Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 4, pp. 833–842
DOI : 10.47974/JSMS-933

Cure rate and statistical inference for the Marshall-Olkin generalized defective Gompertz distribution under censored survival data

Tasnime Hamdeni *
Department of Applied Mathematics
ENIT, University of Tunis El-Manar
Laboratoire Signal
Image et Maitrise de l'Energie (LR13ES03)

And

Departement of Electrical Engineering
Ecole Nationale Supérieure des Ingénieurs de Tunis
5 Avenue Taha Hussein
Montfleury, 1008
Tunis
Tunisia

Soufiane Gasmi †
Department of Mathematics
ENSIT, University of Tunis
Laboratoire Modélisation Mathématique
Analyse Harmonique et théorie de potentiel (LR18ES09)
Tunis
Tunisia

Abstract

Cure fraction modeling usually requires the addition of an extra parameter. Defective modeling, alternatively, allows cure fraction modeling with less complexity. A new extension of the defective version of Gompertz distribution is studied in this paper. This distribution is compared to other well-founded distributions and its distinguished performance to fit censored survival data with and without a cure fraction is presented. A computational study is carried out to highlight the correlation between the censorship level and the proportion of long-term survivors. Interesting mathematical properties are established. The practical relevance of the model is illustrated by applying various real and simulated datasets.

Subject Classification: *62F03, 62F10, 62P10.*
Keywords: *Cure fraction, Defective modeling, Gompertz distribution, Marshall-Olkin, Survival analysis.*

*E-mail: hamdeni.tasnime@gmail.com
† E-mail: soufiane.kasmi3@gmail.com

1. Introduction

Survival data are durations that measure follow-up time from a determined starting point to the occurrence of an event of interest. An example of a survival dataset is the time from the first diagnosis of a disease to death or recurrence of the same disease of several patients. Nevertheless, the event of interest is not always observed [4]. This may be because of two reasons. The first reason is the loss of information which is referred to as censorship or truncation. The second reason is the fact that the event of interest has not occurred yet. That is to say that there exists a cured proportion in the population under study. The cured proportion within the population is also known as the surviving fraction, cure fraction, long-term survivors and immunes proportion. The cure fraction is modeled mathematically in different ways. The first approach to model the cure fraction is by adding an extra parameter to a well-founded distribution. There are two methods to do so. The first method is known as mixture modeling; initiated in 1949 by Boag [1], standard mixture models are commonly used with the adjusted survival function (1):

$$S_1(t) = \lambda + (1-\lambda)S(t), \tag{1}$$

where $S(t)$ is the proper survival function and λ is a parameter in [0, 1] that represents the cure rate. $S(t)$ tends to zero as time t tends to infinity and can be replaced by any proper distribution such as the Exponential distribution, Weibull distribution, and Gompertz distribution, as well as their modified and generalized versions. The mixture model $S_1(t)$ converges to λ as the time approaches infinity [5]. Mixture models were put forward to investigate the heterogeneity between cured and uncured patients.

The second method of adding an extra parameter to a well-founded model to describe a cure fraction is a non-mixture model. The latter describes an asymptote for the cumulative hazard and consequently for the cure proportion, [10]. The survival function of the non-mixture cure rate model is then given by equation (2):

$$S_2(t) = \lambda^{1-S(t)}. \tag{2}$$

There is another approach to model the cure fraction that does not require the addition of an extra parameter. In this approach cure fraction is described simply by modifying the domain of the parameters of a well-founded distribution in such a way that its survival function converges to $\lambda \in]0,1[$. This approach is known as defective modeling. Being an

alternative to (1) and (2) with reduced complexity, several defective models have been published recently. To name just a few, Cantor and Shuster [3] applied the defective Gompertz distribution (and named it the modified Gompertz) on censored survival data of leukemia pediatric patients. Rocha, Nadarajah, Tomazella and Louzada [14] estimated cure fraction using both inverse Gaussian and defective Gompertz distributions, etc.

This work consists of a two-pronged research agenda. The first major line is about the Marshall-Olkin Generalized Defective Gompertz Distribution [8, 9], its properties, its flexibility, and its performance on censored data. The second line is about a computational study resulting in interesting observations about the correlation between censoring level and cured proportion and its manifestation in simulated and real data.

This paper is organized as follows: Section 2 is reserved for the introduction of the Marshall-Olkin Generalized Defective Gompertz Distribution and its properties. Section 3 is devoted to statistical inference with simulated and real type-I right-censored datasets. Finally, we conclude in Section 4.

