Nonparametric estimation of the survival function for ordered multivariate failure time data: a comparative study

Luís Meira-Machado[1], Marta Sestelo[1], Andreia Gonçalves[1]

[1] Centre of Mathematics & Department of Mathematics and Applications, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal

E-mail for correspondence: lmachado@math.uminho.pt

Abstract: In longitudinal studies of disease, patients may experience several events through a follow-up period. In these studies, the sequentially ordered events are often of interest and lead to problems that have received much attention recently. Issues of interest include the estimation of bivariate survival, marginal distributions and the conditional distribution of the second gap time given the first gap time. In this work we consider the estimation for the survival given the first gap time. Different nonparametric approaches will be considered for estimating these quantities, all based on the Kaplan-Meier estimator of the survival function. Real data illustration based on a German breasts cancer study is included.

Keywords: Conditional survival; Gap times; Kaplan-Meier; Nonparametric estimation; Recurrent events.

1 Introduction

In many medical studies individuals can experience several events across a follow-up study. The events of concern can be of the same nature (e.g., cancer patients can experience recurrent disease episodes) or represent different states in the disease process (e.g., alive and disease-free, alive with recurrence and dead). If the events are of the same nature, this is usually referred as recurrent events, whereas if they represent different states they are usually modeled through their intensity functions (Andersen et al., 1993). In this studies several issues are often of interest and lead to problems that have received much attention. Most of the times, one will be interested in describing the distribution of the joint gap times (see e.g., Lin

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

et al., 1999; de Uña-Álvarez and Meira-Machado, 2008; de Uña-Álvarez and Amorim, 2011). In other cases the interest is more focussed in the survival function, such as the estimation of the bivariate survival (Wang and Wells, 1997), the estimation of gap time survival functions or the conditional survival function of the gap times (Wang and Chang, 1999; Schaubel and Cai, 2004). In this work we propose four estimators for the conditional survival function in a three state progressive model. The proposed methods can be easily extended to the k-state progressive model.

2 Nonparametric estimators

Consider n independent and identically distributed pairs of successive failure (gap) times (T_{1i}, T_{2i}), $1 \leq i \leq n$. These pairs of gap times are subject to univariate right-censoring at times C_i which we assume to be independent of (T_{1i}, T_{2i}). Because of this, we only observe $(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_1, \Delta_2)$ where $\widetilde{T}_{1i} = \min(T_{1i}, C_i)$, $\Delta_{1i} = I(T_{1i} \leq C_i)$, $\widetilde{T}_{2i} = \min(T_{2i}, C_{2i})$, $\Delta_{2i} = I(T_{2i} \leq C_{2i})$ where $C_{2i} = (C_i - T_{1i})I(T_{1i} \leq C_i)$. Let $T = T_1 + T_2$ be the total time and put $\widetilde{T} = \min(T, C)$. Since the censoring time is assumed to be independent of the process, the survival function of the first gap time T_1, say S_1, may be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{T}_1, \Delta_1)$. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on the $(\widetilde{T}_i, \Delta_{2i})$'s. In this work we are interested in the estimation of the conditional survival function $S(y \mid x) = P(T > y \mid T_1 > x)$.

Recently de Uña-Álvarez and Meira-Machado (2008) proposed estimators to empirically estimate the bivariate distribution function for censored gap times. The idea behind estimation is to use the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. Since $S(y \mid x) = P(T > y|T_1 > x) = \frac{P(T>y,T_1>x)}{P(T_1>x)}$, a natural estimator for the conditional survival function is obtained using the same ideas (i.e., Kaplan-Meier weights). The proposed estimator (Kaplan-Meier Weighted Estimator, KMW) is given by $\hat{S}^{\texttt{KMW}}(y \mid x) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{1i} > x, \widetilde{T}_i > y)/\hat{S}(x)$, where $\hat{S}(x)$ is the Kaplan-Meier estimator of survival of T and where W_i are Kaplan-Meier weights attached to $\widetilde{T}_i$ when estimating the marginal distribution of T from $(\widetilde{T}_i, \Delta_i)$'s.

The conditional survival will be hard to estimate in the right tail where censoring effects are stronger. Because of this we consider alternative expressions for the conditional survival $S(y \mid x) = 1 - \frac{P(T_1>x,T\leq y)}{1-P(T_1\leq x)}$. The corresponding estimator (transformed Kaplan-Meier Weighted Estimator, tKMW) can be obtained in a similar way as introduced the KMW estimator.

