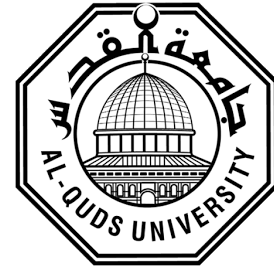


Deanship of Graduate Studies

Al-Quds University



Bayesian and Non-Bayesian Inference for Survival Data

Using Generalized Exponential Distribution:

A Comparison Study

Firyal Mohammad Khalaf Saraheen

M.Sc. Thesis

Jerusalem- Palestine

1437/2016

Bayesian and Non-Bayesian Inference for Survival Data

Using Generalized Exponential Distribution:

A Comparison Study

Prepared By

Firyal Mohammad Khalaf Saraheen

B. Sc. Mathematics, Bethlehem University

Palestine

Supervisor: Dr. Khalid Salah

A thesis Submitted in Partial Fulfillment of the

requirements for the Degree of Master of Mathematics at

Al-Quds University

1437/2016

Al-Quds University
Deanship of Graduate Studies
Graduate Studies / Mathematics



Thesis Approval

Bayesian and Non-Bayesian Inference for Survival Data Using Generalized Exponential Distribution: A Comparison Study


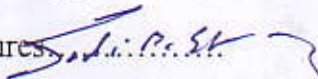

Prepared By : Firyal Mohammad Khalaf Saraheen

Registration No : 21310121

Supervisor: Dr. Khalid Ali Salah

Master thesis submitted and accepted, Date : 9/01/2016.

The names and signatures of the examining committee members are as follows

- | | | |
|-------------------------|--------------------|---|
| 1. Dr. Khalid Ali Salah | Head of Committee, | Signatures.....  |
| 2. Dr. Sa'di Alkronz | Internal Examiner, | Signatures.....  |
| 3. Dr. Bader Aljawadi | External Examiner, | Signatures.....  |

Jerusalem-Palestine

1437/2016

Dedication

I dedicate this thesis to my parents and brothers who led me through this darkness with their light of hope and support.

To my friends who touched my life with their love and passion, especially my dearest friend Bushra, who stands with me when things look bleak.

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study has not been submitted for a higher degree to any other university or institution.

Signature : 

Student's name : Firyal Mohammad Khalaf Saraheen

Date : 9/1/2016

Acknowledgement

I am thankful to all teachers in mathematics department who helped me to know that as lifelong learners of science, I can live more productive, responsible and fulgent. While special thanks go to Dr. Khalid Salah.

Lastly, I offer my regards to all of those who supported me during the completion of this thesis, especially Miss Ibtisam.

Table of Contents

Contents	Page
Declaration	i
Acknowledgement	ii
Table of Contents	iii
List of Tables	iv
List of Figures	vi
List of Abbreviations	vii
Abstract	viii
المخلص	أ
Chapter One: Introduction	1
1.1 Survival data	1
1.2 Survival function	1
1.3 Hazard function	3
1.3.1 Relationship between survival function and hazard function	3
1.4 Censoring in survival data (right- left- interval)	4
1.5 Estimation of survival function	6
1.5.1 Parametric approach	6
1.5.2 Non-parametric approach	8
1.6 Problem statement	12
1.7 Literature review	12
1.8 Objectives	13
1.9 Thesis structure	13
Chapter Two: Non-Bayesian Estimation	14
2.1 Maximum likelihood estimation	14
2.2 Fisher information matrix	18
Chapter Three: Bayesian Estimation	21
3.1 Bayes' theorem	21
3.1.1 From likelihood to Bayesian analysis	22
3.1.2 Marginal posterior distribution	23
3.1.3 Summarizing the posterior distribution	23
3.1.4 The choice of a prior	24
3.2 Markov processes	25
3.2.1 Bayesian-MCMC and Gibbs sampling	27
3.2.2 The Gibbs sampling	34
3.3 Bayesian estimation	36
3.3.1 Lindley Approximation	37
3.3.2 General Entropy Loss Function	39
Chapter Four: Simulation and Real case study	40
4.1 Simulation Study	40
4.2 Real Data Analysis	53
4.3 Conclusion	57
References	62

