REVUE DE STATISTIQUE APPLIQUÉE

S. IM

Analyse d'une variable binaire et de plusieurs variables continues

Revue de statistique appliquée, tome 33, n° 4 (1985), p. 15-28

<http://www.numdam.org/item?id=RSA_1985__33_4_15_0>

© Société française de statistique, 1985, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

ANALYSE D'UNE VARIABLE BINAIRE ET DE PLUSIEURS VARIABLES CONTINUES

S. IM

I.N.R.A., Laboratoire de Biométrie
Centre de Recherches de Toulouse,
B.P. 27, 31326 Castanet Tolosan

RÉSUMÉ

Cet article présente deux modèles pour analyser l'influence des facteurs à effets fixes sur la distribution conjointe d'une variable aléatoire binaire et de plusieurs variables aléatoires continues. Le premier modèle est basé sur la distribution conditionnelle de la variable aléatoire binaire par rapport aux variables aléatoires continues. Le second utilise la distribution des variables aléatoires continues par rapport à la variable aléatoire binaire. Des logiciels existants, GLIM et GENSTAT, peuvent être utilisés pour résoudre les deux modèles par la méthode du maximum de vraisemblance. Les deux modèles sont illustrés par un exemple d'application.

SUMMARY

Two fixed-effects models are considered for the analysis of data on binary and continuous variates. The first one is based on the conditional distribution of the binary variate given the continuous variates. In the second model the conditional distribution of the continuous variates given the binary variate is used. The maximum likelihood estimates can be obtained by using GLIM or GENSTAT. The two models are illustrated with a numerical application.

ANALYSE D'UNE VARIABLE BINAIRE ET DE PLUSIEURS VARIABLES CONTINUES

1. INTRODUCTION

L'étude des effets des facteurs sur la distribution des variables aléatoires continues se fait à l'aide du modèle linéaire gaussien. Pour analyser les données binaires on utilise couramment le modèle logit (COX, 1970) ou le modèle probit (FINNEY, 1971).

Dans certaines applications, les données disponibles sont des observations d'une variable aléatoire (v.a.) binaire et de plusieurs v.a. continues. Des auteurs ont considéré ce type de données en classification et analyse discriminante, par exemple : CHANG et AFIFI (1974), HOSMER *et al.* (1983), KRZANOWSKI (1975, 1980), TU et HAN (1982). Le modèle utilisé est basé sur la distribution conditionnelle des v.a. quantitatives par rapport à la v.a. binaire.

Mots clés : **Données binaires et continues, modèles linéaires généralisés, maximum de vraisemblance, GLIM, GENSTAT.**

Key words : **Binary and continuous data, generalized linear models, maximum likelihood, GLIM, GENSTAT.**

En économétrie, on trouve l'analyse conjointe d'une v.a. binaire et d'une v.a. continue dans MATHOT (1981), OLSEN (1978), SCHMIDT et STRAUSS (1976). La distribution conjointe est définie à partir de la distribution conditionnelle de la v.a. binaire par rapport à la v.a. continue et de la distribution conditionnelle de la v.a. continue par rapport à la v.a. binaire. Cette démarche implique des difficultés dues à une restriction sur les paramètres dans la méthode du maximum de vraisemblance. Elle ne nous semble pas être satisfaisante.

On rencontre également les données binaires et continues en Génétique : FOULLEY *et al.* (1983) ont proposé une méthode bayésienne de prédiction génétique à partir de ce type de données.

Ici on étudie l'influence des facteurs (à effets fixes) sur la distribution conjointe de la v.a. binaire et des v.a. continues. Pour cela on considère deux modèles qui diffèrent par la distribution conditionnelle utilisée. On suppose que la v.a. binaire admet une continuité sous-jacente mais avec un seuil qui lui impose une manifestation dichotomique. Le premier modèle constitue une extension du modèle de HANNAN et TATE (1965) pour tenir compte de la présence des facteurs. Il utilise la distribution conditionnelle de la v.a. binaire par rapport aux v.a. continues. Le second modèle est basé sur la distribution conditionnelle des v.a. continues par rapport à la v.a. binaire.

On utilise la méthode du maximum de vraisemblance pour résoudre le problème d'estimation et de test d'hypothèses.

2. EXEMPLE ET NOTATION

On reprend l'exemple de FOULLEY *et al.* (1983) en considérant que les pères sont les niveaux d'un facteur à effets fixes. La difficulté de vêlage est la v.a. binaire, le poids à la naissance et l'ouverture pelvienne sont les v.a. continues.

Le veau (l'unité statistique) i apporte les observations :

$z_i = \begin{cases} 1 \text{ si vêlage difficile;} \\ 0 \text{ si vêlage facile;} \end{cases}$

y_i^1 : poids à la naissance;

y_i^2 : ouverture pelvienne de la mère, pour $i = 1, \ldots, n$.

