VERSITA

TATRA MOUNTAINS
Mathematical Publications

DOI: 10.2478/v10127-012-0001-4
Tatra Mt. Math. Publ. **51** (2012), 1–9

CURE FRACTION MODELS USING MIXTURE AND NON-MIXTURE MODELS

Jorge A. Achcar — Emílio A. Coelho-Barros –
– Josmar Mazucheli

ABSTRACT. We introduce the Weibull distributions in presence of cure fraction, censored data and covariates. Two models are explored in this paper: mixture and non-mixture models. Inferences for the proposed models are obtained under the Bayesian approach, using standard MCMC (Markov Chain Monte Carlo) methods. An illustration of the proposed methodology is given considering a lifetime data set.

1. Introduction

The mixture cure rate model, also known as standard cure rate model, assumes that the studied population is a mixture of susceptible individuals, who experience the event of interest and non-susceptible individuals that will never experience it. These individuals are not at risk with respect to the event of interest and are considered immune, non-susceptible, or cured, [5]. Different approaches, parametric and non-parametric, have been considered to model cure fraction.

Following Maller and Zhou (1996), let us assume that the population is divided in two groups of individuals: a group of cured individuals with probability π and a group of susceptible individuals with a proper survival function $S_0(t) = P(T > t)$ where T denotes the lifetime of the individual with probability $1 - \pi$.

In this mixture model, we have

$$S(t) = \pi + (1 - \pi) S_0(t), \tag{1}$$

where $\pi \in (0, 1)$ and $S_0(t)$ is the survival function for the susceptible individuals (not-censored observations). Let us denote the mixture model (1) as Model 1.

2010 Mathematics Subject Classification: Primary 62J05; Secondary 62F03, 62F10, 62F30.
Keywords: cure fraction, mixture model, non-mixture model, Weibull distribution, lifetime data.

Considering a random sample (t_i, δ_i), $i = 1, \ldots, n$, the contribution of the ith individual for the likelihood function is given by

$$L_i = \left[f\left(t_i\right)\right]^{\delta_i}\left[S\left(t_i\right)\right]^{1-\delta_i}, \tag{2}$$

where δ_i is a censoring indicator variable, that is, $\delta_i = 1$ for an observed lifetime and $\delta_i = 0$ for a censored lifetime.

Assuming the "mixture model" (1) the probability density function (p.d.f.) for the lifetime T is (from $f\left(t\right) = dF\left(t\right)/dt$) given by

$$f\left(t\right) = \left(1-\pi\right) f_0\left(t\right), \tag{3}$$

where $F\left(t\right) = 1 - S\left(t\right)$ and $f_0\left(t\right)$ is the probability density function for the susceptible individuals. Substitution of the mixture density and survival function in the standard likelihood function yields the likelihood for the mixture cure model

$$L_i = \left(1-\pi\right)^{\delta_i}\left[f_0\left(t_i\right)\right]^{\delta_i}\left[\pi + \left(1-\pi\right) S_0\left(t_i\right)\right]^{1-\delta_i}. \tag{4}$$

Therefore the log-likelihood is given by

$$\begin{aligned} \ln L = {} & r \ln\left(1-\pi\right) + \sum_{i=1}^{n} \delta_i \ln f_0\left(t_i\right) \\ & + \sum_{i=1}^{n} \left(1-\delta_i\right) \ln\left[\pi + \left(1-\pi\right) S_0\left(t_i\right)\right], \end{aligned} \tag{5}$$

where $r = \sum_{i=1}^{n} \delta_i$ is the number of uncensored observations. Common parametrically choices for $S_0\left(t\right)$ are the exponential and Weibull distributions [9].

An alternative non-mixture formulation has been suggested which defines an asymptote for the cumulative hazard and hence for the cure fraction, [4], [8], [10]. In this case, the survival function for non-mixture cure fraction model is

$$S\left(t\right) = \pi^{F_0(t)}, \tag{6}$$

where $0 < \pi < 1$ is the probability of cured individuals and $F_0\left(t\right) = 1 - S_0\left(t\right)$ is the distribution function for the susceptible individuals. Let us denote the model (6) as Model 2.

