J.Univ.Ruhuna 2018 6(1):1-7
Volume 6 Number 1, May 2018
ISSN2345-9387 http://doi.org/10.4038/jur.v6i1.7892

Editorial

Survival time data and their first statistical analysis: Review

Abstract: In the life testing, medical follow-up studies, and other fields, it is often impossible to observe the lifetimes of all experimental units in the study. These types of data are called survival data. Because of the nature of the data, we cannot obtain the full information of the survival data. Therefore, it is not possible to apply the standard statistical techniques to analysis such survival data. In this paper, I mainly focus onright censoring data and explain how to derive the basic nonparametric estimators of cumulative distribution (Kaplan-Meier estimator), hazard, and cumulative hazard function using observed data. In addition to that I discuss how to compare the survival probabilities in two or more groups by using log-rank test. I also introduce the proportional hazard model (Cox's model) to incorporate the other related covariates to the experiment. Finally, I present some simulation study and real data application.

***Keywords**: Cox Model, Kaplan-Meier estimator, Log-rank test, right censoring, Survival data*

Introduction

In the life testing, medical follow-up studies, and other fields, it is often impossible to observe the lifetimes of all experimental units in the study. What makes measuring durations difficult is time itself. In most cases, it is highly likely that all the events have not been observed by the time one wants to make inference about lifetimes. For example, a medical professional will not wait fifty years for each individual in the study to pass away before closing the study. He or she is interested in the effectiveness of improving lifetimes after only a few years. The individuals in the study who have not died by the end of the study period are labeled as right-censored: all the information we have on these individuals are their current lifetime durations which are naturally less than their actual lifetimes. The simplest kind of censoring is that of single censoring which occurs when all observations are censored at the same time. There are two types of single censoring: Type I censoring and Type II censoring. In Type I censoring, the censoring time is predetermined. Type II censoring occurs if an experiment stops when a predetermined number of failures are observed; the remaining subjects are then right censored. In many studies, observations are not censored at the same time, which is frequently referred to as arbitrary censored data. For example, in a clinical trial, censoring occurs because of events due to other causes that are not related to what is being investigated in the study, such as: self-removal from the study, drop out, and death from other factors that are not related to the study. These are known as competing risk factors in literature.

The analysis of the survival data, such as the life time data, is very important in many fields including reliability, engineering, biology, and medicine. Survival data are highly non-normal in nature therefore, the use of standard statistical techniques like linear regression models is problematic.

Under the random censorship model, we assume that $X_1, X_2, \dots, X_n$ are independent nonnegative random variables with the continuous distribution function $F(x) = P(X \leq x)$. The censoring variables $Y_1, Y_2, \dots, Y_n$ are also nonnegative and are assumed to be a random sample, drawn independently of the X_is from a population with the continuous distribution function $G(y) = P(Y \leq y)$. The Y_is right-censor the X_is. The observable random variables are $Z_i = \min\{X_i, Y_i\}$ and $\delta_i = I\{Z_i = X_i\}$ where δ_i indicates whether Z_i is an uncensored observation or not. In this model, the X_is represent times to an endpoint event (e.g., death, relapse, malfunctioning) and the Y_is represent censoring times. In the random censorship model, informative censoring occurs when the distribution function G is informative about the distribution function F.

Some Preliminary Definitions

In this subsection, I introduce some basic concepts and their definitions, namely the survival function and the hazard rate function.

The Survival Function

The survival function of X is denoted by $S(x)$ and defined by $S(x) = 1 - F(x)$. This measures the probability that an individual survives from the time origin to a specific future time x.

The Hazard Rate Function

The hazard rate function is usually denoted by $h(x)$ and is the probability that an individual who is under observation at a time x has an event at that time. This means that $h(x)$ is the instantaneous event rate for an individual who has already survived at x. The hazard rate function is defined mathematically by

$$h(x) = \lim_{dx \to 0} \frac{P(x \ \leq X \leq x + dx \ /X \geq x)}{dx} = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)},$$

where $f(x) = \frac{dF(x)}{dx}$ and $S'(x)$are the probability density function of X and the derivative of $S(x)$, respectively.

