ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

A SURVIVAL ANALYSIS APPROACH ON VARIOUS FIELDS OF SOCIAL DEVELOPMENT: ECONOMICS, DEMOGRAPHY, SOCIAL WORK, PSYCHOLOGY

By

Eleni S. Galiatsatou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
March 2012

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΙΑ ΠΡΟΣΕΓΓΙΣΗ ΤΩΝ ΚΛΑΔΩΝ ΚΟΙΝΩΝΙΚΗΣ ΑΝΑΠΤΥΞΗΣ ΜΕ ΤΗΝ ΧΡΗΣΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΕΠΙΒΙΩΣΗΣ: ΟΙΚΟΝΟΜΙΑ, ΔΗΜΟΓΡΑΦΙΑ, ΚΟΙΝΩΝΙΚΗ ΕΡΓΑΣΙΑ, ΨΥΧΟΛΟΓΙΑ

Ελένη Σ. Γαλιατσάτου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Μάρτιος 2012

DEDICATION

To my parents Mantha and Spyros for their support and effort to make my dreams come alive...

ACKNOWLEDGEMENTS

First of all, I would like to thank all my professors for the chance they gave me to participate in the Postgraduate Program and for their precious advice all these years. I would like to thank specially my supervisor associate professor Katerina Dimaki, for her patience, guidance and recommendations in order to this research come up to completion.

VITA

My name is Eleni Galiatsatou and I was born in Patra, on the 21nd of June in 1983. My distant origin, on the part of my mother, is from Patra. My distant origin on the part from my father is from Valsamata, a small village in Kefalonia, an Island in the Ionian Complex.

I completed the secondary educational stage in 2001, by graduating from the 7th High School of Patra. In 2001 I was admitted to the University of Patra, in the Department of Mathematics and I completed my under-graduate studies in Mathematics in 2007. The same year I was accepted at the Department of Statistics, of the Athens University of Economics and Business, for postgraduate studies in Statistics.

ABSTRACT

Eleni Galiatsatou

March 2012

Survival Analysis is a branch of statistics that examines the time until death of biological organisms and also failures of mechanical systems. Survival Analysis is trying to answer questions such as: Which is the part of a population that will survive after a specified time? Of those who will survive, in what rate will they die or fail? Can the multiple causes of death or failure be taken under consideration? How can the special circumstances or the characteristics raise or drop the probability of survival? For someone to answer these kinds of questions the definition of "lifetime" is necessary. In the occasion of biological survival, death is accurate, but in mechanic trustworthiness, failure may not be defied with total accuracy. Even in biological problems, some facts (such as heart attacks or other organ failure) can have the same inaccuracy.

The reason of this thesis is to present the fields of everyday life on which the analysis of survival data is applied. The thesis is divided into three parts. The first part deals with a historical background of survival analysis, the terminology that is needed to describe the method, while the third part presents some sections of science on which survival analysis is applied. Specifically, theory and applications on Economy, Social Work, Demography and Health are presented. The last part of this report deals with the conclusions that came up during this project.

ΠΕΡΙΛΗΨΗ

Ελένη Γαλιατσάτου

Μάρτιος 2012

Η Ανάλυση Επιβίωσης είναι ένας κλάδος της Στατιστικής που εξετάζει τον χρόνο που μεσολαβεί μέχρι τον θάνατο στους βιολογικούς οργανισμούς και την εμπλοκή ενός μηχανήματος στη μηχανολογία. Η ανάλυση επιβίωσης προσπαθεί να απαντήσει σε ερωτήσεις όπως: ποιο είναι το μέρος ενός πληθυσμού που θα επιζήσει μετά από έναν ορισμένο χρόνο; Από εκείνους που επιζούν, με τι ρυθμό θα "πεθάνουν" ή θα αποτύχουν; Μπορούν οι πολλαπλάσιες αιτίες θανάτου ή η αποτυχία να ληφθούν υπόψη; Πώς οι ιδιαίτερες περιστάσεις ή τα χαρακτηριστικά αυξάνουν ή μειώνουν τις πιθανότητες της επιβίωσης; Για να απαντήσει κανείς σε τέτοιες ερωτήσεις, είναι απαραίτητο να καθοριστεί «*η διάρκεια ζωής*». Στην περίπτωση της βιολογικής επιβίωσης, ο θάνατος είναι σαφής, αλλά για τη μηχανική αξιοπιστία, η αποτυχία μπορεί να μην είναι καθορισμένη με σαφήνεια. Ακόμη και στα βιολογικά προβλήματα, μερικά γεγονότα (παραδείγματος χάριν, καρδιακή προσβολή ή άλλη αποτυχία οργάνων) μπορούν να έχουν την ίδια ασάφεια.

Ο σκοπός της παρούσας διατριβής είναι να παρουσιάσει τομείς της καθημερινότητας στους οποίους εφαρμόζεται η ανάλυση των δεδομένων επιβίωσης. Η διατριβή αποτελείται από 3 μέρη. Στο πρώτο μέρος παρουσιάζεται μια σύντομη ιστορική αναδρομή στην ανάλυση επιβίωσης, την ορολογία που απαιτείται για την περιγραφή της μεθόδου, ενώ το δεύτερο παρουσιάζει μερικούς τομείς της επιστήμης στους οποίους εφαρμόζεται η Ανάλυση Επιβίωσης. Συγκεκριμένα, παρουσιάζεται η θεωρία και οι εφαρμογές της στην Οικονομία, την Κοινωνική Εργασία, την Δημογραφία και την Υγεία. Το τελευταίο μέρος της παρούσας διατριβής περιέχει συμπεράσματα που προέκυψαν κατά την εκπόνηση της εργασίας.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 8

Application of Survival Analysis in Social Work

8.3 The application of the method

CHAPTER 9

Application of Survival Analysis in Event History

9.8 Application of age at first partnership

CHAPTER 10

Application of Survival Analysis in Psychology

10.3 Survival Analysis

LIST OF FIGURES

CHAPTER 7

Application of Survival Analysis in Economics

7.2 Survival Analysis

7.2.1 The Kaplan-Meier Curve

CHAPTER 8

Application of Survival Analysis in Social Work

8.3 The application of the method

CHAPTER 10

Application of Survival Analysis in Psychology

10.3 Survival Analysis

CHAPTER 1

Introduction

In many fields of the natural, medical, and social sciences, there has been much interest in the analysis of data representing the time to occurrence of certain events. Data that measure lifetime or the length of time until the occurrence of an event are called **lifetime**, **failure time** or **survival data**. For instance, variables of interest for engineers might be the lifetime of a specific machine component. Medical scientists are concerned with hospitalizations, visits to a physician, and death or relapse of patients in a clinical trial. In the study of work and careers, attention is given to the length of time a person stayed on a job, job changes, promotions, unemployment or duration of a strike. Criminologists study crimes and arrests while demographers focus on births, deaths, marriages, divorces and migration.

The history of 'survival analysis' begun at 1662 when the book 'Natural and Political Observations upon the bill of Mortality' was published by John Graunt, an English statistician, generally considered to be the founder of the science of demography, the statistical study of human populations. A prosperous haberdasher until his business was destroyed in the London fire of 1666, Graunt held municipal offices and a militia command. While still active as a merchant, he began to study the death records that had been kept by the London parishes since 1532. In his book Graunt classified death rates according to the causes of death among which he included overpopulation, the people's age, the time of the event and the gender: he observed that the urban death rate exceeded the rural. He also found that although the male birth rate was higher than the female, it was compensate by a greater mortality rate for males, so that the population was divided almost evenly between the sexes. His most important innovation was the life table, which presented mortality in terms of survivorship. Using only two rates of survivorship (to ages 6 and 76), derived from actual observations, he predicted the percentage of persons that will live to each successive age and their life expectancy year by year. Many years later, between 1687 and 1691 Edmund Halley created the First Life Table which has a lot of similarities with the Life Tables we use nowadays in demography and event history analysis.

After a sixty year long evolving procedure the level of nowadays' knowledge of survival analysis is well developed. During World War II, survival analysis was used to study failure of military equipment and predicting the probability of response whereas during the last 30 years it's use pertains to clinical trials and the development of a disease. The most remarkable is that we consider the survival time not only as the "time to death" but also time until the occurrence of a failure. (**Dimaki, 2007**)

The purpose of this research is to present various applications of survival analysis, except for the most common field which is clinical trials, used in medicine,

CHAPTER 2

Functions of Survival Time

2.1 Introduction

In this chapter, we present the statistical method for analyzing survival data. We define survival time as a random variable that corresponds to the time from the beginning of the follow-up period of an individual until the failure. In the past, the study of survival data has focused on predicting the probability of response, survival or mean lifetime while in recent years, the identification of risk and prognostic factors related to response, survival, and the development of a disease has become equally important. Nowadays, the survival analysis is suitable for applications in industrial reliability, demography, social science, business and marketing. Examples of survival data in these fields are the lifetime of firms in a market environment, lifetime of electronic devices (reliability engineering), duration of first marriage (sociology), felon's time to parole (criminology), duration of strikes or periods of unemployment in economics.

2.2 Censoring

Many researchers consider survival data analysis to be merely the application of two conventional statistical methods to a special type of problem: a) Parametric if the distribution of the survival time is known to be normal and b) Nonparametric in the case of unknown distribution. This assumption could be true if the survival times of all the subjects were exact and known. However, some survival times are not. Further, the survival distribution is often skewed or far from being normal. Thus, there is a need for new statistical techniques. One of the most important developments is due to a special feature of survival data in the life sciences that occurs when some subjects in the study have not experienced the event of interest at the end of the study or time of analysis. For instance, some patients may still be alive or in remission at the end of the study period. The exact survival times of these subjects are unknown. These are called *censored observations* or *censored times* and can also occur when individuals are lost to follow-up after a period of study. When these are not censored observations, the set of survival times is *complete*. There are three types of censoring:

a. Type I Censoring

Clinical trials usually start with a fixed number of subjects, to which the treatment is given. A familiar difficulty in the analysis of survival data is when we have some information about individual failure time but we do not know the real time to failure. Sometimes the subject does not experience the failure event before the end of the study or the subject is lost to follow-up during the study period, or the subject withdraws from the study because of some reason(e.g. time or cost limitations, death is not the event of interest, adverse drug reaction, etc.) (Lee 1992). First option is to observe for a fixed period of time, after which the animals that survived are sacrificed. This kind of censorship is called *Type I.* In this type of censoring, if there are no accidental losses, all censored observations equal the length of the study period. Survival times recorded for the subjects that died during the study period are the times from the start of the experiment to their death and are called *uncensored observations*. On the other hand, the survival times of the sacrificed subjects are not known but are recorded as at least the length of the study period. These times are called *censored observations*.

For instance, consider leukemia patients followed until they go out of remission (survival time is the time in remission). In case of a patient's death due to a heart disease, the patient's failure time is considered *censored*. Knowing that, the survival time of this person is at least as long as the period that the person has been followed, but we cannot know in any case the full failure time.

b. Type II Censoring

Second option is to wait until a fixed portion of the subjects have failed. This is known as *Type II* censoring. If there are no accidental losses, the censored observations equal the largest uncensored observation. Type I and Type II are also called singly censored data.

For instance, in an experiment of 100 animals, the study terminates when a portion, say 80 out of 100 dies, and the surviving animals are sacrificed. In this case, if there are no accidental losses, the censored observations equal the twenty largest uncensored observations.

c. Type III Censoring

Last but not least, in many clinical studies the period of study is fixed and patients enter the study at different times during that period. For subjects that are lost to follow-up or do not fail until the end of the study, their survival times begin from their entrance until the last contact or the end of the study respectively. This third option is called Type III censoring or progressively censored data (Cohen, 1965). It is also called random censoring.

For example, suppose that six patients with acute leukemia enter a clinical study during a total study period of one year. The remission times of the patients vary according to each organism and leukemia type. If a patient gets into remission in the beginning of the fifth month and he is still in remission at the end of the study, then the observed survival or censor time for the particular patient is seven months.

All of the types of censoring are *right censoring* (or censoring to the right), *left censoring*, *both left and right*, and in some special cases, *censored observations within the observation period* (interval censoring)(Yamaguchi, 1991). For instance, right censored observations can occur in life course histories at the time of the retrospective interview or, both left and right censored observations could be found in a panel study of job mobility (Blossfeld and Rohwer, 1995). When there are no censored observations, the set of survival times is complete.

2.3 Definitions

Survival time is defined as a nonnegative random variable (let assume T), as is has already been mentioned at the beginning of this chapter, which is the time of failure of the entity known to exist at time t=0, and is therefore frequently called the *failure time random variable* and like any random variable forms a distribution. The distribution of survival time is described by three functions: 1. The Survivorship Function, 2. The Probability Function, 3. The Hazard Function. All these functions are mathematically equivalent; if one of them is given, the other two can be derived easily.

- **Survivorship Function (or Survival Function)**

This function is defined as the probability that a subject survives longer than t or the probability that failure (death) will occur after time t, which is the same as the probability that the entity, known to exist at time t=0, will survive to at least time t. It is symbolized as **S(t)**. That is to say:

$$S(t) = P(\text{a subject survives longer than t}) = P(T > t) \qquad (2.1)$$

or

$$S(t) = 1 - P(T \le t) = 1 - P(\text{a subject fails before time t}) = 1 - \int_0^t f(x)dx = 1 - F(t) \qquad (2.2)$$

where the **f(t)** is the *failure density* and the **F(t)** the *failure cumulative probability*. Moreover, S(t) is a non-increasing function of time t with two properties:

$$S(t) = \begin{cases} 1\ , & \text{for t=0} \\ 0\ , & \text{for t=}\infty \end{cases}$$

That is, the probability of surviving at least at the time zero is 1 and that of surviving an infinite time is 0. If T is the time of failure of an entity which exists at t=0, then T is also the future lifetime of this entity measured from t=0. It is important to note that the age of the study unit, animate or inanimate, at time t=0, and hence its attained age at time of failure, is not of interest to us, and might not even be known. The reason for this is that we believe the chance of failure to be a function of time under the study conditions, and not of the attained chronological age of the study unit. For this reason we use the function S(t). This function is also called *cumulative survival rate*. In 1942, Berkson recommended a graphic presentation of S(t) which has been called *survival curve* up to our days.

