STATISTICA, anno LXXXI, n. 3, 2021

REGRESSION ANALYSIS OF CURE MODEL WITH GENERALISED WEIBULL DISTRIBUTION

Parassery Parameswaran Rejani[1]
Department of Statistics ,Cochin University of Science and Technology, Cochin-682022, India
Paduthol Godan Sankaran
Department of Statistics ,Cochin University of Science and Technology, Cochin-682022, India

1. Introduction

The standard survival analysis techniques use an assumption that all subjects in the study population is susceptible to the event of interest and will eventually experience this event if the follow up time is more. But in certain situations, all of the study subjects do not experience the event of interest even after extended follow up time. For example, in drug trials to test the effectiveness of drugs, it is seen that recurrence of the disease does not happen in certain patients due to the influence of the drug. These disease free individuals, or more generally, the event free individuals in an observational window are said to be immune or cured. The presence of immune or cured subjects in a data set is usually suggested by a Kaplan-Meier plot of the survival function, which shows a long and stable plateau with heavy censoring at the extreme right of the plot. Cure models are applied in circumstances where immune are present in a time to event analysis. In studies based on cure models, study subjects are classified into two groups, say, sensitive or susceptible and insensitive or immune. In cure models, the survival distribution of failure time for the uncured patients are studied and the said fraction is taken into account. The cure models are applied in many areas like biomedical studies, finance, criminology, demography, manufacturing, and industrial reliability. The analysis of cure models have been done by many researchers in literature. Boag (1949) first proposed the cure model for the analysis of breast cancer data. He proposed two components in his mixture model, the proportion of immune in the population and latency distribution representing the survival experience of the susceptible population. Nelson (2003) applied cure models to study the association between variation of temperature and length of life of electric motors. Struthers and Farewell (1989) explained the progression of AIDS with cure model in the presence of covariates. For detailed applications in cure models, one can refer to

[1] Corresponding Author. E-mail: rejstat@gmail.com

Peng *et al.* (1998), Yu *et al.* (2004), Mazucheli *et al.* (2009), Roman *et al.* (2012), Ortega *et al.* (2015).

The regression model in survival analysis quantifies the effect of a set of explanatory variables on survival of individuals under study. Apart from ordinary survival regression models, the cure rate regression models allow the chance of occurrence of long-term survivors in the data. These models simultaneously useful to study the effect of covariates on the survival time of study subjects and to estimate the fraction of individuals who are free from the event of interest. In cure models, the failure time distribution of uncured individuals (latency) can be modeled either by parametric or semi-parametric proportional hazards models. Yamaguchi (1992) proposed a regression model to study inter-firm job mobility in Japan. The author used generalised gamma distribution to model the latency part and logistic function to model the cure fraction in terms of covariates. Cure models based on Weibull distribution was explained by Yusuf and Bakar (2016). Naseri *et al.* (2018) explained the application of cure rate model based on generalized modified Weibull distribution. The generalized Weibull distribution has non monotone hazard rate property, which gives its importance in lifetime studies. Even though various forms of generalised Weibull distributions confirmed its applicability in lifetime data modeling, nobody has studied the proposed generalised Weibull distribution for the regression analysis of lifetime data with cured proportion so far. Motivated by this, we propose generalised Weibull distribution developed by Mudholkar *et al.* (1996) for the regression analysis of cure models. The purpose of the present study is to describe the applicability of this distribution in cure rate regression models.

The rest of the paper is organized as follows. We introduce the cure model based on generalised Weibull distribution in Section 2. In Section 3, we explain the likelihood function and its inferential procedures. Section 4 contains the simulation study to assess influence of sample size on bias of estimates. A data analysis is carried out in Section 5 to illustrate the goodness of fit and usefulness of the model. Conclusion of the study is given in Section 6.

2. The model

Let T be a non negative random variable representing time to occurrence of the event. Define the indicator variable Y as

$$Y = \begin{cases} 1, & \text{if the event occurs,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For $Y = 1$, the time T has the probability density function $f(t|Y = 1)$ and survival function $S(t|Y = 1)$.

