Journal of Biomaterials
2017; 1(2): 29-33
http://www.sciencepublishinggroup.com/j/jb
doi: 10.11648/j.jb.20170102.11

Application of Cox Regression and Kaplan Meir Estimates in the Survival Rate of Patients

Amos Langat[1], Joel Koima[2]

[1]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

[2]Department of Mathematics and Informatics, Kabarak University, Nakuru, Kenya

Email address:
moskiplangat@gmail.com (A. Langat), jkoima@kabarak.ac.ke (J. Koima)

To cite this article:
Amos Langat, Joel Koima. Application of Cox Regression and Kaplan Meir Estimates in the Survival Rate of Patients. *Journal of Biomaterials*. Vol. 1, No. 2, 2017, pp. 29-33. doi: 10.11648/j.jb.20170102.11

Received: April 10, 2017; **Accepted**: April 28, 2017; **Published**: July 3, 2017

Abstract: This study aim at focusing on the survival analysis for human subjects, to compare efficacy and safety, controlled experiments which conducted as clinical trials. Sometime it is interesting to compare the survival of subjects in two or more interventions. In situations where survival is the issue then the variable of interest would be the length of time that elapses before some event to occur. In many of the situations this length of time is very long for example in cancer therapy; in such case per unit duration of time number of events such as death can be assessed. The paper is highlighting the two difference estimates in the survival distribution of patients and later explain the strengths of the two estimates when use simultaneously in estimating the survival distribution. The researchers found that, application of the two estimates; Cox regression and Kaplan Meir will result in minimum errors estimates thus producing sufficient and complete survival distribution of patients under study.

Keywords: Survival Analysis, Cox Regression, Kaplan Meir

1. Introduction

Survival analysis techniques employ methods designed to investigate the amount of study time an experimental unit contributes to a study period from entry until event. The term "survival" may be misleading because the techniques are applicable to any well-defined event although traditionally death was the event of interest and the study period consisted of following the subject until death. Events in survival analysis (also referred to as endpoints or outcomes) are defined by a transition from one discrete state to another at an instantaneous moment in time. Examples of events include months until onset of disease, days until remission after cancer therapy, years until stock market crash, and hours until equipment failure, days until unemployment or time until failing or passing an exam.

Although the origin of survival analysis goes back to mortality tables from centuries ago, recent advancements in survival analytic techniques using non-parametric and semi-parametric approaches have allowed researchers flexibility in their work not previously seen within the confines of parametric methods. These methods have become popular over parametric methods due to the relatively robust modeling approaches without distributional assumptions on the survival times.

In the case of medical research for human subjects, to compare efficacy and safety, controlled experiments are conducted which are called as clinical trials [1]. In clinical or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. Sometime it is interesting to compare the survival of subjects in two or more interventions. In situations where survival is the issue then the variable of interest would be the length of time that elapses before some event to occur. In many of the situations this length of time is very long for example in cancer therapy; in such case per unit duration of time number of events such as death can be assessed. In other situations, the duration for how long until a cancer relapses or how long until an infection occurs can be assessed. Sometimes it can even be used for a specific outcome, like how long it takes for a couple to conceive. The time starting from a defined point to the occurrence of a given event is called as the survival time [2] and the analysis of group data as the survival analysis [3].

These analyses are often complicated when subjects under study are uncooperative and refused to be remained in the study or when some of the subjects may not experience the event or death before the end of the study, although they would have experience or died, or we lose touch with them midway in the study. The researchers label these situations as right-censored observations. [2] For these subjects we have partial information. We know that the event occurred (or will occur) sometime after the date of last follow-up. We do not want to ignore these subjects, because they provide some information about survival. We will know that they survived beyond a certain point, but we do not know the exact date of death.

The combination of the two methods for estimating survival distributions will provides modeling focus and thus this research paper is bringing into the focus the beneficial estimates of the two estimates when use in estimating the survival distribution of cancer patients.

Although there are well known methods for estimating unconditional survival distributions, most interesting survival modeling examines the relationship between survival and one or more predictors, usually termed covariates in the survival-analysis literature.