List of Tables

Table	Title	Page
1.1	Construction of the Kaplan-Meier estimator	9
1.2	Kaplan-Meier survival estimates	10
4.1	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 0.75$	42
4.2	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 1.5$	42
4.3	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 2.5$	43
4.4	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 0.75$	43
4.5	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 1.5$	44
4.6	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 2.5$	44
4.7	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 0.75$ and $\theta = 0.5$	46
4.8	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 1.5$ and $\theta = 0.5$	46
4.9	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 2.5$ and $\theta = 0.5$	47
4.10	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 0.75$ and $\theta = 0.5$	47
4.11	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 1.5$ and $\theta = 0.5$	48
4.12	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 2.5$ and $\theta = 0.5$	48

4.13	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 0.75$ and $\theta = 10$	49
4.14	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 1.5$ and $\theta = 10$	50
4.15	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of $(\hat{\theta})$ with $p = 2.5$ and $\theta = 10$	50
4.16	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 0.75$ and $\theta = 10$	51
4.17	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 1.5$ and $\theta = 10$	51
4.18	The average values of MLE, BSE, and BGE along with average mean squared errors, absolute biases and 95% C. I. of (\hat{p}) with $p = 2.5$ and $\theta = 10$	52
4.19	Average parameters estimates and their corresponding standard error	55
4.20	Mean Square Errors (MSE) and (AIC) of the Survival Function	56

List of Figures

Figure	Title	Page
1.1	The survival function	2
1.2	Right-censoring example	5
1.3	Left-censoring example	5
1.4	Interval-censoring example	6
1.5	Exponential Survival Function	8
1.6	Kaplan-Meier survival function for right-censored data	10
1.7	Turnbull survival function for interval-censored data	11
4.1	KM Survival Curves for the Three Data Sets	55
4.2	KM and Survival Curves for the Small Data Under (MLE), (BSE), and (BGE)	59
4.3	KM and Survival Curves for the Moderate Data Under (MLE), (BSE), and (BGE)	60
4.4	KM and Survival Curves for the Large Data Under (MLE), (BSE), and (BGE)	61

List of Abbreviations

Abbreviations	Meaning
K-M	Kaplan-Meier
MLE	Maximum Likelihood Estimation
NPMLE	Nonparametric Maximum Likelihood Estimator
EM	Expectation-Maximization
GE	Generalized Exponential
MPLE	Maximum Partial Likelihood Estimate
LR	likelihood-ratio
CI	Confidence Interval
MCMC	Markov Chain Monte Carlo
M-H	Metropolis-Hastings
Log	Logarithm
BSE	Bayes estimate using Lindley's approximation
BGE	Bayes estimates under the general entropy loss functions
MSE	Mean Squared Errors
DPC	Diagnosis Procedure Combination
QIP	Quality Indicator/Improvement Project
NPML	Non-Parametric Maximum Likelihood
AIC	Akaike's Information Criterion

Abstract

In this thesis we consider the Bayesian and non-Bayesian estimation of the unknown parameters of the Generalized Exponential (GE) distribution. Our aim is to compare the estimates of parameters and to observe the performance of the methods used for estimation.

By the developed methodology for MLE and Bayesian estimation has been demonstrated on a real data set when both the shape (p) and scale (θ) parameters of the GE distribution are unknown under informative set of independent priors. It is observed that the parameter estimates under the classical maximum likelihood method could not be obtained in close form; we therefore employed Newton- Raphson iterative approach via the Hessian matrix.

In this study following *C. Guure and S. Bosomprah (2013)*, we consider the Bayesian estimation of the unknown parameters of the GE distribution. We have also assumed a gamma prior on both parameters, and we provide the Bayesian estimators under the assumptions of squared error and general entropy loss functions. We see that the Bayesian estimators cannot be obtained in explicit forms, due to the complex nature of the posterior distribution of which Bayesian inference is drawn. Therefore, Lindley's numerical approximations procedure is used.

Results show that the Bayesian estimator under general entropy loss function performed quiet better than Bayesian under squared error loss function and that of maximum likelihood estimator for estimating the scale parameter with both MSE and absolute bias.

