REVUE DE STATISTIQUE APPLIQUÉE

S. IM

Modèle log-linéaire et modèle de Cox dans l'analyse d'une table de contingence

Revue de statistique appliquée, tome 34, n° 4 (1986), p. 5-16

<http://www.numdam.org/item?id=RSA_1986__34_4_5_0>

MODÈLE LOG-LINÉAIRE ET MODÈLE DE COX DANS L'ANALYSE D'UNE TABLE DE CONTINGENCE

S. IM

Laboratoire de Biométrie,
INRA, BP 27, 31326 Castanet Tolosan

RÉSUMÉ

Dans cet article, le modèle log-linéaire et le modèle à risques proportionnels sont utilisés pour analyser une table de contingence à trois dimensions. On considère qu'une réponse discrète, obtenue par partition d'une variable continue, est observée sur les combinaisons des niveaux de deux factuers. La méthode d'estimation et de test utilisée est celle du maximum de vraisemblance. Les deux modèles sont comparés sur un exemple d'application.

ABSTRACT

In this paper the log-linear model and the proportional hazards model are used to analyse a three-dimensional contingency table obtained when a response is observed for each treatment combination of two factors. The response variable is obtained by partition of a continuous variable. The estimation and test procedure is derived by maximizing likelihoods. The two models are compared on an example.

Mots clés : **Modèle de Cox, Modèle log-linéaire, Table de contingence, Maximum de vraisemblance, Interaction multiplicative.**

Key words : **Cox model, Log linear model, Contingency table, Maximum likelihood, Multiplicative interaction.**

I. INTRODUCTION

On considère l'analyse d'une grande table de contingence à trois dimensions dans le cas où une variable réponse est observée sur les combinaisons des niveaux de deux facteurs. La variable réponse ordinale est obtenue par discrétisation d'une variable continue positive. L'analyse d'une telle table de contingence a pour but d'étudier l'influence des facteurs sur la distribution de la variable réponse.

On se propose de comparer, à l'aide d'un exemple, deux modèles d'analyse de ce type de table de contingence : modèle log-linéaire et modèle multiplicatif de Cox.

On trouve maintenant une littérature abondante sur le modèle log-linéaire : BISHOP, FIENBERG and HOLLAND (1975), HABERMAN (1974), PLACKETT (1981) et UPTON (1978). Ici, on met l'accent sur l'interprétation des paramètres et sur les difficultés numériques dans l'analyse d'une grande table de contingence.

Le modèle multiplicatif de Cox (COX, 1972), modèle à risques proportionnels, est souvent utilisé dans l'analyse des données de survie. Lorsqu'on l'applique à l'analyse des données groupées l'estimation des paramètres de ce modèle par la méthode du maximum de vraisemblance peut se ramener à un problème d'estimation d'un modèle linéaire généralisé. De bons logiciels existants, GLIM et GENSTAT, peuvent être utilisés pour résoudre ce problème. Cependant, le grand nombre de paramètres ne permet pas de réaliser matériellement le test de non-interaction entre les facteurs. On adopte une solution qui consiste à modéliser l'interaction comme étant proportionnelle au produit des termes des effets principaux.

2. DONNÉES

On a discrétisé en périodes une variable continue positive, délai de retour en chaleur après insémination. On veut étudier les effets de deux facteurs, taureau (F) et inséminateur (G), sur la répartition des retours en chaleur. Pour chaque couple (taureau, inséminateur), on dispose de fréquences observées de retour en chaleur après insémination dans différentes périodes. D'un fichier volumineux communiqué par J.B. DENIS et P. HUMBLOT, on a extrait les données pour les taureaux de la race 2 (26 taureaux et 60 inséminateurs).

Les données sont, en petite partie, représentées dans le tableau 1. L'examen de ce tableau montre qu'il n'est pas raisonnable de supposer que la distribution du délai de retour en chaleur est normale.

Les traitements sont les couples (i, j) constitués d'un niveau i de F et d'un niveau j et G. La distribution de la réponse R connaissant le traitement (i, j) est exprimée par :

$$\text{pr}\,(R = k;\quad F = i,\quad G = j) = p_{ijk}$$
$$(i = 1,..., I;\quad j = 1,..., J;\quad k = 1,..., K).$$

Pour le traitement (i, j), on a expérimenté n_{ij} vaches, et x_{ijk} est la fréquence de l'événement R = k (le délai de retour est dans la période k). Pour chaque (i, j), les variables aléatoires $\{X_{ijk}\}$ suivent conjointement une distribution multinomiale de paramètres n_{ij} et $\{p_{ijk}\}$. La distribution des variables aléatoires $\{X_{ijk}, i = 1,..., I; j = 1,..., J; k = 1,..., K\}$ est un produit de distributions multinomiales.