Les facteurs sont :

1) facteur origine des génisses (OR) ayant 2 niveaux notés i_1 ($i_1 = 1, 2$);

2) facteur père (PE) ayant 6 niveaux notés i_2 ($i_2 = 1, \ldots, 6$);

3) facteur saison de vêlage (SA) ayant 2 niveaux notés i_3 ($i_3 = 1, 2$);

4) facteur sexe (SE) ayant 2 niveaux notés i_4 ($i_4 = 1, 2$).

Les données sont regroupées dans la tableau 1.

On veut étudier les effets des 4 facteurs sur la distribution conjointe des v.a. difficulté de vêlage, poids à la naissance et ouverture pelvienne.

De façon générale, l'unité statistique i apporte une observation binaire z_i et un vecteur d'observations continues $y_i = (y_i^1, \ldots, y_i^p)'$, $i = 1, \ldots, n$.

TABLEAU 1

Données de vélage des génisses de race blonde d'Aquitaine contrôlées à Casteljaloux, France

Origine	Père	Saison	Ouverture pelvienne	Sexe	Poids à la naissance	Difficulté de vêlage
1	1	1	328.0	M	41.0	E
1	1	1	304.5	M	37.5	E
1	1	1	354.8	F	41.5	E
1	1	2	374.0	F	40.0	E
1	1	2	285.0	F	43.0	E
1	1	2	310.0	F	42.0	E
1	1	2	270.0	F	35.0	E
2	1	1	336.0	F	46.0	E
2	1	1	374.0	F	40.5	E
2	1	2	346.0	F	39.0	E
1	2	1	346.5	M	41.4	E
1	2	1	333.3	M	43.0	D
1	2	2	346.5	F	34.0	E
1	2	2	307.5	M	47.0	D
1	2	2	315.0	M	42.0	E
2	2	2	300.0	M	44.5	E
2	2	2	273.0	M	49.0	E
1	3	1	341.0	M	41.6	E
2	3	1	300.0	M	36.0	E
2	3	1	328.0	F	42.7	E
2	3	2	266.0	F	32.5	E
2	3	2	302.3	F	44.4	E
2	3	2	320.0	M	46.0	E
1	4	2	328.0	M	47.0	D
1	4	2	344.0	F	51.0	D
1	4	2	243.0	F	39.0	E
2	4	1	333.3	M	44.5	E
1	5	1	346.5	M	40.5	E
1	5	1	304.5	F	43.5	E
1	5	2	396.0	M	42.5	E
1	5	2	285.0	M	48.8	D
1	5	2	328.0	M	38.5	E
1	5	2	346.5	M	52.0	E
1	5	2	344.0	F	48.0	E
2	5	1	357.0	F	41.0	E
2	5	1	357.0	M	50.5	D
2	5	2	317.8	M	43.7	D
2	5	2	346.5	M	51.0	D
1	6	1	290.0	F	51.6	D
1	6	1	260.0	M	45.3	D
1	6	1	357.0	F	36.5	E
1	6	2	354.8	M	50.5	E
1	6	2	273.0	M	46.0	D
1	6	2	315.0	M	45.0	E
1	6	2	256.5	F	36.0	E
2	6	1	317.8	F	43.5	E
2	6	1	290.0	F	36.5	E

E = Velage facile (Easy en anglais)
D = " difficile (Difficult en anglais)

On note :

Z_i la v.a. dont z_i est une observation,
$Y_i = (Y_i^1, \ldots, Y_i^p)'$ le vecteur des v.a. dont y_i est une observation,
Y_i^0 une v.a. continue telle que : $Z_i = 1 \langle = \rangle Y_i^0 > 0$,
$\mu_i^0 = E(Y_i^0)$,
$\mu_i = (\mu_i^1, \ldots, \mu_i^p)' = E(Y_i)$,
$Y^j = (Y_1^j, \ldots, Y_n^j)'$ pour $j = 0, 1, \ldots, p$,
$Y = (Y^{1\prime}, Y^{n\prime}, \ldots, Y^{n\prime})'$,
$Z = (Z_1, \ldots, Z_n)'$,
$\sigma = (\sigma_1, \ldots, \sigma_p)'$ le vecteur des covariances entre Y_i^0 et Y_i^j $(j = 1, \ldots, p)$,
V la matrice des variances-covariances de $(Y_i^1, \ldots, Y_i^p)$,
Φ la fonction de répartition de la loi normale centrée réduite,
φ la densité de la loi normale centrée réduite,
$\mu_{iz} = E(Y_i | Z_i = z)$ $\quad z = 0, 1.$

3. MODÈLE I

On suppose que la v.a. binaire admet une v.a. sous jacente continue corrélée avec les v.a. continues. On est dans la situation d'un vecteur de v.a. dont l'une est discrétisée. On généralise le modèle de HANNAN et TATE (1965) en appliquant un modèle linéaire au vecteur des $(p + 1)$ v.a. continues.

