UNIVERSITI PUTRA MALAYSIA

THE USE OF SPLIT EXPONENTIAL AND SPLIT WEIBULL ANALYSE SURVIVAL DATA WITH LONG TERM SURVIVORS

DESI RAHMATINA.

THE USE OF SPLIT EXPONENTIAL AND SPLIT WEIBULL MODELS TO ANALYSE SURVIVAL DATA WITH LONG TERM SURVIVORS

By

DESI RAHMATINA

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

December 2005

To my mother Aminah, my father Gafar in memories and my husband Awiskarni

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

THE USE OF SPLIT EXPONENTIAL AND SPLIT WEIBULL MODELS TO ANALYSE SURVIVAL DATA WITH LONG TERM SURVIVORS

By

DESI RAHMATINA

December 2005

Chairman : Associate Professor Mohd Rizam bin Abu Bakar, PhD

Faculty : Science

The split population model is a flexible way of extending the standard survival analytical methods to failure time data in which susceptibles and long-term survivors coexist. Susceptibles would develop the event with certainty if complete follow-up were possible, but the long-term survivors would never experience the event.

A study was conducted to allow the effects of covariates on the probability that an individual is immune, and the immune probability vary from individual to individual. In effect, we are associating with each individual a distinct probability of being immune, which depends on the covariate information specific to that individual. And then fitted a few models using the maximum likelihood estimation to determine whether the covariates are significant or not. Several popular distributions on the survival data analysis as endorsed by graphical techniques were used.

We applied the split exponential and the split Weibull models together with deviance test, a parametric test for the presence of immunes, and a test for outlier, to test for

sufficient follow-up in the samples where there may or may not be immunes presences. We presented the probability of eventual immune for the *i*th individual as the logit model and logistic model. We will work with two data sets, firstly a Clinical Trial in the Treatment of Carcinoma of the Oropharynx and secondly Stanford Heart Transplant data.

The results from the data analyses for a Clinical Trial in the Treatment of Carcinoma of the Oropharynx data show that the simple exponential model produces a fit not significantly worse than the simple Weibull model and the simple split Weibull model no better than the simple split exponential model, also shown that no evidence of immune population and all covariates are not significant.

The results from the data analyses for Stanford Heart Transplant data show that the simple Weibull model is significantly better than the simple exponential model, and the simple split Weibull model is better than the simple split exponential model. We have calculated the maximum log-likelihood function value for both the logit exponential and logistic exponential models. They are exactly similar for both the Clinical Trial in the Treatment of Carcinoma of the Oropharynx and Stanford Heart Transplant data. So, we suggest that both the logit exponential and logistic exponential models are equally superior.

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

PENGGUNAAN MODEL TERPISAH EKSPONEN DAN WEIBULL PADA ANALISIS DATA HAYAT DENGAN MASA HAYAT LAMA

Oleh

DESI RAHMATINA

Disember 2005

Pengerusi: Profesor Madya Mohd Rizam bin Abu Bakar, PhD

Fakulti : Sains

Model populasi terpisah merupakan kaedah perluasan yang anjal dalam kaedah analisis hayat kepada data masa gagal dimana wujud dua kelompok individu iaitu peka dan kebal. Individu yang peka iaitu individu yang mengalami peristiwa ke atas kajian yang dibuat, manakala individu yang kebal iaitu individu yang tak pernah mengalami peristiwa ke atas kajian yang dibuat.

Suatu kajian telah dibuat kepada kesan kovariat keatas kebarangkalian suatu individu kebal dan kebarangkalian kebal berubah dari individu ke individu. Kami menggabungkan dengan setiap individu kebarangkalian wujudnya kebal, yang bergantung kepada maklumat kovariat khusus pada individu tersebut. Dan kemudian menggunakan beberapa model menggunakan anggaran kebolehjadian maksimum untuk menentukan sama ada kovariat bererti atau tidak. Beberapa taburan yang popular dalam analisis data hayat disokong oleh kaedah gambar darjah telahpun digunakan.

