II PhDay-EIO
Santiago de Compostela, 23 de xuño do 2021

Nonparametric kernel estimation in the mixture cure model when the event indicator is available for some censored observations

Wende Clarence Safari[1], Ignacio López-de-Ullibarri[1], María Amalia Jácome[1]

[1]Department of Mathematics. Universidade da Coruña.

ABSTRACT

In time-to-event data it is commonly assumed that all individuals will experience the considered event if observed for enough time. However, cure models have been developed because there might be situations where the standard survival model is not true. Mixture cure models (MCM) assume that population is a mixture of two sub-groups: those where the event is certain not to occur, known to be “statistically cured” (or long-term) survivors, and those who will develop the event, known to be “uncured” (or susceptible) subjects.

Standard cure models typically make inference based on the assumption that the event indicator (cure status) is a latent variable as the event is only known for the uncensored (uncured) subjects, but it is unknown for the censored observations, whether cured or not. In the absence of the cure status information, Beran [1] estimator of the conditional survival function under random right censorship has been used to derive two completely nonparametric estimators: an estimator for the cure probability [9] and an estimator for the conditional survival function (latency) of the uncured subjects [4].

There are situations where the event indicator is known for some of the censored individuals, as they can be identified to be insusceptible to the event of interest, that is, known to be cured. For example, in some cases the individual is assumed to be cured from the event if the event did not happen for a fixed time period called cure threshold. In other cases, medical diagnostic procedures can provide evidence that an individual will not relapse. In this context, the event indicator is not a latent variable any more, and its incorporation to the MCM-based estimators improves the estimation significantly.

The MCM-based estimation of the cure model by introducing the event indicator information has been addressed very recently from a semiparametric point of view [2]. Here, we present kernel methods to estimate the cure probability and the latency function. Moreover, we introduce and study a multiply imputed Nadaraya-Watson (MI-NW) estimator as an alternative estimator of the cure probability when the event indicator is partially observed, based on regression techniques for missing response data. The proposed approach contributes to the state-of-the-art in time-to-event data, as it extends previous works in the mixture cure model.

Let Y be the survival time until the event of interest happens, C^* a random censoring time and, X a covariate. Assume that the survival time Y is subject to random right censoring, so that instead of observing Y, only $T^* = \min(Y, C^*)$ and $\delta = 1(Y < C^*)$ are observed. The random variables Y and C^* are assumed to be conditionally independent given $X = x$. We set $Y = \infty$ if the subject will never experience the event of interest, and therefore is cured. Let $\nu = \mathbf{1}(Y = \infty)$ be the indicator of being cured. In the presence of random-censored observation, δ is observed for all the individuals while ν is observed only for the uncensored ones (which are uncured), with $\nu = 0$. When the cure status is partially known, $\nu = 1$ is also observed for some censored individuals.

To accommodate the cure status information, we include an additional random variable ξ, which indicates whether the cure status is known ($\xi = 1$) or not ($\xi = 0$). Let the censoring distribution be an improper distribution function $G(t|x) = \{1 - \pi(x)\} G_0(t|x)$, so with probability $\pi(x)$ the censoring variable is $C^* = \infty$, and with probability $\{1 - \pi(x)\}$ the value of the censoring variable C^* corresponds to the value of a random variable C with proper continuous distribution function $G_0(t|x)$. In this setup, the data actually observed are $\{(X_i, T_i, \delta_i, \xi_i, \xi_i \nu_i) : i = 1, \ldots, n\}$, where for individuals not identified as cured the observed time is $T_i = T_i^*$ and if an individual is cured, the observed time is not $T_i = \infty$ but $T_i = C_i$. Hence, the observations $\{(X_i, T_i, \delta_i, \xi_i, \xi_i \nu_i) : i = 1, \ldots, n\}$ can be classified into three groups: (a) the individual is observed to have experienced the event and therefore known to be uncured $(X_i, T_i = Y_i, \delta_i = 1, \xi_i = 1, \xi_i \nu_i = 0)$; (b) the event is not observed and the cure status is unknown $(X_i, T_i = C_i, \delta_i = 0, \xi_i = 0, \xi_i \nu_i = 0)$; and (c) the event is not observed and the individual is known to be cured $(X_i, T_i = C_i, \delta_i = 0, \xi_i = 1, \xi_i \nu_i = 1)$. In standard cure models, when the cure status is unknown for all the censored observations, only groups (a) and (b) are considered. The probability of cure is $1 - p(x) = P(Y = \infty | X = x)$, the probability of the event is $p(x)$, and the conditional survival function of the uncured individuals, also known as latency, is the continuous function $S_0(t|x) = P(Y > t | Y < \infty, X = x)$. The mixture cure model specifies the survival function $S(t|x) = P(Y > t | X = x)$ as