Assume that the distribution of lifetime for the susceptible population is generalised

Weibull distribution developed by Mudholkar *et al.* (1996) with survival function

$$S(t|Y=1)=\left(1-\frac{\left(\frac{t}{\nu}\right)^{\alpha}}{\theta}\right)^{\theta},\quad \alpha,\nu>0, \tag{2}$$

where, the parameter θ varies from $-\infty$ to $+\infty$ and the range of the random variable T is $(0,\infty)$ for $-\infty<\theta\leq 0$ and $(0,\nu\theta^{\frac{1}{\alpha}})$ for $0<\theta<\infty$.

It is shown that for the generalised Weibull distribution family (2), the hazard function $h(t)$ is:

(a) bathtub shaped for $0<\alpha<1$ and $\theta>0$,

(b) monotone decreasing for $0<\alpha\leq 1$ and $\theta\leq 0$,

(c) unimodal for $\alpha>1$ and $\theta<0$,

(d) monotone increasing for $\alpha\geq 1$ and $\theta\geq 0$, and

(e) constant for $\alpha=1$ and $\theta\to\infty$.

This flexibility allow us to use the model in wide range of applications.

Let Z be a $(p+1)\times 1$ vector of covariates. The survival function for the regression model with regression coefficient β is defined as

$$S(t|Y=1,Z=z)=\left(1-\frac{\left(\frac{t}{e^{\beta' z}}\right)^{\alpha}}{\theta}\right)^{\theta}. \tag{3}$$

Under logistic regression assumption for incidence part of the model, the probability of occurrence of the event is defined as

$$p=\Pr(Y=1)=p(b,z)=\frac{e^{b'z}}{1+e^{b'z}} \tag{4}$$

and

$$1-p=1-p(b,z)=\frac{1}{1+e^{b'z}}, \tag{5}$$

where $b=(b_0,b_1,...b_p)$ is a vector to model the effect of covariates in the incidence part.

Then, the marginal survival function of T is

$$S(t|z)=1-p+p\left(1-\frac{\left(\frac{t}{e^{\beta' z}}\right)^{\alpha}}{\theta}\right)^{\theta}, \tag{6}$$

where $t<\infty$. Note that as $t\to\infty$, $S(t|z)\to 1-p$.

3. ESTIMATION

In regression modeling, the researcher's primary interest is to estimate the regression parameters of the model. Denote the observations for the i'th individual be (t_i, δ_i, z_i), $i = 1, ...n$, where t_i is observed (survival) time or the censoring time and δ_i is the indicator function given by $\delta_i = 1$, if t_i is uncensored and $\delta_i = 0$, otherwise.We assume that the censoring is statistically independent of Y.

Without loss of generality, we assume that $t_1, ...t_m$ are the survival times and $t_{m+1}, ...t_{m+n}$ are censored times. Obviously, the random variable Y=1 for the first m individuals and is unknown for the remaining $n - m$ individuals. Then the likelihood function for cure model corresponding to the observations (t_i, δ_i, z_i), $i = 1, ...n$ is

$$L = L_1 \times L_2, \tag{7}$$

where

$$L_1 = \prod_{i=1}^{m} p_i f(t_i|Y = 1, z_i) \tag{8}$$

and

$$L_2 = \prod_{m+1}^{n} (1 - p_i) + p_i S(t_i|Y = 1, z_i). \tag{9}$$

The maximum likelihood estimators of the parameters are found by EM algorithm (Dempster *et al.*, 1977) since partial information of random variable Y is missing. The estimation of parameters is carried out using the optimisation function NArgMax in Wolfram Mathematica software. We use the notations L for likelihoods and l for log-likelihoods through out the paper.

3.1. *EM Algorithm*

The following step by step procedure is constituted in the estimation of parameters in EM algorithm.

Step 1 - The E step in the EM algorithm compute the conditional expectation of the complete log-likelihood with respect to Y's, given the observed data and current estimates of the parameters.