In practice, if there are no censored observations, the survival function is estimated as the proportion of patients surviving longer than t:

$$\hat{S}(t) = \frac{\text{number of patients surviving longer than t}}{\text{total number of patients}} \qquad (2.3)$$

where the circumflex denotes an estimate of the function. Thus, when censored observations are present is no longer appropriate for estimating S(t). We will discuss the estimation of S(t) for censored data with the use of nonparametric methods in chapter 4.

- **Probability Density Function (or Density Function)**

For the special case of a continuous random variable, such as the survival time T, its probability density function is defined as the limit of the probability that an individual fails in the short interval t to t+h per unit width h, or the probability of failure in a small interval per unit time. Thus,

$$\begin{aligned} f(t) &= \lim_{h\to 0^+} \frac{P[\text{an idividual dying in the interval (t,t+h)]}}{h},\ \mathrm{t} \geq 0 \\ &= \lim_{h\to 0^+} \frac{P(t \leq T \leq t+h)}{h} = \lim_{h\to 0^+} \frac{P(T \leq t+h) - P(T \leq t)}{h} = \\ &= \lim_{h\to 0^+} \frac{F(t+h) - F(t)}{h} = \lim_{h\to 0^+} \frac{1 - S(t+h) - 1 + S(t)}{h} = \\ &= -\lim_{h\to 0^+} \frac{S(t+h) - S(t)}{h} = -S'(t) \end{aligned}$$

This function indicates the *unconditional instantaneous probability* of event occurrence or episodes ending at the exact time t. By this we mean that it is the density of failure at time t given only that the entity existed at *t=0*. The notion *h* represents a small time interval Δt. As it becomes smaller and smaller, the density function *f(t)* reaches the limit on the right hand side of the equation we set out above. Whereas *F(t)* and *S(t)* are probabilities which relate to certain time intervals, *f(t)* relates to a point of time, and is not a probability, per se. It is an *instantaneous* measure, as opposed to an interval measure. The graph of *f(t)* is called the *density curve*.

In the discrete case, it is just $P(T \leq t) = F(t) = 1 - S(t)$.

If there are no censored observations, the probability density function is estimated as the proportion of patients dying in an interval per unit width.

- **Hazard Function**

Distribution function F(t), survival function S(t) and density function f(t), are mathematical notions which are anticipated to express a process going forward in time. Under a causal view of such a process, temporal aspects as the past, present and future, should be present on the description of the distribution of the duration variable T, with respect to the set of individuals whose behavior generates the process. A complemented description of the distribution of the variable T becomes available when the episode under study has ended for all the individuals. Consequently, to make a causal assessment of *how the process evolves*, we should use a concept that allows *describing the development of the process at every point in time, while the process is going on* (Blossfeld and Rohwer, 1995). This concept is known as *hazard rate h(t)* of survival time T, or the Hazard Rate Function (HRF), or the age-specific failure rate, and gives the conditional density of failure at time t. It is defined as the probability of the subject to fail within the time t given that it had already survived until the time point t. Specifically,

$$h(t) = \lim_{h\to 0}\left[\frac{\begin{array}{r} P(\text{an individual fails in the time interval (t,t+h)} \\ |\text{the individual has survived to t} \end{array}}{h}\right]$$

$$= \lim_{h\to 0}\left[\frac{P(t \le T \le t+h \,|\, T > t}{h}\right]$$

$$= \frac{1}{S(t)}\lim_{h\to 0}\left[\frac{P(t < T \le t+h)}{h}\right] = \frac{f(t)}{1-F(t)}$$

In other words, the hazard rate function can be written in the continuous case, as:

$$h(t) = \frac{f(t)}{S(t)}, \quad \text{where t} \ge 0$$

Whereas in the discrete case,

$$h(t) = \frac{P(T=t)}{P(T \ge t)}, \text{ where t= 0, 1, 2, ...}$$

It is very important to have a clear understanding of the descriptive meanings of h(t) and f(t). They are both instantaneous measures of the density of failure at time t; they differ from each other in that h(t) is *conditional* on survival to time t, whereas f(t) is *unconditional* (i.e. given only existence at time t=0). The hazard rate function h(t) is anyhow, a specialized characteristic of the data. However, is very useful for the study of survival time and thus for the failure distribution, if we consider also that usually the information available is about the diachronic evolution of h(t). In this sense we can choose the functional expression of the hazard rate function for the specific system. For that reason, we end up with a differential equation, or an equation of differences, depending on the type of the random variable. In the actuarial context of human survival models, failure means *death*, or *mortality*, and the hazard rate is normally called the *force of mortality*.

In practice, when there are no censored observations the hazard function is estimated as the proportion of patients dying in an interval per unit time, given that they have survived to the beginning of the interval.

The three above functions are mathematically equivalent. Given any one of them, the other two can be derived. In general, if a conditional measure is multiplied by the probability of obtaining the condition, then the corresponding unconditional measure will result. Thus,

$$h(t)\cdot S(t)=f(t)$$

Or

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

Some important mathematical consequences follow directly from the above equation. Since the probability density function is derivative of the cumulative distribution function,

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

It follows that,

$$h(t)=-\frac{S'(t)}{S(t)}=-\frac{d}{dt}\ln S(t)$$

Integrating, we have

$$\int_0^t h(x)dx=-\ln S(t)$$

Or

$$S(t)=\exp\left[-\int_0^t h(x)dx\right]$$

The Cumulative Hazard Function (CHF) is defined to be

$$H(t)=\int_0^t h(x)dx=-\ln S(t)$$

So that,

$$S(t)=\exp\left[-H(t)\right]$$

Through a similar approach we can define the mathematical relationship between the hazard rate and the density function, which is given by

$$f(t) = h(t) \cdot \exp\left[-H(t)\right]$$

CHAPTER 3

Specific Parametric Survival Distributions

3.1 Introduction

Usually, there are many physical causes that lead to the failure or death of a person at a particular time. It is very difficult, if not impossible to isolate these physical causes and account mathematically for all of them. Parametric approaches are used either when a suitable model or distribution is fitted to the data or when a distribution can be assumed for the population from which the sample is drawn. If a survival distribution is found to fit the data properly, the survival pattern can then be described by the parameters in a compact way. Statistical inference can be based on the distribution chosen. In this chapter we will present several theoretical distributions that have been used widely to describe survival time.

3.2 Exponential Distribution, E(λ)

The exponential distribution is often referred to as a purely random failure pattern. It is famous for its unique "*lack of memory*", which requires that the age of the individual does not affect future survival. This property of lack of fit allows the use of the exponential distribution for the description of the lifetime of a system when there is no actual loss in the system due to the passage of time. Nevertheless, in the framework of survival analysis this situation is unreal as it actual accepts that the working time does not result to damage in the survival time. There are cases although where this situation is found, like in reliability theory, when analysis focuses on the lifetime control with replacement.

When the survival time T follows the exponential distribution with a parameter λ,

✓ The Probability Density Function is defined as

$$f(t) = \lambda e^{-\lambda t}, \quad \mathrm{t} \geq 0, \lambda > 0$$

✓ The Cumulative Distribution Function is

$$F(t) = 1 - e^{-\lambda t}, \quad \mathrm{t} \geq 0$$

✓ The Survival Function is

$$S(t) = e^{-\lambda t}, \quad \mathrm{t} \geq 0$$

✓ The Hazard Function is

$$h(t) = \lambda, \quad \mathrm{t} \geq 0$$

Because the exponential distribution is characterized by a constant hazard rate, independent of the age of the individual, there is no ageing or wearing out, and failure or death is a random event independent of time. This condition is necessary and capable to ensure that any non-negative random variable T is exponentially distributed.

3.3 Weibull Distribution, Weibull(λ,p)

The Weibull Distribution is a generalization of the exponential distribution. It is characterized by two parameters, λ and p. The value of λ determines the shape of the distribution curve and the value of p determines its scaling. Consequently, λ and p are called the *shape* and *scale parameters*, respectively.

✓ The probability density function is

$$f(t) = p\lambda^{p} t^{p-1} \exp\left[-(\lambda t)^{p}\right], \ \mathrm{t} \geq 0, \lambda, \mathrm{p} > 0$$

✓ The Cumulative Distribution Function is

$$F(t) = 1 - \exp\left[-(\lambda t)^{p}\right], \ \mathrm{t} \geq 0, \lambda, \mathrm{p} > 0$$

✓ The Survival Function is

$$S(t) = \exp\left[-(\lambda t)^{p}\right], \ \mathrm{t} \geq 0, \lambda, \mathrm{p} > 0$$

✓ The Hazard Rate is

$$h(t) = \lambda p(\lambda t)^{p-1}, \ \mathrm{t} \geq 0, \lambda, \mathrm{p} > 0$$

When λ=1, the hazard rate remains constant as time increases; this is the exponential case. When λ>1, the hazard rate increases, and when λ<1 it decreases as time t increases.

3.4 Lognormal Distribution, Λ(μ, σ^2)

One of the most commonly used distribution is the Normal Distribution. Since the Normal allows negative value, a plausible way of using it in Survival Analysis is to take *logT* normally distributed. So we assume a lognormal distribution for the failure times.

✓ The probability density function is defined

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} (\log t - \mu)^2 \right], \text{ t>0, σ>2}$$

✓ And the Survivorship Function:

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} \exp\left[-\frac{1}{2\sigma^2} (\log x - \mu)^2 \right] dx$$

The hazard function of the lognormal distribution cannot be written explicitly but only in terms of intervals. For small values of σ, the lognormal density looks very like a Normal one.

3.5 Gamma Distribution, G(α)

The continuous random variable T follows the Gamma Distribution with parameters α, β >0 if

✓ The survival function is

$$S(t) = \int_t^\infty \frac{\beta^a}{\Gamma(a)} e^{-\beta y} y^{a-1} dy, \text{ t, y>0}$$

✓ The probability density function is

$$f(t) = \frac{\beta^a}{\Gamma(a)} e^{-\beta t} t^{a-1}$$

$$\text{With } \Gamma(t) = \begin{cases} \int_0^\infty e^{-y} \cdot y^{t-1} dy, & t > 0 \\ 1 & , t = 1 \\ \sqrt{\pi} & , t = \frac{1}{2} \\ (t-1) \cdot \Gamma(t-1), & t > 1 \\ (n-1)! & , t = n \in \square \end{cases}$$

✓ The hazard rate is

$$h(t) = \frac{\beta (\beta t)^{a-1} \exp(-\beta t)}{\Gamma(a) \left[1 - I_a(\beta t) \right]}$$

Where $I_\alpha(t)$ is the cumulative function of the standardized Gamma which is made by Gamma with parameters α and β =1. In this spot, the cumulative function of the standardized Gamma is given by

$$I_\alpha(t) = F(t) = P(T \le t) = \int_0^t \frac{1}{\Gamma(a)} e^{-y} y^{a-1} dy, \text{ t, y>0}$$

When 0 < α < 1, there is negative ageing and the hazard rate decreases monotonically from infinity to β as time increases from 0 to infinity. When α>1 there is positive aging and the hazard rate increases monotonically from 0 to β as time increases from 0 to infinity. When α=1, the hazard rate equals β, a constant, as in the exponential case. So, varying *a* changes the shape of the distribution while varying *b* changes only the scaling.

3.6 Compound Exponential or Pareto Distribution, Pareto (θ,α)

A continuous random variable T with survival function

$$S(t) = \theta^{a} t^{-a},\ t \geq \theta,\ \alpha,\ \theta>0$$

follows a Pareto Distribution or Compound Exponential Distribution with parameters α and θ.

✓ The Cumulative Distribution Function is

$$F(t) = 1 - \theta^{\alpha} t^{-a},\ t \geq \theta,\ \alpha,\ \theta>0$$

✓ The probability density function is

$$f(t) = \alpha\theta^{a} t^{-(a+1)},\ t \geq \theta,\ \alpha, \theta>0$$

✓ The hazard function is

$$h(t) = \frac{a}{t},\ \ \alpha>0$$

3.7 The Gompertz Distribution

The Gompertz Distribution is also characterized by two parameters λ and γ. Its

✓ Hazard Function is

$$h(t) = \exp(\lambda + \gamma t)$$

When γ>0, there is a positive aging starting from e^{λ}. When γ<0, there is a negative aging and when γ=0, h(t) reduces to a constant, e^{λ}.

✓ The survivorship function is

$$S(t) = \exp\left[-\frac{e^{\lambda}}{\gamma} \cdot \left(e^{\gamma \cdot t} - 1\right)\right]$$

✓ The Probability density function is

$$f(t) = \exp\left[\lambda + \gamma \cdot t - \frac{1}{\gamma}\left(e^{\lambda+\gamma t} - e^{\lambda}\right)\right]$$

3.8 Geometric Distribution, G(p)

A discrete random variable T with

✓ Survival Function

$$S(t) = P(T > t) = q^{t+1}, \;\; t = 0,1,2,..., \; 0 < p < 1, \; q = 1 - p$$

Is said to follow the geometric distribution with parameter p. The survival function is defined on the positive integers, while $\mathrm{p} \in (0,1)$.

