A Koziol-Green Estimator for the conditional distribution function under dependent censoring

A Koziol-Green estimator for the conditional distribution function under dependent censoring.

Braekers, Roel
Hasselt University, Center for Statistics
Agoralaan, Building D
3590 Diepenbeek, Belgium
E-mail: roel.braekers@uhasselt.be

Gaddah, Auguste
Hasselt University, Center for Statistics
Agoralaan, Building D
3590 Diepenbeek, Belgium
E-mail: auguste.gaddah@uhasselt.be

Introduction

At fixed design points $0 \leq x_1 \leq \ldots \leq x_n \leq 1$, we have nonnegative responses $Y_1, \ldots, Y_n$ such as lifetimes. These responses are independent random variables and the distribution function of the response Y_i at x_i will be denoted by $F_{x_i}(t) = P(Y_i \leq t)$. In many clinical or industrial trials, the responses $Y_1, \ldots, Y_n$ are subject to random right censoring. For each response, there is a censoring variable C_i with conditional distribution function $G_{x_i}(t) = P(C_i \leq t)$. The observed random variables at design point x_i are in fact Z_i and δ_i $(i = 1, \ldots, n)$, with

$$Z_i = \min(Y_i, C_i) \qquad \text{and} \qquad \delta_i = I(Y_i \leq C_i).$$

At a given fixed design value $x \in [0, 1]$, we write F_x, G_x, H_x for the distribution function of respectively the response Y_x, the censoring variable C_x and the observed variable $Z_x = \min(Y_x, C_x)$ at x. Also we will write $\delta_x = I(Y_x \leq C_x)$. Note that for the design variables x_i, we write $Y_i, C_i, Z_i, F_i, \ldots$ instead of $Y_{x_i}, C_{x_i}, Z_{x_i}, F_{x_i}, \ldots$.

In order to estimate uniquely the conditional distribution function F_x from the observed data, we have to make an assumption about the dependence between the Y_i and C_i for each i (Tsiatis (1975)). It is very common in survival analysis to assume independence between these random variables (conditional on the covariate). However we see that in some practical situations this assumption clearly does not hold. For example in industrial testing, it may occur that some piece of equipment is taken away (is censored) because it shows some sign of future failure. As in Braekers and Veraverbeke (2005), we assume that at a fixed design value $x \in [0, 1]$, the joint survival function is given by

$$S_x(t_1, t_2) = P(Y_x > t_1, C_x > t_2) = \varphi_x^{[-1]}(\varphi_x(\bar{F}_x(t_1)) + \varphi_x(\bar{G}_x(t_2))) \tag{1}$$

where, for each x, $\varphi_x : [0, 1] \to [0, +\infty]$ is a known continuous, convex, strictly decreasing function with $\varphi_x(1) = 0$. $\varphi_x^{[-1]}$ is the pseudo-inverse of φ_x, as defined in Nelsen (1999) and given by

$$\varphi_x^{[-1]}(s) = \begin{cases} \varphi_x^{-1}(s) & 0 \leq s \leq \varphi_x(0) \\ 0 & \varphi_x(0) \leq s \leq +\infty \end{cases}.$$

From (1), we note that the conditional distribution function of the observed variable Z_x is given by

$$1 - H_x(t) = \bar{H}_x(t) = S_x(t, t) = \varphi_x^{[-1]}(\varphi_x(\bar{F}_x(t)) + \varphi_x(\bar{G}_x(t))). \tag{2}$$

The idea to use a known copula function to describe the dependence structure between the lifetime and the censoring time, was first introduced by Zheng and Klein (1995) and later improved by Rivest

and Wells (2001) for the class of Archimedean copulas. The authors of both papers did not consider covariates.

In the design of some clinical trials, we see another type of informative censoring in which the distribution function of the lifetime and the censoring time are related. Koziol and Green (1976) considered a sub-model for the Kaplan-Meier estimator where they assumed that the survival function of the censoring variable is a power of the survival function of the lifetime. This sub-model has the advantage that the estimator for the distribution function of the lifetime has a simpler form. Veraverbeke and Cadarso Suárez (2000) extended this model for the fixed design regression situation.