From (6), the survival and hazard function for the non-mixture cure rate model can be written, respectively, as

$$\begin{aligned} S\left(t\right) &= \exp\left[F_0\left(t\right) \ln \pi\right], \\ h\left(t\right) &= -\left(\ln \pi\right) f_0\left(t\right). \end{aligned} \tag{7}$$

Since $h(t) = f(t)/S(t)$, the contribution of the ith individual for the likelihood function (see (2)) is given by

$$L_i = \left[h(t_i)\right]^{\delta_i} S(t_i), \tag{8}$$

that is,

$$L_i = \left[-(\ln \pi) f_0(t_i)\right]^{\delta_i} \exp\left[F_0(t_i) \ln \pi\right]. \tag{9}$$

Assuming a random sample of size n, the log-likelihood function is given by

$$\begin{aligned} \ln L = r \ln(-\ln \pi) + \sum_{i=1}^{n} \delta_i \ln f_0(t_i) \\ + (\ln \pi) \sum_{i=1}^{n} \left[1 - S_0(t_i)\right], \end{aligned} \tag{10}$$

where, as before, $r = \sum_{i=1}^{n} \delta_i$.

2. Weibull distribution for the susceptible individuals

As a special case, let us assume a Weibull distribution for the susceptible individuals with probability density function

$$f_0(t) = \gamma \lambda t^{\gamma - 1} \exp\left[-\lambda t^{\gamma}\right], \tag{11}$$

and survival function $S_0(t) = \exp\left[-\lambda t^{\gamma}\right]$.

Assuming the mixture model (1), the log-likelihood function for π, λ and γ (see (5)) is given by

$$\begin{aligned} l(\pi, \lambda, \gamma) = r \ln(1 - \pi) + r \ln \gamma + r \ln \lambda \\ + (\gamma - 1) v - \lambda A_1(\gamma) + A_2(\pi, \lambda, \gamma), \end{aligned} \tag{12}$$

where

$$r = \sum_{i=1}^{n} \delta_i, \qquad v = \sum_{i=1}^{n} \delta_i \ln t_i,$$

$$A_1(\gamma) = \sum_{i=1}^{n} \delta_i t_i^{\gamma}$$

and

$$A_2(\pi, \lambda, \gamma) = \sum_{i=1}^{n} (1 - \delta_i) \ln\left[\pi + (1 - \pi) e^{-\lambda t_i^{\gamma}}\right].$$

Assuming the non-mixture model (6), the log-likelihood function for π, λ and γ (see (10)) is given by

$$\begin{aligned}\ln L = & r \ln(-\ln \pi) + r \ln \gamma + r \ln \lambda \\ & + (\gamma - 1)\, v - \lambda A_1(\gamma) + (\ln \pi)\, A_3(\lambda, \gamma), \end{aligned} \tag{13}$$

where

$$r = \sum_{i=1}^{n} \delta_i, \qquad v = \sum_{i=1}^{n} \delta_i \ln t_i,$$

$$A_1(\gamma) = \sum_{i=1}^{n} \delta_i t_i^{\gamma} \quad \text{and} \quad A_3(\lambda, \gamma) = \sum_{i=1}^{n} \left[1 - e^{-\lambda t_i^{\gamma}}\right].$$

3. A Bayesian analysis

For a Bayesian analysis of the mixture and non-mixture models introduced in Section 1, we assume a uniform $U(0,1)$ prior distribution for the probability of cure π and $Gamma(0.001, 0.001)$ prior distribution for the scale parameter λ and shape parameter γ, where $Gamma(a,b)$ denotes a gamma distribution with mean a/b and variance a/b^2. We further assume prior independence among π, λ and γ. Observe that we are using approximately non-informative priors for the parameters models.