Le modèle est le suivant :

$$\begin{bmatrix} Y_i^0 \\ Y_i \end{bmatrix} = \begin{bmatrix} \mu_i^0 \\ \mu_i \end{bmatrix} + \begin{bmatrix} U_i^0 \\ U_i \end{bmatrix} \qquad i = 1, \ldots, n$$

où $\begin{bmatrix} U_i^0 \\ U_i \end{bmatrix}$ $(i = 1, \ldots, n)$ sont des vecteurs aléatoires indépendants et de même loi normale $N(0, \Sigma)$, avec $\Sigma = \begin{bmatrix} 1 & \sigma' \\ \sigma & V \end{bmatrix}$;

$$\mu_i^j = a_{i1}\beta_1^j + a_{i2}\beta_2^j + \ldots + a_{iq}\beta_q^j$$

pour $j = 0, 1, \ldots, p$; $\quad i = 1, \ldots, n$.

Soient β^0 le vecteur des paramètres β_k^0 $(k = 1, \ldots, q)$;
β le vecteur des paramètres β_k^j $(k = 1, \ldots, q; j = 1, \ldots, p)$.

Les paramètres du modèle, a priori inconnus, sont β^0, β, σ et V. Les coefficients $\{a_{ik}\}$ sont des valeurs réelles connues.

Ce modèle ne comporte pas de paramètres aléatoires. FOULLEY *et al.* (1983) ont considéré un modèle qui contenait des paramètres fixes et des paramètres aléatoires. Supposant connues les variances et covariances, ils ont proposé d'estimer les paramètres par les modes a postériori. Cette méthode bayésienne leur a permis de circonvenir la distinction entre les paramètres fixes et les paramètres aléatoires. Ils ont donné des indications pour résoudre le problème, très difficile, d'estimation des variances et covariances.

Ici la méthode d'estimation et de test d'hypothèses utilisée est celle du maximum de vraisemblance.

3.1. Fonction de vraisemblance

Les données disponibles sont les observations $\begin{bmatrix} z_i \\ y_i \end{bmatrix}$ $(i = 1, \ldots, n)$.

Les vecteurs aléatoires $\begin{bmatrix} Y_i^0 \\ Y_i \end{bmatrix}$ $(i = 1, \ldots, n)$ sont indépendants et la distribution de $\begin{bmatrix} Y_i^0 \\ Y_i \end{bmatrix}$ est normale $N\left(\begin{pmatrix} \mu_i^0 \\ \mu_i \end{pmatrix}, \Sigma\right)$.

La distribution marginale de Z_i est définie par :

$$P(Z_i = 1) = P(Y_i^0 > 0) = 1 - \Phi(-\mu_i^0)$$

La distribution conditionnelle de Y_i^0, sachant $Y_i = y_i$, est normale

$$N(\mu_i^0 + \sigma' V^{-1}(y_i - \mu_i), (1 - \sigma' V^{-1} \sigma)).$$

On en déduit la distribution conditionnelle de Z_i sachant $Y_i = y_i$:

$$P(Z_i = 1 | Y_i = y_i) = P(Y_i^0 > 0 | Y_i = y_i)$$

En posant $$\delta_i = \mu_i^0 + \sigma' V^{-1}(y_i - \mu_i)$$

et $$\gamma = (1 - \sigma' V^{-1} \sigma)^{1/2}$$

On a :

$$P(Z_i = 1 | Y_i = y_i) = \theta_i(y_i) = 1 - \Phi(-\delta_i/\gamma)$$

La densité de $\begin{bmatrix} Z_i \\ Y_i \end{bmatrix}$ est :

$$f_i(y_i, z_i) = g_i(y_i)\, h(z_i | y_i)$$

où $g_i(y_i) = g(y_i ; \mu_i, V)$ est la densité de la distribution normale $N(\mu_i, V)$; et

$$h_i(z_i | y_i) = [\theta_i(y_i)]^{z_i} [1 - \theta_i(y_i)]^{1 - z_i}$$

La fonction de vraisemblance au point d'observation (z, y) est :

$$\ell(z, y; \beta^0, \beta, \sigma, V) = \prod_{i=1}^{n} [f_i(y_i, z_i)]$$

Le logarithme de la vraisemblance est :

$$L(z, y; \beta^0, \beta, \sigma, V) = L_1(y; \beta, V) + L_2(z, y; \beta^0, \beta, \sigma, V)$$

où :

$$L_1(y; \beta, V) = \sum_{i=1}^{n} \log g_i(y_i)$$

ne dépend que de β et V;

$$L_2(z, y; \beta^0, \beta, \sigma, V) = \sum_{i=1}^{n} \log h_i(z_i | y_i)$$

dépend de tous les paramètres.

3.2. Méthode du maximum de vraisemblance

Les estimations du maximum de vraisemblance $\hat{\beta}^0$, $\hat{\beta}$, $\hat{\sigma}$ et $\hat{V}$ des paramètres β^0, β, σ et V sont telles que :

$$L(z, y; \hat{\beta}^0, \hat{\beta}, \hat{\sigma}, \hat{V}) = \underset{\beta^0, \beta, \sigma, V}{\text{Max}} L(z, y; \beta^0, \beta, \sigma, V)$$

Outre les propriétés asymptotiques, la méthode du maximum de vraisemblance possède une autre propriété intéressante, celle d'invariance fonctionnelle. Des auteurs (HOADLEY (1971), PHILIPPOU et ROUSSAS (1975)) ont étudié les propriétés asymptotiques des estimateurs du maximum de vraisemblance dans le cas où l'on dispose d'observations de v.a. indépendantes et non identiquement distribuées.

