

Ricardo Ferreira da Rocha

Defective Models for Cure Rate Modeling

São Carlos, 2016

FEDERAL UNIVERSITY OF SÃO CARLOS
CENTER OF EXACT SCIENCES AND TECHNOLOGY
STATISTICS DEPARTMENT

Ricardo Ferreira da Rocha

Defective Models for Cure Rate Modeling

A thesis submitted to the Statistics
Department at the Federal
University of São Carlos for the
degree of Doctor in Statistics.

Advisor: Dr^a. Vera Lucia Damasceno Tomazella

Co-advisors: Dr. Saralees Nadarajah and
Dr. Francisco Louzada-Neto

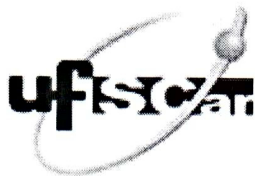
São Carlos, 2016

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

R672d Rocha, Ricardo Ferreira da
Defective models for cure rate modeling / Ricardo
Ferreira da Rocha. -- São Carlos : UFSCar, 2016.
139 p.

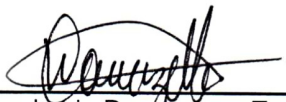
Tese (Doutorado) -- Universidade Federal de São
Carlos, 2016.

1. Cure fraction. 2. Defective models. 3. Inverse
Gaussian distribution. 4. Gompertz distribution. 5.
Kumaraswamy family. I. Título.




Folha de Aprovação


Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Ricardo Ferreira da Rocha, realizada em 01/04/2016:



Profa. Dra. Vera Lucia Damasceno Tomazella
UFSCar



Prof. Dr. Francisco Louzada Neto
USP



Profa. Dra. Juliana Cobre
USP



Prof. Dr. Victor Hugo Lachos Dávila
UNICAMP



Prof. Dr. Vinicius Fernando Calsavara
CIPE

Abstract

Modeling of a cure fraction, also known as long-term survivors, is a part of survival analysis. It studies cases where supposedly there are observations not susceptible to the event of interest. Such cases require special theoretical treatment, in a way that the modeling assumes the existence of such observations. We need to use some strategy to make the survival function converge to a value $p \in (0, 1)$, representing the cure rate. A way to model cure rates is to use defective distributions. These distributions are characterized by having probability density functions which integrate to values less than one when the domain of some of their parameters is different from that usually defined. There is not so much literature about these distributions. There are at least two distributions in the literature that can be used for defective modeling: the Gompertz and inverse Gaussian distribution. The defective models have the advantage of not need the assumption of the presence of immune individuals in the data set. In order to use the defective distributions theory in a competitive way, we need a larger variety of these distributions. Therefore, the main objective of this work is to increase the number of defective distributions that can be used in the cure rate modeling. We investigate how to extend baseline models using some family of distributions. In addition, we derive a property of the Marshall-Olkin family of distributions that allows one to generate new defective models.

Keywords: Cure fraction, Defective models, Inverse Gaussian distribution, Gompertz distribution, Kumaraswamy family, Long-term survivors, Marshall-Olkin family, Survival analysis.

Resumo

A modelagem da fração de cura é uma parte importante da análise de sobrevivência. Essa área estuda os casos em que, supostamente, existem observações não susceptíveis ao evento de interesse. Tais casos requerem um tratamento teórico especial, de forma que a modelagem pressuponha a existência de tais observações. É necessário usar alguma estratégia para tornar a função de sobrevivência convergente para um valor $p \in (0, 1)$, que represente a taxa de cura. Uma forma de modelar tais frações é por meio de distribuições defeituosas. Essas distribuições são caracterizadas por possuírem funções de densidade de probabilidade que integram em valores inferiores a um quando o domínio de alguns dos seus parâmetros é diferente daquele em que é usualmente definido. Existem, pelo menos, duas distribuições defeituosas na literatura: a Gompertz e a inversa Gaussiana. Os modelos defeituosos têm a vantagem de não precisar pressupor a presença de indivíduos imunes no conjunto de dados. Para utilizar a teoria de distribuições defeituosas de forma competitiva é necessário uma maior variedade dessas distribuições. Portanto, o principal objetivo deste trabalho é aumentar o número de distribuições defeituosas que podem ser utilizadas na modelagem de frações de curas. Nós investigamos como estender os modelos defeituosos básicos utilizando certas famílias de distribuições. Além disso, derivamos uma propriedade da família Marshall-Olkin de distribuições que permite gerar uma nova classe de modelos defeituosos.

Palavras-Chave: Análise de sobrevivência, Distribuição inversa Gaussiana, Distribuição Gompertz, Família Kumaraswamy, Família Marshall-Olkin, Fração de cura, Modelos de longa duração, Modelos defeituosos.

Contents

List of Figures	ix
List of Tables	xi
1 Preliminaries	1
1.1 Introduction	1
1.2 Theoretical Background	3
1.2.1 Survival Analysis	3
1.2.2 Kaplan-Meier Estimator	6
1.2.3 Cure Rate Models	7
1.2.4 Maximum Likelihood Estimation	9
1.3 Artificial Data Generation Algorithm	12
1.4 Data Sets	13
1.4.1 Leukemia	13
1.4.2 Melanoma	14
1.4.3 Colon	15
1.4.4 Divorce	15
1.4.5 Second Birth	16
1.5 Objectives and Overview	18
2 Defective Cure Rate Models	19
2.1 Introduction	19
2.2 Methodology	19
2.2.1 Defective Models	19
2.2.2 The Defective Gompertz Distribution	21
2.2.3 The Defective Inverse Gaussian Distribution	22
2.2.4 Inference	23
2.3 Simulation Studies	25
2.4 Applications	27
2.4.1 Leukemia data	28
2.4.2 Melanoma data	29
2.4.3 Second Birth data	31
2.4.4 Divorce data	31

2.5	Conclusions	33
3	Marshall-Olkin Family of Defective Models	35
3.1	Introduction	35
3.2	Methodology	37
3.2.1	The Marshall-Olkin Gompertz distribution	37
3.2.2	The Marshall-Olkin inverse Gaussian distribution	38
3.2.3	Inference	39
3.3	Simulation Studies	41
3.4	Applications	47
3.5	Conclusions	53
4	Kumaraswamy Family of Defective Models	54
4.1	Introduction	54
4.2	Methodology	55
4.2.1	The Kumaraswamy family of distributions	55
4.2.2	The Kumaraswamy Gompertz distribution	58
4.2.3	The Kumaraswamy inverse Gaussian distribution	59
4.2.4	Inference	60
4.2.5	The Kumaraswamy- G regression model	61
4.3	Simulation Studies	62
4.4	Applications	64
4.4.1	Melanoma data	65
4.4.2	Colon data	68
4.4.3	Leukemia data	69
4.4.4	Discussion	71
4.5	Conclusions	73
5	Generalized Extended Class of Defective Models	74
5.1	Introduction	74
5.2	Distribution Families	75
5.2.1	Gamma G	76
5.2.2	Gamma Uniform G	77
5.2.3	Exponentiated G	78
5.2.4	Truncated-Exponential Skew-Symmetric G	79
5.2.5	Beta G	79
5.2.6	Exponentiated Exponential Poisson G	82
5.2.7	Exponentiated Generalized G	82
5.2.8	Weibull- G	83
5.3	Simulation Studies	84
5.4	Applications	86

5.4.1	Leukemia data	88
5.4.2	Melanoma data	91
5.5	Conclusions	94
6	A Special Class of Defective Models Based on the Marshall-Olkin Family	96
6.1	Introduction	96
6.2	Methodology	97
6.2.1	The Marshall Olkin family	97
6.2.2	The extended Weibull distribution	99
6.2.3	Inference	102
6.2.4	Defective Marshall Olkin- G regression model	103
6.3	Simulation studies	104
6.4	Real data applications	109
6.4.1	Leukemia data	109
6.4.2	Colon data	111
6.4.3	Divorce data	113
6.4.4	Melanoma data	115
6.4.5	Discussion	116
6.5	Conclusions	118
7	Final Remarks	119
7.1	Conclusions	119
7.2	Future Works	121
7.3	Acknowledgements	122
	References	139

List of Figures

1.1	Kaplan-Meier and estimated cumulative hazard curves for the leukemia data set.	14
1.2	Kaplan-Meier and estimated cumulative hazard curves for the melanoma data set.	15
1.3	Kaplan-Meier and estimated cumulative hazard curves for the colon data set.	16
1.4	Kaplan-Meier and estimated cumulative hazard curves for the divorce data set.	17
1.5	Kaplan-Meier and estimated cumulative hazard curves for the birth data set.	17
2.1	Example of a cumulative function of a defective distribution.	20
2.2	Density, survival and hazard functions of the defective Gompertz distribution.	22
2.3	Density, survival and hazard functions of the defective inverse Gaussian distribution.	23
2.4	Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the Gompertz distribution with $(a, b, p) = (-1, 1, 0.3678)$	25
2.5	Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the Gompertz distribution with $(a, b, p) = (-2, 1, 0.6065)$	26
2.6	Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the inverse Gaussian distribution with $(a, b, p) = (-1, 5, 0.3296)$	27
2.7	Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the inverse Gaussian distribution with $(a, b, p) = (-1, 1, 0.8646)$	28
2.8	Fitted survival curves of the Gompertz and inverse Gaussian distributions in the leukemia data set.	29

2.9	Fitted survival curves of the Gompertz and inverse Gaussian distributions in the melanoma data set	30
2.10	Fitted survival curves of the Gompertz and inverse Gaussian distributions in the second birth data set	32
2.11	Fitted survival curves of the Gompertz and inverse Gaussian distributions in the divorce data set	33
3.1	Density, survival and hazard functions of the defective Marshall-Olkin Gompertz distribution.	38
3.2	Density, survival and hazard functions of the defective Marshall-Olkin inverse Gaussian distribution.	39
3.3	Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall-Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$	42
3.4	Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall-Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$	43
3.5	In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin Gompertz mixture and defective models, respectively, when the data were generated from a defective model. In the right, the corresponding estimates of p	44
3.6	In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin inverse Gaussian mixture and defective models, respectively, when the data were generated from a defective model. In the right, the corresponding estimates of p	44
3.7	In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin Gompertz mixture and defective models, respectively, when the data were generated from a mixture model. In the right, the corresponding estimates of p	46
3.8	In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin inverse Gaussian mixture and defective models, respectively, when the data were generated from a mixture model. In the right, the corresponding estimates of p	46
3.9	Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the leukemia data set.	48
3.10	Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the second birth data set.	49

3.11	Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the colon data set.	49
3.12	Plots of the Kaplan-Meier estimates of the survival function versus the predicted values from the proposed distributions. The top four plots are for the second birth data set. The middle four plots are for the leukemia data set. The bottom four plots are for the colon data set.	50
4.1	Probability density, survival and hazard functions of the defective Kumaraswamy Gompertz distribution.	58
4.2	Probability density, survival and hazard functions of the defective Kumaraswamy inverse Gaussian distribution.	59
4.3	Survival curves for the fitted distributions for the melanoma data set.	65
4.4	In the left, the fitted Gompertz regression model, in the right, the inverse Gaussian model.	67
4.5	In the left, the fitted Kumaraswamy Gompertz regression model, in the right, the Kumaraswamy inverse Gaussian model.	67
4.6	Survival curves for the fitted distributions for the colon data set.	68
4.7	Survival curves for the fitted distributions for the leukemia data set.	70
4.8	Probability plots for the fit of the four distributions to the three data sets.	71
5.1	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r and p versus n for the Exponentiated Gompertz distribution with $(a, b, r, p) = (-1, 2, 2, 0.2523)$	84
5.2	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r and p versus n for the TESS inverse Gaussian distribution with $(a, b, r, p) = (-2, 2, 2, 0.7257)$	85
5.3	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r , u and p versus n for the Weibull Gompertz distribution with $(a, b, r, u, p) = (-1, 2, 2, 2, 0.3678)$	87
5.4	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r , u and p versus n for the EEP inverse Gaussian distribution with $(a, b, r, u, p) = (-1, 2, 1, 2, 0.3976)$	88
5.5	Fitted survival curves of the proposed models when the baseline distribution is Gompertz, in the leukemia data set.	90
5.6	Fitted survival curves of the proposed models when the baseline distribution is inverse Gaussian, in the leukemia data set.	91
5.7	Fitted survival curves of the proposed models when the baseline distribution is Gompertz, in the melanoma data set.	93

5.8	Fitted survival curves of the proposed models when the baseline distribution is inverse Gaussian, in the melanoma data set.	94
6.1	From the left to the right, from the top to the bottom, the density and survival functions of the proposed distributions, in the same order presented in Table 6.1. The parameter values used are $u = (-0.2, -0.5, -1, -0.2, -0.5, -1)$, $v = (-0.5, -0.5, -0.5, -2, -2, -2)$, $a = (0.5, 0.5, 1, 1, 2, 2)$, $b = (1, 1, 2, 2, 0.5, 0.5)$ and $c = (2, 2, 0.5, 0.5, 1, 1)$. The colors are (black, red, green, blue, light blue, pink).	100
6.2	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v and p versus n for the Marshall Olkin-Lomax distribution with $(r, v) = (-1, -10)$	104
6.3	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Weibull distribution with $(r, v, a) = (-1, -2, 3)$	105
6.4	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Chen distribution with $(r, v, a) = (-1, -2, 2)$	106
6.5	Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Burr XII distribution with $(r, v, a) = (-1, -2, 2)$	107
6.6	Fitted distributions for the leukemia data set.	110
6.7	Fitted distributions for the colon data set.	112
6.8	Fitted distributions for the divorce data set.	114
6.9	From the left to the right, top to bottom, the fitted regression models for the melanoma data set, in the same order as in Table 6.1. The colors black, red, green and blue represents the nodule categories 1, 2, 3 and 4, respectively.	115

List of Tables

2.1	Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the leukemia data set.	28
2.2	Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the melanoma data set	30
2.3	Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the second birth data set	31
2.4	Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the divorce data set	31
3.1	MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the leukemia data set.	51
3.2	MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the second birth data set.	51
3.3	MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the colon data set.	52
3.4	MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the leukemia data set.	52
3.5	MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the second birth data set.	53
3.6	MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the colon data set.	53
3.7	AIC values for the fitted defective distributions compared with their respective mixture models.	53
4.1	Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy Gompertz distribution.	62
4.2	Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy inverse Gaussian distribution.	63
4.3	MLEs for the fitted distributions for the melanoma data set.	65

4.4	MLEs for the fitted regression models for the melanoma data set.	66
4.5	MLEs for the fitted distributions for the colon data set.	68
4.6	MLEs for the fitted distributions for the leukemia data set.	69
4.7	Log-likelihood ratio test for the proposed models and data sets.	72
4.8	95 percent asymptotic confidence intervals for the parameter a for the proposed models and data sets.	72
4.9	AIC values for the fitted distributions and for the standard mixture model. The smaller AIC values are bolded.	72
5.1	Maximum likelihood estimates in the leukemia data set of the proposed models when the baseline distribution is Gompertz.	89
5.2	Maximum likelihood estimates in the leukemia data set of the proposed models when the baseline distribution is inverse Gaussian.	89
5.3	Maximum likelihood estimates in the melanoma data set of the proposed models when the baseline distribution is the Gompertz	92
5.4	Maximum likelihood estimates in the melanoma data set of the proposed models when the baseline distribution is the inverse Gaussian	92
6.1	Some particular cases of the extended Weibull distribution.	99
6.2	MLEs for the fitted distributions and some measures for the leukemia data set.	111
6.3	MLEs for the fitted distributions and some measures for the colon data set.	111
6.4	MLEs for the fitted distributions and some measures for the divorce data set.	113
6.5	MLEs for the fitted regression models and the AIC measure for the melanoma data set.	115
6.6	Comparison of the AIC value of the mixture and defective models.	117
6.7	Asymptotic 95 percent confidence intervals for r	118

Chapter 1

Preliminaries

1.1 Introduction

Modeling of a cure fraction, also known as long-term survivors, is a part of survival analysis. It studies cases where supposedly there are observations not susceptible to the event of interest. Such cases require special theoretical treatment, in a way that the modeling assumes the existence of such observations. In the standard theory of survival analysis the survival function $S(t)$ tends to zero as time increases. We need to use some strategy to make the survival function converge to a value $p \in (0, 1)$, representing the cure rate.

The method most commonly used is the standard mixture model, initially proposed by [Boag \(1949\)](#) and [Berkson & Gage \(1952\)](#). The model is described by $S(t) = p + (1-p)S_0(t)$, where $S_0(t)$ is a proper survival function. Common choices for $S_0(t)$ are the Weibull, Gompertz and lognormal distributions, according to [Ibrahim *et al.* \(2005\)](#). [Tsodikov *et al.* \(2003\)](#) proposed a non-mixture model defined in terms of a cumulative hazard rate function. Its survival function has the form $S(t) = p^{F_0(t)}$, where $F_0(t)$ represents a proper distribution function. More about this method can be found in [Martinez *et al.* \(2013\)](#). Many other methods are known for cure rate modeling, see, for example, [Cooner *et al.* \(2007\)](#), [Rodrigues *et al.* \(2009a\)](#), [Nieto-Barajas & Yin \(2008\)](#) and the book [Maller & Zhou \(1996\)](#).

The literature regarding to cure rate models are very large and have lots of different approaches on how estimate the quantities of interest. In [Chen *et al.* \(1999\)](#) is proposed some Bayesian models to estimate cure fractions. In [Sy & Taylor \(2000\)](#) is discussed maximum likelihood techniques in an Cox proportional hazards structure of cure model. In [Rodrigues *et al.* \(2009b\)](#), its used the Conway-Maxwell Poisson as the distribution of competing causes, as proposed in [Rodrigues *et al.* \(2009a\)](#). In [Yin & Ibrahim \(2005\)](#) an unified approach is presented based in the Box-Cox transformation. In [Peng & Xu \(2012\)](#) an extension of the model presented in [Yin & Ibrahim \(2005\)](#) are done and some model selection criteria are discussed. In [Balakrishnan & Pal \(2012\)](#) is proposed an expectation-maximization algorithm to do estimation in the model proposed in [Rodrigues *et al.* \(2009b\)](#), where the time-to-event is assumed exponential. In [Balakrishnan & Pal \(2013a\)](#), [Balakrishnan & Pal \(2013b\)](#) and [Balakrishnan & Pal \(2015\)](#), the authors keeps developing the EM algorithm but with Weibull, lognormal and generalized gamma distributions to the time-to-event.

Another way to model cure rates is to use defective distributions, as explored in this thesis. Defective distributions are characterized by having probability density functions which integrate to values less than 1 when the domain of some of their parameters is different from that usually defined. There is not so much literature about these distributions. There are at least two distributions in the literature that can be used for defective modeling: the Gompertz and inverse Gaussian distributions. The use of these defective distributions became more appealing after the works of [Balka *et al.* \(2009\)](#) and [Balka *et al.* \(2011\)](#), although some previous papers have used the same idea. In [Whitmore \(1979\)](#), the term *defective* was used to refer to the inverse Gaussian distribution that allows one of its parameters to be negative.

The Gompertz distribution becomes defective when its shape parameter is negative. It first appeared in [Haybittle \(1959\)](#), where it was used to model a breast cancer data set. [Cantor & Shuster \(1992\)](#) applied a modified version of this distribution to a pediatric cancer data set. [Gieser *et al.* \(1998\)](#) extended the distribution to include covariates. More recently, [Rocha *et al.* \(2014\)](#) performed Bayesian estimation of this distribution. In [Marshall & Olkin \(2015\)](#), a bivariate version of the Gompertz distributions is proposed.

In that work, the authors called the gompertz distributions that allows the parameters outside their usual domain of *negative* Gompertz.

The inverse Gaussian distribution was first proposed in [Schrödinger \(1915\)](#) for calculating the first time passage probability of a one-dimensional Brownian motion (Wiener process). More details were studied in [Tweedie \(1945\)](#) and [Whitmore \(1979\)](#). Defective versions were investigated in [Balka et al. \(2009\)](#) and [Balka et al. \(2011\)](#), with classical and Bayesian approaches. Having only two distributions is not enough to provide sufficient flexibility. So, the main goal of this thesis is to provide more distributions with the defective property.

In the next section we present all the basic theoretical components needed to understand and interpret the results in the next chapters. Further, we also present the algorithm to generate artificial data and five real data sets that are used in some chapters of this thesis. In the end, we discuss the objectives and a general overview of this work.

1.2 Theoretical Background

Here we show the first definitions in the survival analysis area, since its basic relations until the unified theory for cure modeling, and its estimation by maximum likelihood.

1.2.1 Survival Analysis

Survival analysis, or reliability analysis, is the branch of statistics that study data normally associated to the duration of time until the occurrence of an event of interest. The time can be from the duration of an electronic component or the lifetime of patients with serious diseases. It does not have to be necessarily a time to event kind of data. For example, we can check how many kilometers can a tire work properly without replacement. The main areas of interest are: medicine, biology, engineering, statistics, economics, social sciences, among others.

What makes the survival analysis a particular area is their specific characteristic to take into account incomplete observations, or also called, censored information. With the

presence of censoring, it is impossible to apply standard statistical for analyzing such data. There are some kinds of censoring, described in the following.

According to [Colosimo & Giolo \(2006\)](#), the type I censorship or right-censorship, occurs when the time to the end of the study is pre-established. Thus, some individuals fail to experience the event of interest in the end of this study, and their lifetimes are right censored. An example of this type of censorship is when a bank want to check the time until the customers of a particular portfolio become a bad payer. It is studied this portfolio for a predetermined amount of time by the institution and at end, some elements will not experience the event of interest (and therefore are not considered bad payers). This censoring also occurs when, for some reason, the subject in study is not available anymore. That could be, for example, someone that quits a drug trial for lack of motivation, or because the event of interest cannot be observed after some point.

The type II censorship occurs when the study is finished after a certain number n of individuals experience the event of interest, that is, after a number n of the research trials is completed and the individuals who have left to experience the event of interest will be considered censored.

The random censorship, unlike the others, is a kind of censorship that is beyond the control of the researcher. It usually occurs when a person leave a given experiment without having experienced the event of interest. The random censorship is a more common case, with the particular case censorship Type I, for example, if the patient die for a different reason of the one considered in the study.

In this work, we represent the data and the censoring by the following: each subject is observed and denoted by (t_i, δ_i) , in which t_i is the time until the fail or censoring and δ_i is the variable that indicates if that observation was as fail or a censoring. If $\delta_i = 1$, then the fail was observed. If $\delta_i = 0$, a censoring occurred.

Suppose now that the random variable T , $T \geq 0$, have density function denoted by $f(t)$. As in [Colosimo & Giolo \(2006\)](#), we can write the density function as the limit of the

probability of a subject fails in the interval of time $[t, t + \Delta t]$:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

Its cumulative function is given by:

$$F(t) = P(T \leq t) = \int_0^t f(u) du.$$

To estimate the probability of an individual survive at least until the time t is one of the major interests of the survival analysis. So, it is defined the survival function, given by:

$$S(t) = P(T > t) = \int_t^{\infty} f(u) du = 1 - F(t).$$

Of course, the properties of this function are quite similar to the cumulative function: $S(t)$ is not increasing; $S(0) = 1$ and $\lim_{t \rightarrow \infty} S(t) = 0$.

Other function of huge importance is the hazard function, also called hazard rate function, that provides the instant rate of fail, that is, knowing that a subject survived until the time t , this function represents the chance of this subject will fail in the time $t + \Delta t$, with $\Delta t \rightarrow 0$. The hazard function is defined by:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Graphically, the hazard function can have several forms. The cases most studied is where the hazard function is increasing, decreasing, constant, unimodal and bathtub shaped. Checking the hazard behavior is important when someone have to choose between parametric models. The cumulative hazard function is defined by:

$$H(t) = \int_0^t h(u) du. \tag{1.1}$$

These equation have a major interest in the survival analysis. Some useful relations

between them are:

$$\begin{aligned}h(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt} \log[S(t)], \\H(t) &= -\log[S(t)], \\S(t) &= \exp[-H(t)].\end{aligned}$$

1.2.2 Kaplan-Meier Estimator

In the survival analysis literature can be found some estimators of a survival function obtained through non-parametric techniques. We can refer, for instance, the Nelson-Aalen estimator, proposed by [Nelson \(1972\)](#) and then reviewed by [Aalen \(1978\)](#), and the one proposed by [Kaplan & Meier \(1958\)](#). This last one is the most important non-parametric estimator and is described next. For that, consider the following:

- $t_{(1)} < t_{(2)} < \dots < t_{(k)}, j = 1, \dots, k$, the k ordered distinct fail times;
- d_j the number of fails in $t_{(j)}, j = 1, \dots, k$;
- n_j the number of individuals at risk in $t_{(j)}$, that is, the individuals that not failed or got censored until the moment instantly previous to $t_{(j)}$.

This way, Kaplan-Meier (KM) estimator is defined by:

$$\widehat{S}(t) = \prod_{j:t_j < t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j} \right).$$

This expression leads to a ladder function with steps in the observed fail times. In the paper where it is proposed, the authors justify the expression by showing that this estimator is the maximum likelihood estimation for $S(t)$. Because of this, one of the most usual ways to check the fit of an proposed parametric model is to compare it to the Kaplan-Meier curve. The better the KM captures the fitted model, the better the model is.

1.2.3 Cure Rate Models

The survival theory has been widely explored by many researchers in various areas, with a major focus on analysis of clinical data. Generally the survival function $S(t) = P(T > t)$ is the function used to represent the random behavior of T . A property of $S(t)$ is that it goes to zero as the time pass, which characterizes an event of interest that eventually always occur.

However, there are situations in which a portion of the population is considered cured and cannot fail. For example, there are cases when it is considered the recurrence of a cancer. Some people can have the recurrence, however, there may be some others that is completely cured from that cancer and, therefore, it would never recur. To solve such problems, [Berkson & Gage \(1952\)](#), based on the work of [Boag \(1949\)](#), proposed the standard mixture model for cured fraction. The survival function is set to

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

in a way that $S_0(t)$ is a proper survival function. Thus, it follows that $S(t)$ converges to p as the time increases. In [Berkson & Gage \(1952\)](#) is made an analysis in patients with stomach cancer, and from there, several other studies of cure rate have been proposed in the literature, focusing on that model standard mixture. The most common choices common to $S_0(t)$ are the Weibull, log-logistic and log-normal distributions. Recently, different models have been proposed for this purpose, as in [Yakovlev & Tsodikov \(1996\)](#), [Chen *et al.* \(1999\)](#) and [Ibrahim *et al.* \(2005\)](#).

In addition to this approach, we have a unified long-term theory, proposed by [Rodrigues *et al.* \(2009a\)](#) that generalizes, among others, the mixture model. Let N be a random variable that represents the number of causes of risk, for a particular event of interest, with probability distribution of

$$p_n = P[N = n],$$

in which $n = 0, 1, 2, \dots$. In this case, N is a latent random variable. Given $N = n$, let Z_v ,

$v = 1, \dots, n$, be independent, non-negative random variables, with distribution function $F(t) = 1 - S(t)$. Consider also that N is independent of Z_v , where Z_v represents the time until the occurrence of an particular event of interest, because of the v -th cause of risk.

The time of occurrence of the event of interest is defined as:

$$T = \min \{Z_1, Z_2 \dots, Z_N\}, \quad (1.2)$$

in which $P[Z_0 = \infty] = 1$, leads to a proportion p_0 of the non-susceptible subjects to the event of interest. The variables Z_v are latent and T is an observable random variable or censoring. The survival function of the random variable T is given by: $S_{pop}(t) = P[T > t]$.

Let $\{a_n\}$ be a sequence of real numbers and $s \in [0, 1]$. Consider then the following:

$$A(s) = a_0 + a_1s + a_2s^2 + \dots .$$

According to [Feller \(1968\)](#), if $A(s)$ converges, then $A(s)$ é defined as the generating function of the sequence $\{a_n\}$. Given a proper survival function $S(t)$, the survival function of the random variable T , as in (1.2), is given by

$$S_{pop}(t) = A[S(t)] = \sum_{n=0}^{\infty} p_n [S(t)]^n. \quad (1.3)$$

The proof is in [Rodrigues et al. \(2009a\)](#). This implies that $\lim_{t \rightarrow \infty} S_{pop} = P[N = 0] = p_0$, with p_0 denoting the cured fraction.

The survival function $S_{pop}(t)$ obtained in (1.3) is not proper. The associated density and hazard function are given, respectively, by:

$$\begin{aligned} f_{pop}(t) &= f(t) \frac{d}{ds} A[S(t)], \\ h_{pop}(t) &= \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{f(t)}{S_{pop}(t)} \frac{d}{ds} A[S(t)]. \end{aligned}$$

Some examples of generating function can be obtained by using the distributions: Bernoulli, binomial, negative binomial, Poisson, geometric, power series, among others. If we assume

the distribution for N is Bernoulli, then S_{pop} is the same proposed in [Berkson & Gage \(1952\)](#). In [Feller \(1968\)](#) we can check that the generating function for the Bernoulli(θ) distribution is $A(u) = \theta + (1 - \theta)u$. Thus, we have the mixture model

$$S_{pop}(t) = A[S(t)] = \theta + (1 - \theta)S(t).$$

If we assume the distribution for N is Poisson, then S_{pop} is the same proposed in [Chen et al. \(1999\)](#), the promotion time cure model.

1.2.4 Maximum Likelihood Estimation

In survival analysis, one of the concerns is to fit parametric models to the observed data, because they have a more natural interpretation and can calculate the needed probabilities more adequately.

Based on results obtained from samples, the maximum likelihood estimator selects the best set of parameters for the alleged distribution of the data. The maximum likelihood method is able to incorporate censorship and has excellent properties for large samples (asymptotic results), and is, therefore, the most widely used method for survival analysis.

As censored data bring us important information, we cannot leave it aside. Its contribution to $L(\boldsymbol{\theta})$ is given by the survival function $S(t)$. Thus, the observations of the random sample can be divided into two sets, the censored and uncensored.

Suppose that the data are independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ denotes a vector of parameters. Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)'$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)'$ are the censored failure times. The δ_i is equal to 1 if a failure is observed and 0 otherwise.

The likelihood function of $\boldsymbol{\theta}$ can be written as (see [Klein & Moeschberger \(2003\)](#))

$$L(\boldsymbol{\theta}; \mathbf{D}) \propto \prod_{i=1}^n \left[f(t_i; \boldsymbol{\theta})^{\delta_i} S(t_i; \boldsymbol{\theta})^{1-\delta_i} \right].$$

The corresponding log-likelihood function is

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^n (1 - \delta_i) \log S(t_i, \boldsymbol{\theta}).$$

This expression is valid for censoring type I, type II, random and when the censor mechanism is not informative. The maximum likelihood estimator is the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, or, equivalently, its log-likelihood function, $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$. The estimators are found by solving the system of equations

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = 0,$$

for $j = 1, \dots, k$.

Normally, the maximum likelihood estimator does not have a closed expression. That is due to the complexity that the equations can get depending on the assumed parametric model for the data in question. So, usually it is necessary to use computational methods to calculate the maximum likelihood estimates numerically. There are various routines available for numerical maximization. We used the routine *optim* in the R software ([R Core Team, 2013](#)). The maximization algorithm used was the *BFGS*, for more information on this method, please see [Liu & Nocedal \(1989\)](#). All estimation procedures by maximum likelihoods done in this thesis was done in R, using *optim* with *BFGS*.

Confidence intervals for the parameters were based on asymptotic normality. If $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimator of $\boldsymbol{\theta}$ then it is well known that the distribution of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ can be approximated by a k -variate normal distribution (where k denotes the length of the vector $\boldsymbol{\theta}$ as defined above) with zero mean and covariance matrix $\mathbf{I}(\hat{\boldsymbol{\theta}})$, where $\mathbf{I}(\boldsymbol{\theta})$

denotes the observed information matrix defined by

$$\mathbf{I}(\boldsymbol{\theta}) = - \begin{pmatrix} \frac{\partial^2 \log L}{\partial \theta_1^2} & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log L}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_k^2} \end{pmatrix}.$$

So, an approximate $100(1 - \alpha)$ percent confidence interval for θ_i is $(\hat{\theta}_i - z_{\alpha/2} \sqrt{I^{ii}}, \hat{\theta}_i + z_{\alpha/2} \sqrt{I^{ii}})$, where I^{ii} denotes the i th diagonal element of the inverse of \mathbf{I} and z_a denotes the $100(1 - a)$ percentile of a standard normal random variable.

In the defective distributions theory, the cured fraction p is calculated as a function of the estimated parameters. To estimate the variance of p is used the delta method with a first order Taylor's approximation. For more on the delta method, please see [Oehlert \(1992\)](#).

We will also consider some measures to check the relative quality of a fitted model: the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and CAIC (Consistent Akaike Information Criterion). They are not a measure of quality by itself, but is useful to compare between fitted models. Therefore, these measures provides a way of model selection. The definitions are:

$$\begin{aligned} \text{AIC} &= 2k - 2 \log(L), \\ \text{BIC} &= k \log(n) - 2 \log(L), \\ \text{CAIC} &= k[\log(n) + 1] - 2 \log(L), \end{aligned}$$

where k is the number of parameters in the model, n is the sample size and L is the likelihood value in the estimated parameters. The better fit is the one with the lowest AIC, BIC or CAIC. For more, see [Bozdogan \(1987\)](#).

1.3 Artificial Data Generation Algorithm

Here we describe the data generation used in order to assess the performance of the maximum likelihood estimates with respect to sample size and to show, among other things, that the usual asymptotes of maximum likelihood estimators still hold for defective distributions. The assessment is based on simulations. In all chapters, the simulation studies are based in this setup. The description of the data generation is given below.

Suppose that the time of occurrence of an event of interest has cumulative distribution function $F(t)$. We want to simulate a random sample of size n containing real times, censored times and a cure fraction of p . An algorithm for this purpose is:

- Determine the desired parameter values, as well as the value of the cure fraction p ;
- Generate $M_i \sim \text{Bernoulli}(1 - p)$;
- If $M_i = 0$ set $t'_i = \infty$. If $M_i = 1$ take t'_i as the root of $F(t) = u$, where $u \sim \text{uniform}(0, 1 - p)$;
- Generate $u'_i \sim \text{uniform}(0, \max(t_i))$, considering only the finite t_i ;
- Calculate $t_i = \min(t'_i, u'_i)$. If $t_i < u'_i$ set $\delta_i = 1$, otherwise set $\delta_i = 0$.

Note that the range of $F(t)$ has been changed and some adjustments made. Instead of $(0, 1)$, we have used $(0, 1 - p)$. Therefore, in the third step of the algorithm, the root of $F(t) - u = 0$ must be for $u \sim \text{uniform}(0, 1 - p)$. In the fourth step, the censoring distribution chosen is a $\text{uniform}(0, \max(t_i))$. The limit $\max(t_i)$ was taken in order to control the censoring regardless of the initial parameter choices. In this way, the censoring rates were kept reasonable, as described above.

In the simulations, we always have to choose the values of the parameters in the distribution that we want to analyze. Also, we always choose the value of $S = 1000$ simulation per sample size. In each sample size, we calculate the bias, mean square error, coverage probability and coverage lengths for each parameter. $\hat{\theta}$ is the average of θ_i , for $i = 1, \dots, S$.

The following equations were used:

$$\begin{aligned}\text{Var}(\hat{\theta}) &= \frac{1}{S} \sum_{i=1}^S (\hat{\theta}_i - \theta)^2, \\ \text{Bias}(\hat{\theta}) &= \hat{\theta} - \theta, \\ \text{MSE}(\hat{\theta}) &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}).\end{aligned}$$

The coverage probability is the frequency in which the real parameter value stays in the confidence region, for each simulation. According to [Calsavara \(2011\)](#), if we consider the S simulations results done as a result of a binomial experiment with parameter 0.95 (the significance level), a test to check proportions equivalence can be performed. This way, we have n_1 and n_2 such that $P(n_1 \leq p \leq n_2) = 0.95$. That is, given a number of simulations S , we can expect that the coverage probability will stay between n_1 and n_2 about 95% of the time. When $S = 1000$, we have $n_1 = 0.936$ and $n_2 = 0.964$. The coverage length is the difference between the upper and lower confidence bounds.

1.4 Data Sets

Here we described the five data sets used in the following chapters. They were chosen to represent a variety of sample sizes, survival and hazard curves. Three of them is of the clinical area and the other two are related to social sciences studies.

1.4.1 Leukemia

This data set relates to a study of recurrence of leukemia in patients who were submitted to a certain kind of transplantation. Leukemia is a type of cancer that affects the white blood cells produced by the bone marrow and can take several forms. The data set has forty four observations with 20.45 percent censoring (nine in total). The maximum observation time was approximately five years. For details of this data set, see [Kersey et al. \(1987\)](#). Figure 1.1 shows the Kaplan-Meier and the cumulative hazard curves. The cumulative hazard is calculated by equation (1.1).

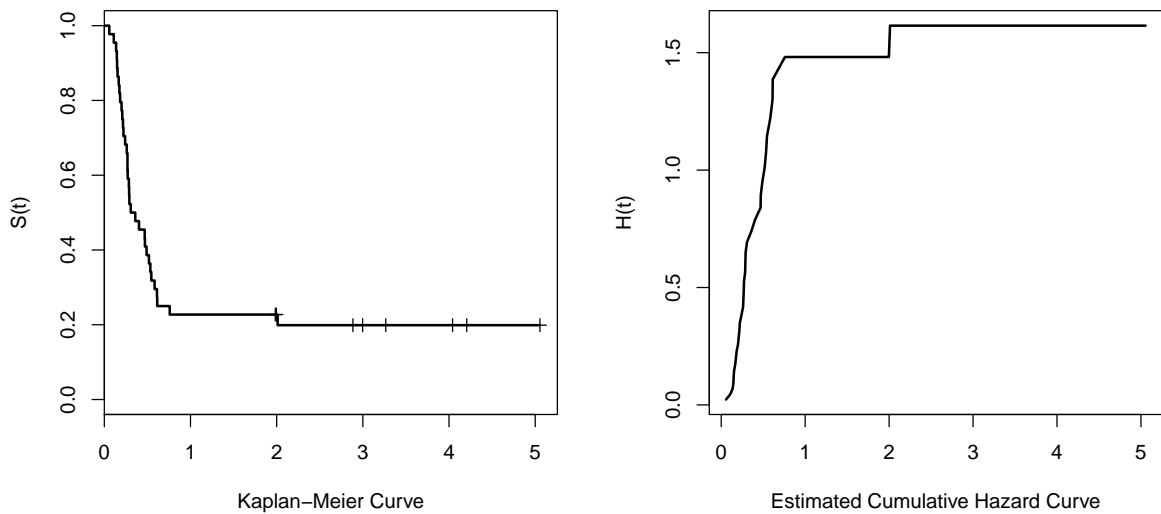


Figure 1.1: Kaplan-Meier and estimated cumulative hazard curves for the leukemia data set.

1.4.2 Melanoma

This data set collected in the period 1991-1998 is related to a clinical study in which patients were observed for recurrence after a removal of a malignant melanoma. Melanoma is a type of cancer that develops in melanocytes, responsible for skin pigmentation. It is a potentially serious malignant tumor that may arise in the skin, mucous membranes, eyes and central nervous system, with a great risk of producing metastases and high mortality rates in the later stages. There are 417 observed times, of which 232 were censored (55.63 percent). For details of this data, see [Ibrahim *et al.* \(2001\)](#).

This data set has covariates information, which is used to illustrate regression models when it is needed. One of the covariate taken represents the nodule category ($n_1 = 82, n_2 = 87, n_3 = 137, n_4 = 111$). Another covariate present is the age of the individuals. The Kaplan-Meier estimates suggest that the survival rate increases with the nodule category. Figure 1.2 shows the Kaplan-Meier and the cumulative hazard curves.

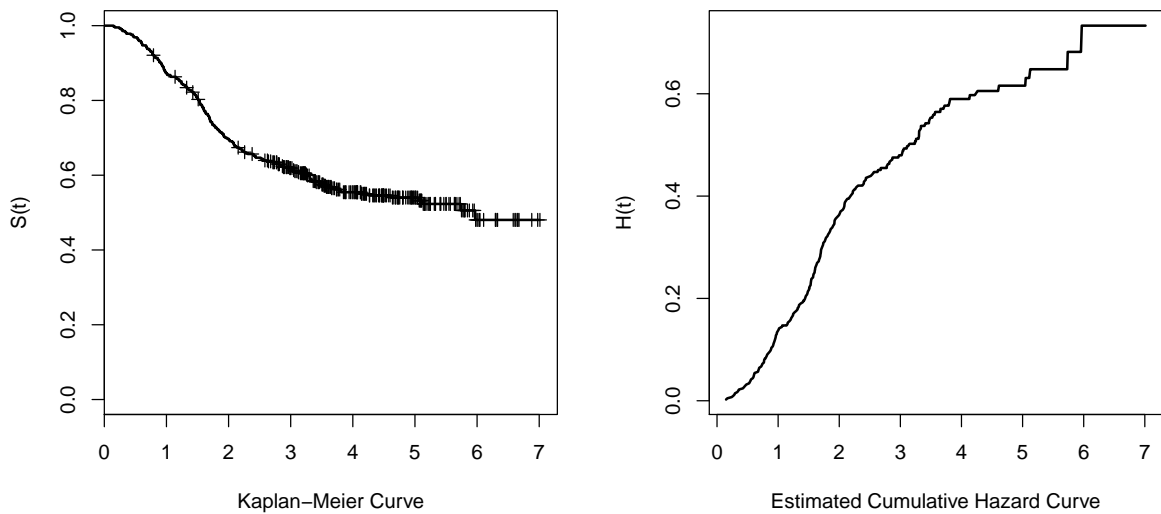


Figure 1.2: Kaplan-Meier and estimated cumulative hazard curves for the melanoma data set.

1.4.3 Colon

This data set arises from one of the first successful trials of adjuvant chemotherapy for colon cancer. The event of interest here is the recurrence or death for the individual under the proposed treatment. The data set has 1858 observations and 50.58 percent censoring (938 in total). The data set is available in R in the survival package. Details of this data set can be found in [Laurie *et al.* \(1989\)](#). Figure 1.3 shows the Kaplan-Meier and the cumulative hazard curves.

1.4.4 Divorce

This data set collected in the USA describes married couples and the event of interest is the divorce. Of course, that event may never occur, there is a high censoring in this data set. The cure elements are those couples who will never divorce. There are 3371 observed times, of which 2339 were censored (69.38 percent). The maximum observed time was 73.07 years and the average observed time was 18.41 years. For details of this data, see [Lillard & Panis \(2000\)](#). The Kaplan-Meier curve for this data stabilizes at 0.5566. It appears quite safe to say that this value is an asymptote of the curve. Almost no failures

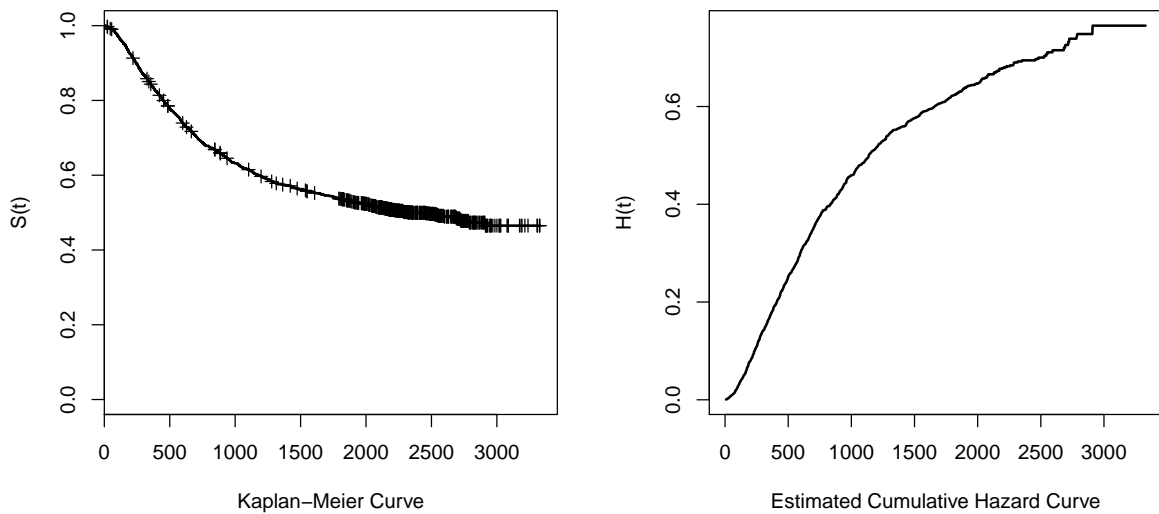


Figure 1.3: Kaplan-Meier and estimated cumulative hazard curves for the colon data set.

were observed in the second half of the period of study. So, we can expect a real cure fraction quite close to the Kaplan-Meier estimate. Figure 1.4 shows the Kaplan-Meier and the cumulative hazard curves.

1.4.5 Second Birth

This data set relates to the time of birth of a second child for a couple and is based on medical records of births in Norway in 1997. The observed time is the gap between the birth of the first child and the birth of the second child for the same couple. The data set consists of 53543 women who had their first child between 1983 and 1997. The censoring indicates whether the woman had a second child, the event of interest, or if she did not before the end of the study. The data set was previously analyzed by [Aalen *et al.* \(2008\)](#). For illustrative purposes, we took a random sample accounting for 2 percent of the data set, totalling 1071 observations with 69.74 percent censoring (747 in total). Figure 1.5 shows the Kaplan-Meier and the cumulative hazard curves for this data.