In the presence of a covariate vector $\mathbf{x} = (x_1, \ldots, x_k)'$ affecting the parameters π and λ, but not affecting the shape parameter γ, let us assume the following regression model

$$\lambda_i = \beta_0 \exp(\beta_1 x_{1i} + \cdots + \beta_k x_{ki})$$

and

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki}. \tag{14}$$

Assuming the mixture and non-mixture models introduced in Section 1, let us consider a gamma prior distribution $Gamma(0.001, 0.001)$ for the regression parameters β_0 and α_0 and a normal prior distribution $N(0, 100)$ for the regression parameters β_l and α_l, $l = 1, \ldots, k$. We also assume prior independence among the parameters.

Posterior summaries of interest are obtained from simulated samples for the joint posterior distribution using standard Markov Chain Monte Carlo (MCMC) methods as the Gibbs sampling algorithm [2] or the Metropolis-Hastings algorithm [1].

4. An application

Bone marrow transplants are standard treatments for acure leukemia. Prognosis of recovery may depend on risk factors known at the time of transplantation, such as patient and/or donor age and sex, the stage of initial disease, the time from prognosis to transplantation, among many others. The final prognosis may change as the patient post transplantation history develops with occurrence of events at random times during the recovery process, such as development of acute or chronic graft-versus-host disease (GVHD), return of the platelet count to normal levels, or development of infections. Transplantation can be considered a failure when a patient leukemia returns (relapse) or when he or she dies while in remission (treatment related to death).

In this study, 137 patients with acute myeloctic leukemia (AML) and acute lymphoblastic leukemia (ALL) received a combination of 16 mg/kg of oral Busulfan (BU) and 120 mg/kg of intravenous cyclopho sphamide (Cy) (99 AML and 38 ALL patients) and were treated at one of four hospitals: 76 at Ohio State University (OSU) in Columbus; 21 at Hahnemann University (HU) in Philadelphia; 23 at St. Vicent's Hospital (SVH) in Sidney, Australia and 17 at Alfred Hospital (AH) in Melbourne, Australia (data set introduced by K l e i n and M o e s c h b e r g e r , 1997)

In the analysis considered in this paper, we assume as lifetimes, the times (in days) to acute graft-versus-host disease (TA) with 111 censored observations and the following covariates: patient age in years; donor age in years; patient sex; donor sex; patient CMV (cytomegalovirus immune status); donor CMV; waiting time to transplant in days and different hospitals.

To analyze this data set, we consider the cure fraction models introduced in Section 1, in the presence or not of covariates. As a first analysis, we assume the cure fraction models not in presence of covariates.

In Table 1, we have the inference results considering Bayesian approaches assuming Models 1 and 2, we also have the inference results considering a standard Weibull distribution not considering cure fraction. The Bayesian estimates were obtained by Proc MCMC [6] available in SAS 9.2.

For all cases considered in this paper, we assume a "burn-in-sample" of size $15,000$ to eliminate the effect of the initial values used in the simulation approach; after this "burn-in-sample" period, we simulated another $200,000$ Gibbs Samples, taking every 100th sample, which gives a final sample of size $2,000$. Monte Carlo estimates for the random quantities of interest are based on this final Gibbs sample of size $2,000$. Convergence of the algorithm was monitored using standard methods, as the trace-plots of the simulated samples.

TABLE 1. Posterior summary (not considering the presence of covariates).

Model	Parameter	Posterior Mean (SD)	95 % Credible Interval	DIC
Standard Weibull	λ	0.0283 (0.0120)	(0.0104; 0.0572)	421.8
	γ	0.3251 (0.0587)	(0.2219; 0.4444)	
Model 1	λ	0.0019 (0.0019)	(0.0001; 0.0072)	352.7
	γ	1.8335 (0.2651)	(1.3766; 2.3809)	
	π	0.7993 (0.0348)	(0.7264; 0.8616)	
Model 2	λ	0.0014 (0.0018)	(0.00002; 0.0064)	347.7
	γ	1.9394 (0.3217)	(1.3941; 2.8134)	
	π	0.7980 (0.0349)	(0.7241; 0.8604)	

In Bayesian context using MCMC methods, we have used the DIC (Deviance Information Criterion) introduced by [7] and given automatically by the SAS software (see, Table 1).