2. The Marshall-Olkin Generalized Defective Gompertz Distribution (MO-GDGD)

Combining the works of Cantor and Shuster [3], El-Gohary, Alshamrani, and Al-Otaibi [6], and Marshall and Olkin [11], the Marshall-Olkin Generalized Defective Gompertz Distribution (MO-GDGD) is specified by the probability density function (PDF) in equation (3):

$$f_{MO}(t;r,\alpha,\beta,\gamma)=\frac{r\gamma\alpha e^{-\beta t}e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma-1}}{(1-(1-r)(1-(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma}))^2}, \tag{3}$$

where $t>0$, $\alpha>0$ is a shape parameter, $r>0$ is the Marshall-Olkin tilt parameter as introduced in Marshall and Olkin [11], $\beta\in R^*$ is a scale parameter and an indicator of defectiveness in the model (if $\beta<0$ then the distribution is proper and if $\beta>0$ then the distribution is defective), $\gamma>0$ is a shape parameter used, as in El-Gohary, Alshamrani and Al-Otaibi, [6] to exponentiate the Gompertz distribution. The probability density function of defective models integrates to values less than 1. The cumulative density function of the MO-GDGD is given by equation (4).

$$F_{MO}(t;r,\alpha,\beta,\gamma)=\frac{(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma}}{(1-r)(1-e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)})^{\gamma}+r}. \tag{4}$$

For scenarios where $\beta>0$, the cure fraction of the MO-GDGD that is the limit of the survival function $S_{MO}(t;r,\alpha,\beta,\gamma)=1-F_{MO}(t;r,\alpha,\beta,\gamma)$ is given by equation (5):

$$\theta_{MO}=\frac{r(1-(1-e^{-\frac{\alpha}{\beta}})^{\gamma})}{1-(1-r)(1-(1-e^{-\frac{\alpha}{\beta}})^{\gamma})}. \tag{5}$$

Among the submodels of the MO-GDGD, we cite the distribution proposed in 1825 by Gompertz [7], the modified Gompertz distribution (MGD) introduced in Cantor and Shuster [3], the modified generalized Gompertz distribution (MGGD) simultaneously in 2017 with Borges [2] and Martinez and Achcar [12], Marshall-Olkin Gompertz distribution (MO-GD) in the same year by Rocha, Nadarajah, Tomazella, and Louzada [14], and in 2020, the generalized defective Gompertz distribution (GDGD) [8].

3. Statistical inference

This section aims to infer the MO-GDGD distribution parameters for a given sample. For this purpose, we choose to use a common statistical method called the maximum likelihood.

Considering a set of survival data of the form $(t_i,\ \delta_i)$ where i an integer in $[1,n]$, n the number of subjects in the study, t_i the independently observed lifetime of the i^{th} patient, and δ_i the censoring indicator that is equal to 1 if the patient has experienced the event of interest (that is death or recurrence of the disease) and 0 if the information is censored (patient is still alive, left the study, died from a reason irrelevant to the study...). Lifetime of the i^{th} patient is censored at time T_i. If all T_i are equal, i.e. censoring time is fixed, then $T_i=T\quad\forall i=1,...,n$. The likelihood function has then the form of equation (6):

$$L(t_i;\Theta)=\prod_{i=1}^{n}f_{MO}(t;\Theta)^{\delta_i}S_{MO}(t;\Theta)^{1-\delta_i}. \tag{6}$$

We apply the logorithm function to equation (6), we obtain log-likelihood function (7):

$$l = \ln L(t_i;\Theta) = \ln(r)\sum_{i=1}^{n}\delta_i + \ln(\alpha)\sum_{i=1}^{n}\delta_i + \ln(\gamma)\sum_{i=1}^{n}\delta_i - \beta\sum_{i=1}^{n}\delta_i t_i + \frac{\alpha}{\beta}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i} - 1)$$
$$+(\gamma - 1)\sum_{i=1}^{n}\delta_i \ln(B(t_i;\Theta)) - 2\sum_{i=1}^{n}\delta_i \ln(C(t_i;\Theta)) + \ln(r)(n - \sum_{i=1}^{n}\delta_i)$$
$$+\sum_{i=1}^{n}(1-\delta_i)\ln(D(t_i)), \qquad (7)$$

To find maximum likelihood estimates, we set the first partial derivatives of the log-likelihood function (7) to zero and numerically solve the system of equations. The major weakness of this method is its sensitivity to the starting values. Nonetheless, suitable starting values can contribute to the achievement of the global optimum approximation.

3.1 *Correlation study between cure fraction and censorship level*

To visualize the mutual relationship between the percentage of incomplete observations and the surviving proportion, we simulate a random sample of size 100 derived from the cumulative density function

Table 1

Correlation study between the estimated cured proportion and the censorship level in the simulated data.