2. Methods and Materials

2.1. Cox Regression

Cox regression (or proportional hazards regression) is method for investigating the effect of several variables upon the time a specified event takes to happen. In the context of an outcome such as death this is known as Cox regression for survival analysis. The method does not assume any particular "survival model" but it is not truly nonparametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale [4].

Provided that the assumptions of Cox regression are met, this function will provide better estimates of survival probabilities and cumulative hazard than those provided by the Kaplan-Meier estimates.

The coefficients in a Cox regression relate to hazard; a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated.

In prospective studies, when individuals are followed over time, the values of covariates may change with time. Covariates can thus be divided into fixed and time-dependent. A covariate is time dependent if the difference between its values for two different subjects changes with time; e.g. serum cholesterol. A covariate is fixed if its values cannot change with time, e.g. sex or race. Lifestyle factors and physiological measurements such as blood pressure are usually time-dependent. Cumulative exposures such as smoking are also time-dependent but are often forced into an imprecise dichotomy, i.e. "exposed" vs. "not-exposed" instead of the more meaningful "time of exposure". There are no hard and fast rules about the handling of time dependent covariates.

Cox Regression builds a predictive model for time-to-event data. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time t for given values of the predictor variables. The shape of the survival function and the regression coefficients for the predictors are estimated from observed subjects; the model can then be applied to new cases that have measurements for the predictor variables.

One of the most popular regression techniques for survival outcomes is Cox proportional hazards regression analysis. There are several important assumptions for appropriate use of the Cox proportional hazards regression model, including

a) independence of survival times between distinct individuals in the sample,
b) a multiplicative relationship between the predictors and the hazard (as opposed to a linear one as was the case with multiple linear regression analysis), and;
c) a constant hazard ratio over time.

The Cox proportional hazards regression model

$$h(t) = h_0(t)\exp(b_1X_1 + b_2X_2 + \ldots + b_pX_p) \quad (1)$$

Where h(t) is the expected hazard at time t, $h_0(t)$ is the baseline hazard and represents the hazard when all of the predictors (or independent variables) X_1, X_2, X_p are equal to zero. Notice that the predicted hazard (i.e., h(t)), or the rate of suffering the event of interest in the next instant, is the product of the baseline hazard ($h_0(t)$) and the exponential function of the linear combination of the predictors. Thus, the predictors have a multiplicative or proportional effect on the predicted hazard.

Suppose we wish to assess the impact of exposure to nicotine and alcohol during pregnancy on time to preterm delivery. Smoking and alcohol consumption may change during the course of pregnancy. These predictors are called time-dependent covariates and they can be incorporated into survival analysis models. The Cox proportional hazards regression model with time dependent covariates takes the form:

$$\ln\left\{\frac{h(t)}{h_0(t)}\right\} = b_1X_1(t) + b_2X_2(t) + \ldots + b_pX_p(t) \quad (2)$$

Notice that each of the predictors, X_1, X_2,..., X_p, now has a time component. There are also many predictors, such as sex and race that are independent of time. Survival analysis models can include both time dependent and time independent predictors simultaneously

2.1.1. Strength of the Cox Regression Estimate

a) Does not require that you choose some particular probability model to represent survival times, and is therefore more robust than parametric
b) *Semi*-parametric

(Kaplan-Meier is non-parametric)

a) Can accommodate both discrete and continuous measures of event times

b) Easy to incorporate time-dependent covariates—covariates that may change in value over the course of the observation period

2.1.2. Weakness of the Cox Regression Estimate

The Cox model relies on the proportional hazards (PH) assumption, implying that the factors investigated have a constant impact on the hazard - or risk - over time. Which emphasize the importance of this assumption and the misleading conclusions that can be inferred if it is violated; this is particularly essential in the presence of long follow-ups.

