Generalized Poisson-Lindely Distribution in Promotion Time Cure Model

Ahmad Reza Baghestani
Affiliation Department of Biostatistics
Shahid Beheshti University of Medical Sciences, Tehran, Iran
Country Iran, Islamic Republic of
baghestani.ar@gmail.com

Mitra Rahimzadeh
Affiliation Alborz University of Medical Sciences, Alborz, Iran
Country Iran, Islamic Republic of
rahimi_1351@yahoo.com

Mohamad Amin Pourhoseingholi
Affiliation Department of Biostatistics
Shahid Beheshti University of Medical Sciences, Tehran, Iran
Country Iran, Islamic Republic of
amin_phg@yahoo.com

Abstract

Long-term survival analysis has been improved in the last decade and most of the models concentrate on the promotion time cure model that proposed by Chen (1999). These models are based on the distribution of latent variable N, number of initiated node cells. In this paper we proposed a Generalized Poisson-Lindely distribution that is another option instead of Negative Binomial distribution when there is overdispersion. The results indicated a better fitness compared to others, because of its more flexibility. Parameter estimation has been done by Bayesian approach, in a real data set and a simulation study has shown the advantages of proposed model.

Keywords: Survival analysis, Long-term survival models, Promotion time cure model, Generalized Poisson-Lindely, Bayesian approach.

1. Introduction

For analyzing count data with over dispersion, it's common to use Negative Binomial (NB) distribution instead of Poisson (P) distribution. It is straightforward that, when the parameter of the Poisson distribution has Gamma distribution, Negative Binomial distribution is obtained. Another choice for Gamma distribution is Lindely or Generalized Lindely Distribution. So the result is called Poisson Lindely (PL) or Generalized Poisson Lindely (GPL) Distribution.

In this paper we used the GPL in the long-term survival analysis. In the Long-term survival analysis two broad classes of Models are used. The first one which has been introduced by Boag (1949) and Brekson and Gage (1952), is called mixture model and the second one has been introduced by Yakovlev and Tsodikov (1996) and Chen et al. (1999) is called non-mixture cure model or promotion time cure model in cancer relapse setting, assume that lymph node cells act as competing causes to produce the detectable tumor cells. Cooner et al., (2007) generalized this approach to a flexible class of cure

models under different latent activation schemes. Several authors considered different distribution for the number of competing cause such as Poisson, Geometric, Negative Binomial, Conway-Maxwell Poisson or generalized power series distribution (see e.g. Chen et al., 1999; Cooner et al., 2007; Cancho et al. 2011; Rodrigues et al. 2009; Borgers et al. 2012).

In this paper, Generalized Poisson-Lindely Distribution was proposed for the number of lymph node cells in the promotion time cure model for obtain a more flexible model to fit a published data set.

A Bayesian framework was assumed for parameter estimation since the posterior distributions do not have a close form and because of complex structure of the model, the Markov chain Monte Carlo (MCMC) methods were employed for the purpose. In other to compare the models, the deviance information criteria (DIC) were applied, as a result of which the smallest value has shown the better fitness.

The rest of this paper is organized as follows. In the next section we introduced Generalized Pisson-Lindely distribution. In third section GPL distribution was proposed in promotion time cure model. Statistical modeling and parameter estimation were discussed in Section 4. Section 5 was devoted to the application of the model in cutaneous melanoma data set and simulation study. Results were discussed and concluded in final Section.

2. Generalized Poisson Lindely Distribution:

The Generalized Lindely distribution has been introduced by Zakerzadeh and Dolati (2010) with the probability density function;

$$f(x;\alpha,\theta)=\frac{\theta^{\alpha+1}}{\theta+1}\frac{x^{\alpha-1}}{\Gamma(\alpha+1)}(\theta+x)\exp(-\theta x)\quad x>0,\ \alpha>0,\ \theta>0 \qquad (1)$$

They mentioned that this distribution can be replaced instead of the Gamma and Weibull distribution for analyzing lifetime or skewed data.