Let the observed data be {O = Observed y's, (t_i, δ_i, z_i); $i = 1, ..., n$}. Now using Eq. (3) and Eq. (4), calculate the missing observations w_1 and w_2, which are the conditional probabilities that the individuals belong to immune group or suspected group given that the individuals survived up to the time t, respectively, as

$$w_1 = \Pr(Y = 1|T > t) = \frac{p(b, z_i)S(t_i|Y = 1, z_i)}{1 - p(b, z_i) + p(b, z_i)S(t_i|Y = 1, z_i)} \tag{10}$$

and

$$w_2 = \Pr(Y = 0|T > t) = \frac{1 - p(b, z_i)}{1 - p(b, z_i) + p(b, z_i)S(t_i|Y = 1, z_i)}. \tag{11}$$

Then, by including the missing observations, the complete log-likelihood function can be written as

$$l = l_1 + l_2, \tag{12}$$

where

$$l_1 = \sum_{i=1}^{m} \log p_i + \sum_{i=m+1}^{n} w_2 \log(1 - p_i) + \sum_{i=m+1}^{n} w_1 \log p_i \tag{13}$$

and

$$l_2 = \sum_{i=m+1}^{n} w_1 \log \left(1 - \frac{\left(t/e^{\beta' z}\right)^{\alpha}}{\theta}\right)^{\theta} + \sum_{i=1}^{m} \log\left(\alpha t^{\alpha-1}\left(\frac{1}{e^{\beta' z}}\right)^{\alpha-1}\right)\left(1 - \frac{\left(t/e^{\beta' z}\right)^{\alpha}}{\theta}\right)^{\theta-1}. \tag{14}$$

Step 2 - M-step maximizes the likelihood function through Eq. (12) with respect to the parameters and find out the maximum likelihood estimates of these parameters. If $\alpha^{(k)}$, $\beta^{(k)}$ and $b^{(k)}$ are estimates of α, β and b at the k^{th} iterate, then the estimates at $(k+1)^{th}$ iterate, $\alpha^{(k+1)}$, $\beta^{(k+1)}$ and $b^{(k+1)}$ can be obtained by maximizing the likelihood function with respect to each parameter α, β and b respectively for fixed values of w_1 and w_2. Then at the $(k+1)^{th}$ stage,

$$w_1^{(k+1)} = \frac{P\left(b^{(k)}, z_i\right) S^{(k)}(t_i|Y = 1, z_i)}{1 - P\left(b^{(k)}, z_i\right) + P\left(b^{(k)}, z_i\right) S^{(k)}(t_i|Y = 1, z_i)} \tag{15}$$

and

$$w_2^{(k+1)} = \frac{1 - P\left(b^{(k)}, z_i\right)}{1 - P\left(b^{(k)}, z_i\right) + P\left(b^{(k)}, z_i\right) S^{(k)}(t_i|Y = 1, z_i)}. \tag{16}$$

Step 3 - The E-Steps and M-Steps are repeated alternatively until the difference between parameter estimates of successive iterations changes by an arbitrarily small quantity.

3.2. *Asymptotic property of estimators*

Let $\hat{\psi} = (\hat{b}, \hat{\beta}, \hat{\theta}, \hat{\alpha})$ denote the maximum likelihood estimates of $\psi = (b, \beta, \theta, \alpha)$. Now consider the following regularity conditions.

(a) The first and second order derivatives with respect to ψ viz., $\frac{\partial l}{\partial \psi}$ and $\frac{\partial^2 l}{\partial \psi^2}$ exist and are continuous functions of ψ in a range R (including the true value ψ_0 of the parameter) for almost all t. For every ψ in R, $\left|\frac{\partial l}{\partial \psi}\right| < H_1(t)$ and $\left|\frac{\partial^2 l}{\partial \psi^2}\right| < H_2(t)$, where $H_1(t)$ and $H_2(t)$ are integrable functions over $(-\infty, \infty)$.

(b) The third order derivative with respect to ψ, $\frac{\partial^3 l}{\partial \psi^3}$ exists such that $\left|\frac{\partial^3 l}{\partial \psi^3}\right| < M(t)$, where $E[M(t)] < K$ and K is a positive quantity.

(c) For every ψ in R, $E\left(-\frac{\partial^2 l}{\partial \psi^2}\right) = \int_{-\infty}^{\infty} \left(-\frac{\partial^2 l}{\partial \psi^2}\right) L\,dt = I(\psi)$ is finite and non-zero.

