ESTIMATION OF THE BASE HAZARD FUNCTION BY BOOTSTRAPPING

RIFINA ARLIN

Master of Science
Universiti Putra Malaysia

December 2004

ESTIMATION OF THE BASE HAZARD FUNCTION BY BOOTSTRAPPING

By

RIFINA ARLIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

December 2004

Dedicated to:

My parents
(Drs. H. Zainal Arifin and Hj. Darlina)

and

My beloved husband,
Danil Junaidy Daulay

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of the Master of Science

ESTIMATION OF THE BASE HAZARD FUNCTION BY BOOTSTRAPPING

By

RIFINA ARLIN

December 2004

Chairman: **Associate Professor Isa Daud, Ph.D.**

Faculty: **Science**

This thesis examines the techniques in estimating the base hazard function by bootstrapping. The base hazard function is a crucial part of survival analysis. It is used to construct an estimate of the proportional hazard model for every individual.

As in many methods for analysing survival data, this thesis utilizes the nonparametric model of Kaplan Meier, the Cox proportional hazard regression of the parametric model and the data validation by bootstrapping.

The Cox proportional hazard regression is used to model failure time data in censored data. Bootstrapping schemes validate the models based on Efron's technique and the data samples are generated using S-Plus programme randomizer.

Assessment of this method is investigated by performing simulation study on generated data. Two simulation studies are carried out to confirm the suitability of the models. Graph obtained from the results indicated that bootstrapping provides an alternative method in constructing estimation for base hazard function. This method is good alternative for a distribution-free approach with a minimal set of data.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Master Sains

**PENGANGGARAN FUNGSI BAHAYA DASAR
DENGAN BOOTSTRAPPING**

Oleh

RIFINA ARLIN

Disember 2004

Pengerus: **Profesor Madya Isa Daud, Ph.D.**

Fakulti: **Sains**

Kajian di dalam tesis ini mengkaji teknik untuk penganggaran fungsi bahaya dasar dengan bootstrapping. Fungsi bahaya dasar merupakan bahagian yang penting didalam analisis mandirian. Ianya digunakan untuk membangun suatu anggaran model bahaya berkadaran untuk setiap individu.

Sebagaimana dalam banyak kaedah untuk menganalisis data mandirian, tesis ini menggunakan Kaplan-Meier model yang tak berparameter, model regresi bahaya berkadaran Cox biasa yang berparametrik dan pengesahan data oleh bootstrapping.

Regresi bahaya berkadaran Cox biasa digunakan untuk memodelkan data masa kegagalan di dalam data tertapis. Bootstrapping mengesahkan model berdasar pada teknik Efron's dan contoh data dihasilkan dengan menggunakan program S-Plus secara rawak.

Penilaian kaedah ini diselidiki dengan melakukan kajian simulasi ke atas data yang dijana. Dua kajian simulasi dijalankan bagi menentukan kesahihan model yang digunakan. Gambarajah yang diperoleh daripada keputusan menunjukkan bootstrapping menyediakan suatu kaedah alternatif dalam membangunkan penganggaran untuk fungsi bahaya dasar. Kaedah ini adalah alternatif yang baik untuk suatu pendekatan bebas taburan dengan suatu data yang minimal.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious, the Most Merciful. He has given me the strength to complete my study. Peace be upon the Prophet Muhammad S.A.W., his family, and his followers.

I would like to express my great appreciation and sincere gratitude to Associate Professor Dr. Hj. Isa Daud, chairman of my supervisory committee, for providing useful suggestions to improve this work. His guidance, encouragement and support had made my dream of pursuing higher education comes true. I am also indebted to my co-supervisor Associate Professor Dr. Hjh. Noor Akma Ibrahim and Associate Professor Dr. Mat Yusof Abdullah.

Thank you to the authorities of IRPA 54064 led by Associate Prof. Dr. Isa bin Daud for granting me the financial assistance through the Graduate Assistantship scheme.