Ici, une transformation judicieuse des paramètres permet de simplifier substantiellement le problème de maximisation précédent. Soient

$$b = (b_1, \ldots, b_p)' = V^{-1} \sigma/\gamma$$

$$\beta_k^* = \sum_{j=1}^{p} b_j \beta_k^j - \beta_k^0/\gamma \qquad k = 1, \ldots, q$$

$$\beta^* = (\beta_1^*, \ldots, \beta_q^*)'$$

Cette transformation permet de séparer la maximisation du logarithme de la vraisemblance en deux maximisations.

Pour le nouveau système de paramètres (β^*, b, β, V), le logarithme de la vraisemblance s'écrit :

$$L(z, y; \beta^*, b, \beta, V) = L_1(y; \beta, V) + L_2(z, y; \beta^*, b)$$

où : $L_1(y; \beta, V)$ est le logarithme de la vraisemblance d'un modèle linéaire gaussien multidimensionnel; $L_2(z, y; \beta^*, b)$ est le logarithme de la vraisemblance d'un modèle binomial avec :

$$\Phi(-\delta_i/\gamma) = \Phi\left(\sum_{k=1}^{q} a_{ik} \beta_k^* - \sum_{j=1}^{p} b_j y_i^j\right) \qquad i = 1, \ldots, n$$

Pour obtenir les estimations du maximum de vraisemblance de (β^*, b, β, V), on maximise $L_1(y; \beta, V)$ par rapport à (β, V) et $L_2(z, y; \beta^*, b)$ par rapport à (β^*, b).

Dans ce modèle, on a le même modèle de plan d'expérience (les mêmes coefficients $\{a_{ik}\}$ pour toute v.a. continue). Par conséquent, chaque $\hat{\beta}^j$ est l'estimation du maximum de vraisemblance de β^j dans le modèle linéaire unidimensionnel

$$Y_i^j = \mu_i^j + U_i^j \qquad i = 1, \ldots, n$$

qui, comme le modèle probit, appartient à la classe des modèles linéaires généralisés définis par NELDER et WEDDERBURN (1972). Cette classe de modèles contient les modèles de régression linéaire et d'analyse de variance, les modèles logit et probit pour les données binomiales et le modèle log-linéaire pour les tables de contingence. Les modèles linéaires généralisés sont largement développés par Mc CULLAGH ET NELDER (1983). Les moyens numériques pour traiter les modèles linéaires généralisés existent dans GENSTAT (ASTIER *et al.*, 1982) et dans GLIM (BAKER et NELDER, 1978).

L'estimation du maximum de vraisemblance de V est :

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$$

Les estimations du maximum de vraisemblance de β^0, σ s'obtiennent en utilisant la transformation réciproque

$$\sigma = Vb\,(1 + b' Vb)^{-1/2}$$

$$\beta_k^0 = \gamma \left(\sum_j b_j \beta_k^j - \beta_k^* \right) \qquad k = 1, \ldots, q$$

Le logarithme de la vraisemblance maximale est :

$$L(z, y; \hat{\beta}^*, \hat{b}, \hat{\beta}, \hat{V}) = L_1(y; \hat{\beta}, \hat{V}) + L_2(z, y; \hat{\beta}^*, \hat{b})$$

GLIM ou GENSTAT donnent la valeur de $-2\,L_2(z, y; \hat{\beta}^*, \hat{b})$.

D'autre part :

$$L_1(y; \hat{\beta}, \hat{V}) = -\frac{np}{2} \text{Log}\,(2\pi) - \frac{n}{2} \text{Log}\,(\text{dét}\,\hat{V}) - \frac{n}{2}$$

Donc, on obtient facilement la valeur de la statistique du test du rapport de vraisemblance.

Dans FOULLEY *et al.* (1983), les paramètres b et V sont supposés connus.

3.3. Prédiction de la valeur Y_i^0

La v.a. Y_i^0 n'est pas observable, on peut vouloir chercher à prédire sa valeur contionnellement à $Z_i = z_i$ et $Y_i = y_i$. Lorsque les paramètres du modèle sont connus, un bon prédicteur de la valeur de Y_i^0 est $E(Y_i^0 \mid Z_i = z_i, Y_i = y_i)$.