In this paper we will further extend this sub-model to the case where the lifetime Y_x depends on the censoring variable C_x. We therefore use the fact that the classical Koziol-Green model is characterized by the conditional independence of Z_x and δ_x. Translating the latter property into our model (1) leads to the following assumption: for each covariate value $x \in [0,1]$,

$$\bar{G}_x(t) = \varphi_x^{[-1]}(\beta_x \varphi_x(\bar{F}_x(t))), \qquad \forall t > 0 \tag{3}$$

where $\beta_x > 0$ is a constant depending only on x. When we consider both types of informative censoring, we rewrite (2) as

$$\bar{H}_x(t) = \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \beta_x \varphi_x(\bar{F}_x(t))\right) = \varphi_x^{[-1]}\left((\beta_x + 1)\varphi_x(\bar{F}_x(t))\right). \tag{4}$$

From this relation we find a conditional distribution estimator F_{xh} for F_x where $x \in]0,1[$ is a fixed design value by rewriting this equation as

$$\bar{F}_x(t) = \varphi_x^{[-1]}\left(\gamma_x \varphi_x(\bar{H}_x(t))\right) \tag{5}$$

with $\gamma_x = \frac{1}{\beta_x + 1} = P(\delta_x = 1)$. In (5), we replace the different quantities $H_x(t)$ and γ_x by estimators. As in other work with non-parametric regression (Veraverbeke and Cadarso Suárez (2000), Braekers and Veraverbeke (2005)), we consider estimators which involve a sequence of smoothing weights $\{w_{ni}(x, h_n)\}$, depending on a positive bandwidth sequence $\{h_n\}$, tending to zero, as $n \to +\infty$. In our present situation of fixed design points, it is customary to take the Gasser-Müller type weights, given by, $\forall i = 1, \ldots, n$,

$$w_{ni}(x, h_n) = \frac{1}{c_n(x, h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} K\left(\frac{x - z}{h_n}\right) dz \quad \text{with} \quad c_n(x, h_n) = \int_0^{x_n} \frac{1}{h_n} K\left(\frac{x - z}{h_n}\right) dz.$$

Here $x_0 = 0$ and K is a known probability density function (kernel). For the conditional distribution function $H_x(t)$, we take a Stone type estimator (Stone (1977)) given by

$$H_{xh}(t) = \sum_{i=1}^{n} w_{ni}(x, h_n) \mathrm{I}(Z_i \leq t).$$

A similar estimator is taken for the exponent γ_x and is given by

$$\gamma_{xh} = \sum_{i=1}^{n} w_{ni}(x, h_n) I(\delta_i = 1).$$

Combining these estimators in (5), we find an estimator for the conditional distribution function $F_x(t)$ by

$$\bar{F}_{xh}(t) = \varphi_x^{[-1]}\left(\gamma_{xh} \varphi_x(\bar{H}_{xh}(t))\right).$$

Note that the estimator $\bar{F}_{xh}(t)$ has a simpler structure than the copula-graphic estimator of Braekers and Veraverbeke (2005) for the more general model under dependent censoring. Furthermore we see that in our estimator the estimator for γ_x only depends on the δ_i while the estimator for $H_x(t)$ only

depends on the Z_i. This result follows from assumption (3), which is equivalent to the assumption that Z_x and δ_x are conditionally independent. If we take $\varphi_x(t) = -\log(t)$, we see that this estimator equals the estimator of Veraverbeke and Cadarso Suárez (2000) as we expected.

Weak convergence result

In this article, we prove, under some mild smoothness conditions on the conditional distribution $F_x(t)$, the weak convergence of process $(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot))$ associated to the Koziol-Green estimator $F_{xh}(t)$. As in the work of Veraverbeke and Cadarso Suárez (2000) and Braekers and Veraverbeke (2005), we first need to derive an almost sure representation for this estimator. This result has already been found by Braekers and Veraverbeke (2007). In this paper we publish their result to clarify the following part of the proof.

Theorem 1. (Almost sure representation) Under some regularity conditions, $h_n \to 0$, $\frac{\log n}{nh_n} \to 0$, $\frac{nh_n^5}{\log n} = O(1)$, $T < T_{F_x} = \inf\{t | F_x(t) = 1\}$. Then, for $t < T_{F_x}$,

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^{n} w_{ni}(x, h_n) g_{tx}(Z_i, \delta_i) + R_n(x, t)$$

where

$$g_{tx}(Z_i, \delta_i) = -\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \left(I(\delta_i = 1) - \gamma_x\right) + \frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \left(I(Z_i \leq t) - H_x(t)\right)$$

and as $n \to +\infty$, $\sup_{0 \leq t \leq T} |R_n(x, t)| = O((nh_n)^{-1} \log n)$ a.s.

Due to this asymptotic representation, we show the weak convergence of $(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot))$ by proving the weak convergence of the main term in this representation which is a weighted sum of independent functions of the observed quantities. After lengthy, but straightforward calculations we also find the asymptotic bias and variance expressions.