$$S(t \mid x) = 1 - p(x) + p(x) S_0(t \mid x). \tag{1}$$

Assuming model (1) and the availability of a suitable estimator of the $S(t|x)$, estimators of the cure probability and the latency can be derived by considering the following relationships:

$$1 - p(x) = \lim_{t \to \infty} S(t \mid x) > 0, \; S_0(t \mid x) = \frac{S(t \mid x) - \{1 - p(x)\}}{p(x)}.$$

The proposed estimator of the cure probability $1 - p(x)$ is

$$1 - \widehat{p}_h^c(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]} B_{h[i]}(x)}{\sum_{j=i}^{n} B_{h[j]}(x) + \sum_{j=1}^{i-1} B_{h[j]}(x) \mathbf{1}\left(\xi_{[j]} \nu_{[j]} = 1\right)} \right), \tag{2}$$

where $X_{[i]}$, $\delta_{[i]}$, $\xi_{[i]}$ and $\nu_{[i]}$ are the concomitants of the ordered observed times $T_{(1)} \leq \cdots \leq T_{(n)}$, $B_{h[i]}(x)$ are the Nadaraya-Watson weights,

$$B_{h[i]}(x) = \frac{K_h\left(x - X_{[i]}\right)}{\sum_{j=1}^{n} K_h(x - X_j)},$$

and $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function $K(\cdot)$ rescaled with bandwidth $h \to 0$ as $n \to \infty$. The proposed latency estimator is

$$\widehat{S}_{0,h_1,h_2}^c(t \mid x) = \begin{cases} 1 & \text{if } t < 0 \\ \frac{\widehat{S}_{h_2}^c(t|x) - \{1 - \widehat{p}_{h_1}^c(x)\}}{\widehat{p}_{h_1}^c(x)} & \text{if } 0 \leq t \leq T_{(n)}^1 \\ 0 & \text{if } t \geq T_{(n)}^1 \end{cases} \tag{3}$$

where $T_{(n)}^1 = \max T_{(i)}$ is the largest uncensored time, $\widehat{S}_{h_2}^c(t|x)$ is the generalized product-limit estimator for the conditional survival function, $S(t|x)$, in [6], and $1 - \widehat{p}_{h_1}^c(x)$ is the nonparametric kernel estimator of the conditional probability $1 - p(x)$, in (2).

The main advantage of these estimators over the available counterparts in the literature is that they benefit from the event indicator information whenever available, while being independent on any prefixed (semi)parametric assumption, hard to be tested in practice, or on the missingness rate as it can be computed even when the event indicator is completely unobserved for the censored observations. The proposed estimator in (2) is a more efficient extended version of well-established estimators: it reduces to the Nadaraya-Watson [5, 8] estimator without censoring, to Xu and Peng [9] estimator when the cure status information is disregarded, or to Laska and Meisner [3] estimator without covariates. The estimator in (3) reduces to López-Cheda et al. [4] estimator when the cure status information is ignored and one single bandwidth $h_1 = h_2$ is used in the estimation.

Theoretical properties have been developed for the estimators and simulation studies were conducted to support the proposed methodology. Further, the bootstrap bandwidth selectors for making feasible inference are proposed.

