RESEARCH ARTICLE OPEN ACCESS

Application of Survival Analysis Model on Drug Addiction Dataset

Runumi Devi
Department of Computer Applications
JSS Academy of Technical Education
Noida, Sector-62, U.P
India

ABSTRACT
Survival analysis methods are common in clinical trials and other types of investigation. It is used predominately in biomedical sciences where the interest is in observing time to death either of patients or of laboratory animals. Time to event analysis has also been used widely in the social sciences where interest is on analyzing time to events such as job changes, marriage, birth of children and so forth. The engineering sciences have also contributed to the development of survival analysis which is called "reliability analysis" or "failure time analysis", since the main focus in this field is in modeling the time it takes for machines or electronic components to break down. This paper discusses the use of Survival Analysis Model using Kaplan Meier method to model time until return to drug use for a set of drug addicted patients. Moreover, Cox Regression Model is also being used to investigate the effect of covariates in drug relapsing for a particular set of drug addicted patients.

Keywords:- Survival Analysis, Kaplan-Meier, Cox Regression Model, Covariate

.

I. INTRODUCTION

Survival Analysis represents a set of statistical methods used to estimate lifetime or length of time between two clearly defined events and is sometimes referred to as time to response or time to failure analysis. Survival data is often analyzed in terms of time to an event. Survival analysis can be performed to explore the occurrence of some events in a population of subjects. The time until the event is of interest, which is called the *survival time* or the *failure time*. More often, subjects are not fully observed. The time at which a subject ceases to be observed for some reasons other than failure is called the *censoring time of the object*. Censoring in the observed population makes survival analysis different with other data analysis approaches. Some regression models are developed to explore the relationship between survival explanatory variables and predict outcomes. The Cox proportional hazards model (Cox PH model) is one of these widely applied models.[3]

A. Survival and hazard probability

The two related probabilities used to describe and model the survival data are the survival probability and the hazard probability. The survival probability *S(t)* is the probability that an individual survives from the start time to a specified future time t. This term focuses on not having an event.

Let T represent survival time. We regard T as a random variable with cumulative distribution function P(t)=Pr(T<t) and probability density function p(t)=dP(t)/dt. The more optimistic *survival function* S(t) is the complement of the distribution function, S(t)=Pr(T>t)=1-P(t)[4]

Another representation of the distribution of survival times is the *hazard function* which assesses the instantaneous risk of demise at time t, conditional on survival to that time. The hazard is expressed as:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\left[(t \le T < t + \Delta t) | T \ge t\right]}{\Delta t} = \frac{f(t)}{S(t)}$$

II. SURVIVAL ANALYSIS METHODS

Separate methods of survival analysis are required to analyze durations mainly for three reasons: (i) durations generally follow a highly skewed distribution – some subjects tend to have large and very large duration whereas most will have relatively short; (ii) study of duration requires follow-up and some subjects tend to be lost – they move away, refuse to cooperate further, do not report for follow-up, etc; and (iii) any follow-up is necessarily for a specified period and by the time you terminate the study, some subjects may not have reached to the end-point of interest – they are still alive, still feeding the child, still not recovered, etc. In the last two cases, the duration is incomplete – you only know that the duration is at least that much but do not know exactly how much. For example, if you are observing infants for duration of breast feeding and decide to follow-up 80 children for 6 months, it is possible that 4 are lost midway to follow-up (drop outs) and another 12 still on breast at the end of six-month period. Such values are called censored values. Because of

censoring,statistics such as mean can not be calculated in a standard manner – neither the standard deviation.[6]

A. Kaplan-Meier Model

The Kaplan-Meier curve, also called the Product Limit Estimator is a popular Survival Analysis method that estimates the probability of survival to a given time using proportion of patients who have survived to that time [1]. Kaplan-Meier methods take into account "censored" or incomplete data. Censored observations are incorporated into the analysis up until the time of censoring. The Kaplan Meier analysis makes the assumption that if subjects had been followed beyond the censored time point they would have had the same survival probabilities as those not censored at that time. [1]

B. Cox regression model

A **Cox model** is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. [5][6] It provides an estimate of the treatment effect on survival after adjustment for other explanatory variables.

Survival analysis is concerned with studying the time between entry to a study and a subsequent event (such as death). A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. In addition, it allows us to estimate the hazard (or risk) of death for an individual, given their prognostic variables.[5]

- A Cox model must be fitted using an appropriate computer program (such as SAS, STATA or SPSS). The final model from a **Cox regression analysis** will yield an equation for the hazard as a function of several explanatory variables.
- Interpreting the Cox model involves examining the coefficients for each explanatory variable. A **positive regression coefficient** for an explanatory variable means that the hazard is higher, and thus the prognosis worse. Conversely, a **negative regression coefficient** implies a better prognosis for patients with higher values of that variable.[5]

III. DATASET DESCRIPTION

The UIS Data has been used as data for the generation of Kaplan Meier Model[2]. The goal of the UIS data is to model time until return to drug use for patients enrolled in two different residential treatment programs that differed in length (**treat**=0 is the short program and **treat**=1 is the long program). The patients were randomly assigned to two different sites (**site**=0 is site A and **site**=1 is site B). The variable **age** indicates age at enrollment, **herco** indicates heroin or cocaine use in the past three months (**herco**=1 indicates heroin and cocaine use, **herco**=2 indicates either heroin or cocaine use or **herco**=3 indicates neither heroine nor cocaine use) and **ndrugtx** indicates the number of previous drug treatments. The variable **time** contains the time until return to drug use and the **censor** variable indicates whether the subject returned to drug use (**censor**=1 indicates return to drug use and **censor**=0 otherwise).[2]