From the fitted survival models, we conclude that the Models 1 and 2 are very well fitted by the survival times. From the results of Table 1, we observe that the Bayesian inferences give similar results; the DIC criteria for the two assumed models also give very close results. Overall, Model 2 (non-mixture model) is better fitted by the data (smaller Monte Carlo estimates for DIC). Model 1 (a first order approximation of Model 2) gives the larger value of DIC.

In the presence of covariates, we assume the following regression models (see (14)),

$$\begin{aligned}\lambda_i &= \beta_0 \exp\big[\beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} \\ &\quad + \beta_6 x_{6i} + \beta_7 (x_{7i} - \bar{x}_7) + \beta_8 x_{8i} + \beta_9 x_{9i} + \beta_{10} x_{10i}\big]; \\ \ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha_0 + \alpha_1 (x_{1i} - \bar{x}_1) + \alpha_2 (x_{2i} - \bar{x}_2) + \alpha_3 x_{3i} + \alpha_4 x_{4i} + \alpha_5 x_{5i} \\ &\quad + \alpha_6 x_{6i} + \alpha_7 (x_{7i} - \bar{x}_7) + \alpha_8 x_{8i} + \alpha_9 x_{9i} + \alpha_{10} x_{10i}, \qquad (15)\end{aligned}$$

were x_{1i} is the patient age; x_{2i} is the donor age; x_{3i} is the patient sex (1 = male; 0 = female); x_{4i} is the donor sex (1 = male; 0 = female); x_{5i} is the patient CMV (1 = CMV positive; 0 = CMV negative); x_{6i} is the donor CMV (1 = CMV positive; 0 = CMV negative); x_{7i} is the waiting time to transplant in days; x_{8i}, x_{9i} and x_{10i} are "dummy" variables related to the different hospitals where $x_{8i} = 1$ for OSU and 0 for the other hospitals; $x_{9i} = 1$ for HU and 0 for the other hospitals; $x_{10i} = 1$ for SVH and 0 for the other hospitals; $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_7$ are samples averages for covariates x_{1i}, x_{2i} and x_{7i}, $i = 1, \ldots, 137$.

To get some preliminary information on the regression parameters of model (15), we initially considered individual regression models with only one covariate assuming non-informative priors for all parameter models (see, Section 3). Using the SAS software, following the same simulation approach used to analyze the data not considering the presence of covariates, we observed that only covariates x_1, x_2 and x_7 showed some effect on the parameters λ_i and π_i, that is, where zero was not included in the 95 % credible intervals for the associated regression parameters. Then, we assume a multiple regression model including only the covariates x_1, x_2 and x_7, that is

$$\begin{aligned} \lambda_i &= \beta_0 \exp\left[\beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + \beta_7 (x_{7i} - \bar{x}_7)\right]; \\ \ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha_0 + \alpha_1 (x_{1i} - \bar{x}_1) + \alpha_2 (x_{2i} - \bar{x}_2) + \alpha_7 (x_{7i} - \bar{x}_7)\,. \end{aligned} \tag{16}$$

In Table 2, we have the inference results considering Bayesian approaches assuming Models 1 and 2. For a Bayesian analysis we assume non-informative priors for all parameters models (see Section 3).

TABLE 2. Posterior summary (multiple regression models on covariates x_1, x_2 and x_7).

Model	Parameter	Posterior Mean (SD)	95 % Credible Interval	DIC
Model 1	α_0	4.8749 (1.1655)	(3.0108; 7.6313)	333.5
	α_1	−0.0400 (0.0370)	(−0.1117; 0.0323)	
	α_2	−0.0250 (0.0362)	(−0.0971; 0.0484)	
	α_7	−0.0005 (0.0006)	(−0.0017; 0.0007)	
	β_0	0.0004 (0.0008)	(0.00001; 0.0021)	
	β_1	0.0269 (0.0288)	(−0.0329; 0.0865)	
	β_2	0.0141 (0.0429)	(−0.0700; 0.0986)	
	β_7	−0.0008 (0.0004)	(−0.0017; −0.00008)	
	γ	2.3440 (0.3298)	(1.6977; 2.9694)	
Model 2	α_0	4.8408 (1.2328)	(2.9245; 7.8466)	338.4
	α_1	−0.0370 (0.0354)	(−0.1080; 0.0303)	
	α_2	−0.0276 (0.0351)	(−0.0934; 0.0409)	
	α_7	−0.0004 (0.0006)	(−0.0016; 0.0008)	
	β_0	0.0002 (0.0004)	(0.000008; 0.0013)	
	β_1	0.0251 (0.0302)	(−0.0346; 0.0848)	
	β_2	0.0133 (0.0448)	(−0.0754; 0.1002)	
	β_7	−0.0009 (0.0004)	(−0.0019; −0.0001)	
	γ	2.4375 (0.3386)	(1.8055; 3.1163)	