✓ The Probability Function is given by

$$f_T(t) = P(T = t) = p \cdot q^t = p \cdot (1-p)^t, \; t = 0,1,2,..., \;\; 0 < p < 1$$

✓ The Hazard Rate is defined as

$$h_T(t) = p$$

3.9 Yule Distribution, Yule(p)

A discrete random variable T is said to follow the Yule distribution with parameter p if

✓ The Survival Function is

$$S(t) = \frac{t+1}{p} \cdot pr(T = t), \quad t = 0,1,2,..., \; p > 0$$

✓ The probability function is

$$P(T = t) = \frac{pt!}{(p+1)(p+2)\cdots(p+t+1)}$$

✓ The Hazard manipulation is

$$h(t) = \frac{pr(T = t)}{pr(T \geq t)} = \frac{pr(T = t)}{pr(T > t) + pr(T = t)} =$$

$$= \frac{pr(T = t)}{\frac{t+1}{p} pr(T = t) + pr(T = t)} = \frac{p}{p + t + 1}$$

CHAPTER 4

Nonparametric Approaches of Estimating

4.1 Introduction

Nonparametric methods are more efficient when no theoretical distribution fits the data sufficiently, or the search of an appropriate model is too time consuming or not economical. For all the above reasons, nonparametric approaches can be suitable to describe the characteristic features included in a substantive process that is under study. Since these methods do not make any assumptions about the distribution of the process, they could be remarkably useful as first explorations in the data analysis before attempting to fit a theoretical model to the data. For instance, this consideration is very helpful to biostatisticians whenever their experiments allow for few if any assumptions about the distribution of event occurrences. In this chapter we will describe two typical nonparametric methods: The *Product-Limit* or *Kaplan-Meier* and the *Life table* estimation method.

These nonparametric estimation methods provide very useful estimates of survival probabilities and graphical presentation of survival distribution as the transition rate or hazard rate (**Blossfeld and Rohwer, 1995**). Moreover, if the sample size is very large, for example in the thousands, or the interest is in a large population, it may be more convenient to perform a life table analysis. Although this method is one of the oldest techniques for measuring mortality and describing the survival experience of a population, many of the actuaries, demographers, governmental agencies, and medical researchers still tend to favor life table (**Kostaki, 1997**).

The PL estimates and life table estimates of the survivorship function are essentially the same. The only difference is that the PL estimate calculates risk sets at any point in time, but needs a large amount of calculations when large data sets are used while in the life table method survival times are grouped into discrete time intervals. However, with the increased availability of computers, PL method can be applicable to small, moderate, and large samples.

4.2 Product - Limit estimation

A non-parametric estimate of the survival function in the case of any right-censored sample, is the product-limit method developed by Kaplan and Meier (1958). A basic characteristic of the product-limit estimator is that it does not require a distribution assumption, such as the uniform or exponential. We will present the process in the simple case of single mutation.

Firstly, we assume that we have *n individuals* under study which all are having the same origin state and *k failures* to occur. Moreover, all episodes of the sample either come to the same destination state, or they have been censored on the right. Also let $\mathbf{m_{(i)}}$ be the number of failures at time $\mathbf{t_i}$ with i=1,2,…,k and assume that t_0=0, which means all episodes have been starting at the time point zero.

Hence, we regard these k points of time in ascending order such as $t_1 \leq t_2 \leq ... \leq t_k$, where at least one of the episodes closes when a failure occurs. Last but not least, we recall $\mathbf{r_i}$ as the number of individuals in risk set at time t_i.

We should notice that the censored episodes ending in the interval $[t_{i-1}, t_i)$ are included in the risk set at the time point t_i. A censored episode in $[t_{i-1}, t_i)$ has no event up to time point t_i, but its duration also includes the same ending time point. It shows that censoring comes about an infinitesimal length of time after the observed ending time point.

The product-limit estimator of the survivor function $\widehat{S}(t)$ is given by

$$\hat{S}(t_i) = \prod_{i:\, t_i < t} \left(1 - \frac{m_{(i)}}{r_i} \right)$$

While in case of uncensored observations the above estimation has the simple expression we have already mentioned in chapter 2 (equation 2.3) and is repeated below:

$$\hat{S}(t) = \frac{\text{number of patients surviving longer than t}}{\text{total number of patients}}$$

If two or more t_i are equal (tied observations), the largest i value is used. Since every individual is alive at the beginning of the study and no one survives longer than t_k, then

$$\hat{S}(t_0) = 1 \quad \text{and} \quad \hat{S}(t_k) = 0 \tag{4.1}$$

Moreover, there is the estimation of standard error of $\hat{S}(t)$, based on the asymptotic theory at a fixed time t and is given by

$$SE[\hat{S}(t)] = \hat{S}(t) \cdot \left[\sum_{i:t_i<t} \frac{m_i}{r_i(r_i - m_i)} \right]^{\frac{1}{2}}$$

An apparent PL estimate of the cumulated transition or transition rate is given by

$$\hat{H}(t) = -\log[\hat{S}(t)]$$

The Cumulative transition rate $\hat{H}(t)$ is very effective for giving simple graphical checks on the assumption that the time, until the event occurrences, follows a particular distribution. On the other hand, the PL estimate method is unable to provide exact estimates of the transition rate.

As far as $\hat{S}(t)$ is concerned, it is computed at every distinct survival time. We do not have to worry about the intervals between the distinct survival times in which no one dies and $\hat{S}(t)$ remains constant. From the definition of $\hat{S}(t)$ and its properties (4.1) we figure that $\hat{S}(t)$ is a step function starting at 1 and decreasing in steps of 1/k (if there are no ties) reaches 0. Also, from the plot of $\hat{S}(t)$ per *t*, we are able to read the various percentiles of survival time or calculate them from $\hat{S}(t)$.

The Kaplan-Meier method provides very useful estimates of survival probabilities and graphical presentation of survival distribution. It is the most widely used method in survival data analysis. Breslow and Crowley (1974) and Meier (1975) have shown that under certain conditions, the estimate is consistent and asymptomatically normal. However, a few critical features should be mentioned:

- The Kaplan-Meier estimates are limited to the time interval in which the observations fall. If the largest observation is *uncensored*, the PL estimate at that time is 0. If the largest observation is *censored*, the PL estimate can never equal 0 and is undefined beyond the largest observation.
- The most commonly used summary statistic in survival analysis is the ***median survival time.*** A simple estimation of the median can be read from survival curves estimated by PL method as the time t at which $\hat{S}(t)$=0.5. However, this solution is not unique. If the survival curve is horizontal at $\hat{S}(t)$=0.5, any t value in the interval [t_{i-1}, t_i) is a reasonable estimate of the median. A practical solution is to

take the midpoint of the interval as the PL estimate of the median. On the other hand there is the case of overestimating the median. A practical way to handle this problem is to connect the points and then locate the median.

- If less than 50% of the observations are uncensored and the largest observation is censored, the median survival time cannot be estimated. A solution to this problem is to use probability of surviving a given length of time, or the mean survival time limited to a given time t.
- The PL method assumes that censoring is independent of the survival times. In other words, the reason an observation is censored is unrelated to the cause of death. This assumption is true if the individual is still alive at the end of the study period. However, the assumption is violated if the patient develops severe adverse effects from the treatment and is forced to leave the study before failure or if the individual died of a cause other than the one under study. In case of inappropriate censoring, the PL method is inappropriate. One solution to this problem is to avoid it or reduce it to a minimum.
- The Standard Error of the Kaplan-Meier estimator of S(t) gives an indication of the potential error of $\hat{S}(t)$. We have to pay more attention to the 95% confidence interval for S(t) which is $\hat{S}(t) \pm 1.96 \cdot SE[\hat{S}(t)]$ rather than the point estimate $\hat{S}(t)$.

4.3 The Life Table

Another non-parametric method to estimate the survivor function, the density function and the transition rate for the time until an event occurs is the life table. It is one of the oldest techniques for measuring mortality and describing the survival experience of a population. Widely used in demographic and actuarial statistics, the life table method is the earliest, best known attempt for studying longitudinal event history data, in the form of single life table as well as in multiple decrement life table (**Namboodiri** and **Suchindran, 1987; Kostaki, 1997**). It has been used in studies of survival, population growth, fertility, migration, length of married life, length of working life, and so on. The life tables, summarizing the mortality experience of a specific population for a specific period of time, are called population life tables. The life tables applied to clinical and epidemiologic research are called clinical life tables.

In using this estimate method we have to take into account two restrictions. Firstly, we have to set up the durations into present time intervals. Because these time intervals are demarcated, the outcome of the process would be based on the ability of the researcher to define the length of the intervals. Secondly, we have to consider the set of episodes put into analysis, in order for the estimates be consistent, within each time interval. So, the life table method needs large data sets to provide good approximations of the survival function, the density function and the transition rate.

Taking everything into consideration, the time axis has to be marked off by a number of split points, say t_1, t_2, …,t_k such as

$$0 \leq t_1 < t_2 < ... < t_k$$

Given that $t_{k+1} = \infty$, the observation period has been divided in k fixed time intervals. Each one of these intervals, denoted by I_i, includes only the left limit. The interval is from t_i up to but not including t_{i+1}. The last interval has an infinite length. Formally,

$$I_i = \{t / t_i \leq t < t_{i+1}\}, \text{ where } i=1, 2, 3, ..., k$$

Afterwards, we show up the process of a single transition life table presented by the formulas used in the calculation of the survivor function, the density function and the transition rate. In the case of a single transition, each episode has only a single origin state and a single destination state. Moreover, to take the analysis straightforward, we should assume that all episodes comprising our sample have the same origin state.

Firstly, we assume that $\boldsymbol{l_i}$ is the number of individuals who are lost to observation and whose survival status is unknown in the *i*th interval. Secondly, let $\boldsymbol{w_i}$ be the number of individuals withdrawn alive in the *i*th interval and those are known to be alive at the closing date of the study. The survival time recorded for such individuals is the length of time from entrance to the closing date of the study. Also, let $\boldsymbol{d_i}$ be the number of individuals who die in the *i*th interval. The survival time of these individuals is the time from entrance to death.

Likewise, for each time interval, we should define a risk set $\boldsymbol{n_i}$ typifying the number of individuals who are exposed to risk in the *i*th interval. It is prospective that a number of episodes would be censored during each time interval. In order to calculate the risk set n_i, we should set firstly the number of individuals entering the *i*th interval, let $\boldsymbol{n_i'}$. Considering that the number of individuals who enter the first interval $\boldsymbol{n_1'}$ is the total sample size, this number is equal to the number of individuals studied at the beginning of the previous interval minus those who are lost to follow-up, are withdrawn alive, or have died in the previous interval, thus we have,

$$n_i^{'} = n_{i-1}^{'} - l_{i-1} - w_{i-1} - d_{i-1}$$

The next step is to determine the n_i number as a result of the operation

$$n_i = n_i^{'} - \frac{1}{2}(l_i + w_i)$$

It is assumed that the times to loss or withdrawal are approximately uniformly distributed in the interval. Therefore, individuals lost or withdrawn in the interval are exposed to risk of death for one-half of the interval. If there are no losses or withdrawals, $n_i = n_i^{'}$.

After the definition of the risk set n_i, we should present two important conditional probabilities. First of all, we should define the *conditional proportional dying* ***q***$\boldsymbol{q_i}$ which is the conditional probability for experiencing an event in the *i*th interval given exposure to the risk of death in the *i*th interval. Thus,

$$q_i = \frac{d_i}{n_i}, \text{ for } i=1,\ldots, k-1$$

Secondly, we define the conditional proportion surviving p_i as the conditional probability of surviving in the *i*th interval, given by

$$p_i = 1 - q_i$$

Using the conditional probability for not having an event in the *i*th interval, or the survivor probability p_i, we could obtain an estimator for the survivor function $S(t_i)$ at time t_i. It is often referred to as the cumulative survival rate. For $i=1$, $S(t_i)=1$ and for $i=2, \ldots, k$ estimates of the survivor function are given by

$$S(t_i) = p_{i-1} \cdot S(t_{i-1})$$

It is the usual life-table estimate and is based on the fact that surviving to the start of *i*th interval means surviving to the start of and then through the *(i-1)*th interval.

Coming up with the estimates of the survivor function $S(t_i)$, we are able to define the estimated probability density function $f(t_{mi})$ at the midpoints t_{mi} of the first k-1 intervals as the probability of dying in the *i*th interval per unit width. To summarize the above we have

$$f(t_{mi}) = \frac{S(t_i) - S(t_{i+1})}{t_{i+1} - t_i}, \text{ where } i=1,\ 2,\ \dots,\ k\text{-}1$$

In case of the last interval being open in the right side, it is impossible to estimate the survivor function *S(t$_k$)* for this interval. The same impediment, which stands for the density function *f(t$_{mi}$)*, goes for the transition rate *h(t$_{mi}$)*. The hazard function for the *i*th interval estimated at the midpoint, is

$$h(t_{mi}) = \frac{d_i}{(t_{i+1} - t_i)\cdot(n_i - \frac{1}{2}d_i)} = \frac{2q_i}{(t_{i+1} - t_i)(1 + p_i)}$$

It is the number of deaths per unit time in the interval divided by the average number of survivors at the midpoint of the interval. That is, *h(t$_{mi}$)* is derived from $\frac{f(t_{mi})}{S(t_{mi})}$ and since *S(t$_i$)* is defined as the probability of surviving at the beginning, not the midpoint, of the *i*th interval, we have

$$h(t_{mi}) = \frac{f(t_{mi})}{S(t_{mi})} = \frac{S(t_i)\,q_i/(t_{i+1} - t_i)}{\frac{1}{2}S(t_i)(p_i + 1)}$$

Searcher (1956) derives an estimate of the hazard function by assuming that hazard is constant within an interval but varies among intervals. His estimate is

$$h(t_{mi}) = \frac{(-\log_e p_i)}{t_{i+1} - t_i}$$

Finally, we need to derive the corresponding standard errors for each of the functions estimated above. Thereby, approximate standard errors for the estimate of survivor function *S(t$_i$)* are given by

$$SE[S(t_i)] = S(t_i)\cdot\left[\sum_{j=1}^{i-1}\frac{q_j}{p_j n_j}\right]^{\frac{1}{2}}$$

Moreover, approximate standard errors for the estimate of density function *f(t$_i$)* are derived from equation

$$SE[f(t_i)] = \frac{q_i S(t_i)}{t_{i+1} - t_i}\cdot\left[\sum_{j=1}^{i-1}\frac{q_i}{p_j n_j} + \frac{p_i}{q_j n_j}\right]^{\frac{1}{2}}$$

While approximate standard errors for the estimate of transition rate are obtained by

$$SE[h(t_{mi})] = \frac{h(t_{mi})}{\sqrt{q_i \cdot n_i}} \left\{ 1 - \left[\frac{h(t_{mi})(t_{i+1} - t_i)}{2} \right]^2 \right\}^{\frac{1}{2}}$$

Under the hypothesis of the treatment of large samples, we could assume that the estimates of the survivor, density and rate functions divided by their standard errors, can be approximately described by the standard normal distribution. In the case of large samples, it is likely to get confidence intervals for the values of above functions.

