, et donc $\boldsymbol{\pi}$.

8.6 On distingue des individus de trois phénotypes. Selon un certain modèle génétique, ces phénotypes sont l'effet d'un seul gène dont les allèles sont *A* et *a*. Les phénotypes 1, 2 et 3 correspondent, respectivement, à *AA*, (*Aa* ou *aA*) et *aa*. Supposons que la probabilité qu'un individu hérite le gène *A* est θ, et qu'il hérite les deux allèles indépendamment l'un de l'autre. Dans un échantillon de taille $n = 109$, on trouve $X_1 = 10$, $X_2 = 53$ et $X_3 = 46$ individus des phénotypes 1, 2 et 3, respectivement, où $X_1 + X_2 + X_3 = n$. On sait alors que le triplet est de loi multinomiale de paramètres n et $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$, où $\pi_1 = \theta^2$, $\pi_2 = 2\theta(1-\theta)$ et $\pi_3 = (1-\theta)^2$.

a) Déterminez une expression pour l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ et évaluez $\hat{\theta}$ à partir des données.

b) Déterminez une expression pour la variance (exacte) de $\hat{\theta}$. Estimez $\mathrm{Var}(\hat{\theta})$ en substituant $\hat{\theta}$ à θ.

c) Déterminez un intervalle de confiance approximatif à 95% pour θ, en supposant que la variable $(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$ est de loi $\mathrm{N}(0 ; 1)$.

d) Évaluez les deux statistiques $G^2 = -2\log\lambda$ et $\chi^2 = \sum (np_i - n\hat{\pi}_i)^2 / n\hat{\pi}_i$ pour tester le modèle, c'est-à-dire, pour tester l'hypothèse que les paramètres π_1, π_2, π_3 ont la structure explicitée en a). [Si $L(\boldsymbol{p} ; \boldsymbol{\pi})$ est la fonction de vraisemblance, où $\boldsymbol{p}$ est le vecteur des fréquences observées, alors $\lambda = L(\boldsymbol{p} ; \boldsymbol{p})/L(\boldsymbol{p} ; \hat{\boldsymbol{\pi}})$, où $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\theta})$.]

e) Que concluez-vous à partir de G^2 et χ^2?

8.7 Le gène qui cause le daltonisme est situé dans la partie du chromosome X qui n'a pas d'homologue dans le chromosome Y. Le garçon ne reçoit donc qu'un gène, et il est daltonien si ce gène est récessif. Une fille, en revanche, reçoit deux gènes, et elle n'est daltonienne que si les deux sont récessifs. Par conséquent, si la probabilité qu'un homme soit daltonien est β, la probabilité qu'une femme le soit est β^2.

Pour tester ce modèle, on prélève un échantillon de 1000 personnes. On obtient les réultats suivants : $X_1 = 6$ femmes daltoniennes; $X_2 = 544$ femmes normales, $X_3 = 45$ hommes daltoniens, et $X_4 = 405$ hommes normaux.

a) Ajustez ces données à un modèle qui incorpore le fait, supposé connu, que la proportion des femmes est ½. En d'autres termes,

i) exprimez les paramètres π_1 , π_2 ,π_3 ,π_4 en fonction de β

ii) estimez β et ensuite π_1 , π_2 ,π_3 ,π_4

iii) estimez la variance de $\hat{\beta}$

iv) testez le modèle

b) Analysez les mêmes données, mais cette fois-ci, en procédant conditionnellement, c'est-à-dire, en supposant que le nombre d'hommes et de femmes n'est pas aléatoire, mais qu'il a été fixé à 450 et 550, respectivement :

i) Estimez β

ii) estimez la variance de $\hat{\beta}$

iii) testez le modèle

c) Supposez maintenant qu'il s'agit d'une population particulière dans laquelle α, la probabilité de tomber sur une femme, est inconnue et doit être estimée.

i) Estimez $\boldsymbol{\theta} = (\alpha , \beta)'$

ii) Estimez la matrice de covariance de $\hat{\boldsymbol{\theta}}$, l'estimateur de $\boldsymbol{\theta}$

iii) Testez le modèle

iv) Utilisez la statistique $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)' \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$, où $\hat{\mathbf{V}}$ est l'estimateur de la matrice de covariance de $\hat{\boldsymbol{\theta}}$, pour tester simultanément les hypothèses $\alpha = ½$ et $\beta = 0{,}1$.

