

ON THE ESTIMATION OF CURE FRACTION USING POWER GOMPERTZ DISTRIBUTION UNDER BAYESIAN APPROACH

KOMAL GOEL, MANOJ KUMAR VARSHNEY, GURPRIT GROVER, AND SEEMA PANT

ABSTRACT. The cure rate models have demonstrated, beyond doubt, their utilitarian value for analysis of data pertaining to long term survivors in diseases like cancer, HIV et al. In the present paper, we have estimated cure fraction using Power Gompertz distribution, in the presence of covariates and censoring under Bayesian framework. Using the developed model, Bayesian analysis of a data set related to patients with breast cancer has been done. The standard MCMC techniques in OpenBUGS Software have been used to analyze the data.

1. Introduction

The standard survival models that are used for analyzing survival time data do not take into account the cure proportion. These methods are based on the assumption that all the patients have the same level of susceptibility to the disease and they all experience the event of interest namely death. But in diseases like cancer, due to the latest advancement in treatment procedures and new medicines, some patients get cured of the disease. They are termed as the long time survivors or immunes and constitute cure fraction. The cured proportion does not experience death during the study period. The patients who do not get cured are called susceptible.

Thus the population can be viewed as consisting of cured (immunes) and uncured patients (susceptible). Kaplan Meier (KM) plot of survival function helps us in deciding if the data has a cured proportion or not. If the KM plot shows a long plateau on the right, then it is indicative of long time survivors.

To analyze the data related to diseases with long time survivors or cure fraction, cure rate models are used. The cure fraction models are mainly classified into two categories namely Mixture and Non Mixture cure rate models. Boag [2] introduced mixture cure rate model. A mixture cure rate model has two components, one each to account for cured and uncured patients. A non-mixture model gives an asymptote for the cumulative hazard and hence for the cure fraction.

Several approaches based on various distributions like Gamma, Weibull and many others, have been proposed to estimate the cure fraction. Achcar et al. [1], Martinez et al. [11] considered two parameter and four parameter generalized modified Weibull distributions respectively. Yamaguchi [14], Yu et al. [15] and

2000 *Mathematics Subject Classification.*

Key words and phrases. Cure fraction models, immunes, OpenBUGS, Power Gompertz Distribution, Susceptible.

several others used different latency distributions to estimate the cure fraction. Kannan et al. [10] considered a cure rate model, with Generalized exponential distribution (with covariates) as latency distribution, to estimate the cure fraction.

One of the important distributions that is being used in survival analysis is Gompertz distribution. It is a generalization of exponential distribution. Chien-Lin Su et al. [4] carried out analysis of survival data with cure proportion by assuming two parameter Gompertz distribution as base line survival function. This was improved upon by Grover et al. [6] for estimating cure fraction assuming generalized Gompertz distribution for latency distribution.

Gompertz distribution is a continuous distribution. It can be skewed both to the left and right. Adding parameters to a distribution makes it more flexible for modeling heavily skewed data. Different extensions to this classical distribution have been proposed, in literature, in order to increase skewness and flexibility. To name a few, Beta Gompertz distribution (Jafaril et al. [9]), Generalized Gompertz Distribution (Gohary et al. [7]), Generalized Exponential-Gompertz (El-Damcese [5]). One such extension of the Gompertz distribution has been proposed by Ieren et al.[8]. They have proposed power transformation approach to obtain Power Gompertz distribution. They have shown that Power Gompertz distribution is a better model as compared to Gompertz distribution for analyzing survival time data.

In this paper we introduce Power Gompertz Distribution as the baseline survival function while estimating cure fraction using cure rate models. We have first compared two-parameter Gompertz distribution, Generalized Gompertz distribution and Power Gompertz Distribution, using Deviance Information Criterion, under Bayesian Setup. Since Power Gompertz Distribution has the minimum value of DIC, we introduce Power Gompertz Distribution as the baseline survival function to estimate cure proportion using mixture and non-mixture cure rate models. To estimate cure fraction in the presence of covariates we have considered mixture cure rate model as it has smaller DIC value as compared to non-mixture model.

The remaining paper has been organized as: Section 2, named as Material and Methods, includes cure rate models, method to compare models, likelihood equations, priors used for Bayesian Analysis of the data considered. It also includes pdf of Power Gompertz Distribution, its survival function etc. In Section 3 we have tabulated and have interpreted the results. Section 4 comprises of discussion which is followed by the references.