L'espérance mathématique de X_{ijk} est :

$$m_{ijk} = n_{ij}\, p_{ijk}$$

Les problèmes que l'on cherche à résoudre sont :

— test d'absence d'effet d'un des facteurs,
— test d'absence d'interaction entre les deux facteurs.

TABLEAU 1
Répartition des retours en chaleur après insémination

Facteurs		Période (*)							
F_1	F_2	1	2	3	4	5	6	7	Totaux
1	1	3	11	3	7	8	1	61	94
1	2	1	19	4	3	5	1	52	85
1	3	0	17	5	8	6	2	104	142
1	4	0	10	5	3	4	2	57	81
⋮									
2	1	0	3	0	0	0	2	4	9
2	2	1	0	0	0	0	0	0	1
2	3	1	2	0	0	0	0	11	14
⋮									
26	57	0	3	1	1	1	1	7	14
26	58	0	0	0	0	0	0	0	0
26	59	0	0	0	0	0	0	2	2
26	60	1	8	2	0	0	0	14	25

(*) Période 1 : inférieur à 18 jours
2 : entre 18 et 24 jours
3 : entre 25 et 35 jours
4 : entre 36 et 48 jours
5 : entre 49 et 89 jours
6 : entre 90 et 400 jours
7 : plus de 400 jours (non retour).

3. MODÈLE LOG-LINÉAIRE

Un modèle log-linéaire spécifie une relation linéaire entre le logarithme de la fréquence espérée, m_{ijk}, et des paramètres inconnues; par exemple, le modèle log-linéaire saturé s'écrit :

$$M_0 : \text{Log } m_{ijk} = u^0 + u_i^1 + u_j^2 + u_k^3 + u_{ij}^{12} + u_{ik}^{13} + u_{jk}^{23} + u_{ijk}^{123}$$

On imposera aux paramètres les mêmes contraintes que celles utilisées dans GLIM et GENSTAT, à savoir :

$$u_1^1 = u_1^2 = u_1^3 = 0;$$
$$u_{i1}^{12} = u_{1j}^{12} = u_{i1}^{13} = u_{1k}^{13} = u_{j1}^{23} = u_{1k}^{23} = 0;$$
$$u_{ij1}^{123} = u_{i1k}^{123} = u_{1jk}^{123} = 0.$$

On a une table de contingence en présence de deux facteurs, tout modèle log-linéaire considéré doit contenir les paramètres u^1, u^2, u^{12}. Cette précaution étant prise, on peut analyser cette table de contingence comme si les variables aléatoires $\{X_{ijk}\}$ suivaient indépendamment les lois de Poisson de paramètres $\{m_{ijk}\}$ (HABERMAN (1974), NELDER et WEDDERBURN (1972)).

3.1. Interprétation des paramètres

On a :

$$p_{ijk} = m_{ijk}/m_{ij+}, \qquad \text{avec} \qquad m_{ij+} = \sum_{k=1}^{K} m_{ijk}$$

Ici, on mesure l'effet du traitement (i, j) sur la distribution de la réponse R par les paramètres

$$\eta_{ijk} = \text{Log}\,(p_{ijk}/p_{ij1}) \qquad k = 2,..., K.$$

C'est la définition de PLACKETT (1981) avec p_{ij1} à la place de p_{ijk} pour conduire à un traitement plus simple compte tenu des contraintes que l'on a imposées aux paramètres.

On a :

$$\eta_{ijk} = u_k^3 + u_{ik}^{13} + u_{jk}^{23} + u_{ijk}^{123}$$

Les paramètres u^3, u^{13}, u^{23} et u^{123} peuvent donc s'interpréter de la même manière que les paramètres d'un modèle d'analyse de variance multivariable (K − 1 variables) :

- u^{13} représente l'effet principal du facteur F,
- u^{23} représente l'effet principal du facteur G,
- u^{123} représente l'interaction entre les deux facteurs.

