International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2018; 3(1): 167-170
© 2018 Stats & Maths
www.mathsjournal.com
Received: 21-11-2017
Accepted: 22-12-2017

Vinay Kumar
Assistant Professor,
Department of Statistics,
Ram Lal Anand College,
University of Delhi,
South Moti Bagh, New Delhi,
Delhi, India

Correspondence
Vinay Kumar
Assistant Professor,
Department of Statistics,
Ram Lal Anand College,
University of Delhi,
South Moti Bagh, New Delhi,
Delhi, India

Some non-parametric survival models for epidemiological studies

Vinay Kumar

Abstract
Over the past few centuries the importance of survival analysis has been felt in many areas of sciences including demography and actuarial sciences. The recent and major developments in survival analysis are due to various clinical trials in chronic diseases and actuarial science. Hence this article is focus on the various non-parametric survival models which are very vital in the epidemiological studies.

Keywords: Kaplan-Meier, weighted empirical survival function, nelson-aalen estimator

1. Introduction
Survival analysis basically refers to a set of statistical procedures that are used to measure the time lapse between the entry of the observations under study and occurrence of event. This collection of statistical procedure is sometimes also known as time-to-event analysis. Time with the occurrence of any particular event eventually has a great significance in biological, medical and reliability studies. This time indicator will basically tell the duration of time that will take from beginning of any follow-up and the occurrence of an event. The time lapse between the starting point and the end point is the outcome variable of interest. In the medical research, the outcome variable or the desired outcome of interest is may be the recurrence of symptoms, death of a patient, relapse from remission, relief from pain, incubation of various diseases like Hepatitis B, AIDS, etc., disease incidence, in clinical traits remission duration of certain disease (Andersen, 1992; Kalbfleisch and Prentice, 1980; Cox and Snell 1968; Cox and Oakes, 1984; Crowley and Hu, 1977; Jenkins, 1997; Miller, 1981; Clayton, 1978) [1, 15, 5, 8, 9, 13, 16, 3]. The survival analysis technique can be used in the fields where data have to analyze regarding the duration between the two events. Therefore, survival analysis is also known as life time data analysis, time to event analysis or event history analysis.
Reports about life tables have been written by many researcher, including Gehan (1969) [10], Berkson and Gage (1950) [2]. The statistical methods related to clinical trials have been suggested by Peto *et al*. (1976). The most thoughtful impact on the development in the field of clinical trials is the Kaplan-Meier method used for estimating the survival function. This method provides an inefficient estimate of the survival probabilities and is not a reliable estimator in case of heavy censoring. Also, at the end points, the Kaplan-Meier survival curve fails to provide the reliable estimates.
In case of heavy censoring, an improved method of Kaplan-Meier estimate, which is known as Weighted Kaplan-Meier method of estimation, is used to have reliable estimates. This method provides reliable estimates when applied as it introduces the weights which are based on non-censored rate. In case of assigning weights to latest observation of censoring, a modified form of Weighted Kaplan-Meier method of estimation is followed up which is named as Modified Weighted Kaplan-Meier method of estimation. Huang (2008) [11] used the Weighted Empirical Survival Function (WESF) in which the choice of weights for obtaining the survival function is introduced. Later on, for obtaining the survival function, the very well remarked Nelson-Aalen estimate is used to establish an interrelationship between cumulative hazard function and the survival function.

2. Statistical Models

Let 'T' be the positive random variable associated with the time of occurrence of an event of interest. In survival analysis, the hazard and the survival functions are the two main functions of investigation by the researchers. Here, S(t) usually denotes the survival function which estimates the probabilities that the subject (or patient) will survive for equal to or greater than some specified time 't'. In this case, the survival function is:

Where T is the time till survival of the subject

$S(t) = P(T > t);\ t > 0$

If $F(t)$ is the cumulative distribution function of t, then,

$$F(t) = P(T \le t) = 1 - S(t)$$

The survival function S(t) is monotonically non-increasing at time S(0) = 1and S(∞) = 0. The function $S(t)$ is also known as the cumulative survival rate. The graph of $S(t)$ on time is called as the survival curve.

The hazard function at time t is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P((t \le T < t + \Delta t) / T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$

The hazard function is sometimes also known as a conditional mortality rate or the instantaneous death rate. Some of the important features of the hazard function are, h(t) may decrease, remain constant or it may increase or can follow up any other pattern. h(t) is not a probability as it depends upon the time and has no upper limit. The type to risk to which the population under consideration is exposed as a function of time depends upon the shape of the hazard function h(t). The cumulative hazard function is denoted by $H(t)$ and is defined as

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

The function $H(t)$ is also known as the integrated hazard function (Collett, 2003) [4]. Some of the main non-parametric estimation methods of survival analysis that have used in this paper which are as discussed:

2.1 Kaplan-Meier Estimation of the Survival Function

The Kaplan-Meier (product limit) method is a special case of the life table technique, in which the series of time intervals are formed in such a way that only one death occurs in each time interval and the death occurs at the beginning of the interval (Collett, 2003) [4].

