Non-parametric estimation of proportion of cured patients and survival probabilities

Gisela Tunes da Silva e Antonio Carlos Pedroso de Lima

Departamento de Estatística, IME-USP

1 Introduction

In the usual survival analysis setup, it is assumed that all individuals are susceptible to the event of interest. However, there are some clinical trials in which a group of patients responds favorably to the treatment. Those patients are usually followed up for a long time and can be regarded as cured. Survival models that consider a cured fraction, known as cure rate models, are becoming more popular to deal with data from cancer clinical trials.

When a cure rate model is assumed, it is often of interest to estimate the proportion of cured patients. A simple non-parametric estimator for the proportion of cured patients was proposed by Maller & Zhou (1992). Parametric regression models have been proposed by Ghitany, Maller & Zhou (1994), Farewell (1982) and Gamel & Vogel (1997) and some authors have considered a semiparametric approach (Peng 2003). Also, in the context of cure rate estimation, the problem of sufficient follow-up arises naturally and it has been addressed by Maller & Zhou (1995) and Klebanova & Yakovlev (2007).

In this work, we consider the problem of non-parametric estimation of the proportion of cured patients and of the survival distribution of the uncured ones. We propose estimators based on the IPW (Inverse Probability Weighting) technique (Robins, Rotnitzky & Zhao 1995) and some properties are obtained.

2 Non-parametric estimator

Assume there are $n$ patients under study and that there is a proportion $\pi$ of non-cured patients (i.e., the probability of a cured patient is $1 - \pi$). Let $S_0(t)$ be the survival function for uncured patients. Let $T_i$ be the event time for the $i$th observation and assume that $T_i = \infty$ for the cured observations. The survival function of $T_i$ is given by

$$S(t) = P(T_i > t) = (1 - \pi) + \pi S_0(t).$$
Let $C_i$ the right censoring time for the $i$th individual and define the failure indicator by $\delta_i = I(T_i \leq C_i)$. We observe data in the form $(Z_i, \delta_i)$, where $Z_i = \min(T_i, C_i)$. Notice that, for the cured patients, we have $Z_i = C_i$ and $\delta_i = 0$. However, if an observation is censored, we do not know whether it belongs to the cured group or the uncured group with survival greater than the censoring time. If this information were available, data could be analyzed with methodologies for multi-state models (Andersen, Klein & Rosthoj 2003). We assume that the event time for the uncured patients is independent of the censoring time.

Let $N_i(t) = I(Z_i \leq t, \delta_i = 1)$ be the counting process associated with the failure time. For a fixed time $t^*$, define $P_* = E(N_i(t^*)) = \pi (1 - S_0(t^*))$.

We also denote by $G(\cdot)$ the survival function associated with the censoring time. The following relationship justifies the use of the weighted response:

$$E \left[ \frac{\delta_i N_i(t^*)}{G(Z_i)} \right] = E \left[ \frac{\delta_i N_i(t^*)}{G(Z_i)} \right] \left| Z_i \right] = E \left[ \frac{1}{G(Z_i)} E(\delta_i N_i(t^*)|Z_i) \right] = E \left[ \frac{1}{G(Z_i)} P(N_i(t^*) = 1, \delta_i = 1|Z_i) \right] = E \left[ \frac{1}{G(Z_i)} P(N_i(t^*) = 1|\delta_i = 1, Z_i) P(\delta_i = 1|Z_i) \right].$$

For uncensored observations, notice that $P(N_i(t^*) = 1|\delta_i = 1, Z_i)$ is either equal to zero (if $Z_i > t^*$) or one (if $Z_i \leq t^*$), i.e., $P(N_i(t^*) = 1|\delta_i = 1, Z_i) = N_i(t)$. Also, we have that $P(\delta_i = 1|Z_i) = P(C_i > Z_i|Z_i) = G(Z_i)$, so that

$$E \left[ \frac{\delta_i N_i(t^*)}{G(Z_i)} \right] = E \left[ \frac{1}{G(Z_i)} P(N_i(t^*) = 1|\delta_i = 1, Z_i) P(\delta_i = 1|Z_i) \right] = E \left[ \frac{1}{G(Z_i)} N_i(t^*) G(Z_i) \right] = E \left[ N_i(t^*) \right] = \pi (1 - S_0(t^*)) = P_*.$$

We first derive estimators for the proportion of cured patients and survival function of the event time distribution for the uncured patients without the inclusion of covariates. In
our method, it is required to estimate the distribution for the censoring times. Following Scheike, Zhang & Gerds (2008), we propose the use of the Kaplan-Meier estimator, which is consistent if the censoring distribution does not depend on covariates.

For the estimation of the probability $\pi$, we let $t^* \to \infty$ in equation (1) so that

$$E \left[ \frac{\delta_i N_i(\infty)}{G(Z_i)} \right] = E \left[ \frac{\delta_i}{G(Z_i)} \right] = \pi.$$ 

The score function is then given by

$$U(\pi) = \sum_{i=1}^{n} \left( \frac{\delta_i}{\hat{G}(Z_i)} - \pi \right).$$

Solving the equation $U(\pi) = 0$, we can estimate $\pi$ by

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(Z_i)}$$

(2)

It can be shown that the estimator given by (2) is equal to the estimator proposed by Maller & Zhou (1992), so that our proposed method can be viewed as an extension of their work.