On s'intéresse au test des hypothèses :

$H_{12} : u^{123} = 0$ $\qquad (u_{ijk}^{123} = 0$ pour tout (i, j, k))

$H_1 : u^{123} = 0$ et $u^{13} = 0$ $\qquad (u_{ijk}^{123} = u_{ik}^{13} = 0$ pour tout (i, j, k))

$H_2 : u^{123} = 0$ et $u^{23} = 0$ $\qquad (u_{ijk}^{123} = u_{jk}^{23} = 0$ pour tout (i, j, k))

On peut utiliser le test du rapport de vraisemblance.

3.2. Estimation des paramètres

La méthode la plus utilisée pour estimer les paramètres d'un modèle log-linéaire est la méthode du maximum de vraisemblance. Le modèle log-linéaire appartient à la classe des modèles linéaires généralisés définie par NELDER et WEDDERBURN (1972), McCULLAGH et NELDER (1983). On

peut, en principe, utiliser GLIM ou GENSTAT pour obtenir l'estimation des paramètres et la matrice des variances-covariances estimée des estimateurs. Mais, GLIM et GENSTAT sont basés sur un algorithme itératif qui implique l'inversion de la matrice d'information dont la dimension est égale au nombre de paramètres du modèle. Ici, tout modèle log-linéaire doit contenir les paramètres u^1, u^2 et u^{12} et en conséquence il n'est pas possible d'utiliser GLIM ou GENSTAT.

On utilise l'algorithme IPF (Iterative Proportional Fitting) décrit par exemple dans BISHOP *et al.* (1975) et HABERMAN (1972). Cet algorithme donne l'estimation du maximum de vraisemblance des fréquences espérées $\{m_{ijk}\}$. Lorsque la résolution des équations du maximum de vraisemblance est possible analytiquement, la convergence est atteinte après une itération. Pour une table à trois dimensions, seul le modèle de non interaction d'ordre 2 :

$$\text{Log } m_{ijk} = u^0 + u_i^1 + u_j^2 + u_k^3 + u_{ij}^{12} + u_{ik}^{13} + u_{jk}^{23}$$

nécessite des itérations.

L'inconvénient de cet algorithme est qu'il ne donne pas la matrice des variances-covariances estimée des estimateurs. LEE (1977) donne les variances asymptotiques des estimateurs pour les modèles dont la résolution est possible analytiquement.

L'algorithme IPF est valable à la fois pour les tables de contingence complètes et incomplètes.

Lorsque la table de contingence à analyser est incomplète, on doit en tenir compte dans le calcul du nombre de degrés de liberté. On trouve dans BISHOP *et al.* (1975) la règle de calcul du nombre de degrés de liberté d'un modèle log-linéaire d'analyse d'une table incomplète.

4. MODÈLE MULTIPLICATIF DE COX

On applique le modèle multiplicatif de COX, modèle à risques proportionnels (COX, 1972), à la variable continue que l'on a discrétisée.

Pour chaque traitement (i, j), on a inséminé n_{ij} vaches dont les délais de retour en chaleur sont considérés comme réalisations indépendantes d'une variable aléatoire T_{ij} de fonction de répartition F_{ij}, de densité f_{ij}.

La fonction de risque instantané h_{ij} associée à F_{ij} est définie par :

$$h_{ij}(t) = \lim_{\Delta t \to 0^+} \frac{\text{pr}\,(t \leqslant T_{ij} < t + \Delta t \mid T_{ij} \geqslant t)}{\Delta t}$$

$$f_{ij}(t) / [1 - F_{ij}(t)] \qquad \text{pour tout } t > 0$$

Dans notre exemple, $h_{ij}(t)$ est le risque ou taux instantané de retour à l'instant t d'une vache ayant subi le traitement (i, j).

On a la relation :

$$F_{ij}(t) = 1 - \exp\left[-\int_0^t h_{ij}(u)\,du\right]$$

Le modèle multiplicatif de COX s'écrit :

$$h_{ij}(t) = h_o(t) \exp(\mu_{ij})$$

où • $h_o(t)$ est une fonction (a priori inconnue) du temps;
• μ_{ij} est une combinaison linéaire des variables explicatives.

On trouve l'utilisation du modèle multiplicatif de COX pour les données groupées dans ARANDA-ORDAZ (1983), BARTLETT (1978), THOMPSON Jr. (1977).

