# A Parametric Cure Model with Covariates

Ana M. Abreu and Cristina S. Rocha



## 1 Introduction

Survival analysis is strongly stimulated by the constant evolution of medicine. In particular, new models were developed to take into account the possibility of cure of certain diseases. It is in this context that cure models appear, because they allow the analysis of survival data in which some subjects can eventually experience, and others never experience, the event of interest. An important property of cure models (mixture and non-mixture) is the fact that they have an improper survival function, which is equivalent to the cumulative hazard function being limited.

Although, frequently, the cure is not observable, the suspicion is based in some features of the data, namely the existence of many censored observations beyond the last observed survival time. Therefore, a long and stable plateau of the Kaplan–Meier survival curve [5] suggests the applicability of the mixture cure model approach [8].

Usually, in a cure model, we want to estimate the proportion of cured individuals, the survival function of the susceptible individuals and the effect of the covariates, if they have been included in the model. There are several ways of modelling the effect of the covariates, $\mathbf{x}$, on the survival of the susceptible individuals for instance, the accelerated failure time model, that is, $S_d(t|\mathbf{x}) = S_{d_0}(te^{\beta'\mathbf{x}})$, where $S_{d_0}(.)$ is independent of the covariates and can be formulated either parametrically [9] or non-parametrically [7]. Another possibility is the proportional odds model, which is used when the hazard functions of individuals with different values of their

AQ1 A.M. Abreu (✉)
AQ2 CCCEE and CCM, University of Madeira, Praca do Municpio 9000-082 Funchal, Portugal
e-mail: abreu@uma.pt

C.S. Rocha
DEIO and CEAUL, Faculty of Science, University of Lisbon, Lisbon, Portugal
e-mail: cmrocha@fc.ul.pt

J.L. da Silva et al. (eds.), *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, Studies in Theoretical and Applied Statistics, DOI 10.1007/978-3-642-34904-1_3,

covariates converge after some time. The most widely used model is undoubtedly the proportional hazards model $S_d(t|\mathbf{x}) = S_{d_0}(t)^{\exp(\beta'\mathbf{x})}$ where, usually, $S_{d_0}(t)$ is non-parametric [10]. Another alternative is to consider a mixture cure model with more than one survival function for susceptible individuals [4]. The logistic regression model is the most common choice to model the effects of the covariates, $\mathbf{z}$, in the cure proportion.

In this chapter, we propose a new mixture cure model with covariates based on the Chen distribution [2]. Section 2 describes the general structure of this model, while in Sect. 3 some parameter estimation details are presented. In Sect. 4 the applicability of our model is illustrated with the analysis of leukaemia data and Sect. 5 is reserved to concluding remarks.

## 2 A Cure Model with Covariates

In this section we describe the structure of the mixture cure model some features of the Chen distribution and present our new model.

### 2.1 The Mixture Cure Model

We denote by $T$ the random variable that represents the survival time in a population where there are susceptible and non-susceptible individuals. Let $Y$ denote a binary random variable indicating that an individual is either susceptible ($Y = 1$) or not ($Y = 0$). The mixture cure model can be formulated through the survival function

$$S(t) = p + (1 - p)S_d(t), \tag{1}$$

where $p = P(Y = 0)$ represents the non-susceptible proportion and $S_d(t) = S(t|Y = 1)$ is the (proper) survival function of the susceptible individuals. As $S(t) \to p$ when $t \to \infty$, then $S(t)$ is an improper survival function. Note that, if an individual has censored survival time, then $Y$ is not observable, so we do not know if that individual is susceptible or not.

If we introduce covariates in model (1), we have

$$S(t_i|\mathbf{x}_i, \mathbf{z}_i) = p(\mathbf{z}_i) + (1 - p(\mathbf{z}_i))S_d(t_i|\mathbf{x}_i), \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the vectors of covariates associated to the $i$th individual ($i = 1, \ldots, n$), $p(\mathbf{z}_i) = P(Y = 0|\mathbf{z}_i)$ is the probability that the $i$th individual is non-susceptible given a covariate vector $\mathbf{z}_i$ and $S_d(t_i|\mathbf{x}_i) = P(T_i > t_i|Y_i = 1, \mathbf{x}_i)$ is the probability that an individual survives longer than $t_i$, given that the individual is susceptible and has a covariate vector $\mathbf{x}_i$. Note that $\mathbf{x}_i$ and $\mathbf{z}_i$ can include the same covariates.