Comparing the life table method which groups the episode durations into fixed time intervals with the product-limit estimation, we conclude that PL estimator has a major advantage. It can use all the information included in the set of episodes put into analysis, by calculating the set of individuals at risk at every point in time, considering that at least one event occurred at the same point in time. According to Blossfeld and Rohwer (1995), the only disadvantage of the method is the large amount of calculations involved because of the large data sets used. This occurs because the method requires the diversity of all episodes according to their ending times. If the episodes have more than one origin states, things become more complicated. Thanks to modern computers with a lot of access memory and computation speed, we simplify this disadvantage.

CHAPTER 5

Comparing Survival Distributions with Non-Parametric Methods

5.1 Introduction

After the estimation of a Survival Function we wish to determine the acceptability of a fitted model as an adequate representation of the true underlying model. The problem of comparing survival distributions arises often in biomedical research. The investigator is interested in comparing the treatment's abilities to prolong life between two and three treatment groups. He seeks a well-known distribution for the remission patterns to compare the two groups. The survival times of the different groups vary, so these differences can easily be delineated by drawing graphs of the estimated survivorship function. The only disadvantage is that this graph does not reveal whether the differences are significant or casual variations. So, a need for a statistical test arises.

There are several parametric and non-parametric tests to compare two survival distributions. Since we have no information of the survival distribution that the data follow, we would continue to use non-parametric methods to compare the two survival distributions. We will present five tests that can be used for data with or without censored observations.

For each test we suppose that we have two groups of patients differing with respect to one factor, whose effect on the survival probability we want to study. Suppose that the n_1 observations in 1st group are samples from distribution with survivorship function $S_1(t)$ and the n_2 observations in 2nd group are samples from a distribution with survivorship function $S_2(t)$. In testing the significance of the difference between the two distributions, we need a hypothesis which concedes that the above survivorship functions are the same. So, we have

The null hypothesis (H_0) : $S_1(t) = S_2(t)$

The alternative two-sided hypothesis (H_1) : $S_1(t) \neq S_2(t)$

(which means that $S_1(t) > S_2(t)$ or $S_1(t) < S_2(t)$) **(Lee, 1992)**

5.2 Gehan's Generalized Wilcoxon Test

In Gehan's Generalized Wilcoxon Test we consider two groups with n_1 and n_2 individuals (observations) respectively differing in a certain factor, whose influence upon the failure time we want to study. The influence is usually reflected on the respective survival curves. Every observation x_i or x_i^+ (in case of censored observations) of the first group is compared with every observation y_i or y_i^+ (in case of censored observations) of the second group and a score U_{ij} is given to the result of every comparison. The scores that are used depend on the kind of the hypothesis test, and in this case we want to test that the survival function $S_1(t)$ is greater than $S_2(t)$. So,

$$H_0: S_1(t) = S_2(t) \quad \text{against} \quad H_1: S_1(t) > S_2(t)$$

We also define the U_{ij} to be:

$$U_{ij} = \begin{cases} +1, & \text{if } x_i > y_j \text{ or } x_i^+ \geq y_j \\ 0, & \text{otherwise} \\ -1, & \text{if } x_i < y_j \text{ or } x_i \leq y_j^+ \end{cases}$$

And the test statistic is calculated by:

$$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij}$$

where the sum is over all $n_1 n_2$ comparisons. Hence, there is a contribution of the test statistic W for every comparison where both observations are failures and for every comparison where a censored observation is equal to or larger than a failure.

In case of large populations the test statistic W is calculated with difficulty. Mantel (1967) shows that it can be calculated in an alternative way by assigning a score to each observation based on its relative ranking. According to his proposal, instead of comparing each observation of the first sample with each one of the second sample (Gehan's Test) we could combine the two samples into a single pooled sample of n_1+n_2 observations and compare each observation ***i*** of our new sample with the rest $\boldsymbol{n_1+n_2-1}$ observations. The n_1+n_2 U_i's define a finite population with mean zero and $W = \sum_{i=1}^{n_1} U_i$ where the summation is over U_i of the first **sample only**. From the above results if $\mathbf{H_1}$ is true, W would be a large positive number and under the null hypothesis $\mathbf{H_0}$ can be consider approximately normally distributed with mean zero and variance

$\mathrm{var}(W) = \dfrac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1+n_2)(n_1+n_2-1)}$. Since W is discrete, an appropriate continuity correction of **1** is used when there are no censored observations. Otherwise, a continuity correction of **0.5** would probably be appropriate.

Since W has an asymptotically normal distribution with mean zero and variance var(W), $Z = \dfrac{W}{\sqrt{\mathrm{var}(W)}}$ has standard normal distribution with mean 0 and variance 1. To complete our test we should present the rejection region, which for null hypothesis

H$_0$: $S_1(t)=S_2(t)$ against H$_1$: $S_1(t) \neq S_2(t)$, is $|\mathbf{Z}| > \mathbf{z_{\alpha/2}}$

where $P(Z > z_\alpha$ *given that* H_0 *is correct*$)=\alpha$.

The number U$_i$ can be computed in two stages. The first stage imputes, for each observation, unity plus the number of remaining observations that it is definitely larger than R_{1i}. The second stage yields R_{2i}, which is unity plus the number of remaining observations that the particular observation is definitely less than. Then $U_i = R_{1i} - R_{2i}$ **(Lee, 1992)**

5.3 The Cox-Mantel Test

Suppose that we have two groups of units, where the units of the first group satisfy normal conditions, while the units of the second group differ as compared to those of the first with respect to a certain feature.

We combine the failure and censored times of units for both groups into a new one and we rank them in ascending order. Let $t_{(1)}<t_{(2)}<\ldots<t_{(n)}$ be the distinct failure times in the two groups together and m$_i$ the number of failure times equal to t$_i$, so that

$$\sum_{i=1}^{n} m_i = r_1 + r_2$$

Further, let *R(t)* -which is called the *risk set* at time t- be the set of units still exposed to the risk of failure at time *t*, whose failure of censoring times are at least *t*. Let r_{1i} and r_{2i} be the number of units that are exposed to risk failure $R(t_i)$ and belong to the first and to the second group respectively. The total number of units in the risk set at each time t$_i$ is given by the sum $r_i=r_{1i}+r_{2i}$.

According to Cox (91972), under the null hypothesis H$_0$: $S_1(t)=S_2(t)$ an asymptotic two-sample test is thus obtained by treating the statistic $C=\frac{U}{\sqrt{I}}$ as a standard normal variate with mean zero and variance 1. The quantities U and I are defined as:

$$U = r_2 - \sum_{i=1}^{n} m_i A_i \ ,$$

where A_i is the proportion of r$_i$ that belong to group 2, that is $A_i = \frac{r_{2i}}{r_i}$ and $I = \sum_{i=1}^{n} \frac{m_i(r_i - m_i)}{r_i - 1} \cdot A_i \cdot (1 - A_i)$. To conclude, we also have to define the rejection region which is **|C| > $\mathbf{Z_{1-\alpha}}$** for the alternative hypothesis H$_1$: $S_1(t) \neq S_2(t)$. (**Lee, 1992**)

5.4 The Logrank Test

The Longrank test is a Mantel's (1966) generalization of the Savage (1956) test which is generally known only as a test for scale. Although the test can be used for the comparison of Survival curves of more than two groups differing respectively to one factor, it is mainly used for the case of two groups only. We assign to each observation of both groups a score w_i, which is the logarithm of the survival function. This score differs in the general form according to whether the observation is censored or not.

Altshuler (1970) estimated the log survival function at t_i using $-e(t_i) = -\sum_{j \le t_i} \frac{m_j}{r_j}$, where m_j is the number of failure until the time t_j while r_j is the number of units exposed to risk failure prior to time t_j.

Peto and Peto suggested that for an uncensored observation t_i the scores are given by $w_i=1 - e(t_i)$, while for a censored observation t_i^+ the score is $w_i= - e(t_j)$ with t_j represent the largest uncensored observation such that $t_j \le t_i^+$. Thus, the larger the uncensored observation becomes, the smaller its score. Censored observations receive negative scores, uncensored observations receive positive scores and the total scores for the two groups sum to zero. The logrank test is based on the sum S of the w scores in one of the two groups. The combined variance of S is then given by:

$$\mathrm{var}(S)=\frac{n_1 n_2\cdot\sum_{i=1}^{n_1+n_2} w_i^2}{(n_1+n_2)(n_1+n_2-1)}=\left[\sum_{j=1}^{n}\frac{m_j(r_j-m_j)}{r_j}\right]\cdot\frac{n_1\cdot n_2}{(n_1+n_2)(n_1+n_2-1)}$$

The two-sided hypothesis test that we are interested in is:

H$_0$: There is no statistically significant difference of the survival of the groups (thus, $S_1(t)=S_2(t)$)

H$_1$: There is statistically significant difference of the survival of the groups (thus, $S_1(t)\neq S_2(t)$)

The test statistic $L=\frac{S}{\sqrt{\mathrm{var}(s)}}$ follows the asymptotically standard normal distribution under the null hypothesis. If S is obtained from group 1, the critical region is $\boldsymbol{L<-Z_\alpha}$. If S is obtained from group 2, the critical region is $\boldsymbol{L>Z_\alpha}$, where $\boldsymbol{\alpha}$ is the significance level for testing **H$_0$** against **H$_1$**. (**Lee, 1992**)

5.5 Peto and Peto's Generalized Wilcoxon Test

Peto and Peto described another generalization of Wilcoxon's Test for two groups in 1972. This test is similar to the Logrank Test, which means that it assigns a score to each observation in both groups, taking into account whether the observation is censored or not. The only difference lies between the scores that are used.

For an uncensored observation *t*, the score is $u_i=\hat{S}(t_i)+\hat{S}(t_{i-1})-1$ and $\hat{S}(t_0)=1$, while for a censored observation t_j^+ the score is $u_j=\hat{S}(t_i)-1$, where $t_i\leq t_j^+$, where $\hat{S}(t)$ is the Kaplan-Meier estimate of the survival function. These generalized Wilcoxon scores sum to zero.

The test procedure after the scores are assigned is the same as for the Logrank Test. In other words, censored observations receive negative scores, uncensored observations receive positive scores and the total scores for the two groups sum to zero. The next step is to calculate the variance of the sum *S* of the *u* scores in one of the two groups which is given by

$$\mathrm{var}(S)=\frac{n_1 n_2\cdot\sum_{i=1}^{n_1+n_2} u_i^2}{(n_1+n_2)(n_1+n_2-1)}=\left[\sum_{j=1}^{n}\frac{m_j(r_j-m_j)}{r_j}\right]\cdot\frac{n_1\cdot n_2}{(n_1+n_2)(n_1+n_2-1)}$$

The two-sided hypothesis test is the same as in the Logrank test and so is the test statistic $L = \frac{S}{\sqrt{\mathrm{var}(s)}}$ which follows the asymptotically standard normal distribution under the null hypothesis. If S is obtained from group 1, the critical region is $\boldsymbol{L<-Z_\alpha}$. If S is obtained from group 2, the critical region is $\boldsymbol{L>Z_\alpha}$, where $\boldsymbol{\alpha}$ is the significance level for testing $\mathbf{H_0}$ against $\mathbf{H_1}$. (**Lee, 1992**)

5.6 Ascendancy over the Tests

All the above tests are based on rank statistics obtained from scores assigned to each observation. They can be further grouped into two categories:

a) Generalization of the Wilcoxon Test (to which category belongs Gehan's and Peto and Peto's)

b) Non-Wilcoxon Test (concerning Cox-Mantel and the Logrank Test). The *S* statistic which equals to the sum of *w* scores in group 2 in the Logrank Test is the same as *U* of the Cox-Mantel Test.

A generalization of the Kruskal-Wallis test, which extends Gehan's generalization test, is proposed by **Breslow** (**1970**) for testing the quality of *k* continuous distribution functions when subjects are subject to arbitrary right censorship. The distribution of the censoring variables is allowed to differ for different populations. An alternative statistic is proposed for use when the censoring distributions may be assumed equal. These statistics have asymptotic chi-squared distributions under their respective null hypothesis, whether the censoring variables are regarded as random or as fixed numbers.

The only reason to choose one test over another in a given circumstance is if it will be more powerful, that is, more likely to reject a false hypothesis. When sample sizes are small (n_1, $n_2 \leq 50$), **Gehan and Thomas (1969)** show that Cox's F test is more powerful than Gehan's generalized Wilcoxon test if samples are from exponential or Weibull distributions and if there are no censored observations or the observations are singly censored. Lee (1975) showed that when samples are from exponential distributions, with or without censoring the Cox-Mantel and Logrank tests are more powerful and more efficient than the generalized Wilcoxon test of Gehan and Peto and Peto.

When the samples are taken from Weibull distributions with constant hazard ratio (i.e. the ratio of the two hazard functions does not vary with time), the results from the Cox-Mantel, the Logrank tests and the two generalized Wilcoxon tests are the same as in the exponential case. However, when the hazard ratio is non-constant, the two generalizations of the Wilcoxon test have more power than the other tests. Thus, the Logrank test is more powerful than the Wilcoxon tests in detecting departures when the two hazard functions are parallel (proportional hazards) or there is random but equal censoring and when there is no censoring in the samples (**Crowley and Thomas, 1975**). The generalized Wilcoxon tests appear to be more powerful than the logrank test for detecting other types of differences, for instance, when the hazard functions are not parallel and when there is no censoring and the logarithm of the survival time follows the normal distribution with equal variance but possibly different means.