Kami telah menggunakan model terpisah eksponen dan model terpisah Weibull bersamaan dengan ujian sisihan, ujian parameter kepada wujudnya kebal, dan satu

ujian kepada data terpencil, untuk menguji kepada tindakan susulan dalam sampel dimana wujud atau tidak wujud kebal. Kami telah membentangkan kebarangkalian kebal kepada setiap individu dalam model logit dan model logistik. Kami menggunakan dua kumpulan data iaitu data "Clinical Trial in the Treatment of Carcinoma of the Oropharynx " dan data "Stanford Heart Transplant".

Keputusan daripada analisis data Clinical Trial in the Treatment of Carcinoma of the Oropharynx menunjukkan bahawa model simpel eksponen menghasilkan signifikan yang tidak lebih buruk dari model simpel Weibull dan model terpisah simpel Weibull tidak lebih baik dari model terpisah simple eksponen, juga ditunjukkan bahawa tidak terbukti populasi kebal dan semua kovariat adalah tidak bererti.

Keputusan daripada analisis data Stanford Heart Transplant pulak menunjukkan bahawa model simpel Weibull menghasilkan signifikan lebih baik dari model simpel eksponen dan model terpisah simpel Weibull adalah lebih baik dari model terpisah simple eksponen. Kami telah mengira nilai bagi fungsi kebolehjadian maksimum kepada model logit eksponen dan model logistik eksponen. Nilai tersebut adalah sama bagi data Clinical Trial in the Treatment of Carcinoma of the Oropharynx dan data Stanford Heart Transplant. Jadi, kami menunjukkan bahawa model logit dan logistik model adalah serupa.

ACKNOWLEDGEMENTS

Praise be to Allah s.w.t who has given me the permission to write this thesis, and peace be upon His messenger, Muhammad s.a.w .

I am very grateful to my thesis advisor, Associate Prof. Dr. Mohd Rizam bin Abu Bakar, for his guidance and encouragement in this study. His encouragement led me to consider this topic despite my hesitation due to my limited knowledge in survival analysis. His many helpful suggestion and ideas have made the completion of this study possible. His guidance and motivation are gratefully appreciated. I want to especially thank him for his kindness and sincerity and for being a very understanding supervisor.

My great thanks also goes to Associate Prof. Dr. Isa bin Daud, the committee member, for all the complementary help that has been crucial to the success of this research. My great thanks is also for Prof. Madya. Dr. Noor Akma Ibrahim, the committee member, for valuable comments during the development of this research.

My gratitude also goes to the authorities of IRPA projects with vote numbers 54207 and 55222 which are led by Associate Prof. Dr. Madya. Noor Akma Ibrahim, for providing financial support through the Graduate Research Assistance scheme. I would like to also thank the authorities of IRPA project with vote number 54429 which is led by Associate Prof. Dr. Mohd. Rizam bin Abu Bakar, for providing financial support through the Special Graduate Research Allowance Scheme.

My heartiest thanks to Prof. Dr. Fachri Ahmad, rector of Universitas Bung Hatta, Padang and Dr. Mukhaiyar, dean of Faculty Teachership and Science Education, Universitas Bung Hatta, Padang, for giving me the letters of recommendation when I applied to study the master degree at University Putra Malaysia. I am also indebted to Dr.Susila Bahri who encouraged me to pursue a master degree at UPM. To all the staff of the Mathematics Department and Graduate School Office (UPM), my sincere thanks.

Special thanks to my parents and my husband whose prayers have helped me through all difficulties, and I thank my brothers for sharing my ups and downs and giving me moral support.

I would also like to thank many friends in the department of Mathematics for all the help and support during my study years.