2. Material and Methods

2.1. Mixture Cure Rate Model. If p represents the proportion of immunes with respect to the event of interest and $S_0(t)$ is the baseline survival function for the patients who remain uncured then under mixture cure rate model, the survival function is given as

$$S(t) = p + (1 - p)S_0(t),$$

$$F(t) = 1 - S(t) \text{ is the cumulative distribution function and}$$

$$f(t) = (1 - p)f_0(t)$$

is the pdf of the life time T , $f_0(t)$ being the baseline density function for the susceptible.

Let t_i be the survival time of i th patient, d_i be a censoring indicator variable, $i = 1, 2, 3, \dots$ and

$$d_i = \begin{cases} 1 & \text{for uncensored lifetime} \\ 0 & \text{for censored lifetime} \end{cases}$$

Then the contribution of i th ($i = 1, 2, \dots, n$) patient to the likelihood function, denoted by L_i is given by

$$L_i = [f(t_i)]^{d_i} [S(t_i)]^{1-d_i}.$$

2.2. Non Mixture Cure Fraction Model. The survival function under non mixture model is expressed as

$$S(t) = p^{1-S_0(t)} = \exp[\ln(p)(1 - S_0(t))].$$

Under this model, the contribution of i th patient to L_i is given by,

$$L_i = [h(t_i)]^{d_i} S(t_i),$$

where $h(t) = -(\ln p)f_0(t)$ is the hazard function.

2.3. Model Comparison Criteria. The Deviance Information Criterion (DIC) has been used for (i) identifying the best distribution among two parameter Gompertz, three parameter Gompertz and Power Gompertz distribution and (ii) for comparison of mixture and non-mixture models (the best distribution in (i) will be taken as the latency distribution). Spiegelhalter [12] had introduced DIC for comparing a set of Bayesian hierarchical models. The distribution/model with smallest value of DIC is considered to be the best fitted distribution/model.

DIC is given by:

$$DIC = \bar{D} + P_D = \hat{D} + 2P_D,$$

where $P_D = \bar{D} - \hat{D}$ is the effective number of parameters in the model, \bar{D} is the posterior mean of deviance and \hat{D} is the deviance calculated at the posterior means.

Based on *DIC* value PGD is the best distribution (Table 2).

2.4. Power Gompertz Distribution. The density function of the Power Gompertz distribution with α (scale), β (shape) and θ (power) as parameters is given as:

$$f(x) = \alpha \theta x^{\theta-1} e^{\beta x^\theta} e^{-\frac{\alpha}{\beta}(e^{\beta x^\theta} - 1)}, \quad x > 0, \alpha > 0, \beta > 0, \theta > 0.$$

This distribution, obtained by applying power transformation on the traditional Gompertz distribution, is more skewed and flexible. The distribution is positively skewed and has increasing failure rate.

Now for Power Gompertz Distribution, we have

$$\begin{aligned} S_0(t) &= \exp \left[-\frac{\alpha}{\beta} (e^{\beta t^\theta} - 1) \right], \\ F_0(t) &= 1 - S_0(t), \\ f_0(t) &= \alpha \theta t^{\theta-1} e^{\beta t^\theta} e^{-\frac{\alpha}{\beta} (e^{\beta t^\theta} - 1)}. \end{aligned}$$

The log likelihood function for PGD under mixture model is given by

$$\begin{aligned}
 l(\gamma) &= [\log(1-p) + \log \alpha + \log \theta] \sum_i \delta_i + (\theta - 1) \sum_i \delta_i \log t_i \\
 &+ \beta \sum_i \delta_i t_i^\theta - \frac{\alpha}{\beta} \left[\sum_i \delta_i e^{\beta t_i^\theta} - \sum_i \delta_i \right] \\
 &+ \sum_i \log(1 - \delta_i) \left[p + (1-p) \exp \left(-\frac{\alpha}{\beta} (e^{\beta t_i^\theta} - 1) \right) \right].
 \end{aligned}$$

The log likelihood function for PGD under non-mixture model can be expressed as

$$\begin{aligned}
 l(\gamma) &= \{\log(-\log p) + \log \alpha + \log \theta\} \sum_i \delta_i + (\theta - 1) \sum_i \delta_i \log t_i \\
 &+ \beta \sum_i \delta_i t_i^\theta - \frac{\alpha}{\beta} \left(\sum_i \exp(\beta t_i^\theta \delta_i) - \sum_i \delta_i \right) \\
 &+ \log p \sum_i \exp \left(-\frac{\alpha}{\beta} (e^{\beta t_i^\theta} - 1) \right).
 \end{aligned}$$

Let x_1, x_2, \dots, x_l be l covariates influencing the parameters, then the cure proportion p can be expressed as:

$$\log \left(\frac{p_i}{1-p_i} \right) = a_0 + a_1 x_{1i} + a_2 x_{2i} + \dots + a_l x_{li}.$$

The Bayesian estimates are obtained using OpenBUGS software. One of the important advantage of using OpenBUGS software is that it requires only the prior distribution and the distribution of the survival data. Throughout the analysis we have used Beta(1, 1) distribution as the prior for cure fraction p and Gamma(1, 1) for other parameters. The prior distribution to find posterior estimates of the parameters of the regression model is assumed to be $N(0, 100)$.

