Un test du log-rank stratifié avec données manquantes

Amel Mezaouer [1] & Jean-François Dupuy [2] & Kamel Boukhetala [3]

[1] *Université Saad Dahleb - Blida (Algérie), mezaouer@univ-blida.dz*
[2] *IRMAR-INSA - Rennes (France), Jean-Francois.Dupuy@insa-rennes.fr*
[3] *Université des Sciences et Technologie Houari Boumédienne - Alger (Algérie), kboukhetala@usthb.dz*

Résumé. Dans [2], les auteurs proposent un test du log-rank stratifié lorsque la variable renseignant l'appartenance aux strates est manquante au hasard et la censure dépend du groupe de traitement. Pour mettre en oeuvre ce test, il est nécessaire de disposer d'un modèle pour la distribution de probabilité des strates. Dans ce travail, nous proposons une nouvelle approche reposant sur une pondération des cas-complets. Nous comparons ces deux approches au moyen de simulations.

Mots-clés. Censure dépendante, comparaison de groupes, étude en cas-complet, simulations.

Abstract. In [2], the authors propose a stratified log-rank test appropriate when the individual strata are missing at random and the censoring depends on the treatment groups. To use this test, one needs to know the probability distribution of the stratum variable. In the present work, we propose a new test statistic based on a weighted complete-case approach. We compare both approaches via simulations.

Keywords. Dependent censoring, complete-case method, groups comparison, simulations.

1 Introduction

Le test du log-rank est souvent utilisé pour comparer des groupes de traitement randomisés en présence de durées censurées. Il se généralise au cas de données stratifiées (voir [1]). Considérons un essai clinique randomisé où n patients sont affectés aléatoirement à K groupes de traitement. On souhaite comparer ces groupes tout en ajustant un facteur S à L modalités (appelées strates). On note $\lambda_{k,l}$ la fonction de risque instantané d'un patient du groupe k, appartenant à la strate l. Les hypothèses à tester peuvent être formulées ainsi:

$$H_0 : \lambda_{1,l} = \ldots = \lambda_{K,l} \quad \text{pour tout } l = 1, \ldots, L$$

et

$$H_a : \text{ il existe } j \text{ et } j' \text{ tels que } \lambda_{j,l} \neq \lambda_{j',l} \text{ pour au moins un } l$$

Lorsque la variable S n'est pas observée pour tous les individus de l'échantillon, une solution consiste à retirer de l'étude les individus sujets à données manquantes (on parle d'analyse en cas-complet). Cette méthode conduit à de substantielles pertes de puissance (voir [2]). Dans [2], les auteurs ont donc proposé une autre méthode, applicable si l'on dispose d'un modèle d'appartenance aux strates. Nous considérons ici une nouvelle approche, qui ne nécessite pas de disposer d'un tel modèle mais pondère simplement les individus de strate connue. De plus, nous considérons le cas où la censure dépend du groupe de traitement. Cette approche et l'approche proposée dans [2] sont comparées au moyen de simulations.

2 Le test du log-rank stratifié

On considère n patients randomisés dans K groupes de traitement. On souhaite comparer les distributions de survie de ces groupes tout en ajustant un facteur S à L modalités (appelées strates). Pour chaque patient i observé sur l'intervalle de temps $[0, \tau]$, on note T_i^0 la durée de vie, C_i un instant de censure aléatoire, $T_i = \min(T_i^0, C_i)$ la durée effectivement observée et $\Delta_i = 1(T_i^0 \leq C_i)$. Les observations consistent en n vecteurs indépendants

$$(T_i, \Delta_i, G_i, S_i), i = 1, \ldots, n$$

où $G_i \in \{1, \ldots, K\}$ et $S_i \in \{1, \ldots, L\}$ indiquent respectivement le groupe et la strate du patient i. $N_i(t) = \Delta_i 1(T_i \leq t)$ et $Y_i(t) = 1(T_i \geq t)$ sont respectivement le processus de comptage et le processus à risque du patient i. On définit alors:

$$E_{k,l}^{(n)}(t) = \frac{\sum_{i=1}^n Y_i(t) 1(G_i = k) 1(S_i = l)}{\sum_{i=1}^n Y_i(t) 1(S_i = l)}.$$

La statistique du log-rank stratifié pour tester H_0 contre H_a est de la forme $(Z_1, \ldots, Z_{K-1})$ $\hat{\Theta}^{-1}(Z_1, \ldots, Z_{K-1})^\top$ où

$$Z_k = \sum_{i=1}^n \int_0^\tau \left\{ 1(G_i = k) - \sum_{l=1}^L 1(S_i = l) E_{k,l}^{(n)}(t) \right\} dN_i(t), \quad k = 1, \ldots, K-1 \qquad (1)$$

et $\hat{\Theta}$ est un estimateur de la matrice de variance asymptotique de $(Z_1, \ldots, Z_{K-1})^\top$. Sous H_0, cette statistique suit asymptotiquement un χ^2 à $(K-1)$ degrés de liberté (voir [1]).