May Allah Subhanahu Wata'ala give a lot of rewards.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Split Population Models

In standard survival analysis, data come in the form of failure times that are possibly censored, along with covariate information on each individual. It is also assumed that if complete follow-up were possible for all individual, each would eventually experience the event. Sometimes however, the failure time data come from a population where a substantial proportion of the individuals does not experience the event at the end of the observation period. In some situations, there is reason to believe that some of these survivors are actually "cured" or "long–term survivors" the sense that even after an extended follow-up, no further events are observed on these individuals. Long-term survivors are those who are not subject to the event of interest. For example, in a medical study involving patients with a fatal disease, the patients would be expected to die of the disease sooner or later, and all deaths could be observed if the patients had been followed long enough. However, when considering endpoints other than death, the assumption may not be sustainable if long-term survivor are present in population. In contrast, the remaining individuals are at the risk of developing the event and therefore, they are called *susceptibles*. Examples in which long-term survivors exist can be found in many different areas.

In the field of radiation research, patients with tumors of the neck and head are frequently treated with radiation. The endpoint of particular interest is local

recurrence of the tumor. It has been observed that only between 5% and 50% of patients will experience local recurrences (Taylor, 1995). It is extremely unlikely, if not impossible, that local recurrences will occur later than 5 years after treatment. Therefore, the patients without experience of local recurrences within 5 years after treatment may be treated as long-term survivors.

In criminology, a criminologist may be interested in the probability that an individual will not return to prison after being released. If recidivism is the event of interest, many individuals who are released from prison will not experience the event because one experience of prison is sufficient (Maller and Zhou, 1996).

Examples can be found even in engineering reliability (Meeker, 1987). Usually, the proportion of defective electronic components from a production process is assessed using a life testing procedure. Electronic components will fail the test if they have manufacturing defects, which cannot be detected in a simple inspection. Only a small fraction of electronic components have such defects. If a component is free of the defects, the chance that it will fail under carefully controlled conditions will be virtually zero.

The above examples suggest that long-term survivor exist in the populations under study. However, the long-term survivors can never be identified and as the result of this, they are manifested as censored observations in the data. Except those long-term survivors who withdraw from the study early and are

censored at the time of their withdrawal, the remaining long-term survivors will be censored at the end of the study. Their large censored survival times will usually make the Kaplan-Meier estimate of the survival function level off at the right extreme, a Kaplan-Meier survival curve that levels off or shows a long and stable plateau is deemed to provide empirical evidence of a cured fraction. The use of standard survival analysis for such data would be inappropriate since not all of the long-term survivors can be considered as censored observations from the same population as those that do experience the event (Pierce, Stewart, and Kopecky, 1979; Farewell, 1982).

Split population models are also known as "cure model". The objective of the cure model is to study the survival distribution and cure rate of such a population. In general term, we have an endpoint or event that we are interested in such as death from a specific cause, disease recurrence, or some other type of failure. The failure time or survival time is the time to the occurrence of such an event. In an individual we are interested in whether the event can occur which we shall call incidence, and when it will occur (given that it can occur) which we shall call conditional latency or simply latency. A cure would correspond to an event-free outcome, and the cure rate would be one minus the incidence probability. As in standard survival analysis, we also want to study the effect of covariates on the outcome. An individual's covariates can affect the incidence probability (more or less probability) and/or the latency (earlier or later occurrence) and the effect of the covariates may be different on these two aspects of the outcome.

Split population models in the biometrics literature, i.e., part of the population is cured and will never experience the event, and have both a long history (e.g. Boag 1949; Berkson and Gage 1952) and widespread applications and extensions in recent years (e.g.Farewell 1982; Aalen 1988; Kuk and Chen 1992). The intuition behind these models is that, while standard duration models require a proper distribution for the density which makes up the hazard (i.e., one which integrates to one; in other words, that all subjects in the study will eventually fail), split population models allow for a subpopulation which never experiences the event of interest. This is typically accomplished through a mixture of a standard hazard density and a point mass at zero (Maller and Zhao 1996). That is, split population models estimate an additional parameter (or parameters) for the probability of eventual failure, which can be less than one for some portion of the data. In contrast, standard event history models assume that eventually all observations will fail, a strong and often unrealistic assumption.