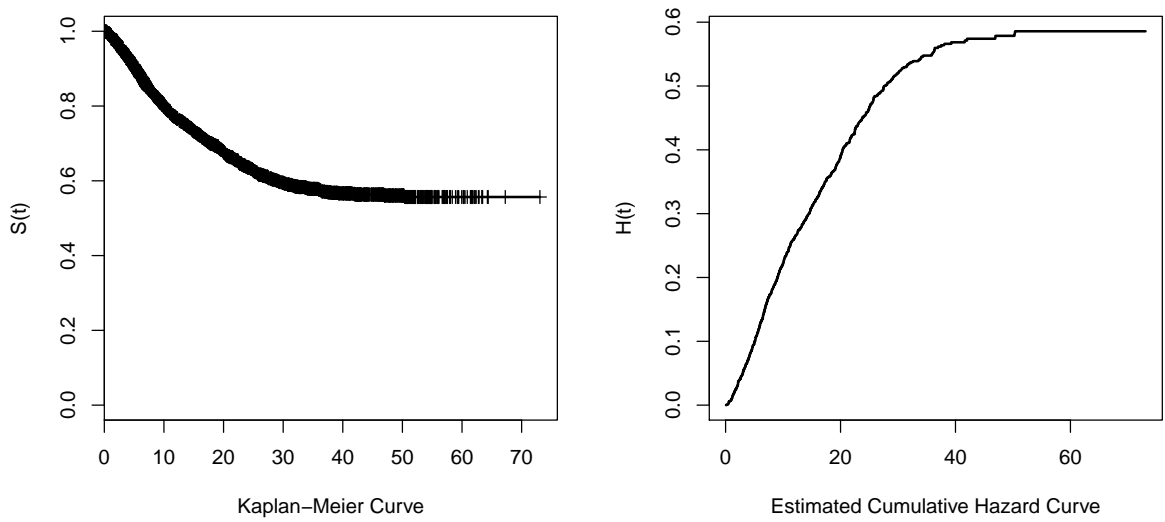


Figure 1.4: Kaplan-Meier and estimated cumulative hazard curves for the divorce data set.

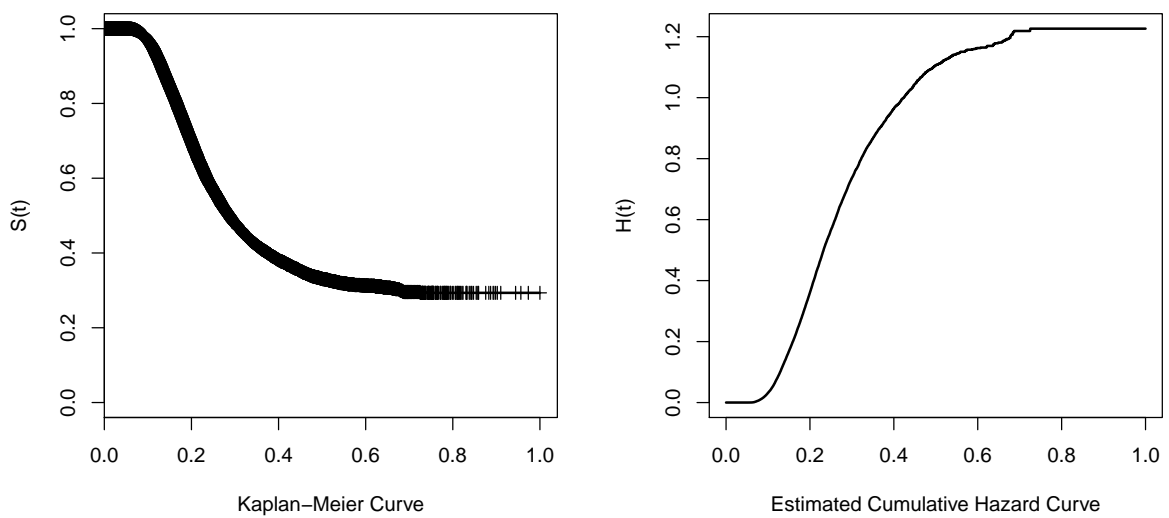


Figure 1.5: Kaplan-Meier and estimated cumulative hazard curves for the birth data set.

1.5 Objectives and Overview

The defective models have the advantage of not need the assumption of the presence of immune individuals in the data set. Because of that, it has one less parameter than the same model in the mixture model approach. The literature provides only two distributions with the defective property. In order to use the defective distributions theory in a competitive way, we need a larger variety of these distributions. Therefore, the main objectives of this work is to increase the number of defective distributions that can be used in the cure rate modeling. We will investigate how to extend baseline models through some family of distributions. In addition, we derive a property of the Marshall-Olkin family of distributions that allows one to generate new defective models.

The overview of this work is as following. In Chapter 2 we investigate the Gompertz and inverse Gaussian distribution as basic defective models and how suitable they are in some scenarios. In Chapter 3 we propose two new defective distributions using the Marshall-Olkin family of distributions. We apply the proposed models in some real data sets in order to reach a improved model in relation to the baseline distributions. In Chapter 4 we propose two more new defective distributions using the Kumaraswamy family of distributions. We apply the proposed models in some real data sets in order to reach a improved model in relation to the baseline distributions. In Chapter 5 we propose a general result that allows one to extend an defective model using any family of distributions. We use eight new families to generate sixteen more new defective distribution, as examples. In Chapter 6 we propose a property of the Marshall-Olkin family that allow one to generate defective distributions without using the Gompertz or the inverse Gaussian as the baseline. We exemplify the result by proposing ten new defective distributions. Finally, in Chapter 7 we discuss the conclusions of this thesis and some proposals for future work. We published the papers [Rocha *et al.* \(2014\)](#), [Rocha *et al.* \(2015a\)](#), [Rocha *et al.* \(2015c\)](#) and submitted [Rocha *et al.* \(2015b\)](#), which is based on the Chapters 2, 3, 4 and 6, respectively. The results in Chapter 5 are about to be submitted.

Chapter 2

Defective Cure Rate Models

2.1 Introduction

The aim of this chapter is to introduce the basic defective distributions found in the literature: the Gompertz and inverse Gaussian. First, we properly define and discuss the defective models. Then, we check the validity of the maximum likelihood estimates through some simulation scenarios. In the application section, we use four different data sets to exemplify the performance of the proposed models.

2.2 Methodology

Here we define what a defective model is and present the known defective distributions present in the literature.

2.2.1 Defective Models

Definition 2.1. *A distribution is called defective if the integral of its density function does not result in 1, but in a value $p \in (0, 1)$, when the domain of the parameters are changed.*

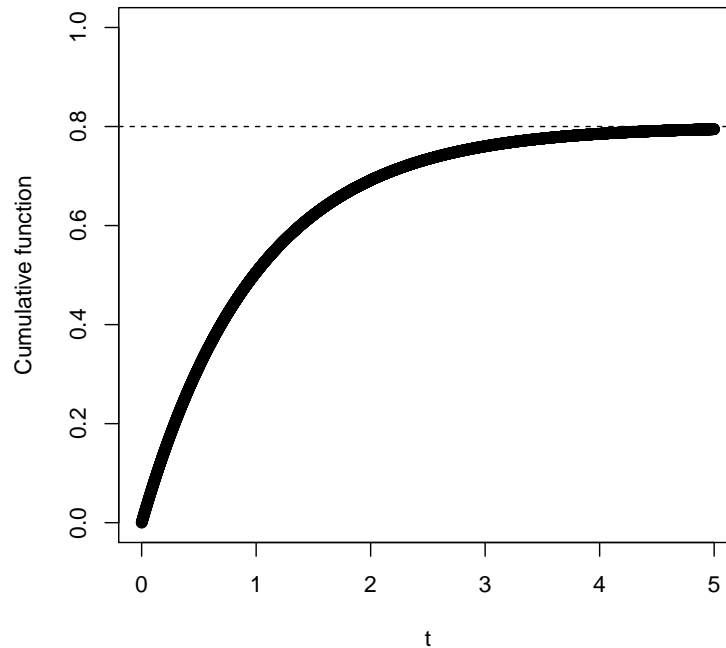


Figure 2.1: Example of a cumulative function of a defective distribution.

A defective model is a model with a defective distribution. In a defective model, it is possible to estimate a cure rate with the use of a naturally improper distribution. Instead of estimating the proportion p directly as a mixture model, we use a distribution by changing the domain of its parameters. And that leads to a model with long-term duration.

In a defective distribution, the cumulative function no longer approaches to 1, but to p and, therefore, the survival function approaches to $1 - p$. Figure 2.1 illustrates the cumulative function of a defective distribution.

Obviously, the defective distribution is not proper. When used as a model for cure fraction, the proportion of the population that is immune is obtained by calculating the limit of the survival function using the estimated parameters. In the literature, there are two known distributions that can be used for this purpose: the inverse Gaussian and Gompertz distribution. Both distribution have two positive parameters. For negative values of the shape parameter, the distribution becomes defective. The parameters that change their domains are called *defective parameters*.

A great advantage of these distributions is that the cured fraction is always estimated using a model with one parameter less than the standard mixture model, which brings plenty of benefits in terms of estimation. And it is easy to calculate because it is a simple function of the estimated parameters.

Other great advantage is that it is not necessary to assume the existence of a cure fraction in your model. Once you have a defective model, it will lead to a cure fraction when the estimation procedure presents a value out of the usual range of parameters. The significance can be tested based on the significance of the defective parameters.

One of the drawbacks is that the model may lose some of its flexibility when we have less parameters. Also, since the cure fraction depends on others parameters, the interval estimation of it is not directly, and need to be approximated using other techniques, for example, the delta method.

In the next section we show the Gompertz and inverse Gaussian models in their defective forms. In Section 2.3 we have some simulation setups in order to verify the properties of the maximum likelihood estimator. In Section 2.4 we show some applications in real data sets.

2.2.2 The Defective Gompertz Distribution

The Gompertz distribution is used for modeling survival data in various areas of knowledge (Gieser *et al.*, 1998), especially where there is a suspicion of exponential hazard. The Gompertz density function is

$$f(t) = be^{at}e^{-\frac{b}{a}(e^{at}-1)} \quad (2.1)$$

for $a > 0$, $b > 0$ and $t > 0$. In this parameterization, a is the shape parameter and b is the location parameter. The survival function is

$$S(t) = e^{-\frac{b}{a}(e^{at}-1)}. \quad (2.2)$$

The defective Gompertz distribution is the one that allows for negative values for the parameter a . The proportion of immunity in the population is calculated as the limit of the survival function when $a < 0$:

$$p = \lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}} \in (0, 1).$$

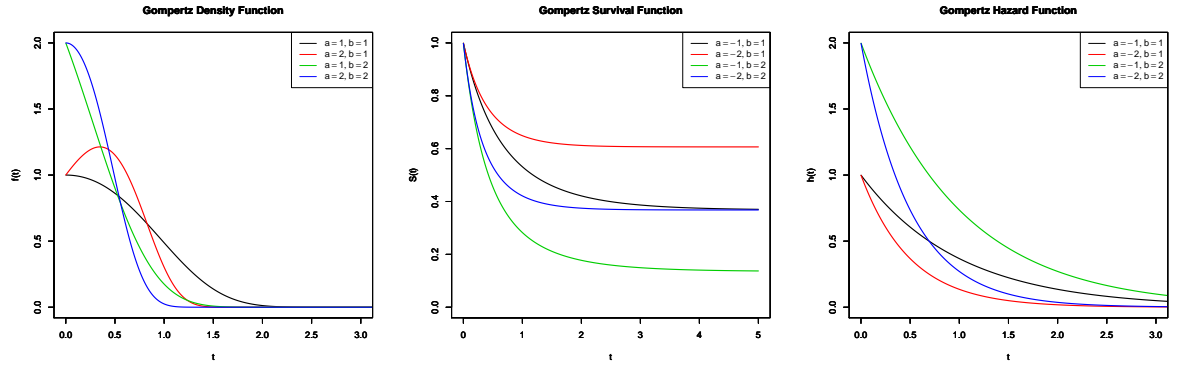


Figure 2.2: Density, survival and hazard functions of the defective Gompertz distribution.

Once the parameter values are estimated, one can easily compute the fraction of cure p . Figure 2.2 illustrates various scenarios for the density, survival and hazard functions of the Gompertz distribution.

2.2.3 The Defective Inverse Gaussian Distribution

The inverse Gaussian distribution arises as the first passage time of a Wiener process (Balka *et al.*, 2009). Lee & Whitmore (2006) noted its potential as models for cure rate. Its density function is

$$f(t) = \frac{1}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{1}{2bt} (1 - at)^2 \right\} \quad (2.3)$$

for $a > 0$, $b > 0$ and $t > 0$. The inverse Gaussian distribution has survival function given by

$$S(t) = 1 - \left[\Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right], \quad (2.4)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable.

The inverse Gaussian distribution can be defective when $a < 0$. The fraction of cure, or the survival function limit, is

$$p = \lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} 1 - \left[\Phi\left(\frac{-1+at}{\sqrt{bt}}\right) + e^{2a/b} \Phi\left(\frac{-1-at}{\sqrt{bt}}\right) \right] = 1 - e^{2a/b} \in (0, 1).$$

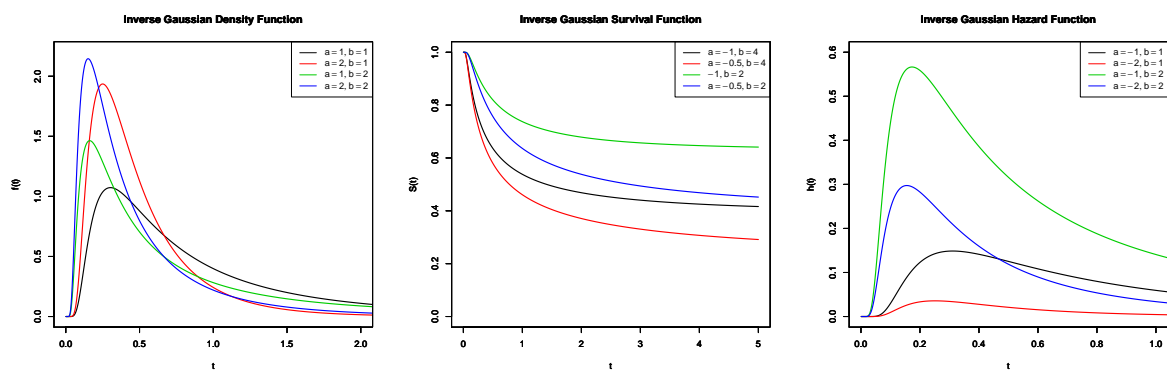


Figure 2.3: Density, survival and hazard functions of the defective inverse Gaussian distribution.

We estimate the cure fraction using the estimated parameters a and b . Figure 2.3 illustrates various scenarios for the density, survival and hazard functions of the inverse Gaussian distribution.

We have been able to find only these two distributions (Gompertz and inverse Gaussian) that can be adapted to being defective. This does not mean there are not others.

2.2.4 Inference

Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)'$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)'$ are the censored failure times. Suppose that the data are independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ denotes a vector

of parameters. The log-likelihood function of $\boldsymbol{\theta}$ can be written as

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^n (1 - \delta_i) \log S(t_i, \boldsymbol{\theta}).$$

For the Gompertz distribution given by (2.1) and (2.2),

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \ln(b) \sum_{i=1}^n \delta_i + a \sum_{i=1}^n \delta_i t_i - \frac{b}{a} \sum_{i=1}^n (e^{at_i} - 1). \quad (2.5)$$

For the inverse Gaussian distribution given by (2.3) and (2.4),

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) = & \text{const} + \sum_{i=1}^n \delta_i \log \left(\frac{1}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{1}{2bt} (1 - at)^2 \right\} \right) + \\ & \sum_{i=1}^n (1 - \delta_i) \log \left(1 - \left[\Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right] \right), \end{aligned} \quad (2.6)$$

where $\boldsymbol{\theta} = (a, b)'$.

The log-likelihood functions, (2.5) and (2.6), can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization.

Confidence intervals for the parameters were based on asymptotic normality. We have supposed the usual asymptotes of the maximum likelihood estimates hold. However, defective distributions like the mixture model are not proper distributions. The checking of regularity conditions for the asymptotes by analytical means is not easy. Such conditions have not been checked even for the standard mixture model.

In the next section, we perform a simulation study to check the asymptotes of the maximum likelihood estimates. Simulations have been used in many papers to assess the behavior of maximum likelihood estimates, especially when an analytical investigation is intractable.

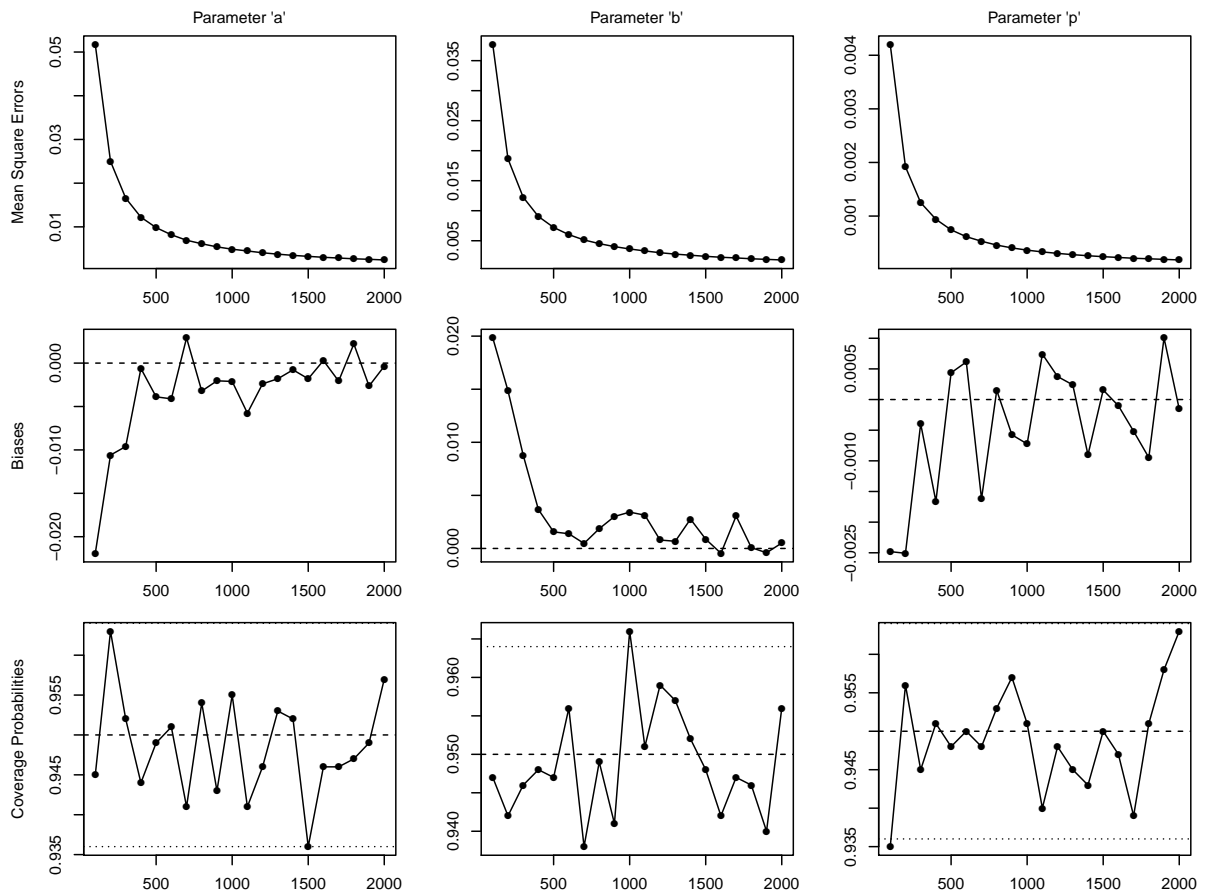


Figure 2.4: Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the Gompertz distribution with $(a, b, p) = (-1, 1, 0.3678)$.

2.3 Simulation Studies

In this section we propose four simulation scenarios in order to check the maximum likelihood estimates when the sample size increases. We generate data from the Gompertz and inverse Gaussian distribution according to Section 1.3.

In the first scenario, we simulated one thousand random samples each of size $n = 100, 200, \dots, 2000$. Random samples were taken to come from the defective Gompertz distribution with $(a, b, p) = (-1, 1, 0.3678)$. We computed the maximum likelihood estimates, \hat{a} , \hat{b} and \hat{p} , and their standard errors for each sample. These were used to compute the bias, the mean squared error and the coverage probability for each parameter. To calculate the standard deviation of the cure fraction, the delta method was used.

The second scenario was simulated from the same distribution, but with parameters

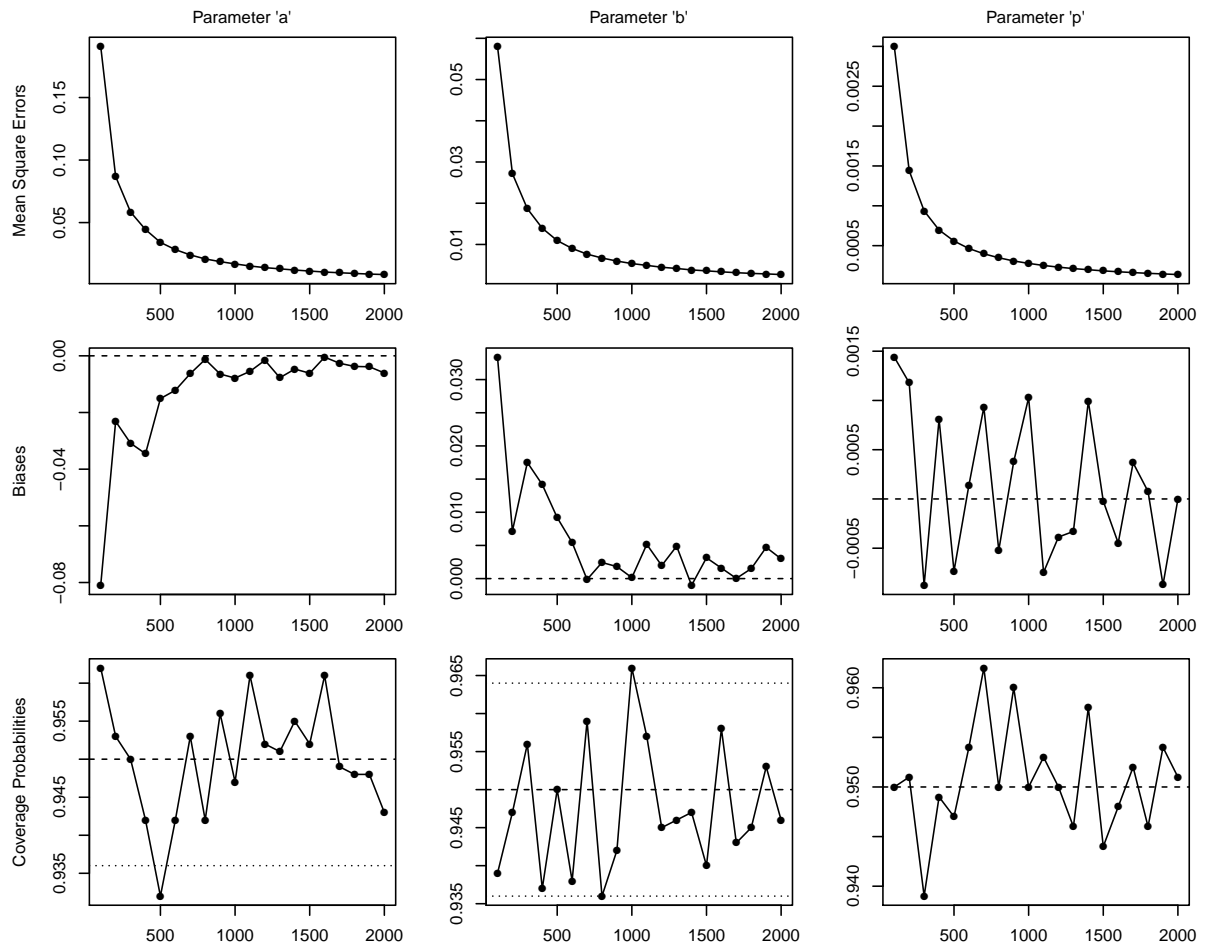


Figure 2.5: Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the Gompertz distribution with $(a, b, p) = (-2, 1, 0.6065)$.

$(a, b, p) = (-2, 1, 0.6065)$. Figure 2.4 and 2.5 show the obtained results.

In third and fourth scenarios we use the defective inverse Gaussian distribution. The random samples are taken from the distribution with parameters $(a, b, p) = (-1, 5, 0.3296)$ and $(a, b, p) = (-1, 1, 0.8646)$, respectively. Figures 2.6 and 2.7 illustrates the results.

The choice of these parameters was taken in order to exemplify cases where we have low and high cure fraction rates (and, therefore, low and high censoring rates).

All four scenarios presented similar results. We can notice the following from them: i) the mean square error decreases very smoothly as the sample size increases and its value is small for any n , specially the parameter p ; ii) the biases are very small for all parameters; iii) the coverage probabilities stays around 95% even for the smallest value of n , for all parameters. This suggests that the delta method provides a good approximation to the

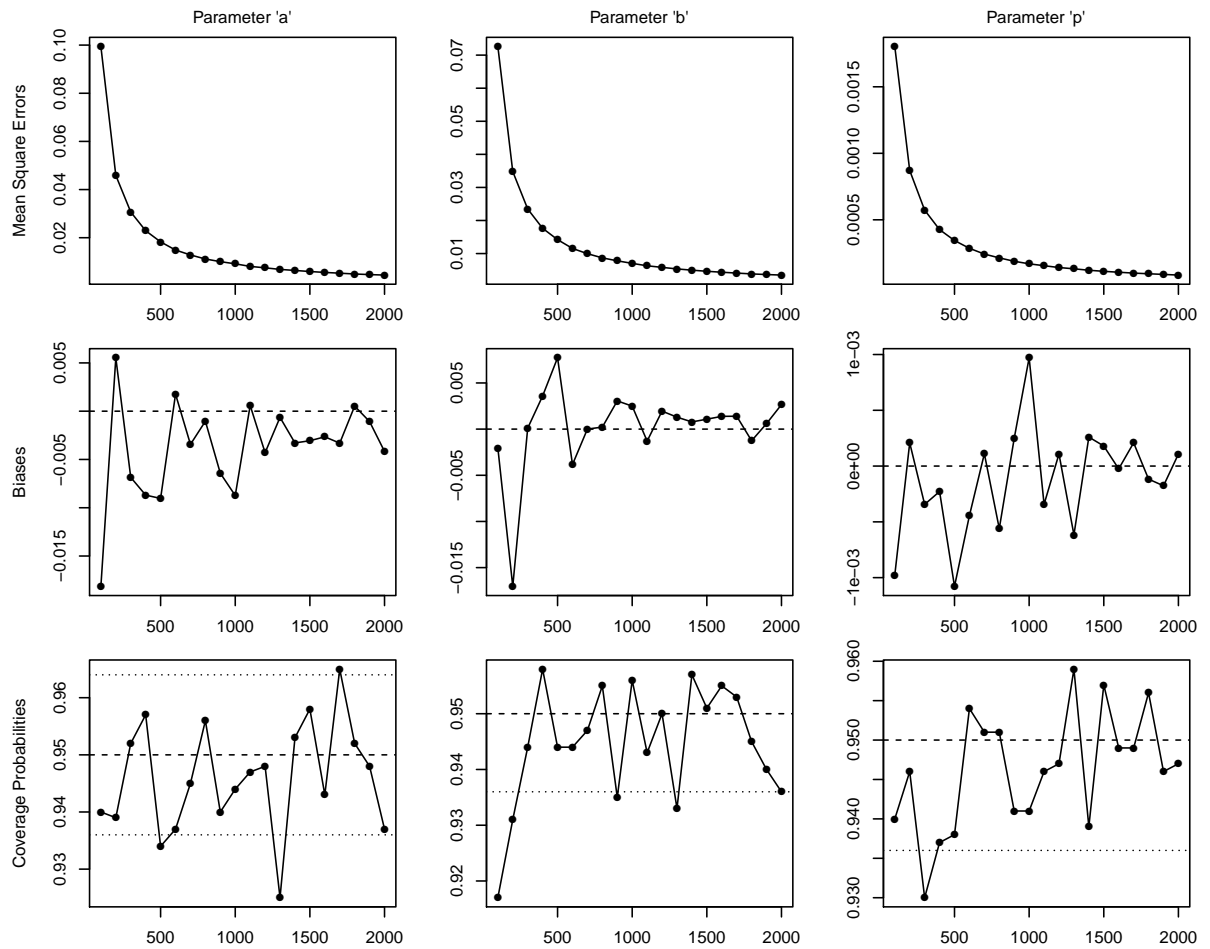


Figure 2.6: Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the inverse Gaussian distribution with $(a, b, p) = (-1, 5, 0.3296)$.

standard deviation of the cure fraction.

Therefore, we can see that these models can give a good point and interval estimation with no need of lots of data. The same results can be observed with different parameter choices.

2.4 Applications

In this section we present the implementation of the defective Gompertz and inverse Gaussian distributions. We fit these distributions in the leukemia, melanoma, second birth and divorce data sets.

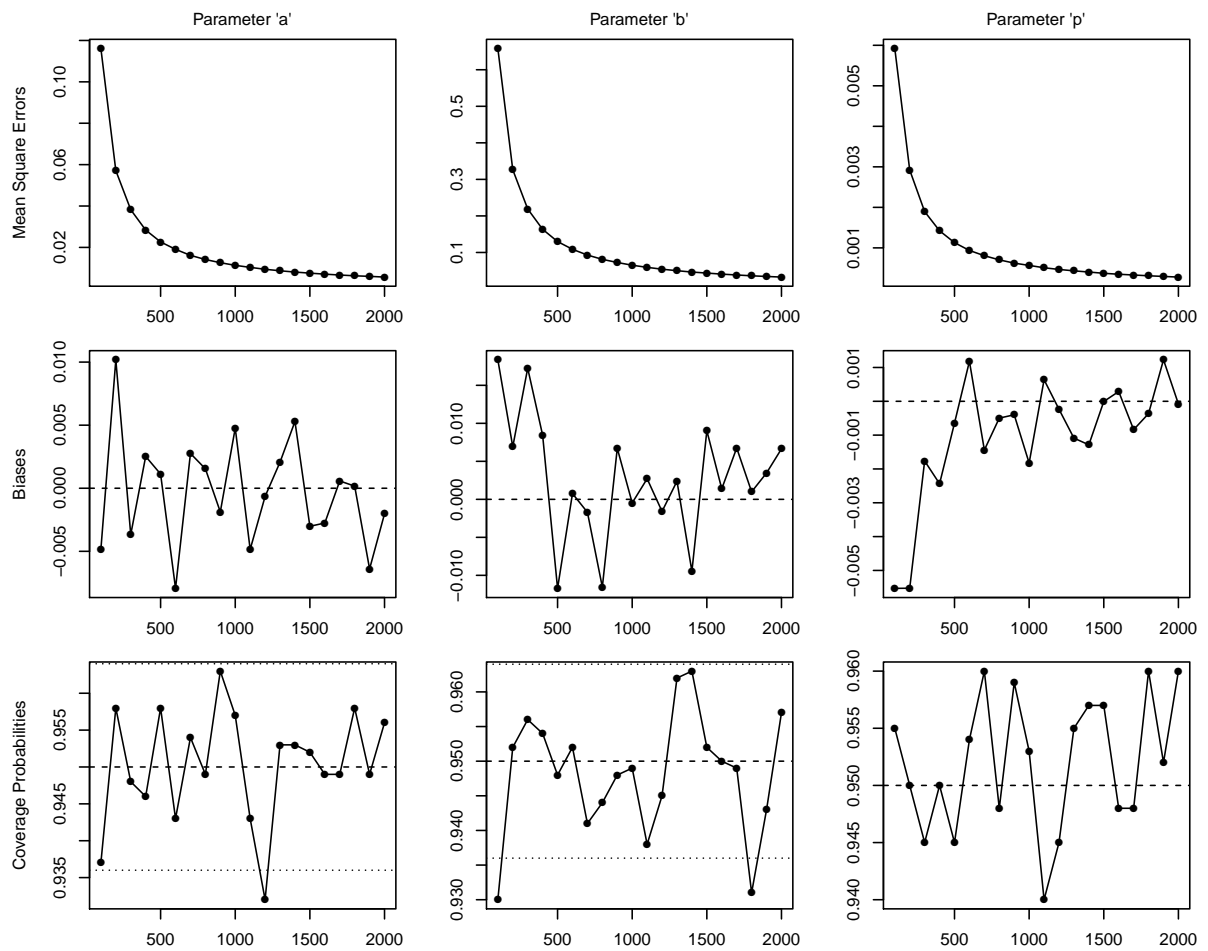


Figure 2.7: Mean squared errors, biases and coverage probabilities of $(\hat{a}, \hat{b}, \hat{p})$ versus n for simulated data from the inverse Gaussian distribution with $(a, b, p) = (-1, 1, 0.8646)$.

This data sets were taken in order to show how the proposed models perform in different kinds of curves given by the Kaplan-Meier estimator.

2.4.1 Leukemia data

Table 2.1: Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the leukemia data set.

Distribution	Parameters	Estimate	Std. Dev.	Lower 95% CI	Upper 95% CI	AIC
Gompertz	a	-1.5103	0.3696	-2.2347	-0.7859	52.58
	b	2.3767	0.5171	1.3633	3.3901	
	p	0.2073	0.0562	0.0971	0.3175	
Inverse Gaussian	a	0.2261	0.3436	-0.4474	0.8996	50.98
	b	3.1391	0.7378	1.6930	4.5852	
	p	-	-	-	-	

The maximum likelihood estimates for the leukemia data set are shown in Table 2.1.

The fitted survival curves are presented in Figure 2.8. We can see that the Gompertz distribution estimate the parameter a in -1.51 , and its confidence interval fully belongs in the negative side of the line. This implies that the Gompertz model suggests a significance of the existence of a cure fraction in this data set. However, this is not confirmed by the inverse Gaussian model. Actually, the last model doesn't capture the presence of a cure fraction at all.

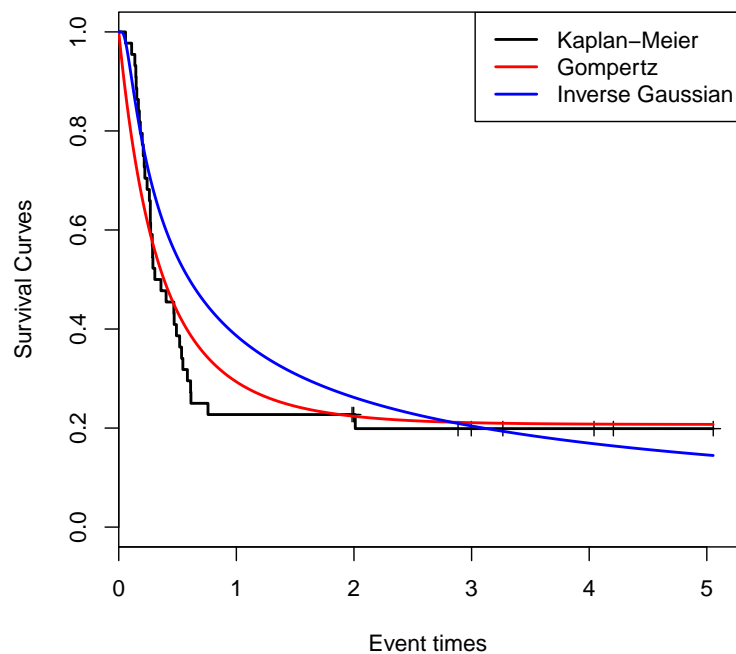


Figure 2.8: Fitted survival curves of the Gompertz and inverse Gaussian distributions in the leukemia data set.

The fitted survival curves confirms the poor fit of the proposed models in this data set. Both model fails to capture the behavior of the Kaplan-Meier curve. The inverse Gaussian distribution shows a small advantage when considering the AIC measure. Probably because of the better fit in the points with small event times.

2.4.2 Melanoma data

In the melanoma data set, we have a similar case to the leukemia one. The maximum likelihood estimates are shown in Table 2.2. The fitted survival curves are presented in

Table 2.2: Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the melanoma data set

Distribution	Parameters	Estimate	Std. Dev.	Lower 95% CI	Upper 95% CI	AIC
Gompertz	a	-0.1313	0.0540	-0.2373	-0.0254	1096.47
	b	0.1792	0.0217	0.1367	0.2217	
	p	0.2555	0.7673	-1.2483	1.7594	
Inverse Gaussian	a	-0.0357	0.0319	-0.0981	0.0268	1062.96
	b	0.4740	0.0427	0.3902	0.5578	
	p	0.1397	0.0000	0.1397	0.1398	

Figure 2.9. We can see that the Gompertz distribution estimate the parameter a in -0.13 , and yet its confidence interval is fully negative. This implies that the Gompertz model suggests a significance of the existence of a cure fraction in this data set. However, this is not confirmed by the inverse Gaussian model. Here, the last model captures the presence of a cure fraction, but with no significance.

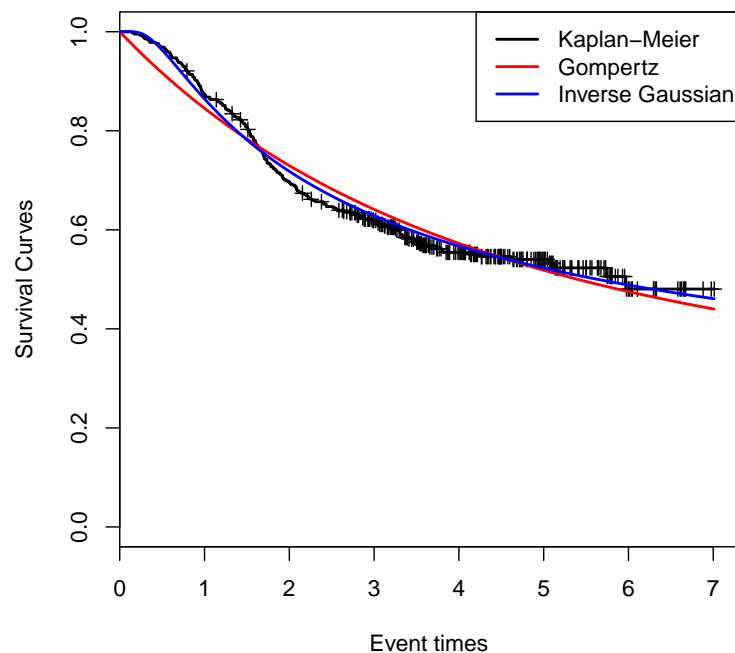


Figure 2.9: Fitted survival curves of the Gompertz and inverse Gaussian distributions in the melanoma data set

The fitted survival curves show that both models do not capture the behavior of the Kaplan-Meier curve as well as one could expect. The inverse Gaussian distribution shows an advantage when considering the AIC measure. It makes sense when considering the fitted curves.

We also notice that the estimated cure fraction by the Gompertz distribution is 0.25, although, its standard deviation is highly estimated by the delta-method. This doesn't match with the results found in the simulation section, and is one more evidence that this distribution is not appropriate for this kind of data .

2.4.3 Second Birth data

Table 2.3: Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the second birth data set

Distribution	Parameters	Estimate	Std. Dev.	Lower 95% CI	Upper 95% CI	AIC
Gompertz	a	1.8924	0.4128	1.0833	2.7014	206.78
	b	0.9402	0.1136	0.7175	1.1628	
	p	-	-	-	-	
Inverse Gaussian	a	1.9184	0.1635	1.5980	2.2388	81.37
	b	1.3080	0.1248	1.0635	1.5525	
	p	-	-	-	-	

In the second birth, we have a case when both models indicate no presence of cure fraction. The maximum likelihood estimates are shown in Table 2.3. The fitted survival curves are presented in Figure 2.10. The proposed distributions estimate the parameter a in the positive range. So, the model suggest no existence of cured elements in this data set. Of course, this is clearly not true.

The fitted survival curves shows that both model fails completely to capture the behavior of the Kaplan-Meier curve. The AIC have a huge difference, but with no practical meaning.

2.4.4 Divorce data

Table 2.4: Maximum likelihood estimates of the Gompertz and inverse Gaussian distributions in the divorce data set

Distribution	Parameters	Estimate	Std. Dev.	Lower 95% CI	Upper 95% CI	AIC
Gompertz	a	-2.5645	0.2294	-3.0140	-2.1149	1517.14
	b	1.9050	0.0875	1.7335	2.0765	
	p	0.4758	0.0119	0.4523	0.4992	
Inverse Gaussian	a	-3.1800	0.2049	-3.5817	-2.7784	1734.44
	b	8.9877	0.2913	8.4169	9.5586	
	p	0.5072	0.0047	0.4980	0.5164	

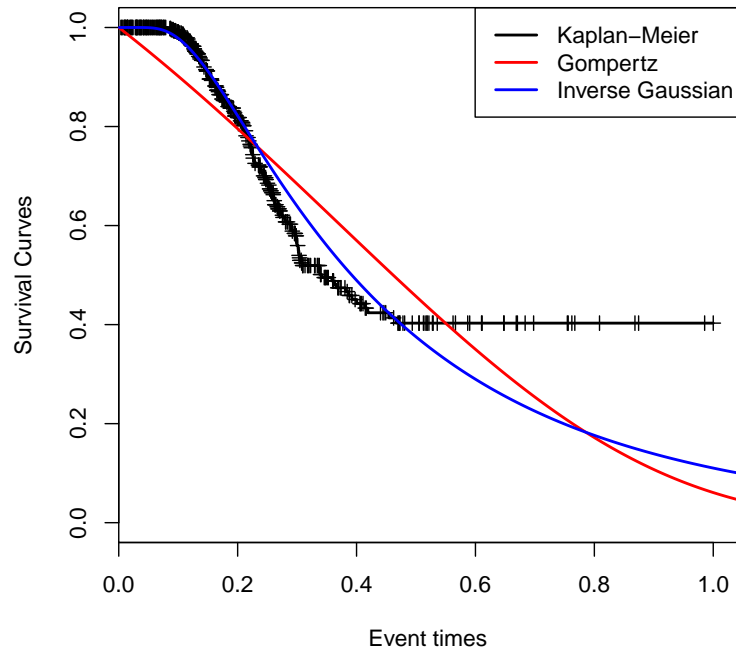


Figure 2.10: Fitted survival curves of the Gompertz and inverse Gaussian distributions in the second birth data set

In the divorce data, we have a case in which both models indicates the presence of cure fraction and gives fairly good models. The maximum likelihood estimates are shown in Table 2.4. The fitted survival curves are presented in Figure 2.11. Both of the proposed distribution estimate the parameter a in the negative range, with confidence interval yet in the negative range. So, the models suggests the existence of cured elements in this data set. The cure fraction is estimated in 0.47 and 0.50 by the Gompertz and inverse Gaussian distributions, respectively.

The fitted survival curves shows that both models do capture the behavior of the Kaplan-Meier curve. The Gompertz distribution captures it better. The inverse Gaussian is better for longer observed times, but for short times, where the massive amount of data is, the Gompertz model capture more closely. This is also evident in terms of AIC.

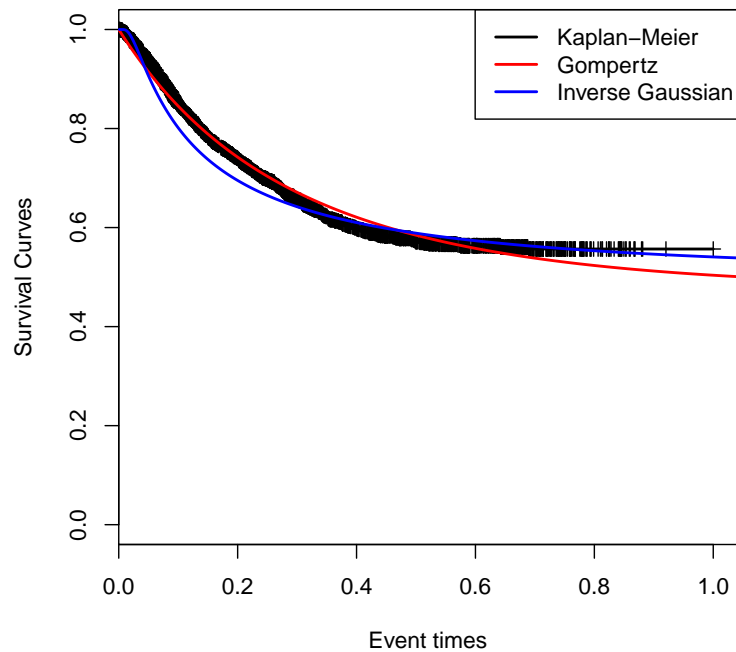


Figure 2.11: Fitted survival curves of the Gompertz and inverse Gaussian distributions in the divorce data set

2.5 Conclusions

In this chapter, we took the two defective models found in literature and performed an simulation study in order to check the validity of its maximum likelihood properties. We noticed that the proposed models can perform well even with very small sample sizes. The cure fraction is estimated precisely, with virtually no bias.