3 Une nouvelle statistique de test

Supposons S manquante au hasard (MAR) chez certains des patients de l'échantillon. On note W une variable auxiliaire de S (W est disponible pour tous les individus) et R

l'indicatrice qui vaut 1 si S est observée et 0 sinon. Les données observées sont alors n vecteurs indépendants

$$(T_i, \Delta_i, G_i, W_i, R_i, R_i S_i), \quad i = 1, \ldots, n$$

On suppose de plus que C et G ne sont pas indépendantes (censure dépendante). Dans [2], la statistique de test suivante est proposée: $\widetilde{U} := (\widetilde{Z}_1, \ldots, \widetilde{Z}_{K-1})\widehat{\Sigma}^{-1}(\widetilde{Z}_1, \ldots, \widetilde{Z}_{K-1})^\top$ où

$$\widetilde{Z}_k = \sum_{i=1}^{n} \int_0^\tau \mu(G_i, t) \left\{ G_i^k - \sum_{l=1}^{L} D_i^l \widetilde{E}_{k,l}^{(n)}(t) \right\} dN_i(t),$$

et $G_i^k = 1(G_i = k)$, $D_i^l = R_i 1(S_i = l) + (1 - R_i)\mathbb{P}(S_i = l|W_i)$, $\mu(G_i, t) = 1/\mathbb{P}(C \geq t|G_i)$, $\widetilde{E}_{k,l}^{(n)}(t) = \widetilde{S}_{k,l}^{(n)}(t)/\widetilde{S}_l^{(n)}(t)$ et

$$\widetilde{S}_{k,l}^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) G_i^k D_i^l \mu(G_i, t), \qquad \widetilde{S}_l^{(n)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) D_i^l \mu(G_i, t).$$

$\widehat{\Sigma}$ est est un estimateur de la variance asymptotique de $(\widetilde{Z}_1, \ldots, \widetilde{Z}_{K-1})^\top$ (voir [2]). Nous proposons de modifier cette statistique de test, qui suppose de pouvoir estimer la distribution conditionnelle de S sachant W, en remplaçant D_i^l et $\widetilde{E}_{k,l}^{(n)}(t)$ dans $\widetilde{Z}_k$ par $D_i^{*l} = R_i 1(S_i = l)/\pi_i$ et $\widetilde{S}_{k,l}^{*(n)}(t)/\widetilde{S}_l^{*(n)}(t)$, avec

$$\widetilde{S}_{k,l}^{*(n)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) G_i^k D_i^{*l} \mu(G_i, t), \qquad \widetilde{S}_l^{*(n)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) D_i^{*l} \mu(G_i, t)$$

et $\pi_i = \mathbb{P}(R_i = 1|W_i, G_i, T_i, \Delta_i)$. Nous notons $\widetilde{U}^*$ la nouvelle statistique de test obtenue.

4 Etude de simulation

Nous rapportons ici une partie des études de simulation que nous avons menées. Dans cette étude, $K = 2$ et $L = 2$. La distribution de survie pour la strate 1 du groupe 1 (respectivement 2) suit une loi de Weibull $W(.5, .75)$ (respectivement $W(.5, .75r_1)$ où r_1 désigne un rapport de risques). La distribution de survie pour la strate 2 du groupe 1 (respectivement 2) suit une loi de Weibull $W(.75, 1.25)$ (respectivement $W(.75, 1.25r_2)$). Dans chaque groupe, la censure C est simulée selon une loi exponentielle assurant un pourcentage de censure c_1 dans le groupe 1 et c_2 dans le groupe 2 (avec $c_1 \neq c_2$). Nous considérons $n = 100, 200$, $(c_1, c_2) = (5, 20)$ et trois cas: (a) $(r_1, r_2) = (1, 1)$, (b) $(r_1, r_2) = (1.5, 1.5)$ et (c) $(r_1, r_2) = (1.25, 2)$. Nous considérons enfin plusieurs pourcentages de données manquantes $(15\%, 25\%, 40\%, 50\%)$. Dans [2], la distribution conditionnelle de S sachant W est estimée par une régression logistique et $\mu(G_i, t)$ est estimée par Kaplan-Meier. Nous estimons π_i par une régression logistique. Le tableau 1 donne un extrait de

Table 1: Niveau et puissance des statistiques de test $\widetilde{U}$, $\widetilde{U}^*$ et U_{CC}.

		$(c_1, c_2) = (5, 20)$											
		15%			25%			40%			50%		
n	RR	$\widetilde{U}$	$\widetilde{U}^*$	U_{CC}	$\widetilde{U}$	$\widetilde{U}^*$	U_{CC}	$\widetilde{U}$	$\widetilde{U}^*$	U_{CC}	$\widetilde{U}$	$\widetilde{U}^*$	U_{CC}
100	(1,1)	.056	.127	.071	.055	.096	.060	.058	.090	.056	.056	.125	.058
	(1.5,1.5)	.356	.527	.467	.361	.350	.400	.362	.202	.314	.369	.088	.228
	(1.25,2)	.331	.494	.468	.337	.332	.367	.331	.180	.306	.333	.082	.190
200	(1,1)	.064	.206	.065	.059	.164	.069	.063	.091	.053	.061	.102	.059
	(1.5,1.5)	.664	.890	.801	.670	.768	.716	.668	.509	.571	.677	.213	.394
	(1.25,2)	.615	.886	.803	.611	.746	.701	.622	.499	.557	.616	.182	.341

nos résultats. Dans ce tableau, nous notons U_{CC} le test du log-rank stratifié usuel obtenu à partir d'une analyse en cas-complets.

Le nouveau test proposé semble plus approprié pour des pourcentages de données manquantes faibles à modérés ($\leq 25\%$). Le test proposé dans [2] semble en revanche plus puissant lorsque le pourcentage de données manquantes est plus élevé.

Bibliographie

[1] Martinussen T., Scheike T.H. (2006). *Dynamic Regression Models for Survival Data.* Springer: New York.

[2] Mezaouer A., Dupuy J.-F., Boukhetala K. (2013). A stratified log-rank test with missing stratum information and dependent censoring. *Submitted.*