Ici, on décompose μ_{ij} de façon à faire apparaître l'effet de chacun des facteurs et l'interaction entre les deux facteurs,

$$\mu_{ij} = \mu + \beta_i^1 + \beta_i^2 + \beta_{ij}^{12},$$

avec les mêmes contraintes sur les paramètres que celles adoptées dans GLIM et GENSTAT.

Ce modèle est assez flexible dans la mesure où la seule supposition faite sur la fonction $h_o(t)$ est celle de son intégrabilité.

4.1. Estimation et test d'hypothèses

La méthode d'estimation et de test utilisée est celle du maximum de vraisemblance.

On ne dispose pas des réalisations T_{ij}, mais de leurs fréquences x_{ijk} dans les intervalles $]t_0, t_1[, [t_1, t_2[, \ldots, [t_{K-1}, t_K[$, avec $t_0 = 0$ et $t_K = +\infty$.

Soit :

$$\pi_{ijk} = \text{pr}\,(t_{k-1} \leq T_{ij} < t_k \mid T_{ij} \geq t_{k-1})$$

$$n_{ijk} = n_{ij} - (x_{ij1} + \ldots + x_{ijk}) \; ; \; k = 1, \ldots, K-1.$$

Dans l'exemple, π_{ijk} est la probabilité conditionnelle qu'une vache, ayant subi le traitement (i, j), ait un délai de retour dans $[t_{k-1}, t_k[$ sachant que ce délai est supérieur à t_{k-1}.

On a

• $\pi_{ijk} = [F_{ij}(t_k) - F_{ij}(t_{k-1})] / [1 - F_{ij}(t_{k-1})]$

$$= 1 - \exp\left[-\exp(\mu_{ij}) \int_{t_{k-1}}^{t_k} h_o(u)\,du\right]$$

On pose

• $\mu_{ijk} = \text{Log}\,(-\text{Log}\,(1 - \pi_{ijk})) = \mu_{ij} + \gamma_k$

où $\gamma_k = \text{Log} \int_{t_{k-1}}^{t_k} h_o(u)\,du$

La vraisemblance pour les observations $\{x_{ijk}\}$ s'écrit :

$$\ell\ (x \mid \pi) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K-1} [\pi_{ijk}^{x_{ijk}}\ (1 - \pi_{ijk})^{n_{ijk} - x_{ijk}}]$$

On a ainsi un modèle linéaire généralisé standard dont la résolution numérique peut se faire au moyen du logiciel GLIM ou GENSTAT.

On considère les hypothèses :

H_{12} : $\beta_{ij}^{12} = 0$ pour tout (i, j)

H_1 : $\beta_{ij}^{12} = 0$ et $\beta_i^1 = 0$ pour tout (i, j)

H_2 : $\beta_{ij}^{12} = 0$ et $\beta_j^2 = 0$ pour tout (i, j)

Le test du rapport de vraisemblance de H_{12} contre non H_{12} nécessite l'estimation des paramètres de deux modèles,

$$M_0 : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 + \beta_{ij}^{12} + \beta_k^3$$

et

$$M_1 : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 \qquad + \beta_k^3$$

où

$\beta^0 = \mu + \gamma_1$, $\beta_k^3 = \gamma_k - \gamma_1$ pour satisfaire la contrainte $\beta_1^3 = 0$.

Il n'est pas matériellement possible d'obtenir l'estimation des paramètres du modèle M_0 car l'interaction comporte (60-1) (26-1) = 1475 paramètres.

Remarque

On a considéré les probabilités conditionnelles $\{\pi_{ijk}\}$ pour résoudre le problème d'estimation et de test d'hypothèses. Cependant, on peut exprimer les probabilités (non conditionnelles)

$$p_{ijk} = \text{pr}\ (t_{k-1} \leq T_{ij} < t_k)$$

en fonction des probabilités conditionnelles π_{ijk} :

$p_{ij1} = \pi_{ij1}$

$p_{ijk} = \pi_{ijk}\ (1 - \pi_{ij1})\ (1 - \pi_{ij2})\ \ldots\ (1 - \pi_{ijk-1})$; $k = 2, \ldots, K - 1$.

On peut en déduire l'estimation du maximum de vraisemblance de p_{ijk}.

4.2. Interaction multiplicative

On a vu dans le paragraphe précédent que le test du rapport de vraisemblance de l'hypothèse d'absence d'interaction entre les deux facteurs n'était pas matériellement réalisable, au moins avec GLIM ou GENSTAT.