The function $\varLambda(x) = \int_0^x h(u)du$ is called the cumulative hazard function for X, it is easy to show that, for continuous $X, S(x) = \exp\{-\varLambda(x)\}$.

Kaplan Meier Survival Estimate

In literature, most of the time the survival function is estimated by using the observed data, both uncensored and censored. This is a nonparametric estimator of $S(x)$ and it is denoted by $\hat{S}(x)$.Consider the right censored data $Z_i = \min\{X_i, Y_i\}$ and $\delta_i = I\{Z_i = X_i\}$ for n number patients in a medical study. Assume that $x_1 < x_2 < \cdots < x_n$are the distinct event times for the above observations. For simplicity, I assume here that there are no ties in the event times. As events are assumed to occur independently of one another, the probabilities of surviving from one interval to the next may be multiplied together to give the cumulative survival probability. That is the probability of being alive at time x_k. $S(x_k)$ is calculated from$S(x_{k-1})$, the probability of being alive at x_{k-1}, n_k, the number of patients alive just before x_k, and d_k, the number of events at x_k, by

$$S(x_k) = S(x_{k-1})(1 - \tfrac{d_k}{n_k}).$$

By using similar arguments, one can reach the following Kaplan-Meier (1958) product limit formula for the survival function,

$$S(x_k) = \prod_{k=0}^{n}(1 - \tfrac{d_k}{n_k}),$$

where$S(0)$ is the probability of survival at time 0. The value of $S(x)$ is constant between event times and, therefore, the estimated survival function is a step function. Confidence intervals for the survival probabilities are also possible. In the next subsection I will show how to obtain the KM-curve and confidence intervals for simulated data set.

In a similar fashion, one can show that the nonparametric estimator for $\varLambda(x)$ is

$$\hat{\varLambda}(x) = \prod_{k=0}^{n} \frac{d_k}{n_k}.`$$

Example 1

Consider an example for survival data. In this example, I simulate the failure data and censored data from the exponential distributions,$Exp(0.2)$and $Exp(0.1)$, respectively. In this case, I generate 50 observed data and compute the Kaplan-Meier (KM) estimator and its95% confidence bands by using R software. One can use the following R codes to generate the KM curve as shown in Figure 1.

```
library(survival)
survtime=rexp(50,0.2)
censtime=rexp(50,0.1)
sex=rep(1:2,c(24,26))
status=as.numeric(survtime<=censtime)
obstime=survtime*status + censtime*(1-status)
fit=survfit(formula=Surv(obstime,status==1)~1)
summary(fit)
plot(fit,xlab="Patient time(Years)",ylab="Survival Probability")
abline(h=0.5)
```

