Statistics and Probability Letters 107 (2015) 101–110

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Copula-graphic estimators for the marginal survival function with censoring indicators missing at random

Yi Liu [a,*], Qihua Wang [a,b]

[a] Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100190, China
[b] Institute of Statistical Science, Shenzhen University, Shenzhen 518006, China

ARTICLE INFO

Article history:
Received 12 May 2015
Received in revised form 11 August 2015
Accepted 11 August 2015
Available online 20 August 2015

Keywords:
Archimedean copula
Dependent censoring
Survival function
Missing at random

ABSTRACT

In this paper, we propose three copula-graphic estimators for the survival function with censoring indicators missing at random in the dependent censoring situation and develop their asymptotic properties. Simulations and data analysis are conducted to evaluate their performances.

1. Introduction

In survival analysis, data are often right censored. Let T denote the failure time with distribution function $(d.f.)$ F and C be the right censoring time with $d.f.$ G. The observed event time is given by $X = T \wedge C$ and the censoring indicator is $\delta = I(T \leq C)$, $\delta = 1$ if the failure time is observed and 0 if it is censored. If T and C are independent, which is a tacit assumption in most current published papers, the marginal survival function can be estimated by the product-limit estimator proposed by Kaplan and Meier (1958). However, if T and C are dependent in a general case, the consistency of the Kaplan–Meier estimator is not ensured.

Assume that the copula function $\mathscr{C}(u_1, u_2)$ relates the joint survival function of (T, C) to their marginal survival functions.

$$P(T > t_1, C > t_2) = \mathscr{C}(\bar{F}(t_1), \bar{G}(t_2)),$$

where $\bar{F}(t) = 1 - F(t)$ and $\bar{G}(t) = 1 - G(t)$. The model has been extensively studied in Nelsen (2006).

Recent research has focused on Archimedean copula class

$$P(T > t_1, C > t_2) = \phi^{-1}(\phi(\bar{F}(t_1)) + \phi(\bar{G}(t_2))). \tag{1}$$

The function $\phi : [0, 1] \to [0, \infty]$ is called the generator of the copular $\mathscr{C}$. It is a known continuous, convex, and strictly decreasing function with $\phi(1) = 0$. In particular, $\phi(t) = -\ln(t)$ leads to $\mathscr{C}(u_1, u_2) = u_1u_2$, which indicates the independence of T and C. Many commendable authors have devoted themselves to the generalization of the Kaplan–Meier estimator to the dependent case, see Zheng and Klein (1995), Rivest and Wells (2001), Braekers and Veraverbeke (2005) and de Uña-Álvarez and Veraverbeke (2013).

[*] Corresponding author.
E-mail address: liuyi@amss.ac.cn (Y. Liu).

http://dx.doi.org/10.1016/j.spl.2015.08.010
0167-7152/© 2015 Elsevier B.V. All rights reserved.

In some practical problems, however, the censoring indicator δ is missing for a variety of reasons, unknown disease cause in competing risk model, loss of information caused by uncontrollable factors, failure on the part of investigator to gather correct information, and so forth. Let ξ be the missing indicator that $\xi = 1$ if δ is observed and $\xi = 0$ otherwise. In this paper, we assume that δ is missing at random (MAR), that is, $P(\xi = 1|X, \delta) = P(\xi = 1|X) := \pi(X)$. MAR is a common assumption for statistical analysis with missing data and is reasonable in many practical situations, see Little and Rubin (2002). There are also many methods to estimate the survival function of T on the missing censoring indicator problem, such as Subramanian (2006), Wang and Ng (2008) and so on.

In this paper, we propose copula-graphic estimators for the marginal survival function with censoring indicators missing at random (MAR). The regression surrogate, imputation and inverse probability weighting methods are used to handle the missing data. Moreover, we circumvent the assumption that failure time and censoring time are independent, and the dependence structure between them is modeled through a known copula function. The asymptotic representation and the corresponding weak convergence result of the estimators are established. A finite sample simulation is conducted to evaluate the performances of the proposed estimators. We also investigate the performances of the proposed estimators if the copula function is misspecified. Furthermore, an application to the breast cancer data set is given to illustrate the methodology.

The rest of this paper is organized as follows. In Section 2, we propose the estimators through regression surrogate, imputation and inverse probability weighting methods. In Section 3, we present the asymptotic properties. In Section 4, some simulations are conducted to evaluate the finite sample performances of the proposed estimators. In Section 5, the proposed methods are illustrated with the data set of breast cancer. We conclude this paper with a brief discussion in Section 6. The regularity conditions and technical proofs are given in the Appendix.

2. Estimation

Let H denote the distribution function of X, $H(t) = P(X \le t)$, $\bar{H}(t) = 1 - H(t)$ and take $H_1(t) = P(X \le t, \delta = 1)$. Then, from Tsiatis (1975), we have

$$\bar{F}(t) = \phi^{-1}\left(-\int_0^t \phi'(\bar{H}(s))dH_1(s)\right). \tag{2}$$

Suppose that we observe n independent and identically distributed data, $(X_i,\ \xi_i,\ \delta_i\xi_i),\ i = 1, \ldots, n$. Let $H_n(t) = 1/n\sum_{i=1}^n I(X_i \le t)$, and denote the estimator of $H_1(t)$ by $H_{1n}(t)$, then $\bar{F}(t)$ can be estimated by

$$\bar{F}_n(t) = \phi^{-1}\left(-\int_0^t \phi'(\bar{H}_n(s))dH_{1n}(s)\right). \tag{3}$$

In the following three subsections, we derive three estimators of $\bar{F}(\cdot)$ by regression surrogate, imputation and inverse probability weighting methods.

2.1. Regression surrogate estimation

It is known

$$\begin{aligned} H_1(t) &= P(X \le t, \delta = 1) = E(I(X \le t)\delta) \\ &= E(I(X \le t)m(X)) = \int_0^t m(x)dH(x), \end{aligned} \tag{4}$$

where $m(x) = E(\delta|X = x)$. Take

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n \xi_i\delta_i K_h(X_i - x)}{\sum_{i=1}^n \xi_i K_h(X_i - x)},$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. It is well known that $\hat{m}_n(x)$ can be a consistent estimator under the MAR assumption. Then, we obtain the estimator of H_1, $H_{1n,R}(t)$ say, by replacing $m(\cdot)$ and $H(\cdot)$ in (4) with $\hat{m}_n(\cdot)$ and $H_n(\cdot)$, where

$$H_{1n,R}(t) = \frac{1}{n}\sum_{i=1}^n \hat{m}_n(X_i)I(X_i \le t).$$

The regression surrogate estimator of $\bar{F}(t)$, say $\bar{F}_{n,R}(t)$, is $\bar{F}_n(t)$ with $H_{1n}(t)$ replaced by $H_{1n,R}(t)$.