CL	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\theta}_{MO}$
0%	2.1315	0.1866	0.0371	0.4877	0.0068
5%	2.1434	0.1739	0.0530	0.4903	0.0391
10%	1.8170	0.1635	0.0631	0.4921	0.0663
15%	1.7069	0.1592	0.0668	0.4910	0.0767
20%	1.8132	0.1526	0.0676	0.4903	0.0917
25%	1.8275	0.1447	0.0698	0.4912	0.1109
30%	1.8724	0.1338	0.0710	0.4855	0.1349
35%	1.7899	0.1217	0.0660	0.5146	0.1423
40%	2.0064	0.1253	0.0673	0.5430	0.1616
45%	1.7172	0.1304	0.0823	0.5566	0.1896
50%	1.8404	0.1035	0.0993	0.5011	0.3094
60%	1.5782	0.0827	0.1390	0.5091	0.4432
70%	1.6582	0.0839	0.1934	0.5190	0.5440
80%	2.3781	0.0875	0.2289	0.5243	0.6622
90%	4.627	0.0816	0.2532	0.5345	0.8222

of the MO-GDGD distribution . Estimated parameters, as well as the estimated cure fraction for each censorship level (CL), are presented in table 1. From the obtained results, we can see that the censorship level and the cure fraction are highly correlated.

3.2 *Pediatric cancer data analysis*

As an illustration of what preceded, we devote this section to real data analysis. For datasets analysis, 1 – parameters of the MO-GDGD, as well as its defective submodels, are estimated using the maximum likelihood approach (see table 3), 2 – efficiency of the MO-GDGD to fit the current datasets is assessed using the parametric likelihood ratio (Γ_{H_0} test statistics) and its associated *P*-value (see table 3), 3 – profiles of the log-likelihood function of each of the model parameters are displayed, 4 – Akaike Information Criterion (AIC) [13] and the Corrected Akaike Information Criterion (AICc) are used the prove the applicability and the superiority of the MO-GDGD in comparison with its defective submodels (see tables 4), 5 – Kaplan-Meier curves of the survival function, as well as its parametric estimation and its lower and upper bounds, are presented.

A set of remission durations in days of a randomized trial in pediatric cancer is studied. The set is composed of 41 observations following the same type of treatment [3]. Only 5 out of 41 observations are complete. Complete observations are highlighted in bold (See table 2).

The Kaplan-Meier estimation of the survival function, as well as the parametric survival curve, are displayed in figure 1. Censored observations, lower and upper bounds are also displayed in the same figure. The estimated survival function converges to 0.8724, which represents the cured proportion as indicated in the figure. The obtained cure fraction matches the cure fraction value calculated using equation (5) after replacing the parameters (r,α,β,γ) by their estimations $(\hat{r},\hat{\alpha},\hat{\beta},\hat{\gamma})$. The superposition of the Kaplan-Meier curve and the estimated survival

Table 2

Duration of recovery of pediatric cancer following the same treatment

327	**385**	**410**	722	**750**	878	890	942	1161	**1215**
1241	1270	1418	1602	1610	1689	1848	1940	2016	2044
2078	2231	2289	2303	2420	2430	2510	2908	2932	2978
2993	3013	3017	3053	3083	3090	3219	3232	3321	3399
3416									

function curve indicate that the estimation process yields satisfactory results.

It should be noted that, for pediatric cancer data, the estimation process yields a value of the Marshall-Olkin tilt parameter r close to 1.

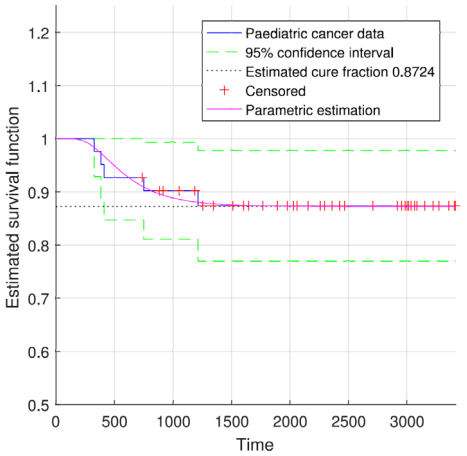

Figure 1

Kaplan-Meier and parametric estimation of the survival function for pediatric cancer data.

Table 3

Maximum Likelihood Estimations of the unknown parameters of MO-GDGD and some of its special cases and their associated P-values and the likelihood ratio Γ_{H_0} values for pediatric cancer data.

