IMPUTATION TECHNIQUES FOR RELIABILITY ANALYSIS BASED ON PARTLY INTERVAL CENSORED DATA

BY

ABDALLAH M T ZYOUD

A thesis submitted in fulfilment of the requirement for the degree of Master of Science (Mechanical Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

March 2017

ABSTRACT

In a conventional statistical analysis the term survival analysis or reliability analysis as it is known in engineering, has been used in a broad sense to describe collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. The time to failure of a particular experimental unit might be censored and this censored can be right, left, and interval (Partly Interval Censored (PIC)). In this thesis the analysis of this particular model was based on non-parametric, semi-parametric Cox model, and parametric accelerated failure time model via PIC data. In these models several imputation techniques are used that is; midpoint, left & right point, random, mean, median, and Multiple Imputations (MI). The maximum likelihood estimate was considered to obtain the estimated survival function. These estimates were then compared to the existing model such as Turnbull and Cox model based on clinical trial data (breast cancer data), for which it showed the validity of our models. In contrast, the data needed to be modified to PIC data for the purpose of the researcher's needs. Likewise, engineering failure rates data was also modified to represent PIC data and then simulation data was generated where the failure rates were taken based on engineering PIC data and was also used to further compare these three methods of estimation. From the simulation study for this particular case, we can conclude that the semi-parametric Cox model proved to be more superior in terms of estimating the survival function, likelihood ratio test and their P-value. In additional to that, based on imputation techniques, the MI, midpoint, random, mean and median showed better results with respect to estimate of the survival function. For the ultimate results, even though the semi-parametric model showed better output compared with the nonparametric and parametric models, all three models can easily be implemented based on engineering data set, medical data and simulation data.

خلاصة البحث

في تحليل البقاء على قيد الحياة أوتحل يل الإعتمادية كما هو معروف في مجال الهندسة، استخدم مصطلح تحليل البقاء على قيد الحيا ة بمعنى واسع لوصف مجموعة من الإجراءات الإحصائية لتحليل البيانات المعنية بوقت حصول حدث معين، إن البيانات التي جمعت من تجربة معينة قد لا تكون مكتملة حيث من الممكن ان تكون (right censored) أو (left censored) أو (interval censored) أو (partly interval censored (PIC)). في هذه الأطروحة، استندنا في تحليل البيانات على نماذج غير حدودية، وشبه حدودية (Cox) وحدودية (accelerated failure time model (AFT)). في هذه النماذج تستخدم أيضا عدة تقنيات لتعويض البيانات المحذوفة هي: استخدام نقطة المنتصف للفترة، او طرف الفترة الأيمن أو الأيسر، أو الوسط الحسابي للفترة أو الوسيط، أو نقطة عشوائية داخل الفترة، او استخدام التعويض المتعدد (MI). لقد اعتمدنا على تقدير احتمال الأقصى (MLE) للحصول على تقديرحدود دالة البقاء على قيد الحياة. هذه التقديرات تمت مقارنتها مع النماذج الحالية مثل تيرنبول وكوكس استنادا إلى بيانات التجارب الطبية (بيانات سرطان الثدي)، حيث أظهرت صحة نماذجنا. في المقابل، احتجنا لتعديل البيانات لتحويلها الى (PIC) لتلبية احتياجات البحث. وبالمثل، يتم استخدام معدلات الفشل في البيانات الهندسية وبيانات المحاكاة حيث كانت معدلات الفشل التي اتخذت على أساس الهندسة تمثل بيانات (PIC) كما تم استخدام هذه البيانات لعقد مقارنات إضافية وتحليل الطرق الثلاث التي استخدمناها لتقدير دالة البقاء على قيد الحياة. من دراسة البيانات و بالذات بيانات المحاكاة، يمكننا أن نستنتج أن النموذج شبه الحدودي لكوكس هو الأكثر تفوقا كما تدل قيمة (P-Value). أما بالنسبة لتقنيات تعويض البيانات الناقصة، فان استخدام التعويض المتعدد (MI) أو نقطة الوسط، أو المتوسط أو الوسيط،أو نقطة عشوائية أظهرت نتائج أفضل في مايتعلق تقدير دالة البقاء على قيد الحياة. على عكس استخدام الطرف الأيمن أو الأيسر كانت أقل فعالية بالغت في التقدير وقللت منه على التوالي. في نهاية المطاف، على الرغم من أن النموذجي شبه حدودي أظهر نتائج أفضل مقارنة مع غيره، إلا أنه يمكن القول أن جميع النماذج الثلاثة أظهرت نتائج مقبولة وأن تطبيقها سهل وغير مكلف.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Mechanical Engineering)

…………………………………..
Faiz Ahmed Mohamed Elfaki
Supervisor

…………………………………..
Meftah Hrairi
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Mechanical Engineering)

…………………………………..
Ari Legowo
Internal Examiner

…………………………………..
Noor Akma Ibrahim
External Examiner

This thesis was submitted to the Department of Mechanical Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Mechanical Engineering)

…………………………………..
Waqar Asrar
Head, Department of Mechanical Engineering

This thesis was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Mechanical Engineering)

…………………………………..
Erry Yulian Triblas Adesta
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Abdallah M. T. Zyoud

Signature ....   Date.......................................

