Analyse des données de survie par un modèle de fragilité paramétrique

Yohann Foucher, Magali Giral, Pierre Dellamonica & Jean-Pierre Daurès

Institut Universitaire de Recherche Clinique, Laboratoire de Biostatistique - 641, avenue du Doyen Gaston Giraud - 34093 Montpellier, France. e-mail: Yohann.Foucher@iurc.montp.inserm.fr

Résumé : Ce papier définit un modèle de fragilité totalement paramétrique, adapté aux problématiques médicales. Nous introduisons une fonction de risque de base en forme de U ou U inversé. Ce modèle est appliqué à deux problématiques cliniques. Nous montrons ainsi l'intérêt de cette approche paramétrique. Elle est assez large pour pouvoir correspondre à la plupart des problématiques cliniques et elle estime des coefficients de régression équivalents à l'approche semi-paramétrique. L'utilisation quasi-systématique de cette dernière peut ainsi être discutée. ***Mots-clés :*** *modèle de fragilité, Gamma, Weibull généralisé, modèle paramétrique*

Summary : This paper deals with a parametrical frailty model, designed for medical analysis. We introduce a generalized Weibull hazard function, with U or U inverse shape. This model is applied to two clinical studies. We demonstrate interest of this parametrical approach which correspond to the large majority of clinical issues. Therefore, the large use of semi-parametrical frailty model can be discussed. ***Key-words :*** *frailty model, Gamma, Weibull generalized, parametrical model*

1 Introduction

Aalen (1988) est un premier a avoir fourni une base théorique et pratique aux modèles de fragilité. Les termes de fragilité pouvant être vus comme des données manquantes, certains auteurs, comme Hougaard (1986) ou Klein (1992), ont appliqué l'algorithme EM pour estimer les paramètres du modèle. Therneau (2003) reprend un certain nombre de ces développements pour mettre en place l'algorithme d'estimation de référence, implémenté sous SAS ou S-plus. Il étend le modèle de Cox semi-paramétrique grâce à l'ajout d'un terme de fragilité.

Nous proposons une alternative à cette modélisation en choisissant une fonction de risque de base paramétrique et une estimation par maximum de vraisemblance. Le méthode possède ainsi plusieurs originalités. Premièrement, elle offre une paramétrisation de la fonction de risque en forme de U ou U inverse, assez large pour correspondre à la plupart des problématiques cliniques. Des estimations et des prédictions, en terme de survie par exemple, sont alors possibles. Deuxièmement, la maximisation de la vraisemblance elle-même offre un cadre théorique plus confortable que l'algorithme EM pour l'estimation et l'inférence statistique.

2 Présentation du modèle de fragilité paramétrique

Considérons un échantillon constitué de N groupes indicés par i, $i = 1, ..., N$. Chacun de ces groupes est constitué de n_i observations repérées par l'indice j, $j = 1, ..., n_i$. Les observations issues d'un même groupe sont considérées dépendantes entre elles. Cette dépendance est contenue dans les termes de fragilité, ω_i, propres à chaque groupe i. Pour l'individu j du groupe i, posons T_{ij} le temps de survie égal au minimum entre le temps de censure et d'apparition de l'événement. Posons I_{ij} l'indicatrice de la censure, $I_{ij} = 1$ si l'événement est observé, 0 sinon. De même, posons $Z_{ij} = (Z_{ij1}, ..., Z_{ijp})$ le vecteur des p covariables, considérées comme indépendantes du temps. Conditionnellement au terme de fragilité ω_i et aux covariables observées Z_{ij}, la fonction de risque de l'observation j du groupe i est de la forme :

$$\lambda(t_{ij}|Z_{ij}, \omega_i) = \omega_i \lambda_0(t_{ij}) exp(\beta z_{ij}) \tag{1}$$

où $\lambda_0(t_{ij})$ est la fonction de risque de base et $\beta = (\beta_1, ..., \beta_p)'$ le vecteur des coefficients de régression associés à Z_{ij}. Comme définie par Aalen [**?**], la vraisemblance marginale s'écrit comme l'espérance sur ω_i du produit des contributions de chaque observation :

$$\mathcal{L} = \prod_{i=1}^{N} E_{\omega_i}\Big\{ exp(-\omega_i \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) exp(\beta z_{ij})) \prod_{j=1}^{n_i} \Big(\omega_i \lambda_0(t_{ij}) exp(\beta z_{ij})\Big)^{I_{ij}} \Big\} \tag{2}$$