The generalized Wilcoxon tests give more weight to early failures than later failures whereas the Logrank test gives equal weight to all failures (**Prentice and Marek, 1979**). Therefore the generalized Wilcoxon tests are more likely to detect early differences in the two survival distributions whereas the Logrank test is more sensitive to differences at the right trails.

There are situations in which neither the Logrank nor the Wilcoxon tests are very effective. When the two distributions differ but the hazard functions or survivorship functions cross, neither the Wilcoxon nor the Logrank test is very powerful and we have to consider other tests. For instance, **Tarone and Ware (1977)** discuss general statistics of similar form (using scores) and **Fleming and Harrington** (**1979**) and **Fleming et al.** (**1980**) present a two-sample test based on Smirnov-type statistic designed to measure the maximum distance between estimates of two distributions. The latter approach is shown to be more effective than the Logrank or Wilcoxon tests when two survival distributions differ substantially for some range of t values but not necessarily elsewhere.

CHAPTER 6

Proportional Hazards Model

6.1 Introduction

Throughout this study, we have only considered survival models that where a function of chronological age, *S(x)*, or those that are function of time since some initial event, *S(t)*. In both cases the model was univariate. The procedure of estimating the survival function for each subgroup and each variable was the Product-Limit method of Kaplan and Meier (1952).

Many cases arise in which survival probabilities are a function of two or more variables, such as those used for insurance premium calculations which depend on age at issue as well as time since issue. For instance, suppose we consider the survival of cancer patients as factor of time since diagnosis. We might believe that type of cancer, sex of patient, and type of treatment all affect survival, so we could estimate a separate *S(t)* for each type/sex/treatment combination. We consider *S(t)* to be univariate, with covariate variables taking into account separately.

We will wish to consider parametric models, in which survival probabilities are determined as a function of both time and the accepted associated variables. In this type of model the associated variables have been taken into account by inclusion rather than by separation. The functional form of the multivariate parametric model should allow the variables to interact in a logical manner, beforehand to testing a proposed model against sample data. That is, the model should be plausible in light of our knowledge of physiology, gerontology and so forth.

Survival analysis typically examines the relationship of the survival distribution to covariates. Most commonly, this examination entails the specification of a linear-like model for the log hazard. In this chapter we discuss a most commonly used model, the Cox proportional hazards model (1972) which does not require knowledge of the underlying distribution. The Cox regression model, or the proportional hazards model, is a statistical theory of counting processes that unifies and extends nonparametric censored survival analysis. The approach integrates the benefits on nonparametric and parametric approaches to statistical inferences. The data in a Cox regression model includes (t_i, z_i), i = 1, 2, …, n, where n is the number of observations in the study, t_i is the time of failure of the ith observation, and z_i is the p-dimensional vector of covariates. In the presence of censoring of data, t_i is replaced by $t_i \wedge c_i$ where c_i is the censoring time for the ith observation. z_i can also be a time-dependent covariate in which case z_i will be replaced by $z_i(t)$. The components of z_i may represent various features thought to affect failure time such as, treatments, virtual properties of the individuals, or, exogenous variables. Further components of z_i may be synthesized to examine interaction effects, in a way that is broadly familiar from multiply regression analysis. Finally, the explanatory variables may be classified also in other ways, in particular as for each individual constant or time dependent. (**Cox DR and Oakes D, 1984**)

6.2 Simple Linear Regression

A simple linear regression model is a model with a single explanatory variable and is represented as $\hat{Y}_i = \beta_0 + \beta_1 \cdot X_i$. In this equation, $\hat{Y}_i$ is the predicted value of Y_i, or the predicted response variable given the value of X_i, the treatment variable. It is worthwhile to consider a simple linear regression model because it captures the essential properties of multiple linear regression models. When making a statistical model it is important to make sure that the underlying assumptions hold. Plotting residuals versus the x values and other residual diagnostics are useful to check the normality of data. Interpretation of censored data must be done carefully because it is not normal and thus complicate the fitting of the distribution. Since failure-time data is almost always censored, one would need to find a model without the underlying assumption of normality.

Consider a set of observations y_i, i = 1, 2,…, n, possibly censored, such that their cumulative distribution functions are $f\left[\frac{y_i - \mu_i}{\sigma_i}\right]$ where μ_i and σ_i are the respective population mean and population standard deviation. Standardized residuals, for this model, are defined to be $\hat{\varepsilon}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}$, where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the maximum likelihood estimates of μ and σ respectively. These residuals should look like a possibly censored random sample from a standardized location-scale distribution, i.e. $\mu = 0$, $\sigma = 1$, and the distribution should be normal or logistic. Residual plots can be made in several different ways. If the residuals are plotted against a certain explanatory variable, the plot can show the variable's explanatory power.

When the observation, y_i, is censored then the residual, $\hat{\varepsilon}_i$, is also censored. Thus for censored data, all we can say is that the actual residual would have been larger than the censored residual. Plotting these standardized residuals, $\hat{\varepsilon}_i$, versus the predicted values of $\hat{y}_i$ will help detect nonlinearity, if the data is not heavily censored.

A cumulative distribution function, c.d.f., is often a common way to summarize and display data. If one plots the c.d.f. versus x, the graph produced will provide information on percentiles, dispersion, and general features of the distribution of the data. The c.d.f. can also be the basis for construction of goodness of fit tests for the hypothesized probability models. (**Lindsay Smith, 2004**)

6.3 Proportional Hazards Model

Assume a set of explanatory variables denoted by z, which represents a collection of predictor variables that is being modelled to predict individual's hazard. Let an arbitrary hazard rate be

$$h(t, z) = h_0(t)\psi(z,\beta)$$ (**Cox & Oakes,1984**)

where, $h_0(t)$ is an arbitrary unspecified base-line hazard function for a continuous t common to all individuals, under the standard conditions z=0, $\psi(z,\beta)$ is the expression of the explanatory variables contained in the vector z, β is a vector of regression parameters expressing the strength of dependence of distribution of t on z, and $h(t, z)$ is the hazard function at time t for an individual with covariates z. The density function, $f(t)$ is

$$f(t,z)=h(t,z)S(t,z)$$

where $S(t)$ is the survival function defined by

$$S(t,z) = \exp(-\int_0^t h_0(u)\psi(z,\beta)du) = \left[S_0(t)\right]^{\psi(z,\beta)}$$

and $S_0(t) = \exp\left[-\int_0^t h_0(u)du\right]$ represents the generator of Lehmann family.

Then we can see that the survivor function of *t* for a covariate value, *z*, is obtained raising the base-line survivor function, S_0 (*t*), to the power $\psi(z,\beta)$.

The regression coefficients, β, may or may not be estimated with assumptions made about the hazard function. If β is estimated with assumptions made about the hazard function then one would maximize the likelihood functions and would consider contributions made to the hazard rate by censored data.

Covariates act multiplicatively on the hazard function. If $h_0=h$, our hazard function reduces to the exponential regression model. The Weibull distribution is a special case where $h_0(t) = hp(ht)^{p-1}$. The conditional density function of *t* given *z* is

$$F(t,z) = h_0(t)\psi(z,\beta)\exp\left[-\psi(z,\beta)\int_0^t h_0(u)du)\right]$$

There are three parameterizations for the expression $\psi(z,\beta)$:

- The log-linear form $\psi(z,\beta)=exp(\beta z)$ and $h(t, z) = h_0(t)\ exp(\beta z)$
- The linear form $\psi(z,\beta)=1+\beta z$ and $h(t, z) = h_0(t)(1+\beta z)$
- The logistic form $\psi(z,\beta)=log(1+exp(\beta z))$ and $h(t, z) = h_0(t)\ log(1+exp(\beta z))$

6.4 Cox's Proportional Hazards Model for Survival Data

Let $x_1,x_2,\dots x_p$ be the possible prognostic variables (covariates or explanatory variables) and for the i_{th} patient observed values of the p variables are $x_{1i},x_{2i},\dots x_{pi}$. In multiple-regression approach, the independent variable is the survival time of the i_{th} patient and a function of t_i and $x_{1i},x_{2i},\dots x_{pi}$ let be $w(t_i)=f(x_{1i},x_{2i},\dots x_{pi})=\exp(\beta_1x_1+\beta_2x_2+\dots+\beta_px_{pi})$

Regression models proposed for survival distributions generally involve the assumption of proportional hazard functions. A proportional hazards model possesses the property that different individuals have hazard functions that are proportional to one another, that is $\frac{h(t/x_1)}{h(t/x_2)}$ the ratio of the hazard functions for two individuals with covariates $x_1=(x_{11},x_{21},\dots,x_{p1})$ and $x_2=(x_{12},x_{22},\dots,x_{p2})$ is invariable to time t. This implies that the hazard function, given a set of covariates $x=(x_1,x_2,\dots,x_p)$ can be written as h(t/x)=h$_0$(t)g(x) where g(x) is a function of x and h$_0$(t) can be considered as a baseline hazard function of an individual for whom g(x)=1. (**Lee,1992**)

When survival times are continuously distributed and the possibility of ties can be ignored, the form of the proportional hazards model is

$$h(t/x)=h_0(t)\exp(\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)=h_0(t)\exp\sum_{j=1}^{p}\beta_jx_j$$

where h$_0$(t) is the hazard function of the underlying survival distribution which expresses the time dependent part, where all the x variables are ignored, that is, all x's equal zero, and β's are regression coefficients.

A particular form of h$_0$(t) is needed for the estimation of its parameters, but it is impossible since the form is unknown. So, it would be more convenient if the assumption of the particular form of h$_0$(t) was unnecessary. This approach used by Cox in 1972 when he proposed the Cox's Proportional Hazards Model, which is non-parametric model with respect to time, but parametric in terms of the covariates, which uses the hazard function as the dependent variables. (**Armitage et al,1994,Lee,1992**)

It is clear that Cox's model assumes that the hazard of the study group is proportional to that of the underlying survival distribution h$_0$(t). It can be shown that it is equivalent to

$$S(t)=[S_0(t)]^{\exp\sum_{j+1}^{p}\beta_jx_j}$$

6.5 Regression Model

If we apply the logarithm to the Cox proportional hazard model

$$\log_e\frac{h_j(t)}{h_0(t)}=\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_px_{pi}=\sum_{j=1}^{p}\beta_jx_j$$

we have a standard multiple-regression model with the prognostic variables as independent variables and a function of the hazard as the dependent variable.

Our target is to identify important prognostic factors, which means to identify from the *p* independent variables a subset of variables that relate significantly to the hazard and consequently the length of survival of the patient. So, we will examine the hypothesis H$_0$: $\beta_i=0$, there is not a significance correlation between the independent variable and the survival of the object over H_1: $\beta i \neq 0$, there is a significance correlation between the independent variable and the survival of the object. In this method, we can select the most related independent variables to survival of the objects, while in the Cox's regression model we can define a prognostic index of the ratio $\log_e \frac{h_j(t)}{h_0(t)}$ that can be used to compare prognosis between objects, that is, to compare the relative risk for objects with different values of the independent variables. In this case the model which can be used is: $\log_e \frac{h_j(t)}{h_0(t)} = \beta_1(x_1 - \overline{x}_1) + \beta_2(x_2 - \overline{x}_2) + ... + \beta_p(x_p - \overline{x}_p)$ where $\overline{x}_j$ is the average of the j^{th} independent variable for all objects, then $h_0(t)$ is the hazard function when all variables are at their average values (**Lee, 1992**).

6.6 Accelerated life model

Suppose that there are two eventualities represented by values 0 and 1 of the explanatory variable z. Let the survival function at z=0 be $S_0(t)$; in the accelerated life model there is a constant ψ such that the survivor function at z=1, written variously $S_1(t)$ is $S_1(t)=S_0(\psi t)$ so that,

$$f_1(t)=\psi \cdot f_0(\psi t) \quad \text{and } h_1(t)=\psi \cdot h_0(\psi t)$$

A stronger version is that any individual having time t/ψ under z=1, i.e. the corresponding random variables are related by $T_1=T_0/\psi$.

In general, with the arbitrary constant vector z of explanatory variables, suppose that there is a function ψ(z) such that the survivor function, density and hazard are respectively

$$S(t,z)=S_0(t\psi(z))$$
$$f(t,z)=f_0[t\psi(z)]\,\psi(z)$$
$$h(t,z)=h_0[t\psi(z)]\,\psi(z)$$

If $S_0(t)$ refers to the standard conditions z=0 then $\psi(0)=1$. A representation in terms of random variables is $T=\frac{T_0}{\psi(z)}$, where T_0 has Survivor function $S_0(t)$. If $\mu_0=E(\ln T_0)$, we can write this as $\ln T=\mu_0-\ln\psi(z)+\varepsilon$, where ε is a random variable of zero mean with distribution not depending on z. In problems with a limited number of distinct values of z, it may be unnecessary to specify $\psi(z)$ further. In other contexts, a parametric form $\psi(z)$ may be needed; we then write $\psi(z;\beta)$. Since $\psi(z;\beta)\geq 0$, $\psi(0;\beta)=1$, a natural candidate is

$$\psi(z;\beta)=e^{\beta^{\mathrm{T}}z}$$

where the parameter of vector β is $q\ x\ 1$. Then $lnT=\mu_0-ln\psi(z)+\varepsilon$, can be written as a linear regression model

$$lnT=\mu_0-\beta^T z+\varepsilon.$$

Note for the comparison of two groups, with a single binary explanatory variable we get

$$\psi(z;\beta)=e^{\beta}$$ **(D.R Cox and D. Oakes, 1984)**

CHAPTER 7

Application of Survival Analysis in Economics

7.1 Introduction

The first use of survival analysis and duration models comes from medical research. Survival analysis involves the modelling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature. Although at the beginning the survival analysis was used to study death as an event specific to medical studies and demographical studies, as from the '70s these statistical techniques have been increasingly used in economics and social sciences. In the area of labor economics, for instance, employment durations are treated as survival times and analyzed accordingly (**Heckman and Singer, 1985; Kiefer, 1988; Lancaster, 1990**). Recently, survival analysis approaches have been proposed for analyzing medical costs. In the survival analysis approach to cost data, individuals' cumulative costs are treated like survival times and analyzed accordingly (**Dudley et al., 1983; Fenn et al., 1995, 1996**).

