JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JULIETTE BLOCH
MICHEL CHAVANCE
JOSEPH LELLOUCH
NADIA TAHRI

Modélisation marginale de délais corrélés

Journal de la société française de statistique, tome 143, n° 1-2 (2002), p. 187-194

<http://www.numdam.org/item?id=JSFS_2002__143_1-2_187_0>

MODÉLISATION MARGINALE DE DÉLAIS CORRÉLÉS

Juliette BLOCH*, Michel CHAVANCE*, Joseph LELLOUCH*
Nadia TAHRI*

RÉSUMÉ

Nous montrons comment un modèle marginal log-linéaire, équivalent du modèle de Cox, permet d'analyser des observations censurées corrélées. Les corrélations sont prises en compte grâce à un estimateur robuste de la variance, comme proposé par Liang et Zeger, 1986, dans le cadre des équations d'estimation généralisées.

ABSTRACT

We show how correlated censored observations can be analyzed by means of a marginal log-linear model, equivalent to the proportional hazards model. Correlations are taken into account thanks to the robust sandwich estimator of the variance, as proposed by Liang and Zeger, 1986, in the framework of the generalized estimating equations.

1. Introduction

Le modèle des risques proportionnels (Cox et Oakes, 1984), fréquemment utilisé pour la modélisation de délais censurés, suppose l'indépendance des observations. Il arrive cependant que cette condition ne soit pas remplie. C'est le cas, bien que l'on s'intéresse au délai de survenue d'un événément unique (décès, incidence d'une maladie chronique,...) lorsque l'échantillon est structuré en groupes (sujets regroupés par hôpitaux ou par familles, yeux regroupés par sujets,...) et que les observations se ressemblent plus à l'intérieur d'un groupe qu'elles ne ressemblent à celles des autres groupes. Dans d'autres situations, les délais étudiés peuvent être corrélés parce qu'un même événement peut se produire plusieurs fois chez chaque sujet (grossesse, récurrence tumorale, accident vasculaire,...) ou parce que l'on s'intéresse à la survenue de plusieurs événements distincts (succession des stades cliniques d'une maladie, infections dues à différents pathogènes, complications,...). Dans ce dernier cas, il est évidemment nécessaire d'envisager autant de distributions distinctes que de types d'événements, mais dans toutes les situations évoquées, des inférences statistiques correctes ne peuvent être effectuées que si les

* INSERM U472, 16 Av. P. Vaillant-Couturier 94807 Villejuif; chavance@vjf.inserm.fr

corrélations entre délais sont prises en compte. Si l'on s'intéresse à l'effet de variables définies au niveau du groupe et que les observations sont si parfaitement corrélées que tous les délais d'un groupe sont identiques, on ne gagne aucune information en augmentant la taille du groupe, et considérer les observations comme indépendantes conduirait à sous-estimer la variance des estimateurs. À l'inverse, ignorer les corrélations positives quand on s'intéresse à l'effet de variables qui fluctuent à l'intérieur des groupes peut conduire à surestimer la variance des estimateurs, comme lorsque l'appariement est ignoré dans une comparaison de moyennes appariées.

Deux généralisations du modèle des risques proportionnels ont été proposées pour prendre en compte les corrélations entre des délais censurés. Dans les modèles de fragilité, la corrélation est modélisée à l'aide d'une ordonnée à l'origine aléatoire, ou fragilité, partagée par toutes les observations d'un même groupe (Rondeau et Commenges, 2002). Ces modèles sont particulièrement adaptés aux problèmes de prédiction individuelle puisqu'ils permettent de prédire les fragilités associées aux différents groupes et donc les risques individuels. Un exemple classique concerne la sélection des meilleurs reproducteurs en génétique animale. À l'inverse, les modèles marginaux permettent de mesurer les effets de variables explicatives au niveau de la population et non des individus puisqu'ils s'intéressent à la distribution marginale des délais, après intégration sur les groupes (Wei *et al.*, 1989). Ils sont donc bien adaptés aux problèmes d'épidémiologie.

Cet article montre comment on peut utiliser un modèle log-linéaire marginal pour estimer les paramètres d'un modèle des risques proportionnels marginal, comme cela a été proposé dans le cas indépendant par Whitehead (1980).