Special thanks to Mr. Kudus and Mr. Fauzi, PhD students in UPM, who have helped me in understanding S-Plus and Bootstrap. My appreciation also goes to my housemates Ms. May, Ms. Tessa, Mrs. Mirna and Mrs. Siffa for their kindness, help and assistance. Thanks to all my fellow country men Dr. Mayastri, Dr. Susila, Ms.Desi, Mrs. Tryana, Mr. Iing and Mr. Sulaiman for the encouragement.

Last, but not least, my greatest appreciation and respect are to my loving parents and my parents-in-law for their endless support and motivation throughout my study. My

sister, my brother, my brother-in-low, my sister-in-low and my unforgettable husband, who have shown endless interest in my study. “Hasbunallah wa ni’mal wakiil ni’mal maulaa wa ni’man nashiir.”

I certify that an Examination Committee met on 21st December 2004 to conduct the final examination of Rifina Arlin on her Master of Science thesis entitled "Estimation of Base Hazard Function by Bootstrapping" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulation 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Habshah Midi, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Mohd Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Kassim Haron, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Zainodin Hj Jubok, PhD
Professor
Centre for Management of Research and Conference
Universiti Malaysia Sabah
(Independent Examiner)

GULAM RUSUL RAHMAT ALI, PhD
Professor/Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee are as follows:

Isa Daud, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Noor Akma Ibrahim, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Mat Yusof Abdullah, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

_____

AINI IDERIS, PhD
Professor/ Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

RIFINA ARLIN

Date: 28st April 2004

TABLE OF CONTENTS

LISTS OF TABLES

LISTS OF FIGURES

CHAPTER 1
INTRODUCTION

Survival analysis deals with the modeling and analysis of data that has as a principal an end point to time until an event occurs. The failure time or survival time is the time elapsed between the entry of a subject into the study and the occurrence of an event thought to be related to the concerned (Anderson *et al*. 1980).

1.1 Background

A fundamental problem in survival analysis is estimating survival probabilities, as either a point estimate or a confidence interval. If a parametric model is assumed, such as the assumption of exponential or lognormal distributions for failure times, then estimation of the model parameters can be done using maximum-likelihood methods, modified to deal with the problem of censored data.

In this research, the data of interest are sets of survival times. These times represent a measured period from some origin t_0 to a particular observation. For example, time from recruiting an individual to a clinical trial to the time of the individual's death.

In some cases, the individual will be lost to the study before a failure time is observed. For example in a clinical trial, the individual may move from the area, or still be alive at the end of the trial. In such cases, the time is noted and referred to as a censoring time. Otherwise, the observation is a failure time or death time.

Deals with methods of estimating these functions from sample of survival data and are said to be non-parametric or distribution-free, since they do not require specific assumptions to be made about the underlying distribution of the survival times. Kaplan and Meier (1958) proposed a nonparametric method, which has been accepted as the standard estimator for such probabilities.

The proportional hazards model was proposed by Cox (1972) and has also come to be known as the Cox regression model. Although the model is based on the assumption of proportional hazards, no particular form probability distribution is assumed for the survival times.

Cox proportional hazards regression model or Cox model has been widely used as a well-known method for the analysis of clinical data (Farewell, 1979). The Cox model has been generalized to include designs, which include multiple failures for the same individual (Andersen and Gill, 1982; Wei, Lin, and Weissfeld, 1989) and cohort studies (Prentice, 1986).

Bootstrap methods resembling techniques for assessing uncertainty. They are useful when inference is to be based on a complex procedure for which theoretical results are unavailable or not useful for the sample sizes met in practice. Or when a standard model is suspect but it is unclear what to replace it with or where a ‘quick and dirty’ answer is required. They can also be used to verify the usefulness of standard approximations for parametric models, and to improve them if they seem to give inadequate inferences.