However, the models works only when the data comes from a distribution very close to the proposed ones. We have shown using four data sets, that only these two defective distribution is not enough to deal with cure rate problems. In some cases, the models can get a reasonably fit, at the best. In others, it is completely not appropriated.

This chapter states clearly that if someone wants to use the defective approach to address problems with cured elements, it is necessary to have more distributions to work with. Only the Gompertz and inverse Gaussian defective distributions are not enough. The next chapter introduces two new defective models based on an extension under the Marshall-

Olkin family.

Chapter 3

Marshall-Olkin Family of Defective Models

3.1 Introduction

The aim of this chapter is to propose two new defective distributions based on the Marshall-Olkin family of distributions ([Marshall & Olkin, 1997](#)). This family is obtained by adding an extra parameter to a known baseline distribution.

Suppose $S(t)$ is a known survival function. Then, the extended survival function by the Marshall-Olkin family, $S^*(t)$, is

$$S^*(t) = \frac{rS(t)}{1 - (1 - r)S(t)}$$

for $r > 0$ and $t \in \mathbb{R}$. Simple algebraic manipulations determine the density function of the extended distribution:

$$f^*(t) = \frac{rf(t)}{[1 - (1 - r)S(t)]^2}. \tag{3.1}$$

Particular Marshall Olkin G distributions studied in the literature include the Marshall-Olkin asymmetric Laplace distribution ([Krishna & Jose, 2011](#)), the Marshall-Olkin beta

distribution (Jose *et al.*, 2009b), the Marshall-Olkin Birnbaum-Saunders distribution (Lemonte, 2013), the Marshall-Olkin Burr type XII distribution (Al-Saiari *et al.*, 2014), the Marshall-Olkin discrete uniform distribution (Sandhya & Prasanth, 2014), the Marshall-Olkin Frechet distribution (Krishna *et al.*, 2013), the Marshall-Olkin gamma distribution (Ristic *et al.*, 2007), the Marshall-Olkin Laplace distribution (George & George, 2013), the Marshall-Olkin Lindley distribution (Zakerzadeh & Mahmoudi, 2012), the Marshall-Olkin log-logistic distribution (Gui, 2013a), the Marshall-Olkin Lomax distribution (Ghitany *et al.*, 2007), the Marshall-Olkin Morgenstern Weibull distribution (Jose & Sebastian, 2013), the Marshall-Olkin q-Weibull distribution (Jose *et al.*, 2010), the Marshall-Olkin Weibull distribution (Ghitany *et al.*, 2005), the Marshall-Olkin uniform distribution (Jose & Krishna, 2011a) and the Marshall-Olkin Zipf distribution (Perez-Casany & Casellas, 2014).

Marshall Olkin G distributions have been used to model: daily ozone measurements in New York (Jose *et al.*, 2009b); daily weighted discharge of Neyyar river in Kerala (Jose *et al.*, 2010); frequency of occurrence of words in the novel Moby Dick by Herman Melville (Perez-Casany & Casellas, 2014); length of time until a breakdown is recorded in electrical insulating (Al-Saiari *et al.*, 2014); number of connections of a total of 225409 electronic mail addresses (Perez-Casany & Casellas, 2014); number of days students attended a class for the whole year (Sandhya & Prasanth, 2014); number of miles to first and succeeding major motor failures of buses operated by a large city bus company (Gui, 2013b); number of times that a given paper is cited in a given database (Perez-Casany & Casellas, 2014); permeability values from horizons of the Dominquez field of Southern California (Jose *et al.*, 2009b); remission times of a random sample of bladder cancer patients (Ghitany *et al.*, 2005, 2007); survival times of guinea pigs injected with different doses of tubercle bacilli (Krishna *et al.*, 2013); vinyl chloride data obtained from clean up gradient monitoring wells (Zakerzadeh & Mahmoudi, 2012); waiting times before service of bank customers (Zakerzadeh & Mahmoudi, 2012).

The main purpose of this chapter is to propose two new defective distributions, extending the Gompertz and inverse Gaussian distributions through the Marshall-Olkin family. The details of these extensions including maximum likelihood estimation and the fact that S^*

is defective if S is defective are shown in the next section. Section 3.3 is a simulation study to assess the performance of the maximum likelihood estimators. Section 3.4 illustrates the proposed distributions using real data sets.

3.2 Methodology

In order to construct defective distributions, we propose the use of the Marshall-Olkin class to generalize a given distribution by adding an extra parameter.

The main result of this chapter is that if a given distribution is defective, then its extension under the Marshall-Olkin family will be defective as well.

Theorem 3.1. *If $S(t)$ is defective then $S^*(t)$ is also defective.*

Proof: Suppose the limit of $S(t)$ is equal to $p_0 \in (0, 1)$. Then

$$\lim_{t \rightarrow \infty} S^*(t) = \lim_{t \rightarrow \infty} \frac{rS(t)}{1 - (1-r)S(t)} = \frac{rp_0}{1 - (1-r)p_0} = \frac{rp_0}{rp_0 + 1 - p_0}. \quad (3.2)$$

Since $1 - p_0$ is positive, it is easy to see that the last expression in (3.2) takes a value in $(0, 1)$. The proof is complete. \square

We propose now two new defective distributions: the Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions.

3.2.1 The Marshall-Olkin Gompertz distribution

Using (3.1) with density function in (2.1) and survival function in (2.2), we obtain the Marshall-Olkin Gompertz density function

$$f(t) = \frac{b \cdot r \cdot \exp\left(\frac{b - b \exp(at)}{a} + at\right)}{\left[r - (r - 1) \exp\left(\frac{b - b \exp(at)}{a}\right)\right]^2} \quad (3.3)$$

for $a > 0$, $b > 0$, $r > 0$ and $t > 0$. The corresponding survival function is

$$S(t) = \frac{r \exp \left[-\frac{b}{a} (\exp(at) - 1) \right]}{1 - (1 - r) \exp \left[-\frac{b}{a} (\exp(at) - 1) \right]}. \quad (3.4)$$

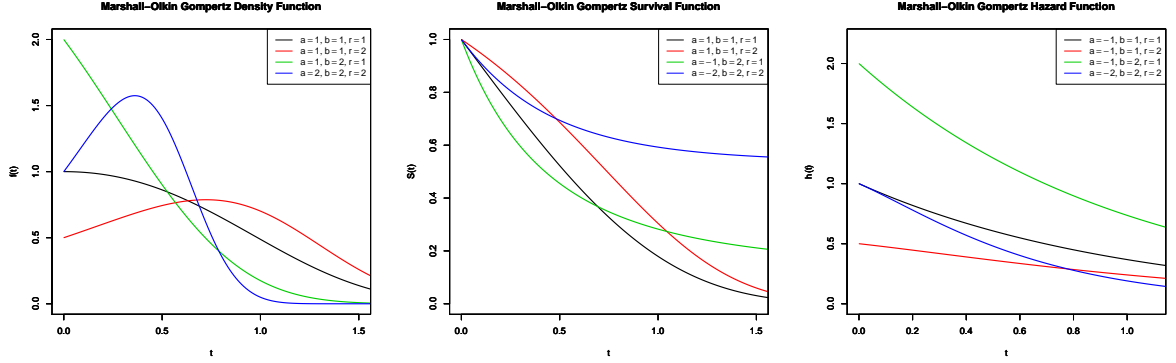


Figure 3.1: Density, survival and hazard functions of the defective Marshall-Olkin Gompertz distribution.

Figure 3.1 illustrates various scenarios for the density, survival and hazard functions of the Marshall-Olkin Gompertz distribution. As in the Gompertz distribution, if $a < 0$ then the Marshall-Olkin Gompertz distribution is defective. Its cure fraction is

$$\lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} 1 - \frac{1}{r e^{\frac{b(e^{at}-1)}{a}} - r + 1} = \frac{rp_0}{1 - (1-r)p_0} = \frac{rp_0}{rp_0 + 1 - p_0} = p,$$

where p_0 is the cure fraction of the defective Gompertz distribution.

3.2.2 The Marshall-Olkin inverse Gaussian distribution

Using (3.1) with density and survival functions of the inverse Gaussian distribution given by (2.3) and (2.4), respectively, we obtain the density function of the Marshall-Olkin inverse Gaussian distribution as

$$f(t) = \frac{r \exp \left(-\frac{(at-1)^2}{2bt} \right)}{\sqrt{2\pi} \sqrt{bt^3} \left[(r-1) \Phi \left(\frac{at-1}{\sqrt{bt}} \right) + (r-1) e^{\frac{2a}{b}} \Phi \left(-\frac{at+1}{\sqrt{bt}} \right) - r \right]^2} \quad (3.5)$$

for $a > 0$, $b > 0$ and $r > 0$. The corresponding survival function is

$$S(t) = \frac{r \left[1 - \Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) - e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right]}{1 - (1 - r) \left[1 - \Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) - e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right]}. \quad (3.6)$$

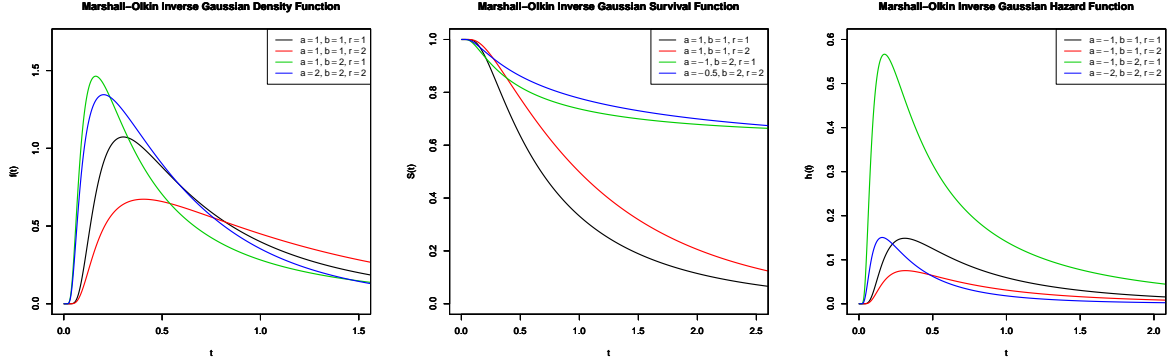


Figure 3.2: Density, survival and hazard functions of the defective Marshall-Olkin inverse Gaussian distribution.

Figure 3.2 illustrates various scenarios for the density, survival and hazard functions of the Marshall-Olkin inverse Gaussian distribution. As in the inverse Gaussian distribution, if $a < 0$ then the Marshall-Olkin inverse Gaussian distribution is also defective. Its cure fraction is

$$\lim_{t \rightarrow \infty} S(t) = \frac{rp_0}{rp_0 + 1 - p_0} = p,$$

where p_0 is the cure fraction of the defective inverse Gaussian distribution.

3.2.3 Inference

Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)'$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)'$ are the censored failure times. Suppose that the data are independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ denotes a vector of parameters. According to Section 1.2.4, the log-likelihood function of $\boldsymbol{\theta}$ can be written

as

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^n (1 - \delta_i) \log S(t_i, \boldsymbol{\theta}).$$

For the Marshall-Olkin Gompertz distribution given by (3.3) and (3.4),

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) &= \text{const} + \sum_{i=1}^n \delta_i \left[\log \left(br \exp \left(\frac{b - b \exp(at)}{a} + at \right) \right) \right] \\ &\quad - \sum_{i=1}^n \delta_i \left[\log \left(\left[r - (r-1) \exp \left(\frac{b - b \exp(at)}{a} \right) \right]^2 \right) \right] \\ &\quad + \sum_{i=1}^n (1 - \delta_i) \left[\log \left(r \exp \left[-\frac{b}{a} (\exp(at) - 1) \right] \right) \right] \\ &\quad - \sum_{i=1}^n (1 - \delta_i) \left[\log \left(1 - (1-r) \exp \left[-\frac{b}{a} (e^{at} - 1) \right] \right) \right]. \end{aligned} \quad (3.7)$$

For the Marshall-Olkin inverse Gaussian distribution given by (3.5) and (3.6),

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) &= \text{const} + \sum_{i=1}^n \delta_i \log \left(r \exp \left(-\frac{(at-1)^2}{2bt} \right) \right) \\ &\quad - \sum_{i=1}^n \delta_i \log \left(\sqrt{bt^3} \left[(r-1) \Phi \left(\frac{at-1}{\sqrt{bt}} \right) \right. \right. \\ &\quad \left. \left. + (r-1) e^{\frac{2a}{b}} \Phi \left(-\frac{at+1}{\sqrt{bt}} \right) - r \right]^2 \right) \\ &\quad + \sum_{i=1}^n (1 - \delta_i) \log \left(r \left[\Phi \left(\frac{at-1}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(-\frac{at+1}{\sqrt{bt}} \right) - 1 \right] \right) \\ &\quad - \sum_{i=1}^n (1 - \delta_i) \log \left(-(r-1) \Phi \left(\frac{at-1}{\sqrt{bt}} \right) \right) \\ &\quad + (r-1) e^{\frac{2a}{b}} \left(\Phi \left(\frac{at+1}{\sqrt{bt}} \right) + 1 \right) - 1. \end{aligned} \quad (3.8)$$

The log likelihoods, (3.7) and (3.8), can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization. In the simulations and real data applications presented in Sections 3.3 and 3.4, the routine `optim` converged all the time, giving unique maximum likelihood estimates. In all cases considered, `optim` did not take more than five seconds for convergence. Confidence

intervals for the parameters were based on asymptotic normality.

We have supposed the usual asymptotes of the maximum likelihood estimates hold. However, defective distributions like the mixture model are not proper distributions. The checking of regularity conditions for the asymptotes by analytical means is not easy. Such conditions have not been checked even for the standard mixture model.

In the next section, we perform an extensive simulation study partly to check the asymptotes of the maximum likelihood estimates. Simulations have been used in many works to assess the behavior of maximum likelihood estimates.

3.3 Simulation Studies

Here, we perform three simulation experiments. The first one is to assess the performance of the maximum likelihood estimates with respect to sample size. The second one is a comparison of defective and mixture models in terms of AIC and cure rate estimates when the data were generated from a defective model. The third one is the same as the second one, but the data were generated from a mixture model. The algorithm to generate according to Section 1.3.

In this first experiment, we simulated one thousand random samples each of size $n = 20, 40, \dots, 1000$. The random samples were taken to come from i) the Marshall-Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$; ii) the Marshall-Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$. We computed the maximum likelihood estimates, \hat{a} , \hat{b} , \hat{r} and \hat{p} , and their standard errors for each sample. These were used to compute the bias, the mean squared error, the coverage probability and the coverage length of \hat{a} , \hat{b} , \hat{r} and \hat{p} for each n .

Figures 3.3 and 3.4 show the plots of the mean squared errors, the biases, the coverage probabilities and the coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions.

We can observe the following from the figures: i) the mean squared errors for all pa-

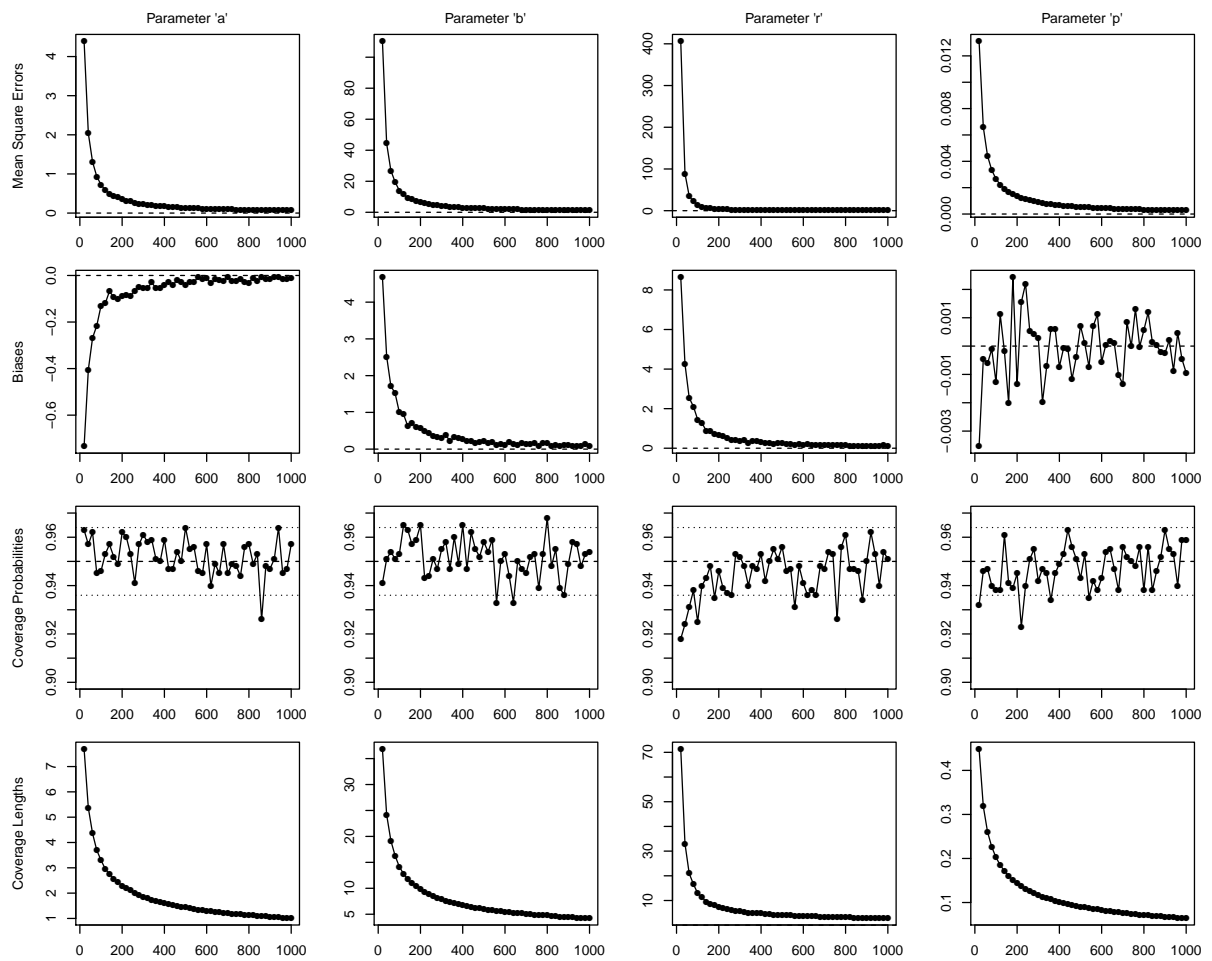


Figure 3.3: Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall-Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$.

rameters generally decrease to zero with increasing n ; ii) the mean squared errors for all parameters appear reasonably close to zero for all $n \geq 600$; iii) the mean squared errors appear smallest for the parameter, p ; iv) the mean squared errors appear largest for the parameters, b and r ; v) the biases for all parameters generally approach zero with increasing n ; vi) the biases for all parameters appear reasonably close to zero for all $n \geq 600$; vii) the biases appear generally negative for the parameter, a ; viii) the biases appear generally positive for the parameter, r ; ix) the biases appear smallest for the parameter, p ; x) the coverage probabilities for all parameters generally approach the nominal level with increasing n ; xi) the coverage probabilities for all parameters appear reasonably close to the nominal level for all $n \geq 800$; xii) the coverage probabilities appear furthest from the nominal level for the parameter, r ; xiii) the coverage lengths for all parameters generally

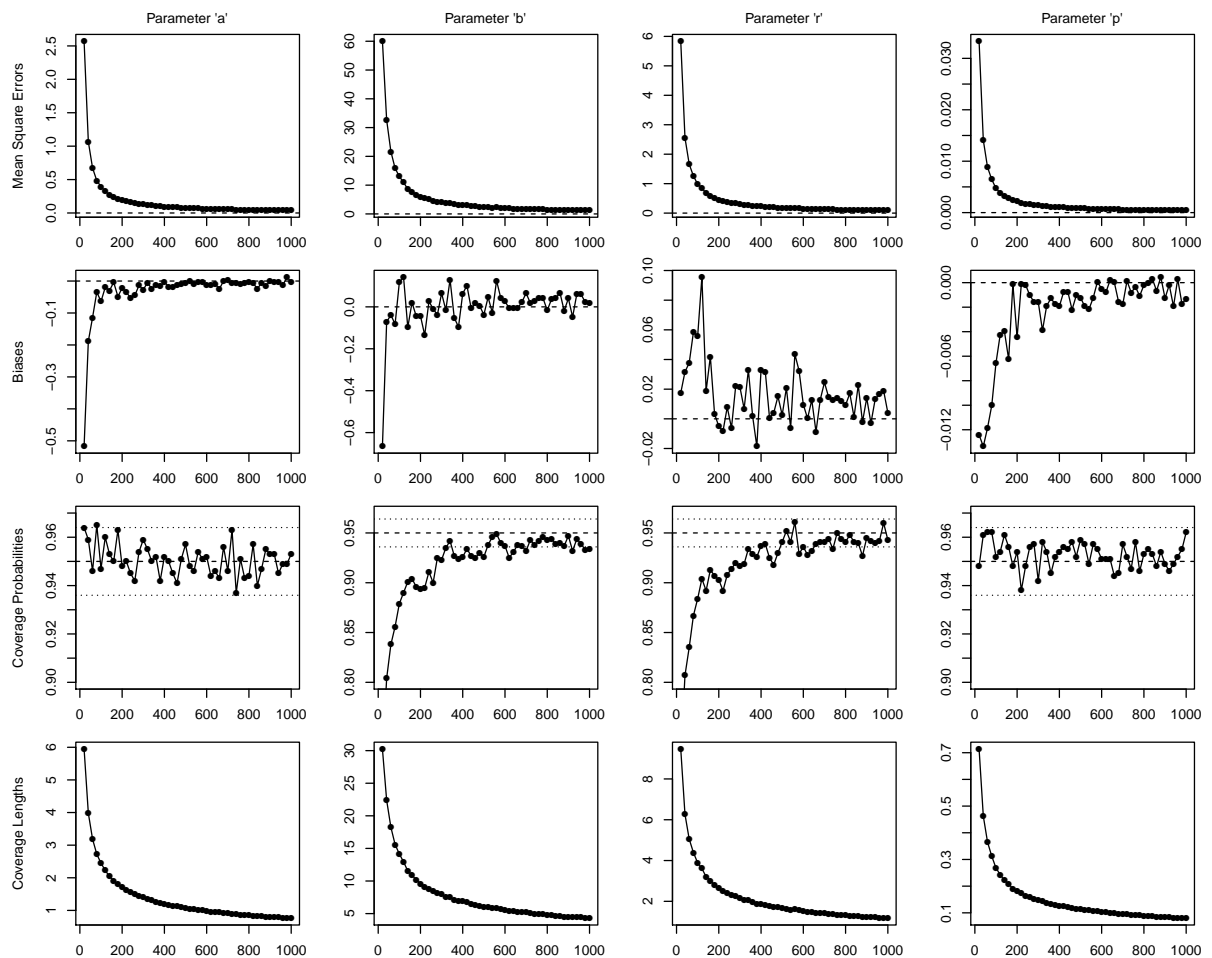


Figure 3.4: Mean squared errors, biases, coverage probabilities and coverage lengths of $(\hat{a}, \hat{b}, \hat{r}, \hat{p})$ versus n for simulated data from the Marshall-Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$.

decrease with increasing n ; xiv) the coverage lengths appear smallest for the parameter, p ; xv) the coverage lengths appear largest for the parameters, b and r .

These observations are for the Marshall-Olkin Gompertz distribution with $(a, b, r, p) = (-3, 4, 2, 0.4172)$ and for the Marshall-Olkin inverse Gaussian distribution with $(a, b, r, p) = (-2, 10, 2, 0.4958)$. But many of the observations were the same when the simulations were repeated for a wide range other values of (a, b, r, p) for both the Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions.

We also noted that the decrease in coverage lengths with increasing n was slow. Indeed, some of the coverage lengths in Figures 3.3 and 3.4 do appear large even for a sample of size 200. Some of the confidence intervals reported in Section 3.4 appear large too.

This suggests a very large sample size may be needed in order to have reliable interval estimates. It is comforting however two of the three real data sets considered in Section 3.4 have sizes over one thousand.

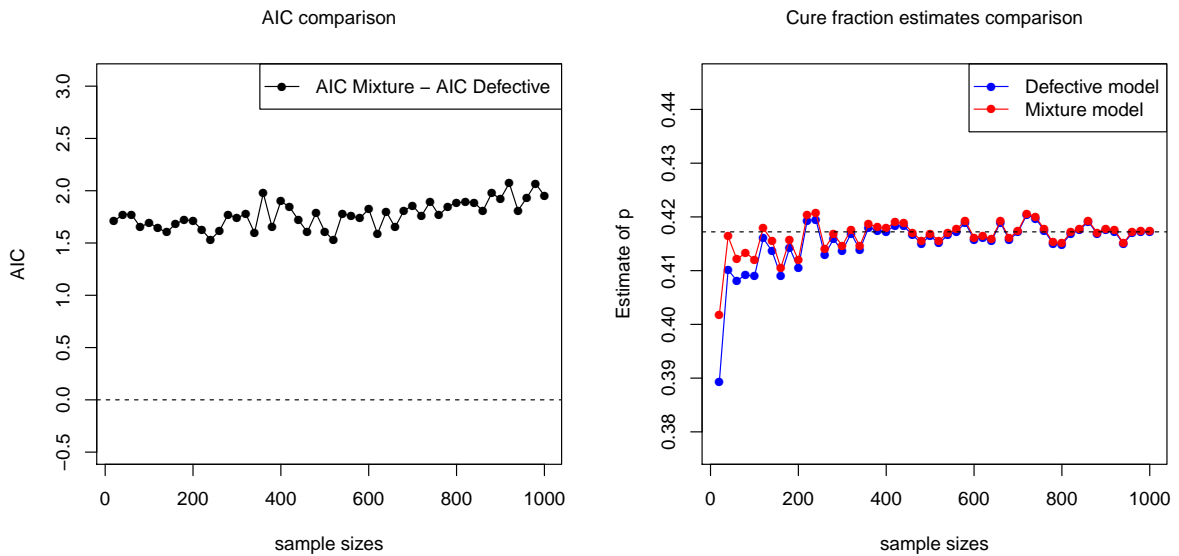


Figure 3.5: In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin Gompertz mixture and defective models, respectively, when the data were generated from a defective model. In the right, the corresponding estimates of p .

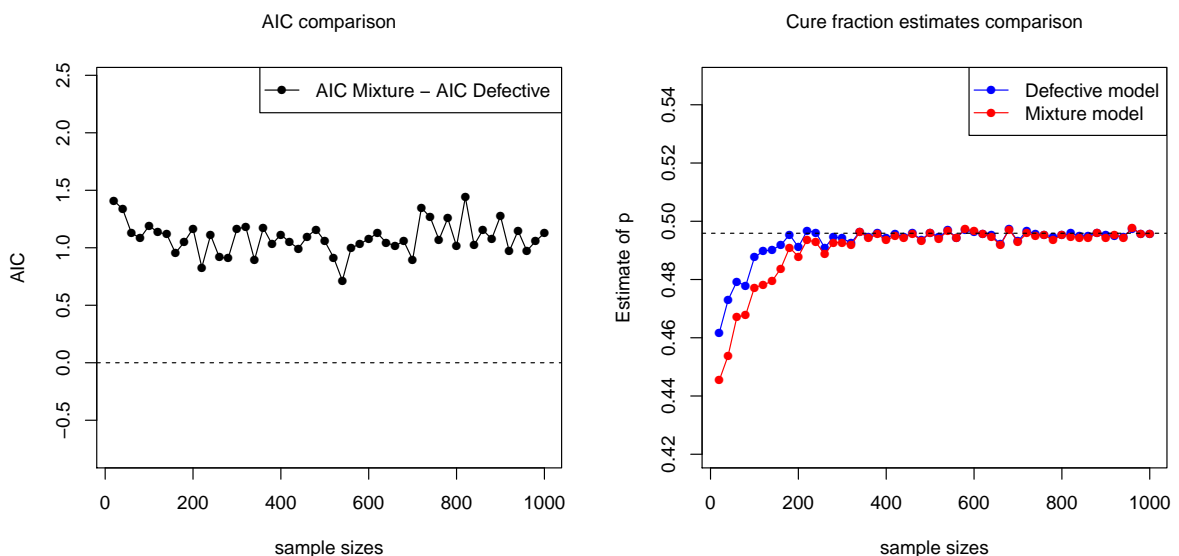


Figure 3.6: In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin inverse Gaussian mixture and defective models, respectively, when the data were generated from a defective model. In the right, the corresponding estimates of p .

The second experiment is to compare the performance of the defective models versus their respective mixture models when the data were generated from defective models. The Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian defective distributions were simulated using $(a, b, r, p) = (-3, 4, 2, 0.4172)$ and $(a, b, r, p) = (-2, 10, 2, 0.4958)$, respectively. They were compared to the corresponding mixture versions. Figures 3.5 and 3.6 (left) provide a comparison in terms of the AIC. The black line represents the difference between the AIC of the mixture model and that of the defective model. The difference is positive for all samples sizes, meaning that the AIC of the defective model is always smaller. On average, the AIC of the defective model is 1.7704 smaller than the AIC of the mixture model for the Marshall-Olkin Gompertz distribution. On average, the AIC of the defective model is 1.0865 smaller for the Marshall-Olkin inverse Gaussian distribution.

Figures 3.5 and 3.6 (right) compare the cure rate estimates for mixture and defective models. We have not compared other parameters since they do not directly relate to the proposed distributions. The estimates of p under both models appear good for the Marshall-Olkin Gompertz distribution, see Figure 3.5. The quadratic error sum for the defective model is 0.00130 and that for the mixture model is 0.00049. This gives a slight advantage for the mixture model. The estimates of p under the defective and mixture models appear good also for the Marshall-Olkin inverse Gaussian distribution, see Figure 3.6. The quadratic error sum for the defective model is 0.00256 and that for the mixture model is 0.00727. Again a small difference but now in favour of the defective model.

The third and the last experiment is to compare the performance of the defective models versus their respective mixture models when the data were generated from mixture models. Mixture versions of the Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions were simulated using $(a, b, r, p) = (0.2, 0.2, 0.2, 0.5)$ and $(a, b, r, p) = (2, 2, 0.5, 0.5)$, respectively. They were compared to the corresponding defective versions. Figures 3.7 and 3.8 (left) compare the models in terms of the AIC. The black line again represents the difference between the AIC of the mixture model and that of the defective model. The differences decrease as n increases for the Marshall-Olkin Gompertz distribution and become less than zero only when $n > 960$, see Figure 3.7. The

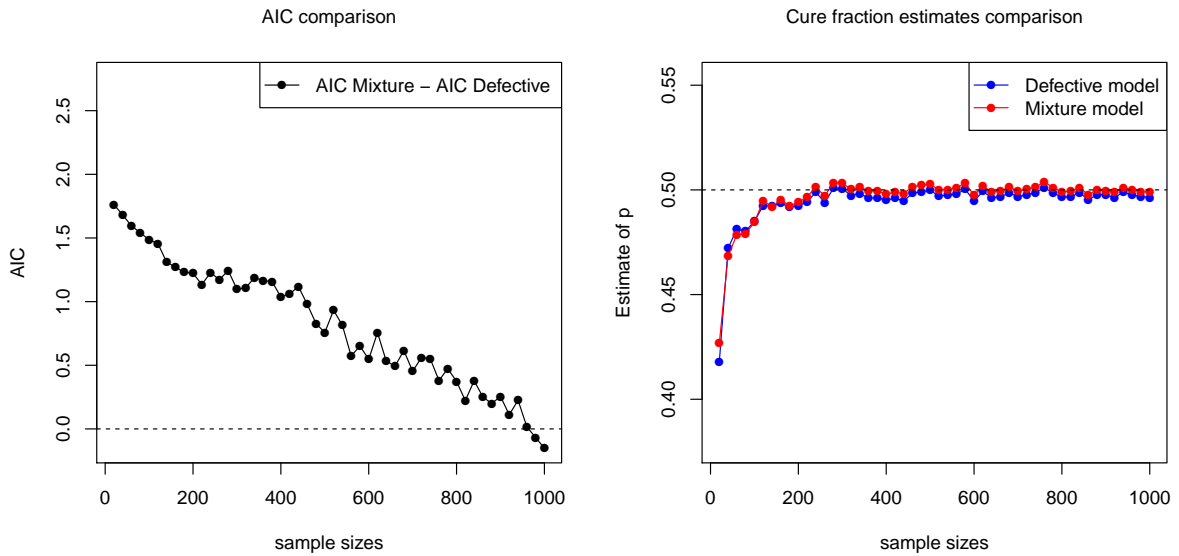


Figure 3.7: In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin Gompertz mixture and defective models, respectively, when the data were generated from a mixture model. In the right, the corresponding estimates of p .

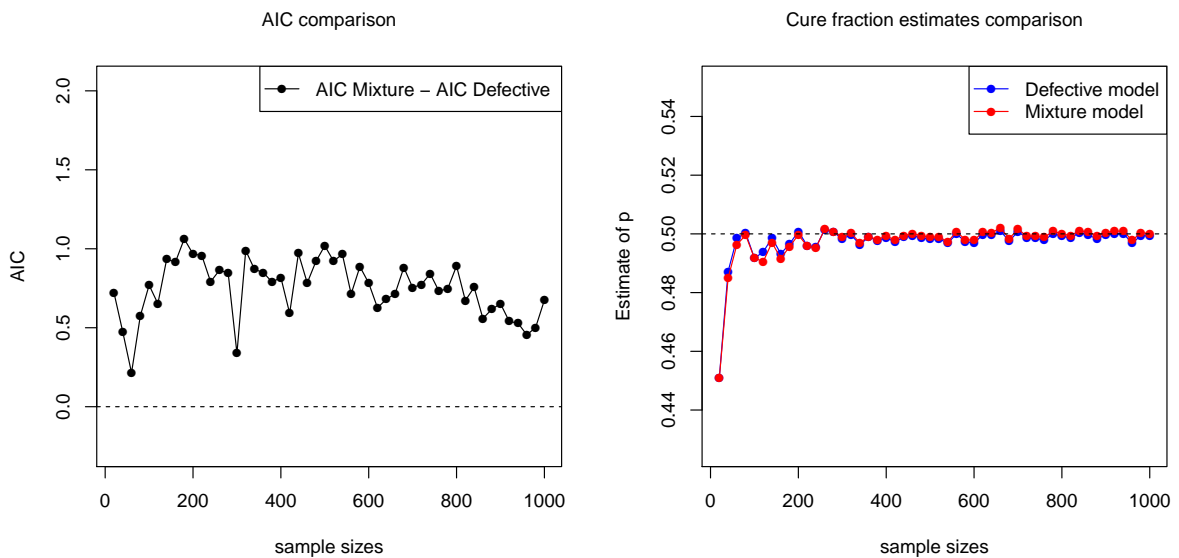


Figure 3.8: In the left, the plotted line represents the difference between the AIC values obtained under the Marshall-Olkin inverse Gaussian mixture and defective models, respectively, when the data were generated from a mixture model. In the right, the corresponding estimates of p .

differences appear positive for all sample sizes for the Marshall-Olkin inverse Gaussian distribution, see Figure 3.8. On average, the AIC of the defective model is 0.8196 smaller than the AIC of the mixture model for the Marshall-Olkin Gompertz distribution. On

average, the AIC of the defective model is 0.7511 smaller for the Marshall-Olkin inverse Gaussian distribution.

Figures 3.7 and 3.8 (right) compare the cure rate estimates for mixture and defective models. The estimates of p under both models appear good for both Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions. The quadratic error sums for the defective and mixture models are 0.00715 and 0.00782, respectively, for the Marshall-Olkin Gompertz distribution. The quadratic error sums for the defective and mixture models are 0.00288 and 0.00302, respectively, for the Marshall-Olkin inverse Gaussian distribution.

The differences found in the second and third experiments are small, but they show clearly that the defective model is better. The results remained the same for a wide range of other parameter choices. That is, the AIC values and the quadratic error sums were smaller for the defective model most of the time for a wide range of parameter choices and for the two distributions. Hence, the defective model can be considered a viable alternative for the mixture model.

Section 3.4 presents three real data applications. The sample size for the first data set is forty four. The sample size for the second data set is over one thousand. The sample size for the third data set is over one thousand eight hundred. Hence, the given point as well as interval estimates for the second and third data sets can be considered accurate enough. But those for the first data set must be treated conservatively.

3.4 Applications

To illustrate the distributions presented we are going to use three data sets: the leukemia, second birth and colon. The three data sets represent three different real scenarios (see Figures 1.1, 1.3 and 1.5). They were chosen carefully to test the flexibility of the proposed distributions under different conditions.

The first and third data sets are about the recurrence of a type of cancer. For these data sets, it is fair to assume that there are individuals who will never have the cancer again,

implying a cure rate. For the second data set, the presence of a cure rate is even more obvious: the immune elements are simply those couples who do not plan to have a second child.

The Gompertz, the Marshall-Olkin Gompertz, the inverse Gaussian and the Marshall-Olkin inverse Gaussian distributions were fitted to the data set via maximum likelihood. The variance of the cure fraction was estimated by using the delta method. The summary of the fitted Gompertz and Marshall-Olkin Gompertz distributions are shown in Tables 3.1, 3.2 and 3.3. The summary of the fitted inverse Gaussian and Marshall-Olkin inverse Gaussian distributions are shown in Tables 3.4, 3.5 and 3.6.

The fitted survival curves of the proposed distributions for the leukemia data set are shown in Figure 3.9. Those for the second birth data set are shown in Figure 3.10. Those for the colon data set are shown in Figure 3.11. Table 3.7 presents the AIC values for all four of the fitted distributions. Figure 3.12 plots the Kaplan-Meier estimates of the survival function versus the predicted values from the proposed distributions. There is a diagonal line in each plot. The closer the points to this line the better the fit.

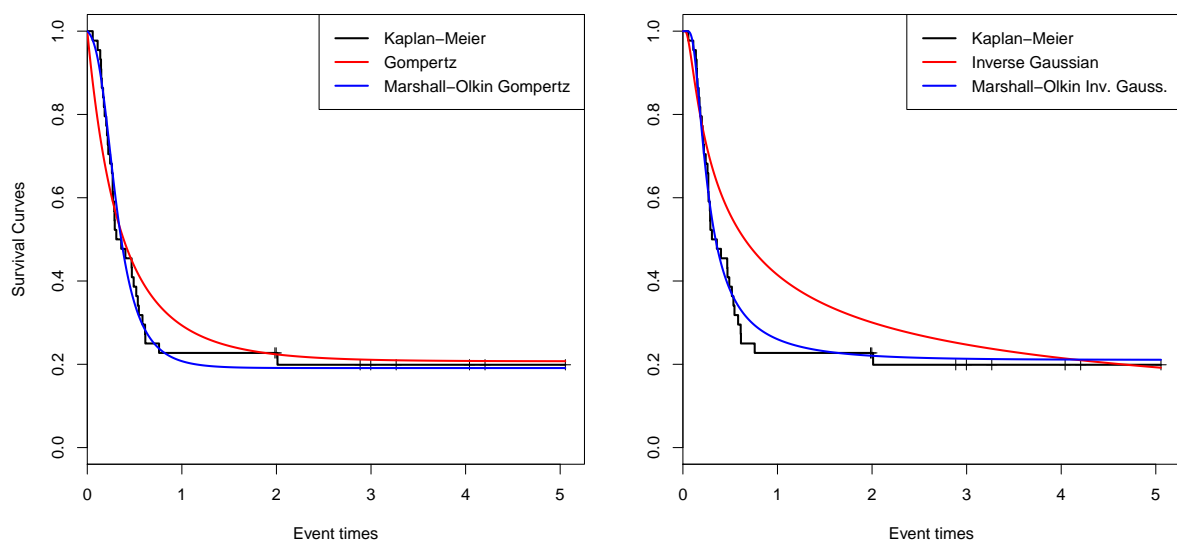


Figure 3.9: Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the leukemia data set.

The Marshall-Olkin Gompertz distribution is a clear improvement over the Gompertz distribution for all three data sets. The fitted survival curve for the former captures the

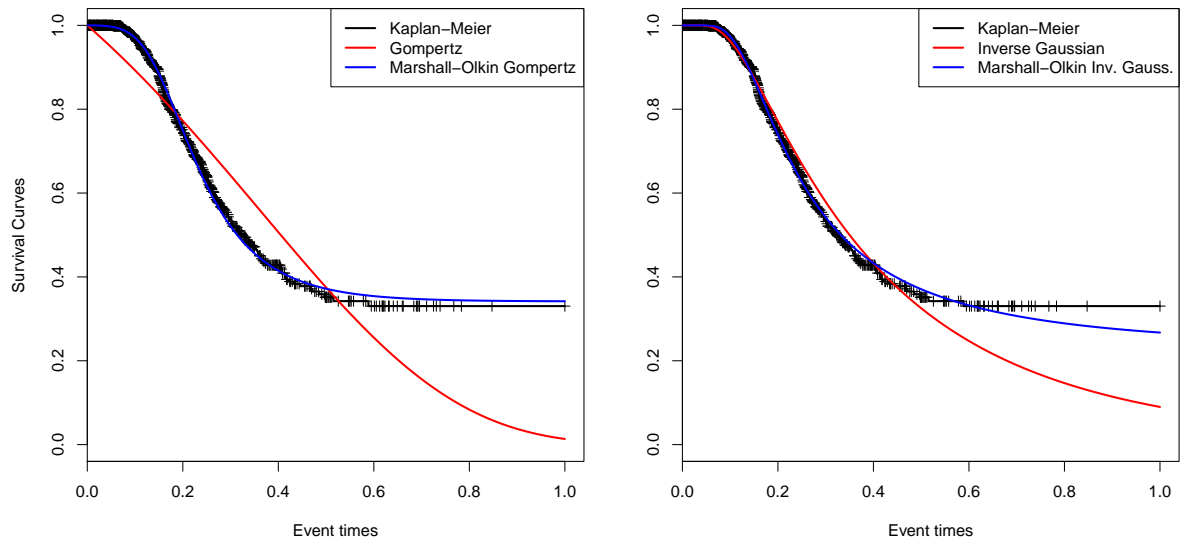


Figure 3.10: Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the second birth data set.

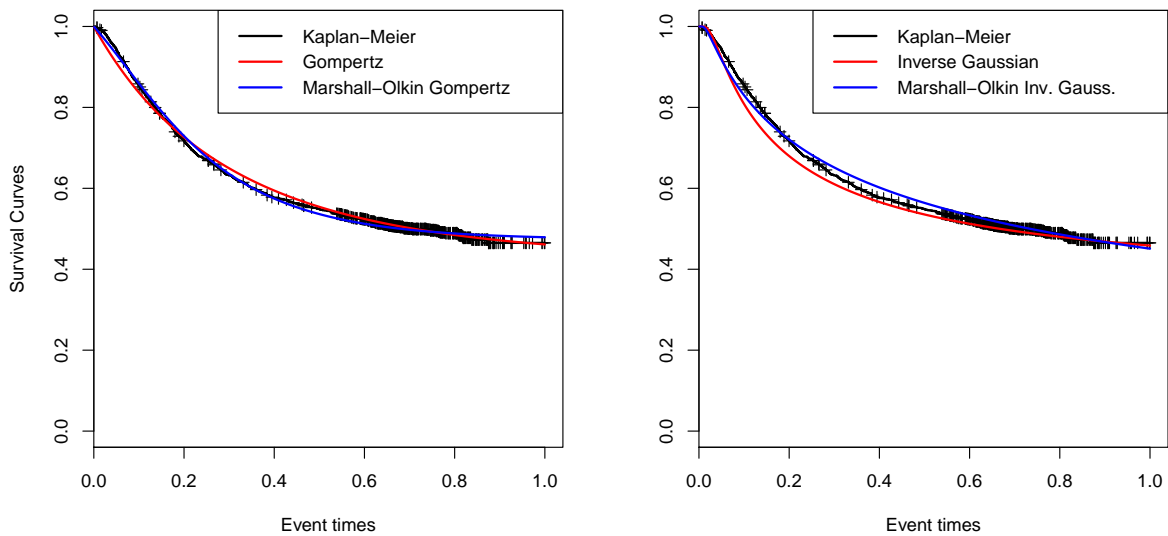


Figure 3.11: Survival curves for the fitted Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions for the colon data set.