2. Analogie entre modèle de Cox et modèle log-linéaire

2.1. Observations indépendantes

Soit T_i la date de l'événement pour le i^{ème} sujet et C_i sa date de censure. Nous supposerons que le temps est continu et qu'il n'y a donc pas d'ex aequo. On observe le vecteur (Z_i, Δ_i) où $Z_i = min(T_i, C_i)$, $\Delta_i = 1$ si T_i est observé $(Z_i = T_i)$ et $\Delta_i = 0$ s'il y a censure $(Z_i = C_i)$. Nous appellerons $f_i(t)$ la densité de la distribution de T_i, $S_i(t) = P[T_i \geqslant t]$ la fonction de survie et $\lambda_i(t) = \frac{f_i(t)}{S_i(t)}$ la fonction de risque de survenue de l'événement chez le sujet i. Cette fonction de risque est définie dans le modèle des risques proportionnels par le produit d'un risque de base $\lambda_0(t)$ et de l'effet d'un vecteur de covariables X_i : $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta)$. Pour des observations indépendantes, la log-vraisemblance partielle de β est

$$L(\beta) = \sum_i \Delta_i \left[X_i\beta - \log \sum_{k \in \Re(t_i)} \exp(X_k\beta) \right] \qquad (1)$$

où $\Re(t)$ représente l'ensemble des sujets à risque en t, c'est-à-dire ceux pour lesquels ni événement ni censure n'ont encore été observés.

On peut aussi choisir une modélisation non paramétrique du risque de base, supposé constant mais quelconque entre deux événements non censurés consécutifs : $t \in]t_{k-1}, t_k] \Rightarrow \lambda_0(t) = \lambda_k$ (Breslow, 1974). La contribution à la log-vraisemblance d'un sujet i est $S_i(t)$ s'il est censuré en t, et $\lambda_i \exp(X_i\beta) S_i(t)$ si l'événement est observé en t pour i. Un sujet censuré dans l'intervalle entre deux événements consécutifs, t_k et t_{k+1} est supposé censuré en t_k. Cela conduit à une log-vraisemblance

$$L(\alpha, \beta) = \sum_i \Delta_i \left[X_i\beta + \alpha_i - \sum_{k \in \Re(t_i)} \exp(X_k\beta + \alpha_i) \right] \tag{2}$$

où α_i represente le logarithme de la contribution de l'intervalle $[t_{i-1}, t_i]$ au risque de base cumulé :

$$\alpha_i = \log[\lambda_i (t_i - t_{i-1})]$$

Cette expression, malgré ses différences avec (1), est la clé de l'équivalence entre le modèle log-linéaire et celui des risques proportionnels.

1) La log-vraisemblance de β est la même en (1) et (2). L'équation du maximum de vraisemblance de α_i est

$$1 - \sum_{k \in \Re(t_i)} \exp(X_k\beta + \alpha_i) = 0$$

en remplaçant α_i par sa racine, $\alpha_i = -\log \sum_{k \in \Re(t_i)} \exp(X_k\beta)$ dans (2), on retrouve (1).

2) La log-vraisemblance (2) est aussi celle d'un modèle log-linéaire ajusté à des pseudo-réponses poissoniennes Y_{ik} définies, à chaque occurence d'événement t_i pour l'ensemble des sujets à risque en t_i par $Y_{ii} = 1$ pour le sujet qui « décède » et $Y_{ik} = 0$ pour les autres. Ce modèle stipule

$$\begin{aligned} \log(E[Y_{ik}]) &= \log \mu_{ik} \\ &= X_k\beta + \alpha_i \end{aligned} \tag{3}$$

Pour estimer β à l'aide d'un modèle log-linéaire, il faut donc générer de pseudo-observations selon l'algorithme suivant (voir tableau 1) :

1) Ordonner les observations d'origine en fonction des dates d'événement ou de censure t_i

2) À chaque date où un événement (non une censure) est observé, construire un ensemble de pseudo-observations comprenant pour chacun des sujets k de l'ensemble à risque $\Re_i$, les variables suivantes :

   - un vecteur des variables explicatives $X^*_{ik} = X_k$

- une variable (e.g. $I = i$) identifiant l'ensemble à risque $\Re(t_i)$ auquel correspond la pseudo-observation ;
- la pseudo-reponse $Y_{ii} = 1$ ou $Y_{ik} = 0$ $(k \neq i)$

Les coefficients α_i associés à la variable qualitative I permettent d'obtenir une estimation non paramétrique du risque de base.