We explain above the assumptions necessary for a survival analysis to be valid and show how they might be violated when survival analysis is applied directly to possibly censored data on cumulative costs. We present some alternative, nonparametric methods that have been developed, and show how the results of these methods differ from the results of survival analysis in a real costs dataset.

7.2 Survival Analysis

7.2.1 The Kaplan-Meier Curve

The *Kaplan–Meier* or *Product Limit* estimator *KM(t)* estimates the probability that the time-to-event or time-to-failure *T* exceeds any given value *t* (Kaplan and Meier, 1958). It is typically plotted as a function of *t* over the range of times of interest and is a decreasing curve with value 1 at time zero and other values given by:

$$KM(t) = \prod_{i:\, t_i < t} \left(1 - r_{t_i}\right) \qquad (1)$$

where $\{t_{1,}\, t_2, \ldots\}$ are the observed failure times and r_{t_i} is the estimated hazard or risk of failure at time t_i, among all individuals at risk of failure at time t_i.

From expression *(1),* it is clear that the key to an unbiased Kaplan–Meier estimator is an unbiased set of estimators of the hazards r_{t_i} at the observed failure times. With censoring, some individuals may be lost to follow-up before a given failure time t_i, in which case we cannot observe the complete at-risk population at this time. In this situation, the survival analyst estimates the hazard of failure at time t_i by the observed failure rate among those at risk and still under observation at t_i. For this to be unbiased, the individuals at risk and still under observation at t_i must be representative of the population at risk at t_i. Equivalently, the individuals censored before t_i cannot be a selectively high or low risk subgroup. If high-risk individuals tend to be censored, then those remaining will constitute a selective, low-risk sample, fewer events than expected will occur, and the estimated hazard will underestimate the true hazard. This is a case of dependent censoring; the selective censoring effectively induces a correlation between the censoring and failure times. From Equation *(1),* it is clear that underestimating hazards will inflate the Kaplan–Meier curve and lead to overestimation of survival. The reverse will occur if low-risk individuals tend to be censored.

Dependent censoring will occur to some extent in practically any cost-to-event analysis (**Hallstrom and Sullivan, 1997; Lin et al., 1997**). The problem arises because individuals tend to accrue costs at different rates, with those in poorer health using more resources and costing more per unit time. Consequently, individuals censored with low costs will tend to be those accumulating costs slowly, who in turn will tend to be those with lower costs-to-event. In practice the correlation between cost at censoring and cost-to-event may not be so extreme as to cause noticeable bias. However, although this correlation is unobservable, its presence in a real application is evidenced by the example in the next section, which shows inflation of the Kaplan–Meier curve. In theory, unless the mapping from time t to cost accumulated by time t is one to one, some degree of bias is to be expected. This can happen even if there is independent censoring on the time scale.

In Figure 1 the magnitude of the bias of the Kaplan–Meier method applied to costs when individuals accumulate costs at different rates is illustrated. The figure represents a 5-year study, with continuous accrual during the follow-up period. Thus, failure times and censoring times are completely independent in the interval 0 to 5 years. Patients accrue costs at a rate of US$1 per month or US$10 per month, each with probability 0.5. *Figure 1a* and *b* show that the Kaplan–Meier method provides an excellent estimate of survival on the time scale, but that the methodology applied to costs can lead to substantial overestimation. The degree of bias is a function of the amount of censoring and the heterogeneity of the cost accrual rates. For instance, if individuals accumulate costs at a rate of either US$1 or US$2 (rather than US$10) per month, then the Kaplan–Meier estimate of the cost-to-event distribution *Figure 1c* shows only slight bias compared with *Figure 1b*. This is a result of the fact that the correlation between costs at censoring and costs at failure is 0.56 in the example depicted by *Figure 1b* and only 0.25 in *Figure 1c.*

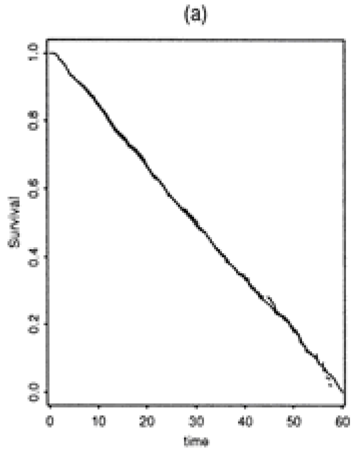

***Figure 1*a** Kaplan–Meier estimate of the survival distribution.

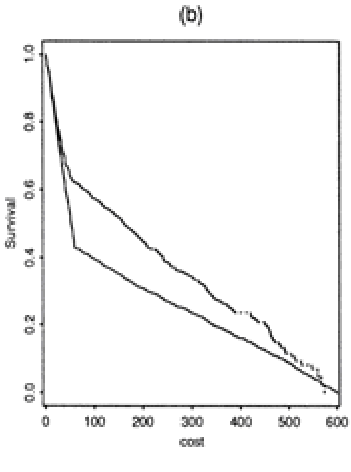

***Figure 1*b** Kaplan–Meier estimate of the distribution of costs to event. Fifty percent of cases accrue costs at a rate of US$1 per month and the rest accrete costs at a rate of US$10 per month.

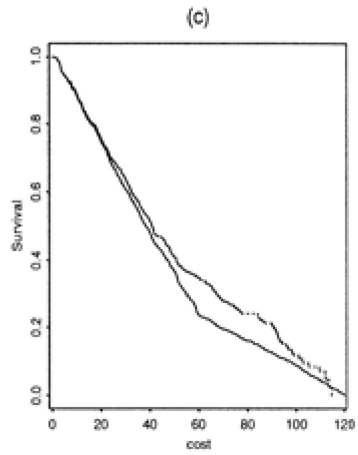

***Figure 1*c** Kaplan–Meier estimate of the distribution of costs to event. Fifty percent of cases accrue costs at a rate of US$1 per month and the rest accrue costs at a rate of US$2 per month.

A key feature of the previous example is that the maximum censoring time (5 years) is at least as large as the maximum failure time. In other words, the follow-up period is sufficient to cover the entire range of possible failure times. In practice, this is not always the case. **(Ruth D. Etzioni et al 1998)**

7.2.3 Cox Regression

The Cox model is based on a modeling approach to the analysis of survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival. The Cox model is a well-recognised statistical technique for analysing survival data. When it is used to analyse the survival of patients in a clinical trial, the model allows us to isolate the effects of treatment from the effects of other variables. The model can also be used, *a priori*, if it is known that there are other variables besides treatment that influence patient survival and these variables cannot be easily controlled in a clinical trial. Using the model may improve the estimate of treatment effect by narrowing the confidence interval. Survival times now often refer to the development of a particular symptom or to relapse after remission of a disease, as well as to the time to death.

The Cox model is a description of the dependence of the risk of failure at any time *t* on the covariates *X*. It is semi-parametric in that no assumptions are made about how the hazard rates vary with time; however, the hazards for different covariate values are assumed to be proportional with a ratio that is constant over time.

Since classical Cox regression relates the hazard at each time t to covariates, the model applied to costs relates the hazard at each cumulative cost c, to covariates. For illustration, consider a binary covariate X taking values 0 and 1. Suppose, for the sake of the of discussion, that X is tumour stage at diagnosis in cancer patients; X=0 is localized and X=1 is metastatic disease. Suppose that the hazard for metastatic disease is a factor a times the hazard for localized disease. The hazard ratio a is termed the '*relative risk*'. A relative risk of 2 in a cost analysis would mean that for metastatic cases, the hazard at any cost c, in terms of events per person-dollars at risk, is twice that for localized cases. This is not in itself a useful quantity, although it indirectly addresses the questions that are usually of interest in cost analyses so long as the Cox regression methodology is valid. These include the following:

1) Overall, how do the costs for localized and metastatic disease compare;
2) For a specific time-to-event, how do the costs compare, and
3) What is an estimate of the marginal cost difference between the two groups?

For Cox regression to be unbiased, independent censoring is required within groups formed by each level of the covariate X so that individuals still under observation are representative of the population at risk in each group, and observed events occur at the correct rate within each group. If censoring is dependent, the observed event rates in each group will be biased. If the dependent censoring mechanism is the same for all levels of X, then the estimate of the relative risk may still be unbiased; the errors caused by dependent censoring within each group may, in a sense, cancel out. However, if, for example, individuals at risk of failure are censored more often when X=1, the observed failure rate for this level of X will be correspondingly lower and as a result, the relative risk a will be underestimated.

In practice, when using Cox regression for cost analysis, the accrual of costs at different rates leads to dependent censoring within subgroups defined by covariate levels. Covariates that affect the rate of cost accrual may lead to differential dependent censoring across groups. To demonstrate the bias that can arise when using Cox regression to analyze costs, we simulated a situation where for Xs0, survival is exponential with mean 20 months, and costs accrue at a rate of either US\$1 or US\$10 per month, each with probability 0.5. For Xs1, survival is exponential with mean 10 months, and costs accrue at a rate of either US\$2 or US\$20 per month, each with probability 0.5. This leads to a proportional hazards model in costs, with a true relative risk of 1.0, since the increased rate of cost accrual is exactly balanced by the higher event rate when *X*=1.

Assuming independent censoring in time with censoring times uniformly distributed in the range 0 to 20 months, the mean estimated relative risk over 100 simulations with 500 subjects per group is 1.2 with a standard deviation of 0.1. Thus, analysis of data from such a model would lead to the conclusion that the costs are lower for *X*=1, which is not the case.

In this example, the different rates of cost increase in the two groups imply differential dependent censoring in costs with independent censoring in time. A confirmation of this is the observation that the correlation between the cost at censoring and the cost at failure is higher in general for *X*=1 than for *X*=0. Consequently, the relative risk estimate is biased. In practice, the degree of bias will differ from one analysis to another, and will depend, among other things on the amount of censoring and the differential in survival and rates of cost accrual in the different groups. When comparing costs in two groups, bias will tend to be greater when the Kaplan–Meier estimate is biased only for one group than when the estimates for both groups are biased in the same direction. For example, bias will occur when rates of cost accrual are highly variable in one group and less so in the other.

Even if it is suspected that dependent censoring will not impact too severely on the bias of the estimated relative risk, the proportional hazards assumption will not in general be satisfied when costs are increasing at different rates. Consider a simple model where, for low *X*, costs accumulate at a rate of US$1 per month with probability *p*, or 10 per month with probability *1-p*.

CHAPTER 8

Application of Survival Analysis in Social Work

8.1 Introduction

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. Statistical analysis of longitudinal data, particularly censored data, lies at the heart of social work research, and many of social work research's empirical problems, such as child welfare, welfare policy, evaluation of welfare-to-work programs, and mental health, can be formulated as investigations of timing of event occurrence. Social work researchers also often need to analyze multilevel or grouped data (for example, event times formed by sibling groups or mother-child dyads or recurrences of events such as reentries into foster care), but these and other more robust methods can be challenging to social work researchers without a background in higher math.

Social work research often involves some measurement of the time elapsed to a particular event, not particular event, not necessarily death, such as the time from admission into residential care to the date of discharge, or from referral to the end of an intervention. In many studies, while researchers would wish to have completed times for all sample members, it is often the case that a report has to be completed or the research terminated before the event of interest (e.g. discharge or termination) has occurred for every sample member. Conventionally, the analyses of such data sets exclude the unfinished cases or simply summarize the number of unfinished cases.

In social work, it is commonplace to find situations where there are many different times of entry into a program and many different times of exit from it, but as indicated in the preceding paragraph, conventional approaches either cannot use all of the information about time that they have available or have to wait, sometimes for lengthy periods, to encompass all of the data. In this respect, time lapse data is rather different from other data, where the absence of the score or value may negate its inclusion in the analysis. This is wasteful and unnecessary. Survival analysis encompasses the basic realization that time, once passes, is a known quantity, thus it can use all available data in the calculation of the overall effectiveness of an intervention.

8.2 The Kaplan-Meier method

The Kaplan-Meier method, as we already know, is the most widely-used method of estimation in survival analysis. It enables a researcher to construct an estimated survival curve using all of the available data including censored times. The survival curve is a plot of the probability of survival (or survival function) against time. The survival curve starts at 0 or 'time zero' where it includes all of the cases, and makes a step change downwards each time an event occurs. However, it is possible, if too crude a measure of unit of time is used, for there to be cases with the same times which compromises the accuracy of the estimate of the true survival curve. As the significance tests between groups are based on ranked time data, it is wise to choose a unit of time small enough to avoid tied observations.

The survival curve is obtained by multiplying together the individual probabilities of an individual case surviving (i.e. not observing an event in a particular time interval) given survival to the beginning of that interval. Consider an individual case observed over a year and let *p1* be the probability of surviving (i.e. remaining registered) the 1st day, *p2* be the conditional probability of surviving (i.e. remaining registered) on the 2nd day given survival in the 1st day, and so on to *p365*. These are known as conditional probabilities, since they are calculated conditional on previous events occurring. The overall probability of surviving (remaining registered) for 100 days is obtained by multiplying these probabilities together to give the survival probability *P(100)*, say, as:

$$P(100) = p1 \times p2 \times p3 \times \ldots \times p100$$

where, for example, p100 is estimated by the number of individuals who survived to 99 days and who also survived to day 100/number of individuals who have survived to day 99. Note, that if no event (deregistration) occurs on day 100, the conditional probability estimate will be 1 and the estimated survival probability will not change. It is important to note also that censored values at 100 days do not affect the conditional probability estimate; however, these are subtracted from the number to 100 days in the estimation of p101. Survival curves are usually summarized by quoting centiles such as the median (50th centile) and quartile (75 per cent, 25 per cent) times to event. Alternatively, a number of days may be fixed and the proportion of surviving individuals can be estimated from the survival curve. This is commonly used in medicine, where it may be used to estimate the five year survival rate for various types of disease.