(d) The range of integration is independent of ψ. This assumption is to make differentiation under the integral sign valid.

Under the above mentioned regularity conditions, as $n \to \infty$, $\sqrt{n}(\psi - \hat{\psi}) \to N_4(0, I^{-1}(\psi))$, where the Fisher information matrix $I(\psi)$ can be replaced by a consistent estimate $I(\hat{\psi}) = \left(\frac{-\partial^2 l}{\partial \psi_i \partial \psi_j}\right)_{\psi=\hat{\psi}}$. The observed information matrix is obtained by applying the method proposed by Louis (1982).

The asymptotic normality property of maximum likelihood estimates is useful to determine the confidence interval of each parameter in the parametric set $\psi = (b, \beta, \theta, \alpha)$ and for survival function of the model.

Let $\hat{b}$ is the maximum likelihood estimator (MLE) of b. Then MLE of cured proportion $1-p = \frac{1}{1+e^{\hat{b}'z}}$ is $1-\hat{p} = g(\hat{b}) = \frac{1}{1+e^{\hat{b}'z}}$ is also asymptotically normally distributed by the invariance property of maximum likelihood estimators. The 95% Confidence interval of the probability of cure can be estimated using the formula $(1-\hat{p}) \pm 1.96SE(1-\hat{p})$.

REMARK 1. *In regression analysis, it is often required to test the statistical significance of regression coefficients in the model. ie, to test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_0 : \beta \neq 0$, The likelihood ratio test statistic is $-2\log\Lambda = 2LogL(\hat{\beta}, \hat{\alpha}, \hat{\theta}) - 2LogL(0, \hat{\alpha}, \hat{\theta})$ which follows Chi-square distribution with p degrees of freedom where p is number of parameters of the model.*

REMARK 2. *The Akaike information criterion (AIC) is a procedure based on likelihood function that can be used for comparing statistical models for a given set of data. Let* $\hat{L}$ *be the maximum value of the likelihood function for the model and let k be the number of estimated parameters in the model. Then the AIC value of given model is* $AIC = 2k - 2\ln(\hat{L})$. *If more than one models are given, AIC values can be computed and the model with minimum AIC value is selected as preferred model for the given set of data.*

4. SIMULATION STUDY

Simulation studies are conducted to evaluate the performance of the proposed model. The data are generated from the model, with probability of cure defined as $1 - p = \frac{1}{1+\exp(b_0+b_1 z)}$.

For the purpose of simulation, the following step by step algorithm is used to generate data from the model and estimation of parameters.

1. Determine the parameter values b_0, b_1, β, θ and α.
2. For the ith subject, generate the covariate z_i from Uniform $(-0.5, 0.5)$.
3. For the ith subject, generate the probability of cure $1 - p_i$.
4. For the ith subject, generate the random variable C_i from Uniform $(0,c)$, where c is a constant set to control the proportion of censored observations.
5. For the ith subject, generate T_i from the model.
6. For the ith subject, find $t_i = \min(T_i, C_i, \tau)$,where τ is the maximum follow up period, $\tau = 10$. If $t_i = T_i$, set $\delta_i = 1$, otherwise $\delta_i = 0$.
7. The data set for the ith subject is (t_i, z_i, δ_i), $i = 1, \ldots, n$.
8. Maximise the likelihood function with the generated data sets using EM algorithm as described in Section 3.1.

Observations are simulated for various sample sizes and for two sets of parametric values of b, β, θ and α. Based on 500 simulations, the maximum likelihood estimates of the parameters are calculated for mild censored and heavy censored data. Average levels of mild and heavy censoring schemes are 16.8%, 41% and 21.5%, 40.5%, respectively, for two given parametric values. The true values, absolute bias, mean squared error (MSE) and average of estimated standard errors (ASE) based on 500 simulated data sets for sample sizes of 50, 100 and 200 are given in Table 1. It shows that the estimates are approximately unbiased and the bias increases as censoring changes from mild to heavy. Also, as sample size increases, both bias and MSE decrease.

TABLE 1
Absolute Bias, MSE and ASE of maximum likelihood estimates of b_0, b_1, β, θ and α.