TABLEAU 1. – Exemple de génération de pseudo-observations pour un échantillon de 4 sujets.

sujet i	censure Δ_i	délai t_i	variable X_i	sujet i	pseudo response Y_{ik}	I k	variable X^*_{ik}
1	1	3	X_1	1	1	1	X_1
				2	0	1	X_2
				3	0	1	X_3
				4	0	1	X_4
2	0	5	X_2				
3	1	6	X_3	3	1	2	X_3
				4	0	2	X_4
4	1	9	X_4	4	1	3	X_4

2.2. Observations corrélées

2.2.1. Un seul risque de base

En présence d'observations corrélées, il faut utiliser un double indiçage afin d'identifier les événements j dans les groupes i. Le modèle marginal le plus simple stipule la proportionalité entre les fonctions de risque marginales et un unique risque de base (Lee *et al.*, 1992)

$$\lambda_{ij}(t) = \lambda_0(t) \exp(X_{ij}\beta) \tag{4}$$

Ce modèle peut s'appliquer à des événements ne se produisant qu'une fois dans une population structurée en groupes ou à des événements susceptibles de se répéter chez un même sujet. Il implique alors que le risque ne dépend pas du rang de l'événement. Le vecteur des paramètres β peut être estimé en faisant « comme si » les observations étaient indépendantes. L'estimateur qui maximise la log-vraisemblance partielle sous cette hypothèse de travail

$$L(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \Delta_{ij} \left[X_{ij}\beta - \log \sum_{(k,l) \in \Re(t_{ij})} \exp(X_{kl}\beta) \right] \tag{5}$$

s'obtient en annulant la dérivée en β de cette expression

$$U(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} U_{ij}(\beta) \qquad (6)$$
$$= 0$$

À cause de l'hypothèse de travail d'indépendance, cet estimateur est le même que celui du modèle de Cox vu au paragraphe 2.1, mais la variance de ces 2 estimateurs n'est pas la même. Un développement de Taylor des scores $U_{ij}(\beta) = X_{ij} - \dfrac{\sum_{(k,l)\in\Re(t_{ij})} X_{kl} \exp(X_{kl}\beta)}{\sum_{(k,l)\in\Re(t_{ij})} \exp(X_{kl}\beta)}$ autour de $\widehat{\beta}$ montre que

$$\widehat{\beta} - \beta \simeq -\left[\frac{\partial U}{\partial \beta}\left(\widehat{\beta}\right)\right]^{-1} U(\beta)$$

Cet estimateur est convergent, mais sa variance n'est pas l'inverse $I^{-1}(\beta)$ de la matrice d'information, elle est donnée par une formule sandwich

$$\Sigma = I^{-1}(\beta)\,\Lambda I^{-1}(\beta) \qquad (7)$$

où la variance des scores Λ peut être estimée par la variance empirique des scores observés

$$\widehat{\Lambda} = \frac{1}{n-1} \sum_{i=1}^{n} U_i\left(\widehat{\beta}\right) U_i\left(\widehat{\beta}\right)^t \qquad (8)$$

avec $U_i\left(\widehat{\beta}\right)^t = \left(U_{i1}(\widehat{\beta}) \ldots U_{in_i}(\widehat{\beta})\right)$.

Considérons le modèle log-linéaire défini sur de pseudo-réponses supposées indépendantes Y_{ijkl} telles que $Y_{ijij} = 1$ si l'événement se produit en t_{ij} pour le sujet j du groupe i et $Y_{ijkl} = 0$ si l'événement ne se produit pas en t_{ij} pour le sujet l du groupe k bien qu'il soit à risque en t_{ij}. Ce modèle est encore équivalent au modèle 4 : sous l'hypothèse de travail d'indépendance, rien n'est changé par rapport au paragraphe 2.1, en revanche si les observations sont corrélées la solution des équations d'estimation de β