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

THE IMPACT OF MOBILE INTERFACE DESIGN ON INFORMATION QUALITY OF M-GOVERNMENT SITES

I declare that the copyright holders of this dissertation are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Abdallah M. T. Zyoud

……..…………………….. ………………………..
Signature Date

ACKNOWLEDGEMENTS

Firstly, it is my utmost pleasure to dedicate this work to my dear parents and my family, who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

I wish to express my appreciation and thanks to those who provided their time, effort and support for this project. To the members of my dissertation committee, thank you for sticking with me.

Finally, a special thanks to Associate Professor Dr Faiz Elfaki who I can honestly say without his continuous support, encouragement and leadership, this work wouldn't have been successful and for that, I will be forever grateful. Special thanks are also extended to my co-supervisor Prof. Dr Meftah Hrairi for his continues support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFT	Accelerated Failure Time
HR	Hazard Ratio
PIC	Partly Interval Censored
MI	Multiple Imputations
MLE	Maximum Likelihood Estimate
PH	Proportional Hazard
LRT	Likelihood Ratio Test
NPMLE	Nonparametric Maximum Likelihood Estimate
AIC	Akaike's Information Criteria

LIST OF SYMBOLS

$\widetilde{X}$	Sample mean
s^2	Sample variance
s	Sample standard deviation
Z	Standard score
$S(t)$	Survival function
β	The regression coefficient
A_0	The cumulative baseline hazard
δ_i	Censoring indicator
N	Sample size

CHAPTER ONE

INTRODUCTION

CHAPTER OVERVIEW

We shall introduce here the background of the research. In addition, we shall describe major key words such as the survival analysis, Cox model, censoring and major types of censoring, imputation techniques. Also the formulation of the problem, the objective of the research, and the scope of the thesis shall be described.

1.1 SURVIVAL ANALYSIS

The term survival analysis has been used in a broad sense to describe collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.

In the past, applications of survival analysis used to focus on biomedical research, an event could have been death, recurrence of a disease, the development of a disease, cessation of smoking, and so forth. Recently the applications have been extended to other fields, such as, criminology, sociology, marketing, health insurance practice, business, economics and last but not least reliability engineering where the event could be the failure of electronic devices, components or systems.

The study of survival data has previously focused on predicting the probability of response, survival, or mean lifetime, and comparing the survival distributions of experimental animals or of human patients. In recent years, the identification of risk and/or prognostic factors related to response, survival, and the development of a disease has become equally important.

Survival models, like other statistical models, can also be considered as situational estimates to a more complex process, and may, therefore, give a less definite result. This can give rise to doubts about the models. A variation study on the results of the analysis with small modifications on the data is then necessary. Therefore, one important factor in statistical analysis is to conduct a study on result suitability. Residual value and Hessian matrix are useful components in detecting extreme points, but, they cannot be used to assess the effect on model suitability in general, and parameter estimate, in particular. In this research, we extend the techniques of studying result suitability of a survival model focusing on imputation techniques based on semiparametric Cox model and other models.

1.2 COX MODEL

The proportional hazards regression model of Cox (Cox, 1972), plays a very important role in the theory and practice of lifetime and duration data analysis. This is because the Cox regression model provides a convenient way to evaluate the influence of one or several covariates on the probability of conclusion of lifetime or duration spells.

In dealing with survival data without any knowledge about the underlying distribution, a semiparametric approach is most suitable to describe the relationship between several variables and the survival probability.

When incorporating explanatory variables, the most popular method is the Cox Proportional Hazard Model. The Cox proportional hazard model given by Cox (1972) is as follows:

$$\lambda(t,z) = \lambda_0(t)\exp(\beta_0 z) \tag{1.01}$$

here $\lambda_0(t)$ is an unknown baseline hazard function, z is a p-vector covariates and β_0 is a vector of regression coefficients.

1.3 CENSORING

Censoring occurs when the information of a failure time of some subjects is incomplete. There are different reasons for censoring which lead to different types of censored data and below are the main types of censoring.

1.3.1 Right Censored

Right censored data occurs when the last observation of a subject is not its failure yet whether it is because the survival study ended before the event of failure of some subjects occurs or because they left the study before it ends. It is the most common type of censored data and the one that received the most attention.

1.3.2 Left Censored

A subject is left censored if it's true survival time is less than the observed time. This happens when some subjects had already failed before the study started. A very common example of left censoring is when conducting Aids studies and some of the subjects test positive in the initial testing.