La distribution conditionnelle de Y_i^0, sachant $Y_i = y_i$, est normale $N(\delta_i, \gamma^2)$. On en déduit :

$$E(Y_i^0 \mid Z_i = z_i, Y_i = y_i) = \delta_i + \gamma\varphi(-\delta_i/\gamma)/[z_i - \Phi(-\delta_i/\gamma)]$$

Si les paramètres sont inconnus on peut retenir

$$\hat{E}(Y_i^0 \mid Z_i = z_i, Y_i = y_i) = \hat{\delta}_i + \hat{\gamma}\varphi(-\hat{\delta}_i/\hat{\gamma})/[z_i - \Phi(-\hat{\delta}_i/\hat{\gamma})]$$

comme prédicteur de Y_i^0.

On peut se servir des prédicteurs $\hat{E}(Y_i^0 \mid Z_i = z_i, Y_i = y_i)$ $(i = 1, \ldots, n)$ pour classer les niveaux des facteurs.

4. MODÈLE II

Ce modèle utilise la distribution conditionnelle des v.a. continues par rapport à la v.a. binaire.

On fait les suppositions suivantes :

1) les vecteurs aléatoires $\begin{pmatrix} Z_i \\ Y_i \end{pmatrix}$ $(i = 1, \ldots, n)$ sont indépendants;

2) la distribution de la v.a. Z_i est définie par :

$$P(Z_i = z) = \theta_i^z (1 - \theta_i)^{1-z} \qquad z = 0, 1$$

avec $$\theta_i = \Phi (a_{i1} \alpha_1^0 + a_{i2} \alpha_2^0 + \ldots + a_{iq} \alpha_q^0) ;$$

3) sachant $Z_i = z$, la distribution conditionnelle de Y_i est normale $N(\mu_{iz}, Q)$,
avec

$$\mu_{iz} = (\mu_{iz}^1, \ldots, \mu_{iz}^p)'$$

$$\mu_{iz}^j = a_{i1} \alpha_{z1}^j + a_{i2} \alpha_{z2}^j + \ldots + a_{iq} \alpha_{zq}^j$$

pour $i = 1, \ldots, n$; $z = 0, 1$; $j = 1, \ldots, p$.

Les constantes $\{a_{ik}\}$ sont des valeurs réelles connues. Les paramètres inconnus du modèle sont $\{\alpha_k^0\}$, $\{\alpha_{zk}^j\}$ et Q.

Comme dans le modèle I, on peut utiliser la méthode du maximum de vraisemblance pour résoudre le problème d'estimation et de test d'hypothèses. Soient

α^0 le vecteur des paramètres $\{\alpha_k^0\}$

α le vecteur des paramètres $\{\alpha_{zk}^j\}$.

La fonction de vraisemblance s'écrit :

$$\ell (z, y; \alpha^0, \alpha, Q) = \prod_{i=1}^{n} [\theta_i^{z_i} (1 - \theta_i)^{1-z_i} g (y_i ; \mu_{iz_i}, Q)]$$

Le logarithme de la vraisemblance est :

$$L (z, y; \alpha^0, \alpha, Q) = \sum_i (z_i \operatorname{Log} \theta_i + (1 - z_i) \operatorname{Log} (1 - \theta_i + \sum_i \operatorname{Log} g (y_i ; \mu_{iz_i}, Q)$$

Soient

$$L_1(z, \alpha^0) = \sum_{i=1}^{n} [z_i \operatorname{Log} \theta_i + (1 - z_i) \operatorname{Log} (1 - \theta_i$$

$$L_{20}(z, y; \alpha_0, Q) = \sum_{i=1}^{n} [(1 - z_i) \operatorname{Log} g (y_i ; \mu_{i0}$$

$$L_{21}(z, y; \alpha_1, Q) = \sum_{i=1}^{n} [z_i \operatorname{Log} g (y_i ; \mu_{i1}$$

On a :

$$L (z, y; \alpha^0, \alpha, Q) = L_1(z; \alpha^0) + L_{20}(z, y; \alpha_0, Q) + L_{21}(z, y; \alpha_1$$

On peut utiliser GLIM ou GENSTAT pour obtenir les estimations du maximum de vraisemblance des paramètres du modèle II.

$L_1(z; \alpha^0)$ est le logarithme de la vraisemblance d'un modèle linéaire généralisé standard, de distribution binomiale et de fonction de lien probit.

$\hat{\alpha}_z$ s'obtient de la même façon que $\hat{\beta}$ dans le modèle I pour les observations y_i telles que $z_i = z$, pour $z = 0, 1$.

L'estimation du maximum de vraisemblance de Q est :

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_{iz_i}) (y_i - \hat{\mu}_{iz_i})'$$

où μ_{iz_i} est l'estimation du maximum de vraisemblance de μ_{iz_i}.

GENSTAT offre une grande variété d'opérations matricielles qui ne sont pas disponibles dans GLIM3.

Donc, si le nombre de v.a. continues est grand, GENSTAT est préférable à GIM3 pour traiter les modèles I et II.

La qualité d'ajustement des deux modèles peut être mesurée par la quantité $-2\hat{L}$ (déviance), où $\hat{L}$ est le maximum de logarithme de la vraisemblance.

5. RÉSULTATS NUMÉRIQUES

Les modèles I et II sont appliqués à l'exemple de FOULLEY *et al.* (1983), présenté dans le paragraphe 2.