8.3 The application of the method

Survival analysis was used in a study which sought to indentify patterns of variation in child protection registration practice (**Pugh, 2003)**. Data were collected on the length of time spent on the register of every child in Wales who was on the register at any point during the year from 1 April 1999 until 31 March 2000. The unit of measurement used was the number of days that a child was registered. In addition, further information was collected about each child, including age, gender and category of registration. The sample comprised children already on the register at the start of the study period together with those who were newly registered during the period. Thus, the data set included the:

- Total times of registration for those children who were already on the registers at the start of the year and were then registered during the year
- Total times of those children who were registered and subsequently deregistered within the period of study
- Elapsed times since registration for those children who remained on the register at the end of the year.

Thus, total time refers to the time registered in a completed case, that is one where deregistration has taken place, while elapsed time refers to those unfinished cases where the child remained registered at the end of the study period.

In this study, total times were available for 1,627 children, while the elapsed (censored) times were available for the remaining 2,042 children. Thus, the inclusion of the elapsed times into the analysis considerably increases the size of the sample to 3,669 children registered at some point during the year.

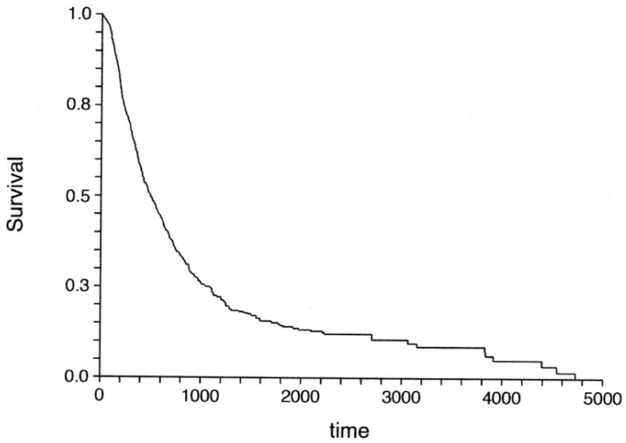

***Figure 1**: Survival curve for all child protection registrations*

The plot of the estimated survival curve with its initially steep drop shows that approximately 70 per cent of all registrations end before 1,000 days. In comparison, the latter section of the curve, with its marked steps, indicates a comparatively small number of cases, with the short plateau representing periods of time in which no de-registrations took place.

A comparison of the quartiles and medians obtained by using only the available total times and those estimated using all of the data (total and elapsed times), provides a striking illustration of the effect of this inclusion.

For example, the median for the total times is 256 days, whereas the median for the survival curve is considerably higher at 494 days.

	25 per cent quartile	***Median***	***75 per cent quartile***
Total times	147	256	473
Survival curve	223	494	1067

***Table 1**: Difference between quartiles and median calculated only on completed registrations and those estimated using all registrations*

In this study the number of completed cases with total times represented approximately 44 per cent of the survival sample of 3,669 cases and the effect of including the elapsed times for the other 56 per cent of registrations arguably provided a better picture of current registration activity, where the majority of registrations are significantly longer than might be expected if the calculations were based solely upon completed cases. **(R.Pug 2004)**

CHAPTER 9

Application of Survival Analysis in Event History

9.1 Introduction

Event history analysis is used to study the duration until the occurrence of an event of interest, where the duration is measured from the time at which an individual becomes exposed to a ‘risk’ of experiencing the event. Therefore, an event history is a longitudinal record of the timing of the occurrence of one or more types of event. Examples include employment histories which typically include dates of any changes in job or employment status, and partnership histories which usually include the start and end dates of co-residential relationships. In an analysis of employment histories events of interest might be the end of an employment or unemployment spell, while a study of partnership histories, such as demography, might examine entry into marriage and marital dissolution. Demographers focus more specific on births , child mortality of children in the same family, occupational careers of spouses, educational careers of brothers, marriages, divorces and migration. In a marriage example, an event history model concerns a person’s marriage rate during the period that he or she is in the state of never having been married.

The techniques used in event history analysis are also commonly known as *survival analysis*, *duration analysis* or *hazard modeling*. Although often used in turn with survival analysis, the term *event history analysis* is used primarily in social science applications where events may be repeatable and an individual’s *history* of events is of interest.

9.2 State, event, duration and risk period

In order to understand the nature of event history data, during a marriage for instance, and the purpose of the analysis we have to understand the: state, event, duration and risk period.

State: Is the category of the “dependent” variable, the dynamics of which we want to explain. At every particular point in time, each person occupies exactly one state. In the analysis of marital histories, four states are generally distinguished: never married, married, divorced, and widowed. The set of possible states is sometimes also called the *state space*.

<u>Event</u>: Is a passage from one state to another, that is, from an origin state to a destination state. In our example or marital history, a possible event is “first marriage”, which can be defined as the transition from the origin state, never married, to the destination state, married. Other possible events are: a divorce, becoming a widow(er), and a non-first marriage. It is important to note that the states which are prominent determine the definition of possible events. If only the states married and not married were distinguished, none of the above-mentioned events could have been defined. In that case, the only events that could be defined would be marriage and marriage breakup.

<u>Risk Period</u>: Is the period that someone is at risk of a particular event, or exposed to a particular risk. Not all persons can experience each of the events under study at every point in time. To be able to experience a particular event, one must occupy the origin state defining the event, that is, one must be at risk of the event concerned. For example, someone can only experience a divorce when he or she is married. Thus, only married persons are at risk of a divorce. Furthermore, the risk period(s) for a divorce are the period(s) that a subject is married. Moreover, another related concept is the *risk set*. The *risk set* at a particular point in time is formed by all subjects who are at risk of experiencing the event concerned at that point in time.

<u>Duration</u>: Is the duration of the nonoccurrence of an event during the risk period. For instance, when our interest focuses on “first marriage”, the analysis concerns the duration of nonoccurrence of a first marriage, in other words, the time that individuals remained in the state of never being married. In practice, the dependent variable in event history models is a transition rate.

9.3 Censoring

An observation is called censored if it is known that it did not experience the event of interest during some time, but it is not known when it did experience the event. We have two types of censoring, right and left. As far as the right censoring is concerned, let assume we have a first-birth situation, a censored case could be a woman who is 30 years old at the time of interview (and has no follow-up interview) and does not have children. For such a woman, it is known that she did not have a child until age 30, but it is not known whether or when she will have her first child. In left censoring we do not have information on the duration of nonoccurrence of the event before the start of the observation period.

9.4 Time-varying covariates

In hazard models we may also have data on changes in individual characteristics or circumstances over time. For instance, from employment histories collected in the British cohort studies it is possible to determine whether a person is in full-time education at a given point in time. In an analysis of age at first association, we might be interested in the relationship between an individual's probability of partnering at time t and their educational status at that time. Educational record is an example of a *time-varying covariate*. While one approach would be to take the value of such variables at one point in time, such as the start of the observation period, this is wasteful and does not allow us to explore how the timing of an event relates to a *change* in the value of a covariate.

9.5 Proportional hazards model

The important goal of most event history analysis is to identify factors that are associated with the timing of the event of interest. The values of covariates may be fixed over time or time varying. One distinction between models is based on whether event times are assumed to be measured in continuous or discrete time. In this section we consider continuous-time models. Models can also be classified as either proportional hazards or accelerated life models, according to the way in which covariates are assumed to affect the timing of events. The most important consideration when choosing an appropriate model is the nature of the distributional assumption for event times. The most flexible and well known continuous-time model is the Cox proportional hazards model. For each individual i we observe a vector of covariates with values fixed across time. The hazard at time is now a function of t and x_i, which we denote by $h(t,x_i)$. Denote by $h_0(t)$ the hazard at $x_i=0$. If all covariates are categorical, *$h_0(t)$* is the hazard for individuals in the reference (baseline) category of each variable. For this reason *$h_0(t)$* is often referred to as the *baseline hazard*. A proportional hazards model is written as

$$h(t,x_i)=h_0(t)g(x_i),$$

where g(x) is some function of the covariates. If the values of the covariates are changed from their reference categories (or, more generally, from zero) to a value x_j, then the hazard is multiplied by $g(x_j)$. Therefore, the covariates are assumed to have a multiplicative effect on the hazard. The proportional hazard assumption implies that the effect of a change in **x** on the hazard is the same for all values of t. Consider the hazard functions for two different sets of covariate values, x_1 and x_2. In that case, the ratio of the hazards at these two values is

$$\frac{h(t,x_1)}{h(t,x_2)}=\frac{g(x_1)}{g(x_2)}$$

which is independent on *t*.

9.6 Accelerated life model

An accelerated life model is based on the idea that individuals experience time in different units. For example, suppose we wish to compare mortality risks of humans (x=0) and dogs (x=1). Dogs have a shorter lifespan than humans, so dogs are said to age faster than humans. If a year of human life is approximately equal to seven dog years, the relationship between the survivor functions for humans and dogs can be expressed as

$$S(t,x=1)=S(7t,x=0)$$

The Accelerated life model assumes that a change in covariate values from 0 to x_j accelerates time by a factor $g(x_j)$ or, equivalently, reduces the median survival time by a factor $g(x_j)$. While the proportional hazards model assumes that the covariates have a multiplicative effect on the *hazard*, the accelerated life model assumes that covariates have a multiplicative effect on the *timescale*.

In practice proportional hazard models are used far more frequently than Accelerated life models, to the extent that **Hosmer** and **Lemeshow** (1999) state that "It is now accepted as the standard method for regression analysis of survival times in many applied settings." However, it is not always necessary to make the distinction between these two types of models. Both the Weibull model and the exponential model (a special case of the Weibull) can be viewed as a proportional hazard or Accelerated life model, and their parameters interpreted as covariate effects on the hazard or the timescale.

9.7 Cox proportional hazards model

The most commonly applied event history model is the Cox proportional hazards model. In the Cox model, $g(x)=\exp(\beta' x)$, so that

$$h(t,x_i)=h_0(t)\exp(\beta' x)$$

where β is a vector of regression coefficients. One reason for the popularity of the Cox model is its flexibility, the baseline hazard function $h_0(t)$ is left completely unspecified. Another attractive feature of the model is that the exponents of the regression coefficients **β** can be interpreted as *relative risks*.

9.8 Application of age at first partnership

In the example of age at first partnership, we consider the effects of educational inscription and the following time-constant categorical covariates: gender, region of residence, and father's social class. The results from fitting a Cox model are given below. The 95% confidence interval for each relative risk, exp(β), provides a test of the null hypothesis of no effect. The null is rejected at the 5% level if the confidence interval does not contain the value 1. As a result, we find that the effects of gender and educational inscription are both significant at the 5% level. In general, a likelihood ratio test is preferred. A likelihood ratio test is used to compare a pair of nested models. For example, to test the significance of father's social class, we compare the following two models: the model shown in ***Table 1*** and the model without social class. The test statistic is the difference between the -2 log-likelihood values for the two models, which is compared to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between.

	$\hat{\beta}$	exp($\hat{\beta}$)	95% CI for exp(β)
Female	0.398	1.489	(1.220, 1.817)
Region			
Scotland and the North	0.238	1.268	(0.939, 1.712)
Wales and Midlands	0.155	1.168	(0.850, 1.606)
Southern and Eastern	0.081	1.084	(0.765, 1.536)
South East, including London	0	1	-
Father's social class			
I or II (professional and managerial)	-0.288	0.749	(0.549, 1.023)
III	-0.148	0.863	(0.674, 1.104)
IV or V (manual)	0	1	-
Enrolled in full-time education	-1.089	0.337	(0.225, 0.505)
-2 log-likelihood	4253.1		

***Table 1** Results of a Cox proportional hazards analysis of age at first partnership*

The test statistic for comparing models with and without social class is 3.3 on 2 degrees of freedom (p=0.192). We therefore conclude that social class has no effect on age at first partnership. **(Yamaguchi, 1991)**

The hazard of first partnership is almost 1.5 times higher for women than for men, which implies that women partner at an earlier age. Enrolment in full-time education is associated with delayed partnership formation: being in education reduces the hazard of partnering by a factor of (1-0.337)×100%=66.3%. However, we should hesitate to interpret the effect of educational enrolment as causal because the decisions about when to leave education and when to partner are likely to be jointly determined, i.e. enrolment is potentially endogenous.

CHAPTER 10

Application of Survival Analysis in Psychology

10.1 Introduction

Psychiatrists' major aim, as far as the public health is concerned, is the prediction and outcome from the start of severe and potentially repeated or chronic affective illness with psychosis. Suicide is the most venturous consequence of depressive illness (**Jamison KR, 1990**). Mood disorder, attempted suicide and suicide have a lot of common features. Mood disorder, particularly depression (**Guze SB., Robins E., 1970),** psychiatric patient status (**Roy A., 1982**) and attempted suicide (**Nordstrom et al., 1995, Nordentoft et al., 1993**) are well used predictors of suicide risk. Further studies suggest that the suicide risk after attempted suicide is 5-10% within a few years (**Nordstrom et al., 1995, Nordentoft et al., 1993**) and that the long-term suicide risk in depression is 10-15% **(Jamison, 1990 and Guze, 1970**). The assessment and prediction of suicide risk within the group of depressed psychiatric in patients with a high suicide risk is of great clinical concern.

Survival analysis and especially an application of life-table analysis (**Coltont, 1974**) is used to study prediction of suicide risk over time in patient groups with and without a current suicide attempt. Predicting potential suicide in those who have attempted suicide is difficult. There are many risk factors associated with completed suicide, for example: male sex, previous suicide attempts, psychiatric illness, abuse of alcohol, planned attempt and high lethality and intention of suicide attempt (**Arensman E.,Kerkhof A., 1996**). It is not known whether the short-term risk factors are the same as those which can be used to predict the suicide risk in the long term, but there is evidence that they may differ somewhat (**Fawcett J, Scheftner WA, Fogg L. et al., 1990**).