2.2. Kaplan-Meier

The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals.[3] There are three assumptions used in this analysis. Firstly, it assumes that at any time patients who are censored have the same survival prospects as those who continue to be followed. Secondly, it assumes that the survival probabilities are the same for subjects recruited early and late in the study. Thirdly, it assumes that the event happens at the time specified. This creates problem in some conditions when the event would be detected at a regular examination. All we know is that the event happened between two examinations. Estimated survival can be more accurately calculated by carrying out follow-up of the individuals frequently at shorter time intervals; as short as accuracy of recording permits i.e. for one day (maximum). The Kaplan-Meier estimate is also called as "product limit estimate". It involves computing of probabilities of occurrence of event at a certain point of time

$$S_t = \frac{\text{Number of subjects living at the start} - \text{Number of subjects died}}{\text{Number of subjects living at the start}}$$

It is often of interest to assess whether there are statistically significant differences in survival between groups between competing treatment groups in a clinical trial or between men and women, or patients with and without a specific risk factor in an observational study. There are many statistical tests available; we present the log rank test, which is a popular non-parametric test. It makes no assumptions about the survival distributions and can be conducted relatively easily using life tables based on the Kaplan-Meier approach.

There are several variations of the log rank statistic as well as other tests to compare survival curves between independent groups.

The researchers use the following test statistic which is distributed as a chi-square statistic with degrees of freedom k-1, where k represents the number of independent comparison groups:

$$\chi^2 = \Sigma \frac{(\Sigma O_{jt} - \Sigma E_{jt})^2}{\Sigma E_{jt}} \quad (3)$$

Where ΣOjt represents the sum of the observed number of events in the jth group over time and ΣE_{jt} represents the sum of the expected number of events in the jth group over time. The observed and expected numbers of events are computed for each event time and summed for each comparison group over time. To compute the log rank test statistic, we compute for each event time t, the number at risk in each group, N_{jt} (e.g., where j indicates the group) and the observed number of events O_{jt} in each group. We then sum the number at risk, N_t, in each group over time to produce ΣN_{jt}, the number of observed events O_t, in each group over time to produce ΣO_{jt}, and compute the expected number of events in each group using $E_{jt} = N_{jt}*(O_t/N_t)$ at each time. The expected numbers of events are then summed over time to produce ΣE_{jt} for each group.

For time to event without competing risks, each patient under study will either fail, or survive without failure to their last contact. Such a failure may be death, disease-free, etc. The specific time of failure depends on the end point analyzed. A patient without failure at last contact is said to be censored due to lack of follow-up beyond this time, where it is known that such a patient has not failed by last contact but failure could occur at a later time.

The most reasonable and natural estimate of the probability of failure by a pre-specified time is the simple ratio of the number of failures divided by the total number of patients, provided that all patients without failure have follow-up to this time. This simple ratio is appropriately interpreted as an estimate of the probability of failure. This estimate is not only intuitive but is also unbiased when all patients who have not failed have follow-up through the specified time.

If one or more patients are censored before the specified time the simple ratio is no longer adequate, and methods that take into account data from censored patients are required to obtain an estimate consistent with this ratio. The KM method was developed for precisely this purpose, and when competing risk are not present this method leads to an estimate that is consistent with the desired simple ratio. The resulting estimate is also exactly equal to this ratio when all patients have either failed or been followed through the specified follow-up time.

2.2.1. Strength of the Kaplan- Meir (KM) Estimates

The benefit of the Kaplan–Meier (KM) estimate is the possibility to include censored data, which means that the information about patients who are lost at any point in time, for any reason, can be used for the analysis. The outcome measurements and the survival curves are results of a study and, therefore, depend on the design of the study and the group of patients included. The estimate also is useful in examining recovery rates, the probability of death, and the effectiveness of treatment. In a review of survival analysis published in cancer journals the quality of graphs was felt to be poor in 37% of the papers that included at least one survival curve [6].

2.2.2. Weakness of the Kaplan- Meir (KM) Estimates

Weakness of K-M method are firstly, the vertical drop at

each actual failure, draws undue visual attention to those particular "danger times," with the K-M estimate of the survival function remaining unchanged until the next failure is encountered. In reality, no practitioner would believe that patients are only at risk at specific times; rather, they are in continuous danger of failure, with a degree of danger perhaps changing with time.

Secondly, as time progresses, there are fewer remaining patients at risk. This has two direct effects on the K-M curve: The interval between failures grows, and the effect of each individual failure on the size of the step-down increases. Thus, the visual impact of a single failure is unjustifiably magnified in both the horizontal and the vertical directions if it occurs at a later time.