## 2.2 The Chen Distribution

The distribution function proposed by Chen [2] is

$$F(t) = 1 - \exp[\lambda_1(1 - \exp(t^{\lambda_2}))], \quad t > 0, \quad \lambda_1, \lambda_2 > 0, \tag{3}$$

where $\lambda_1$ is the scale parameter and $\lambda_2$ is the shape parameter. The corresponding survival and hazard functions are, respectively,

$$\overline{F}(t) = \exp[\lambda_1(1 - \exp(t^{\lambda_2}))], \quad t > 0, \tag{4}$$

$$h^*(t) = \lambda_1\lambda_2 t^{\lambda_2 - 1} \exp(t^{\lambda_2}), \quad t > 0.$$

The author refers that $h^*(t)$ can be bathtub-shaped when $\lambda_2 < 1$ and that it increases when $\lambda_2 \geq 1$, which is unusual in most distributions used in survival analysis. In fact, as

$$h^{*'}(t) = \lambda_1\lambda_2 t^{\lambda_2 - 2} \exp(t^{\lambda_2})((\lambda_2 - 1) + \lambda_2 t^{\lambda_2}),$$

for $\lambda_2 < 1$ we have $h^*(t)$ decreasing for $t \in [0, (\frac{1}{\lambda_2} - 1)^{\frac{1}{\lambda_2}}]$ and, for $t \geq (\frac{1}{\lambda_2} - 1)^{\frac{1}{\lambda_2}}$, $h^*(t)$ is an increasing function. Hence, the range of the interval where $h^*(t)$ is decreasing will increase as $\lambda_2$ decreases. Therefore, if $\lambda_2$ is near zero, for example, $\lambda_2 = 0.1$, the interval is so large that, from the practical point of view, it is just like having a decreasing hazard function. Reciprocally, as $\lambda_2$ approaches 1, the interval where the hazard function is decreasing is so small that it is almost like if the hazard function was always increasing.

## 2.3 The Cure Model Based on the Chen Distribution with Covariates

Admit that the survival time of susceptible individuals follows the Chen distribution, given by Eq. (3). As stated by Abreu and Rocha [1], the cure model obtained by substituting in Eq. (1) $S_d(t)$ by the expression (4) is

$$S(t) = p + (1 - p)\exp[\lambda_1(1 - \exp(t^{\lambda_2}))], \quad t > 0, \ \lambda_1, \lambda_2 > 0. \tag{5}$$

If the model is defined in terms of hazard function, we have

$$h(t) = \frac{(1 - p)\lambda_1\lambda_2 t^{\lambda_2 - 1} \exp(t^{\lambda_2}) \exp[\lambda_1(1 - \exp(t^{\lambda_2}))]}{p + (1 - p)\exp[\lambda_1(1 - \exp(t^{\lambda_2}))]}.$$

Consider the proportional hazards model for the survival time of susceptible individuals. Then we have

$$S_d(t|\mathbf{x}) = S_d(t|\boldsymbol{\beta}'\boldsymbol{x}, \lambda_1, \lambda_2) = S_{d_0}(t|\lambda_1, \lambda_2)^{\exp(\boldsymbol{\beta}'\boldsymbol{x})},$$

where $\lambda_1$ and $\lambda_2$ are the parameters of the Chen distribution corresponding to the baseline survival function, that is,

$$S_d(t|\mathbf{x}) = [\exp[\lambda_1(1 - \exp(t_i^{\lambda_2}))]]^{\exp(\boldsymbol{\beta}'\boldsymbol{x})}. \quad (6)$$

Let

$$p(\mathbf{z}) = P(Y = 0|\mathbf{z}) = \frac{1}{1 + \exp(\boldsymbol{\gamma}'z)} \quad (7)$$

be the function that models the effect of the covariates in the proportion of non-susceptible individuals. In fact, in this context, the logistic regression model is the most commonly used binary regression model.

The mixture cure model of proportional hazards specified by Eqs. (2), (6) and (7) can be written in the form

$$S(t|\mathbf{x}, \mathbf{z}) = \frac{1}{1 + \exp(\boldsymbol{\gamma}'z)} + \frac{\exp(\boldsymbol{\gamma}'z)}{1 + \exp(\boldsymbol{\gamma}'z)}[\exp[\lambda_1(1 - \exp(t^{\lambda_2}))]]^{\exp(\boldsymbol{\beta}'\boldsymbol{x})}. \quad (8)$$

## 3 Parameters Estimation

In this section, the parameters estimation process for the proposed model is presented. With this purpose, we apply the maximum likelihood method, making use of the EM algorithm [3], since here we are dealing with missing data.

