International Journal of Pure and Applied Mathematics

Volume 5 No. 1 2003, 93-101

A NOTE ON PARTITIONING ESTIMATE OF CONDITIONAL DISTRIBUTION UNDER CENSORING

Ali Gannoun[1 §], Jérome Saracco[2], George E. Bonney[3]

[1,3]Statistical Genetics and Bioinformatics Unit
National Human Genome Center at Howard University
2216 6th Street, Suite 206
Washington D.C. 20059, USA
[1]e-mail: gannoun@howard.edu
[3]e-mail: ge_bonney@howard.edu

[1,2]Laboratory of Probabilities and Statistics, cc 051
University of Montpellier II
34095 Montpellier Cedex 05, FRANCE
[1]e-mail: gannoun@stat.math.univ-montp2.fr
[2]e-mail: saracco@stat.math.univ-montp2.fr

Abstract: Let X be a random variable taking values in $\mathbb{R}$ and let Y be a nonnegative bounded random variable. Assume a right censoring random variable C, with continuous distribution function, operating on Y such that Y and C are conditionally independent on given X. In this randomly censored situation, we want to estimate the conditional distribution of Y given X. For this purpose, we construct a nonparametric partitioning estimate $F_n(y|x)$ which is regressogram-like mean regression function estimate, and prove its uniform consistency using Dvoretzky-Kiefer-Wolfwitz [1] type inequality under censoring.

AMS Subject Classification: 62N01, 62N02

Key Words: censoring, conditional distribution, Kaplan-Meier estimator, partitioning estimate

Received: February 11, 2003

[§] Correspondence author

1. Introduction

In life time data analysis, nonparametrically estimated conditional survival curves (such as the conditional Kaplan-Meier estimate) are useful for assessing the influence of risk factors, predicting survival probabilities, and checking goodness-of-fit of various survival regression models. Let Y and C be nonnegative random variables having a joint distribution that depends on a covariate X, and assume that Y is independent of C given X. Set $F(y|x) = P(Y \leq y|X = x)$ and $G(y|x) = P(C \leq y|X = x)$. In survival analysis, Y and C are referred to as the survival time (or failure time) and censoring time, respectively. Let $Z = \min(Y, C)$ and $\delta = \mathbb{I}(Y \leq C)$, where $\mathbb{I}(.)$ is the indicator function.

It is well known that in medical studies the observation on the survival time of a patient is often incomplete due to right censoring. Classical examples of the causes of this type of censoring are that the patient was alive at the termination of the study, that the patient withdrew alive during the study, or that the patient died from other causes than those under study.

We observe (X_i, Z_i, δ_i), $i = 1, ..., n$ which are n independent replications of (X, Z, δ). Then our goal is to estimate nonparametrically the conditional distribution function $F(y|x)$ from the censored data. The first fully nonparametric approach was given by Beran [2], who introduced a class estimators for the conditional survival functions in the presence of right censoring, and proved their strong consistency. Dabrowska [3] proved weak convergence results for these estimators. More development about these estimators can be found also in McKeague and Utikal [4], Dabrowska [5], Li and Doss [6] and Li [7], among others. An alternative approach was taken by Horváth [8], who proposed an estimator of the conditional survival function by integrating an estimator of the conditional density. Here, we introduce a so-called partitioning Kaplan-Meier estimate $F_n(y|x)$, which generalizes the notion of regressogram estimate to the situation under censoring, and prove its uniform consistency. Our approach is similar to that developed by Carbonez et al [9] for estimating the conditional expectation. We emphasize that our model is fully nonparametric like the ones treated by Beran and Dabrowska.

The next section recalls the classical Kaplan-Meier estimator and provides the explicit expression for the partitioning proposed here. The main result regarding the consistency of this estimator is also given in this section. Finally, Section 3 contains the proof of this main result.

2. Estimation and Main Result

Let us first recall the Kaplan-Meier estimator. Then we exhibit the partitioning estimator of conditional distribution and study its consistency.

2.1. The Kaplan-Meier Estimator

The survival functions of Y and C are defined by $S(t) = 1 - F(t) = P(Y > t)$ and $R(t) = 1 - G(t) = P(C > t)$, where F and G are the distribution function of Y and C, respectively. Set $T_S = \sup\{t : S(t) > 0\}$, $T_R = \sup\{t : G(t) > 0\}$ and $T_K = \min\{T_S, T_R\}$.