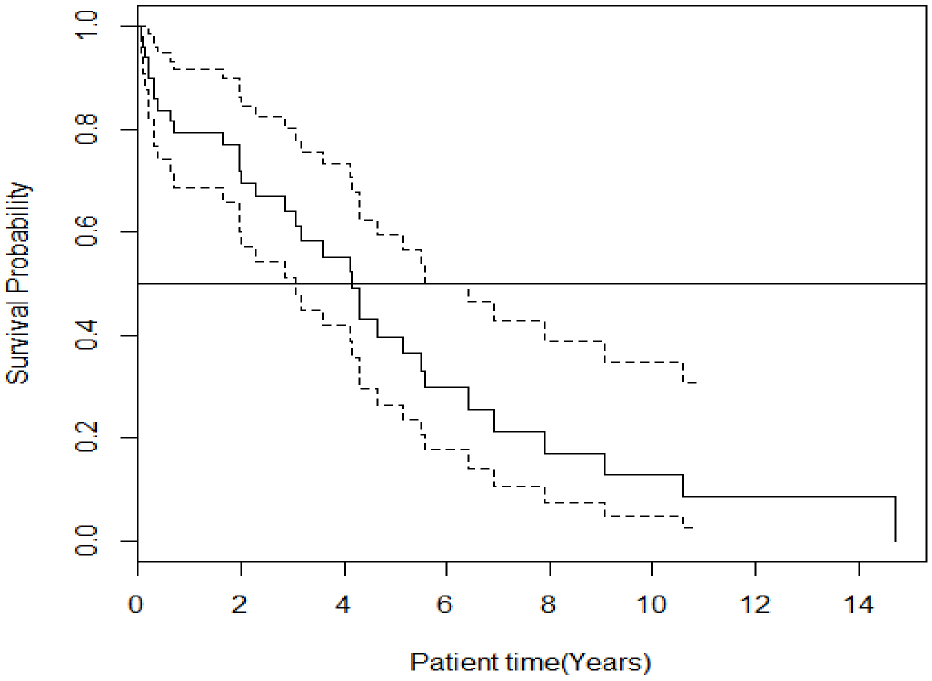
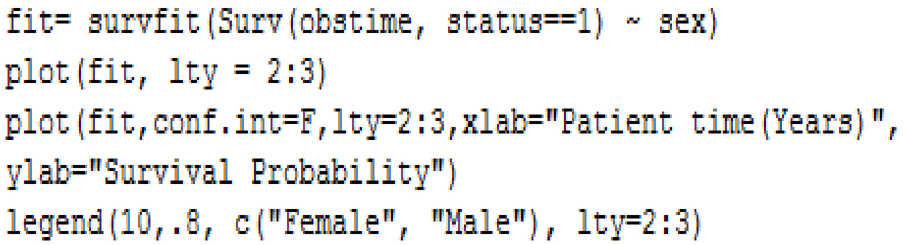


Figure 1: The KM curve, 95% confidence brand and median survival probability for simulated data

To obtain the median survival time and its95% confidence interval, one can use the KM curve in Figure 1. In order to build a confidence interval with a different confidence level, say 90%, for $S(x)$, youshould to use *conf.int = 0.9* in the R*survfit()* function. Next, one can try to compare the survival curves for two groups by extending the R codes.



```
fit= survfit(Surv(obstime, status==1) ~ sex)
plot(fit, lty = 2:3)
plot(fit,conf.int=F,lty=2:3,xlab="Patient time(Years)",
ylab="Survival Probability")
legend(10,.8, c("Female", "Male"), lty=2:3)
```