$$\sum_{i,j} U_{ij}(\beta) = \sum_{i,j} \Delta_{ij} \left[\sum_{(k,l)\in\Re(t_{ij})} Y_{ijkl} X_{ij}\beta - \log \sum_{(k,l)\in\Re(t_{ij})} \exp(X_{ijkl}\beta) \right] \qquad (9)$$
$$= 0$$

maximise une quasi-vraisemblance et non une vraisemblance. En raison des corrélations, il faut utiliser un estimateur sandwich pour obtenir la variance de l'estimateur. Le même développement limité de $U_{ij}(\beta)$ conduit à la même approximation de $\widehat{\beta} - \beta$ et à la même estimation de la variance que l'on raisonne sur la log-vraisemblance partielle du modèle de Cox marginal ou sur la log-vraisemblance du modèle log-linéaire.

Les pseudo-observations auxquelles est ajusté le modèle log-linéaire sont générées comme en 2.1, en incluant une indicatrice de groupe afin d'identifier les pseudo-réponses corrélées. On peut ensuite utiliser tout logiciel permettant d'estimer les paramètres d'un modèle log-linéaire et d'obtenir la variance sandwich de l'estimateur, comme la procédure GENMOD de SAS version 8.

2.2.2. Plusieurs risques de base

Quand on s'intéresse à plusieurs types d'événements susceptibles de se produire chez un même sujet, il est peu raisonnable de leur associer un même risque de base. Si l'on suppose la proportionalité des risques associés aux différents événements, il suffit d'introduire dans le vecteur de variables explicatives des indicatrices des différents types d'événement. L'écriture du modèle reste (4) et chaque événement observé contribue à l'estimation du risque de base commun. On peut préférer un modèle plus général et supposer un risque de base quelconque pour chacun des types d'événements (Wei *et al.*, 1989). Le modèle devient

$$\lambda_{ij}(t) = \lambda_j(t) \exp(X_{ij}\beta) \tag{10}$$

chaque événement observé ne contribue qu'à l'estimation de son risque de base spécifique et pour la génération des pseudo-observations, il faut considérer des ensembles à risque séparés pour chaque type d'événement.

Avec (4) comme avec (10), il est possible d'imposer ou non que le risque relatif associé à une variable soit le même pour chaque type d'événement.

3. Application : risque de retrait de chambres implantables chez le mucoviscidosique

Les chambres implantables (CI) sont utilisées pour l'administration chronique d'antibiotiques chez des patients mucoviscidosiques. L'événement étudié est ici le retrait d'une chambre en raison de complications. Les observations pouvaient être censurées en raison de la fin de l'étude, d'un retrait de la CI pour une autre cause (greffe pulmonaire, amélioration clinique,...) ou du décès du patient. Au total 96 retraits ont été observés dont 90 pour complications. L'enquête concerne le devenir de 265 chambres implantées chez 200 patients et les facteurs de risque sont des caractéristiques de la chambre ou du patient. Un quart des patients (N=49) a eu au moins deux CI, avec un maximum de 4 CI pour 4 patients. La durée de vie médiane des CI était de 60 mois.

Le tableau 2 donne les résultats de deux modèles des risque proportionnels appliqués à ces données, le modèle de Cox, pour observations indépendantes et un modèle marginal utilisant l'hypothèse de travail d'indépendance, un seul risque de base et l'estimateur sandwich de la variance. Les estimations ponctuelles des risques relatifs sont identiques car effectuées dans les deux cas en supposant l'indépendance. En revanche, les intervalles de confiance sont différents et pour deux variables (utilisation de la contre-pression et colonisation par Pseudomonas) les conclusions du test de Wald au seuil de 5% sont modifiées. En présence de données dont la structure suggère,

TABLEAU 2. – Estimation des risques relatifs (RR), de leurs intervalles de confiance à 95% (IC_{95}), et test de leur nullité (p) dans le modèle de Cox pour données indépendantes et dans le modèle log-linéaire utilisant l'estimateur sandwich de la variance.