Kaplan-Meier curve much better, see Figures 3.9, 3.10, 3.11 and 3.12. For all data sets, the Marshall-Olkin Gompertz distribution estimates a by a negative value with a negative confidence interval. The Gompertz distribution gives a negative interval for a for the leukemia and colon data sets but estimates a by a positive value for the second birth data set. So, the second birth data set is an example, where the baseline distribution does not yield a defective model, while the Marshall-Olkin extension gives a much better fit as a

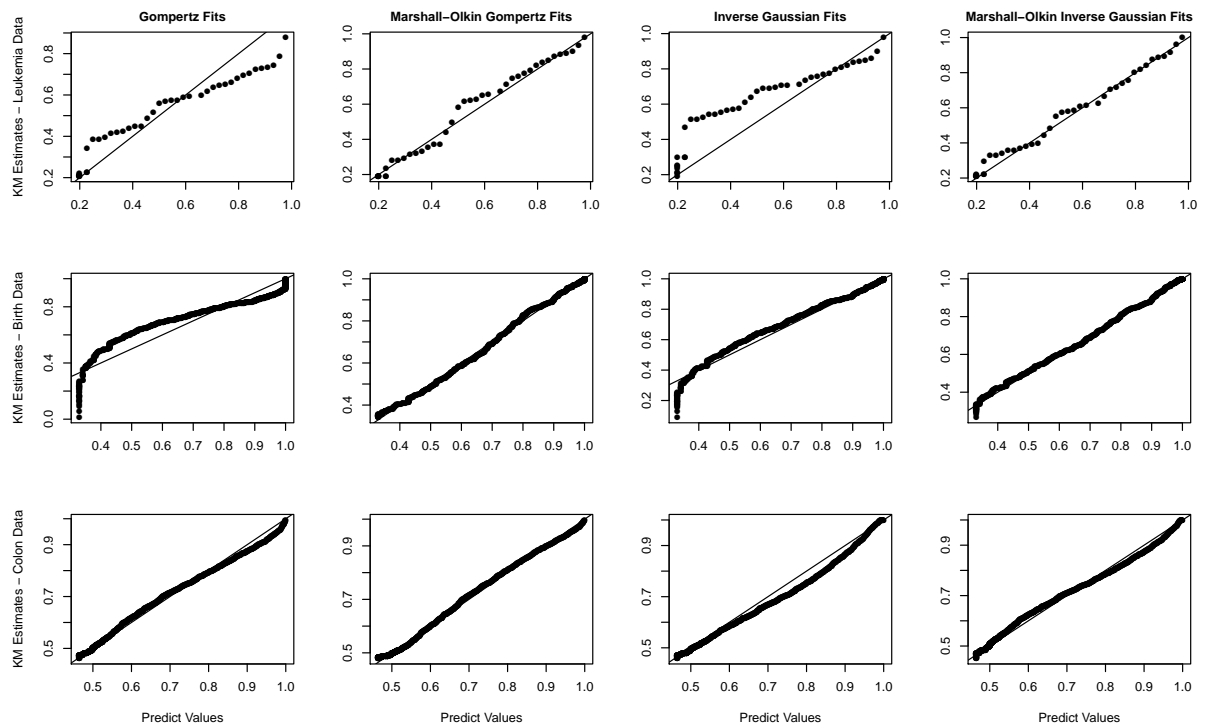


Figure 3.12: Plots of the Kaplan-Meier estimates of the survival function versus the predicted values from the proposed distributions. The top four plots are for the second birth data set. The middle four plots are for the leukemia data set. The bottom four plots are for the colon data set.

defective model. All but the Marshall-Olkin inverse Gaussian distribution appear to estimate the cure fraction in the expected range in relation to the Kaplan-Meier curve. The Marshall-Olkin inverse Gaussian distribution appears to underestimate the cure fraction for the second birth and colon data sets.

The Marshall-Olkin inverse Gaussian distribution is a clear improvement over the inverse Gaussian distribution for all data sets, especially for the leukemia data set. The fitted survival curve for the former captures the Kaplan-Meier curve much better. For the second birth data set, both distributions appear to perform equally well at first, but as time increases the tail of the inverse Gaussian distribution gets distanced from the Kaplan-Meier curve while that of the Marshall-Olkin inverse Gaussian distribution keeps close. For the leukemia data set, the inverse Gaussian distribution estimates a by a very small negative value, giving a very small estimate of the cure fraction not significantly different from zero. For the leukemia and second birth data sets, the Marshall-Olkin inverse Gaussian distribution estimates a by a negative value with a negative confidence

Table 3.1: MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the leukemia data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Gompertz	a	-1.5103	0.3696	-2.4399	-0.9349
	b	2.3767	0.5171	1.5517	3.6405
	p	0.2073	0.0611	0.0875	0.3271
Marshall-Olkin Gompertz	a	-4.0973	0.7898	-5.9783	-2.8082
	b	25.6059	9.1558	12.7051	51.6061
	r	121.9638	150.2286	10.9085	1363.6302
	p	0.191	0.0593	0.0748	0.3071

Table 3.2: MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the second birth data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Gompertz	a	2.4401	0.3178	1.8172	3.063
	b	1.0025	0.0865	0.8329	1.172
	p	-	-	-	-
Marshall-Olkin Gompertz	a	-8.6164	0.6121	-9.9036	-7.4965
	b	84.5282	11.5758	64.6298	110.5529
	r	9449.995	7210.3028	2118.2085	42159.4024
	p	0.3416	0.0145	0.3132	0.37

interval. The estimate of a for the colon data set is close to zero, leading to a very small cure fraction.

The estimate of r for Marshall-Olkin distributions is significantly different from 1, meaning that those distributions provide better fits. This can also be checked in Figure 3.12. The Marshall-Olkin distributions have points closer to the diagonal line than the baseline distributions.

Table 3.7 shows there is a big reduction in AIC values when the Marshall-Olkin Gompertz and Gompertz distributions are compared and when the Marshall-Olkin inverse Gaussian and inverse Gaussian distributions are compared.

For the leukemia and second birth data sets, the best fitting defective model is the Marshall-Olkin inverse Gaussian distribution, the second best fitting model is the Marshall-Olkin Gompertz distribution, the third best fitting model is the inverse Gaussian distribution and the worst fitting model is the Gompertz distribution. For the colon data set, the best fitting defective model is the Marshall-Olkin Gompertz distribution, the second best

Table 3.3: MLEs for the fits of the Gompertz and Marshall-Olkin Gompertz distributions for the colon data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Gompertz	a	-2.3372	0.1772	-2.7117	-2.0145
	b	2.0014	0.1025	1.8103	2.2127
	p	0.4247	0.0115	0.4022	0.4472
Marshall-Olkin Gompertz	a	-4.6989	0.3527	-5.4436	-4.0560
	b	11.1570	1.7778	8.1642	15.2469
	r	8.7515	2.2139	5.3304	14.3685
	p	0.4732	0.0116	0.4505	0.4959

Table 3.4: MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the leukemia data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Inverse Gaussian	a	-0.0003	0.0141	-0.0279	0.0273
	b	3.3612	0.7169	2.2128	5.1057
	p	0.0002	0.0021	0.0000	0.0044
Marshall-Olkin inverse Gaussian	a	-1.3387	0.4147	-2.4567	-0.7294
	b	1.0507	0.2182	0.6993	1.5786
	r	0.0226	0.0247	0.0027	0.1918
	p	0.2107	0.0615	0.0902	0.3312

fitting model is the Gompertz distribution, the third best fitting model is the Marshall-Olkin inverse Gaussian distribution and the worst fitting model is the inverse Gaussian distribution.

Table 3.7 also compares the AIC values between the defective and mixture models based on the Gompertz, Marshall-Olkin Gompertz, inverse Gaussian and Marshall-Olkin inverse Gaussian distributions. The bold value represents the smaller value in the comparison. The defective model performs better for the leukemia data set when based on the Marshall-Olkin Gompertz and Marshall-Olkin inverse Gaussian distributions. The defective model is better for the second birth data set when based on all but the Gompertz distribution. The defective model is better for the colon data set when based on all but the Marshall-Olkin Gompertz distribution.

Table 3.5: MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the second birth data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Inverse Gaussian	a	2.1169	0.1277	1.8666	2.3673
	b	1.5312	0.101	1.3332	1.7293
	p	-	-	-	-
Marshall-Olkin inverse Gaussian	a	-1.5842	0.6094	-3.3668	-0.7454
	b	1.063	0.09	0.9004	1.2549
	r	0.0161	0.017	0.002	0.1274
	p	0.2318	0.0129	0.2065	0.2571

Table 3.6: MLEs for the fit of the Marshall-Olkin inverse Gaussian distribution for the colon data set.

Distribution	Parameter	Point estimate	Std. dev.	Low 95% CI	Upper 95% CI
Inverse Gaussian	a	-1.6688	0.1568	-2.0063	-1.3881
	b	7.3406	0.2901	6.7936	7.9317
	p	0.3653	0.0112	0.3435	0.3872
Marshall-Olkin inverse Gaussian	a	-0.0012	0.0160	-0.0326	0.0302
	b	12.3160	1.0183	10.4734	14.4827
	r	2.8375	0.2190	2.4392	3.3009
	p	0.0005	0.0005	-0.0005	0.0016

Table 3.7: AIC values for the fitted defective distributions compared with their respective mixture models.

Distribution	Leukemia		Birth		Colon	
	Defective	Mixture	Defective	Mixture	Defective	Mixture
Gompertz	52.58	50.74	321.74	197.17	1518.02	1520.02
MO Gompertz	37.16	37.88	80.56	136.75	1488.64	1484.79
Inv. Gaussian	51.38	36.43	99.94	114.36	1597.47	1668.36
MO Inv. Gaussian	35.35	38.34	72.54	109.73	1529.34	1601.22

3.5 Conclusions

We have proposed two new distributions by using an idea due to Marshall and Olkin. These distributions can assume a defective form. In this way, the cure rate can be estimated by models having one less parameter than the usual standard mixture models.

Three real data applications have shown that Marshall-Olkin distributions perform much better than known deflection distributions in terms of likelihood values, proximity to the Kaplan-Meier curve and AIC values. Further investigations are needed to verify the potential of such distributions as cure fraction models.

Chapter 4

Kumaraswamy Family of Defective Models

4.1 Introduction

Here we use the Kumaraswamy family of distributions proposed by [Cordeiro & de Castro \(2011\)](#) as a means for developing distributions that can be defective. Similarly to the Marshall-Olkin family, this family adds two parameters to a baseline distribution, making it more flexible.

The aims of this chapter are to show that if a distribution is defective, then its extension under the Kumaraswamy family of distributions is also defective. Based on that, we propose two new defective distributions by extending the Gompertz and inverse Gaussian distribution under the Kumaraswamy family. In the next section we discuss the Kumaraswamy family of distributions and the two distributions, the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions, generated based on the family. Maximum likelihood estimation and a regression model are also discussed. Section [4.3](#) assesses the finite sample performance and checks the asymptotes of the maximum likelihood estimators. Section [4.4](#) illustrates the usefulness of these two distributions as defective models in three real cancer data sets. The fit of these two distributions is compared to

those of the Gompertz and inverse Gaussian distributions, the baseline distributions, and the standard mixture model.

4.2 Methodology

4.2.1 The Kumaraswamy family of distributions

The Kumaraswamy distribution due to [Kumaraswamy \(1980\)](#) has the probability density and cumulative distribution functions specified by

$$f(x) = rux^{r-1}(1-x^r)^{u-1},$$

$$F(x) = 1 - (1-x^r)^u$$

for $r > 0$, $u > 0$ and $0 < x < 1$. Here, both r and u are shape parameters. This distribution is closely related to the beta distribution, but it is simpler. Its hazard function can be unimodal, uniantimodal, increasing, decreasing and constant. This shows that the Kumaraswamy distribution can model a wide variety of data sets.

[Cordeiro & de Castro \(2011\)](#) proposed the Kumaraswamy family of distributions: given a baseline cumulative distribution function $G(x)$ with $g(x) = dG(x)/dx$ and $S(x) = 1 - G(x)$, they define the Kumaraswamy- G distribution as the one having the probability density, cumulative distribution, survival and hazard rate functions specified by

$$g^*(x) = urg(x)G(x)^{r-1}[1 - G(x)^r]^{u-1}, \quad (4.1)$$

$$G^*(x) = 1 - [1 - G(x)^r]^u, \quad (4.2)$$

$$S^*(x) = [1 - G(x)^r]^u = \{1 - [1 - S(x)]^r\}^u, \quad (4.3)$$

$$h^*(x) = g^*(x)/S^*(x). \quad (4.4)$$

For every given G , this defines a family of distributions. Clearly, the Kumaraswamy- G distribution for $r = u = 1$ is the baseline distribution.

Particular Kumaraswamy G distributions studied in the literature include the Kumaraswamy Birnbaum-Saunders distribution (Saulo *et al.*, 2012), the Kumaraswamy Burr XII distribution (Paranaiba *et al.*, 2013), the Kumaraswamy exponentiated Pareto distribution (Elbatal, 2013a), the Kumaraswamy generalized exponentiated Pareto distribution (Shams, 2013a), the Kumaraswamy generalized gamma distribution (de Pascoa *et al.*, 2011), the Kumaraswamy generalized half normal distribution (Cordeiro *et al.*, 2012d), the Kumaraswamy generalized linear failure rate distribution (Elbatal, 2013b), the Kumaraswamy generalized Lomax distribution (Shams, 2013b), the Kumaraswamy generalized Pareto Distribution (Nadarajah & Eljabri, 2013), the Kumaraswamy generalized Rayleigh distribution (Gomes *et al.*, 2014), the Kumaraswamy geometric distribution (Akinsete *et al.*, 2014), the Kumaraswamy Gumbel distribution (Cordeiro *et al.*, 2012a), the Kumaraswamy half-Cauchy distribution (Ghosh, 2014), the Kumaraswamy inverse exponential distribution (Oguntunde *et al.*, 2014), the Kumaraswamy inverse Rayleigh distribution (Roges *et al.*, 2014), the Kumaraswamy inverse Weibull distribution (Shahbaz *et al.*, 2012), the Kumaraswamy Kumaraswamy distribution (El-Sherpieny & Ahmed, 2014), the Kumaraswamy Lindley distribution (Cakmakyapan & Kadilar, 2014), the Kumaraswamy log-logistic distribution (de Santana *et al.*, 2012), the Kumaraswamy modified inverse Weibull distribution (Aryal & Elbata, 2015), the Kumaraswamy modified Weibull distribution (Cordeiro *et al.*, 2014b), the Kumaraswamy Pareto distribution (Bourguignon *et al.*, 2013), the Kumaraswamy quasi Lindley distribution (Elbatal & Elgarhy, 2013) and the Kumaraswamy Weibull distribution (Cordeiro *et al.*, 2010).

Kumaraswamy G distributions have been used to model: breaking strengths of glass fibers (Paranaiba *et al.*, 2013); breaking strengths of polyester/viscose yarns (Aryal & Elbata, 2015); breaking stress of carbon fibers (Shams, 2013a,b); carbon monoxide levels from several cigarette brands (Gomes *et al.*, 2014); exceedances by the river Nidd at Hunsingore Weir (Nadarajah & Eljabri, 2013); exceedances of flood peaks of the Wheaton river near Carcross in Yukon Territory, Canada (Bourguignon *et al.*, 2013); failure times for epoxy insulation specimens (Gomes *et al.*, 2014); failure times of mechanical components (Cordeiro *et al.*, 2012d); flood data for the Floyd river located in James, Iowa, USA (Cordeiro *et al.*, 2012d); flood discharge of at least seven consecutive days and return

period of 10 years in the Brazilian Pantanal (Cordeiro *et al.*, 2012a); frequencies of the purchases of a brand X breakfast cereals (Akinsete *et al.*, 2014); lifetimes of industrial devices put on life test at time zero (Cordeiro *et al.*, 2014b; de Pascoa *et al.*, 2011); number of absences among shift-workers in a steel industry (Akinsete *et al.*, 2014); stress-rupture life of kevlar epoxy strands subjected to constant sustained pressure (Paranaiba *et al.*, 2013); survival times of cutaneous melanoma (a type of malignant cancer) patients (de Santana *et al.*, 2012); survival times of guinea pigs injected with different doses of tubercle bacilli (Cordeiro *et al.*, 2012d); survival times of patients given radiation therapy and radiation plus chemotherapy (Cordeiro *et al.*, 2014b); the number of millions revolutions reached by ball bearings before fatigue failure (Ghosh, 2014); times of failure and running times of devices from a field-tracking study of a larger system (Cordeiro *et al.*, 2010); times to serum reversal of children exposed to HIV by vertical transmission (de Pascoa *et al.*, 2011; de Santana *et al.*, 2012; Paranaiba *et al.*, 2013); times until bulls reach the weight of 160kg since birth (Roges *et al.*, 2014).

We now state and prove the result that if G is defective then G^* is also defective.

Theorem 4.1. *If $S(t)$ is a survival function of a defective distribution, then $S^*(t)$ is also a survival function of a defective distribution.*

Proof: Suppose the limit of $S(t)$ is equal to $p_0 \in (0, 1)$. Then

$$\begin{aligned} \lim_{t \rightarrow \infty} S^*(t) &= \lim_{t \rightarrow \infty} \{1 - [1 - S(t)]^r\}^u \\ &= \left\{1 - \left[1 - \lim_{t \rightarrow \infty} S(t)\right]^r\right\}^u = \{1 - [1 - p_0]^r\}^u = p \in (0, 1). \end{aligned}$$

Since $0 < 1 - p_0 < 1$, it is easy to see that the last expression in (4.5) takes a value in $(0, 1)$. The proof is complete. \square

Now we propose two new defective distributions: the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions.

4.2.2 The Kumaraswamy Gompertz distribution

Substituting (2.1) and (2.2) into (4.1), (4.3) and (4.4), we obtain the Kumaraswamy Gompertz distribution specified by

$$g^*(t) = urbe^{at} \left(e^{\frac{b-be^{at}}{a}} \right) \left[1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right]^{r-1} \left[1 - \left\{ 1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right\}^r \right]^{u-1},$$

$$S^*(t) = \left\{ 1 - \left[1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right]^r \right\}^u$$

and

$$h^*(t) = urbe^{at} \left(e^{\frac{b-be^{at}}{a}} \right) \left[1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right]^{r-1} \left[1 - \left\{ 1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right\}^r \right]^{-1}$$

for $a > 0$, $b > 0$, $r > 0$, $u > 0$ and $t > 0$. Figure 4.1 illustrates possible shapes of these functions.

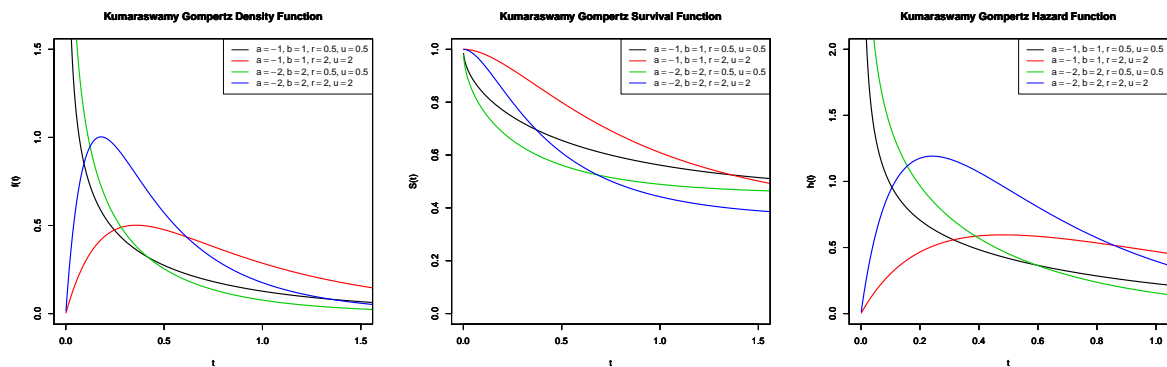


Figure 4.1: Probability density, survival and hazard functions of the defective Kumaraswamy Gompertz distribution.

The Kumaraswamy Gompertz distribution is defective if $a < 0$. According to Theorem 2.1, its cure fraction is

$$\lim_{t \rightarrow \infty} S^*(t) = \{1 - [1 - p_0]^r\}^u = p,$$

where p_0 is the cure fraction of the defective Gompertz distribution.

4.2.3 The Kumaraswamy inverse Gaussian distribution

Substituting (2.3) and (2.4) into (4.1), (4.3) and (4.4), we obtain the Kumaraswamy inverse Gaussian distribution specified by

$$g^*(t) = \left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \\ \times \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{u-1},$$

$$S^*(t) = \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^u$$

and

$$h^*(t) = \left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \\ \times \left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{-1}$$

for $a > 0$, $b > 0$, $r > 0$, $u > 0$ and $t > 0$. Figure 4.2 illustrates possible shapes of these functions.

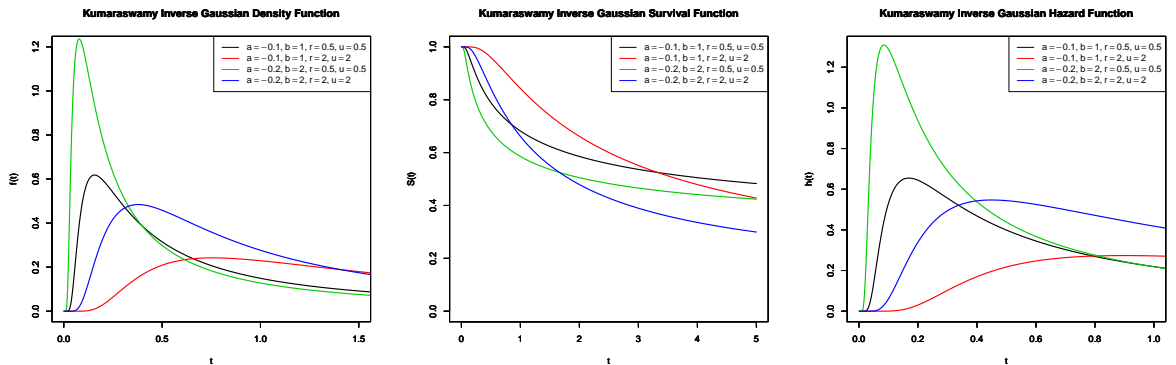


Figure 4.2: Probability density, survival and hazard functions of the defective Kumaraswamy inverse Gaussian distribution.

The Kumaraswamy inverse Gaussian distribution is defective if $a < 0$. According to

Theorem 2.1, its cure fraction is

$$\lim_{t \rightarrow \infty} S^*(t) = \{1 - [1 - p_0]^r\}^u = p,$$

where p_0 is the cure fraction of the defective inverse Gaussian distribution.

Therefore, now we have two new distributions that can assume a defective form. They have a lot more flexibility and capacity to model different kinds of data sets, at the cost of having two extra parameters.

4.2.4 Inference

Here, we discuss estimation issues. Consider a data set $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)'$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)'$ are the censored failure times.

Suppose that the data are independently and identically distributed and come from a distribution with probability density and cumulative distribution functions specified by $f(\cdot, \boldsymbol{\theta})$ and $F(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ denotes a vector of parameters. The log-likelihood function of $\boldsymbol{\theta}$ is

$$\log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + \sum_{i=1}^n (1 - \delta_i) \log [1 - F(t_i, \boldsymbol{\theta})].$$

For the Kumaraswamy Gompertz distribution,

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) &= \text{const} + \sum_{i=1}^n \delta_i \left[\log \left(urbe^{at} \left(e^{\frac{b-be^{at}}{a}} \right) \left[\left(e^{\frac{b-be^{at}}{a}} \right) \right]^{r-1} \left[1 - \left\{ 1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right\}^r \right]^{u-1} \right) \right] \\ &\quad + \sum_{i=1}^n (1 - \delta_i) \left[\log \left(\left\{ 1 - \left[1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right]^r \right\}^u \right) \right]. \end{aligned}$$

For the Kumaraswamy inverse Gaussian distribution,

$$\begin{aligned} \log L(\boldsymbol{\theta}, \mathbf{D}) &= \text{const} + \sum_{i=1}^n \delta_i \left[\log \left(\left(\frac{ur}{\sqrt{2b\pi t^3}} \exp \left\{ -\frac{(1-at)^2}{2bt} \right\} \right) \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^{r-1} \right) \right] \\ &\quad + \sum_{i=1}^n \delta_i \left[\log \left(\left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{\frac{2a}{b}} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^{u-1} \right) \right] \\ &\quad + \sum_{i=1}^n (1 - \delta_i) \left[\log \left(\left\{ 1 - \left[\Phi \left(\frac{-1+at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1-at}{\sqrt{bt}} \right) \right]^r \right\}^u \right) \right]. \end{aligned}$$

The log-likelihoods functions can be maximized numerically to obtain the maximum likelihood estimates. There are various routines available for numerical maximization. Confidence intervals for the parameters were based on asymptotic normality.

4.2.5 The Kumaraswamy- G regression model

Here, we briefly discuss a possible approach to use the Kumaraswamy family with covariate information. Suppose that $\mathbf{x}' = (1, x_1, \dots, x_p)$ is a vector of covariates from a data set and $S(t)$ is the survival function of a baseline distribution. Then, the Kumaraswamy- G regression model is given by

$$S^*(t|\mathbf{x}) = \{S^*(t)\}^{\exp(\beta'\mathbf{x})} = \{1 - [1 - S(t)]^r\}^{u \exp(\beta'\mathbf{x})}$$

for $t > 0$, $r > 0$, $u > 0$ and $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ a vector of regression coefficients.

This approach has a simple interpretation when the Kumaraswamy- G distribution is used as a defective model. Suppose that p is the cure fraction of a Kumaraswamy- G defective model, then the limit of the survival function is

$$\lim_{t \rightarrow \infty} S^*(t|\mathbf{x}) = p^{\exp(\beta'\mathbf{x})}.$$

It is easy to check that the cured proportion increases when $\beta'\mathbf{x} < 0$ ($\exp(\beta'\mathbf{x}) < 1$) and decreases when $\beta'\mathbf{x} > 0$ ($\exp(\beta'\mathbf{x}) > 1$). Therefore, negative coefficients in β contribute to increase the cure fraction and positive coefficients contribute to decrease it.

The survival function of the Kumaraswamy Gompertz regression model is given by

$$S^*(t|\mathbf{x}) = \left\{ 1 - \left[1 - \left(e^{\frac{b-be^{at}}{a}} \right) \right]^r \right\}^{u \exp(\beta'\mathbf{x})}.$$

The survival function of the Kumaraswamy inverse Gaussian regression model is given by

$$S^*(t|\mathbf{x}) = \left\{ 1 - \left[\Phi \left(\frac{-1 + at}{\sqrt{bt}} \right) + e^{2a/b} \Phi \left(\frac{-1 - at}{\sqrt{bt}} \right) \right]^r \right\}^{u \exp(\beta'\mathbf{x})}.$$

An application of these models to a melanoma data set is described in Section 4.4.1.

4.3 Simulation Studies

Here, we assess the performance of the maximum likelihood estimates with respect to sample size to show, among other things, that the usual asymptotes of maximum likelihood estimators still hold for defective distributions. The assessment is based on simulations. The description of data generation is in Section 1.3.

Table 4.1: Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy Gompertz distribution.

Sample sizes (Censoring rates)	Mean estimates					Standard deviation estimates				
	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	\hat{a}	b	\hat{r}	\hat{u}	\hat{p}
$(a, b, r, u, p) = (-1, 5, 5, 0.5, 0.1823)$										
100 (35.80%)	-0.9887	7.0934	8.3434	0.7775	0.1697	0.444	6.1046	7.3722	2.5008	0.2493
250 (33.28%)	-1.0001	6.0915	6.761	0.8496	0.1787	0.2449	3.6785	3.8297	1.6763	0.186
500 (31.36%)	-0.9994	5.6122	5.8102	0.7077	0.1809	0.1571	2.4126	2.0546	0.8084	0.1499
1000 (29.89%)	-1.0076	5.1474	5.2758	0.636	0.1819	0.1045	1.5693	1.1801	0.43	0.1234
2000 (28.92%)	-1.0037	5.101	5.1547	0.5495	0.1824	0.0703	1.0868	0.7892	0.2142	0.1023
5000 (27.65%)	-1.0009	5.0181	5.042	0.5195	0.1822	0.0419	0.6662	0.4727	0.1153	0.0799
$(a, b, r, u, p) = (-2, 2, 2, 2, 0.3605)$										
100 (50.68%)	-1.9845	6.4412	3.1071	1.6974	0.3506	1.0041	13.1274	2.3258	13.5832	0.2808
250 (48.40%)	-1.8958	6.6877	2.8202	1.3912	0.356	0.5	8.3259	1.2862	6.1321	0.2035
500 (47.21%)	-1.8916	5.2443	2.4304	1.5544	0.3579	0.3246	5.4028	0.6895	4.5243	0.1663
1000 (46.16%)	-1.9455	3.7424	2.2102	2.2102	0.3597	0.2323	3.4466	0.4052	5.0514	0.1361
2000 (45.11%)	-1.9719	2.7868	2.0903	2.374	0.3598	0.148	2.0106	0.2358	3.4315	0.1122
5000 (44.16%)	-1.9881	2.3317	2.0379	2.3209	0.3604	0.0943	1.264	0.1468	2.1978	0.0879
$(a, b, r, u, p) = (-3, 11, 2, 0.2, 0.5503)$										
100 (66.01%)	-2.9369	24.1308	3.9845	0.156	0.5254	1.572	58.2769	5.3884	0.9326	0.2893
250 (64.09%)	-3.0368	20.6565	2.99	0.2086	0.5457	0.8143	26.4515	1.7769	0.7126	0.2031
500 (62.89%)	-2.9784	19.0204	2.5784	0.209	0.5479	0.5328	19.116	1.246	0.5011	0.1667
1000 (62.29%)	-2.9769	16.7299	2.3478	0.2251	0.5502	0.3483	12.936	0.6584	0.399	0.1363
2000 (61.46%)	-2.9955	13.4688	2.1343	0.2534	0.5504	0.237	8.4298	0.3923	0.3186	0.1128
5000 (60.71%)	-3.0013	11.6581	2.0375	0.2321	0.5502	0.1463	5.0363	0.2252	0.166	0.0889

Tables 4.3 and 4.3 describe the results for three different parameters values. The parameter values were selected in order to show the simulation results for small, medium and large cure rates. We took the sample size as $n = 100, 250, 500, 1000, 2000, 5000$. Each sample was replicated 1000 times. In each replication, we computed the maximum likelihood estimates of the parameters, maximum likelihood estimate of the cure fraction, and the standard deviation of the cure fraction (obtained using the delta method). The averages of these estimates over the 1000 replications are reported in Tables 4.3 and 4.3. The tables also report the censoring rates in addition to the sample size. Note that censoring rates decrease as sample size increases, but they are still in an appropriate range

Table 4.2: Simulation of the maximum likelihood estimates for mean and standard deviation of the Kumaraswamy inverse Gaussian distribution.

Sample sizes (Censoring rates)	Mean estimates					Standard deviation estimates				
	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}
$(a, b, r, u, p) = (-1, 3, 1, 1, 0.2834)$										
100 (34.38%)	-0.491	4.1373	1.0232	0.9545	0.2771	1.033	13.598	2.6446	1.0137	0.2759
250 (32.53%)	-0.647	6.4553	1.4828	1.055	0.2811	0.8367	13.587	2.3245	0.6701	0.193
500 (31.48%)	-0.632	6.1289	1.4636	1.0436	0.2819	0.5705	9.1473	1.6305	0.4533	0.1514
1000 (31.01%)	-0.555	4.3093	1.1954	1.0104	0.2835	0.3652	4.602	1.0116	0.3298	0.1245
2000 (30.48%)	-0.534	3.7087	1.1205	1.0092	0.2832	0.2406	2.733	0.6694	0.2239	0.1034
5000 (30.10%)	-0.504	3.1797	1.022	0.9942	0.2836	0.139	1.4721	0.3886	0.133	0.0815
$(a, b, r, u, p) = (-0.25, 1, 2, 2, 0.3996)$										
100 (47.56%)	-0.233	1.4029	1.9374	2.3161	0.3897	0.5344	5.4392	5.6379	4.9044	0.2851
250 (45.71%)	-0.289	1.9843	2.5624	2.3638	0.3983	0.4737	6.0378	5.2413	2.9732	0.1967
500 (44.60%)	-0.315	2.2856	2.8774	2.2603	0.3985	0.3737	5.2263	4.2558	1.8689	0.1585
1000 (43.67%)	-0.293	1.8463	2.5862	2.1577	0.3981	0.2275	2.6732	2.5578	1.2013	0.1304
2000 (43.19%)	-0.282	1.5384	2.4238	2.1246	0.3991	0.1508	1.5268	1.6917	0.8215	0.1084
5000 (42.66%)	-0.264	1.1796	2.1783	2.0629	0.3994	0.0831	0.6675	0.9237	0.4833	0.0854
$(a, b, r, u, p) = (-1, 10, 2, 0.5, 0.5741)$										
100 (62.00%)	-0.629	5.6421	1.0865	0.4459	0.5608	1.7284	24.336	3.6897	0.6395	0.3045
250 (60.97%)	-0.744	8.4841	1.4547	0.4219	0.5694	1.217	22.068	2.8561	0.3205	0.2099
500 (59.86%)	-0.975	12.473	2.0029	0.4701	0.5718	1.075	23.028	2.6607	0.228	0.158
1000 (59.51%)	-1.095	14.577	2.2936	0.4948	0.5735	0.851	19.398	2.1499	0.1635	0.1296
2000 (59.18%)	-1.077	13.452	2.2451	0.4961	0.5744	0.5836	12.844	1.5027	0.1097	0.1077
5000 (58.87%)	-1.044	11.539	2.1227	0.5002	0.5747	0.3412	6.7388	0.8807	0.0657	0.085

considering the cure fraction (not too low when the cure fraction is small and not too high when the cure fraction is large).

The following can be observed from the tables: i) the biases and standard deviations generally decrease as sample size increases; ii) large sample sizes are needed for the biases of a, b, r, u to become smaller than 10^{-2} , the biases of b and r do not become smaller than 10^{-2} even for $n = 5000$; iii) the biases of p become smaller than 10^{-3} for much smaller sample sizes; iv) although the standard deviations decrease as n increases, their values are not so small even for n as large as 5000. Sample sizes larger than 5000 are not realistic in practical survival analysis.

These observations are for specific parameter values and $n = 100, 250, 500, 1000, 2000, 5000$. But the same observations held for a wide range of other parameter values and a wide range of the values of $n < 5000$.

4.4 Applications

We illustrate the distributions in Section 2 using three real cancer data sets. The data sets have different sample sizes and exhibit distinct Kaplan-Meier curves.

The following distributions were fitted to each of the data sets: the Gompertz distribution, the inverse Gaussian distribution, the Kumaraswamy Gompertz distribution and the Kumaraswamy inverse Gaussian distribution. This allows use to see if the Kumaraswamy Gompertz distribution provides a better fit than the Gompertz distribution or if the Kumaraswamy inverse Gaussian distribution provides a better fit than the inverse Gaussian distribution.

For each fitted distribution, we provide the maximum likelihood estimates, their standard errors (in parenthesis), the values of the AIC (Akaike Information Criterion), the values of the BIC (Bayesian Information Criterion) and the values of the CAIC (Consistent Akaike Information Criterion). As discussed in the methodology section, the point estimate of p was obtained using the estimates of the other parameters. The standard deviation of the estimator of p was calculated using the delta method with a first-order Taylor's approximation.

An estimate of the survival times was also obtained using the Kaplan-Meier estimator. As an experiment comes to an end, some of the observed elements do not fail. However, it is not known if these elements would fail at some given point, after the end of the observed period. Depending on what is being analyzed, there are two possibilities; that the elements will never fail, or will fail at some point in the future if given enough time. The end of a Kaplan-Meier curve is an estimate of the cure fraction, with the assumption that elements in the study which did not fail, will never fail at all at any point in the future, after the end of the study. It should be noted that this is a difficult assumption to make.

We also compared the parametric curves with the Kaplan-Meyer curve. Closer they are to each other, the better the fit.

4.4.1 Melanoma data

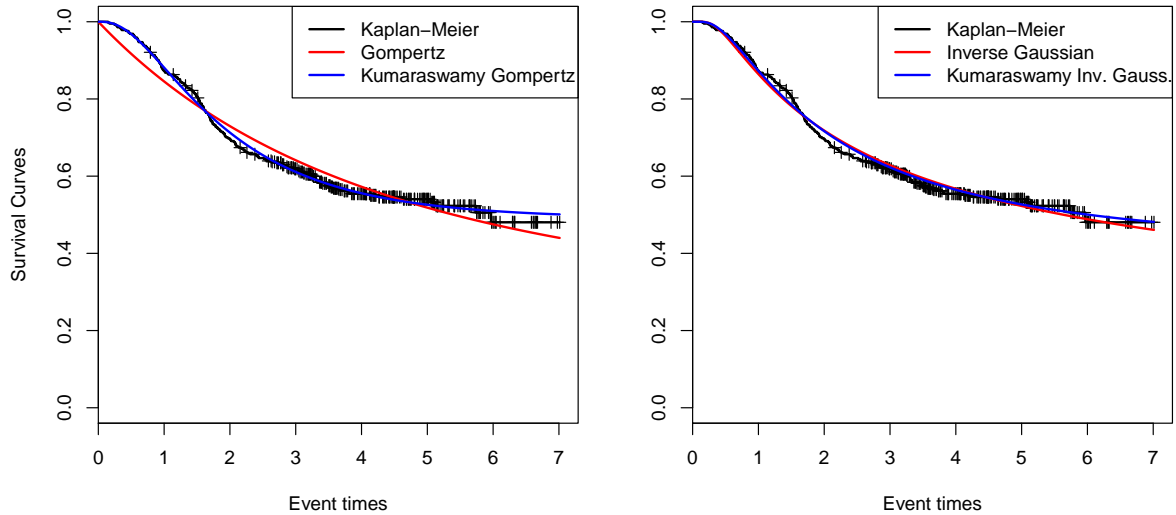


Figure 4.3: Survival curves for the fitted distributions for the melanoma data set.

Table 4.3: MLEs for the fitted distributions for the melanoma data set.

Model	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC	BIC	CAIC
Gompertz	-0.1314 (0.0540)	0.1793 (0.0217)	1 -	1 -	0.2556 (0.3359)	1096.467	1104.534	1096.496
Kumaraswamy Gompertz	-0.5929 (0.1441)	0.6040 (0.9382)	2.8708 (0.7902)	2.2039 (6.6481)	0.4901 (0.1872)	1057.318	1073.45	1057.415
Inverse Gaussian	-0.0357 (0.0319)	0.4740 (0.0427)	1 -	1 -	0.1398 (0.1084)	1062.956	1071.022	1062.985
Kumaraswamy inverse Gaussian	-2.5544 (0.8850)	32.6713 (18.6450)	22.0545 (7.5425)	27.2590 (18.2056)	0.4146 (0.0475)	1061.043	1077.175	1061.14

Here we consider the melanoma data set. The survival curves of the fitted distributions in the melanoma data set are shown in Figure 4.3. The corresponding maximum likelihood estimates are given in Table 6.5.

The estimated value of a is negative for all of the distributions, so all of the fitted distributions are defective. The cure fraction estimates for the Gompertz and inverse Gaussian distributions are 25.56 and 13.98 percents, respectively. These estimates are smaller than the end of the Kaplan-Meier curve, which stabilizes at 48.04 percent. The cure fraction estimates for the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 49.01 and 41.46 percents, respectively.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan-Meier

curve more accurately than the survival curves of the fitted baseline distributions. This is especially the case for the fitted Kumaraswamy Gompertz distribution.

We can also see that the AIC and CAIC values are smaller for the Kumaraswamy distributions than for the baseline distributions. The AIC and CAIC values for the inverse Gaussian distribution are a little larger than those for the Kumaraswamy inverse Gaussian distribution. Both these distributions outperform the Gompertz distribution, but do not perform as well as the Kumaraswamy Gompertz distribution. In terms of BIC, the best fit is given by the inverse Gaussian distribution, followed by the Kumaraswamy Gompertz distribution, then the Kumaraswamy inverse Gaussian distribution and then the Gompertz distribution.

Now we present an application of the regression model described in Section 4.2.5. We are going to consider a covariate that represents the age of the individuals in the data set. For a simple illustrative example, we categorize this variable into two classes. The variable was classified as zero when the age was below the average of all individuals and as one otherwise.

Table 4.4: MLEs for the fitted regression models for the melanoma data set.

Model	\hat{a}	\hat{b}	\hat{r}	\hat{u}	$\hat{\beta}_0$	$\hat{\beta}_1$	\hat{p}_0	\hat{p}_1
Gompertz	-0.8988 (0.3799)	0.0264 (0.0388)	- -	- -	3.7657 (1.4764)	0.1893 (0.1472)	0.2812 (0.3480)	0.2158 (0.3380)
Kumaraswamy Gompertz	-4.3401 (0.8070)	2.6954 (3.4839)	2.7143 (0.5119)	0.6281 (1.3449)	2.0631 (3.4039)	0.1719 (0.1472)	0.5215 (0.2017)	0.4616 (0.2075)
Inverse Gaussian	-0.6542 (1.0402)	3.1659 (0.4712)	- -	- -	0.1386 (0.6018)	0.1789 (0.1472)	0.2881 (0.4972)	0.2258 (0.4832)
Kumaraswamy inverse Gaussian	-0.8113 (1.8714)	2.3698 (4.2916)	0.8062 (1.3570)	3.3050 (7.1690)	-0.9812 (2.5830)	0.1753 (0.1473)	0.3456 (0.4744)	0.2819 (0.4697)

Table 4.4 reports the maximum likelihood estimates of the proposed regression models. We can see that the Gompertz and inverse Gaussian distributions estimate the cure fraction of both groups very closely. The cure fraction for the group with age below the average is around 0.28. That for the group with age above the average is 0.22. The Kumaraswamy inverse Gaussian distribution gives higher estimates, 0.34 and 0.28. The Kumaraswamy Gompertz distribution also gives higher estimates, 0.52 and 0.46.

Figures 4.4 and 4.5 show the fitted survival curves of the baseline and Kumaraswamy

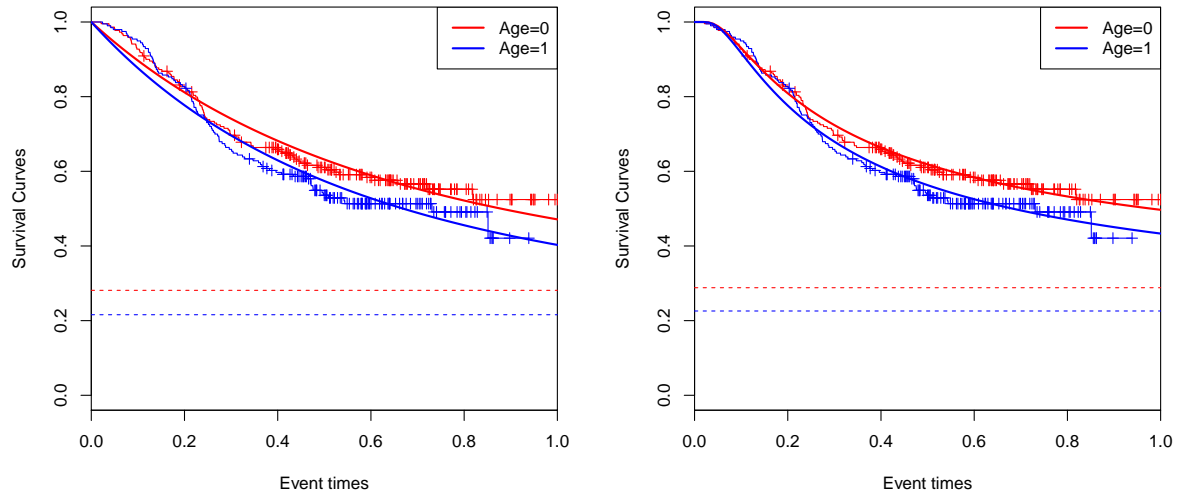


Figure 4.4: In the left, the fitted Gompertz regression model, in the right, the inverse Gaussian model.

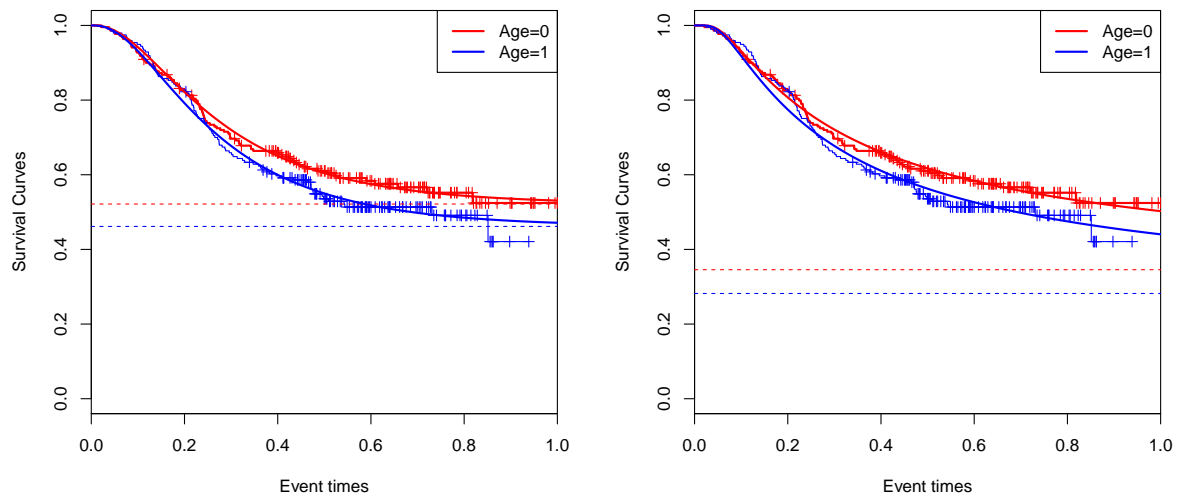


Figure 4.5: In the left, the fitted Kumaraswamy Gompertz regression model, in the right, the Kumaraswamy inverse Gaussian model.

distributions together with Kaplan-Meyer curves, plotted in the same color. The dashed lines represent the estimated cure fractions. As we can see, the baseline distributions do not capture the Kaplan-Meyer curve as well as the Kumaraswamy distributions do. Kumaraswamy distributions fit the end of the curve better, leading to higher cure fraction estimates. The baseline distributions give more conservative estimates.