### 3.1 Maximum Likelihood Function

Let us assume that censoring is noninformative. Denote the observed survival time for the $i$th individual by $t_i$, $i = 1, \ldots, n$. Suppose we have data in the form $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, where $\delta_i = 1$ if $t_i$ is uncensored and $\delta_i = 0$ otherwise, and $\mathbf{x}_i$ and $\mathbf{z}_i$ are two covariate vectors. Without loss of generality, suppose that the first $m$ ($m < n$) survival times are censored. Then $\delta_i = 0$ if $1 \leq i \leq m$ and $\delta_i = 1$ if $m + 1 \leq i \leq n$.

The contribution to the likelihood of an individual for whom the event of interest was observed at $t_i$ is $(1 - p(\mathbf{z}_i))\, f_d(t_i|\mathbf{x}_i)$, where $f_d(t_i|\mathbf{x}_i)$ represents the density function of the susceptible individuals, conditional on the corresponding covariates. If the event of interest is not observed until time $t_i$, then the contribution of the individual to the likelihood is $p(\mathbf{z}_i) + (1 - p(\mathbf{z}_i))S_d(t_i|\mathbf{x}_i)$.

Then, the observed likelihood function is

$$L_O = \prod_{i=1}^{n} \left\{\left[1 - p(\mathbf{z}_i)\right] f_d(t_i|\mathbf{x}_i)\right\}^{\delta_i} \left\{p(\mathbf{z}_i) + \left[1 - p(\mathbf{z}_i)\right] S_d(t_i|\mathbf{x}_i)\right\}^{1-\delta_i},$$

which can be written as

$$L_O = \prod_{i=1}^{n} \Big\{[1-p(\mathbf{z}_i)]\lambda_1\lambda_2 t_i^{\lambda_2-1}\exp(t_i^{\lambda_2}+\boldsymbol{\beta}'\boldsymbol{x}_i)\big\{\exp[\lambda_1(1-\exp(t_i^{\lambda_2}))]\big\}^{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\Big\}^{\delta_i}$$
$$\times\Big\{\big\{\exp[\lambda_1(1-\exp(t_i^{\lambda_2}))]\big\}^{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}\Big\}^{1-\delta_i}$$

when the Chen distribution is used for the survival time of susceptible individuals.

Let $y_1,\ldots,y_n$ be such that $y_i = 1$ if the individual is susceptible and $y_i = 0$ otherwise. If all $y_i'$s were observed, the complete likelihood would be

$$L_C = \prod_{i=1}^{n}\Big\{\big[(1-p(\mathbf{z}_i))\,f_d(t_i|\mathbf{x}_i)\big]^{y_i}\Big\}^{\delta_i}\Big\{p(\mathbf{z}_i)^{1-y_i}\big[(1-p(\mathbf{z}_i))S_d(t_i|\mathbf{x}_i)\big]^{y_i}\Big\}^{1-\delta_i}. \tag{13}$$

Considering $q(\mathbf{z}_i) = 1-p(\mathbf{z}_i)$, after some calculations the previous expression can be rewritten as

$$L_C = \prod_{i=1}^{n} q(\mathbf{z}_i)^{y_i}[1-q(\mathbf{z}_i)]^{1-y_i}\prod_{i=1}^{n} h_d(t_i|\mathbf{x}_i)^{y_i\delta_i}S_d(t_i|\mathbf{x}_i)^{y_i}. \tag{9}$$

The logarithm of Eq. (9) is given by

$$\begin{aligned}\log L_C = {} & \textstyle\sum_{i=1}^{n}[y_i\log q(\mathbf{z}_i) + (1-y_i)\log(1-q(\mathbf{z}_i)) + \\ & \textstyle\sum_{i=1}^{n} y_i\delta_i\log h_d(t_i|(\mathbf{x}_i)) + y_i\log S_d(t_i|(\mathbf{x}_i))].\end{aligned} \tag{10}$$

## 3.2 EM Algorithm

The fact that in most cases cure is not observable, gives origin to an incomplete data situation. In this context, the EM algorithm is a widely used tool for maximizing the likelihood function. In general terms, the maximization of the likelihood is replaced by maximizing its expectation conditional to the current parameter values and the observed data. Thus, the missing values are identified with the corresponding conditional expected value.