Suppose that there are n individuals, with observed survival times $t_1, t_2, \ldots\ldots\ldots, t_n$ and r death times amongst the individuals, where $r \le n$. The ordered death times $t_{(j)},\ j = 1, 2, \ldots\ldots, r$ are $t_{(1)} < t_{(2)} < \ldots\ldots\ldots < t_{(r)}$/ Let $n_j,\ j = 1, 2, \ldots\ldots, r$ be the number of individuals who are alive just before the time $t_{(j)}$ and let $d_{(j)}$ be the number of individuals who die at time $t_{(j)}$. The probability for an individual dies during the interval $t_{(j+1)}$ to $t_{(j)}$ is estimated by $\frac{d_j}{n_j}$. Therefore the corresponding estimated survival probability in that interval is $\frac{(n_j - d_j)}{n_j}$. If the Censored survival times and one or more death times are same, then in this case it is assumed that the censored survival time(s) is taken to occur immediately after the death time. So, the estimated survival function for any time t in the j^{th} constructed time interval from $t_{(j)}$ to $t_{(j+1)},\ j = 1, 2, \ldots\ldots, r$, and all preceding time intervals is leads to the following Kaplan-Meier of the survivor function, $S(t) = \prod_{t \le t_j} \left(\frac{n_j - d_j}{n_j} \right)$ for $t_j \le t < t_{(j+1)},\ j = 1, 2, \ldots\ldots\ldots, r$ with $S(t) = 1$ for $t < t_{(1)}$ and $S(t) = 0$ for $t > t_{(r)}$ if $t_{(r)}$ is the last observation. It is noted that the Kaplan-Meier estimate is unspecified for the largest censored survival time. The plot of the Kaplan-Meier estimate of the survivor function is a step-function, where the estimated survival probabilities are constant between adjacent death times and decreases at each death time. Also, it is to be noted that the Kaplan-Meier estimate is a generalization of the empirical survival function, which includes both censored and complete observations. It is pertinent to note the following points about the Kaplan-Meier estimator in spite of its simplicity and wide applicability. Miller (1981) [16] in a path breaking research paper "what price Kaplan-Meier?" has stated the following.

The asymptotic efficiency of the Kaplan-Meier estimator is low when compared to the efficiency obtained under the parametric setup. He has examined this aspect by taking two parametric models namely exponential and Weibull.

2.2 Nelson-Aalen Estimator

The Nelson-Aalen Estimator, an alternative estimate of the survival function which is based on the individual event times and of cumulative hazard at time t is defined as:

$$H(t) = \sum_{t_j \le t} \frac{d_j}{n_j}, \text{ for } t > 0$$

Using this integrated function the estimate of the $S(t)$ is given by $S(t) = \exp(-H(t))$. The Kaplan-Meier estimate of the survivor function is given by

$$S(t) = \prod_{t \le t_j} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{t \le t_j} \left(1 - \lambda_j\right)$$

Now, instead of using KM estimate, we adopt the method proposed by Nelson-Aalen through the integrated hazard function $H(t)$ to estimate the survivor function.

2.3 Modified Forms of Kaplan-Meier Method of Estimation

Weighted Kaplan-Meier Estimator of Survival Function Jan *et al.* (2005) [12] claims that in life threatening diseases when some portion of the data is censored, the Kaplan Meier estimator becomes unreliable and inefficient. To deal with the situation they define a weight W_j at time $t_{(j)}$ as

$W_j = \frac{n_j - d_j}{n_j}$, which is known as non-censored rate.

Where, $W_j = 1$ if there is no censoring and in case of censoring at time t_j. Then, the Weighted Kaplan-Meier estimator is defined as

$$S^{**}(t) = \begin{cases} 1; \text{ for } t = 0 \\ \prod_{j;t_{(j)} \le t} W_j \left(\frac{n_j - d_j}{n_j} \right); \text{ for } t_{(j)} \le t; j = 1, 2, ..., n \\ 0; \text{ for } t > t_n \end{cases}$$

In this case $S^{**}(t)$ will reach zero, if the last observed survival time is censored. In the case of no censoring $n_j - d_j = n_{j+1}$ as there are no censoring times then $W_j = 1$. This leads to

$$S^{**}(t) = \begin{cases} 1; \text{ for } t < t_{(1)} \\ \frac{n_k}{n_{k+1}}; \text{ for } t_{(k)} \le t < t_{(k+1)} \text{ and } k = 1, 2, ..., n-1 \\ 1; \text{ for } t \ge t_n \end{cases}$$

2.4 Modified Weighted Kaplan-Meier Estimator

The main defect in the above Weighted Kaplan-Meier method is that it gives zero weight to the last censored observation and probability is equal to zero. To overcome this difficulty. Shafiq *et al*. (2007) [18] proposed a new weight which gives a non-zero weight to the last censored observation. The proposed Modified Weighted Kaplan-Meier estimator is then,

$$S^{**}(t) = \prod_{T \le t} W_j \left\lceil 1 - \frac{d_j}{n_j} \right\rceil$$

Where, the weight functions $W_j = \left\lceil 1 - \text{Sin} \left(\frac{c_j * P_j}{n_j} \right) \right\rceil$ is known as a non-censoring rate.