Sample		True	Mild			Heavy		
size	Parameter	value	Bias	MSE	ASE	Bias	MSE	ASE
50	b_0	1.0	0.055	0.019	0.169	0.111	0.061	0.222
	b_1	2.0	0.016	0.099	0.314	0.154	0.512	0.701
	β	1.0	0.012	0.019	0.137	0.082	0.250	0.493
	θ	1.0	0.053	0.067	0.259	0.200	0.076	0.486
	α	1.0	0.046	0.015	0.122	0.195	0.070	0.181
100	b_0	1.0	0.038	0.005	0.062	0.111	0.035	0.177
	b_1	2.0	0.003	0.095	0.308	0.157	0.285	0.510
	β	1.0	0.014	0.003	0.059	0.070	0.021	0.013
	θ	1.0	0.010	0.054	0.225	0.074	0.067	0.247
	α	1.0	0.019	0.012	0.108	0.193	0.067	0.172
200	b_0	1.0	0.009	0.004	0.047	0.035	0.032	0.150
	b_1	2.0	0.001	0.089	0.292	0.082	0.250	0.493
	β	1.0	0.014	0.003	0.053	0.052	0.019	0.127
	θ	1.0	0.002	0.010	0.099	0.031	0.038	0.236
	α	1.0	0.009	0.009	0.092	0.169	0.058	0.171
50	b_0	0.9	0.024	0.031	0.146	0.051	0.004	0.078
	b_1	2.1	0.024	0.033	0.145	0.107	0.049	0.222
	β	0.7	0.088	0.032	0.0179	0.263	0.072	0.127
	θ	1.1	0.109	0.109	0.238	0.269	0.238	0.242
	α	0.65	0.056	0.032	0.137	0.165	0.056	0.173
100	b_0	0.9	0.001	0.014	0.006	0.030	0.007	0.076
	b_1	2.1	0.015	0.010	0.106	0.071	0.048	0.219
	β	0.7	0.055	0.012	0.097	0.070	0.021	0.083
	θ	1.1	0.023	0.178	0.166	0.057	0.178	0.191
	α	0.65	0.054	0.023	0.097	0.150	0.044	0.167
200	b_0	0.9	0.000	0.012	0.009	0.025	0.002	0.060
	b_1	2.1	0.011	0.007	0.010	0.011	0.043	0.198
	β	0.7	0.015	0.011	0.053	0.014	0.009	0.077
	θ	1.1	0.019	0.021	0.139	0.022	0.021	0.143
	α	0.65	4.2E-07	0.012	0.032	0.086	0.028	0.152

5. Data analysis

We illustrate the applicability of the proposed model with a real life data set. We consider the data on survival times (in months) of 101 patients with advanced acute myelogenous leukemia reported to the International Bone Marrow Transplant Registry. The data are given in Klein and Moeschberger (2003). Out of 101 patients, fifty one had received an autologous (auto) bone marrow transplant and fifty patients had an allogeneic (allo) bone marrow transplant. The plot of the Kaplan-Meier estimator of the data is displayed in Figure 1 and shows a large plateau at about 0.47. Furthermore, a large proportion of the censored observations is in the plateau, which suggests that a cure model is appropriate for these data. We considered the type of treatment (allo-auto) as a covariate in our regression model.

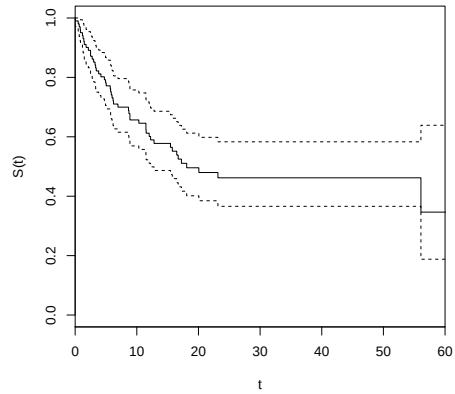

Figure 1 – Kaplan-Meier Survival curve of data set.