5.1. Modèle I

Le modèle linéaire de mon interaction de tous ordres entre les facteurs s'écrit :

$$\mu_i^j = m^j + or^j(i_1) + pe^j(i_2) + sa^j(i_3) + se^j(i_4)$$

avec :

$$or^j(1) = pe^j(1) = sa^j(1) = se^j(1) = 0$$

pour $j = 0, 1, 2$.

Soient

$$m^* = b_1 m^1 + b_2 m^2 - m^0/\gamma$$
$$or^*(i_1) = b_1 or^1(i_1) + b_2 or^2(i_1) - or^0(i_1)/\gamma$$
$$pe^*(i_2) = b_1 pe^1(i_2 + b_2 pe^2(i_2) - pe^0(i_2)/\gamma$$
$$sa^*(i_3) = b_1 sa^1(i_3 + b_2 sa^2(i_3) - sa^0(i_3)/\gamma$$
$$se^*(i_4) = b_1 se^1(i_4 + b_2 se^2(i_4) - se^0(i_4)/\gamma$$

Les probabilités conditionnelles $\theta_i(y_i)$ s'écrivent :

$$\theta_i(y_i) = 1 - \Phi [m^* + or^*(i_1) + pe^*(i_2) + sa^*(i_3) + se^*(i_4) - b_1 y_i^1 - b_2 y_i^2]$$

L'estimation des paramètres b_1 et b_2 ne pose aucun problème particulier. Les estimations du maximum de vraisemblance des paramètres m*, or*, pe*, sa*, se*, b_1 et b_2 ainsi que les écarts-types estimés des estimateurs sont donnés dans le tableau 2. On remarque que l'estimation du maximum de vraisemblance de b_2 est voisine de la vapeur proposée par FOULLEY *et al.* (1983) : $b_2 = -0{,}0184$.

a) Classement des pères

Dans le modèle précédent les probabilités conditionnelles $\theta_i(y_i)$ sont fonctions décroissantes des paramètres $pe^*(i_2)$. Le classement des pères selon les valeurs décroissantes de $\hat{pe}^*(i_2)$ est : père 3, père 1, père 6, père 2, père 5, père 4.

TABLEAU 2

Estimation du maximum de vraisemblance

Paramètres	Estimations	Ecarts-types estimés
m^*	6,758	9,524
$or^*(2)$	0,533	0,7344
$pe^*(2)$	- 2,550	8,667
$pe^*(3)$	0,190	14,45
$pe^*(4)$	- 3,499	8,682
$pe^*(5)$	- 3,079	8,667
$pe^*(6)$	- 2,412	8,680
$sa^*(2)$	0,859	0,673
$se^*(2)$	1,220	0,837
b_1	0,2319	0,0902
b_2	- 0,0189	0,0112

La moyenne des probabilités conditionnelles de vêlage difficile pour le père ℓ est :

$$\pi_\ell = \sum_i \delta_i(\ell)\, \theta_i(y_i) \,/\, \sum_i \delta_i(\ell) \qquad \text{pour} \qquad \ell = 1, \ldots, 6;$$

où $\delta_i(\ell) = \begin{array}{l} 1 \text{ pour tout i tel que } i_2 = \ell \\ 0 \text{ pour tout i tel que } i_2 \neq \ell \end{array}$

Les estimations du maximum de vraisemblance des π_ℓ sont :

$$\begin{array}{ll} \hat{\pi}_1 = 0{,}3745 \times 10^{-4} & \hat{\pi}_4 = 0{,}4990 \\ \hat{\pi}_2 = 0{,}2751 & \hat{\pi}_5 = 0{,}3593 \\ \hat{\pi}_3 = 0{,}3178 \times 10^{-4} & \hat{\pi}_6 = 0{,}3349 \end{array}$$

Le classement des pères selon les valeurs des $\hat{\pi}_\ell$ est : 1 $\simeq$ père 3, père 2, père 6 $\simeq$ père 5, père 4.

La moyenne estimée des espérances conditionnelles $E(Y_i^0 \mid Z_i = z_i, Y_i = Y_i)$ pour le père ℓ est :

$$\hat{V}_\ell = \sum_i \delta_i(\ell)\, \hat{E}(Y_i^0 \mid Z_i = z_i, Y_i = y_i) \,/\, \sum_i \delta_i(\ell)$$

Les valeurs de $\hat{V}_\ell$ sont :

$$\begin{array}{ll} \hat{V}_1 = -0{,}5495 & \hat{V}_4 = 0{,}0000 \\ \hat{V}_2 = -0{,}2355 & \hat{V}_5 = -0{,}1499 \\ \hat{V}_3 = -0{,}5495 & \hat{V}_6 = -0{,}1832 \end{array}$$

Le classement des pères selon les valeurs des $\hat{V}_\ell$ est le même que celui basé sur les $\hat{\pi}_\ell$.

b) Test du rapport de vraisemblance

La valeur de la statistique de test du rapport de vraisemblance de l'hypothèse $b_1 = b_2 = 0$ contre $b_1 \neq 0$ ou $b_2 \neq 0$ est de 11,27. Au niveau 0,05 on rejette l'hypothèse $b_1 = b_2 = 0$.