TABLE I : DATASET DESCRIPTION

Obs	ID	Age	Ndrugtx	Treat	Site	Time	Censor	Herco
1	1	39	1	1	0	188	1	3
2	2	33	8	1	0	26	1	3
3	3	33	3	1	0	207	1	2
4	4	32	1	0	0	144	1	3
5	5	24	5	1	0	551	0	2
6	6	30	1	1	0	32	1	1
7	7	39	34	1	0	459	1	3
8	8	27	2	1	0	22	1	3
9	9	40	3	1	0	210	1	2
10	10	36	7	1	0	184	1	2

The probability of replase is

time	probability of relapse
0.0	1.0
0.06027397	0.9
0.07123288	0.8
0.087671235	0.7
0.39452055	0.6
0.50410956	0.5
0.5150685	0.4
0.5671233	0.3
0.5753425	0.20000002
1.2575343	0.10000001

Fig1: KAPLAN MEIER RESULT

This interface displays the probability of drug relapse for the dataset used. Here time is in years. Here the probability of drug relapse decreases with time that is, if the treatment is done for more time, chances of drug relapse are less. If the graph of this data is plotted, it is a decreasing slope graph.

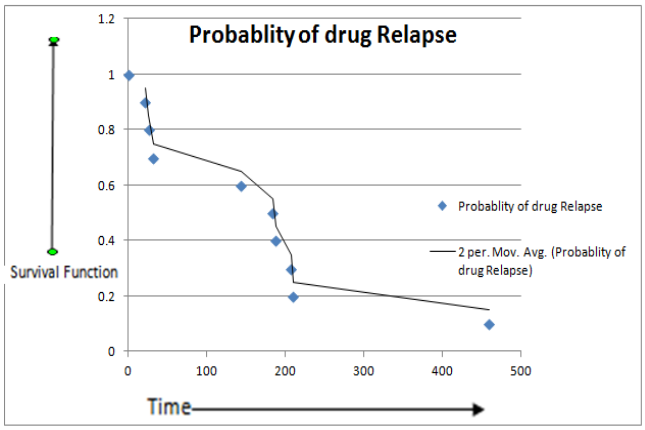

Fig2 : KAPLAN MEIER GRAPH

The graph when plotted taking time on x-axis and probability of drug relapse on y-axis, we get a decreasing slope graph as shown above.

Effect of covariate

The experiment shows the effect of various factors like age, kind of drug and no of drug treatments on the probability of drug relapse. Larger the value of regression coefficient more are the chances of drug relapse.

TABLE II : EFFECT OF COVARIATE

Covariate	Value	Regression coefficient
Age	<=25	0.003709488481062083
	25< age<= 30	0.015623518459456436
	30< age <= 35	0.01624007343523183
	age> 35	0.016383927714389936
Kind of drug used	both heroine and cocaine	0.016765484814097623
	heroine alone	0.016794042260972714
	neither heroine nor cocaine	0.016866903149809487

IV. RESULT ANALYSIS

More is age more is the probability of relapse

The effect of number drug treatment used:

More is the kind of drug used more is the probability of relapse

The effect of various kinds of drug used:

Kaplan Meier model has been implemented successfully to calculate the relapse time of drug addiction after the treatment is stopped for a particular group of drug addicted patient. The graph obtained also has a decreasing slope as in the standard Kaplan Model Graph[fig3]. According to the graph at 100 days the probability of relapse is 70%, at 200 days its around 35%, at 300 days it decreases to about 28% while at 500 days its around 10%. Cox Regression Model is also being implemented to estimate the effects of covariates like age, no. of drug treatments and type of drug intake on the probability of drug relapse. More age more probability, less no. of drug treatment more probability. More is the kind of drug used more is the probability

V. FUTURE APPLICATIONS

Survival analysis model can have varied applications not only in the field of medical science but also engineering as well as population dynamics. In medical sciences it can also be used to find the survival rate of other life threatening diseases such as Brain cancer, Cervical Tumor, etc. from medical data. In engineering it can be used to find the fatigue limit of several machines

REFERENCES

[1] John Ventre, United Biosource Corporation, Blue Bell, PA and Lisa Fine, United Biosource Corporation, Ann Arbor, MI, "A Programmer's Introduction to Survival Analysis with Kaplan model", Paper CC16 – PharmaSUG 2011

[2] Statistical Computing Seminars Survival Analysis with Stata, "http://www.ats.ucla.edu/stat/sas/seminars/stata_survival/.[UIS data].

[3] John Fox(Lecture Notes), "Introduction to Survival Analysis", Sociology 761, 2006.

[4] John Fox, "Cox Proportional-Hazards Regression for Survival Data", Appendix to An R and S-PLUS Companion to Applied Regression, February 2002

[5] Stephen J Walters, "What Is Cox Model?" , Statistics, Second Edition, May 2009.

[6] Indrayan, A., and A. K. Bansal. "The methods of survival analysis for clinicians."Indian pediatrics 47.9 (2010): 743-748..