In all fits, the group consisting of younger individuals has a survival probability higher

than the group consisting of older individuals. That makes sense and it is fair to expect this behaviour in all data sets of the kind.

4.4.2 Colon data

The second data considered here is the colon data set.

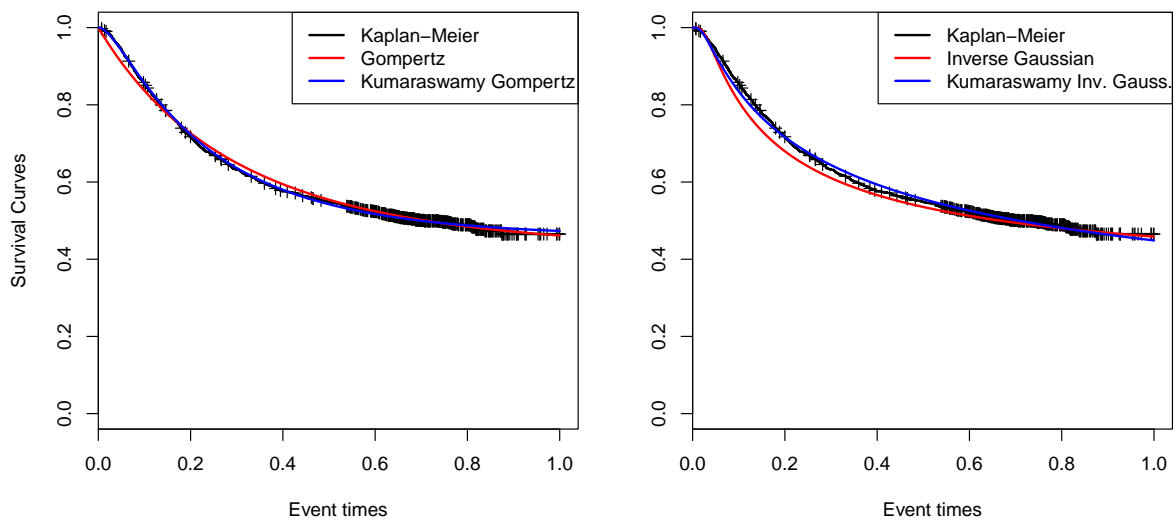


Figure 4.6: Survival curves for the fitted distributions for the colon data set.

Table 4.5: MLEs for the fitted distributions for the colon data set.

Model	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC	BIC	CAIC
Gompertz	-2.3375 (0.1772)	2.0018 (0.1025)	1 -	1 -	0.4247 (0.1339)	1518.022	1529.076	1518.028
Kumaraswamy Gompertz	-3.3598 (0.2727)	17.3085 (6.4465)	2.0508 (0.2476)	0.1757 (0.0803)	0.4586 (0.1215)	1452.173	1474.282	1452.195
Inverse Gaussian	-1.6688 (0.1568)	7.3406 (0.2901)	1 -	1 -	0.3653 (0.0219)	1597.472	1608.527	1597.479
Kumaraswamy inverse Gaussian	-1.3366 (2.861)	91.5128 (29.5785)	5.6371 (1.3563)	0.9717 (0.3392)	0.1602 (0.2075)	1511.714	1533.823	1511.735

The survival curves of the fitted distributions are shown in Figure 4.6. The corresponding maximum likelihood estimates are given in Table 4.5.

Again, the estimated value of a is negative for all of the distributions, and so all of the fitted distributions are defective. The cure fraction estimates for the Gompertz and inverse Gaussian distributions are 42.47 and 36.53 percents, respectively. The cure fraction

estimates for the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 45.86 and 16.02 percents, respectively. The cure fraction estimates for the inverse Gaussian and Kumaraswamy inverse Gaussian distributions are smaller than the end of the Kaplan-Meier curve, which stabilizes at 46.51 percent. The cure fraction estimates for the Gompertz and Kumaraswamy Gompertz distributions are a lot closer to the end of the Kaplan-Meier curve.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan-Meier curve more accurately than the survival curves of the fitted baseline distributions. However, it must be said that all four distributions provide reasonable fits.

The AIC and CAIC values show that the Kumaraswamy Gompertz distribution gives the best fit by far. The second smallest values for AIC and CAIC are for the Kumaraswamy inverse Gaussian distribution. The third smallest values (little larger than the second smallest values) for AIC and CAIC are for the Gompertz distribution. The largest values (lot larger than the third smallest values) for AIC and CAIC are for the inverse Gaussian distribution. The BIC values show that the Kumaraswamy Gompertz distribution gives the best fit, followed by the Gompertz distribution, then the Kumaraswamy inverse Gaussian distribution and then the inverse Gaussian distribution.

4.4.3 Leukemia data

The third data considered here is the leukemia data set.

Table 4.6: MLEs for the fitted distributions for the leukemia data set.

Model	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC	BIC	CAIC
Gompertz	-1.5103 (0.3696)	2.3767 (0.5171)	1 -	1 -	0.2073 (0.2557)	52.5763	56.1447	52.869
Kumaraswamy Gompertz	-2.9825 (0.9399)	12.1742 (11.5263)	7.1974 (6.0582)	0.7541 (1.3356)	0.1961 (0.2448)	36.4139	43.5507	37.4396
Inverse Gaussian	0.2261 (0.3436)	3.1393 (0.7379)	1 -	1 -	0 -	50.9801	54.5485	51.2728
Kumaraswamy inverse Gaussian	-1.6449 (1.2734)	0.8895 (1.6488)	0.7226 (1.2366)	22.3123 (42.1862)	0.2025 (0.0612)	36.5291	43.6659	37.5548

The survival curves of the fitted distributions are shown in Figure 4.7. The corresponding maximum likelihood estimates are given in Table 4.6.

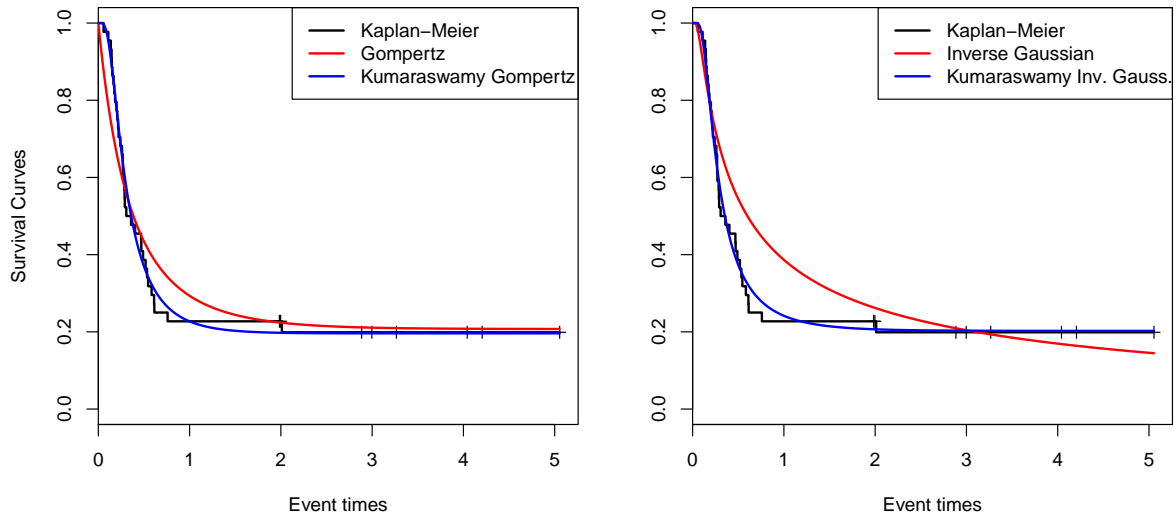


Figure 4.7: Survival curves for the fitted distributions for the leukemia data set.

Not all of the fitted distributions estimate a by a negative value. The inverse Gaussian distribution is not estimated as being defective, leading to an estimated cure rate of 0. The cure fraction estimate for the Gompertz distribution is 20.73 percent. The cure fraction estimates for the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are 19.61 and 20.25 percents, respectively. The cure fraction estimates for the Gompertz, Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions are close to the end of Kaplan-Meier curve, which stabilizes at 19.88 percent.

The survival curves of the fitted Kumaraswamy distributions capture the Kaplan-Meier curve more accurately than the survival curves of the fitted baseline distributions. However, it must be said that the Gompertz distribution provides a reasonable fit as well. The inverse Gaussian distribution does not fit well.

The AIC, BIC and CAIC values show that the Kumaraswamy Gompertz distribution gives the best fit. Gompertz distribution gives the largest values for AIC, BIC, CAIC in spite of its good fit to the Kaplan-Meier curve. This appears a little strange. But closeness of the fitted survival curves to the Kaplan-Meier curve is only a “measure” just like the AIC, BIC and CAIC. It just happens that the inverse Gaussian distribution does not perform so well with respect to this measure but the Gompertz distribution does so.

4.4.4 Discussion

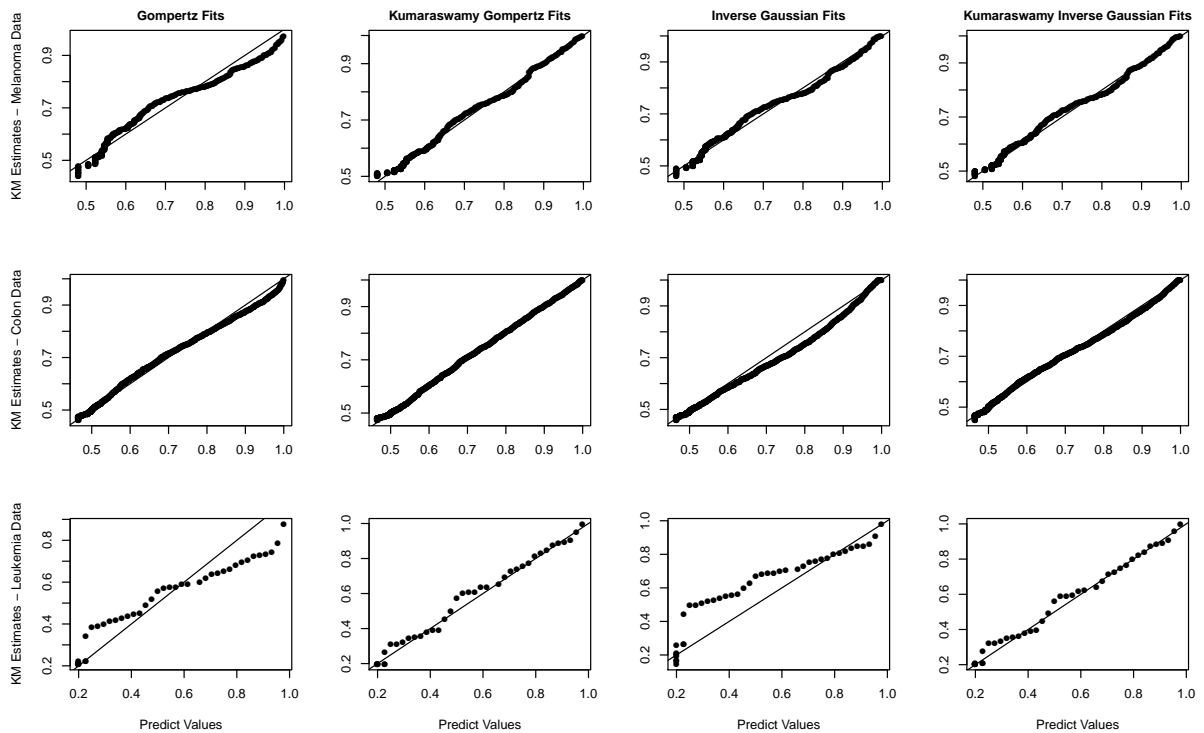


Figure 4.8: Probability plots for the fit of the four distributions to the three data sets.

Figure 4.8 plots the observed probabilities versus the expected probabilities for the fitted Gompertz, inverse Gaussian, Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions for the three data sets. We see that the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions have the points closer to the diagonal line for each of the data sets (especially when the Kumaraswamy Gompertz and Gompertz distributions are compared for the melanoma and leukemia data sets and when the Kumaraswamy inverse Gaussian and inverse Gaussian distributions are compared for the colon and leukemia data sets), showing their better fits.

Table 4.7 reports the statistic and p -value of the log-likelihood ratio test. The test compares the baseline distribution with the corresponding Kumaraswamy distribution. The p -values are very small except for the Kumaraswamy inverse Gaussian distribution with a p -value of 0.052. These are compatible with the visual assessment of the survival curves in Figure 4.8.

Table 4.8 reports the 95 percent asymptotic confidence intervals for the parameter a .

The objective here is to see if the cure rate p is different from zero. If the parameter a is significantly negative, we can say that p is significantly different from zero, This is because a is free to take any value on the real line, while the other parameters are positive by definition. The Gompertz and Kumaraswamy Gompertz distributions have fully negative intervals for all data sets. The inverse Gaussian distribution has a fully negative interval for only the colon data set. The Kumaraswamy inverse Gaussian distribution has a fully negative interval for only the melanoma data set. Therefore, the cure fraction is significantly different from zero in most cases. For every data set, there are at least two distributions indicating that the cure fraction p is significantly different from zero.

Table 4.7: Log-likelihood ratio test for the proposed models and data sets.

Data set	Model	LR test statistic	p -value
Melanoma	Gompertz/Kum-Gompertz	43.149	< 0.0001
	Inverse Gaussian/Kum-inverse Gaussian	5.913	0.052
Colon	Gompertz/Kum-Gompertz	69.849	< 0.0001
	Inverse Gaussian/Kum-inverse Gaussian	89.758	< 0.0001
Leukemia	Gompertz/Kum-Gompertz	20.1624	< 0.0001
	Inverse Gaussian/Kum-inverse Gaussian	18.451	0.0001

Table 4.8: 95 percent asymptotic confidence intervals for the parameter a for the proposed models and data sets.

Model	Melanoma	Colon	Leukemia
Gompertz	(-0.2372; -0.0256)	(-2.6848; -1.9902)	(-2.2347; -0.7859)
Kumaraswamy Gompertz	(-0.8753; -0.3105)	(-3.8943; -2.8253)	(-4.8247; -1.1403)
Inverse Gaussian	(-0.0982; 0.0268)	(-1.9761; -1.3615)	(-0.4474; 0.8996)
Kumaraswamy inverse Gaussian	(-4.2890; -0.8198)	(-6.9442; 4.2710)	(-4.1408; 0.8510)

Table 4.9: AIC values for the fitted distributions and for the standard mixture model. The smaller AIC values are bolded.

Model	Melanoma		Colon		Leukemia	
	Defective	Mixture	Defective	Mixture	Defective	Mixture
Gompertz	1096.467	1085.459	1518.022	1520.023	52.5763	50.741
Kumaraswamy Gompertz	1057.318	1058.741	1452.173	1451.823	36.4139	36.432
Inverse Gaussian	1062.956	1063.200	1597.472	1668.362	50.9801	37.139
Kumaraswamy inverse Gaussian	1061.043	1058.741	1511.714	1496.993	36.5291	37.473

The AIC values for the fitted distributions and those for the standard mixture model are compared in Table 4.9. The defective model was compared with its own version as a standard mixture model. In this way, the mixture model has always one extra parameter than the defective one. As noted before, we can see that the Kumaraswamy distributions have generally smaller AIC values, in the defective form or in the mixture model form.

For the melanoma data set, all but the Gompertz and Kumaraswamy inverse Gaussian distributions have smaller AIC values. For the colon data set, the baseline distributions perform better under the defective strategy while the mixture models have better values in the Kumaraswamy forms. For the leukemia data set, the baseline distributions perform better under the mixture models strategy while the Kumaraswamy distributions are better under the defective strategy.

We have presented three different examples on defective models. The examples show that there are cases where the baseline distributions perform well and some cases where they do not. The examples also show that the Kumaraswamy distributions lead to better fits in most cases. We also see that there are cases, where the defective models outperform the mixture models.

4.5 Conclusions

We have proposed two new defective distributions: the Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions. Each of these distributions has one less parameter than its version under the standard mixture approach. We have assessed the finite sample performance of their maximum likelihood estimators. We have illustrated the use of the new distributions to three real cancer data sets containing cure fractions. We have shown that the new distributions perform better than the baseline distributions (the Gompertz and inverse Gaussian distributions) in terms of the AIC, BIC, CAIC and proximity to the Kaplan-Meier curves. In some cases, they outperform the standard mixture model in terms of AIC. We have also illustrated the use of regression models based on the new distributions to one of the three data sets.

Chapter 5

Generalized Extended Class of Defective Models

5.1 Introduction

In this chapter, we generalize the results obtained in Chapter 3 and 4. We propose a theorem that guarantee that an extended distribution under a family is also defective, if the baseline distribution is also defective. In literature, obtain new classes to generate new distributions in the form $G^*(t) = f[G(t)]$ have been the focus for many researcher, with a high increase in the number of these families lately.

The first approach proposed in recent years was that due to [Marshall & Olkin \(1997\)](#). Since that, many other approaches have been proposed. For example: exponentiated G distributions due to [Gupta *et al.* \(1998\)](#), beta G distributions due to [Eugene *et al.* \(2002\)](#), gamma G distributions due to [Zografos & Balakrishnan \(2009\)](#), Kumaraswamy G distributions due to [Cordeiro & de Castro \(2011\)](#), generalized beta G distributions due to [Alexander *et al.* \(2012\)](#), beta extended G distributions due to [Cordeiro *et al.* \(2012c\)](#), gamma G distributions due to [Ristić & Balakrishnan \(2012\)](#), gamma uniform G distributions due to [Torabi & Montazeri \(2012\)](#), beta exponential G distributions due to [Alzaatreh *et al.* \(2013b\)](#), Weibull G distributions also due to [Alzaatreh *et al.* \(2013b\)](#),

log gamma G I distributions due to [Amini *et al.* \(2014\)](#), log gamma G II distributions also due to [Amini *et al.* \(2014\)](#), exponentiated generalized G distributions due to [Cordeiro *et al.* \(2013d\)](#), exponentiated Kumaraswamy G distributions due to [Lemonte *et al.* \(2013\)](#), geometric exponential Poisson G distributions due to [Nadarajah *et al.* \(2013a\)](#), truncated-exponential skew-symmetric G distributions due to [Nadarajah *et al.* \(2013b\)](#), modified beta G distributions due to [Nadarajah *et al.* \(2013c\)](#), and exponentiated exponential Poisson G distributions due to [Ristić & Nadarajah \(2013\)](#).

We will take the Gamma, Gamma uniform, exponentiated, truncated-exponential skew-symmetric, Beta, exponentiated exponential, exponentiated generalized and Weibull families to exemplify the proposed theorem.

The main purposes of this chapter are: i) introduce the results that can generate new defective distributions, given an extended family of distributions; ii) explore a variety of these families to show that they can properly be used to fit different kinds of data sets. iii) provide a full literature review regarding to the considered families.

The next section contains the methodology details. In Section 5.3 we provide simulations for four different scenarios. In Section 5.4 we apply the proposed models in the leukemia and melanoma data set, considering the Gompertz and inverse Gaussian as the defective baseline distribution.

5.2 Distribution Families

In Chapter 3, we proposed two new defective distributions using the Marshall-Olkin family. Similarly, in Chapter 4, we proposed two new defective distributions using the Kumaraswamy family. However, it is easy to see that the only specific characteristic needed in a family is to be continuous in relation to the baseline distribution. Therefore, we enunciate:

Theorem 5.1. *If $S(t)$ is a survival function of a defective model and the extension function $S^*(t)$ is continuous in relation to $S(t)$, then $S^*(t)$ is also a defective model.*

Proof: Let $S^*(t) = g(\mathbf{v}, S(t))$, where g is the extension function and \mathbf{v} is the extra vector of parameter from the extension. Consider also that $\lim_{t \rightarrow \infty} S(t) = p_0$. If g is a continuous function then $\lim_{t \rightarrow \infty} S^*(t) = \lim_{t \rightarrow \infty} g(\mathbf{v}, S(t)) = g(\mathbf{v}, \lim_{t \rightarrow \infty} S(t)) = g(\mathbf{v}, p_0)$. We know that a function that extends a survival model must keep its basic properties. So, for any value of $S(t)$ in the interval $(0, 1)$, $S^*(t)$ must also return a value in this interval. Therefore, $g(\mathbf{v}, p_0) = p \in (0, 1)$, showing that g comes from a defective distribution. The proof is complete. \square

This theorem now extend the basics defective models to a full new variety of distributions. For each family, it is possible to generate two new distributions when considering the directly application of the baseline into the extended family. We choose eight families to illustrate the theorem, as following.

5.2.1 Gamma G

The Gamma G family was introduced by [Zografos & Balakrishnan \(2009\)](#). The density and cumulative functions of the extended distribution are

$$g^*(x) = \Gamma(a)^{-1} g(x) \{-\log [1 - G(x)]\}^{a-1},$$

$$G^*(x) = Q(a, -\log [1 - G(x)]),$$

for x in the range of g and $a > 0$, the shape parameter, where $Q(a, x) = \int_0^x t^{a-1} \exp(-t) dt / \Gamma(a)$ denotes the regularized incomplete gamma function, $\Gamma(a) = \int_0^\infty t^{a-1} \exp(-t) dt$ denotes the gamma function.

These distributions were constructed as the distribution of the a th upper record value for a random sample from the cumulative distribution function G .

Particular gamma G I distributions studied in the literature include the gamma Dagum distribution ([Oluyede et al., 2014](#)), the gamma exponentiated Weibull distribution ([Castellares & Lemonte, 2014](#)), the gamma extended Frechet distribution ([da Silva et al., 2013](#)), the gamma half normal distribution ([Alzaatreh & Knight, 2013](#)), the gamma inverse

Weibull distribution (Pararai *et al.*, 2014), the gamma linear failure rate distribution (Cordeiro *et al.*, 2014a), the gamma log-logistic distribution (Ramos *et al.*, 2013), the gamma logistic distribution (Castellares *et al.*, 2015), the gamma Lomax distribution (Cordeiro *et al.*, 2015) and the gamma normal distribution (Alzaatreh *et al.*, 2014).

Gamma G distributions have been used to model: breaking stress of carbon fibers (Alzaatreh *et al.*, 2014; Cordeiro *et al.*, 2014a); flood levels for the Susquehanna river at Harrisburg, PA (Alzaatreh & Knight, 2013); gene expression levels on human cancer cells (Castellares *et al.*, 2015); number of million of revolutions before failure of ball bearings in a life testing experiment (Pararai *et al.*, 2014); number of successive failures for the air conditioning system of each member in a fleet of Boeing 720 jet airplanes (Oluyede *et al.*, 2014); remission times of a random sample of bladder cancer patients (Castellares & Lemonte, 2014; Cordeiro *et al.*, 2015; Oluyede *et al.*, 2014); salaries of professional baseball players (Oluyede *et al.*, 2014); strengths of glass fibers (Alzaatreh *et al.*, 2014); survival times of breast cancer patients (Ramos *et al.*, 2013); survival times of cutaneous melanoma (a type of malignant cancer) patients (Cordeiro *et al.*, 2014a); survival times of guinea pigs injected with different doses of tubercle bacilli (Pararai *et al.*, 2014); tensile strength for single-carbon fibers (Alzaatreh & Knight, 2013); the cDNA microarray data of the NC160 cancer cell lines (Castellares *et al.*, 2015); waiting times between consecutive eruptions of the Kiama Blowhole (da Silva *et al.*, 2013).

5.2.2 Gamma Uniform G

The Gamma uniform G distributions was recently introduced by Torabi & Montazeri (2012). The density and cumulative functions of the extended distribution are

$$g^*(x) = \frac{1}{\Gamma(a)} \frac{g(x)}{[1 - G(x)]^2} \left[\frac{G(x)}{1 - G(x)} \right]^{a-1} \exp \left[-\frac{G(x)}{1 - G(x)} \right],$$

$$G^*(x) = Q \left(a, \frac{G(x)}{1 - G(x)} \right),$$

for x in the range of g and $a > 0$, the shape parameter.

These distributions were constructed by considering the distribution of $G^{-1}(W/(1+W))$, where W is a gamma random variable. These distributions have been used to model survival times of leukemia patients [Torabi & Montazeri \(2012\)](#).

5.2.3 Exponentiated G

The Exponentiated G family was introduced by [Gupta *et al.* \(1998\)](#). The density and cumulative functions of the extended distribution are

$$g^*(x) = ag(x)G^{a-1}(x),$$

$$G^*(x) = G^a(x),$$

for x in the range of g and $a > 0$, the shape parameter.

These distributions were motivated to model the failure of time of a system having a units functioning in parallel the failure times of which are assumed to be independent and identical with cumulative distribution function G .

Particular exponentiated G distributions studied in the literature include the exponentiated Frechet distribution ([Nadarajah & Kotz, 2003](#)), the exponentiated gamma distribution ([Nadarajah & Gupta, 2007](#)), the exponentiated generalized inverse Weibull distribution ([Elbatal & Muhammed, 2014](#)), the exponentiated Gumbel distribution ([Nadarajah, 2006](#)), the exponentiated Lomax distribution ([Abdul-Moniem & Abdel-Hameed, 2012](#); [Salem, 2014](#)), the exponentiated Pareto distribution ([Shawky & Abu-Zinadah, 2009](#)) and the exponentiated transmuted Weibull distribution ([Hady & Ebraheim, 2014](#)).

Exponentiated G distributions have been used to model: annual maximum daily rainfall from Orlando, Florida ([Nadarajah, 2006](#)); drought data from Nebraska ([Nadarajah & Gupta, 2007](#)); remission times of a random sample of bladder cancer patients ([Elbatal & Muhammed, 2014](#)).

5.2.4 Truncated-Exponential Skew-Symmetric G

The Truncated-exponential skew-symmetric G (TESS- G) family was introduced by [Nadarajah *et al.* \(2013b\)](#). The density and cumulative functions of the extended distribution are

$$g^*(x) = \frac{\lambda}{1 - \exp(-\lambda)} g(x) \exp\{-\lambda G(x)\},$$

$$G^*(x) = \frac{1 - \exp\{-\lambda G(x)\}}{1 - \exp(-\lambda)},$$

for x in the range of g , and $-\infty < \lambda < \infty$, the skewness parameter.

These distributions were constructed as modifications of the skew-symmetric distributions proposed in [Azzalini \(1985\)](#). They have been used to model annual maximum daily rainfall data for 14 locations in west central Florida: Clermont, Brooksville, Orlando, Bartow, Avon Park, Arcadia, Kissimmee, Inverness, Plant City, Tarpon Springs, Tampa International Airport, St Leo, Gainesville and Ocala ([Nadarajah *et al.*, 2013b](#)).

5.2.5 Beta G

The Beta G family was introduced by [Eugene *et al.* \(2002\)](#). The density and cumulative functions of the extended distribution are

$$g^*(x) = \frac{1}{B(a, b)} g(x) [G(x)]^{a-1} [1 - G(x)]^{b-1},$$

$$G^*(x) = I_{G(x)}(a, b),$$

for x in the range of g , $a > 0$, the first shape parameter, and $b > 0$, the second shape parameter. $I_x(a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt / B(a, b)$ denotes the incomplete beta function ratio, $B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$ denotes the beta function.

These distributions were motivated to model the failure time of a a -out-of- $a+b-1$ system when the failure times of the components are independent and identical random variables with cumulative distribution function G .

Particular beta G distributions studied in the literature include the beta Birnbaum-

Saunders distribution (Cordeiro & Lemonte, 2011a), the beta Burr III distribution (Gomes *et al.*, 2013), the beta Burr XII distribution (Parainaba *et al.*, 2011), the beta Cauchy distribution (Alshawarbeh *et al.*, 2014), the beta Dagum distribution (Domma & Condino, 2013), the beta exponential distribution (Nadarajah & Kotz, 2006), the beta exponential geometric distribution (Bidram, 2012; Nassar & Nada, 2012), the beta exponentiated Pareto distribution (Zea *et al.*, 2012), the beta exponentiated Weibull distribution (Cordeiro *et al.*, 2013c), the beta Frechet distribution (Barreto-Souza *et al.*, 2011), the beta gamma distribution (Kong *et al.*, 2007), the beta generalized exponential distribution (Barreto-Souza *et al.*, 2010), the beta generalized gamma distribution (Cordeiro *et al.*, 2013a), the beta generalized half normal geometric distribution (Ramires *et al.*, 2013), the beta generalized Lindley distribution (Oluyede & Yang, 2014), the beta generalized logistic distribution (Morais *et al.*, 2013), the beta generalized normal distribution (Cintra *et al.*, 2014), the beta generalized Pareto distribution (Mahmoudi, 2011; Nassar & Nada, 2011), the beta generalized Rayleigh distribution (Cordeiro *et al.*, 2013b), the beta generalized Weibull distribution (Singla *et al.*, 2012), the beta Gompertz distribution (Jafari *et al.*, 2014), the beta Gumbel distribution (Nadarajah & Kotz, 2004), the beta half-Cauchy distribution (Cordeiro & Lemonte, 2011b), the beta inverse Rayleigh distribution (Leao *et al.*, 2013), the beta inverse Weibull distribution (Hanook *et al.*, 2013), the beta linear failure rate distribution (Jafari & Mahmoudi, 2014), the beta Laplace distribution (Cordeiro & Lemonte, 2011c; Kozubowski & Nadarajah, 2008), the beta Lindley distribution (Merovci & Sharma, 2014), the beta lognormal distribution (Montenegro & Cordeiro, 2013), the beta Lomax distribution (Rajab *et al.*, 2013), the beta modified Weibull distribution (Silva *et al.*, 2010), the beta Moyal distribution (Cordeiro *et al.*, 2012b), the beta Nakagami distribution (Shittu & Adepoju, 2013), the beta normal distribution (Eugene *et al.*, 2002), the beta Pareto distribution (Akinsete *et al.*, 2008), the beta power distribution (Cordeiro & Brito, 2012), the beta power exponential distribution (Adepoju *et al.*, 2014), the beta skew normal distribution (Mameli & Musio, 2013), the beta transmuted Weibull distribution (Pal & Tiensuwan, 2014), the beta truncated Pareto distribution (Lourenzutti *et al.*, 2014), the beta Weibull geometric distribution (Bidram *et al.*, 2013; Cordeiro *et al.*, 2013e), the beta Weibull Poisson distribution (Percontini *et al.*, 2013) and

the beta weighted Weibull distribution (Badmus & Bamiduro, 2014; Idowu & Ikegwu, 2013).

Beta G distributions have been used to model: adult numbers for *Tribolium Castaneum* and *Tribolium Confusum* (Eugene *et al.*, 2002; Kong *et al.*, 2007); breaking strength of glass fibers (Adepoju *et al.*, 2014; Alshawarbeh *et al.*, 2014; Barreto-Souza *et al.*, 2010, 2011; Cordeiro & Lemonte, 2011a; Cordeiro *et al.*, 2013a; Domma & Condino, 2013); breaking stress of carbon fibers (Alshawarbeh *et al.*, 2014; Barreto-Souza *et al.*, 2011; Cordeiro & Lemonte, 2011a; Leao *et al.*, 2013; Oluyede & Yang, 2014); carbon monoxide measurements in several brands of cigarettes (Cordeiro *et al.*, 2013c); daily ozone level measurements in New York (Cordeiro *et al.*, 2013e); exceedances of flood peaks of the Wheaton river in Yukon Territory, Canada (Akinsete *et al.*, 2008; Alshawarbeh *et al.*, 2014; Cordeiro *et al.*, 2012b; Mahmoudi, 2011); failure times of a polyester/viscose yarn in a textile experiment (Pal & Tiensuwan, 2014); failure times of motorettes with a new insulation (Cordeiro *et al.*, 2013c; Pal & Tiensuwan, 2014); failure times of turbocharger of one type of engine (Singla *et al.*, 2012); fatigue life of 6061-T6 aluminum coupons cut parallel with the direction of rolling (Bidram, 2012; Bidram *et al.*, 2013; Mahmoudi, 2011); fatigue life of bearings of a certain type (Montenegro & Cordeiro, 2013); flood data for the Floyd river located in James, Iowa, USA (Akinsete *et al.*, 2008); household income and consumption in Italy (Domma & Condino, 2013); lifetimes of mechanical components (Badmus & Bamiduro, 2014; Jafari *et al.*, 2014; Silva *et al.*, 2010); maximum values of monthly flood rates of the Castelo river, Brazil (Lourenzutti *et al.*, 2014); monthly actual taxes revenue in Egypt (Nassar & Nada, 2011); national index of consumer prices of Brazil corresponding to health and personal care (Cordeiro & Lemonte, 2011c); number of successive failures of the air-conditioning system of each number of a fleet of Boeing 720 jet airplanes (Bidram *et al.*, 2013; Nassar & Nada, 2012); remission times of a random sample of bladder cancer patients (Merovci & Sharma, 2014; Oluyede & Yang, 2014; Zea *et al.*, 2012); repair times for an airborne communication transceiver Cordeiro *et al.* (2012b, 2013b); Percontini *et al.* (2013); SAR image processing (Cintra *et al.*, 2014); short-term and long-term outcomes of constraint induced movement therapy after stroke (Nassar & Nada, 2012); strength of ball bearings (Nassar & Nada, 2012); stress-rupture life of

kevlar epoxy strands subjected to constant sustained pressure (Cordeiro *et al.*, 2013b); survival times of cutaneous melanoma (a type of malignant cancer) patients (Parainaba *et al.*, 2011); survival times of guinea pigs injected with different doses of tubercle bacilli (Cordeiro & Lemonte, 2011b; Merovci & Sharma, 2014); survival times of myelogenous leukemia patients (Mahmoudi, 2011); times to first failure of devices (Jafari & Mahmoudi, 2014).

5.2.6 Exponentiated Exponential Poisson G

The Exponentiated exponential Poisson G (EEP- G) family was introduced by Ristić & Nadarajah (2013). The density and cumulative functions of the extended distribution are

$$g^*(x) = a\lambda \{1 - \exp(-\lambda)\}^{-1} g(x)G^{a-1}(x) \exp[-\lambda G^a(x)],$$

$$G^*(x) = \{1 - \exp(-\lambda)\}^{-1} \{1 - \exp[-\lambda G^a(x)]\},$$

for x in the range of g , $\lambda > 0$, the scale parameter, and $a > 0$, the shape parameter.

These distributions were motivated to model the time to failure of the first out of a Poisson number of systems functioning independently where each system has a fixed number of parallel units and their failure times are independent and identical random variables with cumulative distribution function G . These distributions have been used to model the daily average air temperature (F) in Cairo (Ristić & Nadarajah, 2013).

5.2.7 Exponentiated Generalized G

The Exponentiated generalized G (EG- G) family was introduced by Cordeiro *et al.* (2013d). The density and cumulative functions of the extended distribution are

$$g^*(x) = abg(x) [1 - G(x)]^{a-1} \{1 - [1 - G(x)]^a\}^{b-1},$$

$$G^*(x) = \{1 - [1 - G(x)]^a\}^b,$$

for x in the range of g , $a > 0$, the first shape parameter, and $b > 0$, the second shape parameter.

These distributions were motivated to model the failure of time of a system having b units functioning in parallel and each of these units have a subunits functioning in series. The failure times of the subunits are assumed to be independent and identical with cumulative distribution function G .

Particular exponentiated generalized G distributions studied in the literature include the exponentiated generalized Birnbaum-Saunders distribution (Cordeiro & Lemonte, 2014).

Exponentiated generalized G distributions have been used to model: breaking stress of carbon fibers (Cordeiro *et al.*, 2013d); effects of mechanical damage on banana fruits (Cordeiro *et al.*, 2013d); exceedances of flood peaks of the Wheaton river near Carcross in Yukon Territory, Canada (Cordeiro & Lemonte, 2014; Cordeiro *et al.*, 2013d); lifetimes for industrial devices put on life test at time zero (Cordeiro & Lemonte, 2014); stress-rupture life of kevlar epoxy strands subjected to constant sustained pressure (Cordeiro *et al.*, 2013d).

5.2.8 Weibull-G

The Weibull G family was introduced by Alzaatreh *et al.* (2013b). The density and cumulative functions of the extended distribution are

$$g^*(x) = \frac{c}{\beta^c} \frac{g(x)}{1 - G(x)} \left\{ -\frac{\log [1 - G(x)]}{\beta} \right\}^{c-1} \exp \left\{ - \left[-\frac{\log [1 - G(x)]}{\beta} \right]^c \right\},$$

$$G^*(x) = 1 - \exp \left\{ \left[-\frac{\log [1 - G(x)]}{\beta} \right]^c \right\},$$

for x in the range of g , $\beta > 0$, the scale parameter, and $c > 0$, the shape parameter.

Particular Weibull G distributions studied in the literature include the Weibull exponentiated exponential distribution (Salem & Selim, 2014) and the Weibull Pareto distribution (Alzaatreh *et al.*, 2013a).

Weibull G distributions have been used to model: adult numbers for *Tribolium Confusum* and *Tribolium Castaneum* cultured at 24C and *Tribolium Confusum* strain (Alzaatreh *et al.*, 2013a); breaking stress of carbon fibers (Salem & Selim, 2014).

5.3 Simulation Studies

Here, we assess the performance of the maximum likelihood estimates with respect to sample size to show, among other things, that the usual asymptotes of maximum likelihood estimators still hold for defective distributions. The assessment is based on simulations. The description of data generation is in Section 1.3.

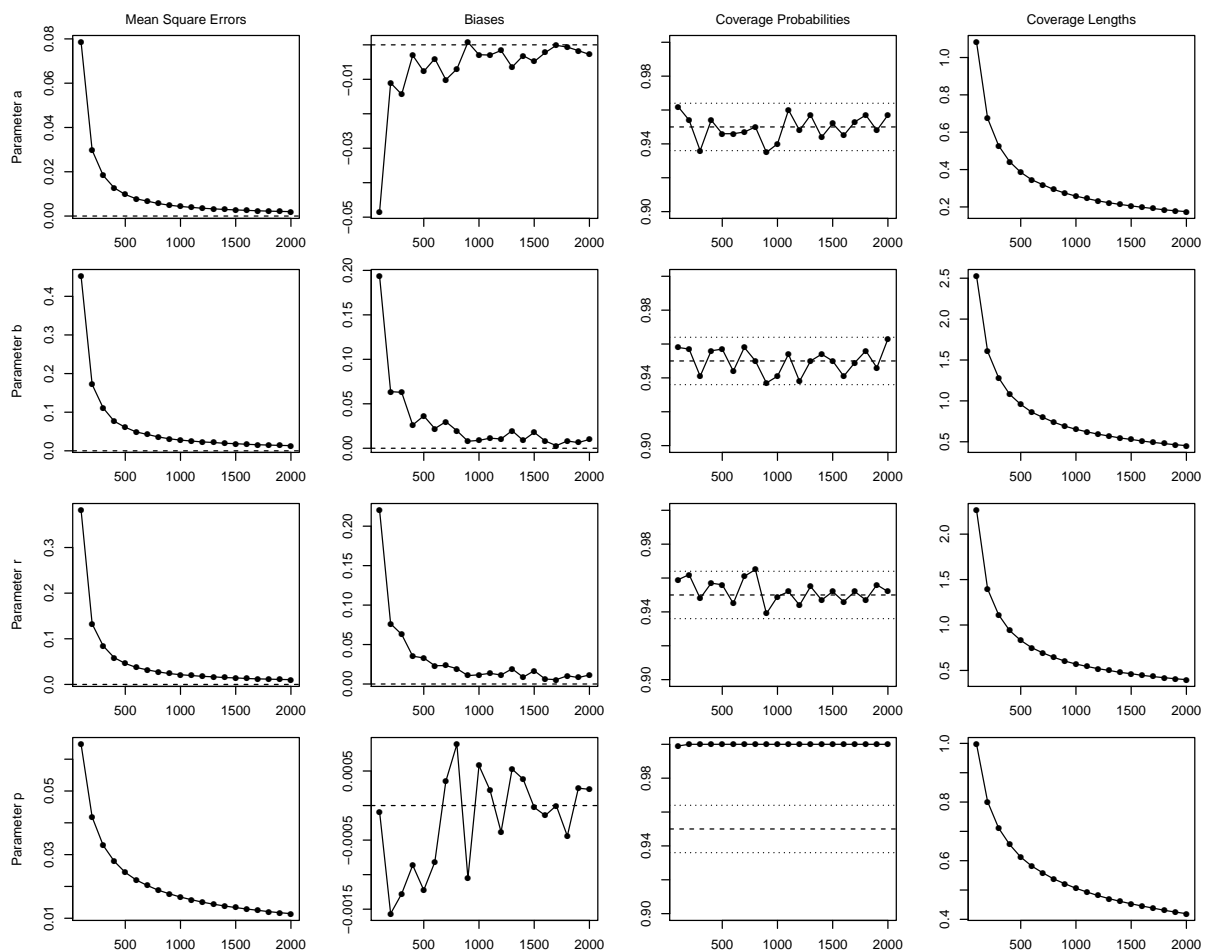


Figure 5.1: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r and p versus n for the Exponentiated Gompertz distribution with $(a, b, r, p) = (-1, 2, 2, 0.2523)$.

We took the sample size to vary from 100 to 2000 in steps of 100. Each sample was

replicated 1000 times. We chose only four of our proposed distributions: the Exponentiated Gompertz distribution with $(a, b, r, p) = (-1, 2, 2, 0.2523)$, the simulation results for which are shown in Figure 5.1; the Weibull Gompertz distribution with $(a, b, r, u, p) = (-1, 2, 2, 2, 0.3678)$, the simulation results for which are shown in Figure 5.3; the TESS inverse Gaussian distribution with $(a, b, r, p) = (-2, 2, 2, 0.7257)$, the simulation results for which are shown in Figure 5.2; the EEP inverse Gaussian distribution with $(a, b, r, u, p) = (-1, 2, 1, 2, 0.3976)$, the simulation results for which are shown in Figure 5.4. The parameter was chosen to represent a very simple set of parameters which have small and large cured rates.

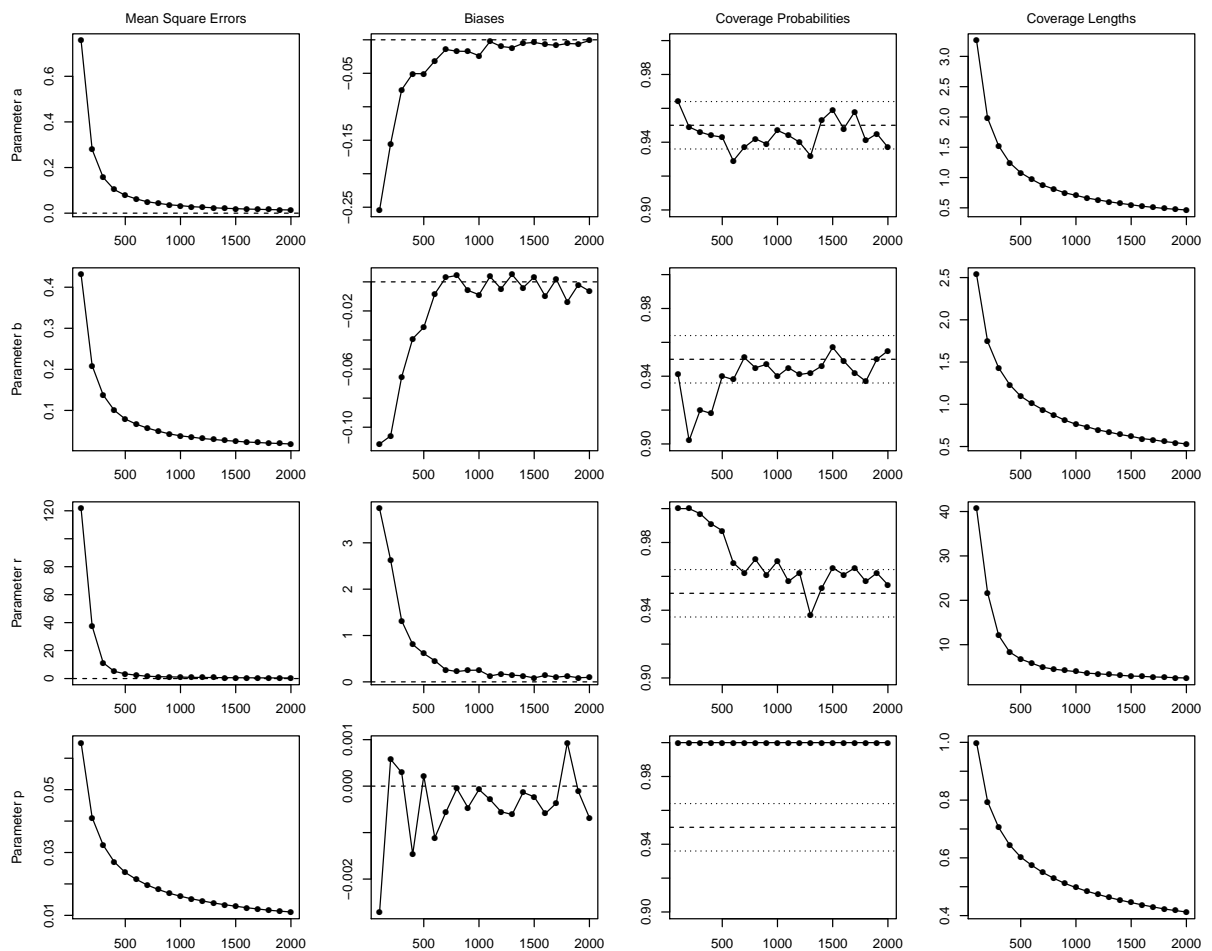


Figure 5.2: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r and p versus n for the TESS inverse Gaussian distribution with $(a, b, r, p) = (-2, 2, 2, 0.7257)$.