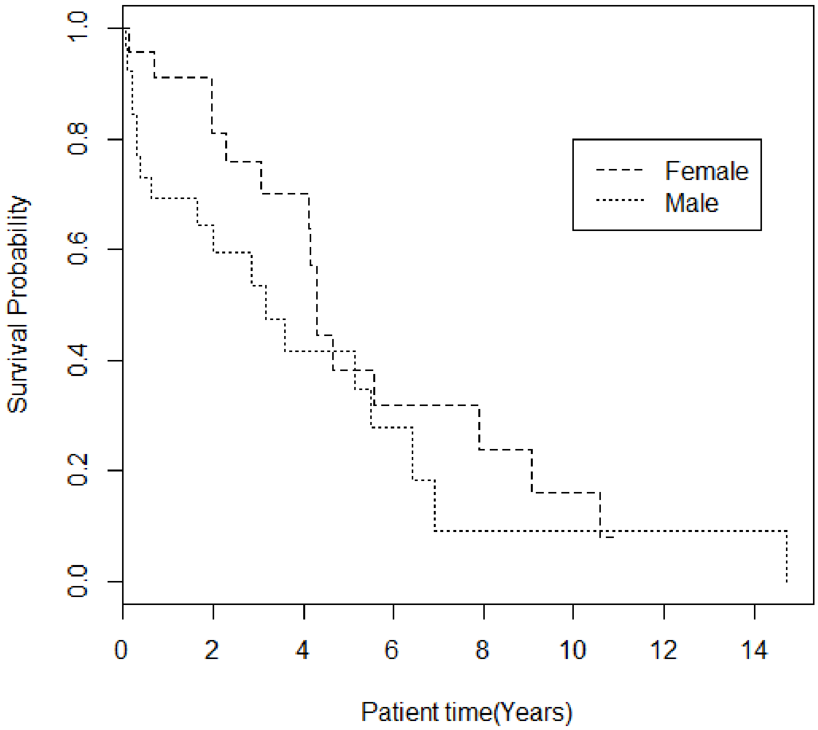


Figure 2: Comparisons of survival function for two groups

From the curves, it is clear that the survival probability functions are not much different in the middle part of the curves for both groups. But they differ on both ends. The above conclusion can be justified by using standard statistical tests like log-rank test.

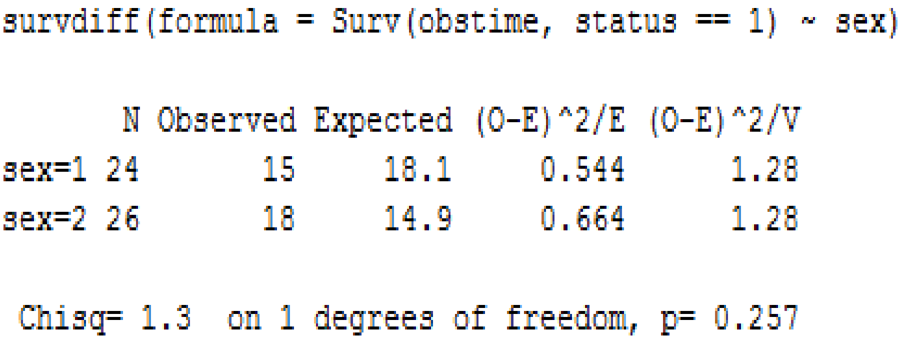
The log-rank Test

As in the standard statistics, two or more survival curves can be compared by conducting hypotheses testing. Because of the nature of the survival data here, the standard hypotheses testing like t-test for two sample case cannot be used. In such instances, the log-rank test can be used to check whether two or more survival curves are identical or not. Here we can test the hypothesis $H_0 : S_F(x) = S_M(x)$, for $x \geq 0$,where $S_F(x)$and $S_M(x)$ are survival functions for two groups. We can consider the composite hypothesis instead of the simple one. The log-rank test statistic for this hypothesis is given by

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected number of events for group i.In example 1, $k = 2$, where the two groups are females and males. This test statistic has a chi-square distribution with $k-1$ degrees of freedom if null hypothesis is true.

We use the following R codes to check the above hypothesis for example 1. The following output results for the log-rank test justifies the conclusion from the graph in Figure 2.

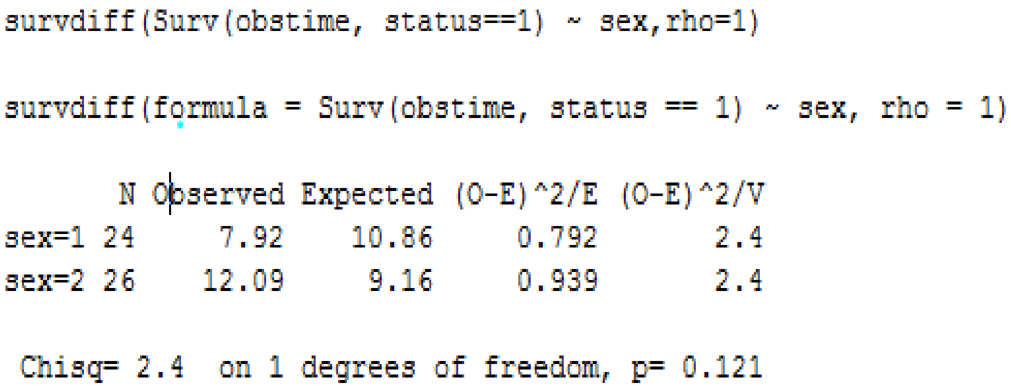


```
survdiff(formula = Surv(obstime, status == 1) ~ sex)

       N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 24       15     18.1     0.544      1.28
sex=2 26       18     14.9     0.664      1.28

 Chisq= 1.3  on 1 degrees of freedom, p= 0.257
```