où $\Lambda_0(t_{ij}) = \int_0^{t_{ij}} \lambda_0(x)dx$ est la fonction de risque de base cumulée. La formule (2) peut être réécrite en utilisant la transformée de Laplace de ω_i, définie par $L(a) = E_{\omega_i}\{exp(-a\omega_i)\}$. La r$^{\text{ième}}$ dérivée est égale à $L^{(r)}(a) = (-1)^r E_{\omega_i}\{\omega_i^r exp(-a\omega_i)\}$. En reprenant (2) et en passant au logarithme , on obtient :

$$ln(\mathcal{L}) = \sum_{i=1}^{N} \Big\{ \sum_{j=1}^{n_i} I_{ij} \Big(ln(\lambda_0(t_{ij})) + \beta z_{ij} \Big) + ln\Big((-1)^{d_i} L^{(d_i)}\Big(\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) exp(\beta z_{ij}) \Big) \Big) \Big\} \tag{3}$$

où $d_i = \sum_{j=1}^{n_i} I_{ij}$ est le nombre d'événements dans le groupe i.

Nous supposerons les ω_i indépendants et identiquement distribués selon une loi Gamma, $G(\delta^{-1}, \delta^{-1})$, dont la densité est donnée par :

$$g(\omega_i) = \frac{\omega_i^{(\delta^{-1}-1)} exp(-\omega_i \delta^{-1})}{\Gamma(\delta^{-1}) \delta^{\delta^{-1}}} \quad , \forall \delta \geq 0 \quad , \quad i = 1, ..., N \tag{4}$$

Cette distribution permet d'obtenir une espérance égale à 1 et une variance égale à δ. Une grande valeur du paramètre δ reflète ainsi une hétérogénéité importante entre les groupes constituant l'échantillon. Sa transformée de Laplace est égale à $L(a) = (1 + \delta a)^{-\delta^{-1}}$.

De plus, pour la fonction de risque de base, nous avons choisi une forme en U ou U inverse. Elle généralise ainsi la loi de Weibull :

$$\lambda(t) = (\theta+1)\left(1+\frac{\sigma t^{\nu+1}}{\nu+1}\right)^{\theta}\sigma t^{\nu} \quad , \forall\sigma>0 \quad , \quad \nu>-1 \quad , \quad \theta>-1 \tag{5}$$

Si $\theta = 0$, alors on retrouve une fonction de risque de type Weibull basée sur seulement deux paramètres. A partir de (4), (5) et (3), on obtient la logvraisemblance suivante :

$$ln(\mathcal{L}) = \sum_{i=1}^{N}\Bigg\{\sum_{j=1}^{n_i} I_{ij}\Big(ln(\theta+1)+\theta ln\Big(1+\frac{\sigma t_{ij}^{\nu+1}}{\nu+1}\Big)+ln(\sigma)+\nu ln(t_{ij})+\beta z_{ij}\Big)+$$

$$\sum_{j=1}^{d_i}\Big(ln(1+\delta(j-1))\Big)-(\delta^{-1}+d_i)ln\Big(1+\delta\Big[\sum_{j=1}^{n_i}\Big(\Big(1+\frac{\sigma t_{ij}^{\nu+1}}{\nu+1}\Big)^{\theta+1}-1\Big)exp(\beta z_{ij})\Big]\Big)\Bigg\}$$

Plusieurs tests d'inférence sont importants. D'abord, tester δ égal à 0 permet d'estimer l'homogénéité des groupes, et donc d'évaluer la pertinence de ce modèle de fragilité par rapport au modèle de survie paramétrique classique. Aalen (1991) montre que δ peut être étendu à des valeurs légérement négatives, afin que la vraisemblance sous l'hypothèse nulle devienne un point intérieur de son espace de définition. Ensuite, il est intéressant de tester l'hypothèse nulle θ égale à 0 et donc de justifier une loi Weibull généralisée.