From Figures 5.1 and 5.2, we can notice: i) the mean square error decreases smoothly, reaching reasonable small values for $n > 500$; ii) for the parameters a , b and r , the biases

are very small for almost every n , for p , the bias is virtually non-existent; iii) the coverage probabilities stay in the proper confidence region for every n , for a , b and r , however, the parameter p doesn't stay, being estimated in 1 for every n . iv) the coverage lengths decrease smoothly with the increase of the sample sizes, but present large values for the parameter p .

It seems that the delta method over-estimates the standard deviation of the cure fraction in these models. Besides that, every thing looks adequate, with reasonable results in small sample sizes. Nevertheless, the models give a very precise point estimate for p , but the interval estimation is not so good.

From Figures 5.3 and 5.4, we can notice: i) the mean square error decreases smoothly for the Weibull Gompertz, but not so much for the EEP inverse Gaussian. In both cases, the values seem to be too large for the parameters b and u ; ii) biases are small, but doesn't show a behavior around the mean for all parameters. Besides that, the bias of the cure rate is very small; iii) the coverage probabilities stay in the proper confidence region only for a and r , b , u and p , however, stays out of the confidence region. iv) the coverage lengths are higher for the b and u parameters and do not show a decreasing behavior, as expected.

When considering the models with two extra parameters (other than the baseline ones), it gets more complex in terms of estimation. Looks like 2000 samples are not enough to reach the maximum likelihood estimator properties. Another conclusion here is that the delta method seems to over-estimate the deviation of the cure fraction. However, the point estimate is quite good.

5.4 Applications

Here we apply the models proposed in the methodology section in the leukemia and melanoma data sets. One represents a small sample data and the other a large sample. They both have Kaplan-Meier curves completely distinct, so they represent quite distinguished scenarios. We present the point estimation by maximum likelihood and their

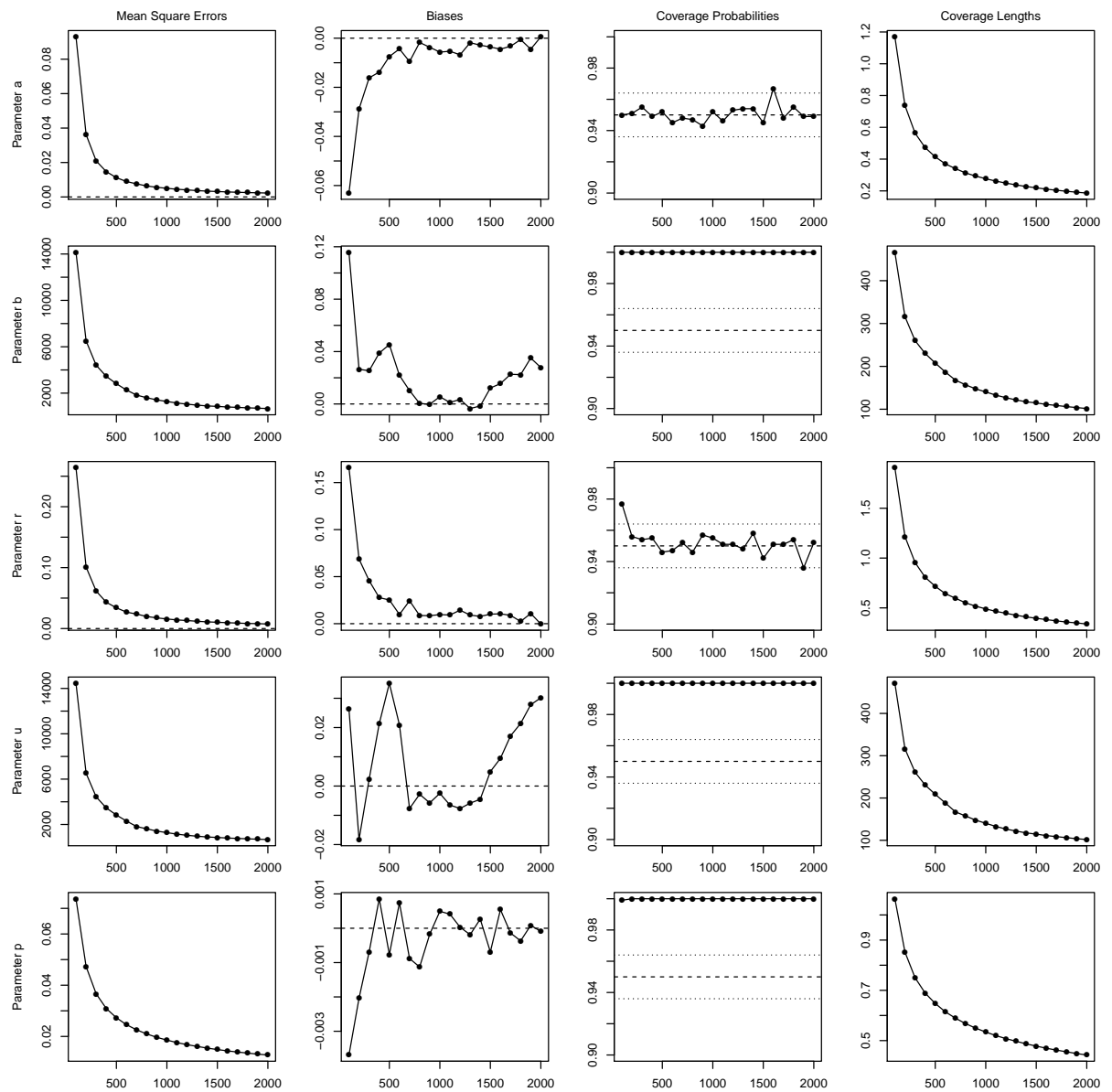


Figure 5.3: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r , u and p versus n for the Weibull Gompertz distribution with $(a, b, r, u, p) = (-1, 2, 2, 2, 0.3678)$.

respective AIC.

We will also consider the Marshall-Olkin Gompertz, Marshall-Olkin inverse Gaussian, Kumaraswamy Gompertz and Kumaraswamy inverse Gaussian distributions, as stated in Chapter 3 and 4, respectively. In total, we are analyzing 20 different defective models.

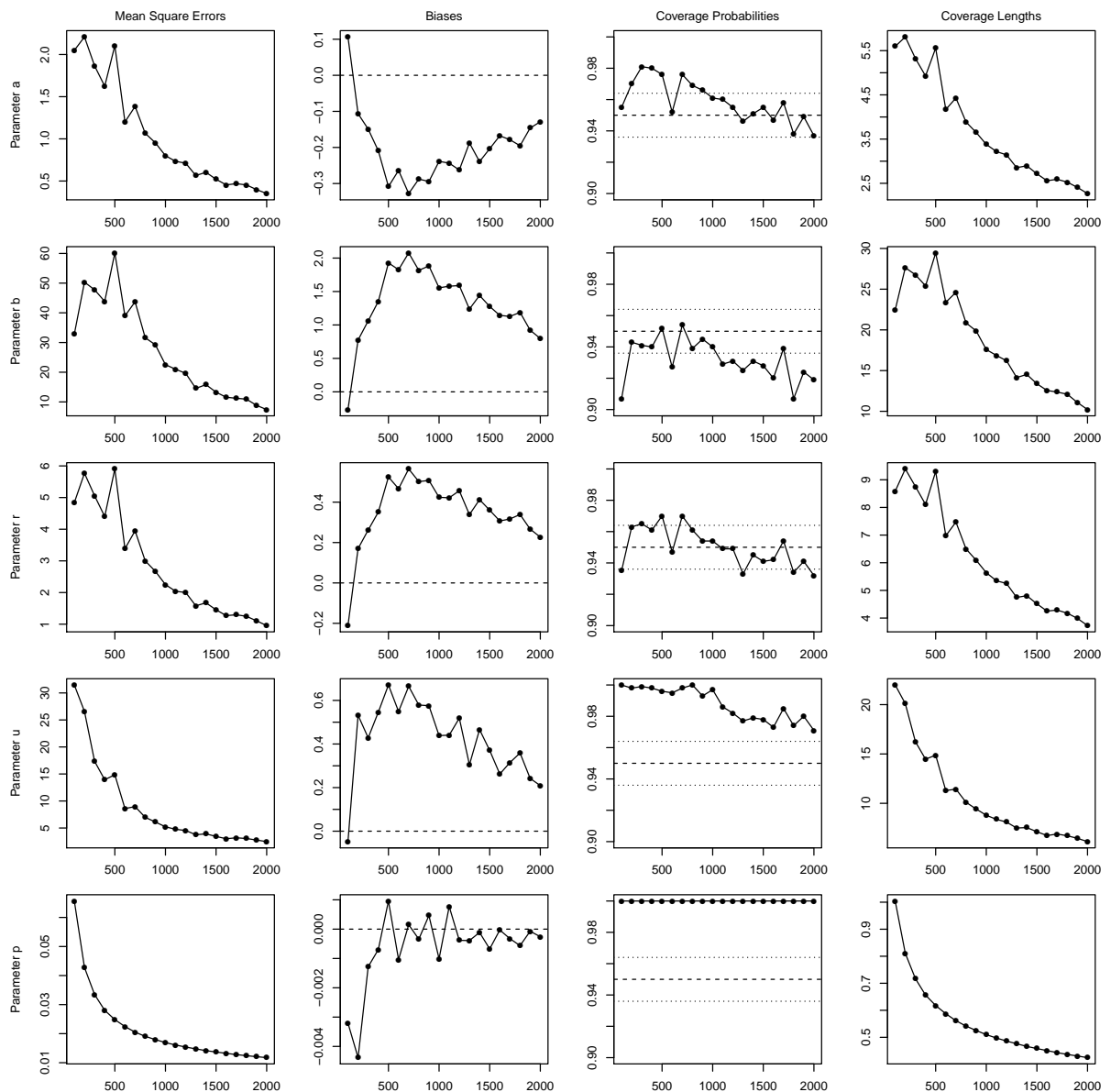


Figure 5.4: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of a , b , r , u and p versus n for the EEP inverse Gaussian distribution with $(a, b, r, u, p) = (-1, 2, 1, 2, 0.3976)$.

5.4.1 Leukemia data

The fitted results for the distributions with Gompertz as baseline are presented in Table 5.1. Figure 5.5 illustrate the respective survival curves. Every single models were lead to be defective. The parameter a is estimated negative for all distributions. The cure fraction was estimated 0.2073 by the Gompertz model, which is known to be a bad fit. The other models estimated the cure fraction between 0.1905 and 0.1953. We can see also that all models captures the Kaplan-Meier curve very well and very similarly, with just some little

Table 5.1: Maximum likelihood estimates in the leukemia data set of the proposed models when the baseline distribution is Gompertz.

Distribution	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC
Gompertz	-7.6348	12.0135	-	-	0.2073	-60.85
Gamma Gompertz	-15.2425	94.3266	4.5005	-	0.1930	-78.50
Gamma uniform Gompertz	-20.4555	33.7584	2.9104	-	0.1946	-77.47
Exponentiated Gompertz	-15.6102	53.2379	6.4700	-	0.1953	-78.99
TESS Gompertz	-16.8292	64.5223	-9.9434	-	0.1935	-78.93
Marshall-Olkin Gompertz	-18.2322	96.1874	46.7848	-	0.1939	-75.58
Beta Gompertz	-16.7286	13.6681	5.4789	7.0361	0.1945	-76.79
EEP Gompertz	-15.0173	33.428	5.1822	2.4156	0.1905	-77.19
EG Gompertz	-15.6302	7.3042	7.3042	6.4897	0.1953	-76.99
Weibull Gompertz	-19.6219	7.6312	3.1278	0.3322	0.1944	-75.65
Kumaraswamy Gompertz	-16.9566	28.1259	4.5168	3.3681	0.1943	-76.65

differences. We can conclude that all extended models can give an reasonable fit for this data and with this baseline distributions. The Exponentiated Gompertz, however, have the smallest AIC between them. It is quite clear that all of the extended distributions outperforms the Gompertz baseline distribution.

Table 5.2: Maximum likelihood estimates in the leukemia data set of the proposed models when the baseline distribution is inverse Gaussian.

Distribution	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC
Inverse Gaussian	1.1426	15.8676	-	-	-	-62.44
Gamma inv. Gaussian	-1.4757	1.4910	0.1332	-	0.1875	-68.18
Gamma uniform inv. Gaussian	-1.3740	0.0693	0.0069	-	0.2377	-71.46
Exponentiated inv. Gaussian	-1.3219	0.0720	0.0072	-	0.2331	-71.38
TESS inv. Gaussian	-9.6568	6.3184	34.0854	-	0.2012	-78.95
Marshall-Olkin inv. Gaussian	-6.7423	5.3253	0.0230	-	0.2104	-78.08
Beta inv. Gaussian	-9.2717	5.7933	0.8741	34.0850	0.2013	-76.98
EEP inv. Gaussian	-7.2200	3.4363	0.5596	16.7902	0.2021	-76.84
EG inv. Gaussian	-9.3250	6.0378	32.7396	0.9103	0.1999	-76.96
Weibull inv. Gaussian	-15.2314	15.7859	2.1717	0.1258	0.1989	-77.16
Kumaraswamy inv. Gaussian	-7.0169	3.1174	0.5037	14.6852	0.2008	-76.83

The fitted results for the distributions with inverse Gaussian as baseline are presented in Table 5.2. Figure 5.6 illustrate the respective survival curves. Every single models were lead to be defective, but the baseline one. The parameter a is estimated negative for all distributions but in the inverse Gaussian in positive, leading to a proper model. Here we have a wider range for the cure fraction, the models estimated the cure fraction between 0.1875 and 0.2331. The Gamma uniform inverse Gaussian and Exponentiated inverse Gaussian estimates the cure around 0.23, but their AIC shows that their are far from the best ones and their survival curves failed to complete capture the Kaplan-Meier curve. We can say that the best model here is the TESS inverse Gaussian, because it have the

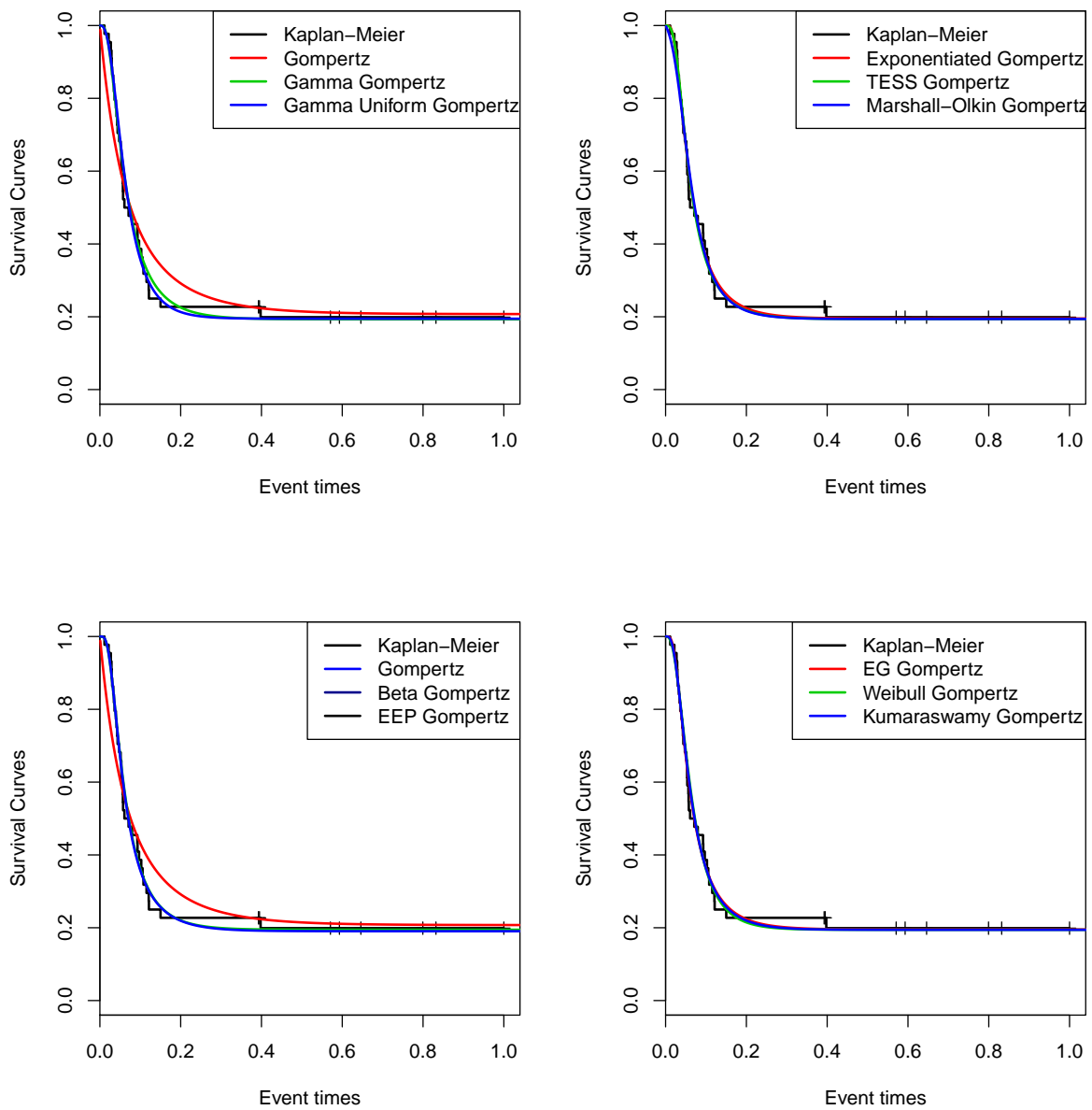


Figure 5.5: Fitted survival curves of the proposed models when the baseline distribution is Gompertz, in the leukemia data set.

smallest AIC between those whose capture the Kaplan-Meier curve properly.

Again, it is quite clear that all of the extended distributions outperforms the inverse Gaussian baseline distribution.

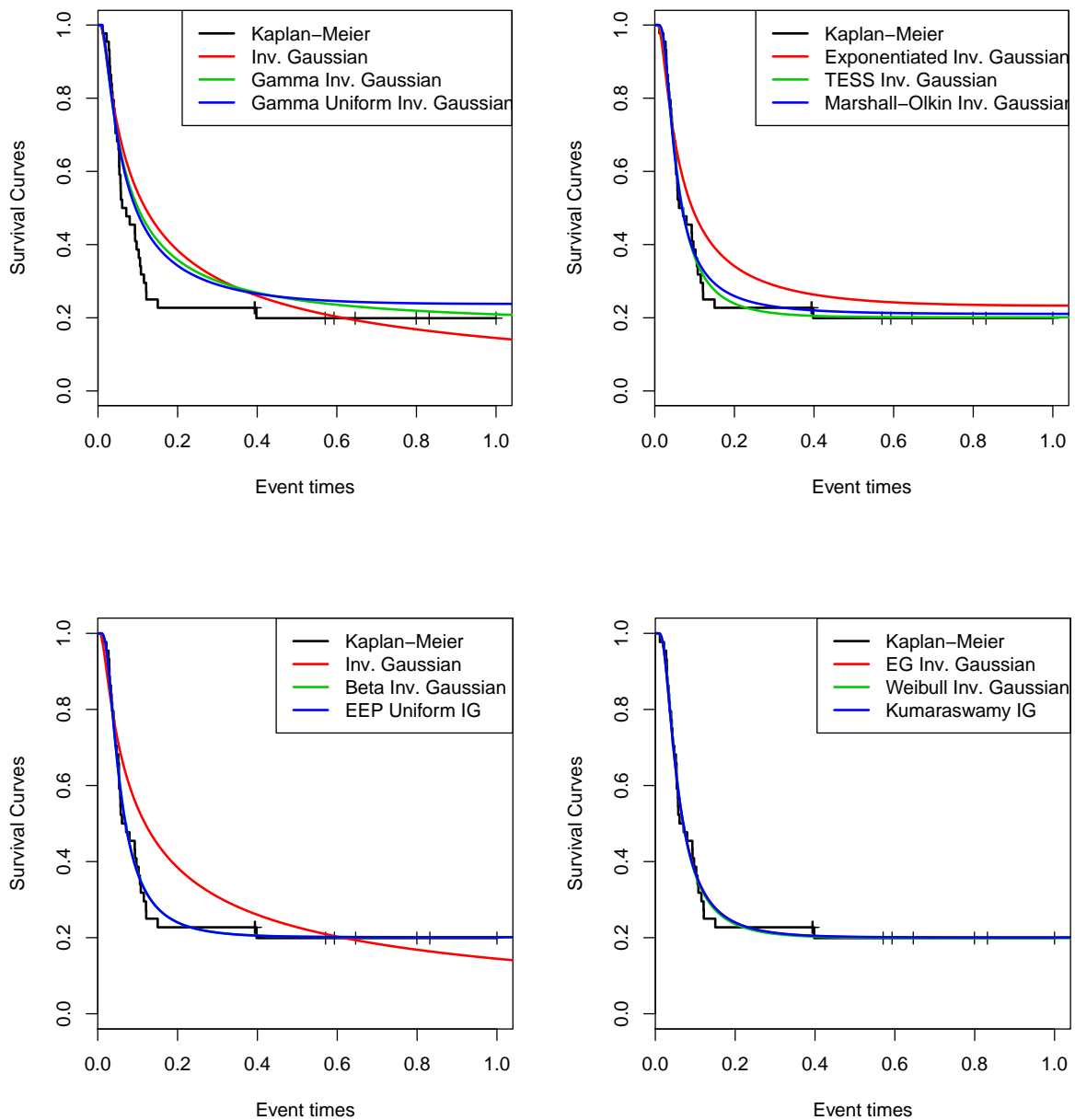


Figure 5.6: Fitted survival curves of the proposed models when the baseline distribution is inverse Gaussian, in the leukemia data set.

5.4.2 Melanoma data

The fitted results for the distributions with Gompertz as baseline are presented in Table 5.3. Figure 5.7 illustrate the respective survival curves. Here all models were lead to be defective. The parameter a is estimated negative for all distributions. The cure fraction was estimated 0.2555 by the Gompertz model and 0.2277 by the TESS Gompertz. These

Table 5.3: Maximum likelihood estimates in the melanoma data set of the proposed models when the baseline distribution is the Gompertz

Distribution	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC
Gompertz	-0.9209	1.2566	-	-	0.2555	375.87
Gamma Gompertz	-4.0571	11.0445	3.0008	-	0.4884	334.76
Gamma uniform Gompertz	-5.0514	5.5032	2.3005	-	0.5013	335.29
Exponentiated Gompertz	-3.8982	6.3941	3.0864	-	0.4859	334.78
TESS Gompertz	-0.7624	0.1594	7.8367	-	0.2277	378.12
Marshall-Olkin Gompertz	-5.6831	24.7232	78.6506	-	0.5069	340.19
Beta Gompertz	-4.0316	2.5272	3.0144	3.4883	0.4880	336.76
EEP Gompertz	-3.9707	3.6496	2.8226	2.7426	0.4880	336.61
EG Gompertz	-3.8982	2.5286	2.5286	3.0864	0.4859	336.78
Weibull Gompertz	-4.7873	2.8033	2.4541	0.6785	0.4983	337.01
Kumaraswamy Gompertz	-4.1656	4.1641	2.8638	2.2782	0.4902	336.72

two have the worst fit, they failed to capture the Kaplan-Meier curve and have the highest AIC values. The other models estimated the cure fraction between 0.4859 and 0.5069, which makes way more sense. Their survival curves properly models the inflections point that came up with this data. The Gamma Gompertz have the smallest AIC between them.

Table 5.4: Maximum likelihood estimates in the melanoma data set of the proposed models when the baseline distribution is the inverse Gaussian

Distribution	\hat{a}	\hat{b}	\hat{r}	\hat{u}	\hat{p}	AIC
Inverse Gaussian	-0.2498	3.3231	-	-	0.1396	342.35
Gamma inv. Gaussian	-0.1934	3.5422	1.0477	-	0.1121	344.35
Gamma uniform inv. Gaussian	-1.6541	10.1062	1.8089	-	0.2261	343.43
Exponentiated inv. Gaussian	0.0195	6.4939	1.6730	-	-	344.15
TESS inv. Gaussian	-0.2907	3.2899	0.0675	-	0.1575	344.35
Marshall-Olkin inv. Gaussian	-0.2804	3.2981	0.9751	-	0.1531	344.35
Beta inv. Gaussian	-2.7494	32.5666	6.8146	2.2961	0.2738	344.39
EEP inv. Gaussian	-9.4365	70.3245	10.8933	16.4963	0.4120	341.43
EG inv. Gaussian	-1.3819	25.8939	1.4396	5.7427	0.1946	345.11
Weibull inv. Gaussian	-9.0216	43.6478	5.4965	1.1178	0.4316	339.34
Kumaraswamy inv. Gaussian	-6.9996	43.3280	7.8071	10.7587	0.4065	341.97

The fitted results for the distributions with inverse Gaussian as baseline are presented in Table 5.4. Figure 5.8 illustrate the respective survival curves. In this case, it looks like none of the proposed models fits the data very closely. They all have some spots in the survival curve that stays quite distant than expected. It seems that is hard for the proposed models to fit the data when its Kaplan-Meier curve changes the curvature too fast. That happens because the baseline distribution does not provide the needed mathematical properties to get the extended families more flexible, in this case. This shows that the extended models have its limitations.

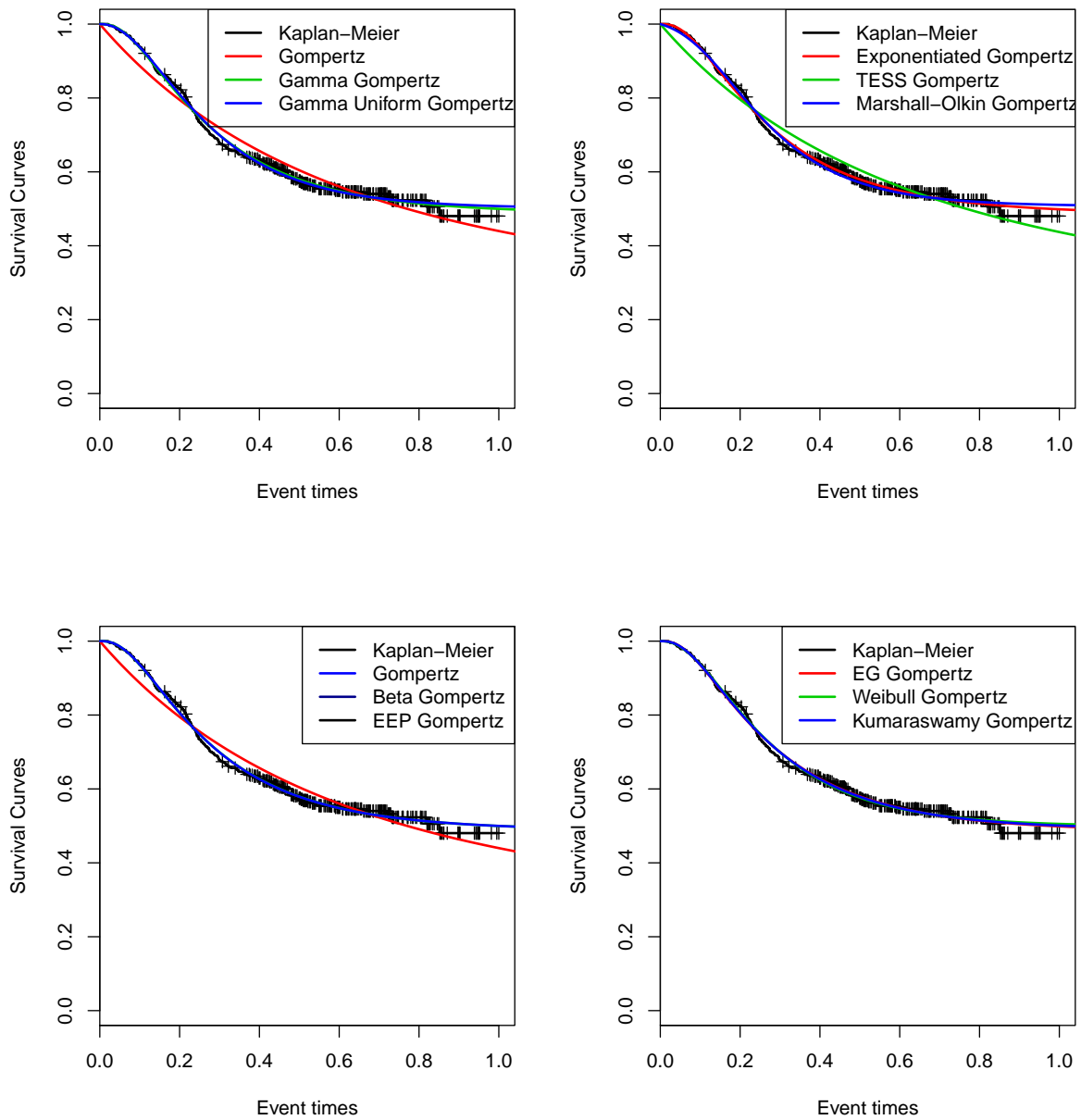


Figure 5.7: Fitted survival curves of the proposed models when the baseline distribution is Gompertz, in the melanoma data set.

The Weibull inverse Gaussian have the smallest AIC. But is a little far from the one obtained with Gompertz as the baseline distribution.

Here, the extended distributions does not outperforms the inverse Gaussian baseline distribution, as in the other cases.

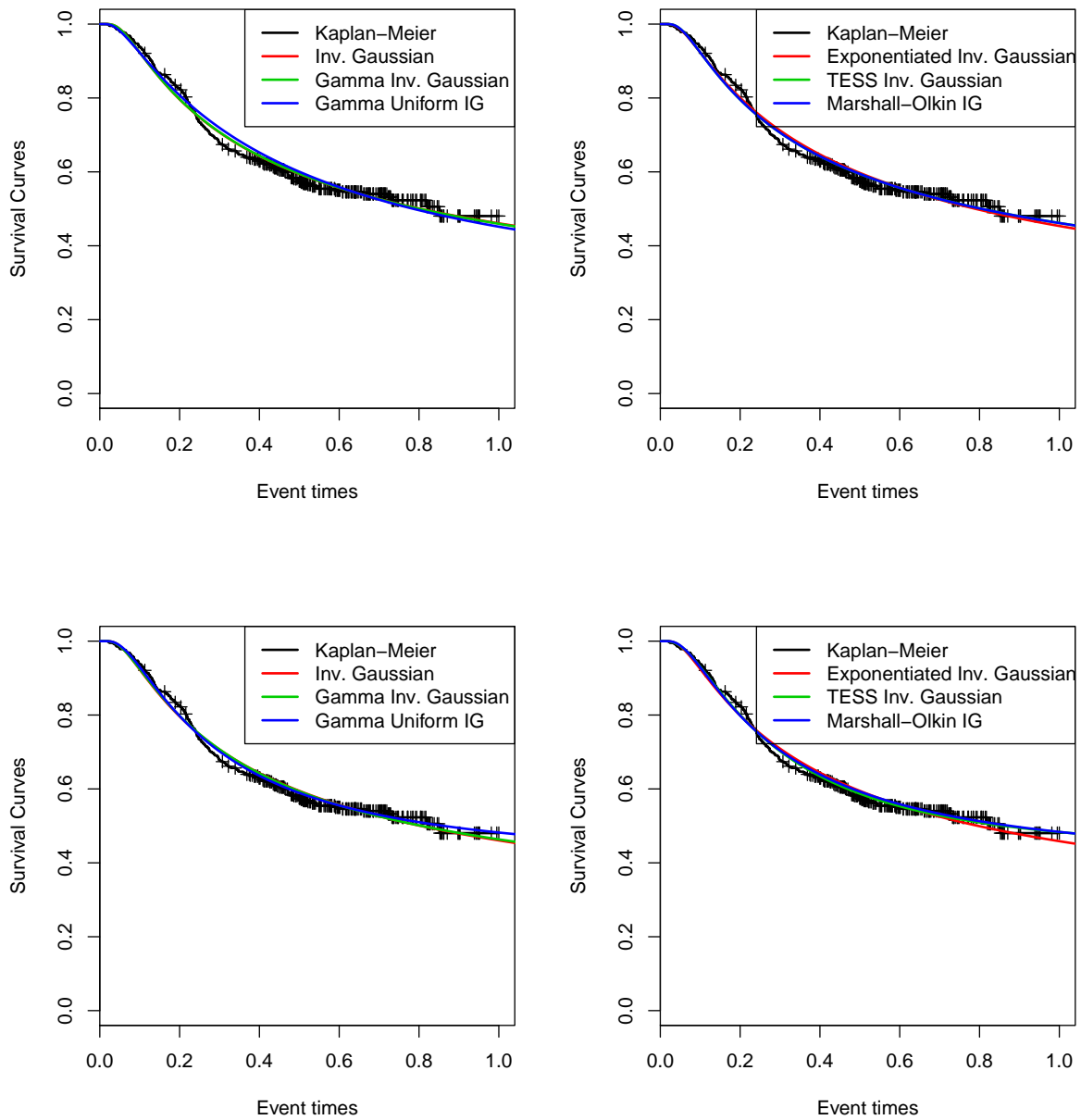


Figure 5.8: Fitted survival curves of the proposed models when the baseline distribution is inverse Gaussian, in the melanoma data set.

5.5 Conclusions

In this chapter we generalized the results obtained in Chapters 3 and 4. This result can lead to an series of new defective models. We present here 16 more extended distributions, besides those 4 already presented in the previous chapters. We did some simulations scenarios to check how the maximum likelihood estimator performs regarding to finite

sample sizes.

We have illustrated the proposed models in two real data sets and showed that the extended defective distributions can be very effective when dealing with cure rate problems, outperforming the baseline distributions. And more important, now with these new defective distributions we can fill the emptiness that was to have only two distributions to work with. Now we have a lot of different models that can properly fit almost any kind of data.

Chapter 6

A Special Class of Defective Models Based on the Marshall-Olkin Family

6.1 Introduction

In this chapter, we derive a useful property of the Marshall Olkin family of distributions which allows one to generate new defective distributions. The details are given in Section 2, including estimation by the method of maximum likelihood and an approach to include covariate information. Simulation studies are performed in Section 6.3 in order to check the usual asymptotic properties of maximum likelihood estimators and to assess the quality of maximum likelihood estimators. Four real data applications of the proposed methodology are illustrated in Section 6.4. Some concluding remarks are given in the last section.

In short, the contributions of this chapter to the literature are: i) derive a new property of the Marshall Olkin family of distributions which allows for the construction of numerous defective distributions; ii) propose ten new defective distributions in order to exemplify the derived property; iii) illustrate the performance of such distributions through simulations and applications to real data sets.

6.2 Methodology

In this section, we present details about the Marshall Olkin family of distributions and derive a new property of this family that can be very useful for cure rate modeling. We also discuss the extended Weibull family of distributions, which together with the Marshall Olkin family can generate a whole set of new distributions for cure fraction estimation. Furthermore, we discuss details of maximum likelihood estimation and an approach to use the proposed distributions as regression models.

6.2.1 The Marshall Olkin family

Let $f(t)$, $S(t)$ and $\lambda(t)$ denote, respectively, the density, survival and hazard rate functions associated with a baseline distribution. The Marshall Olkin (MO) family, proposed in [Marshall & Olkin \(1997\)](#), extends the baseline distribution by adding an extra shape parameter, leading to a more flexible distribution often capable of providing better fits. The density, survival and hazard rate functions of the Marshall Olkin family are

$$f_{MO}(t; r) = \frac{rf(t)}{[1 - (1 - r)S(t)]^2}, \quad (6.1)$$

$$S_{MO}(t; r) = \frac{rS(t)}{1 - (1 - r)S(t)}, \quad (6.2)$$

$$\lambda_{MO}(t; r) = \frac{\lambda(t)}{1 - (1 - r)S(t)} \quad (6.3)$$

for $t > 0$ and $r > 0$.

There has been much work on the Marshall Olkin family of distributions. Many authors have derived details for particular Marshall Olkin distributions. For some examples, see [Jose & Krishna \(2011b\)](#) for the Marshall Olkin-uniform distribution, [Ghitany \(2005\)](#) for the Marshall Olkin-Pareto distribution, [Ghitany et al. \(2005\)](#) for the Marshall Olkin-Weibull distribution, [Ristic et al. \(2007\)](#) for the Marshall Olkin-gamma distribution and [Jose et al. \(2009a\)](#) for the Marshall Olkin-beta distribution.

Theorem 2.1 derives a new property of the Marshall Olkin family that relates to the theory

of defective distributions. This new property allows one to generate of new defective distributions.

Theorem 6.1. *Suppose $S(t)$ is an improper survival function satisfying $\lim_{t \rightarrow \infty} S(t) = \infty$. Then the Marshall Olkin distribution given by (6.1) and (6.2) for $r < 0$ is a defective distribution.*

Proof: If $\lim_{t \rightarrow \infty} S(t) = \infty$ then

$$\begin{aligned} \lim_{t \rightarrow \infty} S_{MO}(t; r) &= \lim_{t \rightarrow \infty} \frac{rS(t)}{1 - (1 - r)S(t)} \\ &\stackrel{L'H}{=} \frac{rS'(t)}{(r - 1)S'(t)} \\ &= \frac{r}{r - 1}, \end{aligned}$$

where $L'H$ indicates the use of the L'Hôpital rule. If $r < 0$ then $\frac{r}{r-1} \in (0, 1)$, so the proof is complete. \square

For example, the exponential distribution has survival function $S(t) = \exp(-at)$, $a > 0$. If $a < 0$, then $\lim_{t \rightarrow \infty} S(t) = \infty$ which satisfies the condition of Theorem 2.1. Therefore, the Marshall Olkin-exponential distribution is a defective distribution when $a < 0$ and $r < 0$.

Theorem 2.1 still holds if $\lim_{t \rightarrow \infty} S(t) = -\infty$. If $\lim_{t \rightarrow \infty} S(t) = M < \infty$ and $M \in (-\infty, 0) \cup (1, \infty)$ then Theorem 2.1 still holds for $r < 0$. The limiting cure rate in this case will be $rM/(rM + 1 - M)$. If $\lim_{t \rightarrow \infty} S(t) = M < \infty$ and $M \in (0, 1)$ then Theorem 2.1 still holds for $r > 0$. In this case, the distribution becomes a defective by definition. If $M = 0$ then $S(t)$ is a proper survival function and therefore there is no cure rate. $M = 1$ corresponds to a degenerate distribution with the cure rate of 1 (no one would be susceptible to the event of interest).

Section 6.2.2 shows that a known family of extended Weibull distributions can give ideal choices for $S(t)$.

6.2.2 The extended Weibull distribution

The extended Weibull (EW) distribution, firstly proposed in [Gurvich *et al.* \(1997\)](#), generalizes the Weibull distribution by means of a non-negative monotonically increasing function $H(t, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a vector of k parameters. Its density, survival and hazard rate functions are

$$f_{EW}(t; v, \boldsymbol{\gamma}) = v h(t, \boldsymbol{\gamma}) \exp[-vH(t, \boldsymbol{\gamma})], \quad (6.4)$$

$$S_{EW}(t; v, \boldsymbol{\gamma}) = \exp[-vH(t, \boldsymbol{\gamma})], \quad (6.5)$$

$$\lambda_{EW}(t; v, \boldsymbol{\gamma}) = v h(t, \boldsymbol{\gamma}) \quad (6.6)$$

for $t > 0$, $v > 0$ and $h(t, \boldsymbol{\gamma}) = dH(t, \boldsymbol{\gamma})/dt$.

Different choices for $H(t, \boldsymbol{\gamma})$ lead to different extended Weibull distributions. [Table 6.1](#) lists ten extended Weibull distributions which will be used to illustrate [Theorem 6.1](#). They were selected from [Santos-Neto *et al.* \(2014\)](#).

Table 6.1: Some particular cases of the extended Weibull distribution.

Distribution	$H(t, \boldsymbol{\gamma})$	Parameters in $\boldsymbol{\gamma}$
Exponential	t	\emptyset
Rayleigh	t^2	\emptyset
Lomax	$\log(1 + t)$	\emptyset
Weibull	t^a	$a > 0$
Gompertz	$[\exp(at) - 1] / a$	$a > 0$
Burr XII	$\log(1 + t^a)$	$a > 0$
Chen	$\exp(t^a) - 1$	$a > 0$
Modified Weibull	$t^a \exp(bt)$	$a \geq 0, b > 0$
Weibull extension	$a \{ \exp[(t/a)^b] - 1 \}$	$a > 0, b > 0$
Traditional Weibull	$t^b [\exp(at^c) - 1]$	$a \geq 0, b \geq 0, c > 0$

Some more distributions for positive data can be obtained from the extended Weibull family: the Pareto distribution for $H(t, \boldsymbol{\gamma}) = \log(t/a)$, $t \geq a$; the log-logistic distribution for $H(t, \boldsymbol{\gamma}) = \log(1 + t^a)$; the Fréchet distribution for $H(t, \boldsymbol{\gamma}) = t^{-a}$; the exponential power distribution for $H(t, \boldsymbol{\gamma}) = \exp[(at)^b] - 1$; the Pham distribution for $H(t, \boldsymbol{\gamma}) = (at)^b - 1$, among others. For more details, see [Santos-Neto *et al.* \(2014\)](#).

Note that some distributions in [Table 6.1](#) are generalizations of others: the exponential

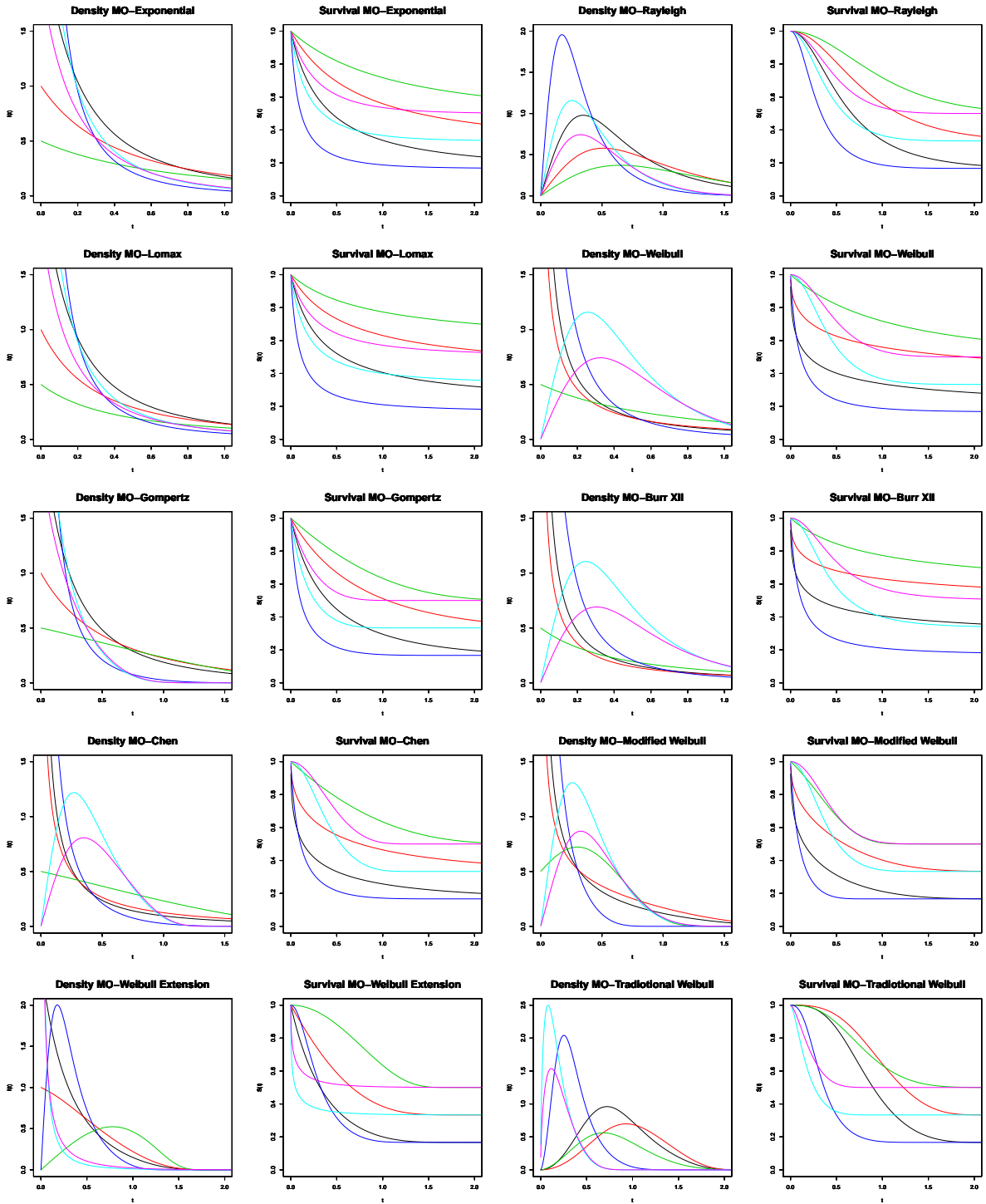


Figure 6.1: From the left to the right, from the top to the bottom, the density and survival functions of the proposed distributions, in the same order presented in Table 6.1. The parameter values used are $u = (-0.2, -0.5, -1, -0.2, -0.5, -1)$, $v = (-0.5, -0.5, -0.5, -2, -2, -2)$, $a = (0.5, 0.5, 1, 1, 2, 2)$, $b = (1, 1, 2, 2, 0.5, 0.5)$ and $c = (2, 2, 0.5, 0.5, 1, 1)$. The colors are (black, red, green, blue, light blue, pink).

and Rayleigh distributions are particular cases of the Weibull distribution for $a = 1$ and $a = 2$, respectively; the Lomax distribution is the particular case of the Burr XII distribution for $a = 1$; the Weibull distribution is the particular case of the modified Weibull distribution for $b = 0$; the Chen and Gompertz distributions are particular cases of the Weibull extension distribution for $a = 1$ and $a' = a^{-1}$, $b = 1$, respectively; the Weibull distribution is the particular case of the Chen distribution for $a = 1$, $b = 0$.