Credible Interval has been used to find the significant prognostic factors.

3. Results

We have used the methodology discussed above to estimate cure fraction on the basis of a data set of 85 patients suffering from breast cancer (period of study is January, 2009 to December, 2010). About 17.6% patients were censored during follow up time. The median age of patients at the time of diagnosis is 49 years. Table 1 shows a summary of the breast cancer data.

TABLE 1. Descriptive characteristics of the data

Covariate	Frequency	Mean	Std. Dev.	Min	Max
Age	85	50.09	12.82	25	85
Tumor Size	85	3.72	1.62	1	8
Nodal Metastasis	85	4.36	4.70	0	15
Tumor Grade	85	1.96	0.71	1	3
NPI	85	4.81	1.34	2.12	7.6
CA-15	85	32.08	6.16	15.2	46

The Kaplan-Meier plot for overall survival (Figure 1) shows a “plateau” on the right. A plateau on the right indicates that there is a cure fraction among the patients. This suggest that a cure rate model rather than traditional methods of analyzing life time data is more appropriate in this case.

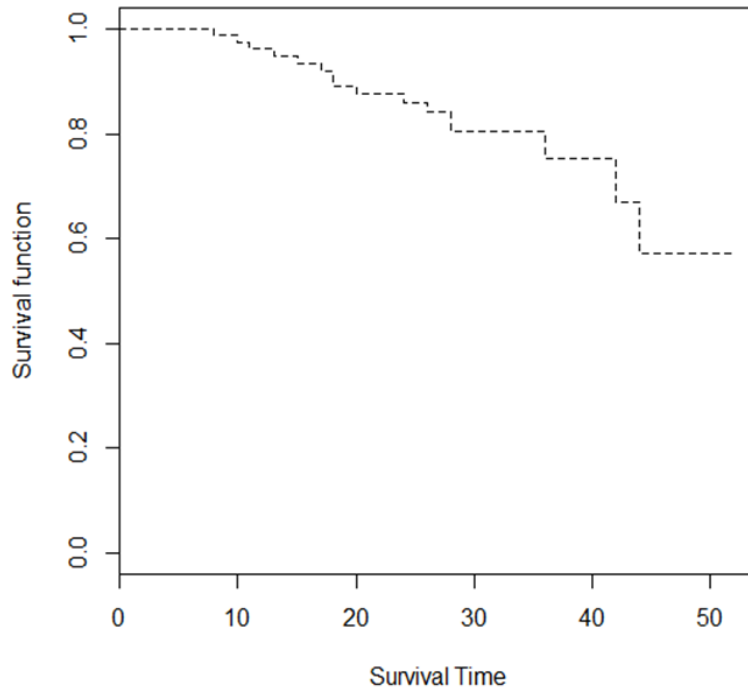


FIGURE 1. Kaplan-Meier Survival Curve indicates suitability of cure rate models

We first obtained the posterior summaries (Table 2) under traditional Gompertz, Generalized Gompertz and Power Gompertz distributions. The DIC values in Table 2 indicate that Power Gompertz Distribution is the best among the three distributions as it has the smallest DIC.

TABLE 2. Posterior Summaries without including p

Model	Parameter	Posterior Mean (SD)	95% credible interval	DIC
Gompertz distribution (2-parameters)	λ	0.0689(0.0091)	(0.0515,0.086)	565.7
	θ	0.1314(0.053)	0.0022,0.1219)	
Generalized Gompertz (3-parameters)	λ	0.058(0.025)	(0.0152,0.0834)	561.4
	θ	3.947(1.477)	(0.1385,3.897)	
	c	0.0165(0.0182)	(0.0004,0.056)	
Power Gompertz	α	0.075(0.0037)	(0.0002,0.0073)	533
	β	0.6759(0.4096)	(0.0412,0.5168)	
	θ	0.585(0.1009)	(0.3892,0.7208)	

Next we fit the cure rate model on the best fitted distribution (Power Gompertz Distribution) on our data. Initially, we consider the cure fraction models without covariates. Table 3 shows the posterior estimates of parameters based on Power Gompertz Distribution under cure rate models.