In fact, the E step of the EM algorithm consists in obtaining the expectation of the logarithm of the complete likelihood with respect to the distribution of the unobserved $Y_i$'s, given the current parameter values and the observed data $\mathscr{O}$, where $\mathscr{O} = \{\text{observed } y_i'\text{s}, (t_i,\delta_i,\mathbf{x}_i,\mathbf{z}_i), i = 1,\ldots,n\}$. However, since $\log L_C$ is linear in $Y_i$, to compute the expected value of $\log L_C$, we only need to replace in Eq. (10) each unobserved $Y_i$ by its expected value, denoted by $\tau_i$. Therefore, we have

$$\tau_i = E(Y_i|\mathscr{O}) = P(Y_i = 1|T_i > t_i,\delta_i = 0,\boldsymbol{\theta}) = \frac{[1-p(\mathbf{z}_i)]S_d(t_i|\mathbf{x}_i)}{S(t_i|\mathbf{x}_i,\mathbf{z}_i)} \tag{11}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is the vector parameter of model (8) and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. Thus, in the logarithm of the complete likelihood, each $y_i$ is replaced by $\omega_i$, the probability of the $i$th individual being susceptible, where $\omega_i = 1$ if $\delta_i = 1$ and $\omega_i = \tau_i$ if $\delta_i = 0$.

At the M step, we need to maximize the following two components of the expected log-likelihood:

$$\begin{aligned}\log L_{E_1} &= \textstyle\sum_{i=1}^{n}[\omega_i \log q(\mathbf{z}_i) + (1-\omega_i)\log(1-q(\mathbf{z}_i))] \\ &= (n-m)\log q(\mathbf{z}_i) + m\log(1-q(\mathbf{z}_i)) + \textstyle\sum_{i=1}^{m}\tau_i[\log q(\mathbf{z}_i) - \log(1-q(\mathbf{z}_i))],\end{aligned}$$

$$\begin{aligned}\log L_{E_2} &= \textstyle\sum_{i=1}^{n}[\delta_i\omega_i \log h_d(t_i|\mathbf{x}_i) + \omega_i \log S_d(t_i|\mathbf{x}_i)] \\ &= \textstyle\sum_{i=1}^{m}\tau_i \log S_d(t_i|\mathbf{x}_i) + \sum_{i=m+1}^{n}[\log h_d(t_i|\mathbf{x}_i) + \log S_d(t_i|\mathbf{x}_i)].\end{aligned}$$

From $\log L_{E_1}$, after some algebra, we obtain the following explicit expression for the estimate of $q(\mathbf{z}_i)$ at the $(k+1)$th iteration:

$$q(\mathbf{z}_i)^{(k+1)} = \frac{1}{n}\Big[(n-m) + \sum_{i=1}^{m}\tau_i^{(k)}\Big],$$

but only in the case where the covariates are not included in the cure proportion. Making use of the Chen distribution for the survival time of the susceptible individuals, by Eq. (11), we get

$$\tau_i = \frac{q(\mathbf{z}_i)\big\{\exp[\lambda_1(1-\exp(t_i^{\lambda_2}))]\big\}^{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}}{1-q(\mathbf{z}_i)+q(\mathbf{z}_i)\big\{\exp[\lambda_1(1-\exp(t_i^{\lambda_2}))]\big\}^{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}}. \tag{12}$$

In what concerns $\log L_{E_2}$, since it can be written as

$$\begin{aligned}\log L_{E_2} = {}& \lambda_1 \textstyle\sum_{i=1}^{m}\tau_i\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)[1-\exp(t_i^{\lambda_2})] + (n-m)(\log\lambda_1 + \log\lambda_2) + \\ & (\lambda_2 - 1)\textstyle\sum_{i=m+1}^{n}\log t_i + \sum_{i=m+1}^{n}(\exp(\boldsymbol{\beta}'\boldsymbol{x}_i) + t_i^{\lambda_2}) + \\ & \lambda_1 \textstyle\sum_{i=m+1}^{n}\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)[1-\exp(t_i^{\lambda_2})],\end{aligned}$$

after some algebra, we obtain an explicit formula for the estimator of $\lambda_1$,

$$\hat{\lambda}_1 = \frac{n-m}{\sum_{i=1}^{m}\tau_i\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)\big[\exp(t_i^{\lambda_2})-1\big] + \sum_{i=m+1}^{n}\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)\big[\exp(t_i^{\lambda_2})-1\big]},$$

where $\tau_i$ is given by Eq. (13). No explicit formula was obtained for the estimator of $\lambda_2$. Therefore, we recommend using simultaneously another maximization procedure, such as the Newton–Raphson method.