Another way to introduce a nonparametric estimator for the conditional survival is by considering specific subsamples or portions of data at hand.

For example, given the time point x, to estimate $S(y \mid x) = P(T > y|T_1 > x)$ the analysis can be restricted to the individuals with a first gap time greater than x. Let $n_x = \#\{i : T_{1i} > x\}$ and introduce $\hat{S}^{\texttt{cKMW}}(y \mid x) = 1 - \sum_{i=1}^{n_x} W_i^x I(\widetilde{T}_i \leq y)$, the survival function of T computed from such a subset.

The standard error of the cKMW estimator may be large when the censoring is heavy, particularly with a small sample size. Interestingly, the variance of this estimator may be reduced by presmoothing (de Uña-Álvarez and Amorim 2011). The corresponding presmoothed estimator (Kaplan-Meier presmooth weighted estimator, cKMPW) involves replacing the censoring indicators in the building of the Kaplan-Meier weights, W_i^x, by a smooth fit (e.g. using logistic regression). In the limit case of no presmoothing, the cKMPW estimator reduces to the cKMW estimator.

3 Example of application

To illustrate our methods we will use data from a German Breast cancer study. In this dataset, a total of 686 woman with primary node positive Breast cancer were recruited in the period between 1984 and 1989. From this total 299 developed a recurrence and among these 171 died. For each patient, the two gap times (time to recurrence and time from recurrence to death) and the corresponding indicator status is recorded. Other covariates were also recorded. The covariate recurrence is the only time-dependent covariate, while the other covariates included are fixed. Recurrence can be considered as an intermediate transient state and modeled using a three-state progressive model with states "Alive and disease-free", "Alive with Recurrence" and "Dead". For illustration purposes we show in Figure 1 the plot for $S(y \mid x)$ for all four methods by fixing $T_1 = 1084$ and $T_1 = 1684$. From this plot we can see the behavior of all methods. With the exception of the KMW estimator, all perform similarly. As expected, the cKMPW estimator has less variability.

4 Conclusions

In this paper, the problem of estimating the conditional survival function for ordered multivariate failure time data has been reviewed, and four estimators has been considered. Two new sets of estimators have been proposed. Simulation results, not reported here, reveal that a new proposals perform favorably when compared with the competing methods.

Acknowledgments: The authors acknowledge financial support from FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência

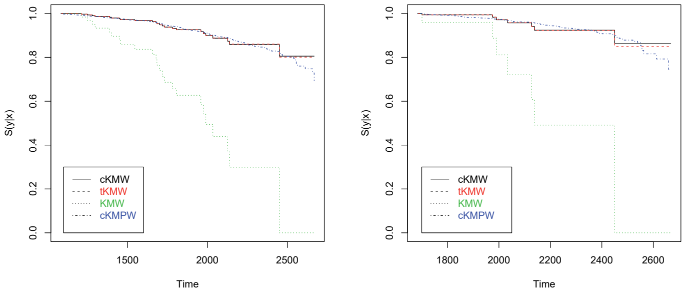

FIGURE 1. Estimated conditional survival for $S(y|x)$, $x = 1084$ (left) and $x = 1684$ right. Breast cancer data.

e a Tecnologia”, in the form of grant PEst-OE/MAT/UI0013/2014. Marta Sestelo acknowledges financial support from Grant SFRH/BPD/93928/2013 of Portuguese “Fundação para a Ciência e a Tecnologia” and by FEDER Funds through “Programa Operacional Factores de Competitividade - COMPETE”.

References

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer: New York.

de Uña-Álvarez, J. and Meira-Machado, L. (2008). A simple estimator of the bivariate distribution function for censored gap times. *Statistics and Probability Letters*, **78**, 2440 – 2445.

de Uña-Álvarez, J. and Amorim, A.P. (2011). A semiparametric estimator of the bivariate distribution function for censored gap times. *Biometrical Journal*, **53**, 113 – 127.

Lin, D., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the time distributions for serial events with censored data. *Biometrika*, **86**, 59 – 70.

Schaubel, D.E. and Cai, J. (2004). Non-parametric estimation of gap time survival functions for ordered multivariate failure time data. *Statistics in Medicine*, **23**, 1885 – 1900.

Wang, M. and Chang, S. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, **94**, 146 – 153.

Wang, W. and Wells, M. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika*, **84**, 863 – 880.