Suppose X|λ~P(λ) & $\lambda\sim GL(\alpha,\theta)$ Then $X\sim GPL(\alpha,\theta)$ with the density function is given by;

$$f(x;\theta,\alpha)=\frac{\Gamma(\mathrm{x}+\alpha)}{x!\Gamma(\alpha+1)}\frac{\theta^{\alpha+1}}{(\theta+1)^{x+\alpha+1}}\left(\alpha+\frac{x+\alpha}{\theta+1}\right)\quad ,x=0,1,2,\ldots\ \theta>0,\alpha>0 \quad (2)$$

While θ is the scale and α the shape parameter. It is obvious that if $\alpha=1$, this distribution reduce to the Lindely Poisson distribution that have been shown by Ghitany et al., (2008) which in many ways, is a better distribution to model count data. Ghitany and Al-Mutari have done a complete comparison between Negative Binomial and Poisson Lindely distribution. For more detail see (12).

To comparison with the NB distribution proposed by Cancho et al., (2011) and well known P distribution, please refer to table 1 for their properties.

Table 1: Mean, Variance and Over-dispersion of P, NB and GPL distribution

Distribution	Mean	Variance	overdispersion
Poisson	θ	θ	1
Negative-Binomial	θ	$\theta+\eta\theta^2$	$1+\eta\theta$
Generalized Poisson-Lindely	$\frac{\alpha(\theta+1)+1}{\theta(\theta+1)}$	$\frac{\alpha(\theta+1)^3+\theta^2+3\theta+1}{\theta^2(\theta+1)^2}$	$1+\frac{\alpha(\theta+1)^2+2\theta+1}{\alpha\theta(\theta+1)^2+\theta(\theta+1)}$

So the GPL (α,θ) is over-dispersed for all values of α and θ and equal-dispersed $(\mu=\sigma^2)$ if $\frac{\alpha(\theta+1)^2+2\theta+1}{\alpha\theta(\theta+1)^2+\theta(\theta+1)}$=0 and the limitation of it when the $\theta\to\infty$ is $\frac{\alpha}{\alpha+1}$ so this distribution is equal-dispersed for large enough amount of θ. Negative Binomial is over-dispersed for all value of η and θ and equal-dispersed if $\eta=0$.

3. Promotion time cure Model with Generalized Poisson-Lindely Distribution:

In the promotion time cure rate model has been introduced by Cooner (2007), in the first activation scheme, If N is the number of competing causes (lymph nodes that remain actives after treatment), $Z_1,Z_2,\ldots,Z_N$ are time for the jth competing causes to produce the detectable tumor cells and the observable time to event is defined as $T=min\{Z_1,Z_2,\ldots,Z_N\}$, with $P(T=\infty|N=0)=1$ also N is independent of $Z_1,Z_2,\ldots,Z_N$, the survival function for population can be obtained by;

$$S_p=P(N=0)+\sum_{n=1}^{\infty}P(Z_1>t,\ldots,Z_n>t)P(N=n)=\sum_{n=0}^{\infty}S(t)^nP(N=n)=G_N(S(t)) \quad (3)$$

While $G_N(.)$ is the Probability Generating Function of the N, so the survival function for population is given by;

$$S_{p(t)}=\left(\frac{\theta}{\theta-S(t)+1}\right)^{\alpha+1}\left(\frac{\theta-S(t)+2}{\theta+1}\right) \quad (4)$$

And the density function is given by;

$$f_p(t)=\frac{\theta f(t)}{\theta+1}\left(\frac{\theta}{\theta-S(t)+1}\right)^{\alpha}\left[\frac{(\alpha+1)(\theta-S(t)+2)-(\theta-S(t)+1)}{(\theta-S(t)+1)^2}\right] \quad (5)$$

Where S(t) is the survival function of promotion time of N lymph nodes that can be any of the common survival function like Weibull, Piece Wise ,…. In this model the cure fraction should be $P(N=0)=\left(\frac{\theta}{\theta+1}\right)^{\alpha+1}\left(\frac{\theta+2}{\theta+1}\right)$, where $\theta=\exp(\beta'X)$ so the relation between the covariates and the cure rate like the Poisson model is direct. For example with increased in the coefficient covariate the cure rate is increasing.