	Modèle de Cox pour données indépendantes		Modèle log-linéaire et variance sandwich	
	RR (IC_{95})	p	RR (IC_{95})	p
utilisation d'une contre-pression	1,63 (0,96-2,76)	0,07	1,63 (1,03-2,59)	0,04
réalisations de prélèvements sanguins				
parfois/jamais	1,14 (0,70-1,86)	0,60	1,14 (0,69-1,90)	0,59
souvent/jamais	2,20 (1,20-3,95)	0,01	2,20 (1,32-3,67)	0,002
cathéter en polyuréthane	2,04 (1,28-3,25)	0,003	2,04 (1,31-3,17)	0,002
Pseudomonas Aeruginosa	2,92 (0,72-11,95)	0,13	2,92 (1,0-8,56)	0,05

comme ici, la possibilité de corrélations, on préfère utiliser l'estimateur sandwich, qui converge vers la matrice de covariance de l'estimateur même lorsque la structure de covariance, ici l'indépendance, utilisée pour estimer les paramètres du modèle n'est pas vérifiée. Cette robustesse a évidemment un prix, une moins bonne efficacité que l'estimateur « naïf » quand cette hypothèse est correcte.

4. Discussion

Le modèle des risques proportionnels pour données censurées corrélées peut être mis en oeuvre à l'aide d'un programme fortran (Lin, 1993) ou à l'aide d'un programme pour données indépendantes incluant une option pour le calcul de la variance sandwich (option covsandwich de la procédure PHREG de SAS). L'approche log-linéaire sous hypothèse d'indépendance conduit à des résultats strictement identiques en ce qui concerne l'estimation des paramètres et leur variance. Elle produit en outre une estimation du risque de base. Un avantage théorique est de permettre également une estimation des paramètres en utilisant une autre structure de corrélation que l'indépendance. Un inconvénient est la taille du fichier des pseudo-observations beaucoup plus importante que celle des données initiales, de l'ordre de $d(n-\frac{d-1}{2})$ où n est le

nombre de sujets à risque à l'entrée dans l'étude et d le nombre d'événements observés (le nombre exact de pseudo-observations est plus ou moins inférieur à cette valeur selon le nombre et la répartition des censures). Dans notre exemple pour 265 chambres et 90 retraits, 13195 pseudo-observations ont été générées.

Nous n'avons traité ici que le cas où le temps est une variable continue. Lorsque la mesure du temps est discrète et que plusieurs événements peuvent être observés simultanément, l'écriture de la log-vraisemblance se complique, mais des approximations ont été proposées. Celle de Peto (Peto, 1972) se retrouve dans l'approche log-linéaire pour données corrélées en générant les pseudo-réponses $Y_{ijkl} = 1$ pour tous les sujets (kl) pour lesquels un événement est observé au même instant t_{ij}. D'autres approches sont envisageables, comme le recours à une stratégie d'imputation multiple (Bloch *et al.*, 1999).

RÉFÉRENCES

BLOCH J., CHAVANCE M., LELLOUCH J., TAHRI N., MUNCK A., MALBEZIN S. (1999), Utilisation des GEE pour la modélisation de données censurées corrélées : application à l'étude des facteurs de risque de retrait des chambres implantables chez le mucoviscidosique, *Revue d'Epidémiologie et de Santé Publique*, 47, 585-591.

BRESLOW N. (1974), Covariance analysis of censored survival data, *Biometrics*, 30, 89-99.

COX D.R., OAKES D. (1984), Analysis of survival data, *Chapman & Hall London.*

LIANG K.Y., ZEGER S. (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, 73, 13-22.

LEE E.W., WEI L.J., AMATO D.A. (1992), Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in J.P. Klein and P.K. Goel (eds), *Survival analysis : State of the art*, Kluwer Academic Publisher, Dordrecht, 237-247.

LIN D.Y. (1993), MULCOX2, a general computer program for the Cox regression analysis of multivariate failure time data, *Comput, Methods Programs Biomed*, 40, 279-293.

PETO R. (1972), Contribution to the discussion of paper by D.R. Cox, *Journal of the Royal Statistical Society Series B*, 34, 205-207.

RONDEAU V., COMMENGES D. (2002), Modélisation de la fragilité en survie, *Journal de la Société Française de Statistique*, 143, 1-2, 103-119.

WEI L.J., LIN D.Y., WEISSFELD L. (1989), Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *J.A.S.A.*, 84, 1065-1073.

WHITEHEAD J. (1980), Fitting Cox's regression model to survival data using GLIM, *Applied Statistics*, 29, 268-275.