Let S_n be the estimator of S introduced by Kaplan and Meier [10]. As often in the literature we treat the largest observation as uncensored. In this case, S_n is expressed as

$$S_n(t) = \begin{cases} \prod_{Z_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{if } \ t \leq T_{K,n}, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $Z_{(1)} \leq \cdots \leq Z_{(n)}$ are the order statistics of $Z_1, \ldots, Z_n$, $\delta_{(i)}$ is the δ_i associated with the $Z_{(i)}$, and $T_{K,n} = \max\{Z_1, \ldots, Z_n\}$. It is easy to show that $T_{K,n} \longrightarrow T_K$.

Obviously, from (1) an estimator of the distribution function F is given by

$$F_n(t) = 1 - S_n(t) = \begin{cases} 1 - \prod_{Z_{(i)} \leq t} \left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{if } \ t \leq T_{K,n}, \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Note that, if S is arbitrarily chosen, some of Z_i may be identical. So that the ordering is not unique. However, the Kaplan-Meier estimator is unique (see Peterson [11]).

Let $\varepsilon = \inf\{t \in \mathbb{R}; (1 - F(t))(1 - G(t)) = 0\}$. For $\tau < \varepsilon$, Földes and Rejtö [12] proved the following Dvoretzky-Kiefer-Wolfwitz exponential inequality:

$$P\left(\sqrt{n} \sup_{t \leq \tau} |F_n(t) - F(t)| \geq \lambda)\right) \leq \alpha e^{-\eta((1-F(\tau))(1-G(\tau)))^4 \lambda^2}, \tag{3}$$

where α and η are absolute constants.

A Glivenko-Cantelli Lemma was obtained e.g. by Stute and Wang [13]. Their results imply that, for continuous S,

$$\sup_{t \leq T_K} |S_n(t) - S(t)| \longrightarrow 0 \ \ (n \longrightarrow \infty).$$

Remark 1. Bitouzé et al [14] improved Földes and Rejtö result and gave more explicit bound in (3). Nevertheless, their result is not usable in this note.

Let us now introduce the partitioning estimator.

2.2. The Partitioning Kaplan-Meier Estimator

Suppose that the covariate X is available for predicting Y. Let $(\pi_{n,q})_{q\in Z}$ a sequence of partitions of $\mathbb{R}$ such that

$$\pi_{n,q} = [(q-1)h_n, qh_n[\ , \tag{4}$$

where h_n is a positive real number. For a given x, find the $\pi_{n,j}$ to which x belongs (this set is denoted by $\pi_n(x)$) and select the Y_i, whose corresponding X_i's fall in $\pi_n(x)$. In the uncensored case, the partitioning estimator $F_{1,n}(t|x)$ of $F(t|x)$ is defined as follows:

$$F_{1,n}(t|x) = \begin{cases} \dfrac{\sum\limits_{i=1}^{n} \mathbb{I}(Y_i \le t)\mathbb{I}(X_i \in \pi_n(x))}{\sum\limits_{i=1}^{n} \mathbb{I}(X_i \in \pi_n(x))} & \text{if } \sum\limits_{i=1}^{n} \mathbb{I}(X_i \in \pi_n(x)) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The a.s. consistency of $F_{1,n}$ was proved by Gannoun [14].

In the censored case (where instead of the (X_i, Y_i) only the (X_i, Z_i, δ_i) are available), let us denote $q(x) = \text{card}\,(\pi_n(x))$, i.e. the number of X_i in $\pi_n(x)$, and by $Z_1^\pi, ..., Z_{q(x)}^\pi$ (respectively $\delta_1^\pi, ..., \delta_{q(x)}^\pi$) the observations Z_i (respectively δ_i) such that X_i belong to $\pi_n(x)$. As in (2), we also denote $F_{q(x)}^\pi$ the Kaplan-Meier estimator obtained with $(Z_i^\pi, \delta_i^\pi), i = 1, ..., q(x)$. Then, the partitioning estimate $F_n(t|x)$ of $F(t|x)$, under censoring, is defined by

$$F_n(t|x) = \begin{cases} F_{q(x)}^\pi(t) & \text{if} \quad q(x) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We are able now to formulate our main result.

2.3. Main Result

Let us first write $\varepsilon_x = \inf\{t \in \mathbb{R}; (1 - F(t|x))(1 - G(t|x)) = 0\}$, and τ such that $0 < \tau < \varepsilon_x$.