Theorem 2. (Weak convergence) Under some regularity conditions, $T < T_{F_x}$,

(a) if $nh_n^5 \to 0$ and $(nh_n)^{-1/2} \log n \to 0$, then as $n \to \infty$,

$$(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to W(\cdot|x) \text{ in } l^\infty[0, T]$$

(b) If $h_n = Cn^{-1/5}$ for some $C > 0$, then, as $n \to \infty$,

$$(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to \tilde{W}(\cdot|x) \text{ in } l^\infty[0, T]$$

where $W(\cdot|x)$ and $\tilde{W}(\cdot|x)$ are Gaussian processes with covariance function given by

$$\Gamma_x(t, s) = \frac{||K||_2^2 \{\varphi_x(\bar{H}_x(t))\varphi_x(\bar{H}_x(s))\gamma_x(1 - \gamma_x) + \gamma_x^2 \varphi'_x(\bar{H}_x(t))\varphi'_x(\bar{H}_x(s))(H_x(s \wedge t) - H_x(t)H_x(s))\}}{\varphi'_x(\bar{F}_x(t))\varphi'_x(\bar{F}_x(s))}.$$

and for $\tilde{W}(\cdot|x)$, mean function given by

$$b_{tx} = \frac{1}{2}\mu_2^K C^{5/2} \left\{ \frac{-\varphi_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \ddot{\gamma}_x + \frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(\bar{F}_x(t))} \ddot{H}_x(t) \right\}.$$

In any fixed time point, we note that the asymptotic variance of the Koziol-Green estimator is always smaller than the asymptotic variance of the copula-graphic estimator of Braekers and Veraverbeke (2005), when applied under this sub-model. For the copula-graphic estimator, the asymptotic

variance has an extra term given by

$$-\gamma_x(1-\gamma_x)\left\{\varphi_x(\bar{H}_x(t))^2-\int_0^t\varphi_x'(\bar{H}_x(w))^2dH_x(w)\right\}$$

which is always positive by the Cauchy-Schwartz inequality.

An application: A confidence band

The weak convergence result established in the previous section, can be used as a starting point to derive some practical applications. In this section we develop an asymptotic confidence band for the Koziol-Green estimator $F_{xh}(t)$. This result is given in the following theorem.

Theorem 3. (Confidence band) Under some regularity conditions with $T < T_{F_x}$, $nh_n^5 \to 0$, $(nh_n)^{-1/2}\log n \to 0$. For each $0<\alpha<1$, let $c_{\alpha xh}$ be such that, as $n \to +\infty$,

$$P\left(\sup_{0\le t\le T}\left|B_1(H_{xh}(t))+\frac{\varphi_x(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi_x'(\bar{H}_{xh}(t))}B_2(\gamma_x)\right|\le c_{\alpha xh}\right)\to 1-\alpha,$$

Then, as $n \to +\infty$,

$$P(F_{xh}(t)-c_{\alpha xh}D_{xh}(t)\le F_x(t)\le F_{xh}(t)+c_{\alpha xh}D_{xh}(t), \text{for all } 0\le t\le T)\to 1-\alpha$$

where $B_1(s)$ and $B_2(s)$ are independent Brownian bridges and $D_{xh}(t)=(nh_n)^{-1/2}\dfrac{\gamma_{xh}\varphi_x'(\bar{H}_{xh}(t))}{\varphi_x'(\bar{F}_{xh}(t))}$.

References

R. Braekers and N. Veraverbeke, (2005) A copula-graphic estimator for the conditional survival function under dependent censoring, *The Canadian Journal of Statistics*, **33**, 429-447.

R. Braekers and N. Veraverbeke, (2007) A conditional Koziol-Green model under dependent censoring, Accepted to be published by *The Probability and statistics letters*.

J.A. Koziol and S.B. Green, (1976) A Cramér-von Mises statistic for randomly censored data, *Biometrika* **63**, 465-474.

R.B. Nelsen, (1999) *An introduction to copulas* (Springer-Verlag, New York).

C.J. Stone, (1977) Consistent nonparametric regression, *Ann. Statist.* **5**, 595-645.

A. Tsiatis, (1975) A nonidentifiability aspect of the problem of competing risks, *Proc. Nat. Acad. Sci. USA.* **72**, 20-22.

L. Rivest and M.T. Wells, (2001) A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *J. Multivariate Analysis* **79**, 138-155.

N. Veraverbeke and C. Cadarso Suárez, (2000) Estimation of the conditional distribution in a conditional Koziol-Green model, *Test*, **9**, 97-122.

M. Zheng and J.P. Klein, (1995) Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika* **82**, 127-138.