Thirdly, if the last remaining patient at risk fails, the K-M estimate of the survival function drops to zero at that time, whereas the true survival function will never reach zero in any physically sensible model.

The fourth drawback of the K-M method is somewhat more subtle. The K-M estimate of the probability of surviving each "danger time" depends only on the number of patients at risk at that time; for each censored patient, it disregards the time between the last failure and the time of censoring. [7] Again, the amount of censored subjects and the distribution of censored subjects are also important. If the number of censored subjects is large, one must question how the study was carried out or if the treatment was ineffective, resulting in subjects leaving the study to pursue different therapies. A curve that does not demonstrate censored patients should be interpreted with caution. [8]

And finally it's is limited in its ability to estimate survival adjusted for covariates

3. Results and Discussion

Cox regression and Kaplan Meir estimates works well in estimating the survival distribution of patients under study suffering from a particular disease. No estimates is superior that the other since, each estimates has its own statistical strenghth and weakness. Application of the two estimates Cox Regression and Kaplan -Meier is one of the best options to be used to measure the subjects of living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time and the analysis of group data as survival analysis. This can be affected by subjects under study that are uncooperative and refused to be remained in the study or when some of the subjects may not experience the event or death before the end of the study, although they would have experienced or died if observation continued, or we lose touch with them midway in the study. We label these situations as censored observations. The survival curve can be created assuming various situations.

Both the Kaplan-Meier method and the Cox proportional hazards (PH) model allow one to analyze censored data [7,8], and to estimate the survival probability, S(t), that is the probability that a subject survives beyond some time t. Statistically, this probability is provided by the survival function.

$S(t) = P(T > t)$, where T is the survival time. The Kaplan Meier method estimates the survival probability non-parametrically, that is, assuming no specific underlying function [19]. Several tests are available to compare the survival distributions across groups, including the log-rank and the Mann-Whitney-Wilcoxon tests [9, 10]. The Cox PH model accounts for multiple risk factors simultaneously. It does not posit any distribution, or shape for the survival function, however, the instantaneous incidence rate of the event is modeled as a function of time and risk factors.

4. Conclusion

Cox regression estimator advanced to prediction of survival time in individual subjects by only utilizing variables covarying with survival and ignoring the baseline hazard of individuals. Cox did this by making no assumptions about the baseline hazard of individuals and only assumed that the hazard functions of different individuals remained proportional and constant over time.

The Kaplan-Meier (KM) estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step approach. The KM estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. When there is no censoring, the estimator is simply the sample proportion of observations with event times greater than t. The technique becomes a little more complicated but still manageable when censored times are included.

In this study, the researchers found that the applications of the two estimates in measuring the survival distribution will result in minimum errors estimates thus produce sufficient and complete survival distribution of patients.

References

[1] Armitage P, Berry G, Matthews JN. 4th ed. Oxford (UK): Blackwell Science; 2002. Clinical trials. Statistical methods in medical research; p. 591

[2] Berwick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care. 2004; 8: 389–94

[3] Altman DG. London (UK): Chapman and Hall; 1992. Analysis of Survival times. In: Practical statistics for Medical research; pp. 365–93

[4] Cox DR, Oakes D. Analysis of Survival Data, Chapman and Hall, 1984

[5] Hosmer, DW and Lemeshow, S. Applied Survival Analysis: Regression Modeling of Time to Event Data. New York: John Wiley and Sons; 1999

[6] Altman DG, De Stavola BL, Love SB, Stepniewska KA (1995) Review of survival analyses published in cancer journals. Br J Cancer 72:511–518

[7] Carter RE, Huang P. Cautionary note regarding the use of CIs obtained from Kaplan-Meier survival curves. J Clin Oncol 2009; 27:174-5

[8] Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. Otolaryngol Head Neck Surg 2010;143:331-6

[9] Cox D. Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B. 1972;34:187–220

[10] Kaplan E, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc. 1958; 53:457–81. doi: 10.2307/2281868

[11] GEHAN EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965; 52:203–23

[12] Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966; 50:163–70