Theorem. *Assume that:*

(H1) the marginal distribution of X admits a density f satisfying:

$$\text{for any } x \in \mathbb{R}, \text{ there exists } \rho > 0 \\ \text{such that } f(x) > \rho \text{ in a neighborhood of } x,$$

(H2) $h_n \longrightarrow 0$ *and* $\dfrac{nh_n}{\log n} \longrightarrow \infty$.
Then, for all $x \in \mathbb{R}$,

$$\sup_{t \leq \tau} |F_n(t|x) - F(t|x)| \longrightarrow 0 \quad \text{as} \quad (n \longrightarrow \infty).$$

Remark 2. It is very easy to extend the result to multivariate X. In this case, the support of X should be partitioned, for example, on hypercubes. The hypotheses are still the same except $\frac{nh_n}{\log n} \to \infty$, which is replaced by $\frac{nh_n^d}{\log n} \to \infty$, where d is the dimension of X.

Remark 3. It is also possible to use another partition on $\mathbb{R}$ (or on the support of X). This partition should satisfy some regularity conditions (see Carbonez et al [9] or Bosq and Lecoutre [16] for more details).

Remark 4. If the values of X belong to a compact set, it is easy to obtain the uniform convergency on both x and y. The idea is to cover this compact set by a finite number of intervals and to proceed as in Berlinet et al [17], for example.

Remark 5. The idea of our estimator is of course similar to that of Beran, but its construction is based on a fixed partitioning. Moreover, our proof of the uniform consistency presented below is different and uses more recent exponential inequality. Similar result is also given by Van Keilegom and Veraverbeke [18].

3. Proof of the Theorem

Let $F(x,y)$ denote the joint distribution function of (X,Y), and $p(x,h_n) = P\left((X,Y) \in \pi_n(x)\right)$. It is clear that

$$p(x,h_n) = F(q^* h_n, +\infty) - F((q^*-1)h_n, +\infty)\,, \tag{6}$$

where q^* is such that $\pi_n(x) = \pi_{n,q^*}$.

Set

$$d((F_n(t|x) - F(t|x)) = \sup_{t \le \tau} |F_n(t|x) - F(t|x)|\,. \tag{7}$$

By Total Probability Theorem and (6), we get

$$\begin{aligned} &P\left(d((F_n(t|x) - F(t|x)) > \lambda\right) \\ &\qquad = \sum_{j=0}^{n} \tbinom{n}{j} p(x,h_n)^j \left(1 - p(x,h_n)\right)^{n-j} P\left(d((F_n(t|x) - F(t|x))\right. \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. > \lambda | q(x) = j\right). \end{aligned} \tag{8}$$

From (5), it follows that

$$P\left(d((F_n(t|x) - F(t|x)) > \lambda | q(x) = j\right) = P(d(F_j^{\pi}(t), F(t|x)) > \lambda). \tag{9}$$

By (3) and (9), (8) becomes

$$\begin{aligned} & P\left(d((F_n(t|x) - F(t|x)) > \lambda\right) \\ \le\ & \alpha \tbinom{n}{j} p(x,h_n)^j \left(1 - p(x,h_n)\right)^{n-j} e^{-\eta((1-F(\tau|x))(1-G(\tau|x)))^4 j \lambda^2} \\ =\ & \alpha \sum_{j=0}^{n} \tbinom{n}{j} (p(x,h_n) e^{-\eta((1-F(\tau|x))(1-G(\tau|x)))^4 \lambda^2})^j \\ \times\ & (1 - p(x,h_n))^{n-j} \\ =\ & \alpha \left((1 - p(x,h_n)(1 - e^{-\eta((1-F(\tau|x))(1-G(\tau|x)))^4 \lambda^2})\right)^n . \end{aligned} \tag{10}$$

Now, let $\beta = \eta((1-F(\tau|x))(1-G(\tau|x)))^4$, and observe that

$$\left((1 - p(x,h_n)(1 - e^{-\beta\lambda^2})\right)^n = \exp\left(n \log\left(1 - p(x,h_n)(1 - e^{-\beta\lambda^2})\right)\right).$$

By the *Mean Value Theorem* and using (6), there exists $\xi \in\,](q^*-1)h_n, q^* h_n[$ such that

$$p(x,h_n) = h_n f(\xi). \tag{11}$$

Moreover, $\log(1-x) \leq -x$, for $x \ll 1$. Therefore, using *(H1)* and (11), we get for n sufficiently large

$$n \log\left(1 - p(x,h_n)(1-e^{-\beta\lambda^2})\right) \leq -nh_n\rho(1-e^{-\beta\lambda^2}) \\ \leq -nh_n\rho\beta\lambda^2, \quad (12)$$

because $1-e^{-\beta\lambda^2} \leq \beta\lambda^2$. Now, from (10), (11) and (12), it follows

$$P\left(d((F_n(t|x) - F(t|x)) > \lambda\right) \leq \alpha \exp(-nh_n\rho\beta\lambda^2), \quad (13)$$

which is very similar to Dvoretzky-Kiefer-Wolfwitz inequality obtained by Földes and Rejtö [12].

