ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Robustness of estimation methods in a survival cure model with mismeasured covariates

A. Bertrand [a,*], C. Legrand [a], D. Léonard [b], I. Van Keilegom [a]

[a] Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[b] Colorectal Surgery Unit, Department of Abdominal Surgery and Transplantation, Cliniques universitaires Saint-Luc, Brussels, Belgium

ARTICLE INFO

*Article history:*
Received 29 January 2016
Received in revised form 21 November 2016
Accepted 23 November 2016
Available online xxxx

*Keywords:*
Bias correction
Cure fraction
Measurement error
Promotion time cure model
Semiparametric method

ABSTRACT

In medical applications, time-to-event data is frequently encountered. While classical survival methods are well known and broadly used to analyze such data, they do not take into account two phenomena which appear quite often in practice: the presence of individuals who will never experience the event of interest (they are cured from this event) and of measurement error in the continuous covariates. Two approaches exist in the literature to estimate a model, taking these features into account. However, they require information about the distribution of the measurement error which is rarely fully known in practice. A theoretical study of bias motivates the need to take the measurement error into account. The conclusions of an extensive simulation study investigating the robustness of both correction approaches with respect to their assumptions then provide some practical recommendations for similar situations. Finally, the time until recurrence after surgery for rectal cancer patients is analyzed, taking into account the results from the simulations. Both correction methods were implemented in the R package `miCoPTCM`.



## 1. Introduction

Time-to-event data occur frequently in medicine. While death is, of course, an example of such an event, others might be of interest. For instance, in the study we consider here, the focus lies on the time until recurrence of rectal cancer for patients having previously received an operation. With such an event, it is known that some patients will never experience this recurrence, even if they are followed for an infinite time. These nonsusceptible subjects are considered as cured from, or immune to, the event of interest. This phenomenon appears regularly when studying non-lethal events in medicine, e.g. the contraction of the flu or the onset of age-related macular degeneration. For such data, classical survival analysis techniques are not suitable since they assume that, if the follow-up is sufficiently long, everyone will experience the event. This is why cure models appeared in the literature: they are specific models that take the presence of "cured" subjects in the population of interest into account.

In medical applications such as the one we consider here, we are often interested in the impact of patient or disease characteristics on the outcome. However, we generally tend to ignore the fact that such biological variables may be measured with error. This error can appear when the device or the method used to measure the quantity of interest is not precise; an example is the measurement of the maximal diameter of a tumor by an imaging technique. This feature is also present when

* Corresponding author.
*E-mail addresses:* aurelie.bertrand@uclouvain.be (A. Bertrand), catherine.legrand@uclouvain.be (C. Legrand), daniel.leonard@uclouvain.be (D. Léonard), ingrid.vankeilegom@uclouvain.be (I. Van Keilegom).

http://dx.doi.org/10.1016/j.csda.2016.11.013
0167-9473/

Please cite this article in press as: Bertrand, A., et al., Robustness of estimation methods in a survival cure model with mismeasured covariates. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.013

the quantity we want to measure fluctuates over time around its true value, as is the case, for example, with blood pressure. In the analysis of the time until recurrence in rectal cancer patients, one of the covariates of interest is the hemoglobin level in the blood, which is known to be measured with some error. The presence of mismeasured covariates in a statistical model can have several consequences (Carroll et al., 2006), including bias in the estimated effects of the covariates. This can lead to incorrectly concluding that a covariate has no significant effect on the response when it actually does (Cook and Stefanski, 1994).

In this paper, we address both features, cured individuals and mismeasured covariates, as they are often present together in medical problems, as illustrated by our example about rectal cancer, which we will study in this paper.

We assume that our data suffer from a problem which is very classical in survival analysis: right censoring. For some subjects, we are not able to observe the actual time point at which the event of interest occurs, since another event (called censoring) takes place before the event of interest. For such individuals, we only know that the actual event time is greater than the censoring time. Because of this right censoring, we are not able to distinguish between cured subjects (who are always censored) and non-cured subjects that are censored. This is why specific techniques are required to deal with such data.

More formally, with right-censored data, we observe $(Y_i, \Delta_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$ with $T_i$ the survival time and $C_i$ the right-censoring time. $\Delta_i = I(T_i \leq C_i)$ is the censoring indicator (taking the value 1 for uncensored subjects, for whom $T_i$ is observed, and the value 0 for censored subjects) and $\boldsymbol{X}_i$ is a $P$-dimensional vector of covariates. The vectors $(T_i, C_i, \boldsymbol{X}_i)$ are independent and identically distributed, with the same distribution as a generic vector $(T, C, \boldsymbol{X})$.