Chapter One

Introduction

1.1 Survival Data

Survival analysis is a branch of statistics which includes a variety of "statistical methods designed to describe, explain or predict the occurrence of events". It is widely applied in many fields such as biology, medicine, public health, and epidemiology. In survival analysis, our objective is to model the survival time, i.e. the time to the occurrence of a given event. The event could be just about anything. Within the medical field, common examples are the time to development of a disease, response to a treatment, and of course death. The available data often include the survival time, patient characteristics (such as gender, age, and blood pressure), disease information, treatment information, examination data and much more. Often we attempt to predict the probability of survival, response, or mean lifetime given a set of observed variables and compare survival distributions.

1.2 Survival Function

For matters of simplicity we assume time T (where T is the random variable representing survival time) to be continuous. The distribution of survival times is described by three mathematically equivalent functions: survival , hazard and cumulative hazard functions . A very simple way to specify the probability distribution of continuous durations T is the distribution function

$$F(t) = P(T \leq t) \tag{1.1}$$

The distribution function of t represents the probability that a realization of the random variable T is less than a value t . Furthermore $f(t)$ is the density function corresponding to (1.1) and thus can be written as

$$f(t) = dF(t)/dt \quad (1.2)$$

An alternative specification of the probability distribution of duration and an important concept in survival analysis is the survivor function, $S(t)$, defined as

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_{-\infty}^t f(x)dx = \int_t^{\infty} f(x)dx \quad (1.3)$$

which is the probability that a realization of the random variable T is greater than or equals to t . Or in other words: the probability that the event has not yet occurred by time t . Theoretically, the survival curve $S(t)$ can be plotted graphically to represent the probability of an individual's survival at varying time points. As t ranges from 0 to ∞ all survival curves have the following properties:

- i. $S(t)$ is monotone
- ii. $S(t)$ is non-increasing
- iii. At time $t = 0$, $S(t) = 1$ (i.e. the probability of surviving past time 0 is 1)
- iv. At time $t = \infty$, $S(t) = 0$ (i.e. as time goes to infinity, the survival curve goes to 0)

(See Figure 1.1).

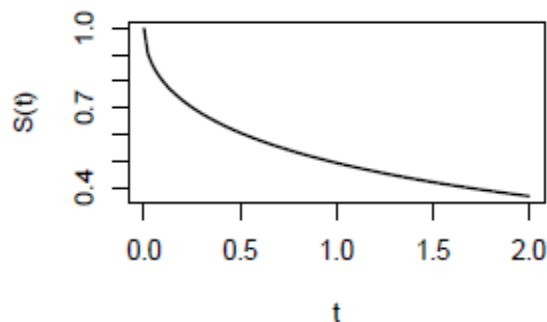


Figure 1.1: The survival function

1.3 Hazard Function

The hazard function $h(t)$ is the instantaneous rate at which events occur, given no previous events, defined as:

$$\begin{aligned}
 h(t) &= \lim_{dt \rightarrow 0} \frac{\Pr\{t \leq T < t + dt / T \geq t\}}{dt} \\
 &= \lim_{dt \rightarrow 0} \frac{\Pr\{t \leq T < t + dt\}}{dt \Pr(T \geq t)} \\
 &= \frac{1}{S(t)} \lim_{dt \rightarrow 0} \frac{F(t + dt) - F(t)}{dt} \\
 &= \frac{f(t)}{S(t)} \\
 &= -\frac{d}{dt} \log(S(t))
 \end{aligned} \tag{1.4}$$

from the definition; the hazard function is the ‘chance’ of failure (though it is a normalized probability, not a probability) at time t , given that the individual has survived until time t .

We see that the hazard function is similar to the density in the sense that it is a positive function. However it does not integrate to one. Indeed, it is not integrable.

The cumulative hazard function, $H(t)$, define as:

$$H(t) = \int_0^t h(u) du = -\log S(t) \tag{1.5}$$

1.3.1 Relationship between survival function and hazard function

From (1.3) and (1.4), we get the relationship

$$h(t) = \frac{f(t)}{S(t)} \tag{1.6}$$

Furthermore, since the density function is defined as the derivative of the cumulative distribution function, we get

$$f(t) = \frac{d}{dt} [1 - S(t)] = -S'(t) \tag{1.7}$$