From the results of Table 2, we observe similar results considering the Models 1 and 2. We also observe similar Monte Carlo estimates for the posterior means, standard deviations and credible intervals considering the two models, and similar DIC values. The covariate x_7 affect the scale parameter λ since zero is not included in the 95 % credible intervals for β_7.

5. Concluding remarks

Usually in the analysis of lifetime data we could have the presence of cure fraction and covariates, especially in medical applications. To analyze this kind of data, we have different parametrical formulations, as the mixture and non-mixture models given by equations (1) and (6). Computationally, especially using the Bayesian paradigm, the obtained results are very similar as observed in the application of the bone transplant lifetime data introduced in Section 4. The great advantage of the mixture model (1) is related to the simple interpretations, especially for medical researchers, where we have the proportion of cured and non-cured individuals given directly in the survival function expression. Further, generalizations have been obtained considering bivariate lifetime data in the presence of cured fraction and covariates. In this way, we are using standard one-parameter copula functions to derive parametrical formulations for the joint bivariate survival functions.

REFERENCES

[1] CHIB, S.—GREENBERG, E.: *Understanding the metropolis-hastings algorithm*, Amer. Statist. **49** (1995), 327–335.

[2] GELFAND, A. E.—SMITH, A. F. M.: *Sampling-based approaches to calculating marginal densities*, J. Amer. Statist. Assoc. **85** (1990), 398–409.

[3] KLEIN, J. P.—MOESCHBERGER, M. L.: *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, NY, 1997.

[4] LAMBERT, P. C.—DICKMAN, P. W.—WESTON, C. L.—THOMPSON, J. R.: *Estimating the cure fraction in population-based cancer studies by using finite mixture models*, J. Roy. Stat. Soc. Ser. C **59** (2010), 35–55.

[5] MALLER, R. A.—ZHOU, X.: *Survival Analysis with Long-Term Survivors*. John Wiley & Sons Ltd., Chichester, 1996.

[6] SAS: *The MCMC Procedure, SAS/STAT® User's Guide, Version 9.22*, Cary, NC: SAS Institute Inc., 2010, 4102–4326.

[7] SPIEGELHALTER, D.—BEST, N.—CARLIN, B.—VAN DER LINDE, A.: *Bayesian measures of model complexity and fit*, J. R. Stat. Soc. Ser. B, Methodol. **64** (2002), 583–639.

[8] TSODIKOV, A. D.—IBRAHIM, J. G.—YAKOVLEV, A. Y.: *Estimating cure rates from survival data: an alternative to two-component mixture models*, J. Amer. Statist. Assoc. **98** (2003), 1063–1078.

[9] WIENKE, A.—LOCATELLI, I.—YASHIN, A. I.: *The modelling of a cure fraction in bivariate time-to-event data*, Austrian J. Statist. **35** (2006), 67–76.

[10] YAKOVLEV, A.—TSODIKOV, A. D.: *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, World Scientific, Singapore, 1996.

Received October 31, 2011

Jorge A. Achcar
Department of Social Medicine—FMRP
University of São Paulo
Ribeirão Preto, SP
BRAZIL
E-mail: achcar@fmrp.usp.br

Emílio A. Coelho-Barros
Josmar Mazucheli
Department of Statistics
University of Maringá
Paraná
BRAZIL
E-mail: eacbarros@uem.br
jmazucheli@uem.br