If $v < 0$ then $\lim_{t \rightarrow \infty} S_{EW}(t; v, \gamma) = \infty$ provided that $H(t, \gamma)$ is non-negative and monotonically increasing. So, any member of the extended Weibull family that uses v as a parameter can be used to generate a defective distribution.

The Marshall Olkin-extended Weibull (MOeW) distributions are obtained by combining (6.4), (6.5), (6.6) and (6.1), (6.2), (6.3), i.e., by using the extended Weibull distribution as a baseline distribution for the Marshall Olkin family. The resulting density, survival and hazard rate functions are

$$f_{MOeW}(t; r, v, \gamma) = \frac{r v h(t, \gamma) \exp[-vH(t, \gamma)]}{\{1 - (1 - r) \exp[-vH(t, \gamma)]\}^2}, \quad (6.7)$$

$$S_{MOeW}(t; r, v, \gamma) = \frac{r \exp[-vH(t, \gamma)]}{1 - (1 - r) \exp[-vH(t, \gamma)]}, \quad (6.8)$$

$$\lambda_{MOeW}(t; r, v, \gamma) = \frac{v h(t, \gamma)}{1 - (1 - r) \exp[-vH(t, \gamma)]}.$$

The ten different functions in Table 6.1 lead to ten different defective distributions. Figure 6.1 plots the density and survival functions of all distributions proposed in Table 6.1. This collection of distributions can be very flexible. The black and blue curves in the figure have the same cure rate of $-0.2/(-0.2 - 1) = 1/6$. The red and light blue curves have the curve rate of $-0.5/(-0.5 - 1) = 1/3$. The green and pink curves have the cure rate of $-1/(-1 - 1) = 1/2$.

We have used the extended Weibull family to generate defective distributions. The generated distributions give good fits to the data considered in this chapter. But other distributions could have been used to generate defective versions via Theorem 6.1. As an example, consider the Maxwell-Boltzmann distribution specified by the density and

survival functions

$$f_{MB}(t; a) = a^{-3}t^2 \exp\left(-\frac{t^2}{2a^2}\right) \sqrt{2\pi^{-1}},$$

$$S_{MB}(t; a) = 1 - \operatorname{erf}\left(\frac{t}{\sqrt{2}a}\right) + \frac{t \exp\left(-\frac{t^2}{2a^2}\right) \sqrt{2\pi^{-1}}}{a}$$

for $t > 0$, where $a > 0$ is a scale parameter and $\operatorname{erf}(t) = 2\pi^{-\frac{1}{2}} \int_0^t e^{-x^2} dx$ denotes the error function. The error function approaches 1 as $t \rightarrow \infty$ and approaches -1 as $t \rightarrow -\infty$. If $a < 0$ we have

$$\lim_{t \rightarrow \infty} S_{MB}(t; a) = 1 - (-1) + 0 = 2.$$

So, this distribution under the Marshall Olkin family is defective when $a < 0$ and $r < 0$. Its cure rate is $2r/(2r + 1 - 2) = r/(r - 0.5)$.

6.2.3 Inference

Here, we present a procedure to obtain maximum likelihood estimates for the MOeW distribution, when considering data with right-censored information. Let $\mathbf{D} = (\mathbf{t}, \boldsymbol{\delta})$, where $\mathbf{t} = (t_1, \dots, t_n)'$ are the observed failure times and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)'$ are the right-censored times. The δ_i is equal to 1 if a failure is observed and 0 otherwise. Suppose that the data are independently and identically distributed and come from a distribution with density and survival functions specified by $f(\cdot, \boldsymbol{\theta})$ and $S(\cdot, \boldsymbol{\theta})$, respectively, where $\boldsymbol{\theta} = (r, v, \gamma)'$ denotes a vector of $k + 2$ parameters. The log-likelihood function of $\boldsymbol{\theta}$ can be written as

$$l(\boldsymbol{\theta}, \mathbf{D}) = \log L(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + \sum_{i=1}^n \delta_i \log f(t_i, \boldsymbol{\theta}) + (1 - \delta_i) \log S(t_i, \boldsymbol{\theta}). \quad (6.9)$$

By (6.7) and (6.8), the log-likelihood function for the MOeW distribution is

$$l(\boldsymbol{\theta}, \mathbf{D}) = \text{const} + n \log(r) - v \sum_{i=1}^n H(t_i, \boldsymbol{\gamma}) - \sum_{i=1}^n (1 + \delta_i) \log \{1 - (1 - r) \exp[-vH(t_i, \boldsymbol{\gamma})]\} + \sum_{i=1}^n \delta_i \log [vh(t_i, \boldsymbol{\gamma})].$$

The maximum likelihood estimates are the simultaneous solutions of $\frac{\partial l(\boldsymbol{\theta}, \mathbf{D})}{\partial r} = 0$, $\frac{\partial l(\boldsymbol{\theta}, \mathbf{D})}{\partial v} = 0$ and $\frac{\partial l(\boldsymbol{\theta}, \mathbf{D})}{\partial \gamma_j} = 0$. Asymptotic normality of the maximum likelihood estimates holds only under certain regularity conditions. These conditions are not easy to check analytically for our models. Section 6.3 performs a simulation study to see if the usual asymptotes of the maximum likelihood estimates hold. Simulations have been used in many papers to check the asymptotic behavior of maximum likelihood estimates, especially when an analytical investigation is not trivial.

6.2.4 Defective Marshall Olkin- G regression model

The use of covariate information is essential when analysing survival data. Here, we discuss an approach on how to include covariate information to the proposed models. The approach has a simple interpretation as we shall see.

Suppose $\mathbf{x}' = (1, x_1, \dots, x_p)$ is a vector of covariates from a data set and $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$ a vector of regression coefficients. We are going to set $r(\mathbf{x}) = -\exp(\boldsymbol{\beta}'\mathbf{x})$ to link the cure rate to the covariates. In this way, the Marshall Olkin- G regression model is given by

$$S(t|\mathbf{x}) = \frac{r(\mathbf{x})S(t)}{1 - [1 - r(\mathbf{x})]S(t)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})S(t)}{[1 + \exp(\boldsymbol{\beta}'\mathbf{x})]S(t) - 1},$$

for $t > 0$. If $S(t)$ has a cure rate of p then that of $S(t|\mathbf{x})$ is

$$p = \lim_{t \rightarrow \infty} S(t|\mathbf{x}) = \frac{r(\mathbf{x})}{r(\mathbf{x}) - 1} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}. \quad (6.10)$$

In this way, the cure fraction is easily calculate throw the logit function. This approach is

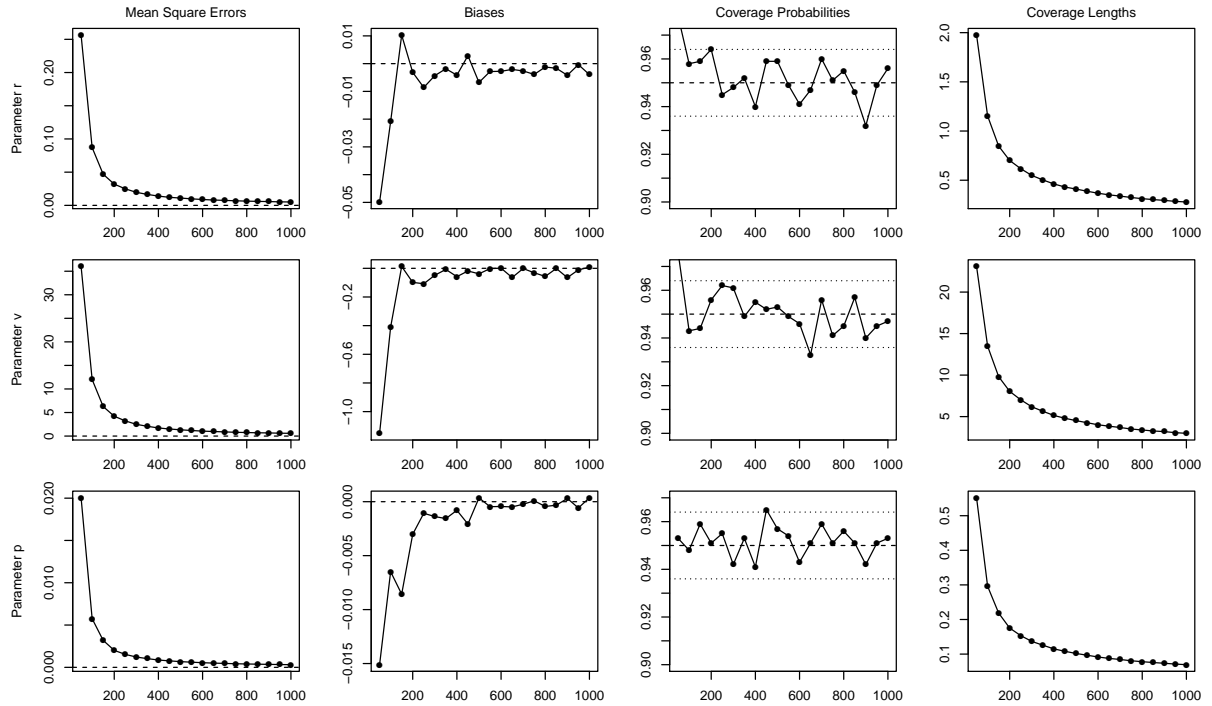


Figure 6.2: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v and p versus n for the Marshall Olkin-Lomax distribution with $(r, v) = (-1, -10)$.

attractive because of the way the cure rate depends on the regression coefficients, making it very easy to interpret. If $\beta'x$ increases its value, so does the cure rate (towards 1). If $\beta'x$ decreases its value, so does the cure rate (towards 0).

The MOeW regression model is given by

$$S(t|x) = \frac{r(x) \exp[-vH(t, \gamma)]}{1 - [1 - r(x)] \exp[-vH(t, \gamma)]} = \frac{\exp(\beta'x) \exp[-vH(t, \gamma)]}{[1 + \exp(\beta'x)] \exp[-vH(t, \gamma)] - 1}$$

An application is presented in Section 6.4.4.

6.3 Simulation studies

Here, we assess the performance of the maximum likelihood estimates with respect to sample size to show, among other things, that the usual asymptotes of maximum likelihood estimators still hold for defective distributions. The assessment is based on simulations.

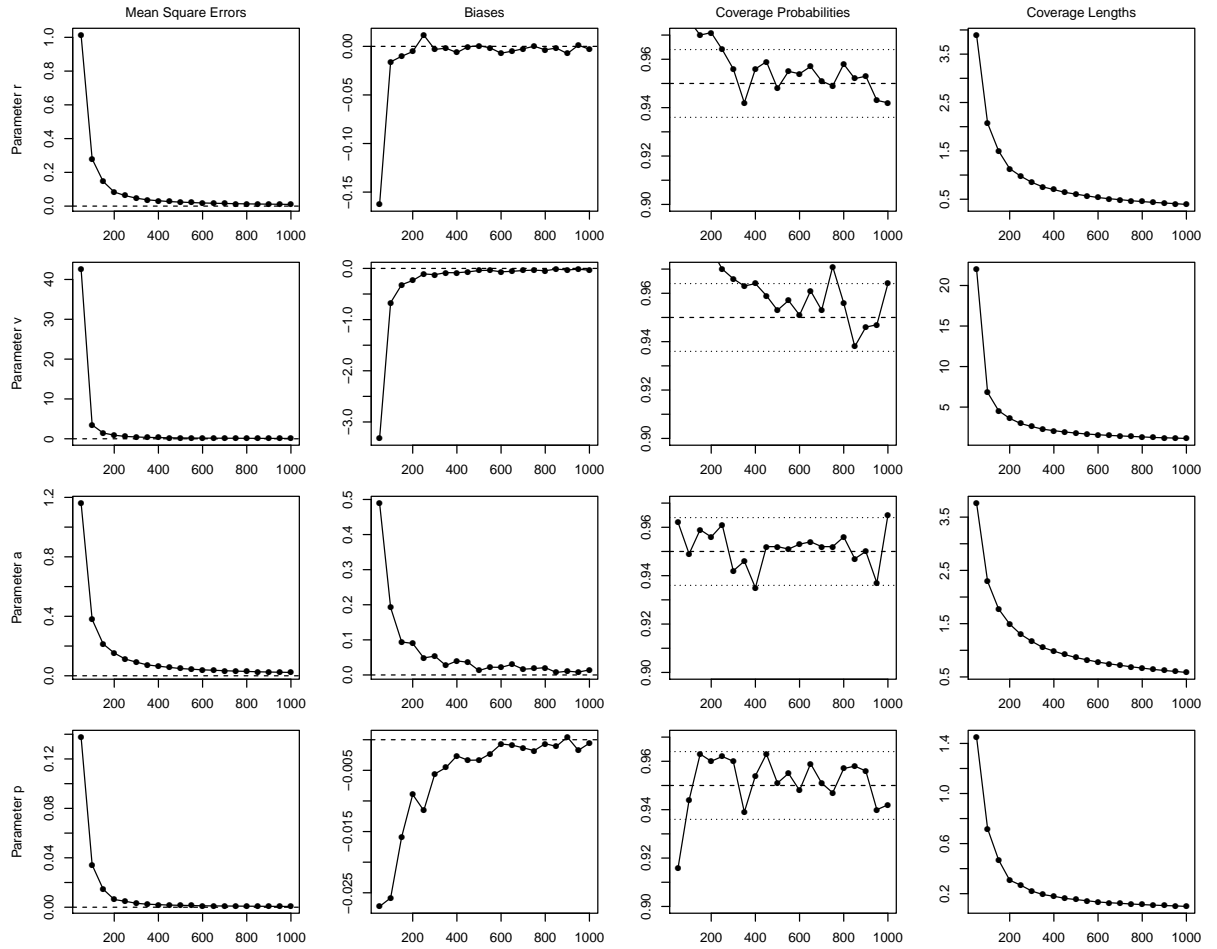


Figure 6.3: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Weibull distribution with $(r, v, a) = (-1, -2, 3)$.

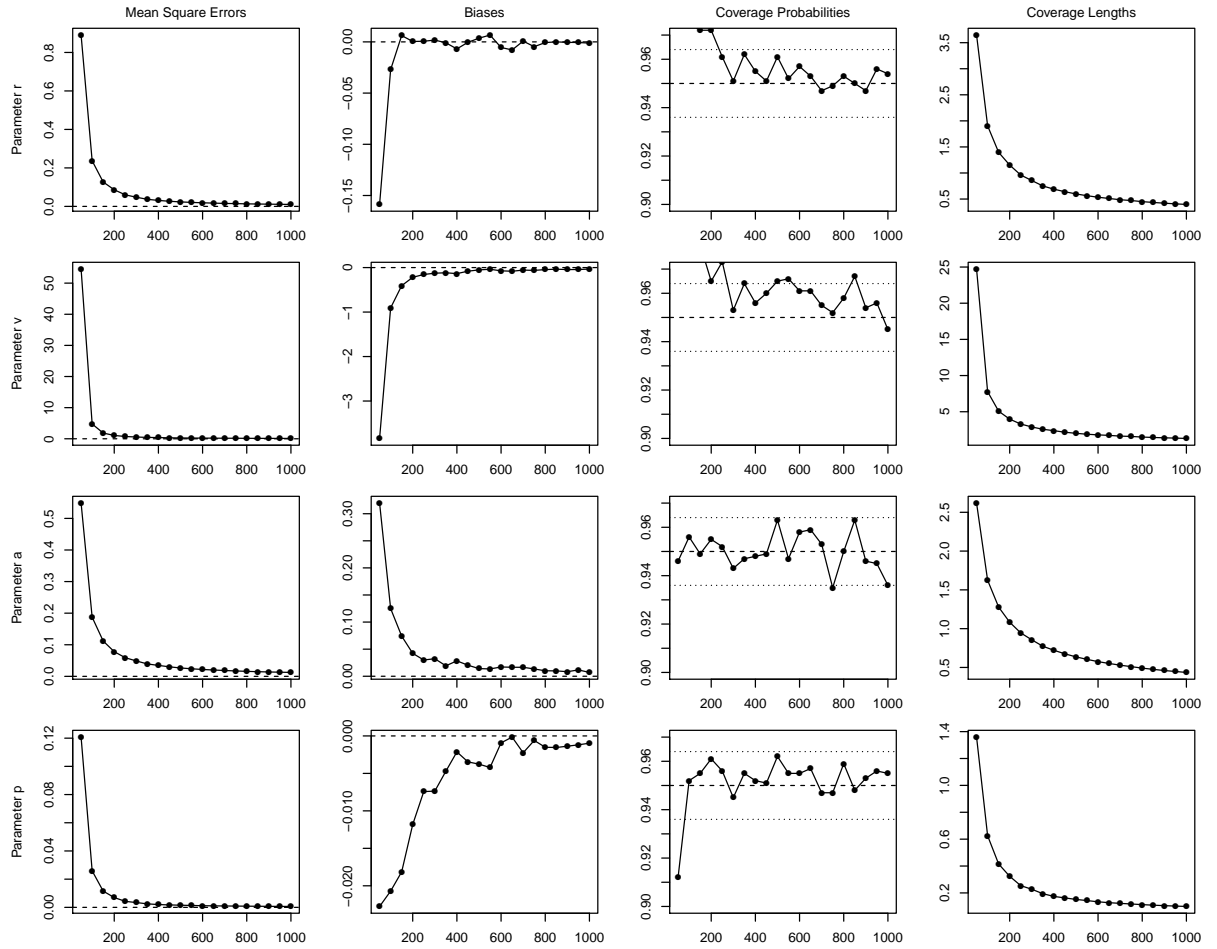


Figure 6.4: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Chen distribution with $(r, v, a) = (-1, -2, 2)$.

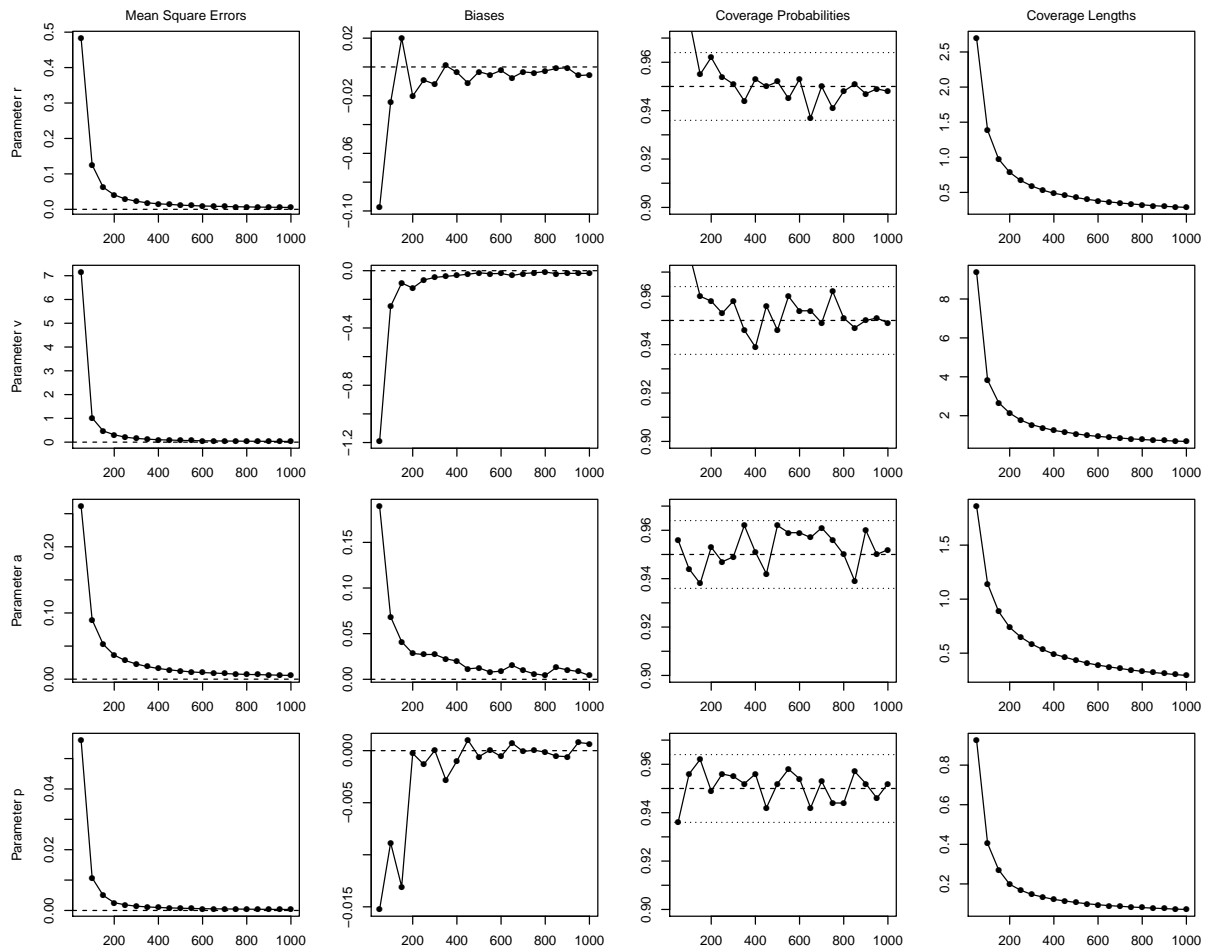


Figure 6.5: Mean squared errors, biases, coverage probabilities and coverage lengths of the estimators of r , v , a and p versus n for the Marshall Olkin-Burr XII distribution with $(r, v, a) = (-1, -2, 2)$.

We took the sample size to vary from 50 to 1000 in steps of 50. Each sample was replicated 1000 times. The variance of the cure rate p was estimated using the delta method with first order Taylor's approximation. We chose only four of our proposed distributions: the Marshall Olkin-Lomax distribution with $(r, v) = (-1, -10)$, the simulation results for which are shown in Figure 6.2; the Marshall Olkin-Weibull distribution with $(r, v, a) = (-1, -2, 3)$, the simulation results for which are shown in Figure 6.3; the Marshall Olkin-Chen distribution with $(r, v, a) = (-1, -2, 2)$, the simulation results for which are shown in Figure 6.4; the Marshall Olkin-Burr XII distribution with $(r, v, a) = (-1, -2, 2)$, the simulation results for which are shown in Figure 6.5. For the purpose of comparison, we have fixed $r = -1$ for all simulations, which leads to a cure rate of 0.5.

We can observe the following from the figures: the biases for each parameter approach zero as sample size increases; the biases for each parameter appear small enough for all $n \geq 600$; the mean squared errors for each parameter decrease to zero as sample size increases; the mean squared errors for each parameter appear small enough for all $n \geq 600$; the coverage probabilities for each parameter stay mostly in the interval $(0.936, 0.964)$; the coverage lengths for each parameter decrease fast to zero as sample size increases; the coverage lengths for each parameter appear small enough for all $n \geq 600$.

Similar observations held when the simulations were repeated for other defective distributions and for a wide range of parameter values under the Marshall Olkin family. In particular, the biases always approached zero as sample size increased, the biases for each parameter always appeared small enough for all $n \geq 600$, the mean squared errors always approached zero as sample size increased, the mean squared errors for each parameter always appeared small enough for all $n \geq 600$, the coverage probabilities always stayed mostly in the interval $(0.936, 0.964)$, the coverage lengths always decreased fast to zero as sample size increased and the coverage lengths for each parameter always appeared small enough for all $n \geq 600$.

6.4 Real data applications

Here, we present applications to four real data sets. In the first three data sets, we are only considering the event times and censoring information, with no covariates. The fourth data set contains covariate information and is used to illustrate the model proposed in Section 6.2.4. The ten defective distributions discussed in Section 6.2.2 are fitted to each data set. The following are used to distinguish between the fitted distributions: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the consistent Akaike information criterion (CAIC) and visual comparison of the fitted survival curves and the Kaplan-Meier curve. For computational stability, the observed times in each data set were divided by their maximum value. The parameters r and v were set free to take any value on the real line. Negative estimates of r and v correspond to a defective model. Positive estimates of r and v correspond to a proper survival model.

The four data sets were chosen to show a variety of survival curves and sample sizes. Each data set is supposed to contain observations not susceptible to the event of interest. In practice, it is unknown if the event of interest could be observed if enough time was given. An evidence of existence of cured individuals is when the Kaplan-Meier curve reaches a plateau between zero and one. In some cases that is more clear than others, as one can see in our examples. We can assume that some of the censored observations at the end of the study belong to the cured group. If everyone censored at the end are indeed cured, then the plateau reached by the Kaplan-Meier curve is a good estimate of the cure fraction. In general, a lower value of this plateau or a value close to it is an acceptable estimate.

6.4.1 Leukemia data

Here we consider the leukemia data. The fitted results are summarized in Table 6.2 and Figure 6.6. Every distribution is estimated as a defective distribution. The cure rate estimates are around 0.02 lower than the value suggest by the Kaplan-Meier curve. The Marshall Olkin-Rayleigh distribution gives the smallest values for AIC, BIC and CAIC, suggesting it fits better than the others. Its estimate of the cure fraction is furthest from

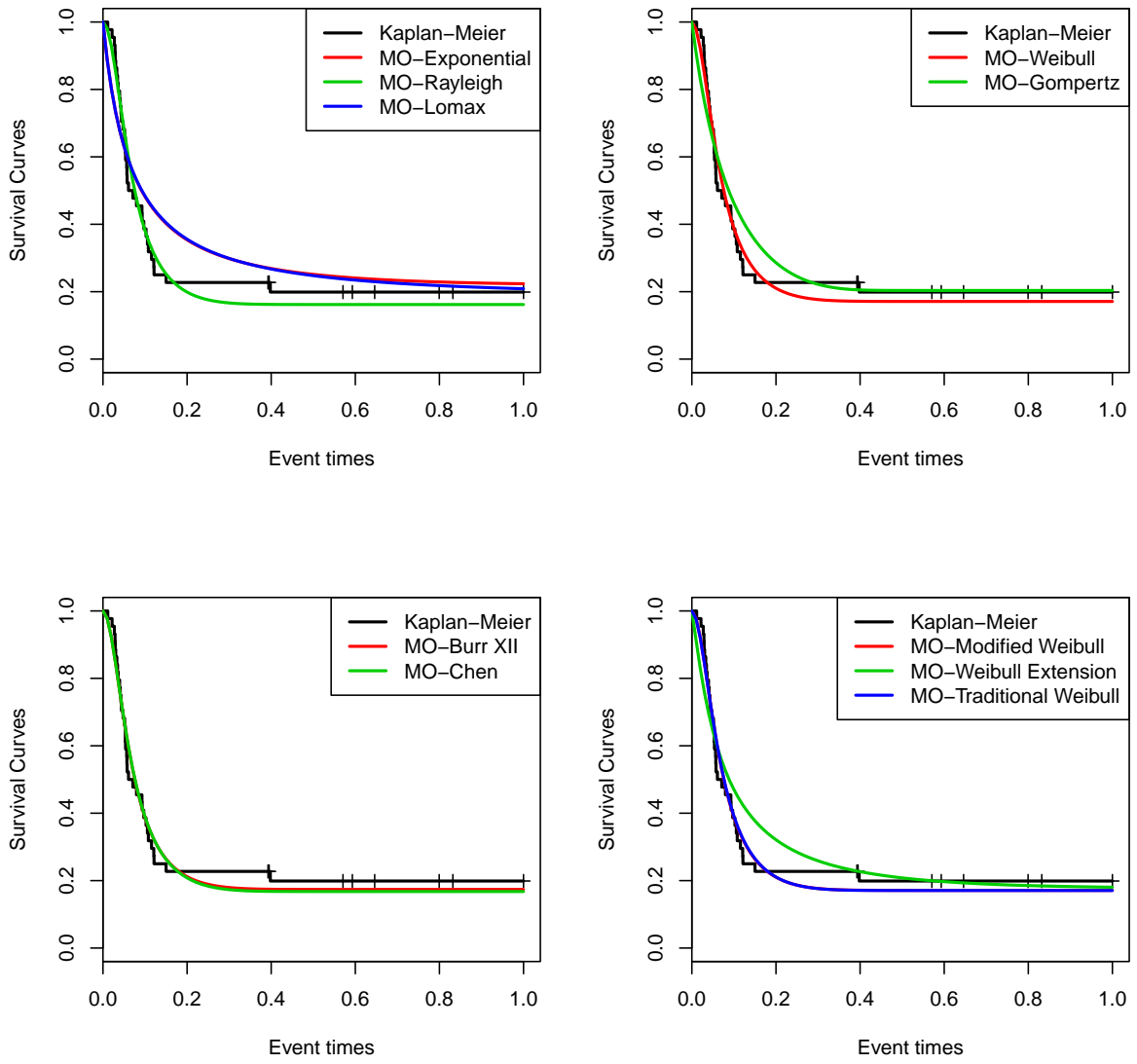


Figure 6.6: Fitted distributions for the leukemia data set.

Table 6.2: MLEs for the fitted distributions and some measures for the leukemia data set.

MO-Distribution	$\hat{\tau}$	$\hat{\nu}$	\hat{a}	\hat{b}	\hat{c}	\hat{p}	AIC	BIC	CAIC
Exponential	-0.2798	-3.5807	-	-	-	0.2186	-54.86	-51.29	-54.57
Rayleigh	-0.1932	-37.6941	-	-	-	0.1619	-75.39	-71.82	-75.10
Lomax	-0.2296	-2.9506	-	-	-	0.1867	-54.03	-50.46	-53.74
Weibull	-0.2064	-33.0993	1.9246	-	-	0.1711	-73.50	-68.15	-72.90
Gompertz	-0.2557	-2.5202	6.3571	-	-	0.2036	-58.31	-52.96	-57.71
Burr XII	-0.2103	-35.3517	1.9384	-	-	0.1738	-73.93	-68.58	-73.33
Chen	-0.2015	-30.8096	1.9111	-	-	0.1677	-73.00	-67.65	-72.40
Modified Weibull	-0.2064	-33.224	1.9259	0.0002	-	0.1711	-71.50	-64.37	-70.48
Weibull extension	-0.2053	-56.4223	1.935	1.9045	-	0.1703	-71.36	-64.22	-70.33
Traditional Weibull	-0.2057	-1.6413	3.1087	1.5997	0.1237	0.1706	-69.09	-60.17	-67.51

the one suggested by the Kaplan-Meier curve, 0.1619, but still an acceptable estimate. The Marshall Olkin-Weibull, Marshall Olkin-Burr XII, Marshall Olkin-Chen and Marshall Olkin-Modified Weibull distributions also provide reasonable fits. All other distributions perform poorly. Visual comparison of the fitted survival curves and the Kaplan-Meier curve shows that the Marshall Olkin-Exponential, Marshall Olkin-Lomax, Marshall Olkin-Gompertz and Marshall Olkin-Weibull extension distributions provide the worst fits.

6.4.2 Colon data

Table 6.3: MLEs for the fitted distributions and some measures for the colon data set.

MO-Distribution	$\hat{\tau}$	$\hat{\nu}$	\hat{a}	\hat{b}	\hat{c}	\hat{p}	AIC	BIC	CAIC
Exponential	-0.5871	-1.2272	-	-	-	0.3699	1531.10	1542.15	1531.10
Rayleigh	-0.8655	-8.6495	-	-	-	0.464	1668.31	1679.36	1668.32
Lomax	-0.2282	-0.4812	-	-	-	0.1858	1537.00	1548.06	1537.01
Weibull	-0.8805	-3.6376	1.367	-	-	0.4682	1462.36	1478.94	1462.38
Gompertz	-0.9101	-1.6598	1.9054	-	-	0.4765	1516.68	1533.27	1516.70
Burr XII	-0.8381	-3.9545	1.4167	-	-	0.456	1456.53	1473.11	1456.54
Chen	-0.9114	-3.0808	1.2918	-	-	0.4768	1474.74	1491.32	1474.75
Modified Weibull	-0.8809	-3.6404	1.3672	0.0014	-	0.4683	1464.39	1486.50	1464.41
Weibull extension	-0.8805	-14.1338	40.9902	1.366	-	0.4682	1464.42	1486.52	1464.44
Traditional Weibull	-0.8805	-1.3754	1.2936	1.355	0.0068	0.4682	1466.37	1494.00	1466.40

Here we consider the colon data set. The fitted results are summarized in Table 6.3 and Figure 6.7. All of the fitted distributions are estimated to being defective. The Marshall Olkin-Lomax distribution estimates the cure fraction as 0.1858, far lower than the Kaplan-Meier plateau. The Marshall Olkin-Exponential distribution gives the estimate 0.3699 and the Marshall Olkin-Weibull distribution gives the estimate 0.4198. All others give a value very close to the Kaplan-Meier estimate. The Marshall Olkin-Burr XII distribution has the smallest values for AIC, BIC and CAIC. This distribution gives a cure rate of 0.456, slightly lower than the Kaplan-Meier estimate and is probably the best for this data. Note that the Marshall Olkin-Weibull, Marshall Olkin-Modified Weibull, Marshall Olkin-Weibull extension and Marshall Olkin-Traditional Weibull distributions give practically

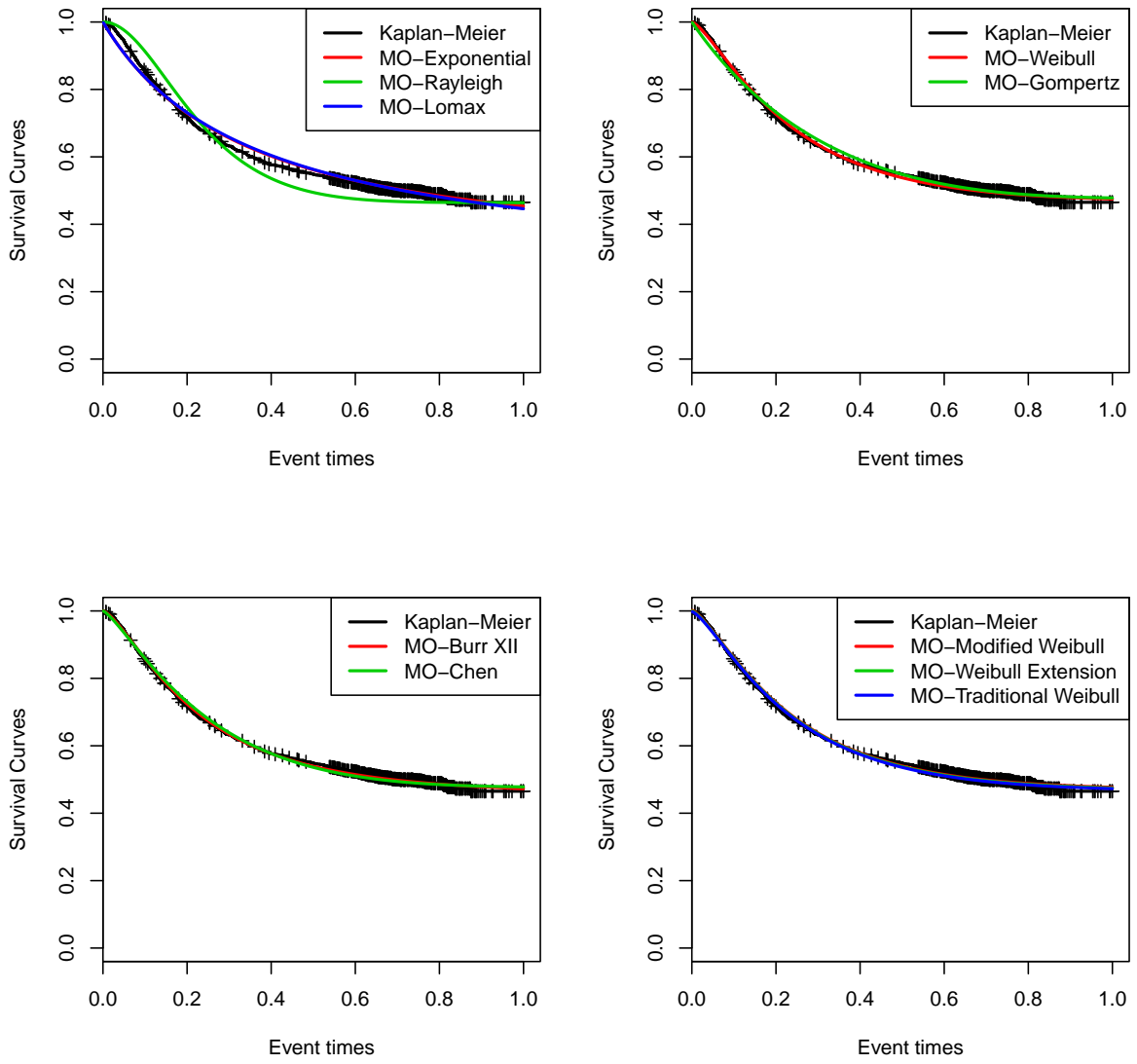


Figure 6.7: Fitted distributions for the colon data set.

the same cure rate estimate of 0.4682, very close to the Kaplan-Meier estimate.

Visual comparison of the fitted survival curves and the Kaplan-Meier curve shows that the Marshall Olkin-Rayleigh distribution gives the worst fit (and the worst measures for AIC, BIC and CAIC too). The Marshall Olkin-Exponential and Marshall Olkin-Lomax distributions provide a better comparison, but their fits are worst than all others (also in agreement with the AIC, BIC and CAIC values). The remaining distributions seem to fit the Kaplan-Meier curve well. The cure rate asymptotes for the Marshall Olkin-Modified Weibull, Marshall Olkin-Exponential and Marshall Olkin-Lomax distributions are after the end of the study.

6.4.3 Divorce data

Table 6.4: MLEs for the fitted distributions and some measures for the divorce data set.

MO-Distribution	$\hat{\tau}$	$\hat{\nu}$	\hat{a}	\hat{b}	\hat{c}	\hat{p}	AIC	BIC	CAIC
Exponential	-0.7037	-1.3674	-	-	-	0.4130	1532.16	1544.41	1532.17
Rayleigh	-1.2283	-16.5389	-	-	-	0.5512	1633.88	1646.13	1633.89
Lomax	-0.2604	-0.5045	-	-	-	0.2066	1538.82	1551.07	1538.83
Weibull	-1.2215	-5.8018	1.4083	-	-	0.5499	1435.90	1454.27	1435.90
Gompertz	-1.2622	-1.9082	3.9220	-	-	0.5579	1471.46	1489.82	1471.46
Burr XII	-1.1819	-6.1051	1.4355	-	-	0.5417	1437.27	1455.64	1437.28
Chen	-1.2498	-5.2752	1.3694	-	-	0.5555	1435.01	1453.38	1435.02
Modified Weibull	-1.2350	-5.3745	1.3820	0.2300	-	0.5526	1437.73	1462.22	1437.74
Weibull extension	-1.2497	-5.2844	1.0033	1.3696	-	0.5555	1437.01	1461.51	1437.03
Traditional Weibull	-1.2499	-5.2052	1.0112	0.0036	1.3651	0.5555	1439.02	1469.64	1439.04

Here we consider the divorce data set. The fitted results are summarized in Table 6.4 and Figure 6.8. The Marshall Olkin-Chen distribution has the smallest values for AIC, BIC and CAIC. Its cure estimate is 0.5555, the closest to the Kaplan-Meier estimate. Its fit captures the Kaplan-Meier curve very well. Therefore, we can consider Marshall Olkin-Chen distribution as giving the most adequate fit. The Marshall Olkin-Modified Weibull, Marshall Olkin-Weibull extension and Marshall Olkin-Traditional Weibull distributions also give very close fits as the Marshall Olkin-Chen distribution. Their measures differ basically because of the difference in the number of parameters. The simplest Marshall Olkin-Exponential, Marshall Olkin-Rayleigh and Marshall Olkin-Lomax distributions all give poor fits. The remaining distributions provide reasonably good fits with respect to AIC, BIC and CAIC measures as well as visual comparison to the Kaplan-Meier curve. Their cure rate estimates are quite close to the value suggested by the Kaplan-Meier curve.

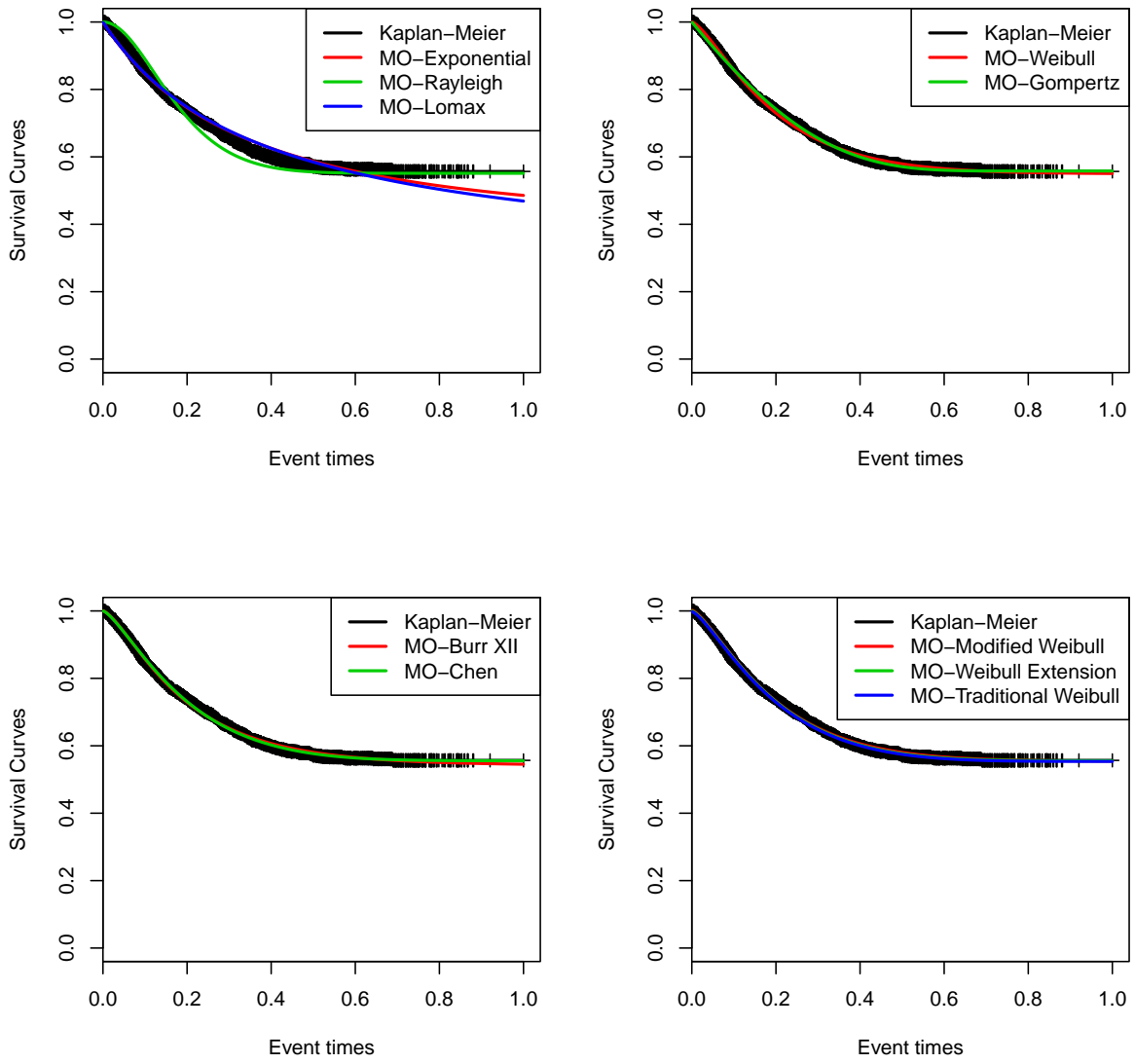


Figure 6.8: Fitted distributions for the divorce data set.