If we want to perform Peta-Prentice's Wilcoxon test, we need to specify rho=1 in the above R code.



```
survdiff(Surv(obstime, status==1) ~ sex,rho=1)

survdiff(formula = Surv(obstime, status == 1) ~ sex, rho = 1)

       N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 24     7.92    10.86     0.792       2.4
sex=2 26    12.09     9.16     0.939       2.4

 Chisq= 2.4  on 1 degrees of freedom, p= 0.121
```

So, we have a similar conclusion as the log-rank test. Finally, we discuss the Cox's proportional hazard model in this paper. This is a regression type model known as Cox's regression model (1972).

Cox's Proportional Hazard Model

We can use the log-rank test to compare the survival times in different groups. But it does not allow other covariates to be taken into account in our analysis. Cox's proportional hazard model is analogous to the multiple regression model. This model allows the analysis of survival data by regression model similar to those of linear models and generalized linear models. The scale on which linearity is assumed is the log-hazard scale. Therefore, in the Cox's model, the dependent variable is the hazard rate. This model enables one to compare the survival times of particular groups by taking into account other relevant factors. These factors are sometimes known as covariates.

The model can be written as

$$\ln h(t) = \ln h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

or

$$h(t) = h_0(t)\exp(\sum_{i=1}^{p} \beta_1 X_1)$$

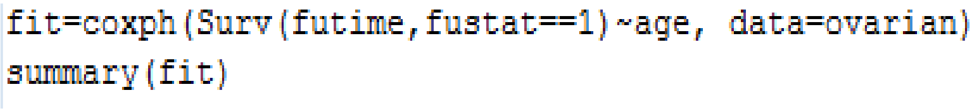
where$h(t)$ is the hazard rate at time t, $h_0(t)$ is the baseline hazard when values of all covariates are zero, and the $X_1, X_2, \dots, X_p$ are the covariates.

Most of the time, the coefficients $\beta_1, \beta_2, \dots, \beta_p$ are estimated by likelihood methods using observed data.

Real Data Application: Ovarian Data

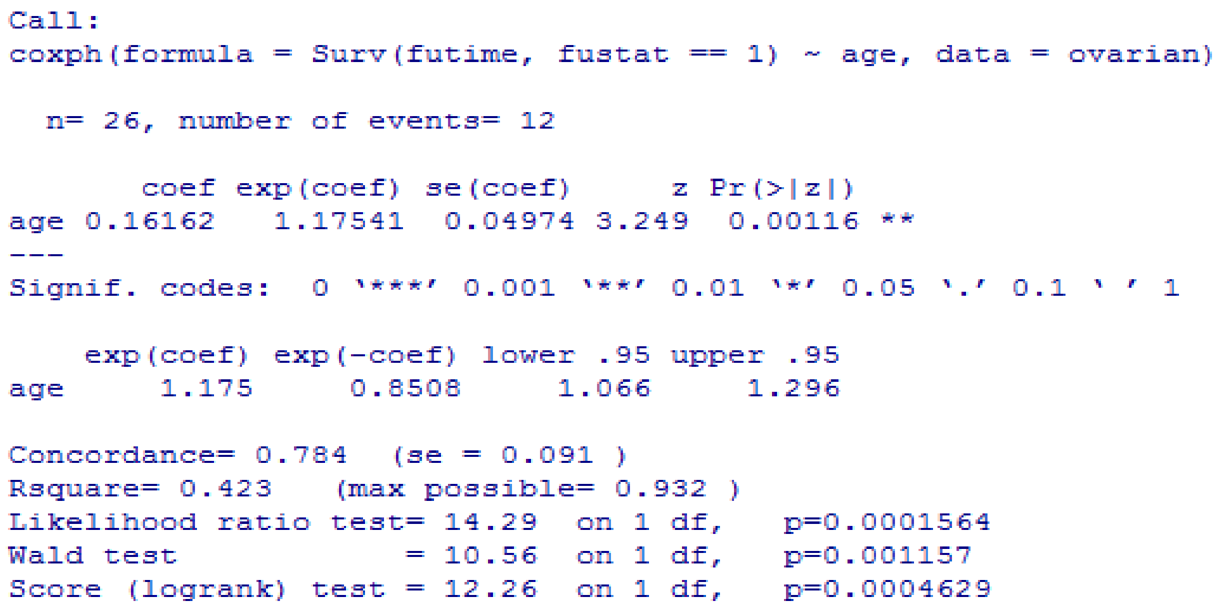
In this subsection, we consider an example from literature. An investigator collected data related to 845 patients with primary epithetical ovarian carcinoma between January 1990 and December 1999 at the Western General Hospital in Edinburg. Follow-up data were available up to the end of December 2000. By this time 550(75.9%) subjects had died (Clark et al, 2001).