We illustrate the performance of the new model using a COVID-19 data for patients requiring admission to the intensive care unit (ICU) during the first wave of the pandemic in Galicia. The database contains 2380 hospitalized COVID-19 patients reported by the Galician Healthcare Service between March 6 and May 7, 2020. The time of interest is the length of stay in hospital ward until admission to ICU, and the aim of this analysis was to estimate the probabilities of need for ICU (probability of experiencing the event) and time from admission on hospital ward until admission to the ICU of the patient in the ICU (latency function) given age as a covariate of interest. Among hospitalized patients, 197 (8.3%) patients required admission to ICU and 2183 (91.7%) patients were censored. In this censored group, 1638 (68.8%) patients were discharged alive before entering ICU, and 328 (13.8%) had died before entering ICU. Therefore, considered to be "cured" from the event of interest, which is admission to ICU. Note that "cure" means being free of experiencing admission to ICU, not cured in medical terms.

Figure 1 shows the estimated probability of requiring admission to ICU depending on the age, obtained using the proposed estimator $1 - \widehat{p}_h^c(x)$, Xu and Peng's estimator, both of them computed using the bootstrap bandwidth, the semiparametric estimator, and the MI-NW estimator computed using the improved cross-validation [7] bandwidth selector. Although the semiparametric estimator suggests a uniformly decreasing effect of the age on the probability of admission to the ICU, the other three estimators indicate that the logistic assumption for the cure probability might not be acceptable, as the curve patterns are characterized by a constant to a slightly increasing probability of admission to the ICU for younger patients (below 55 years), a sharp increase of the probability for middle age patients (from 55 to 69 years) and a decrease for elderly patients (70 years or older). Xu and Peng's estimator seems to overestimate the probability of ICU admission as it dismisses the significant information given by the the observed event indicator. Regarding the MI-NW estimator, the pattern of the estimated probability is consistent with that of the proposed estimator. However, it seems to underestimate the probability of admission to ICU for young-to-middle age patients. This is due to the low percentage of observed admissions to ICU in patients of those ages, resulting in an estimation with a high percentage of missing response.

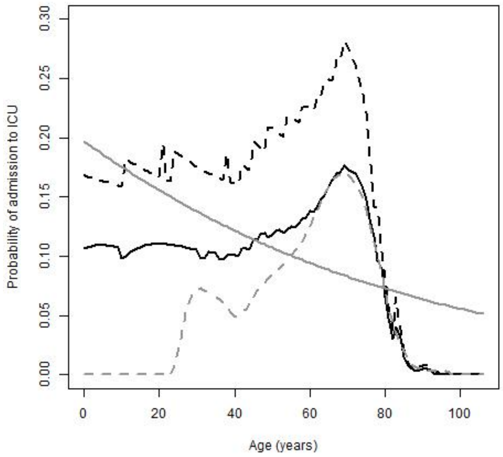

Figure 1: Estimation of the probability of admission to ICU for COVID-19 patients estimated using the proposed estimator $1 - \widehat{p}_h^c(x)$ (solid black line), Xu and Peng's estimator (dashed black line), both of them computed with the bootstrap bandwidth, the MI-NW estimator (dashed grey line) using the cross-validation bandwidths, and the semiparametric estimator (solid grey line).

References

[1] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

[2] Bernhardt, P. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine*, **25**, 4607 − 4623.

[3] Laska, E.M. and Meisner, M.J. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, **48(4)**, 1223 − 1234.

[4] López-Cheda, A. Jácome, M.A. and Cao, R. (2017). Nonparametric latency estimation for mixture cure models. *TEST*, **2**, 353 − 376.

[5] Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9(1)**, 141 − 142.

[6] Safari W.C. and López-de-Ullibarri I. and Jácome M. A. (2021). A product-limit estimator of the conditional survival function when cure status is partially known. *Biometrical Journal* `https://doi.org/10.1002/bimj.202000173`

[7] Tristen, H. and Jeffrey, R. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, **27(5)**, 1 − 32.

[8] Watson G.S. (1964). Smooth regression analysis. *The Indian Journal of Statistics, Series A,* **26(4)**, 359 − 372.

[9] Xu, J., and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, **42(1)**, 1 − 17.