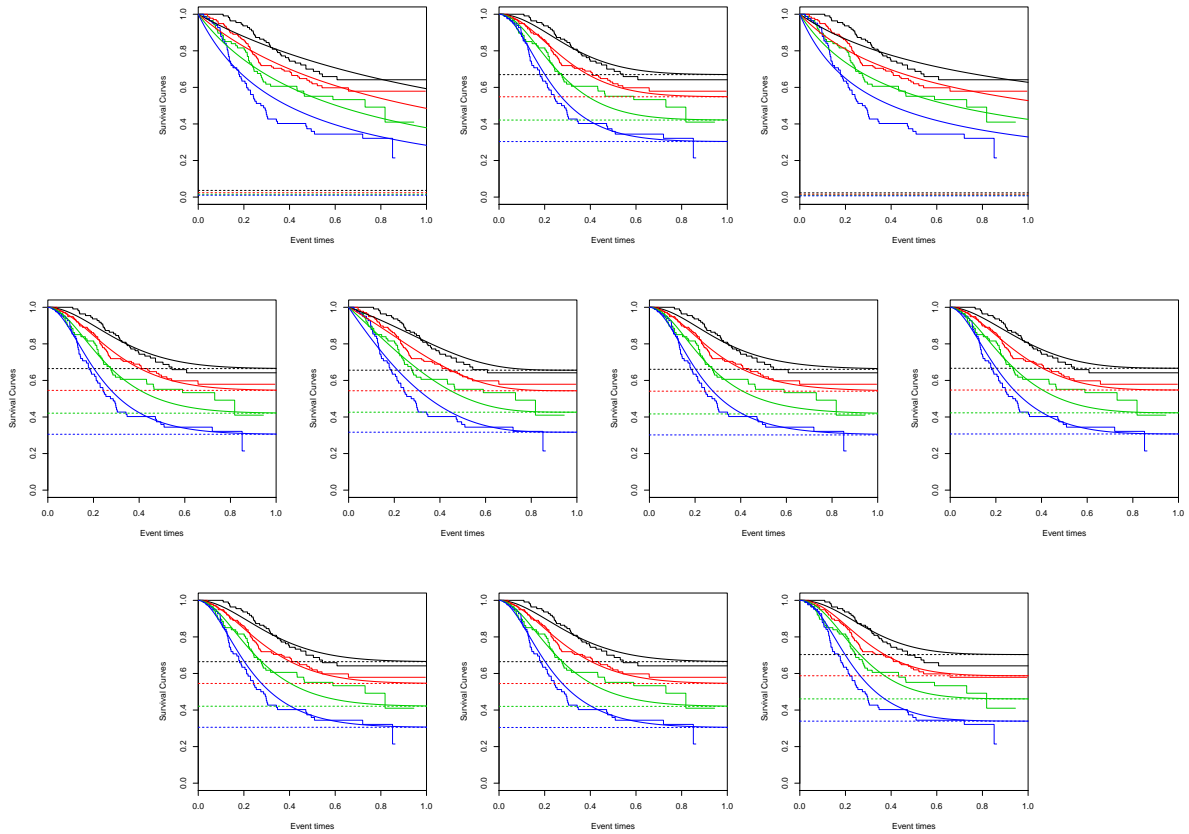


Figure 6.9: From the left to the right, top to bottom, the fitted regression models for the melanoma data set, in the same order as in Table 6.1. The colors black, red, green and blue represents the nodule categories 1, 2, 3 and 4, respectively.

6.4.4 Melanoma data

Here we consider the melanoma data. There are 417 observed times, of which 232 were censored (55.63 percent). This data set has covariate information. The covariate taken represents the nodule category ($n_1 = 82, n_2 = 87, n_3 = 137, n_4 = 111$). The Kaplan-Meier estimates suggest that the survival rate increases with the nodule category.

Table 6.5: MLEs for the fitted regression models and the AIC measure for the melanoma data set.

MO-Distribution	$\hat{\nu}$	\hat{a}	\hat{b}	\hat{c}	$\hat{\beta}_0$	$\hat{\beta}_1$	\hat{p}_1	\hat{p}_2	\hat{p}_3	\hat{p}_4	AIC
Exponential	-0.03	-	-	-	-2.84	-0.43	0.0365	0.0240	0.0156	0.0102	354.12
Rayleigh	-5.99	-	-	-	1.22	-0.51	0.6697	0.5485	0.4213	0.3036	306.38
Lomax	-0.02	-	-	-	-3.34	-0.41	0.0230	0.0154	0.0102	0.0068	363.34
Weibull	-5.16	1.89	-	-	1.19	-0.50	0.6647	0.5455	0.4209	0.3056	307.57
Gompertz	-0.85	3.74	-	-	1.12	-0.47	0.6560	0.5434	0.4263	0.3169	338.02
Burr XII	-5.82	1.96	-	-	1.17	-0.50	0.6609	0.5413	0.4167	0.3019	305.87
Chen	-4.26	1.79	-	-	1.19	-0.50	0.6665	0.5474	0.4227	0.3071	310.96
Modified Weibull	-5.16	1.89	0.00	-	1.18	-0.50	0.6645	0.5454	0.4209	0.3058	309.58
Weibull Extension	-31.25	7.66	1.89	-	1.19	-0.50	0.6645	0.5450	0.4201	0.3047	309.63
Traditional Weibull	-73.10	0.10	0.98	1.11	1.37	-0.51	0.7034	0.5875	0.4610	0.3393	335.90

The fitted results are summarized in Table 6.5 and Figure 6.9. The estimated cure rates

\hat{p}_1 , \hat{p}_2 , \hat{p}_3 and \hat{p}_4 for groups 1, 2, 3 and 4, respectively, are calculated by (6.10). The Marshall Olkin-Lomax and Marshall-Olkin Exponential distribution gives cure rates very close to zero and they have the worst AIC. Better AIC values are given by the Marshall Olkin-Rayleigh, Marshall Olkin-Weibull, Marshall Olkin-Burr XII, Marshall Olkin-Chen, Marshall Olkin-Modified Weibull and Marshall Olkin-Modified Weibull Extension distributions. The lowest AIC found was in the Marshall-Olkin Chen distribution, with 305.87. The distributions giving the best AIC values capture the Kaplan-Meier curve relatively well, but not so well for nodule category 1 and nodule category 3 near the tails.

The estimates of β_0 and β_1 are in agreement in all models. For β_0 , the value lies around 1.20 by most of models (except for Marshall Olkin-Lomax and Marshall-Olkin Exponential) and for β_1 , the value is around -0.50. That means that the cure rate decreases when the nodule category increases.

The estimated cure rates for the nodule category 1 is around 0.66. In the nodule category 2 is around 0.54. In the nodule category 3 is 0.42. In the nodule category 4 is 0.30. The standard deviation of these cure rates can be estimated using the standard deviation of β_0 e β_1 by the delta method. In the Marshall-Olkin Chen model, we have the standard deviation of p_1 , p_2 , p_3 e p_4 given by 0.0379, 0.0305, 0.0319 and 0.0395, respectively. Taking the asymptotic 95% confidence region, those values leads to the intervals (0.59, 0.74), (0.48,0.60), (0.36,0.48) and (0.23,0.38), respectively. This indicates a significant difference between nodules categories 1 and 3, 1 and 4 and 2 and 4. Similar results can be found in the other models that performed well. This results agrees with the results founded in [Rodrigues *et al.* \(2009b\)](#), [Balakrishnan & Pal \(2013a\)](#) e [Balakrishnan & Pal \(2013b\)](#).

6.4.5 Discussion

Here, we discuss some of the results in Sections 6.4.1, 6.4.2, 6.4.3, 6.4.4, a non-zero cure rate testing approach and compare the fitted distributions with their respective mixture model versions.

Table 6.6 compares the results in Tables 6.3, 6.4, 6.5 to the standard mixture model given by $S_{\text{mix}} = p + (1 - p)S(t)$, where $S(t)$ is the same baseline distribution as in the Marshall Olkin defective distributions. The distributions were compared in terms of the AIC and have the same number of parameters. The bold numbers represent the smaller AIC value. In all data sets, the defective approach performs better in seven out the ten cases. The baseline distributions performing better under a chosen approach are the same, regardless of the data analysed. The following distributions performed better under the defective approach for each of the three data sets: the Marshall Olkin-Rayleigh, Marshall Olkin-Weibull, Marshall Olkin-Burr XII, Marshall Olkin-Chen, Marshall Olkin-Modified Weibull, Marshall Olkin-Weibull extension and Marshall Olkin-Traditional distributions. The remaining performed better under the standard mixture approach. We can conclude that the defective distributions are good competitors for modelling cure rates. They provide better fits more often than the mixture model.

Table 6.7 gives 95 percent asymptotic confidence intervals for r based on the normal approximation. We check this table to see r is significantly lower than zero. Since the cure rate p only depends on r , the cure rate is significantly greater than zero, implying the existence of cure fraction, if r is significantly lower than zero. Almost all of the confidence intervals in Table 6.7 are in the negative side of the real line. The only exception is that for the Marshall Olkin-Lomax distribution fitted to the leukemia and divorce data sets. Even this confidence interval is almost all negative. We can conclude therefore that the leukemia, colon and divorce data sets have non-zero cure rates.

Table 6.6: Comparison of the AIC value of the mixture and defective models.

Baseline distribution	Leukemia data		Colon data		Divorce data	
	Mixture	Defective	Mixture	Defective	Mixture	Defective
Exponential	-64.41	-54.86	1509.62	1531.10	1503.66	1532.16
Rayleigh	-55.71	-75.39	1879.82	1668.31	1770.51	1633.88
Lomax	-63.77	-54.03	1518.33	1537.00	1518.61	1538.82
Weibull	-68.54	-73.50	1481.29	1462.36	1439.79	1435.90
Gompertz	-62.20	-58.31	1512.46	1516.68	1469.76	1471.46
Burr XII	-69.58	-73.93	1470.14	1456.53	1439.09	1437.27
Chen	-67.18	-73.00	1503.11	1474.74	1442.66	1435.01
Modified Weibull	-63.19	-71.50	1464.69	1464.39	1441.13	1437.73
Weibull extension	-66.50	-71.36	1483.39	1464.42	1456.22	1437.01
Traditional Weibull	-64.54	-69.09	1485.29	1466.37	1443.79	1439.02

All of the examples provided here show that the newly introduced defective distributions can be used to provide adequate fits to several different kinds of data sets. The Marshall

Table 6.7: Asymptotic 95 percent confidence intervals for r .

Marshall Olkin distribution	Leukemia data		Colon data		Divorce data	
	Lower CI	Upper CI	Lower CI	Upper CI	Lower CI	Upper CI
Exponential	-0.5486	-0.0109	-0.7249	-0.4493	-0.9083	-0.499
Rayleigh	-0.3302	-0.0562	-0.9467	-0.7844	-1.3366	-1.1199
Lomax	-0.5359	0.0767	-0.4271	-0.0294	-0.5376	0.0167
Weibull	-0.3702	-0.0426	-0.9776	-0.7834	-1.3542	-1.0889
Gompertz	-0.4619	-0.0495	-1.0096	-0.8107	-1.3825	-1.1419
Burr XII	-0.3767	-0.0440	-0.9445	-0.7318	-1.3252	-1.0385
Chen	-0.3620	-0.0409	-1.0032	-0.8197	-1.3752	-1.1244
Modified Weibull	-0.3702	-0.0426	-0.978	-0.7838	-1.3744	-1.0956
Weibull extension	-0.3679	-0.0397	-0.9777	-0.7833	-1.3797	-1.1197
Traditional Weibull	-0.3652	-0.0418	-0.9776	-0.7834	-1.3796	-1.1203

Olkin-Rayleigh distribution gives the best fit for the leukemia data set, but it does not perform so well for the colon data set. The Marshall Olkin-Burr XII distribution gives the best fit for the colon data set, while the Marshall Olkin-Chen distribution gives the most adequate fit for the divorce data set and melanoma data set, as a regression model. This shows how competitive the newly proposed distributions can be, even when competing with the standard mixture models. More investigations are needed for these new distributions, but we hope we have provided strong evidence of the competitiveness of the proposed distributions.

6.5 Conclusions

The theory on defective distributions has been quite limited. In this chapter, we have derived a new property of the Marshall Olkin family of distributions, allowing one to generate many new defective distributions as possible models for a wide variety of data sets. We have constructed ten new defective distributions based on the new property. The usual asymptotes of the maximum likelihood estimators for these distributions have been checked by simulation. An approach to include covariate information has been proposed and illustrated in one of the applications. In total, applications to four real data sets have been illustrated. We have presented sufficient evidence of the relevance and competitiveness of the proposed distributions, covering a range of different scenarios and showing that they can provide adequate fits. We have also shown that the proposed distributions can perform better than the standard mixture models.

Chapter 7

Final Remarks

7.1 Conclusions

In this thesis, we worked in a way to explore the distributions that can be used to model survival data with presence of a cured rate. At the beginning of this thesis, the only two defective distributions known was the Gompertz and inverse Gaussian. There was only a few references about it with no further efforts to increase the numbers of defective distributions. Now, at the end of this thesis, we have developed two different ways to generate defective distribution, as stated in Chapters 5 and 6.

We started in Chapter 2 defining the Gompertz and inverse Gaussian defective distribution. In none of the previous works were evidenced or proved the suitability of maximum likelihood estimators for defective models. We presented, for the first time, a simulation work where it is possible to check the validity of the maximum likelihood estimators and analyze its needs regarding to sample sizes. We also pointed out how limited the two defective distributions are to fit several kinds of data.

With that in mind, we proved in Chapter 3 that if a baseline distribution is defective, then the respective distribution extended by the Marshall-Olkin family is also defective. This was the starting point in the generation of the new defective distributions. The Marshall-Olkin family increases the flexibility of a baseline distribution by adding a new

shape parameter to it. We checked the estimation procedure through some simulations and evidenced using real data sets how better the extended family is comparing to the baseline distributions. We finished showing that those two new three parameter defective distributions can be properly used to fit more data sets with cured fraction and its competitiveness in relation to the mixture models.

We went further and extended the basic defective distributions using the Kumaraswamy family, in Chapter 4. The Kumaraswamy family increases the flexibility of a baseline distribution by adding two new shape parameter to it. The new distributions are even more flexible than the ones presented in the previous chapter. We also propose a regression approach in order to incorporate covariates into the modeling. Therefore, two extra new four parameter defective distribution to model cure rate problems with a background for covariates modeling.

In Chapter 5, we generalized the results obtained in Chapters 3 and 4. We showed that any family of extended distributions provides defective ones if the baseline is also defective, since the extension is continuous in relation to the baseline distribution. There, we presented a full literature review regarding to families of distributions and chose another 8 to exemplify the results. We showed that the new distributions can improve the baseline ones in most cases, and in each scenario, there is a better distribution that fits the data. With the result of this chapter, we state the first method to generate new defective distributions.

In Chapter 6, we took another turn and propose a way that generates defective distribution based in a peculiar property of the Marshall-Olkin family. We show that, if a distribution have it survival function going to the infinity when some of its parameters domains is changed, then the Marshall-Olkin extension of this distribution can assume a defective form. This result leads to several different defective distributions, without the Gompertz or inverse Gaussian as the baseline. We used the extended Weibull to exemplify our result, taking 10 especial model as sub-cases. A regression approach is also proposed and exemplified.

Summarizing, this thesis is a work related to the cure rate modeling in survival studies,

using the defective distributions approach. At first, we had only two of these distribution to work with, now we have two different methods to generate defective distributions. We showed 20 new distributions in Chapter 5 and 11 in Chapter 6, but many more can be generated. We also showed that this variety of models can properly fit very different kinds of data sets and, often enough, outperform the standard mixture model.

This thesis is based in five papers developed in my doctoral period. Three of it are already published ([Rocha et al. \(2014\)](#), [Rocha et al. \(2015a\)](#) and [Rocha et al. \(2015c\)](#)), one is being reviewed [Rocha et al. \(2015b\)](#) and a last one is almost ready for submission.

7.2 Future Works

From this point, we see some paths to continue to develop the defective distributions theory. One of them is to work with these models in a Bayesian point a view. Changing the estimation method may lead to a better result regarding to the interval estimation os these models. Also, we can incorporate informative prioris for the first time using defective distributions. Will be interesting to compare with the work of [Balka et al. \(2011\)](#) and the ones proposed in here.

Another way to go is develop models with a frailty term, to estimate the influence of the non-observed covariates. In [Price & Manatunga \(2001\)](#) it is considered models with cured fraction, frailty term, and cured fraction with frailty term. They show that the frailty models are useful the model cure rate data. So, the idea here is to properly incorporate a frailty term into the defective models and check estimators properties, simulations scenarios and real data cases.

Since all the computation done here was in R. One of our future works is to create a new R package to provide functions to help with the modeling with defective distributions. These functions can generate random samples of defective distributions and calculate density, cumulative and quantiles values for a given defective distribution. These function may also cover some post-modeling issues, like probability calculating, plot of figures, analysis of the influent data points, among others functionalities.

This package will be essential to spread the methodology proposed in this work. As a means of dissemination, we intend to create an educational material with examples and guides for applications in R by following the functions listed in the developed package. The aim would be to write a book to publish in accordance with the new partnership between the Brazilian Statistical Association (ABE) and the international publisher Springer. The SpringerBriefs in Statistics is a series created especially for this partnership and aims to internationally publicize studies conducted by brazilian researchers and other Latin American countries, in various fields of statistics. Such series consists of short texts, around 120 pages, in which it is treated the latest topics in a particular area.

7.3 Acknowledgements

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil) for financial support during the course of this doctorate. All R codes used in this thesis are available with the author.

Bibliography

- Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, pages 534–545.
- Aalen, O., Borgan, O. & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer Verlag, New York.
- Abdul-Moniem, I. & Abdel-Hameed, H. (2012). On Exponentiated Lomax Distribution. *International Journal of Mathematical Archive*, **3**, 2144–2150.
- Adepoju, K., Chukwu, A. & Wang, M. (2014). The Beta Power Exponential Distribution. *Journal of Statistical Science and Application*, **2**, 37–46.
- Akinsete, A., Famoye, F. & Lee, C. (2008). The Beta Pareto Distribution. *Statistics*, **42**, 547–563.
- Akinsete, A., Famoye, F. & Lee, C. (2014). The Kumaraswamy-Geometric Distribution.
- Al-Saiari, A., Baharith, L. & Mousa, S. (2014). Marshall-Olkin Extended Burr Type XII Distribution. *International Journal of Statistics and Probability*, **3**.
- Alexander, C., Cordeiro, G., Ortega, E. & Sarabia, J. (2012). Generalized Beta-Generated Distributions. *Computational Statistics and Data Analysis*, **56**, 1880–1897.
- Alshawarbeh, E., Famoye, F. & Lee, C. (2014). Beta-Cauchy Distribution: Some Properties and Applications.
- Alzaatreh, A. & Knight, K. (2013). On the Gamma-Half Normal Distribution and Its Applications. *Journal of Modern Applied Statistical Methods*, **12**, 103–119.

- Alzaatreh, A., Famoye, F. & Lee, C. (2013a). Weibull-Pareto Distribution and Its Applications. *Communications in Statistics—Theory and Methods*, **42**, 1673–1691.
- Alzaatreh, A., Lee, C. & Famoye, F. (2013b). A New Method for Generating Families of Continuous Distributions. *Metron*, **71**, 63–79.
- Alzaatreh, A., Famoye, F. & Lee, C. (2014). The Gamma-Normal Distribution: Properties and Applications. *Computational Statistics and Data Analysis*, **69**, 67–80.
- Amini, M., MirMostafaei, S. & Ahmadi, J. (2014). Log-gamma-generated families of distributions. *Statistics*, **48**(4), 913–932.
- Aryal, G. & Elbata, I. (2015). Kumaraswamy Modified Inverse Weibull Distribution: Theory and Application. *Applied Mathematics and Information Sciences*, **9**, 651–660.
- Azzalini, A. (1985). A Class of Distributions which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**, 171–178.
- Badmus, N. & Bamiduro, T. (2014). Life Length of Components Estimates with Beta-Weighted Weibull Distribution. *Journal of Statistics: Advances in Theory and Applications*, **11**, 91–107.
- Balakrishnan, N. & Pal, S. (2012). Em algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice*, **6**(4), 698–724.
- Balakrishnan, N. & Pal, S. (2013a). Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on com-poisson family. *Computational Statistics & Data Analysis*, **67**, 41–67.
- Balakrishnan, N. & Pal, S. (2013b). Expectation maximization-based likelihood inference for flexible cure rate models with weibull lifetimes. *Statistical methods in medical research*. doi: 0962280213491641.
- Balakrishnan, N. & Pal, S. (2015). An em algorithm for the estimation of exible cure rate model parameters with generalized gamma lifetime and model discrimination using likelihood- and information-based methods. *Computational Statistics*, **30**, 151–189.

- Balka, J., Desmond, A. F. & McNicholas, P. D. (2009). Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime data analysis*, **15**(2), 147–176.
- Balka, J., Desmond, A. F. & McNicholas, P. D. (2011). Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, **38**(1), 127–144.
- Barreto-Souza, W., Santos, A. & Cordeiro, G. (2010). The Beta Generalized Exponential Distribution. *Journal of Statistical Computation and Simulation*, **80**, 159–172.
- Barreto-Souza, W., Cordeiro, G. & Simas, A. (2011). Some Results for Beta Frechet Distribution. *Communications in Statistics—Theory and Methods*, **40**, 798–811.
- Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.
- Bidram, H. (2012). The Beta Exponential-Geometric Distribution. *Communications in Statistics—Theory and Methods*, **41**, 1606–1622.
- Bidram, H., Behboodjan, J. & Towhidi, M. (2013). The Beta Weibull Geometric Distribution. *Journal of Statistical Computation and Simulation*, **83**, 52–67.
- 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.
- Bourguignon, M., Silva, R., Zea, L. & Cordeiro, G. (2013). The Kumaraswamy Pareto Distribution. *Journal of Statistical Theory and Applications*, **12**, 129–144.
- Bozdogan, H. (1987). Model selection and akaike’s information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, **52**(3), 345–370.
- Cakmakyapan, S. & Kadilar, G. (2014). A New Customer Lifetime Duration Distribution: The Kumaraswamy Lindley Distribution. *International Journal of Trade, Economics and Finance*, **5**.

- Calsavara, V. F. (2011). Modelos de sobrevivencia com fracao de cura usando um termo de fragilidade e tempo de vida weibull modificada generalizada. *Master Thesis*.
- 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.
- Castellares, F. & Lemonte, A. (2014). A New Generalized Weibull Distribution Generated by Gamma Random Variables. *Journal of the Egyptian Mathematical Society*.
- Castellares, F., Santos, M., Montenegro, L. & Cordeiro, G. (2015). A Gamma-Generated Logistic Distribution: Properties and Inference. *American Journal of Mathematical and Management Sciences*, **34**, 14–39.
- Chen, M.-H., Ibrahim, J. G. & Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.
- Cintra, R., Rego, L., Cordeiro, G. & Nascimento, A. (2014). Beta Generalized Normal Distribution with An Application for SAR Image Processing. *Statistics*, **48**, 279–294.
- Colosimo, E. A. & Giolo, S. R. (2006). *Analise de sobrevivencia aplicada*. Edgard Blucher.
- Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**(478).
- Cordeiro, G. & Brito, R. (2012). The Beta Power Distribution. *Brazilian Journal of Probability and Statistics*, **26**, 88–112.
- Cordeiro, G. & Lemonte, A. (2011a). The Beta Birnbaum-Saunders Distribution: An Improved Distribution for Fatigue Life Modeling. *Computational Statistics and Data Analysis*, **55**, 1445–1461.
- Cordeiro, G. & Lemonte, A. (2011b). The Beta-Half-Cauchy Distribution. *Journal of Probability and Statistics*, **2011**.

- Cordeiro, G. & Lemonte, A. (2011c). The Beta Laplace Distribution. *Statistics and Probability Letters*, **81**, 973–982.
- Cordeiro, G. & Lemonte, A. (2014). The Exponentiated Generalized Birnbaum-Saunders Distribution. *Applied Mathematics and Computation*, **247**, 762–779.
- Cordeiro, G., Ortega, E. & Nadarajah, S. (2010). The Kumaraswamy Weibull Distribution with Application to Failure Data. *Journal of the Franklin Institute*, **347**, 1399–1429.
- Cordeiro, G., Nadarajah, S. & Ortega, E. (2012a). The Kumaraswamy Gumbel Distribution. *Statistical Methods and Applications*, **21**, 139–168.
- Cordeiro, G., Nobre, J., Pescim, R. & Ortega, E. (2012b). The Beta Moyal: A Useful Skew Distribution. *Int. Journal Res. Rev. Appl. Sci.*, **10**, 171–192.
- Cordeiro, G., Ortega, E. & Silva, G. (2012c). The Beta Extended Weibull Family. *Journal of Probability and Statistical Science*, **10**, 15–40.
- Cordeiro, G., Pescim, R. & Ortega, E. (2012d). The Kumaraswamy Generalized Half-Normal Distribution for Skewed Positive Data. *Journal of Data Science*, **10**, 195–224.
- Cordeiro, G., Castellares, F., Montenegro, L. & de Castro, M. (2013a). The Beta Generalized Gamma Distribution. *Statistics*, **47**, 888–900.
- Cordeiro, G., Cristino, C., Hashimoto, E. & Ortega, E. (2013b). The Beta Generalized Rayleigh Distribution. *Statistical Papers*, **54**, 133–161.
- Cordeiro, G., Gomes, A., da Silva, C. & Ortega, E. (2013c). The Beta Exponentiated Weibull Distribution. *Journal of Statistical Computation and Simulation*, **83**, 114–138.
- Cordeiro, G., Ortega, E. & da Cunha, D. (2013d). The Exponentiated Generalized Class of Distributions. *Journal of Data Science*, **11**, 1–27.
- Cordeiro, G., Silva, G. & Ortega, E. (2013e). The Beta Weibull Geometric Distribution. *Statistics*, **47**, 817–834.

- Cordeiro, G., Ortega, E. & Popovic, B. (2014a). The Gamma-Linear Failure Rate Distribution: Theory and Applications. *Journal of Statistical Computation and Simulation*, **84**, 2408–2426.
- Cordeiro, G., Ortega, E. & Silva, G. (2014b). The Kumaraswamy Modified Weibull Distribution: Theory and Applications. *Journal of Statistical Computation and Simulation*, **84**, 1387–1411.
- Cordeiro, G., Ortega, E. & Popovic, B. (2015). The Gamma-Lomax Distribution. *Journal of Statistical Computation and Simulation*, **85**, 305–319.
- Cordeiro, G. M. & de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, **81**(7), 883–898.
- da Silva, R., de Andrade, T., Maciel, D., Campos, R. & Cordeiro, G. (2013). A New Lifetime Model: The Gamma Extended Frechet Distribution. *Journal of Statistical Theory and Applications*, **12**, 39–54.
- de Pascoa, M., Ortega, E. & Cordeiro, G. (2011). The Kumaraswamy Generalized Gamma Distribution with Application in Survival Analysis. *Statistical Methodology*, **8**, 411–433.
- de Santana, T., Ortega, E., Cordeiro, G. & Silva, G. (2012). The Kumaraswamy-Log-Logistic Distribution. *Journal of Statistical Theory and Applications*, **11**, 265–291.
- Domma, F. & Condino, F. (2013). The Beta-Dagum Distribution: Definition and Properties. *Communications in Statistics—Theory and Methods*, **42**, 4070–4090.
- El-Sherpieny, E. & Ahmed, M. (2014). On the Kumaraswamy Kumaraswamy Distribution. *International Journal of Basic and Applied Sciences*, **3**, 372–381.
- Elbatal, I. (2013a). The Kumaraswamy Exponentiated Pareto Distribution. *Economic Quality Control*, **28**, 1–8.
- Elbatal, I. (2013b). Kumaraswamy Generalized Linear Failure Rate Distribution. *Indian Journal of Computational and Applied Mathematics*, **1**, 61–78.

- Elbatal, I. & Elgarhy, M. (2013). Statistical Properties of Kumaraswamy Quasi Lindley Distribution. *International Journal of Mathematics Trends and Technology*, **4**, 237–246.
- Elbatal, I. & Muhammed, H. (2014). Exponentiated Generalized Inverse Weibull Distribution. *Applied Mathematical Sciences*, **8**, 3997–4012.
- Eugene, N., Lee, C. & Famoye, F. (2002). Beta-Normal Distribution and Its Applications. *Communications in Statistics—Theory and Methods*, **31**, 497–512.
- Feller, W. (1968). *An Introduction to Probability Theory, vol. I, vol. II*. John Wiley, New York.
- George, D. & George, S. (2013). Marshall-Olkin Esscher Transformed Laplace Distribution and Processes. *Brazilian Journal of Probability and Statistics*, **27**, 162–184.
- Ghitany, M. (2005). Marshall-olkin extended pareto distribution and its application. *International Journal of Applied Mathematics*, **18**(1), 17.
- Ghitany, M., Al-Hussaini, E. & Al-Jarallah, R. (2005). Marshall-olkin extended weibull distribution and its application to censored data. *Journal of applied Statistics*, **32**(10), 1025–1034.
- Ghitany, M., Al-Awadhi, F. & Alkhalfan, L. (2007). Marshall-Olkin Extended Lomax Distribution and Its Application to Censored Data. *Communications in Statistics—Theory and Methods*, **36**, 1855–1866.
- Ghosh, I. (2014). The Kumaraswamy Half-Cauchy Distribution: Properties and Applications. *Journal of Statistical Theory and Applications*, **13**, 122–134.
- Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J. & Pullen, J. (1998). Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, **17**(8), 831–839.
- Gomes, A., da Silva, C., Cordeiro, G. & Ortega, E. (2013). The Beta Burr III Model for Lifetime Data. *Brazilian Journal of Probability and Statistics*, **27**, 502–543.

- Gomes, A., da Silva, C., Cordeiro, G. & Ortega, E. (2014). A New Lifetime Model: The Kumaraswamy Generalized Rayleigh Distribution. *Journal of Statistical Computation and Simulation*, **84**, 290–309.
- Gui, W. (2013a). Marshall-Olkin Extended Log-Logistic Distribution and Its Application in Minification Processes. *Applied Mathematical Sciences*, **7**, 3947–3961.
- Gui, W. (2013b). A Marshall-Olkin Power Log-Normal Distribution and Its Applications to Survival Data. *International Journal of Statistics and Probability*, **2**.
- Gupta, R., Gupta, P. & Gupta, R. (1998). Modeling Failure Time Data by Lehman Alternatives. *Communications in Statistics—Theory and Methods*, **27**, 887–904.
- Gurvich, M., Dibenedetto, A. & Ranade, S. (1997). A new statistical distribution for characterizing the random strength of brittle materials. *Journal of Materials Science*, **32**(10), 2559–2564.
- Hady, A. & Ebraheim, N. (2014). Exponentiated Transmuted Weibull Distribution: A Generalization of the Weibull Distribution. *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, **8**.
- Hanook, S., Shahbaz, M., Mohsin, M. & Golam Kibria, B. (2013). A Note on Beta Inverse-Weibull Distribution. *Communications in Statistics—Theory and Methods*, **42**, 320–335.
- Haybittle, J. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *The British journal of radiology*, **32**(383), 725–733.
- Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, **57**(2), 383–388.
- Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2005). *Bayesian survival analysis*. Wiley Online Library.
- Idowu, B. & Ikegwu, E. (2013). The Beta Weighted Weibull Distribution: Some Properties and Application to Bladder Cancer Data. *Journal of Applied and Computational Mathematics*, **2**.

- Jafari, A. & Mahmoudi, E. (2014). Beta Linear Failure Rate Distribution and Its Applications. *arXiv preprint 1212.5615*.
- Jafari, A., Tahmasebi, S. & Alizadeh, M. (2014). The Beta Gompertz Distribution. *Revista Colombiana de Estadística*, **37**, 139–156.
- Jose, K. & Krishna, E. (2011a). Marshall-Olkin Extended Uniform Distribution. *ProbStat Forum*, **4**, 78–88.
- Jose, K. & Sebastian, R. (2013). Marshall-Olkin Morgenstern Weibull Distribution: Generalisations and Applications. *Economic Quality Control*, **28**, 105–116.
- Jose, K., Ancy, J. & Ristić, M. M. (2009a). A marshall-olkin beta distribution and its application. *Journal of Probability and Statistical Science*, **7**(2), 173–186.
- Jose, K., Joseph, A. & Ristic, M. (2009b). A Marshall-Olkin Beta Distribution and Its Applications. *Journal of Probability and Statistical Science*, **7**, 173–186.
- Jose, K., Naik, S. & Ristic, M. (2010). Marshall-Olkin q-Weibull Distribution and Max/Min Processes. *Statistical Papers*, **51**, 837–851.
- Jose, K. K. & Krishna, E. (2011b). Marshall-olkin extended uniform distribution. *Probability Statistics and Optimization*, **4**, 78–88.
- Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282), 457–481.
- Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B. *et al.* (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, **317**(8), 461–467.
- Klein, J. P. & Moeschberger, M. L. (2003). Survival analysis: Statistical methods for censored and truncated data. *Springer Verlag, New York*.

- Kong, L., Carl, L. & Sepanski, J. (2007). On the Properties of Beta Gamma Distribution. *Journal of Modern Applied Statistical Methods*, **6**.
- Kozubowski, T. & Nadarajah, S. (2008). The Beta Laplace Distribution. *Journal of Computational Analysis and Applications*, **10**, 305–318.
- Krishna, E. & Jose, K. (2011). Marshall-Olkin Generalized Asymmetric Laplace Distributions and Processes. *Statistica*, **71**, 453–467.
- Krishna, E., Jose, K., Alice, T. & Ristic, M. (2013). Marshall-Olkin Frechet Distribution. *Communications in Statistics—Theory and Methods*, **42**, 4091–4107.
- Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46**(1), 79–88.
- Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G. W., Gerstner, J. B., Krook, J. E. & Malliard, J. (1989). Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, **7**(10), 1447–1456.
- Leao, J., Saulo, H., Bourguignon, M., Cintra, R., Rego, L. & Cordeiro, G. (2013). On some properties of the beta inverse rayleigh distribution. *Chilean Journal of Statistics*, **4**, 111–131.
- Lee, M.-L. T. & Whitmore, G. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, pages 501–513.
- Lemonte, A. (2013). A New Extension of the Birnbaum-Saunders Distribution. *Brazilian Journal of Probability and Statistics*, **27**, 133–149.
- Lemonte, A., Barreto-Souza, W. & Cordeiro, G. (2013). The Exponentiated Kumaraswamy Distribution and Its Log-Transform. *Brazilian Journal of Probability and Statistics*, **27**, 31–53.

- Lillard, L. A. & Panis, C. W. (2000). aml multilevel multiprocess statistical software, release 1.0. *Los Angeles: EconWare*.
- Liu, D. C. & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, **45**(1-3), 503–528.
- Lourenzutti, R., Duarte, D. & Azevedo, M. (2014). The Beta Truncated Pareto Distribution.
- Mahmoudi, E. (2011). The Beta Generalized Pareto Distribution with Application to Lifetime Data. *Mathematics and Computers in Simulation*, **81**, 2414–2430.
- Maller, R. A. & Zhou, X. (1996). *Survival analysis with long-term survivors*. Wiley New York.
- Mameli, V. & Musio, M. (2013). A Generalization of the Skew-Normal Distribution: The Beta Skew-Normal. *Communications in Statistics—Theory and Methods*, **42**, 2229–2244.
- 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.
- Marshall, A. W. & Olkin, I. (2015). A bivariate gompertz–makeham life distribution. *Journal of Multivariate Analysis*, **139**, 219–226.
- Martinez, E. Z., Achcar, J. A., Jácome, A. A. & Santos, J. S. (2013). 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**(3), 343–355.
- Merovci, F. & Sharma, V. (2014). The Beta Lindley Distribution: Properties and Applications. *Journal of Applied Mathematics*, **2014**.
- Montenegro, L. & Cordeiro, G. (2013). The Beta Lognormal Distribution. *Journal of Statistical Computation and Simulation*, **83**, 203–228.

- Morais, A., Cordeiro, G. & Cysneiros, A. (2013). The Beta Generalized Logistic Distribution. *Brazilian Journal of Probability and Statistics*, **27**, 185–200.
- Nadarajah, S. (2006). The Exponentiated Gumbel Distribution with Climate Application. *Environmetrics*, **17**, 13–23.
- Nadarajah, S. & Eljabri, S. (2013). The Kumaraswamy GP Distribution. *Journal of Data Science*, **11**, 739–766.
- Nadarajah, S. & Gupta, A. (2007). The Exponentiated Gamma Distribution with Application to Drought Data. *Calcutta Statistical Association Bulletin*, **59**, 29–54.
- Nadarajah, S. & Kotz, S. (2003). The Exponentiated Frechet Distribution. *InterStat*.
- Nadarajah, S. & Kotz, S. (2004). The Beta Gumbel Distribution. *Mathematical Problems in Engineering*, **2004**, 323–332.
- Nadarajah, S. & Kotz, S. (2006). The Beta Exponential Distribution. *Reliability Engineering and System Safety*, **91**, 689–697.
- Nadarajah, S., Cancho, V. & Ortega, E. (2013a). The Geometric Exponential Poisson Distribution. *Statistical Methods and Applications*, **22**, 355–380.
- Nadarajah, S., Nassiri, V. & Mohammadpour, A. (2013b). Truncated-Exponential Skew-Symmetric Distributions. *Statistics*. to appear.
- Nadarajah, S., Teimouri, M. & Shih, S. (2013c). Modified Beta Distributions. *Sankhyā B*. to appear.
- Nassar, M. & Nada, N. (2011). The Beta Generalized Pareto Distribution. *Journal of Statistics: Advances in Theory and Applications*, **6**, 1–17.
- Nassar, M. & Nada, N. (2012). A New Generalization of the Exponential-Geometric Distribution. *Journal of Statistics: Advances in Theory and Applications*, **7**, 25–48.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**(4), 945–966.

- Nieto-Barajas, L. E. & Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian Journal of Statistics*, **35**(3), 540–556.
- Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, **46**(1), 27–29.
- Oguntunde, P., Babatunde, O. & Ogunmola, A. (2014). Theoretical Analysis of the Kumaraswamy-Inverse Exponential Distribution. *International Journal of Statistics and Applications*, **4**, 113–116.
- Oluyede, B. & Yang, T. (2014). A New Class of Generalized Lindley Distributions with Applications. *Journal of Statistical Computation and Simulation*.
- Oluyede, B., Huang, S. & Pararai, M. (2014). A New Class of Generalized Dagum Distribution with Applications to Income and Lifetime Data. *Journal of Statistical and Econometric Methods*, **3**, 125–151.
- Pal, M. & Tiensuwan, M. (2014). The Beta Transmuted Weibull Distribution. *Austrian Journal of Statistics*, **43**, 133–149.
- Parainaba, P., Ortega, E., Cordeiro, G. & Pescim, R. (2011). The Beta Burr XII Distribution with Application to Lifetime Data. *Computational Statistics and Data Analysis*, **55**, 1118–1136.
- Paranaiba, P., Ortega, E., Cordeiro, G. & de Pascoa, M. (2013). The Kumaraswamy Burr XII Distribution: Theory and Practice. *Journal of Statistical Computation and Simulation*, **83**, 2117–2143.
- Pararai, M., Warahena-Liyanage, G. & Oluyede, B. (2014). A New Class of Generalized Inverse Weibull Distribution with Applications. *Journal of Applied Mathematics and Bioinformatics*, **4**, 17–35.
- Peng, Y. & Xu, J. (2012). An extended cure model and model selection. *Lifetime data analysis*, **18**(2), 215–233.
- Percontini, A., Blas, B. & Cordeiro, G. (2013). The Beta Weibull Poisson Distribution. *Chilean Journal of Statistics*, **4**, 3–26.

- Perez-Casany, M. & Casellas, A. (2014). Marshall-Olkin Extended Zipf Distribution.
- Price, D. L. & Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in medicine*, **20**(9-10), 1515–1527.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rajab, M., Aleem, M., Nawaz, T. & Daniya, M. (2013). On Five Parameter Beta Lomax Distribution. *Journal of Statistics*, **20**, 102–118.
- Ramires, T., Ortega, E., Cordeiro, G. & Hamedani, G. (2013). The Beta Generalized Half-Normal Geometric Distribution. *Studia Scientiarum Mathematicarum Hungarica*, **50**, 523–554.
- Ramos, M., Cordeiro, G., Marinho, P., Dias, C. & Hamedani, G. (2013). The Zografos-Balakrishnan Log-Logistic Distribution: Properties and Applications. *Journal of Statistical Theory and Applications*, **12**, 225–244.
- Ristić, M. & Balakrishnan, N. (2012). The Gamma Exponentiated Exponential Distribution. *Journal of Statistical Computation and Simulation*, **82**, 1191–1206.
- Ristić, M. & Nadarajah, S. (2013). A New Lifetime Distribution. *Journal of Statistical Computation and Simulation*. doi: 10.1080/00949655.2012.697163.
- Ristic, M. M., Jose, K. & Ancy, J. (2007). A marshall-olkin gamma distribution and minification process. *Stress Anxiety Res Soc*, **11**, 107–117.
- Rocha, R., Tomazella, V. & Louzada, F. (2014). Inferencia classica e bayesiana para o modelo de fracao de cura gompertz defeituoso. *Revista Brasileira de Biometria*, **32**(1), 104–114.
- Rocha, R., Nadarajah, S., Tomazella, V. & Louzada, F. (2015a). Two new defective distributions based on the marshall-olkin extension. *Lifetime data analysis*, pages 1–25.

- Rocha, R., Nadarajah, S., Tomazella, V. & Louzada, F. (2015b). A new class of defective models based on the marshall olkin family of distributions. *Submitted paper to Computation Statistics and Data Analysis*.
- Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F. & Eudes, A. (2015c). New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical methods in medical research*. doi: 0962280215587976.
- Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009a). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**(6), 753–759.
- Rodrigues, J., de Castro, M., Cancho, V. G. & Balakrishnan, N. (2009b). Com–poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**(10), 3605–3611.
- Roges, D., de Gusmao, F. & Diniz, C. (2014). The Kumaraswamy Inverse Rayleigh Distribution.
- Salem, H. (2014). The Exponentiated Lomax Distribution: Different Estimation Methods. *American Journal of Applied Mathematics and Statistics*, **2**, 364–368.
- Salem, H. & Selim, M. (2014). The Generalized Weibull-Exponential Distribution: Properties and Applications. *International Journal of Statistics and Applications*, **4**, 102–112.
- Sandhya, E. & Prasanth, C. (2014). Marshall-Olkin Discrete Uniform Distribution. *Journal of Probability*, **2014**.
- Santos-Neto, M., Bourguignon, M., Zea, L. M., Nascimento, A. D. & Cordeiro, G. M. (2014). The marshall-olkin extended weibull family of distributions. *Journal of Statistical Distributions and Applications*, **1**(1), 9.
- Saulo, H., Leao, J. & Bourguignon, M. (2012). The Kumaraswamy Birnbaum-Saunders Distribution. *Journal of Statistical Theory and Practice*, **6**, 745–759.
- Schrödinger, E. (1915). Zur theorie der fall-und steigversuche an teilchen mit brownscher bewegung. *Physikalische Zeitschrift*, **16**, 289–295.

- Shahbaz, M., Shahbaz, S. & Butt, N. (2012). The Kumaraswamy Inverse Weibull Distribution. *Pakistan Journal of Statistics and Operation Research*, **8**, 479–489.
- Shams, T. (2013a). The Kumaraswamy Generalized Exponentiated Pareto Distribution. *International Journal of Statistics and Applications*, **5**, 92–99.
- Shams, T. (2013b). The Kumaraswamy Generalized Lomax Distribution. *Middle-East Journal of Scientific Research*, **17**, 641–646.
- Shawky, A. & Abu-Zinadah, H. (2009). Exponentiated Pareto Distribution: Different Methods of Estimations. *International Journal of Contemporary Mathematical Sciences*, **4**, 677–693.
- Shittu, O. & Adepoju, K. (2013). On the Beta-Nakagami Distribution. *Progress in Applied Mathematics*, **5**, 49–58.
- Silva, G., Ortega, E. & Cordeiro, G. (2010). The Beta Modified Weibull Distribution. *Lifetime Data Analysis*, **16**, 409–430.
- Singla, N., Jain, K. & Sharma, S. (2012). The Beta Generalized Weibull Distribution: Properties and Applications. *Reliability Engineering and System Safety*, **102**, 5–15.
- Sy, J. P. & Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. *Biometrics*, **56**(1), 227–236.
- Torabi, H. & Montazeri, N. (2012). The Gamma-Uniform Distribution and Its Applications. *Kybernetika*, **48**, 16–30.
- Tsodikov, A., Ibrahim, J. & Yakovlev, A. (2003). Estimating cure rates from survival data. *Journal of the American Statistical Association*, **98**(464).
- Tweedie, M. (1945). Inverse statistical variates. *Nature*, **155**(3937), 453–453.
- Whitmore, G. A. (1979). An inverse gaussian model for labour turnover. *Journal of the Royal Statistical Society. Series A (General)*, pages 468–478.
- Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic models of tumor latency and their biostatistical applications*, volume 1. World Scientific.

- Yin, G. & Ibrahim, J. G. (2005). Cure rate models: a unified approach. *Canadian Journal of Statistics*, **33**(4), 559–570.
- Zakerzadeh, H. & Mahmoudi, E. (2012). A New Two Parameter Lifetime Distribution: Model and Properties. *arXiv preprint 1204.4248*.
- Zea, L., Silva, R., Bourguignon, M., Santos, A. & Cordeiro, G. (2012). The Beta Exponentiated Pareto Distribution with Application to Bladder Cancer Susceptibility. *International Journal of Statistics and Probability*, **1**.
- Zografos, K. & Balakrishnan, N. (2009). On Families of Beta- and Generalized Gamma-Generated Distributions and Associated Inference. *Statistical Methodology*, **6**, 344–362.