4. Statistical method

4.1 Likelihood function

Suppose that there are n subjects and let N_i be the number of lymph nodes representing the number of competing causes that can produce a detectable tumor cells for the ith subject. Let Tj and Cj denote, respectively, the observable lifetime and the censored time for the ith subject, such that $T_i = min\{Y_i, C_i\}$ and δ_i is an indicator function that $\delta_i = 1$ if $T_{i=}Y_i$ and $\delta_i = 0$ if $T_{i=}C_i$. So for the *i*th individual, our observed data $\mathbf{D_{obs}}$ = {T*i, δi ,Xi* } where X_i is a matrix containing covariates.

We assumed that the N_i , i= 1,2,…,n, are independent generalized Poisson Lindely variables with probability function given by (2), with $\theta > 0$, $\alpha > 0$ and given N_i=n_i the promotion times $Z_1, Z_2, ..., Z_N$ are independent with Weibull distribution described by (7). The corresponding Likelihood function under right censor is given by;

$$L(\alpha,\theta,\beta,\tau,\nu|D_{obs})=\prod_{i=1}^{n} S_p(t)^{1-\delta_i} \times f_p(t)^{\delta_i}=$$

$$\left\{\left(\frac{\theta}{\theta-S(t)+1}\right)^{\alpha+1}\left(\frac{\theta-S(t)+2}{\theta+1}\right)\right\}^{1-\delta_i} \times\left\{\frac{\theta f(t)}{\theta+1}\left(\frac{\theta}{\theta-S(t)+1}\right)^{\alpha}\left[\frac{(\alpha+1)(\theta-S(t)+2)-(\theta-S(t)+1)}{(\theta-S(t)+1)^2}\right]\right\}^{\delta_i} \quad (6)$$

Where the survival function of the Weibull distribution is as the following;

$$S(t) = \exp(-e^{\tau}\nu^{t}) \quad \upsilon > 0 \;\&\; -\infty < \tau < \infty \qquad (7)$$

4.2 Parameter Estimation:

For parameter estimation we employed the Bayesian approach using the MCMC methods. We take non-informative prior in order to the likelihood function dominate the posterior distribution. Without lost generality, we supposed that the prior distributions are independent.

For β we consider $\Pi(\beta) \propto 1$ a uniform improper prior, for τ has considered Normal distribution which $\tau \sim N(0, \sigma^2)$ and for ν and α have considered Gamma distribution which $\nu \sim G(\gamma_\nu, \eta_\nu)$ and $\alpha \sim G(\gamma_\alpha, \eta_\alpha)$. Combining these prior distribution with the likelihood function of the posterior distribution of $(\beta, \tau, \nu, \alpha)$ obtains to be;

$$\Pi(\beta,\tau,\nu,\alpha) \propto L(\alpha,\theta,\beta,\tau,\nu|D_{obs}) \times \pi(\beta)\pi(\tau)\pi(\nu)\pi(\alpha) \qquad (8)$$

Because of analytically intractable of the joint posterior density in equation (8), we applied the Markov Chain Monte Carlo (MCMC) simulations, carried out with Metropolis Hastings algorithm. [6]

For comparison, we considered the numbers of lymph nodes have Negative Binomial and Poisson distribution that are discussed by Cancho et al., (2011) and Chen et al. (1999).

Bayesian estimates were calculated for each parameter using the samples drawn from conditional posterior distributions, which usually derived from the marginal distributions obtained from the joint distribution of parameters given the observations. In this model, posterior joint distribution of the parameters takes a complicated form and it is too difficult to derive the posterior marginal distribution of each parameter. Hence, a Markov chains are good tools to approximate the distribution of interest. Sampling from such a Markov chain after an adequate burn-in period yields good approximations of model parameters. In this study, the Metropolis algorithm and Gibbs sampling method are implemented by a specific Winbugs (1.4) program [14].