Model	$\hat{r}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	***P*-value**	Γ_{H_0}
MGD (α,β)	--	0.0002	0.0012	--	0.0446	4.0342
$MO-MGD$ (r,α,β)	55.2482	0.0035	0.0016	--	0.0724	3.2272
$GDGD$ (α,β,γ)	--	0.0007	0.0018	1.7296	0.1294	2.2994
$MO-GDGD$ (r,α,β,γ)	1.0031	0.0067	0.0035	12.8901	--	--

Table 4

The null hypothesis H_0, the value of the log-likelihood function l under H_0 and the information criteria AIC and AICc for pediatric cancer data.

Model	H_0	l	AIC	AICc
MGD (α,β)	$r=\gamma=1,$ $\beta>0$	-52.0495	108.0990	105.3917
$MO-MGD$ (r,α,β)	$\gamma=1,\ \ \beta>0$	-51.6460	109.2920	105.8774
$GDGD$ (α,β,γ)	$r=1,$ $\beta>0$	-51.1821	108.3642	104.9496
$MO-GDGD$ (r,α,β,γ)	--	-50.0324	108.0648	104.0404

This means that if we set $r=1$ from the start (i.e. use the GDGD instead of MO-GDGD) can give very close results. This remark is confirmed in tables 3 and 4 where the information criteria are close, the likelihood ratio has a small value. It should also be noted that the high censorship level has yielded an important cure fraction which is in harmony with the results obtained in the simulation study.

In table 3, the Γ_{H_0} test and its corresponding P-value for pediatric cancer data prove the efficiency of the MO-GDGD model to fit current data. From the P-values, we reject all the three null hypotheses in favor of the MO-GDGD at a 5% level of significance. This is also proved by the AIC and AICc criteria in table 4.

4. Conclusion

This paper has been mainly focused on defective modeling and surviving proportions of type-I censored survival data. The flexibility of the Marshall-Olkin Generalized Defective Gompertz Distribution to fit medical survival data is highlighted. An exhaustive simulation study is also conducted. Statistical inference using maximum likelihood estimation is provided. Cure fraction estimation can be misleading in the case of censored data as patients with censored information are all considered cured. Taking this into account in future studies will pave the way for obtaining more accurate predictions.

References

[1] Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society.* Series B (Methodological), 11 (1), 15–53.

[2] Borges, P. (2017). Em algorithm-based likelihood estimation for a generalized gompertz regression model in presence of survival data with long-term survivors: an application to uterine cervical cancer data. *Journal of Statistical Computation and Simulation*, 87 (9), 1712–1722.

[3] Cantor, A. B., & Shuster, J. J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. Statistics in Medicine, 11 (7), 931–937.

[4] Dey, S., Moala, F. A., & Kumar, D. (2018). Statistical properties and different methods of estimation of gompertz distribution with application. *Journal of Statistics and Management Systems*, 21 (5), 839–876.

[5] Dhar, M., & Bhattacharya, P. (2018). Comparison of the logistic and the gompertz curve under different constraints. *Journal of Statistics and Management Systems,* 21 (7), 1189–1210.

[6] El-Gohary, A., Alshamrani, A., & Al-Otaibi, A. N. (2013). The generalized gompertz distribution. *Applied Mathematical Modelling*, 37 (1-2), 13–24.

[7] Gompertz, B. (1825). Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. Philosophical transactions of the Royal Society of London, 115, 513–583.

[8] Hamdeni, T., & Gasmi, S. (2020a). The marshall–olkin generalized defective gompertz distribution for surviving fraction modeling. *Communications in Statistics-Simulation and Computation*, 1–14.

[9] Hamdeni, T., & Gasmi, S. (2020b). A proportional-hazards model for survival analysis and long-term survivors modeling: Application to amyotrophic lateral sclerosis data. *Journal of Applied Statistics*. doi: 10.1080/02664763.2020.1830954

[10] Lambert, P. C., Dickman, P. W., Weston, C. L., & Thompson, J. R. (2010). Estimating the cure fraction in population-based cancer studies by using finite mixture models. J*ournal of the Royal Statistical Society: Series C* (Applied Statistics), 59 (1), 35–55.

[11] Marshall, A. W., & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. Biometrika, 84 (3), 641–652.

[12] Martinez, E. Z., & Achcar, J. A. (2017). The defective generalized gompertz distribution and its use in the analysis of lifetime data in presence of cure fraction, censored data and covariates. *Electronic Journal of Applied Statistical Analysis*, 10 (2), 463–484.

[13] Nissas, W., & Gasmi, S. (2019). A hybrid decision dependent maintenance model of failure rate and virtual age classes using modified weibull intensity. *Communications in Statistics-Simulation and Computation*, 1–15.

[14] Rocha, R., Nadarajah, S., Tomazella, V., & Louzada, F. (2017). A new class of defective models based on the marshall–olkin family of distributions for cure rate modeling. *Computational Statistics & Data Analysis*, 107, 48–63.

Received June, 2021

Revised December, 2021