8.8 Deux gènes liés se présentent sous la forme de deux allèles, A et *a*, et B et *b*, respectivement. Considérons un couple donc chaque membre est du génotype AB/*ab* (c'est-à-dire, un chromosome avec les gène A et B, et un chromosome avec les gènes *a* et *b*). Le descendant ne peut donc être que du phénotype AB ou *ab*, car A et B sont dominants. Supposons, cependant, qu'au moment de la formation des gamètes, les chromosomes peuvent se croiser, de sorte que le parent, au lieu de transmettre AB ou *ab*, transmet plutôt A*b* ou *a*B. Soit θ la probabilité qu'un croisement se produise. Le problème est d'estimer θ. On prélève un échantillon de 197 descendants, et on trouve la répartition suivante des phénotypes : AB : 125; A*b* : 18; *a*B : 20; *ab* : 34.

a) Estimez θ

b) Testez le modèle

c) Estimez la variance de $\hat{\theta}$.

d) Déterminez un intervalle de confiance approximatif pour θ.

8.9 Supposons que le nombre d'orignaux dans un ravage d'orignaux suit une loi de Poisson de paramètre θX, où X est la superficie du ravage. On prélève un échantillon de 6 ravages et on observe les couples (y_i , x_i) suivants, où y_i est le nombre d'orignaux dans le i^e ravage et x_i la superficie, en hectare, du ravage : (5 ; 0,6) , (12 ; 1,0) , (15 ; 2,1) , (8 ; 0,07) , (9 ; 0,4) , (13 ; 1,1).

a) Déterminez l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ , en utilisant la fonction de vraisemblance conditionnelle, étant donné les x_i.

b) Déterminez un estimateur sans biais (conditionnellement) de Var($\hat{\theta}$ |X_1, … , X_6), et montrez que c'est un estimateur sans biais de Var($\hat{\theta}$) .

8.10 Considérez la table de contingence 2×2

x_{11}	x_{12}	x_{1+}
x_{21}	x_{22}	x_{2+}
x_{+1}	x_{+2}	n

où les x_{ij} suivent une loi multinomiale de probabilités π_{ij}. Considérons l'hypothèse que les π_{ij} satisfont les contraintes suivantes (l'hypothèse d'indépendance) : $\pi_{11} = \alpha\beta$; $\pi_{12} = \alpha(1-\beta)$; $\pi_{21} = (1-\alpha)\beta$; $\pi_{22} = (1-\alpha)(1-\beta)$.

a) Montrez que les sommes marginales x_{1+} et x_{+1} sont des statistiques exhaustives pour α et β.

b) Déterminez les estimateurs du maximum de vraisemblance de α et β.

c) Déterminez la statistique du rapport de vraisemblance pour tester l'hypothèse d'indépendance.

8.11 On prélève un échantillon de 100 familles de 3 enfants. Voici les résultats :

FFF	130	FGG	32
FFG	52	GFG	41
FGF	30	GGF	50
GFF	55	GGG	110

Le problème est de tester l'hypothèse que le nombre de filles est vraiment de loi binomiale. Faites-le de deux façons :

a) En considérant la distribution du nombre de filles dans les 500 familles ;

b) En considérant la distribution du nombre d'enfants avant la première fille.