4.3 Model Comparison Criteria:

In order to compare these models, the DIC was computed for each model. DIC, which was proposed by Spiegelhalter et al. [15], is one of the best criterions for the comparison of Bayesian models [25]. Let θ be the vector of model parameters DIC defined by the expression DIC $=D(\theta)+p_D =2D(\theta)+p_D$, where $D(\theta)$ is the deviance of the model which evaluated at the posterior mean estimate .θ and $D(\theta)$ is the posterior mean of the deviance which is derived from the average of the logarithm of likelihood after the burn-in period and denote the goodness of fitness. Where $P_D = \overline{D(\theta)} - D(\bar{\theta})$ difference between the posterior mean of the deviance and the deviance of the posterior mean of the vector of parameters of interest, which represents the number of parameters effectiveness in the model, so it is an indicator of model complexity. Based on this measure, the model with a smallest DIC value is known to be the best one.

5. Application

5.1 Cutaneous Melanoma

We used the cutaneous melanoma data set that is available in the homepage of Ibrahim book (2001)[10]. This data set contains 427 patients for the evaluation of postoperative treatment with a high dose of interferon alfa-2beta in order to prevent recurrence in the period 1991 until 1998. 10 subjects were excluded because tumor thickness data were missing. The observed time (T) ranges from 0.15 to 7.01 years (3.18 ± 1.69). In this data set 55.6 percent of observation was censored. The most important covariate that is important and significant in several models was nodule category (1: n=82; 2: n=87; 3: n=137; 4: n=111). We considered this covariate as a categorical variable and defined 3 dummy variables to handle this covariate.

Table 2: Posteriors Summaries of the P,NB and GPL models

Model	parameter	Mean	SD	2.5 Percentile	97.5 Percentile
Poisson	τ	-1.62	0.132	-1.892	-1.369
	ν	1.71	0.112	1.502	1.93
	β_0	0.26	0.140	-0.021	0.532
	β_1	-1.11	0.213	-1.525	-0.715
	β_2	-0.83	0.188	-1.202	-0.452
	β_3	-0.55	0.205	-0.962	-0.141
Negative Binomial	τ	-2.45	0.427	-3.392	-1.73
	ν	2.03	0.202	1.668	2.452
	β_0	1.52	0.600	0.463	2.807
	β_1	-1.85	0.434	-2.764	-1.087
	β_2	-1.30	0.358	-2.036	-0.628
	β_3	-0.79	0.358	-1.495	-0.089
	η	2.02	0.910	0.445	3.942
Generalized Poisson-Lindely	τ	-2.14	0.274	-2.668	-1.657
	ν	1.89	0.144	1.620	2.181
	β_0	-0.77	0.656	-1.764	0.711
	β_1	1.23	0.230	0.777	1.688
	β_2	0.90	0.209	0.491	1.318
	β_3	0.59	0.228	0.144	1.034
	α	0. 7	0.762	0.015	2.843

For parameter estimation we proposed for scale and shape parameters of promotion time a normal prior with $\mu = 0 \ \ and \ \sigma^2 = 1$ and gamma prior with $\gamma_\nu = \eta_\nu = 0.1$. For shape parameter of the Generalized Poisson-Lindely, gamma prior with $\gamma_\alpha = \eta_\alpha = 1$.

The self write codes were written in WinBugs. The 50000 iterations were run and a sample was recorded every 10 iteration to reduction of autocorrelation within chain after 10,000 burn-ins. The results of this analysis of 3 models (Poisson, Negative Binomial and Generalized Lindely-Poisson distribution) have been shown in table 2.

The credible intervals for the $\beta_1, \beta_2, \beta_3$ does not include zero, so there is evidence that the cure rate is different in different categorical. To Compare these model we used the DIC criteria. According to this criterion, the best model should have fewer amounts. These criteria for the P, NB and GPL are 1036.9, 1029.9, and 1026.6; therefore the GPL is the best model. The percent of cure rate based on the categorical nodule parameter of these models have shown in table 3.

Table 3: Cure rate estimation based on the P,NB and GPL

Cure rate	P_0	P_1	P_2	P_3
Poisson	27.3	47.3	56.8	65.2
Negative Binomial	31.6	44.2	53.7	64.1
Generalized Poisson-Lindely	23.8	40.5	50.3	60.4

The figure 1 has showed the K-M estimates of the survival function and the Bayesian estimation of the survival function based on the different models. The best fitness of the GPL model has been emphasized.