Both the Modified Weighted Kaplan-Meier and the Weighted Kaplan-Meier estimators give same weight to all censored observations. They also give same probability of survival, but the important point is that Weighted Kaplan-Meier estimator gives zero weight to the last censored observation while the modified weighted estimator gives it some non-zero weight and has a small probability of survival.

2.5 Hung's Estimator of Survivor Function

Hung (2008) [11] studied a Weighted Empirical Survivor Function (WESF). It has been shown that by choosing appropriate weights the estimator proposed by him is more efficient than the Kaplan-Meier estimate in both censored and uncensored data.

For the censoring case, the idea of the above WESF for the uncensored case has been applied to obtain a weighted KME as given below:

$$S_{KMW}(t) = \begin{cases} 1, & t < \tau_1 \\ \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i} \right) (1 - p_{k,i}), \tau_j \le t \le \tau_{j+1}, & j = 1, ..., k-1 \end{cases}$$

The weights suggested by Huang are,

$$w \equiv p_{n,i} = \frac{1}{\sqrt{n(n-1)}}; i = 2, ..., n-1$$

$$w_{1,n} = p_{n,1} = p_{n,n} = \frac{1}{2} \left(1 - \frac{n-2}{\sqrt{n(n-1)}} \right)$$

3. Conclusion

The standard method of estimating survival probabilities is the Kaplan-Meier estimator but this method of estimating probabilities is based on the assumption of censoring. When the data have high rate of censoring Kaplan-Meier estimator gives biased results and the reliability of the results get affected. To deal with the problem of censoring, proper weights have been assigned to Kaplan-Meier estimator and Weighted Kaplan-Meier estimator is very useful in the epidemiological studies.

4. References

1. Andersen PK. Repeated assessment of risk factors in survival analysis. Statistical Methods in Medical Research, 1992; 1:297-315.
2. Berkson, Gage RP. Calculation of survival rates for cancer. Proceedings of the Staff Meetings of the Mayo Clinic. 1950; 25:270-286.
3. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 1978; 65:141-151.
4. Collett D. Modeling of Survival Data in Medical Research. Chapman Hall, London, U.K, 2003.
5. Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society, Series B, 1968; 30:248-275.
6. Cohen AC. Jr. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. Technometrics, 1965; 3:327-339.
7. Cox DR. Regression model and life tables (with discussion). Journal of the Royal Statistical Society (B), 1972; 34:187-220.
8. Cox DR, Oakes D. Analysis of Survival Data. Chapman and Hall, London, 1984.
9. Crowley J, Hu M. Covariate analysis of heart transplant survival data. Journal of American Statistical Association. 1977; 78:27-36.
10. Gehan EA. Estimating Survival functions from the life table. Journal of Chronic Diseases. 1969; 21:629-644.
11. Huang ML. A Weighted Estimation Method for Survival Function. Applied Mathematical Sciences. 2008; 2:753-762.
12. Jan B, Ali Shah WS, Shah S, Qudir FM. Weighted Kaplan-Meier estimation of survival function in heavy censoring. Pakistan Journal of Statistics. 2005; 21(1):55-63.
13. Jenkins SP. Discrete time proportional hazard regression, Stata Technical Bulletin, 1997; 39:17-32.
14. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 1958; 53:457-481.
15. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, John Wiley and Sons, London, 1980.
16. Miller RG. Survival Analysis, John Wiley and Sons, New York, 1981.
17. Peto R, Peto J. Asymptotically efficient rank invariant procedures. Journal of the Royal Statistical Society, A. 1972; 135:185-207.

18. Shafiq M, Shah S, Alamgir. Modified Weighted Kaplan-Meier Estimator. Pakistan Journal of Statistics Operation Research. 2007; 3(1):39-44.
19. Susan M. Using Weighted Kaplan Meier statistics in non-parametric comparisons of paired censored survival outcomes. Biometrics. 2001; 57:361-368.
20. Ramadurai M, Ponnuraja C. Non-Parametric Estimation of the Survival probability of children affected by TB Meningitis, Journal of Arts, Science and Commerce, 2011; 2:216-228.