TABLE 3. Posterior Estimates including the cure fraction p (without the covariates)

Model	Parameter	Posterior Mean (SD)	95% credible Interval	DIC
Mixture Cure Model	α	0.0082(0.0028)	(0.0002,0.0077)	542
	β	0.08198(0.0497)	(0.0115,0.1473)	
	θ	1.025(0.1733)	(0.8414,1.343)	
	p	0.0181(0.0203)	(0.0006,0.0863)	
Non Mixture Cure Model	α	0.0017(0.0006)	(0.0006,0.0032)	562.8
	β	0.3522(0.0245)	(0.3172,0.4163)	
	θ	0.7209(0.02258)	(0.671,0.7616)	
	p	0.0262(0.0203)	(0.0034,0.0830)	

On the basis of the values of the parameters (Table 3), we can say that in the absence of the covariates the mixture and non-mixture approach under Power Gompertz Distribution fits well to the data. It is evident from the results in Table 3 that the cure fraction is significant under both the models. The DIC values indicate that mixture model (DIC= 542) is better as compared to the non-mixture cure model (562.8). Therefore to estimate cure fraction in the presence of covariates (Table 4) we have used the mixture model under Bayesian approach. We have considered Age, Tumor size, Tumor grade, NPI, Nodal metastasis and CA-15 as the covariates.

TABLE 4. Posterior summaries including both the cure proportion p and the covariates

Model	Parameter	Posterior	SD	95% credible interval
		Mean		
Mixture Cure Model (Power Gompertz Distribution)	Intercept	0.1276	0.137	(-0.091,0.375)
	Age	-0.0562	0.045	(-0.135,0.023)
	Tumor Size	0.0671	0.024	(0.017,0.107)
	Tumor grade	-0.0767	0.029	(-0.144,-0.033)
	NPI	-0.0106	0.044	(-0.086,0.070)
	Nodal M	0.1275	0.034	(0.0398,0.0197)
	CA-15	-0.0248	0.055	(-0.341,-0.163)
	α	0.0104	0.003	(0.005,0.017)
	β	0.1952	0.087	(0.058,0.317)
	θ	0.81	0.105	(0.673,1.03)

Table 4 gives the posterior summaries in the presence of covariates for mixture cure rate model with Power Gompertz Distribution. The values of 95% Credible Interval for Tumor size, Tumor grade, Nodal metastasis and CA-15 does not include zero suggesting that they are the significant prognostic factors under cure rate model, which match with the study of Grover et al. [13] (2016). This implies that these covariates effect the survival and cure probability significantly.

The objective of this paper is to analyze the importance of Power Gompertz Distribution (PGD) under mixture and non-mixture cure rate models. The Cure Rate models are an important tool to analyze the survival data having a proportion of cured patients (cure fraction). A proportion of patients get cured after taking the treatment and they would not experience the event of interest. Our objective here is in estimating this proportion among patients of breast cancer using cure rate models. We are also interested in finding out the significant prognostic factors of breast cancer. The covariates that we have considered are Age, Tumor Grade, Tumor Size, NPI, Nodal metastasis and CA-15. Earlier these models have been used to estimate cure fraction assuming traditional Gompertz Curve and Generalized Gompertz Curve as susceptible distribution. In this paper we have proposed PGD as the latency distribution. This distribution is an extension of Gompertz Curve developed by using power transformation approach. It is positively skewed. Its hazard function has increasing failure rate and has been proved to give better results. On the basis of DIC value we could infer that PGD has smaller DIC value as compared totraditional Gompertz and Generalized Gompertz distributions and hence it is the best model. Taking PGD as the distribution of susceptible we have developed Mixture and Non-Mixture Cure rate models .The DIC of mixture model is smaller. Therefore to study the effect of covariates we have considered mixture model. Among all the covariates that we have considered, Tumor size,

Tumor grade, Nodal metastasis and CA-15 have been found to be the significant prognostic factors.