## 4 Application to Leukaemia Data

Kersey et al. [6] reported data on patients with refractory acute lymphoblastic leukaemia. Patients receive either an allogeneic transplant (group 1) or an autologous transplant (group 2) and are followed until a recurrence occurs.

If we fit model (5) for each group separately, the estimated survival functions are

$$\hat{S}_1(t) = 0.2714 + 0.7286 \times \exp(0.76112 \times (1 - \exp(t^{0.61397})))$$

for group 1 and

$$\hat{S}_2(t) = 0.1799 + 0.8201 \times \exp(1.15842 \times (1 - \exp(t^{0.6853})))$$

for group 2. We can consider the data from the two groups jointly and fit the same model. The result is

$$\hat{S}(t) = 0.22739 + 0.77261 \times \exp(0.92261 \times (1 - \exp(t^{0.63706}))).$$

For the moment, we restrict our analysis to the case of one binary covariate. So, defining a covariate, $x$, as the indicator of the patients group, we obtain

$$\hat{S}(t|x) = 0.22821 + 0.77179 \times (\exp(1.15379 \times (1 - \exp(t^{0.65037}))))^{\exp(-0.42x)}. \quad (13)$$

This covariate had no significant effect on the non-susceptible proportion, something expected given the proximity of the values in the two previous models. Note that the survival time of the susceptible individuals follows a Chen distribution with parameters $\lambda_1$ and $\lambda_2$ when $x = 0$ and with parameters $\lambda_1 \times e^{\beta}$ and $\lambda_2$ when $x = 1$. Due to difficulties in the implementation of the EM algorithm, namely convergence problems, the estimate of $\beta$ was obtained making use of this characteristic.

## 5 Concluding Remarks

The aim of this article is to increase the options for survival distributions when the use of cure models is relevant. The Chen distribution is very versatile, resulting in a good fit in many cases where other parametric models were unsatisfactory. We introduced covariates in the model in order to make it more suitable for practical situations. So far, some issues in the estimation process are not completely solved. Nevertheless, we obtained significant correlation coefficients (r=0.9946, p=0.000

for group 1 and r=0.9512, p=0.000 for group 2) between the Kaplan–Meier estimates and the fitted values obtained using model (13), indicating a good fit for both groups.

**Acknowledgements** Ana Abreu's research was supported by FCT, POCTI-219, FEDER, and Cristina Rocha's research is partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal—FCT under the project PEst-OE/MAT/UI0006/2011. We would like to thank the reviewers for their thorough and insightful review of the manuscript.

## References

1. Abreu, A.M., Rocha, C.S.: Um novo modelo de cura paramétrico. In: Castro, L.C., Martins, E.G., Rocha, C., Oliveira, M.F., Leal, M.M., Rosado, F. (eds.) Ciência Estatística, pp. 151–162. Edições SPE, Lisboa (2006)
2. Chen, Z. : A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. Stat. Probab. Lett. **49**, 155–161 (2000)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39**, 1–38 (1977)
4. Hunsberger, S., Albert, P.S., London, W.B.: A finite mixture survival model to characterize risk groups of neuroblastoma. Stat. Med. **28**, 1301–1314 (2009)
5. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **57**, 457–481 (1958)
6. Kersey, J.H., Weisdorf, D., Nesbit, M.E., LeBien, T.W., Woods, W.G., McGlave, P.B., Kim, T., Vallera, D.A., Goldman, A.I., Bostrom, B., Hurd, D., Ramsay, N.K.C.: Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractary acute lymphoblastic leukaemia. N. Engl. J. Med. **317**, 461–467 (1987)
7. Li, C–S., Taylor, J.M.G.: A semi-parametric accelerated failure time cure model. Stat. Med. **21**, 3235–3247 (2002)
8. Maller, R.A., Zhou, S.: Survival Analysis with Long-Term Survivors. Wiley, New York (1996)
9. Peng, Y., Dear, K.B.G., Denham, J.W.: A generalized F mixture model for cure rate estimation. Stat. Med. **17**, 813–830 (1998)
10. Sy, J.P., Taylor, J.M.G.: Estimation in a Cox proportional hazards cure model. Biometrics **56**, 227–236 (2000)

AUTHOR QUERIES

AQ1. First author has been treated as corresponding author. Please check.
AQ2. Please check whether the inserted affiliation details are correct.