1.3.3 Interval Censored

While in the previous two types the event of interest occurred either before the beginning of the study or after it ended, in this type of censored data the event occurs within the time of the study but it is not exactly observed, it is only known to fall in an interval [A,B] for example.

Interval censored data arises in many areas such as demography, epidemiology, finance, medicine and engineering but its importance is not confined to that but also to its flexibility.

The left censored data can be treated as interval censored data where A is 0 and B is the first observed time while right censored data can be treated as interval censored data where A is the last observed time and B is infinity. There are many types of interval censoring data and here is a summary of the most common ones.

Case 1 Interval Censored

By case 1 interval censoring we mean that there is only one random observation time T that divides the study time into two intervals. So all we know is whether the event occurred before or after that observation time.

Case 2 Interval Censored

In case 2 interval censored data we have two observation times, T_1 and T_2, which divide the study period into three intervals $[0,T_1]$, $[T_1,T_2]$ and $[T_2, \infty)$. And generally case k interval censored data has exactly k observations.

Mixed Case Interval Censored

Mixed case interval censored data means that different objects in the study may have different number of observations. Each object is observed n times where n is an integer $n \in [1,k]$ instead of being exactly k in "case k interval censored data".

There are two main reasons why mixed case interval censoring appears; first, in many cases the nature of the experiment produces different number of observations, for example, it is common that in medical follow up studies different patients may have different number of observations (follow ups). Second, we may find out that the event occurred before the kth observation and in that case continuing until the kth

observation is a waste of time and resources which makes mixed case interval censoring preferred to case k interval censoring especially when k is large.

1.3.4 Partly Interval Censored

One of the most important types of interval censored data is partly interval censored data which means that for some of the subjects the event of interest is exactly observed while for others it lies within an interval (Kim 2003).

Not many researchers used partly interval censored data in their study compared with other types that mentioned early in this chapter. In this thesis, analysis will be based on partly interval censored via engineering and medical data.

1.4 IMPUTATION

Imputation methods can be classified into:

1. Probability-based imputation method.
2. Simple imputation methods.

1.4.1 Probability-Based Imputation Methods

Probability-based imputation requires estimating the distribution of the partly interval censored data based on the observed intervals and using our knowledge of the distribution to impute the missing data. More detailed discussion of this probability based imputation techniques and references of past work are given in the next two chapters.

1.4.2 Simple Imputation Methods

There are three main types of simple imputation methods:

1. *Right-point imputation* where the event time is imputed by the right limit of the interval.

2. *Left-point imputation* where the event time is imputed by the left limit of the interval.

3. *Mid-point imputation* which refers to imputing the event time by the midpoint of the interval.

1.5 PROBLEM STATEMENT

Cox's proportional hazard model is one of the most important statistical methods. It is widely used in medical, engineering, economical researches and etc. Many researchers addressed Cox model from several angles, among others; Kim (2003) discussed the maximum likelihood estimation in the present of partly interval censored data under the Cox model. Elfaki (2012) used Cox model with Weibull distribution in the present of partly interval censored data and applied it to AIDS studies. Elfaki et al (2013) presented the estimating functions for partly interval censored data using the semi-parametric Cox's model of the sub-distribution function. Alharpy and Ibrahim (2013a) used parametric Weibull distribution for score test and likelihood ratio test based partly interval censored data and Alharpy and Ibrahim (2013b) used piecewise exponential distribution with non-proportional hazard for partly interval censored data.

For imputation techniques, Liu et al. (1988) used midpoint imputation to estimate of the mean incubation period of AIDS. Mariotto et al., (1992) used midpoint imputation to estimate the acquired immune deficiency syndrome incubation period in intravenous drug users. Law and Brookmeyer (1992) used midpoint imputation for Kaplan-Meier to estimate the survival function based on wide interval censoring. Xiang et al. (2001) used right-point imputation on survival of patients with HIV. Tillmann et al., (2001) also used the right-point imputation method for HIV-infected patients. Zhang et al. (2009) compared right-point, midpoint, conditional mean, conditional median, conditional mode, multiple and random methods for doubly censored HIV data. Alharpy and Ibrahim (2013a & 2013b) used multiple imputations for parametric and nonparametric based on partly interval censored data.

As there are few studies that focus on the partly interval censored data and even fewer applied it to engineering related applications, this research will tackle partly interval censored data for reliability analysis and apply a model that is significantly applicable to be used in engineering and medical data via Cox proportional hazard model in the present of imputation techniques which is used to simplify the procedure.

1.6 RESEARCH OBJECTIVES

The main objectives of the study are:

- To modify a model suitable for engineering partly interval censored data.
- To compare the survival functions of the proposed model with the existing model.
- To investigate the performance of Cox's model on partly interval censored data using imputation techniques.
- To compare the imputation techniques based on partly interval censored data using both secondary data and simulation data.