An improved bootstrap method that is second-order correct in a wider class of problem concerns setting approximate confidence intervals for a real-valued parameter θ in a multi parameter family (Efron, 1987). Beran, 1985 developed a higher-order approximate confidence intervals based on Edgeworth expansions which sometimes use bootstrap methods to reduce the theoretical computations.

1.2 Objective of the Research

The general objective of the research is to estimate the hazard function of survival data. These estimates can then be used to summarize the survival experience of individuals in the study. Therefore, the main objective of the study is to estimate the base hazard function in the proportional hazards model by bootstrapping through the estimation of the parameters, whereas the specific objectives are as follows:

a) To estimate the hazard function based on the Kaplan-Meier estimator in small sample;

b) To determine the confidence intervals of the hazard function based on the Kaplan-Meier estimator by bootstrapping;

c) To obtain the parameters in the proportional hazards model based on maximum likelihood theory by bootstrapping.

CHAPTER 2
LITERATURE REVIEW

Survival analysis is the study of experiments which are performed to measure the amount of time that elapses until a particular event occurs. The time event can only occur once for a given subject and is usually described as the subject's failure time.

According to Miller (1981) survival analysis is a loosely defined statistical term that encompasses a variety of statistical techniques for analyzing positive-valued random variables. Typically the value of the random variable is the time to the failure of a physical component (mechanical or electrical unit) or the time to the death of a biological unit of patient, animal, cell etc.

Ideally, each subject would be observed until they fail, however, this is not always possible. For example, some subject may not have reached their failure time when the study is terminated or some patients in a medical study may die from causes unrelated to the disease being studied. The time at which a subject ceases to be observed for some reason other than failure is called the subject's censoring time. So, survival analysis is used for data in which censoring is present.

2.1 Censored Data

A distinctive characteristic of survival data is that the event of interest may not be observed for every experimental unit. This feature is known as censoring. For example when monitoring the life span of an electric bulb, the time to failure may not be

observable due to it is un availability before it actually fails. The time to this event can be considered as censored time. Censoring may arise due to time limits and other restrictions depending on the nature of the experiment.

Survival data can arise from many of situation, particularly in problems of reliability and life time analysis (Basu, 1984).

2.1.1 Type I censoring

In clinical and epidemiological studies, censoring is mainly caused by a time restriction, and is therefore of type I (Lagakos, 1982).

Let t_c be some fixed number that is the fixed censoring time. Instead of observing $T_1, T_2, ..., T_n$ (the random variable of interest) we can only observe $Y_1, Y_2, \ldots, Y_n$ where

$$Y_{\cdot} = \begin{cases} T_{\cdot} & if \quad T_{\cdot} \le t_{\cdot} \\ t_{\cdot} & if \quad T_{\cdot} > t_{\cdot} \end{cases} \qquad \begin{cases} \text{(uncensored)} \\ \text{(censored)} \end{cases} \tag{2.1}$$

The distribution function of Y has positive mass P {T > $t_{\cdot}$ } > 0 at Y = $t_{\cdot}$

2.1.2 Type II censoring

Let $r < n$ be fixed, and let $T_1 < T_2 < T_3 < ... < T_n$ be the order statistic of failure times $T_1, T_2, T_3, ..., T_n$. Observation ceases after the r-th failure, so we can observe $T_1, T_2, T_3, ..., T_r$. In this design, the number of failures is not a random variable.

Both types of censoring designs, or generalizations of them, may be usefully adopted, provided that n, r and T are adequately chosen and appropriate methods are adopted for the analysis (Lawless, 1982).

2.1.3 Type III censoring

Let $C_{1,}\ C_{2,}\ C_{3,}\ ...,C_n$ be independent and identically distributed (i.i.d) each with degrees of freedom (df) G. C_i is the censoring time associated with T_i. We observe (Y_1, δ_1), (Y_2, δ_2), …, $(Y_{\cdot}, \delta_{\cdot})$, where

$$\cdot\ \cdot = \min (T_{\cdot}, C_{\cdot})$$