In addition to the estimation of the cured proportion, we may be interested to estimate the probability of a patient that is not in the cured group be alive at some fixed time, say $t^*$. Also, in practical applications, the researcher may be interested in the probability of a patient being alive at $t^*$, given by $(1 - \pi) + \pi S_0(t^*)$. We note that $\pi S_0(t^*)$ can be interpreted as the probability of a patient being alive at $t^*$ and eventually die, since he/she is not in the cured group.

If one wants to estimate $S_0(t^*)$, using (1), the score function is then given by

$$U(\pi, S_0(t^*)) = \sum_{i=1}^{n} D_{\eta^*}^{T} \left( \frac{\delta_i}{\hat{G}(Z_i)} - \frac{\delta_i N_i(t^*)}{\hat{G}(Z_i)} - P_s \right)$$

$$= \sum_{i=1}^{n} D_{\eta^*}^{T} \left( \frac{\delta_i N_i(t^*)}{\hat{G}(Z_i)} - \pi \left( 1 - S_0(t^*) \right) \right),$$

where $D_{\eta^*} = \frac{\partial \mu}{\partial \eta^*}$, $\mu = (\pi, \pi(1 - S_0(t^*)))^T$ and $\eta^*$ is the vector of unknown parameters to be estimated, i.e., $\eta^* = (\pi, S_0(t^*))^T$ in this particular situation. Since

$$D_{\eta^*} = \begin{pmatrix} 1 & 0 \\ 1 - S_0(t^*) & -\pi \end{pmatrix},$$
the score equation is then given by

$$U(\pi, S_0(t^*)) = \sum_{i=1}^{n} \left( \left( \frac{\delta_i}{G(Z_i)} - \pi \right) + (1 - S_0(t^*)) \left( \frac{\delta_i N_i(t^*)}{G(Z_i)} - \pi (1 - S_0(t^*)) \right) \right).$$  \quad (3)$$

The solution for the equation $U(\pi, S_0(t^*)) = 0$ leads to an estimator for $\pi$ given by (2) and

$$\hat{S}_0(t^*) = 1 - \sum_{i=1}^{n} \frac{\delta_i N_i(t^*)}{\delta_i G(Z_i)}.$$  \quad (4)$$

It can be easily seen that if it is needed to estimate $S(t)$ for several time points simultaneously, one gets the same estimator given by (4). Therefore, it is possible to construct a plot of the estimated $S(t)$. It is a step function with jumps at the time points when a failure occurred.

### 2.1 Properties of the estimator

We briefly describe some properties of our proposed estimators. The properties can be established following Scheike et al. (2008). Assume initially the problem of non-parametric estimation of the proportion of cured patients. It can be shown that our proposed estimator $\hat{\pi}$ is consistent and asymptotically normal with variance

$$\text{var}(\hat{\pi}) = \frac{1}{n^2} \sum_{i=1}^{n} U_i^2 = \frac{1}{n^2} \sum_{i=1}^{n} \left( \frac{\delta_i}{G(Z_i)} - \pi \right)^2.$$  \quad (5)$$

Assume now the problem of jointly estimating the proportion of cured patients and the survival probability at a fixed time $t^*$ for the uncured patients. From Scheike et al. (2008), it can be shown that our proposed estimator is consistent and also asymptotically normal with variance given by

$$\text{Var}(\hat{\pi}, S_0(t^*)) = \frac{1}{n^2} \sum_{i=1}^{n} U_i^T \left( \sum_{i=1}^{n} U_i U_i^T \right) \mathcal{I}_{n^*}^{-1},$$

where

$$U_i = \left( \left( \frac{\delta_i}{G(Z_i)} - \pi \right) + (1 - S_0(t^*)) \left( \frac{\delta_i N_i(t_c)}{G(Z_i)} - \pi (1 - S_0(t^*)) \right) \right).$$ \quad (6)$$
and
\[
I_{n^*} = n \left( \frac{1 + (1 - S_0(t^*))^2}{-\pi (1 - S_0(t^*))^{\frac{1}{2}}} - \pi (1 - S_0(t^*)) \right).
\]

Similar expressions are obtained when estimating simultaneously the proportion of cured patients and survival probabilities at several fixed time points.

3 Discussion

We have proposed in this work a consistent estimator for the proportion of cured patients and survival probabilities based on the IPW technique. Our method provided the same estimator for the proportion of cured patients proposed by Maller & Zhou (1992). In addition to the estimation of cured proportion, our method has the advantage of providing non-parametric estimator of the survival distribution for uncured patients. The graphical representation of the survival distribution of uncured patients may be useful for descriptive purposes in some studies when there is evidence of cured patients on the study. The results were applied in a real data analysis about anemia in blood donors, where the main interest is to study the time until anemia for frequent blood donors.

Our proposed method can be extended to allow for covariates. We can include covariates effects either in the proportion of cured patients and in survival probabilities for uncured patients for fixed time points and this problem is currently being addressed.

Referências