5.2 Simulation

To assess the performance of our new model, we conducted a simulation study and generated a data set that was subsequently analyzed by fitting the model. We employed these steps for simulation as the following:

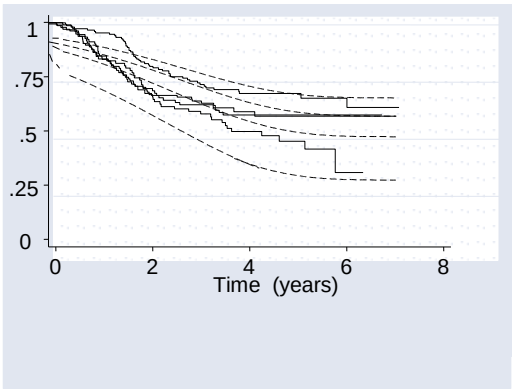
Step1, Generate a dummy variable from the Bernoulli distribution with p=0.5.

For $\beta_0 = -0.5 \; and \; \beta_1 = 1$ we have $\theta = \exp(\beta_0)$ when x=0 and $\theta = \exp(\beta_0 + \beta_1)$ when x=1 so that the cure rate in each group are 23.1 and 63.4 if $\alpha = 2$.

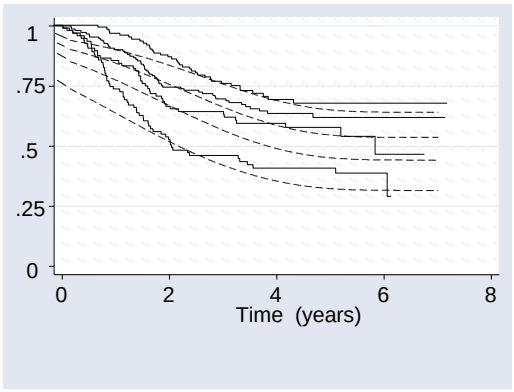
Step2, generate data from the GPL with parameters θ & α obtain from step1.

Step3, if N=0 so T=∞ otherwise for N=n generate $Z_1, Z_2, \ldots, Z_n$ from the Weibull distribution with parameters (-2,2) and take T=min{ $Z_1, Z_2, \ldots, Z_n$ }.

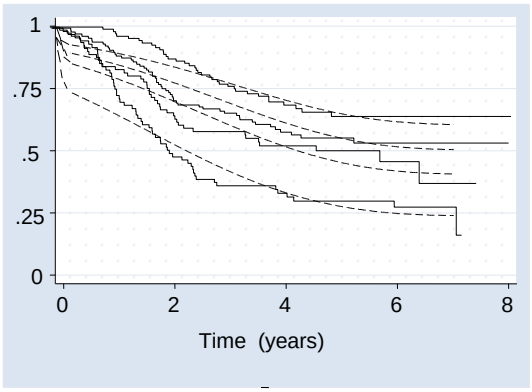
Figure 1



Poisson



Negative Binomial



Generalized Poisson Lindely

We generated 500 samples with 50 times repetitions and estimated the parameters. The simulation program has been written in R Package and then recorded them to Winbugs in order to obtain the parameters estimation. In Table 4, the posterior mean and standard deviation averaged was shown for each regression parameter. We can see that the posterior means of the parameters are quite close to the true values, indicating that the MCMC chains converged properly.

Table 4: Mean and Square root of the mean square error of the parameters estimation in the GPL with n=500

Parameter	Mean	SRMSE
τ	-2.09	0.137
ν	2.03	0.066
β_0	-0.57	0.101
β_1	0.97	0.071
α	2.02	0.380

6. Discussion:

In this paper we introduced another option instead of Negative Binomial distribution to overcome over-dispersion problem in promotion time cure model. This distribution is called generalized Lindely-Poisson distribution that introduce by Mahmoudi and Zakerzadeh (2010). Not only the variance of this distribution related to the shape parameter, but also the mean of its. This cause more flexible model to analyze complex data set.

This data set was analyzed by Cancho et al. (2011, 2012) which used the Negative Binomial and Conway-Maxwell poisson distribution. They mentioned that when using the nodule covariate like a categorical covariate, the DIC is increasing a little. Due to this fact that cure model is providing the cure rate, in this study we considered the covariates as the categorical variables to aim this purpose.