The promotion time cure model (Tsodikov, 1998) is one of the two survival cure models, which are models that take into account the existence of cured subjects. Compared to the mixture cure model (Taylor, 1995), the promotion time cure model assumes a proportional structure for the hazard, similarly to the classical (Cox, 1972) model. The conditional survival function of the whole population, giving the probability of surviving up to time $t$, i.e. $S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{X} = \boldsymbol{x})$, is modeled as

$$S(t|\boldsymbol{x}) = \exp\{-\theta(\boldsymbol{x})F(t)\}, \tag{1}$$

which is equivalent, for the conditional hazard function of $T$ given $\boldsymbol{X} = \boldsymbol{x}$, to

$$h(t|\boldsymbol{x}) = \theta(\boldsymbol{x})F'(t), \tag{2}$$

where $F$ is a proper baseline cumulative distribution function, $\theta$ is a known link function with an intercept, usually $\theta(\boldsymbol{x}) = \exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})$ for some $P$-dimensional vector of regression coefficients $\boldsymbol{\beta}$, and $\boldsymbol{x}$ is the vector of covariates. We work with the semiparametric version of this model, in which no assumptions are made on the distribution of $F$.

Zeng et al. (2006) and Ma and Yin (2008) propose two different methods (based on respectively profiling and backfitting the likelihood function to be maximized) to estimate the model parameters of the promotion time cure model when there is no measurement error in the covariates. These methods provide estimates for the regression parameters $\boldsymbol{\beta}$, as well as for the baseline cumulative distribution function: it can be shown that its nonparametric maximum likelihood estimator is a step function which increases only at the observed event times. We denote by $\widehat{p}_i$ the estimated jump size of $F$ at $Y_i$. For identifiability reasons, this estimated function is constrained to reach 1 at a predefined threshold (usually the largest observed event time), which amounts to considering as cured (for the estimation) the censored individuals with a censoring time larger than this threshold.

As motivated previously and studied in Bertrand et al. (in press), we suppose that (some of) the continuous covariates are not correctly measured. Assuming the classical additive model for the error, we observe

$$\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U}, \tag{3}$$

where $\boldsymbol{W}$ is the vector of observed covariates and $\boldsymbol{U}$ is the vector of measurement errors. We assume that $\boldsymbol{U}$ is independent of $\boldsymbol{X}$ and that it follows a continuous distribution with mean zero and known covariance matrix $\boldsymbol{V}$. It is also assumed that $(T, C)$ and $\boldsymbol{W}$ are independent given $\boldsymbol{X}$.

When (some of) the covariates are mismeasured, Zeng et al.'s (2006) technique and Ma and Yin's (2008) non-corrected approach yield biased estimators. This is a well-known fact which holds for many statistical models; however, the form of this bias in the context of a promotion time cure model has, to the best of our knowledge, never been presented. This paper is the first one to study the form of the bias in this context.

Ma and Yin (2008) were the first authors to address the problem of measurement error in the covariates of the promotion time cure model. Their approach requires a Gaussian error with a known variance, and a specific form for $\theta$, i.e. $\theta(\boldsymbol{x}) = \exp(\beta_0 + \boldsymbol{x}^T\boldsymbol{\beta})$. They proposed a corrected score strategy with a backfitting procedure which consists in solving (for $\boldsymbol{\beta}$, the $p_i$'s (the jump sizes of $F$) and a Lagrange multiplier $\lambda_{MY}$) the score equations of the model in which the terms containing $\boldsymbol{X}$ have been replaced by terms containing $\boldsymbol{W}$ and $\boldsymbol{V}$:

$$\frac{1}{p_{(i)}} = \sum_{j=1}^{n} I\left(Y_{(i)} \leq Y_j < \infty\right) \exp(\boldsymbol{W}_j^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{V}\boldsymbol{\beta}/2) + n\lambda_{MY}, \quad i = 1, \ldots, m,$$

$$\sum_{i=1}^{m} p_{(i)} = 1$$

Please cite this article in press as: Bertrand, A., et al., Robustness of estimation methods in a survival cure model with mismeasured covariates. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.11.013