We fit a Cox model to ovarian data with futime as a dependent variable and age as a covariate. The following R code can be used to fit this model.



```
fit=coxph(Surv(futime,fustat==1)~age, data=ovarian)
summary(fit)
```

The results are as follows:



```
Call:
coxph(formula = Surv(futime, fustat == 1) ~ age, data = ovarian)

  n= 26, number of events= 12

       coef exp(coef) se(coef)     z Pr(>|z|)
age 0.16162   1.17541  0.04974 3.249  0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
age     1.175     0.8508     1.066     1.296

Concordance= 0.784  (se = 0.091 )
Rsquare= 0.423   (max possible= 0.932 )
Likelihood ratio test= 14.29  on 1 df,   p=0.0001564
Wald test            = 10.56  on 1 df,   p=0.001157
Score (logrank) test = 12.26  on 1 df,   p=0.0004629
```

According to the output, the likelihood ratio test, the Wald test, and the log-rank test reveal that the model is significant. These are all equivalent in the large samples but may differ a little in small sample cases. The *coef*,0.16162, is the hazard ratio between two groups in log scale and $\exp(coef)$, 1.175, is the actual hazard ratio.

Discussion

In survival analysis, the statistical inference techniques that can be used are different from the standard statistical techniques. This difference is mainly because of the nature of the survival data, especially the right censoring. This preventsthe full information of the event interested of some subjects in the study from being obtained. In this review, the Kaplan-Meier estimator is used to get the cumulative distribution function, the log-rank test to compare the survival functions of two groups, and Cox's regression model to incorporate other covariates. These techniques can be used with time dependent covariates. Finally, one can extend these techniques to analyze the recurrent event data, where we observed more than one event of a subject. For example, for a subject with a cancer, he or she has multiple occurrences in the observation window.

References

Clark T. G., Stewart M. E., Altman D. G., Gabra H., Smyth J. (2001) A prognostic model for ovarian cancer, Br J Cancer 85: 944-952.

Cox, D.R. (1972) Regression models and life-tables (with discussion), Journal of Royal Statistical Society, B (34), pp. 187-200.

De Mel, Withanage Ajith Raveendra, "On some inferential problems with recurrent event models" (2014). Doctoral Dissertations. 2340, http://scholarsmine.mst.edu/doctoral_dissertations/2340.

Kaplan, E.L, Meier, P. (1958) Nonparametric estimation from incomplete observations, Journal of the American Statistical Association, 153(282), pp. 457-481.

Withanage Ajith Raveendra De Mel
Department of Mathematics, Faculty of Science,
University of Ruhuna, Matara, Sri Lanka.