We proposed a new way to simulate cure rate data, that was different from the way of Cancho et. al (2011). In this method we used the latent mechanism in which, the initial model has produced from it.

As mentioned before, Generalized Lindely-Poisson reduces to the Lindely-Poisson when the $\alpha = 1$, since in table 1 the estimate of α isn't different from one 1, but when we used the Poisson-Lindely distribution, the DIC increased to 1032.3, so that we take the GPL to interpret this data.

Mahmoudi and Zakerzadeh (2010) have mentioned that the Generalized-Lindely distribution is a two component mixture gamma distribution is given by;

$$\frac{\theta}{1+\theta} \sim G(\alpha, \theta) + \frac{1}{1+\theta} \sim G(\alpha + 1, \theta)$$

Therefore it's not so amazing that the result of the NB and GPL is not so different.

References

1. J.W. Boag, 1949. *Maximum likelihood estimates of the proportion of patients cured by cancer therapy*, J. R. Stat. Soc. Ser.B 11, 15–44.
2. P. Borges, J. Rodrigues, and N. Balakrishnan, 2012. Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data, Comput. Stat. Data. Anal. 56, 1703–1713
3. J. Brekson and R.P. Gage, 1952. *Survival curve for cancer patients following treatment*, J. Am. Stat. Assoc. 47, 501–515.
4. V.G. Cancho, J. Rodrigues and M. de Castro, 2011. A flexible model for survival data with a cure rate: a Bayesian approach, Appl Stat. 38, 57–70
5. M.H. Chen, J.G. Ibrahim, and D. Sinha, 1999. A new Bayesian model for survival data with a surviving fraction, J. Am.Stat. Assoc. 94, 909–919.
6. S. Chib and E. Greenberg, 1995. Understanding the Metropolis-Hasting Algorithm. Am. Stat. 49, 327-335.
7. F. Cooner, S. Banerjee, B.P. Carlin, and D. Sinha, 2007. Flexible cure rate modeling under latent activation schemes, J. Amer. Statist. Assoc. 102, 560–572.

8. M.E. Ghitany, D.K. Al-Mutari, and S. Nadarajah, 2008, Zero-truncated Poisson-Lindley distribution and its application, Math. Comput. Simul. 79, 279–287.
9. M.E. Ghitany, and D.K. Al-Mutairi, 2009. Estimation methods for the discrete Poisson-Lindley distribution. J. Stat. Comput. Sim .79, 1–9.
10. J.G. Ibrahim, M.H. Chen, and D. Sinha, 2001. Bayesian Survival Analysis, Springer, New York.
11. E. Mahmudi, and H. Zakerzadeh, 2010. Generalized Poisson–Lindley Distribution. Commun. Statist-Theor. Method. 39, 1785–1798
12. R Development Core Team, 2009. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.
13. J. Rodrigues, M. Castrode, V.G. Cancho, and N. Balakrishnan, 2009. COM–Poisson cure rate survival models and an application to a cutaneous melanoma data, J. Statist. Plann. Inference 139, 3605–3611.
14. D.J. Spiegelhalter, A. Thomas, N. Best, and D. Lunn, (2003), WinBUGS User Manual, Version 1.4, MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK, available at: http://www.mrc-bsu.cam.ac.uk/bugs
15. D.J. Spiegelhalter, N.G. Best, B.P Carlin, and V. der Linde, 2002, A Bayesian measures of model complexity and fit, J. R. Statist. Soc. Ser. B 64, 583–639.
16. A.D. Tsodikov, J.G. Ibrahim, and A.Y. Yakovlev, 2003. Estimating cure rates from survival data: An alternative to two-component mixture models, J. Amer. Statist. Assoc. 98, 1063–1078.
17. A.Y. Yakovlev and A.D. Tsodikov, 1996. Stochastic Models of Tumor Latency and Their Biostatistical Applications,World Scientific, Singapore, Hackensack, NJ.
18. H. Zakerzadeh and A. Dolati, 2010. Generalized Lindley distribution, J. Math. Exten, 3, 13-25.