3 Applications

3.1 Données sur la transplantation rénale

L'objectif clinique est de modéliser la survie d'un greffon rénal en fonction de facteurs propres au donneur et au receveur. Les donneurs sont tous des patients en état de mort cérébrale. Ils sont donc potentiellement impliqués dans deux greffes différentes (deux reins disponibles). Cliniquement, il est pertinent de penser que certains facteurs prédictifs du donneur non-observés soient à l'origine d'un lien entre les futurs greffons. Nous étudions ainsi le temps de survie du greffon depuis la transplantation. La base de données est constituée de 262 donneurs et de 447 greffes issues de la cohorte DIVAT du CHU de Nantes (France). Il semble que la distribution Weibull généralisée à 3 paramètres ne soit pas la plus parcimonieuse. Selon le test du rapport de vraisemblance, l'hypothèse nulle selon laquelle θ est égale à 0 ne peut pas être rejetée ($p = 0,0634$). Le tableau (1) donne les résultats obtenus, en fixant $\theta = 0$ et en utilisant le modèle de Cox fragilisé. Après une stratégie de sélection de variable en univarié (p¡0,20) et multivarié (p¡0,5), les covariables explicatives suivantes ont été retenues : l'âge de donneur (1 si supérieur à 55 ans et 0 sinon), le poids du donneur (1 si supérieur à 70 kg et 0 sinon), l'année de la greffe (1 si

supérieur à 1997 et 0 sinon) et le délai de reprise du greffon (1 si supérieur à 24 jours et 0 sinon).

	Modèle paramétrique ($ln(\mathcal{L}) = -521,76$)			Modèle semi-paramétrique ($ln(\mathcal{L}) = -417,72$)		
	Estimation	Ecart-type	p-value	Estimation	Ecart-type	p-value
$ln(\sigma)$	-5,28	0,38	<0,0001	.	.	.
ν	-0,04	0,10	0,6505	.	.	.
θ	.	.	.	.	.	.
δ	<0,0001	.	.	<0,0001	.	.
Age donneur	1,10	0,30	0,0003	1,11	0,31	0,0003
Poids donneur	-0,68	0,25	0,0063	-0,69	0,25	0,0058
Année greffe	-1,95	0,60	0,0011	-1,88	0,60	0,0018
Délai reprise	1,23	0,47	0,0092	-1,23	0,47	0,0094

Table 1: Estimations des modèles de fragilité pour la survie des greffons rénaux

Les résultats sur les effets des covariables sont strictement identiques quelque soit le modèle utilisé. Néanmoins, comme le montre la figure (1), un modèle paramétrique permet non seulement d'évaluer l'impact de facteurs sur la survie, mais aussi de calculer la survie elle-même. Notons enfin que la variance de l'effet aléatoire est nulle. Aucune hétérogénéité entre les greffes n'est due au donneur si son âge et son poids sont pris en compte. Le modèle paramétrique est donc ici équivalent à un modèle de régression de type Weibull.

3.2 Données sur le VIH

Deux marqueurs sont importants pour identifier l'état d'avancement de l'infection par le virus de l'immunodéficience acquise humaine (VIH) : la charge virale et le taux de lymphocytes CD4. L'état le plus grave de l'infection est défini par une charge virologique supérieure à 400 copies/ml et un nombre de CD4 inférieur à 200. L'analyse porte ainsi sur le temps entre la sortie d'un état grave et le retour dans cet état. Il s'agit donc d'un processus de renouvellement. L'étude est basée sur une cohorte de 293 patients suivis au CHU de Nice (NADIS). Un patient peut transiter plusieurs fois vers cet état s'il est suivi assez longtemps et si son état clinique s'améliore. Sur un total de 1055 observations, 554 transitions sont observées. La covariable principale est le mode de contamination. Nous avons choisi, par des analyses préliminaires, de la coder en trois classes : infection par toxicomanie, par rapport hétérosexuel, ou autre (homosexualité, accident, etc.). Les résultats obtenus sont présentés dans le tableau (2).

Comme précédemment, l'effet des covariables est inchangé entre les deux modèles. Cependant, pour cette application, il semble que les patients forment des groupes hétéro-