10.2 Data set and variables

The study was based on a database, which was obtained from the emergency unit of Helsinki University Central Hospital. The <u>*sample*</u> was 1018 patients who made 1207 suicide attempts in 1983. Patients were categorized according to their *age*, 54%

were under 35 years old, their *sex*, 53% were women and their *marital status,* 47% were unmarried while 65% belonged to the lowest social classes. With the term "suicide attempt" we mean "an act with non-fatal outcome", that is an individual will cause a self-harm by taking overdose from the prescribed therapeutic dosage (**World Health Organization, 1986**). For the term "physical health" three grades were used: *good*, if no disability was reported, *satisfactory*, if there was a tinny defect which may cause subjective sadness, and *physical disease*, if a chronic disease demands a daily attention.

The estimation of physical severe event, such as death, after a suicide attempt, was based on the patients' physical condition in the emergency room and the medical records until that time of occurrence. A suicide attempt was categorized according to *planned* (non-spontaneous) and *not planned* (spontaneous) into three groups: *certain*, *probable* and *undetermined* based on the intention to die (**Lonnqvist J., 1977).** The common reason in both cases was "the wish to die".

10.3 Survival Analysis

In order to identify the long-term risk factors for suicide among all possible risk factors, a Cox regression model was used. Covariates used were: sex, age, social class, marital status, physical disease, previous suicide attempt, psychiatric treatment and alcohol consumption during the attempt, psychiatric surveillance at the time of attempt, spontaneity of the attempt, physical harm, motive and intention to die.

For selecting the model the forward stepwise (conditional) method was used, the model begins as the baseline model without any variables in it and the variables are added into the model if they meet the selection criterion based on the P-value for the score statistic. We also used several variables, such as sex, age, motive, previous suicide attempt, previous psychiatric treatment and impulsiveness of the attempt as strata.

The cumulative survival curve plot for the patients surviving is given below for each sex:

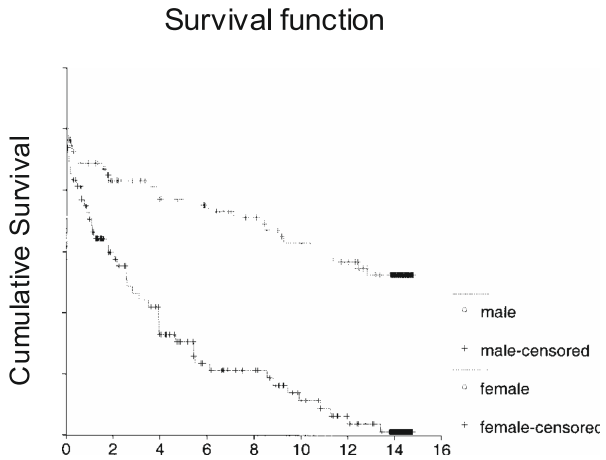

Figure 1: Cumulative proportion of patients surviving (not commiting suicide) after the suicide attempt, males have less prognosis than females. Kaplan-Meier survival curve shows p-value significance 0.015, long-rank statistic 10.07.

From the sample of 1018 patients, 221 individuals who had attempted suicide in 1983, had died. From the rest 797 patients, 68 had committed suicide, 24 women, and 44 men. Most of those had attempted suicide before the research, except for 8 men and 8 women who hadn't done so. In other words, 9.6% of the last had committed suicide during the last years (x^2=14.562, df=1, p=0.000), a frequency of suicide which had increased during the research. More than 50% patients, especially 57.4% had changed the suicide method to a fatal attempt as shown below:

Suicide method	**Both sexes**	**Men**	**Women**
Drug overdose	29	18	11
Carbon monoxide	4	4	0
Hanging	15	10	5
Drowning	1	0	1
Firearms	2	1	1
Jumping under vehicles	12	9	3
Jumping from high places	5	3	2
Total	**68**	**44**	**24**

Table 1: Suicide method for later suicide after suicide attempt by self-poisoning(no gender difference)

A Cox multiple regression model was used to examine the long-term relative suicide risk factors by the end of the follow-up period, as a function of explanatory variables. As indicated in Table 2, the 5 variables predicted suicide are: male sex, previous suicide attempt, somatic disease, subjective motive and previous psychiatric treatment, with their estimated coefficients and their estimated standard errors. The Cox regression model gave the same risk factors using sex as a strata rather than a covariate for the model.

Risks factors	Regression Coefficient	standard error SE	Wald statistic	df	significance	relative risk	95% confidence interval
Male sex	1.036	0.296	12.245	1	0.001	2.82	1.58-5.04
Previous suicide attempt	0.688	0.341	4.066	1	0.044	1.99	1.02-3.88
Somatic disease	1.200	0.597	4.034	1	0.045	3.32	1.03-10.71
Subjective motive: wish to die	1.073	0.287	13.987	1	0.000	2.92	1.67-5.13
Previous psychiatric treatment	0.877	0.409	4.594	1	0.032	2.40	1.08-5.36

Table 2: Cox regression model showing long-term risk factors for suicide

The research was based on the suggestion that there was high suicide risk more than ten years while there may be factors for suicide which may affect the study in the long term period. The suicides continued to occur long after the first episode, especially among women. Over half of the patients changed their fatal method for the last attempt. That suicide risk is higher among males, a general rule in suicidology, which means that male gender may predict suicide in the follow-up study. However, there are studies that have found equal risk **(NIELSEN B. et al 1990, Nordebtoft M. et al., 1993).** The subjective motive “wish to die” also predicted future suicide. We should focus on the reason patients give for their suicide attempt. Moreover, somatic disease also appeared to be a risk factor for suicide. It is possible that these factors may vary depending on the time interval from the first attempt. In other words, we found that during the first year after the first attempt, the risk factors were: male, sex, previous suicide attempt and non-spontaneous attempt. As time goes by, male, sex and previous attempt were still risk factors but there were three more factors to be added: somatic disease, subjective motive and previous psychiatric treatment.

As a result of the study, males have a high suicide risk, especially during the first years, while there was a continuity of suicidal risk for both sexes. Somatic disease appeared to be a long-term risk factor for suicide. Treatment in association with somatic disease and history of previous attempt seems to be very important though the risk remain high for over a decade **(Cavanagh et al, 1999).**

Conclusion

This work presents many applications of survival analysis. Survival analysis provides special techniques that are required to compare the risks for death (or of some other event) associated with different treatments or groups, where the risk changes over time. In measuring survival time, the start and end-points must be clearly defined and the censored observations noted. Kaplan–Meier provides a method for estimating the survival curve and Cox's proportional hazards model allows additional covariates to be included.

REFERENCES

Δημάκη Αικατερίνη (2004) *Ανάλυση Επιβίωσης, Αθήνα, Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών.*

Armitage P, B. G, 1959 *Statistical Methods in Medical Research. Blackwell.*

Armitage P. and Gehan E.A., 1974, Lee, 1992, *Statistical Methods for the Identification and Use of Prognostic Factors*. International Journal of Cancer, 13, p.16-35

Balakrishnan, N., 1991 *Handbook of the Logistic Distribution. Marcel Dekker, Inc*

Berkson J, **1942**, *The Calculation of Survival Rates, in Carcinoma and Other Malignant Lesions of the Stomach*, Edited by W. Walters, H.K. Gray, and J. T. Priestley. W.B. Sauders, Philadelphia

Blossfeld H.P. and Rohwer G., **1995** *Techniques of Event History Modeling. New Approaches to Causal Analysis*, Hillsdale, New Jersey: Lawrence Erlbaum Associates

Breslow N., 1970, *A Generalized Kruskal-Wallis Test for Comparing K Samples Subject To Unequal Pattern of Cencorship*. Biometrika, 57, p.579-594

Cavanagh J., Owens D., Johnstone E. *Suicide and undetermined death in south east Scotland. A case-control study using the psychological autopsy method. Psychol Med 1999;29:1141-1149*

Carroll Nick, **2005** *Explaining Unemployment Duration in Australia http://econpapers.repec.org/article/blaecorec/.*

Cox, D. R., **1972**, *Regression Models and Life Tables. Journal of the Royal Statistic Society, Series B*, p. 34, 187—220.

Cox, D. R., and Oakes, D., 1984, Analysis *of Survival Data. Chapman & Hall, New York.*

Crowley J. **and Thomas D.R., 1975**, *Large Sample Theory for the Log Rank Test*. Technical Report 415. Department of Statistics, University of Wisconsin, Madison, WI

Dudley, R.A., Harrell, F.E., Smith, L.R., Mark, D.B., Califf, R.M., Pryor, D.B., Glower, D., Lipscomb, J., Hlatky, M., 1983. *Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery*. Journal of Clinical Epidemiology 46, p.171–261.

Elisa T. Lee, John Wenyu Wang, 1992, *Statistical Methods for Survival Data Analysis third Edition, Wiley*

Etzioni, R., Urban, N., Baker, M., 1996, Estimating *the costs attributable to a disease with application to ovarian cancer*. Journal of Clinical Epidemiology 49,p. 95–103.

Fenn, P., McGuire, A., Phillips, V., Backhouse, M., Jones, D., 1995. *The analysis of censored treatment cost data in economic evaluation*. Medical Care 33, p.851–863

Fleming T.R. and Harrington D.P., 1979, *Non-parametric Estimation of the Survival Distribution in Censored Data*. Unpublished manuscript

Fleming T.R., O'Fallon J.R., O'Brian P.C. and Hariington D.P., 1980, *Modified Kolmogorov-Smirnov Test Procedures with Appliccation to Arbitarily Right Censored Data.* Biometrics, 36, p.607-626

Fiona Steel, 2005, *Event History Analysis*. Centre for Multilevel Modelling, Graduate School of Education, University of Bristol

Foley, M.C. (1997*).* *Determinants of Unemployment Duration in Russia, Yale Economic Growth Center Discussion Paper* 779:39.

Gehan E.A. and Thomas D.G. 1969, *The performance of some two-sample tests in small samples with and without censoring*. Biometrica, 56, p.127-132

Hallstrom, A.P., Sullivan, S.D., 1997*,* *On estimating costs for economic evaluation in failure time studies*. Medical Care, in press

Heckman, J.J., Singer, D.D., 1985. Longitudinal Analysis of Labor Market Data. Cambridge UK Econometric Society Monographs, 10. Cambridge Univ. Press, Cambridge, UK.

Hosmer, D.W. and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, p.273. New York: Wiley.

Jeroen K. Vermunt and Guy Moors, *Event history analysis,* Department of Methodology and Statistics Tilburg University

Kaplan, E. L., and Meier, P., 1958, *Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association,*p. 53, 457—481.

Kiefer, N.M., 1988, *Economic duration data and hazard functions*. Journal of Economic Literature 26, p.646–679.

Kostaki, **1997** *Διερεύνηση της επίδρασης των δημογραφικών παραγόντων στη μακροχρόνια ανεργία*". Επιθεώρηση Κοινωνικών Ερευνών, τεύχος 94, σ.185-200.

Lancaster, 1990, *The Econometric Analysis of Transition Data. Econometric Society Monographs*, 17. Cambridge Univ. Press, Cambridge, UK

Lin, D.Y., Feuer, E.J., Etzioni, R., Wax, Y., 1997, *Estimating medical costs from incomplete follow-updata*. Biometrics 53, p.113–128

Yamaguchi, K., 1991, *Event History Analysis*. Applied Social Research Methods Series, Vol. 28. Sage: Newbury Park.

Jamison KR, 1990, *Suicide in manic-depressive illness*. **In:** GOODWIN F.K, JAMISON KR, ed. Manic-depressive illness. London: Oxford University Press, p. 227-244.

Guzes B. Robins E., 1970, Suicide and primary affective disorders. Br J Psychiatry 117 p.37-438.

Coltont, 1974, *Statistics in medicine*. Boston: Little, Brown

Roy A.,1982, *Risk factors for suicide in psychiatric patients*. Arch Gen Psychiatry 39, p.1089-1095.

Narendranathan, W and M. Stewart (1993). *Modelling the Probability of Leaving Unemployment:Competing Risks Models with Flexible Baseline Hazards, Journal of the Royal Statistical Society, Series C, Applied Statistics, 42(1*), pp. 63-83.

Nielsen B,Wang AG, Bille-Brahe U. *Attempted suicide in Denmark. IV. A five-year follow-up. Acta Psychiatr Scand 1990;81:250-254*

Nordentoft M, Breum L ,Munnch L. K. et al. *High mortality by natural an unnatural causes: a 10-year follow-up study of patients admitted to a poisoning treatment centre after suicide attempts. BMJ 1993;306:1637-1641*

Nordstrom P., Samuelssom N., Asberg M.,1995, *Survival analysis of suicide risk after attempted suicide*. Acta Psychiatr Scand 91, p.336-340.

Nordentoft M., Rubin P., 1993, *Mental illness and social integration among suicide attempters in Copenhagen*. Acta Psychiatr Scand 88: 278-285.

Arensman E.,Kerkhof A., 1996, *Classification of attempted suicide: a review of empirical studies 1963-1993*. Suicide Life- Threatening Behav, 26, p:46-67.

Fawcett J, Scheftner WA, Fogg L. et al., 1990, *Time-related predictors of suicide in major affective disorders*. Am J Psychiatry, 147, p:1189-1194

Jaana Suokas, Kirsi Suominen, Erkki Isometsa, Aini Ostamo, Jouko Lonnqvist, 2001, *Long-term risk factors for suicide mortality after attempted suicide Findings of a 14-year follow-up study*, p.117

World Health Organization, 1986, Working Group on Preventive Practices in Suicide and Attempted Suicide: summary report. Copenhagen: World Health Organization Regional Office for Europe

Lonnqvist J (1977*)*. *Suicide in Helsinki: an epidemiological and socialpsychiatric study of suicide in Helsinki in 1960-1970*. Monogr Psychiatr Fenn, p.8.