Using *Borell-Cantelli Lemma*, the theorem is established if we prove the following:

$$\sum_{n=0}^{\infty} P\left(d((F_n(t|x) - F(t|x)) > \lambda\right) < \infty.$$

From (13), and using *(H2),* we get

$$\sum_{n=0}^{\infty} P\left(d((F_n(t|x) - F(t|x)) > \lambda\right) \\ \leq \alpha \sum_{n=0}^{\infty} \exp(-nh_n\rho\beta\lambda^2) = \alpha \sum_{n=0}^{\infty} n^{-(\frac{nh_n\rho\beta\lambda^2}{\log n})} < \infty. \quad \square$$

Acknowledgment

The work was supported by the United States Public Service Grant No AG 16996 from the National Institutes of Health.

References

[1] A. Dvoretzky, J.C. Kiefer, J. Wolfowitz, Asymptotic minimax caracter of the sample distribution function and of the classical multinomial estimator, *Ann. Math. Statist.*, **33** (1956), 642-669.

[2] R. Beran, Nonparametric regression with randomly censored survival data, *Technical Report*, Univ. California, Berkeley (1981).

[3] D.M. Dabrowska, Non-parametric regression with censored survival time data, *Scand. J. Stat.*, **14** (1987), 181-197.

[4] I.W. McKeague, K.J. Utikal, Inference for a nonlinear counting process regression model, *Ann. Stat.*, **18** (1990), 1172-1187.

[5] D.M. Dabrowska, Uniform consistency of kernel conditional Kaplan-Meier estimate, *Ann. Stat.*, **14** (1989), 1157-1167.

[6] G. Li, H. Doss, An approach to non parametric regression for life history data using local linear fitting, *Ann. Stat.*, **23** (1995), 787-823.

[7] G. Li, Optimal rate local smoothing in a multiplicative intensity counting process model, *Math. Methods. Stat.*, **6** (1997), 224-244.

[8] L. Horváth, On nonparametric regression with randomly censored data, In: *Proc. Third Pannonian Symp. on Math. Statat* (Ed-s: J. Mogyorodi, I. Vincze, W. Wertz), Visegrad, Hungary (1981), 105-112.

[9] A. Carbonez, L. Györfi, E.C. van der Meulen, Partitioning-estimates of a regression function under random censoring, *Stat. Decis.*, **13** (1995), 21-37.

[10] E. Kaplan, P. Meier, Nonparametric estimation from incomplete observation, *J. Amer. Stat. Assoc.*, **53** (1958), 457-481.

[11] A.V. Jr Peterson, Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions, *J. Amer. Statist. Assoc.*, **72** (1977), 854-858.

[12] A. Földes, L. Rejtö, A LIL type result for the product limit estimator, *Z. Wahrscheinlichkeitsheor. Verw. Geb.*, **56** (1981), 75-86.

[13] W. Stute, J.L. Wang, The strong law under random censorship, *Ann. Stat.*, **21** (1993), 1591-1607.

[14] D. Bitouzé, B. Laurent, P. Massart, A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator, *Ann. Inst. Henri Poincaré, Probab. Stat.*, **35** (1999), 735-763.

[15] A. Gannoun, *Estimation non Paramétrique de la Médiane Conditionnelle*, PHD, Paris VI (1989).

[16] D. Bosq, J.P. Lecoutre, *Théorie de L'estimation Fonctionnelle*, Economica (1987).

[17] A. Berlinet, A. Gannoun, E. Matzner-Løber, Asymptotic normality of convergent estimates of conditional quantiles, *Statistics*, **35** (2001), 139-169.

[18] I. Van Keilegom, N. Veraverbeke, Uniform strong convergence results for the conditional Kaplan-Meier estimator and its quantiles, *Commun. in Stat., Theory and Methods*, **25** (1996), 2251-2265.