Suppose that $F_R(t)$ is the usual cumulative distribution function for recidivists only, and ω is the probability of being subject to reconviction, which is also usually known as the eventual recidivism rate. The probability of being immune is (1-ω), which is sometimes described as the rate of termination. This second group of immune individuals will never reoffend. Therefore their survival times are infinite (with probability one) and so their associated

cumulative distribution function is identically zero, for all finite $t > 0$. If we now define $F_S(t)=\omega\, F_R(t)$, as the new cumulative distribution function of failure for the split-population, then this is an improper distribution, in the sense that, for $0 < \omega < 1$, $F_S(\infty)= \omega < 1$.

Let Y_i be an indicative variable, such that

$$Y_i = \begin{cases} 0; & ith \quad individual\, will \quad never\, fail \quad (immunity) \\ 1; & ith \quad individual \quad will\, eventually \quad fail\, (recidivist) \end{cases}$$

and follows the discrete probability distribution

$$\Pr[Y_i = 1] = \omega$$

and

$$\Pr[Y_i = 0] = (1-\omega).$$

For any individual belonging to the group of recidivists, we define the density function of eventual failure as $F_R(t)$ with corresponding survival function $S_R(t)$, while for individual belonging to the other (immune) group, the density function of failure is identically zero and the survival function is identically one, for all finite time t.

Suppose the conditional probability density function for those who will eventually fail (recidivists) is

$$f(t \mid Y = 1) = f_R(t) = F_R'(t)$$

wherever $F_R(t)$ is differentiable. The unconditional probability density function of the failure time is given by

$$f_s(t) = f(t \mid Y = 0)\Pr[Y = 0] + f(t \mid Y = 1)\Pr[Y = 1]$$

$$= 0\,(1-\omega) + f_R(t)\,\omega = \omega\, f_R(t).$$

Similarly, the survival function for the recidivist group is defined as

$$S_R(t) = \Pr[T > t \mid Y = 1] = \int_t^{\infty} f(u \mid Y = 1)du$$

$$= \int_t^{\infty} f_R(u)du = 1 - F_R(t).$$

The unconditional survival time is then defined for the split population as

$$S_S(t) = \Pr[T > t] = \int_t^{\infty} \{f(u \mid Y = 0)\Pr[Y = 0] + f(u \mid Y = 1)\Pr[Y = 1]\}du$$

$$= (1-\omega) + \omega\, S_R(t)$$

which corresponds to the probability of being a long-term survivor plus the probability of being a recidivist who reoffends at some time beyond t. In this case,

$$F_S(t) = \omega F_R(t)$$

is again an improper distribution function for $\omega < 1$. The likelihood function can then be written as

$$L(\omega, \theta) = \prod_{i=1}^{n} [\omega\, f_R(t_i)]^{\delta_i} [(1-\omega) + \omega\, S_R(t_i)]^{1-\delta_i}$$

and the log-likelihood function becomes

$$l(\omega,\theta) = \ln L(\omega,\theta) = \sum_{i=1}^{n} \{\delta_i [\ln \omega + \ln f_R(t_i)] + (1-\delta_i) \ln[(1-\omega) + \omega S_R(t_i)]\}$$

where δ_i is an indicator of the censoring status of observation t_i, and θ is vector of all unknown parameters for $f_R(t)$ and $S_R(t)$. The existence of these two types of release, one type that simply does not reoffend and another that eventually fails according to some distribution, leads to what may be described as simple split-population model. When we modify both $f_R(t)$ and $S_R(t)$ to include covariate effects, $f_R(t \mid z)$ and $S_R(t \mid z)$ respectively, then these will be referred to as *split-population models.*

Several authors have fitted the model in $F_S(t) = \omega F_R(t)$ to recidivist data through various parametric forms of $F_R(t)$. Schmidt and Witte (1988) consider a great number of possible parameterisations to model their North Carolina datasets, including the log-normal, the exponential and the Weibull distributions. They also consider "standard" parametric survival model, i.e. when all individuals are assumed to be susceptibles (ω=1). They find that all split-population models fit their data far better than the standard model. Rhodes (1989) and Farewell (1986), however, emphasize that, to use the split-population model, the dataset should be extensive enough to distinguish between desisters and persisters.