On se propose d'étudier l'interaction entre les deux facteurs en supposant qu'elle est proportionnelle au produit des termes des effets principaux,

$$\beta_{ij}^{12} = \psi\ \beta_i^1\ \beta_j^2 \text{ pour tout (i, j).}$$

Dans le cadre du modèle linéaire, on trouve une étude de l'interaction entre deux facteurs dans DENIS (1983). La notion d'interaction multiplicative a été introduite par TUKEY (1949).

Sous la supposition d'interaction multiplicative, l'absence d'interaction entre les deux facteurs est traduite par l'hypothèse $H_0 : \psi = 0$.

Le test du rapport de vraisemblance de l'hypothèse nulle H_0 contre son contraire nécessite l'estimation des paramètres sous chacune de ces deux hypothèses. Il est asymptotiquement équivalent au test basé sur les scores qui n'exige que l'estimation des paramètres sous l'hypothèse H_0. Ce dernier test est donc préférable au test du rapport de vraisemblance; il est défini par exemple dans COX et HINKLEY (1974) et MORAN (1970).

Le test des scores de l'hypothèse H_0 contre son contraire est décrit en annexe.

On a donc résolu le problème de test de non interaction entre les deux facteurs en supposant que l'interaction est proportionnelle au produit des termes des effets principaux.

5. RÉSULTATS NUMÉRIQUES

Dans l'exemple considéré, 94 des traitements (i, j) ne sont pas expérimentés. On en tient compte en analysant la table incomplète obtenue après élimination de ces 94 traitements.

5.1. Modèle log-linéaire

Dans le cas d'une table incomplète, la règle de calcul du nombre de degrés de liberté d'un modèle log-linéaire est donnée dans BISHOP *et al.* (1975).

Soient :

- z_e le nombre de zéros dans la table espérée $\{m_{ijk}\}$;
- z_{12} le nombre de zéros dans la table marginale espérée $\{m_{ij+}\}$;
- z_{13} le nombre de zéros dans la table marginale espérée $\{m_{i+k}\}$;
- z_{23} le nombre de zéros dans la table marginale espérée $\{m_{+jk}\}$;

Le nombre de degrés de liberté du modèle :

$$M_1 : \text{Log } m_{ijk} = u^0 + u_i^1 + u_j^2 + u_k^3 + u_{ij}^{12} + u_{ik}^{13} + u_{jk}^{23}$$

est (I-1) (J-1) (K-1) $- z_e + (z_{12} + z_{13} + z_{23})$; celui du modèle

$$M_2 : \text{Log } m_{ijk} = u^0 + u_i^1 + u_j^2 + u_k^3 + u_{ij}^{12} + u_{ik}^{13}$$

est I (J-1) (K-1) $- z_e + (z_{12} + z_{13})$.

Le modèle

$$M_3 : \text{Log } m_{ijk} = u^0 + u_i^1 + u_j^2 + u_k^3 + u_{ij}^{12} + u_{jk}^{23}$$

a (I-1) J (K-1) $- z_e + (z_{12} + z_{23})$ degrés de liberté.

Dans l'exemple, $z_e = 658$, $z_{12} = 94$, $z_{13} = z_{23} = 0$.

Le tableau 2 donne la valeur de la déviance (moins deux fois le logarithme du maximum de vraisemblance).

TABLEAU 2

Modèle	D	Degré de liberté
M_1	8065.59	8286
M_2	8600.70	8640
M_3	8608.60	8436

$$D = 2 \sum_{ijk} \{x_{ijk} \text{ Log } (x_{ijk} / \hat{m}_{ijk}) \; ; \; m_{ijk} \neq 0\}$$

et le nombre de degrés de liberté pour chacun des 3 modèles M_1, M_2 et M_3.

La valeur de la statistique de test du rapport de vraisemblance de :

(i) M_1 contre M_0 est de 8065.59 pour 8286 degrés de liberté;
(ii) M_2 contre M_1 est de 535.11 pour 354 degrés de liberté;
(iii) M_3 contre M_1 est de 543.01 pour 150 degrés de liberté.

En utilisant l'approximation d'une loi de χ^2 à ν (grand) degrés de liberté par la loi normale de moyenne ν et de variance 2ν, le degré de signification du test est 0.956 pour (i) et pratiquement nul pour (ii) et (iii).

Donc on peut conclure :

(i) il y a un effet significatif du taureau et de l'inséminateur;
(ii) il n'y a pas interaction entre les deux facteurs.