Pour étudier les effets des facteurs, on utilise les notations formelles de GLIM (ou de GENSTAT) pour écrire les modèles linéaires sur les espérances mathématiques μ_i^j.
Par exemple, le modèle de non interaction de tous ordres entre les facteurs s'écrit OR + PE + SA + SE.

Le tableau 3 donne la déviance $(-2L(z, y; \hat{\gamma}, \hat{b}, \hat{\beta}, \hat{V}))$ et le nombre de paramètres pour différents modèles linéaires sur les espérances mathématiques μ_i^j.

Au niveau 0,05 le test du rapport de vraisemblance conduit à l'acceptation de l'hypothèse que les facteurs sont sans effets.

TABLEAU 3

Déviance pour le modèle I

Sous-modèle	Déviance	Nombre de paramètres
OR + PE + SA + SE	701,0	32
OR + PE + SA	706,0	29
OR + PE	711,2	26
PE	711,6	23
	732,2	8

5.2. Modèle II

a) Classement des pères

On peut classer les pères selon la moyenne des probabilités conditionnelles de vêlage difficile. Pour le veau i, on a :

$$\theta_i(y_i) = P(Z_i = 1 \mid Y_i = y_i)$$
$$= \theta_i g(y_i ; \mu_{i1}, Q)/[\theta_i g(y_i ; \mu_{i1}, Q) + (1 - \theta_i) g(y_i ; \mu_{i0}, Q)]$$
$$= 1/\left[1 + \frac{1 - \theta_i}{\theta_i} \exp \frac{1}{2} [(y_i - \mu_{i1})' Q^{-1} (y_i - \mu_{i1}) - (y_i \mu_{i0})' Q^{-} (y_i - \mu_{i0})]\right]$$

La moyenne pour les descendants du père ℓ est :

$$\pi_\ell = \sum_i \delta_i(\ell)\, \theta_i(y_i) / \sum_i \delta_i(\ell), \qquad \ell = 1, \ldots, 6.$$

Sous le modèle de non interaction de tous ordres entre les facteurs, les estimations du maximum de vraisemblance des π_ℓ sont :

$$\hat{\pi}_1 = 0{,}6629 \times 10^{-4} \quad \hat{\pi}_4 = 0{,}5246$$
$$\hat{\pi}_2 = 0{,}2199 \quad \hat{\pi}_5 = 0{,}2912$$
$$\hat{\pi}_3 = 0{,}3057 \times 10^{-4} \quad \hat{\pi}_6 = 0{,}3070$$

Le classement des pères selon les valeurs de $\hat{\pi}_\ell$ est : père 1 $\simeq$ père 3, père 2, père 5 $\simeq$ père 6, père 4. Ce classement est pratiquement le même que celui donné par le modèle I.

Pour les mêmes coefficients $\{a_{ik}\}$, le modèle II comporte plus de paramètres que le modèle I et peut poser des problèmes d'identifiabilité. Dans l'exemple considéré, on a imposé des contraintes supplémentaires : les paramètres représentant les effets des pères 3 et 6 sur les espérances mathématiques $\mu_{i\cdot}^{j}$ sont nuls.

b) Test du rapport de vraisemblance

Le tableau 4 donne la déviance et le nombre de paramètres pour différents modèles linéaires sur les espérances mathématiques $\mu_{iz_i}^{j}$.

Au niveau 0,05 le test de OR + PE + SA contre OR + PE + SA + SE est légèrement significatif. Les autres tests conduisent aux mêmes conclusions que dans le modèle I.

Pour les deux modèles, le test du rapport de vraisemblance conduit à l'acceptation de l'hypothèse que les facteurs sont sans effets. Le classement des pères n'est donné qu'à titre d'illustration. En comparant la déviance dans le tableau 3 et 4, le choix entre le modèle I et le modèle II n'apparaît pas de façon évidente. Cependant, si on s'intéresse au classement des niveaux d'un facteur basé sur les probabilités conditionnelles $\theta_i(y_i)$ on peut préférer le modèle I au modèle II. En effet le modèle I contient moins de paramètres et d'après les résultats de ALTHAM (1984), l'estimation d'une fonction des $\theta_i(y_i)$ y est plus précise.

TABLEAU 4

Déviance pour le modèle II

Sous-modèle	Déviance	Nombre de paramètres
OR + PE + SA + SE	678,6	44
OR + PE + SA	690,8	39
OR + PE	698,4	34
PE	702,2	29
	732,4	8

DISCUSSION

Dans cet article nous avons présenté deux modèles pour étudier l'influence des facteurs à effets fixes sur la distribution conjointe d'une v.a. binaire et de plusieurs v.a. continues. Il existe des techniques de tests d'hypothèses non emboîtées (COX (1961, 1962), GOURIEROUX *et al.* (1983)) qui permettent de choisir entre les deux modèles. Mais la mise en œuvre de ces techniques est très laborieuse.