8.12 Un échantillon de 100 ampoules a été observé pendant 4 mois. 28 ont duré moins d'un mois; 20 ont duré entre 1 et 2 mois; 15 ont duré entre 2 et 3 mois; 11 entre 3 et 4 mois. Les 26 autres fonctionnaient encore à la fin des 4 mois.

a) Supposez que la durée d'une ampoule, en mois, suit une loi exponentielle de moyenne β, et estimez β par la méthode du maximum de vraisemblance.

b) Testez le modèle.

8.13 Supposons que le nombre de garçons dans une famille de n enfants suit une loi binomiale de paramètre p mais que la valeur de p varie d'un couple à l'autre. Supposons que pour un couple choisi au hasard, la probabilité p est une variable aléatoire de fonction de densité $f(p) = \alpha p^{(\alpha-1)}$, $0 < p < 1$, $\alpha > 0$.

Supposons que parmi 100 couples avec 2 enfants chacun, on trouve 30 couples avec 2 filles, 35 avec 2 garçons et 35 avec un garçon et une fille.

a) Estimez α dans le modèle

b) Testez le modèle.

8.14 Répétez l'exercice précédent en supposant cette fois-ci que p est une variable aléatoire de loi bêta de paramètres α et $\beta = \alpha$, c'est-à-dire, de densité $f(p) = [\Gamma(2\alpha)/\Gamma^2(\alpha)][p(1-p)]^{(\alpha-1)}$.

8.15 Reprenez les deux exercices précédents sans supposer que $\alpha = \beta$.

8.16 Les jumeaux humains sont de deux types : homozygotes et hétérozygotes. Soit α la probabilité que des jumeaux soient homozygotes, et β la probabilité qu'un enfant soit un garçon. La distribution du nombre de garçons est donc : 2 garçons : $\beta(\beta+\alpha\phi)$; 1 garçon : $2\beta\phi(1-\alpha)$; aucun garçon : $\phi(\alpha\beta+\phi)$, où $\phi = 1 - \beta$.

a) Supposons que dans un échantillon de 198 paires de jumeaux, on trouve 58 paires de deux garçons, 89 paires d'une fille et un garçon, et 51 paires de 2 filles. Estimez α et β, et la matrice de covariance des estimateurs.

b) Qu'est-ce qui empêche qu'on teste le modèle?

c) Estimez le paramètre α en supposant que β est connu et vaut 0,513, et testez le modèle.

8.17 Soit $y_1, \ldots, y_k$ k variables aléatoires indépendantes de loi de Poisson de paramètres $\lambda_1, \ldots, \lambda_k$, respectivement.

a) Montrer que conditionnellement étant donné $\sum_{i=1}^{k} y_j = n$, le vecteur $\boldsymbol{y} = [y_1, \ldots, y_k]$ est de loi multinomiale de paramètres n et $p_1, \ldots, p_k$, ou $p_i = \lambda_i \Big/ \sum_{j-1}^{k} \lambda_j$.

b) Soit $\boldsymbol{y} = [y_1, \ldots, y_k]$ un vecteur de loi multinomiale de paramètres n et $p_1, .., p_k$, $\boldsymbol{y}_1 = [y_1,..., y_{k_1}]$ et $\boldsymbol{y}_2 = [y_{k_1+1},..., y_k]$ ($k_1 < k$-1). Montrer que conditionnellement, étant donné $\sum_{i=1}^{k_1} y_i = n_1$ et $\sum_{i=k_1+1}^{k} y_i = n_2$, $\boldsymbol{y}_1$ et y_2 sont indépendantes, $\boldsymbol{y}_1$ de loi multinomiale de paramètres n_1 et $p_1 \Big/ \sum_{i=1}^{k_1} p_i$, … , $p_{k_1} \Big/ \sum_{i=1}^{k_1} p_i$; et $\boldsymbol{y}_2$ de loi multinomiale de paramètres n_2 et $p_{k_1+1} \Big/ \sum_{i=k_1+1}^{k} p_i$, … , $p_k \Big/ \sum_{i=k_1+1}^{k} p_i$.