L'effet du taureau semble être plus important que celui de l'inséminateur.

5.2. Modèle de Cox

On considère les modèles :

$$M_0 : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 + \beta_{ij}^{12} + \beta_k^3$$
$$M_0' : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 + \psi \beta_i^1 \beta_j^2 + \beta_k^3$$
$$M_1 : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 + \beta_k^3$$
$$M_2 : \mu_{ijk} = \beta^0 + \beta_i^1 + \beta_k^3$$
$$M_3 : \mu_{ijk} = \beta^0 + \beta_j^2 + \beta_k^3$$

Le test de M_1 contre M_0 n'est pas matériellement réalisable. On substitue à M_0 le modèle M_0'. La valeur de la statistique W_0 étant de 0.801, on accepte l'hypothèse $H_0 : \psi = 0$. Ainsi, en admettant que l'interaction entre taureau et inséminateur est proportionnelle au produit des termes des effets principaux, on ne rejette pas l'hypothèse d'absence d'interaction.

La déviance associée au modèle M_1 est 8616 pour 8637 degrés de liberté. La qualité d'ajustement du modèle M_1, mesurée par la probabilité qu'une variable aléatoire de loi de χ^2 à 8637 degrés de liberté dépasse la valeur 8616, est 0.563. Ce modèle M_1 est de bonne qualité.

La valeur de la statistique du rapport de vraisemblance de

(i) M_2 contre M_1 est de 176 pour 59 degrés de liberté;

(ii) M_3 contre M_1 est de 356 pour 25 degrés de liberté.

Les degrés de signification sont pratiquement nuls.

On aboutit donc aux mêmes conclusions que dans le modèle log-linéaire.

6. DISCUSSION

Deux modèles, modèle log-linéaire et modèle de COX, sont considérés pour étudier l'influence de deux facteurs sur une variable réponse obtenue par discrétisation d'une variable continue. Le modèle log-linéaire nécessite beaucoup plus de paramètres que le modèle de COX pour décrire l'influence des facteurs sur la variable réponse. Entre deux modèles de bonne qualité d'ajustement, selon ALTHAM (1984), on préfère celui qui a moins de paramètres car il permet d'estimer toute fonction des probabilités des cellules avec plus de précision. D'autre part, le modèle log-linéaire ne tient pas compte du fait que la variable réponse est obtenue par discrétisation d'une variable continue.

Cependant, lorsque les facteurs ont chacun un grand nombre de niveaux, le modèle de COX ne permet pas d'étudier l'interaction sans supposition supplémentaire. On a pu étudier l'interaction entre deux facteurs en supposant qu'elle est proportionnelle au produit des termes des effets principaux.

Le modèle log-linéaire permet d'étudier l'interaction entre les deux facteurs au moyen de l'algorithme IPF. Mais on ne peut pas avoir la précision de l'estimation des paramètres sous l'hypothèse (acceptée) de non interaction entre les deux facteurs. Alors que les écarts-types estimés des estimateurs des paramètres sont obtenus dans le modèle de COX sans interaction.

Le modèle log-linéaire et le modèle de COX sont deux modèles non emboîtés. Théoriquement, il existe des techniques de tests permettant de choisir entre deux modèles non emboîtés (GOURIEROUX *et al.*, 1983). Mais leur mise en œuvre se heurte au grand nombre de paramètres qui existent dans le modèle log linéaire et le modèle de COX.

Ainsi, il est difficile de choisir entre les deux modèles. Dans le traitement des données, les deux modèles conduisent aux mêmes conclusions : il y a effet des deux facteurs et absence d'interaction entre les deux facteurs.

On peut s'orienter vers un compromis. On utilise d'abord le modèle log-linéaire pour tester l'absence d'interaction entre les deux facteurs. On applique ensuite, en cas de non interaction, le modèle de COX pour tester les effets principaux et estimer les probabilités des cellules.

RÉFÉRENCES BIBLIOGRAPHIQUES

P.M.E. ALTHAM (1984). — Improving the Precision of Estimation by Fitting a Model. *Journal of the Royal Statistical Society B46,* 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.

F.J. ARANDA-ORDAZ (1983). An extension of the Proportional-Hazards Model for Grouped Data. *Biometrics 39,* 109-117.

R.J. BAKER and J.A. NELDER (1978). — *The GLIM system,* Release 3. Oxford : Numerical Algorithms Group.