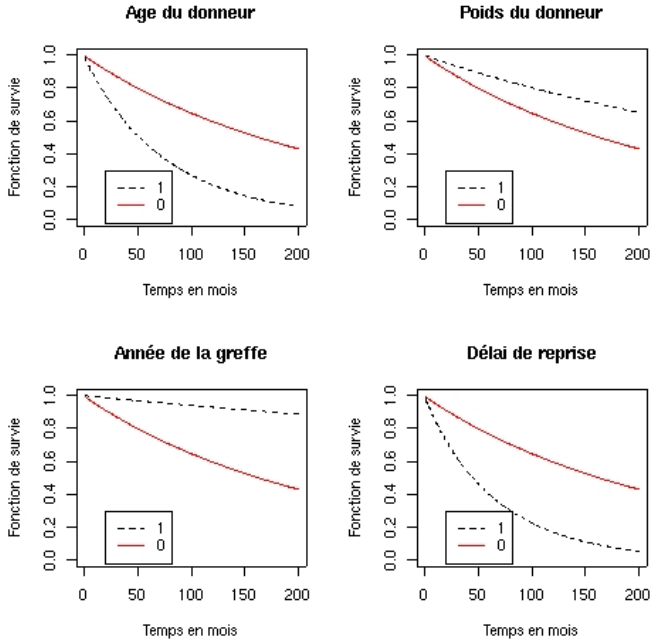

Figure 1: Taux de survie du greffon rénal par un modèle de fragilité paramétrique

gènes ($p = 0,0002$). La prise en compte d'une fragilité semble donc justifiée. Enfin, notons que la fonction de risque de base correspond à une distribution de type Weibull généralisée en forme de U inverse.

4 Discussion

Les deux exemples précédents nous montrent les avantages de ce modèle de fragilité paramétrique. Premièrement, l'intérêt d'utiliser une distribution Weibull généralisée est de pouvoir tester l'apport en information d'une forme en U ou U inverse de la fonction de risque. Nous constatons d'ailleurs pour le VIH qu'une distribution Weibull simple est peu informative par rapport à sa généralisation.

Deuxièmement, en prenant comme référence ces modèles semi-paramétriques, puisqu'ils ne font aucune hypothèse sur la distribution de la fonction de risque, les applications sur les greffons et sur le VIH montrent que le modèle paramétrique estime les mêmes coefficients de régression. Autrement dit, dans un but d'identification des facteurs explicatifs, les deux méthodes sont équivalentes. Excepté pour le nombre de paramètres à estimer, on peut alors se poser la question de l'intérêt de la non-paramétrisation des fonctions de risque, où aucune représentation de la survie et aucune prédiction ne sont possibles.

Le troisième intérêt de la méthode est l'ajout d'effets aléatoires multiplicatifs perme-

	Modèle paramétrique ($ln(\mathcal{L}) = -1895,43$)			Modèle semi-paramétrique ($ln(\mathcal{L}) = -3193,55$)		
	Estimation	Ecart-type	p-value	Ecart-type	Variance	p-value
$ln(\sigma)$	0,60	0,11	<0,0001	.	.	.
ν	1,84	0,26	<0,0001	.	.	.
θ	-0,84	0,02	<0,0001	.	.	.
δ	0,19	0,05	0,0002	.	.	.
Autre	0	.	.	.	.	.
Toxicomanie	-0,46	0,12	0,0002	-0,51	0,14	0,0003
Hétérosexualité	-0,21	0,13	0,1129	-0,26	0,16	0,0970

Table 2: Estimations des modèles de fragilité pour la dynamique du VIH

ttant l'analyse de données non-indépendantes. Seul le cas du VIH permet de mesurer l'apport en information de la fragilité. Notre estimation, basée sur la maximisation de la vraisemblance elle-même, permet d'utiliser la théorie du maximum de vraisemblance sans les approximations de l'algorithme EM.

Ce travail ouvre d'autres portes de recherches, en particulier à travers le choix de différentes distributions des effets aléatoires, même si la fonction de vraisemblance théorique (3) reste identique. La question se pose aussi de l'ajout de termes aléatoires additifs dans le predicteur linéaire des covariables. Cependant, la méthode des transformées de Laplace ne peut alors pas être appliquée et une des seules alternatives est l'estimation par EM. Enfin, ce travail peut être généralisé à la problématique multi-états avec la théorie semi-Markovienne. Des travaux sont en cours dans ce sens.

Références

[1] OO Aalen. Heterogeneity in survival analysis. Statistics in Medicine, 8:1121-1137, 1988.
[2] P Hougaard. Survival models for heterogeneous populations drived from stable distributions. Biometrika, 73:387-396, 1986.
[3] JP Klein. Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. Biometrics, 48:795-806, 1992.
[4] TM Therneau, PM Grambsch, and VS Pankratz. Penalized survival models and frailty. Journal of Computational and Graphical Statistics, 1:156-175, 2003.
[5] OO Aalen and E Husebye. Statistical analysis of repeated events forming renewal processes. Statistics in Medicine, 10:1227-1240, 1991.