The maximum likelihood estimators of regression parameter were found out via EM Algorithm as described in Section 3.1. The estimator of regression parameter and its standard error obtained as $\hat{\beta} = 2.195$ (0.451) and other parameters of the model and corresponding standard errors estimated to be $\hat{b}_0 = 0.328$ (0.201), $\hat{b}_1 = -0.107$ (0.284), $\hat{\theta} = -1.140$ (0.332) and $\hat{\alpha} = 0.672$ (0.077). The statistical significance of the regression coefficient was tested by likelihood ratio test procedure as mentioned in Remark 1 of Section 3 and the result is found to be significant ($p < 0.001$) and it is inferred that the covariate, type of treatment has a significant positive effect on survival time in this study. The present model was compared with Weibull cure models using Akaike information criteria (AIC) as described in Remark 2 of Section 3 and it is found out that our model is best fit compared to Weibull model. The results are described in Table 2.

TABLE 2
Comparison of Weibull and generalised Weibull models.

Model	Parameter	Estimate (SE)	LL	AIC
Weibull	b_0	0.351 (0.200)	-	-
	b_1	-0.264 (0.283)	-	-
	β	1.364 (0.505)	-271.609	551.218
	θ	0.422 (0.037)	-	-
generalised Weibull	b_0	0.328 (0.201)	-	-
	b_1	-0.107 (0.285)	-	-
	β	2.195 (0.451)	-	-
	θ	-1.140 (0.332)	-216.833	443.667
	α	0.672 (0.077)	-	-

The survival curves for the model along with Kaplan-Meier estimates are drawn for two transplant groups and are given in Figure 2. The curves are close to Kaplan- Meier curve suggesting the proposed parametric model fits well.

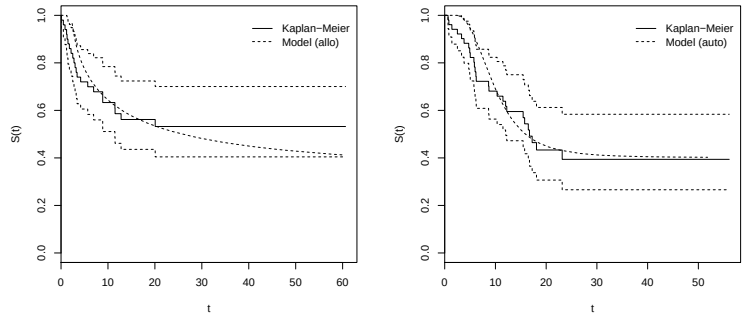

Figure 2 – Fitted curves of transplant groups 'allo' and 'auto'.

Figure 3 shows plots of survival functions for the model parameters. From the curves, we can see that the survival of auto group has more survival rate at early stage of recovery. This is due to the fact that the patients in this group are not at risk of developing complications like acute graft-versus-host disease, which is commonly seen in the allo transplant group. After this period, about one year, we can see the advantage for allogeneic transplants, due to the decreased relapse rate in these patients. The survival curves for both groups of data crosses each other due to the early advantage for autologous transplants group and due to the progress of allogeneic group in later stages.

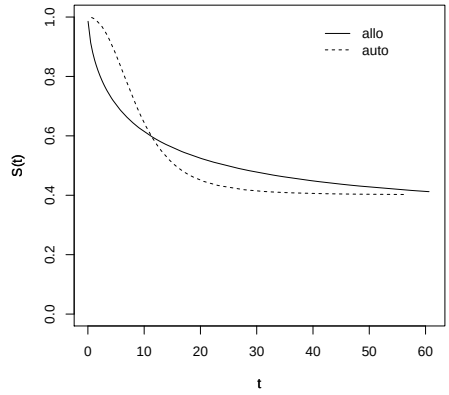

Figure 3 – Survival curves of allo and auto transplants.