N.R. BARTLETT (1978). — A survival Model for a Wood Preservative Trial. *Biometrics 34,* 673-679.

Y.M.M. BISHOP, S.E. FIENBERG and P.W. HOLLAND (1975). — *Discrete Multivariate Analysis : Theory and Practice.* Cambridge, Mass. : MIT Press.

D.R. COX (1972). — Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society, B34,* 45-58.

D.R. COX and D.V. HINKLEY (1974). — *Theoretical Statistics.* London : Chapman and Hall.

J.B. DENIS (1983). — Interaction entre deux facteurs. *Thèse de Docteur Ingénieur.* Paris : Institut National Agronomique.

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

S.J. HABERMAN (1972). — Loglinear fit for contingency tables (Algorithm AS 51). *Applied Statistics 21;* 218-225.

S.J. HABERMAN (1974). — *The Analysis of Frequency Data.* Chicago, Univ. of Chicago Press.

S.K. LEE (1977). — On the asymptotic variances of û terms in loglinear models of multidimensional contingency tables. *Journal of the American Statistical Association 72,* 412-419.

P. McCULLAGH and J.A. NELDER (1983). — *Generalized Linear Models.* London : Chapman and Hall.

P. McCULLAGH (1980). — Regression Models of Ordinal Data. *Journal of the Royal Statistical Society B42,* 109-142.

P.A.P. MORAN (1970). — On asymptotically optimal tests of Composite statistical hypotheses. *Biometrika 57,* 47-55.

J.A. NELDER and R.W.M. WEDDERBURN (1972). — Generalized linear models. *Journal of the Royal Statistical Society A 135,* 370-384.

R.L. PLACKETT (1981). — *The Analysis of Categorical Data.* London : Griffin.

J.R. W.A. THOMPSON (1977). — On the treatment of grouped observations in life-tables. *Biometrics 33,* 463-470.

J.W. TUKEY (1949). — One degree of freedom for non-additivity. *Biometrics, 5,* 232-242.

C.J.G. UPTON (1978). — *The Analysis of Cross-Tabulated Data.* New York : Wiley & Sons.

ANNEXE

Test des scores

Soient :

$\mu_{ijk} = \beta^0 + \beta_i^1 + \beta_j^2 + \psi\, \beta_i^1\, \beta_j^2 + \beta_k^3$

$\lambda = (\beta^0, \beta_2^1, \ldots, \beta_I^1, \beta_2^2, \ldots, \beta_J^2, \beta_2^3, \ldots, \beta_{K-1}^3)$

$\theta = (\psi, \lambda)$

$$L(\psi, \lambda) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K-1} [x_{ijk} \text{ Log } \pi_{ijk} + (n_{ijk} - x_{ijk}) \text{ Log } (1 - \pi_{ijk})]$$

le logarithme de la vraisemblance

$\hat{\lambda}_0$ l'estimation du maximum de vraisemblance de λ sous l'hypothèse H_0

$\mu_{ijk}^0 = \mu_{ijk}(\hat{\lambda}_0)$

$\pi_{ijk}^0 = \pi_{ijk}(\hat{\lambda}_0)$

$$U_0 = \left[\frac{\partial L}{\partial \psi} \mid \psi = 0\right] \lambda = \hat{\lambda}_0$$

$$= \sum_{ijk} \frac{(x_{ijk} - n_{ijk}\, \pi_{ijk}^0)}{\pi_{ijk}^0} \exp(\mu_{ijk}^0)\, \hat{\beta}_{oi}^1\, \hat{\beta}_{oj}^2$$

$$I_0 = \left[E\left(-\frac{\partial^2 L}{\partial \psi^2}\right) \mid \psi = 0\right] \lambda = \hat{\lambda}_0$$

$$= \sum_{ijk} \frac{n_{ijk}}{\pi_{ijk}^0} \exp(2\mu_{ijk}^0)(1 - \pi_{ijk}^0)(\hat{\beta}_{oi}^1\, \hat{\beta}_{oj}^2)^2$$

$W_0 = U_0^2 / I_0$

$\chi_{\nu,\alpha}$ la valeur réelle qui est dépassée avec une probabilité α par une variable aléatoire de loi de χ^2 à ν degrés de liberté.

Le test des scores, de niveau asymptotique α, de H_0 contre non H_0 est défini par :

$$\text{Rejet de } H_0 \Longleftrightarrow W_0 > \chi_{1,\alpha}$$