Nous avons montré que de bons logiciels existants, GLIM et GENSTAT, permettaient de traiter numériquement les deux modèles par la méthode du maximum de vraisemblance. Les difficultés d'obtention des estimations du maximum de vraisemblance dans le modèle utilisé par MATHOT (1981) n'apparaissent pas dans les deux modèles I et II.

Les modèles I et II sont appliqués à l'exemple de FOULLEY *et al.* (1983) en considérant que les pères sont les niveaux d'un facteur à effets fixes. Il peut être plus approprié de considérer que les effets des pères sont aléatoires comme l'ont fait FOULLEY *et al.* (1983). Cependant, la version « modèle mixte » du modèle I est beaucoup plus difficile à résoudre. FOULLEY *et al.* (1983) n'ont pas complètement résolu le problème. En particulier, on ne possède pas de bons estimateurs des composantes de variances et on ne connaît pas l'influence d'une variation de ces composantes de variances sur l'estimation des effets fixes et aléatoires.

Dans les modèles I et II, le problème de test d'hypothèses est résolu en utilisant le test du rapport de vraisemblance. Les techniques de test ne sont pas encore disponibles dans la version « modèle mixte » du modèle I.

RÉFÉRENCES

P.M.E. ALTHAM (1984). — Improving the precision of estimation by fitting a model. *J. Roy. Stat. Soc. B,* 46, 118-119.

R. ASTIER, A. BOUVIER, J. COURSOL, J.B. DENIS, C. DERVIN, E. JOLIVET, E. LESQUOY, O. PONS, R. TOMASSONE et J.P. VILA (1982). — Genstat : un langage statistique. INRA, Versailles.

R.J. BAKER, J.A. NELDER (1978). — The GLIM system, Release 3. Oxford : Numerical algorithms group.

P.C. CHANG, A.A. AFIFI (1974). — Classification based on dichotomous and continuous variables. *J. Am. Stat. Assoc.,* 69, 333-339.

D.R. COX (1961). — Tests of separate families of hypotheses. In *Proceedings of the fourth Berkeley symposium,* Vol. 1, 105-123.

D.R. COX (1962). — Further results on tests of separate families of hypotheses. *J. Roy. Stat. Soc. B,* 24, 406-424.

D.R. COX (1970). — *The analysis of Binary Data.* Chapman and Hall, London.

D.J. FINNEY (1971). — *Probit analysis,* 3rd ed., Cambridge University Press.

J.L. FOULLEY, D. GIANOLA, R. THOMPSON (1983). — Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Génét. Sél. Evol.,* 15, 401-424.

Ch. GOURIEROUX, A. MONFORT, A. TROGNON (1983). — Testing nested or non-tested hypotheses. *Journal of Econometrics,* 21, 83-115.

J.F. HANNAN, R.F. TATE (1965). — Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized. *Biometrika,* 52, 664-668.

B. HOADLEY (1971). — Asymptotic properties of maximum likelihood estimations for the independant not identically distributed case. *Ann. Math. Stat.,* 42, 1977-1981.

T. HOSMER, D. HOSMER, L. FISHER (1983). — A comparison of the maximum likelihood and discrimination function estimators of the coefficients of the logistic regression model for mixed continuous and discrete variables. *Comm. in Stat., Simul. and Comp.,* 12, 23-43.

W.J. KRZANOWSKI (1975). — Discrimination and classification using both binary and continuous variables. *J. Am. Stat. Assoc.,* 70, 782-790.

W.J. KRZANOWSKI (1980). — Mixtures of continuous and categorical variables in discriminant analysis. Biometrics, 36, 493-499.

F. MATHOT (1981). — L'utilisation du crédit lors de l'achat d'une voiture. *Annales de l'INSEE,* 44, 121-147.

P. McCULLAGH, J.A. NELDER (1983). — Generalized linear models. Chapman and Hall, London.

J.A. NELDER, R.W.M. WEDDERBURN (1972). — Generalized linear models. *J. Roy. Stat. Soc. A,* 135, 370-384.

I. OLKIN, R.F. TATE (1961). — Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.,* 32, 448-465.

R.J. OLSEN (1978). — Comment on "The effect of unions on earnings and earnings on unions : A mixed logit approach". *International Economic Review,* 19, 259-261.

A.N. PHILIPPOU, G.G. ROUSSAS (1975). — Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case. *Ann. Inst. Stat. Math.,* 27, 45-55.

P. SCHMIDT, R.P. STRAUSS (1976). — The effect of unions on earnings and earnings on unions : A mixed logit approach. *International Economic Review,* 17, 204-212.

R.F. TATE (1955). — The theroy of correlation between two continuous variables when one is dichotomized. *Biometrika,* 42, 205-216.

C.T. TU, C.P. HAN (1982). — Discriminant analysis based on binary and continuous variables. *J. Am. Stat. Assoc.,* 77, 447-454.