The graphical method based on Total Time on Test (TTT) plot is used to illustrate the shape of hazard function of the model. The scaled TTT transform for the generalised Weibull distribution model is defined as

$$\phi_F(u) = \frac{H_F^{-1}(u)}{H_F^{-1}(1)}, \quad 0 < u < 1, \tag{17}$$

where

$$H_F^{-1}(u) = \int_0^{F^{-1}(u)} [1 - F(x)]\mathrm{d}x. \tag{18}$$

The function $F^{-1}(u)$ is obtained through Eq. (3). $H_F^{-1}(u)$ can easily be found out using the estimated values of model parameters. Aarset (1987) showed that the scaled TTT transform is convex (concave) if the hazard rate is decreasing (increasing), and for bathtub (unimodal) hazard rates, the scaled TTT transform is first convex (concave) and then concave (convex). Figure 4 shows the scaled TTT plot of the data. The plot drawn give the evidence of decreasing hazard and it agree with the characteristics of the hazard function of the generalised Weibull distribution explained in Section 2. Hence the distribution assumption is appropriate to the given data.

The 95% Confidence interval of the probability of cure is estimated as $(0.403, 0.717)$ and $(0.264, 0.583)$ for allo and auto group respectively. The overlapping of confidence interval occurred due to the complications of allo transplants in early stage of recovery period. The comparison of upper limit of both confidence intervals is more considered. It can be seen that the proportion of cure in allo transplant group is more compared to

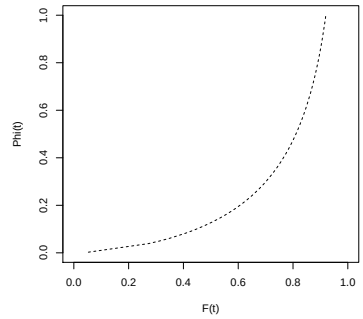

Figure 4 – Scaled TTT transform plot showing the shape of hazard function of the distribution.

the auto transplant group giving the evidence of decreased recurrence in allo group compared to auto group and our results coincides with many clinical study results conducted in this area (Fenske *et al.*, 2016).

The graphical check of overall fit of the proposed regression model is assessed by Cox-Snell residuals (Klein and Moeschberger, 2003). The Cox-Snell residual r_i, is defined by $r_i = \hat{H}(T_i|Z_i)$, where $\hat{H}$ is the fitted model. If the model fits the data, then the r_i's should follow a standard exponential distribution so that the hazard plot of r_i versus the Nelson-Aalen estimator of the cumulative hazard of the r_i's should be a straight line with slope one. In our example we plot the r_i versus Nelson- Aalen estimator of cumulative hazard plot for two groups of data and shown in Figure 5. We see from these plots that the model give reasonable fits to the data.

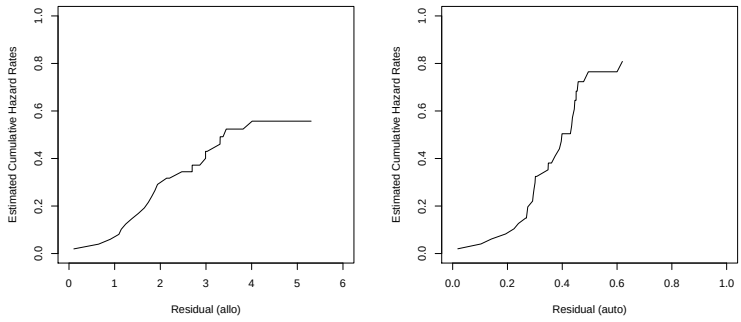

Figure 5 – Cox-Snell residuals of transplant groups 'allo' and 'auto'.

6. Conclusion

In this paper we proposed a parametric regression model with generalised Weibull distribution for the analysis of survival data with cured fraction. The probability of cure was modeled using logistic distribution assumption and the model parameters were estimated using maximum likelihood method via EM algorithm. The proposed model was applied to a real data set on bone marrow transplantation and it was found to be good fit. The proposed model was compared with Weibull cure model using the Akaike information criterion and better results found out with the new model.

Acknowledgements

We thank the reviewers and editors for their constructive comments that greatly improved this paper.

References

M. V. Aarset (1987). *How to identify a bathtub hazard rate*. IEEE Transactions on Reliability, 36, no. 1, pp. 106–108.

J. W. Boag (1949). *Maximum likelihood estimates of the proportion of patients cured by cancer therapy*. Journal of the Royal Statistical Society, Series B, 11, no. 1, pp. 15–53.