4. Discussion

In this research article efforts have been made to highlight the essence and benefit of Power Gompertz Distribution (PGD) under Mixture and Non-Mixture cure rate models. Traditional survival models do not take into account the fact that with the latest advanced treatments, in chronic diseases like cancer and AIDS, not all the patients meet the event of interest (death). Hence the population consists of two types of patients: immunes or long term survivors and susceptible or uncured patients. To deal with such data, special models called cure rate models have been developed. In this article we have used the mixture and non-mixture cure rate models to analyze a breast cancer data set with long time survivors. Here we assume the baseline survival function of cured individuals to follow PGD and try to estimate cure rate under Mixture and Non Mixture cure models both in presence and absence of covariates. The cure fraction is found to be significant under both the models. The covariate tumor size, tumor grade, nodal metastasis and CA-15 are found to be significant for cancer patients. Other directions of model extension by adding some more parameters may be worthy for further investigation.

References

- [1] Achcar JA, Coelho-Barros, E and Mazucheli, J: Cure fraction models using mixture and non-mixture models, *Tatra Mountains Mathematical Publications* **51** (2012) 1–9.
- [2] Berkson J and Gage RP: Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47** (1952) 501–515.
- [3] Boag JW: Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society B* **11** (1949) 15–53.
- [4] Chien-Lin S, Adam Ding A and Wang W: Analysis of clustered survival data in the presence of cure with the Gompertz distribution model, (2013), [http://www.wseas.us/e-library/conferences/2013/Cambridge USA/ MATHECO/MATHECO-26.PDF](http://www.wseas.us/e-library/conferences/2013/Cambridge%20USA/MATHECO/MATHECO-26.PDF)
- [5] El-Damcese MA, Mustafa A, El-Desouky BS and Mustafa ME: The odd generalized exponential Gompertz distribution, *Applied Mathematics* **6** (2015) 2340–2353.
- [6] Grover G, Goel K and Swain PK: Mixture and non-mixture cure fraction models based on generalized Gompertz distribution under Bayesian approach, *Tatra Mountains Mathematical Publications* (2016) 121–135.
- [7] Gohary E, Alshamrani A and Al-Otaibi: The generalized Gompertz distribution, *Applied Mathematical Modelling* **37** (2013) 13–24.
- [8] Ieren TG, Kromtit FM, Agbor BU, Eraikhuemen IB and Koleoso PO: A power Gompertz distribution: Model, properties and application to bladder cancer data, *Asian Research Journal of Mathematics* **15**(2) 2019 1–14, Article no. ARJOM.52035.
- [9] Jafaril AA, Tahmasebi S and Alizadeh M: The Beta-Gompertz distribution, *Revista Colombiana de Estadística Junio* **37**(1) (2014) 139–156.
- [10] Kannan N, Kundu D, Nair P and Tripathi R C: The generalized exponential cure rate model with covariates, *Journal of Applied Statistics* **37** (2010) 1625–1636.
- [11] Martinez EZ, Achcar JA, Jacomeb AA and Santos SJ: Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data, *Computer Methods and Programs in Biomedicine* **112** (2012) 343–355.
- [12] Spiegelhalter DJ, Best NG, Carlin BP and Linde V: A Bayesian measures of model complexity and fit, *J. R. Statist. Soc. Ser. B* **64** (2002) 583–639.

- [13] Swain PK, Grover G, Chakravorty S, Goel K and Singh V: Estimation of number of involved lymph nodes in breast cancer patients using bayesian regression approach, *Journal of Statistics Applications and Probability Letters* **4** (2016) 17–25.
- [14] Yamaguchi K: Accelerated failure-time regression model with a regression model for the surviving fraction: an application to the analysis of ‘permanent employment’ in Japan, *Journal of the American Statistical Association* **87** (1992) 284–292.
- [15] Yu B, Tiwari, RC, Cronin KA and Feuer, EJ: Cure fraction estimation from the mixture cure models for grouped survival data, *Stat. Med.* **23** (2002) 1733–1747.

KOMAL GOEL: DEPARTMENT OF STATISTICS, MATA SUNDRI COLLEGE FOR WOMEN, UNIVERSITY OF DELHI, DELHI, INDIA

Email address: komalstats@gmail.com

MANOJ KUMAR VARSHNEY: DEPARTMENT OF STATISTICS, HINDU COLLEGE, UNIVERSITY OF DELHI, DELHI, INDIA

Email address: mkvarshney.du@yahoo.com

GURPRIT GROVER: DEPARTMENT OF STATISTICS, FACULTY OF MATHEMATICAL SCIENCES, UNIVERSITY OF DELHI, DELHI, INDIA

Email address: gurpritgrover@yahoo.com

SEEMA PANT: DEPARTMENT OF STATISTICS, RAM LAL ANAND COLLEGE, UNIVERSITY OF DELHI, DELHI, INDIA

Email address: joshi.seema1224@gmail.com