A. P. Dempster, N. M. Laird, D. B. Rubin (1977). *Maximum likelihood from incomplete data via the* EM *algorithm*. Journal of the Royal Statistical Society, Series B, 39, no. 1, pp. 1–22.

T. S. Fenske, M. Hamadani, J. B. Cohen, L. J. Costa, B. S. Kahl, A. M. Evens, P. A. Hamlin, H. M. Lazarus, E. Petersdorf, C. Bredeson (2016). *Allogeneic hematopoietic cell transplantation as curative therapy for patients with non-*Hodgkin *lymphoma: increasingly successful application to older patients*. Biology of Blood and Marrow Transplantation, 22, no. 9, pp. 1543–1551.

J. P. Klein, M. L. Moeschberger (2003). *Survival analysis: techniques for censored and truncated data*, vol. 1230. Springer, New York.

T. A. Louis (1982). *Finding the observed information matrix when using the* EM *algorithm*. Journal of the Royal Statistical Society, Series B, 44, no. 2, pp. 226–233.

J. Mazucheli, J. Achcar, E. Coelho-Barros, F. Louzada-Neto (2009). *Infant mortality model for lifetime data*. Journal of Applied Statistics, 36, no. 9, pp. 1029–1036.

G. S. Mudholkar, D. K. Srivastava, G. D. Kollia (1996). *A generalization of the* Weibull *distribution with application to the analysis of survival data*. Journal of the American Statistical Association, 91, no. 436, pp. 1575–1583.

P. NASERI, A. R. BAGHESTANI, N. MOMENYAN, M. ESMAEIL AKBARI (2018). *Application of a mixture cure fraction model based on the generalized modified* Weibull *distribution for analyzing survival of patients with breast cancer*. International Journal of Cancer Management, 11, no. 5.

W. B. NELSON (2003). *Applied life data analysis*, vol. 521. John Wiley & Sons, New York.

E. M. ORTEGA, G. M. CORDEIRO, A. K. CAMPELO, M. W. KATTAN, V. G. CANCHO (2015). *A power series beta* Weibull *regression model for predicting breast carcinoma*. Statistics in medicine, 34, no. 8, pp. 1366–1388.

Y. PENG, K. B. DEAR, J. DENHAM (1998). *A generalized* F *mixture model for cure rate estimation*. Statistics in medicine, 17, no. 8, pp. 813–830.

M. ROMAN, F. LOUZADA, V. G. CANCHO, J. G. LEITE, *et al.* (2012). *A new long-term survival distribution for cancer data*. Journal of Data Science, 10, no. 2, pp. 241–258.

C. A. STRUTHERS, V. T. FAREWELL (1989). *A mixture model for time to* AIDS *data with left truncation and an uncertain origin*. Biometrika, 76, no. 4, pp. 814–817.

K. YAMAGUCHI (1992). *Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan*. Journal of the American Statistical Association, 87, no. 418, pp. 284–292.

B. YU, R. C. TIWARI, K. A. CRONIN, E. J. FEUER (2004). *Cure fraction estimation from the mixture cure models for grouped survival data*. Statistics in medicine, 23, no. 11, pp. 1733–1747.

M. U. YUSUF, M. R. A. BAKAR (2016). *Cure models based on* Weibull *distribution with and without covariates using right censored data*. Indian Journal of Science and Technology, 9, no. 28.

SUMMARY

Cure models are of special attention when all of the study subjects do not experience the event of interest even after long follow-up time. Many researchers have used exponential, gamma and Weibull distribution in the latency part of parametric cure models. In this article, we propose a new regression model with cured fraction, in its latency part is explained by the generalised Weibull distribution (Mudholkar *et al.*, 1996). The estimation of the parameters of the proposed model is done using maximum likelihood method *via* EM algorithm. Simulations are carried out to study the effect of sampling fluctuations and to know the efficiency of estimators. The proposed model is applied to real data on acute myelogenous leukaemia. The statistical significance of the regression parameter is checked by likelihood ratio (LR) test and the new model was compared with Weibull cure model using Akaike information criterion (AIC).

Keywords: Cure models; EM algorithm; Likelihood ratio; Akaike information criterion; Generalised Weibull distribution.