Survival Analysis of UIS patients under Parametric and Non-Parametric Approach using R software

Deepapriya. S and Ravanan. R

Department of Statistics, Presidency college, Chennai - 600 005

Abstract- The study of survival analysis involves censoring which is an important feature of the clinical data. This paper deals with the analysis of non-parametric and parametric estimates of survival function and median survival time of UMARU Impact Study (UIS) data. The Survival probabilities S(t) are estimated by Kaplan Meier product Limit method under non-parametric approach and the survival functions are estimated through Weibull distribution, Exponential distribution and lognormal distribution under parametric approach. In modelling the survival data, most of the time we have no prior information about the theoretical distribution of survival time and graphical tools are employed to fit a better distribution using R Software.

Index Terms- censoring, Kaplan Meier estimate, parametric survival function, survival probability

I. Introduction

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. A special feature of survival data, which renders standard methods inappropriate, is that survival times are frequently censored. The exact survival times of the study subjects are unknown. That are called censored observation or censored times and can also occur when people are lost to follow up after a period of study. This paper deals with the survival analysis of University of Massachusetts AIDS Research Unit (UMARU) Impact Study (UIS) data under non Parametric and parametric method. Survival probabilities are estimated by Kaplan Meier Product limit method under non-parametric approach and the survival functions are estimated through Weibull distribution, exponential distribution and lognormal distribution under parametric approach. The non parametric technique (Kaplan-Meier, 1958) is applied to estimate the survival functions and hazard rates of the survival data. Estimating the distribution of the dependent variable without making assumptions about its shape is an important first step in analyzing a dataset [8]. Given the importance of the distribution of the dependent variable it is valuable to "let the data speak for itself" first. In biomedicine, prior knowledge about the distribution of survival function is appropriate to compare the survival functions [4]. Although non parametric methods play an important role in survival studies, parametric techniques cannot be ignored. Once an appropriate statistical model for survival time has been constructed and its parameters estimated, its information can help to predict survival, develop optimal treatment regimens, plan future clinical or laboratory studies [7]. R software is used to estimate survivor function Non-parametrically, Parametrically by Weibull, exponential, Log-normal distribution for the UIS data.

1.1 Survivorship function

Let T be the survival time and S(t) is the probability that an individual survives longer than t. i.e.,
$S(t) = p(T > t) = 1- p(T \leq t) = 1- F(t)$,
Where F (t) is the distribution function of the time t,
S(t) is non increasing function of time t with the properties
S (t) =1 for t = 0.
S (t) =0 for t = ∞.
S (t) is known as the cumulative survival rate. The graph of S (t) is called the survival curve.

II. Non Parametric Method

Non parametric analysis, estimating the probabilities without making any assumptions on its shape is called non-parametric analysis [8].

2.1 Kaplan- Meier method of estimation

Let n be the total number of individuals whose survival time, censored or not, are available. Relabeling the survival times in order of increasing magnitude such that $t_1 \leq t_2 \leq \ldots \leq t_n$ and the values of r are consecutive integers 1,2,….,n if there are no censored observation. If there censored observations, they are not. Then the survival probabilities are calculated using $s(t) = \prod_{t_r - t} \frac{n-r}{n-r+1}$ where r runs through those positive integers for which $t_r \leq t$ and t_r is uncensored. The variance of s(t) is approximated by $var[s(t)] = [s(t)]^2 \sum_r \frac{1}{(n-r)(n-r+1)}$ where r includes those positive integer for which $t_{(r)} \leq t$ and t_r corresponds to a death. Estimated standard error is $\sqrt{var(s(t))}$ .95% confidence interval for s(t) is s(t)±1.96S.E[s(t)].

III. Parametric Approach

Parametric approaches are used either when a suitable model or distribution is fitted to the data or when a distribution can be assumed for the population from which the sample is drawn [8]. Commonly used survival distributions are the exponential, Weibull, lognormal, and gamma. In this paper AIC (Akaike Information Criterion)[4] is used for the selection of a Distribution that fit the data among the given.

3.1 Exponential distribution

The simplest and most important distribution in survival studies is the exponential distribution. The exponential distribution is characterized by a constant hazard rate λ, its only parameter. A high λ value indicates high risk and the short survival: a low λ value indicates low risk and long survival [4,5]. When the survival time T follows the exponential distribution with a parameter λ, the probability density function is referred as

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0, \lambda > 0 \\ 0 & t < 0 \end{cases}$$

and survivorship function is then $S(t) = \exp[-\lambda t]$ $t \geq 0$, and the hazard function is $h(t) = \lambda$.

3.2 Weibull Distribution

Weibull distribution is one of these lifetime distributions, named after the Swedish physicist Waloddi Weibull(1936)[4,8]. The weibull distribution with two parameters has hypothetical confirmations in technical as well as biomedical applications as a purely empirical model.

The continuous random variable t has a weibll distribution with two parameters λ,γ and the density function is given by

$f(t) = \gamma\lambda(\lambda t)^{\gamma-1}\exp[-(\lambda t)^{\gamma}]$ $t \geq 0, \lambda > 0, \gamma > 0$

Survivorship function is $S(t) = \exp[-(\lambda t)^{\gamma}]$

$h(t) = \gamma\lambda(\lambda t)^{\gamma-1}$

weibull distribution has two parameters, where γ is the shape parameter and λ is scale parameter. At $\gamma < 1$, the failure rate decreases over time, at $\gamma = 1$ failure rate remains constant over time, and at $\gamma > 1$ failure rate increases over time.

3.3 Log-normal

Its origin may be traced as far back as 1879, when McAlister(1879) described explicitly a theory of the distribution. Most of its aspects have since been under study. Gaddum(1945) gave a review of its application in biology, followed by Boag's(1949) application in cancer research[4,5]. The Survival time T such that log T is normally distributed with mean μ and variance σ^2. We then say that T is log normally distributed and write T as $\Lambda(\mu,\sigma^2)$. The log normal distribution is suitable for survival patterns with an initially increasing and then decreasing hazard rate [8].

The probability density function and the survivorship function are, respectively,

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log t-\mu)^2} \qquad t > 0, \sigma > 2$$

$$S(t) == \frac{1}{\sigma\sqrt{2\pi}} \int_t^{\infty} \frac{1}{x} e^{-\frac{1}{2\sigma^2}(\log x-\mu)^2} dx$$

The popularity of the distribution is due in part to the fact that the cumulative values of y = log t can be obtained from the tables of the standard normal distribution and corresponding values of t are then found by taking antilogs.

3.4 R Software

R software is used to find survival probabilities under non parametric approach [11], survival curves and fitting the survival distributions parametrically. First install package survival using
>install.packages('survival')
To load libraries, use
>library(survival)

IV. Computation and Calculation

The data set contains 628 subjects are considered without incorporating the covariates a null model is being fitted for the survival times of the data.

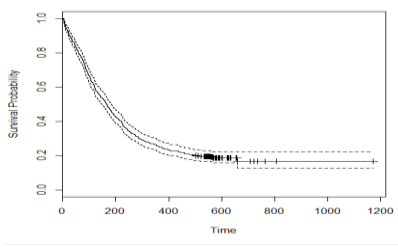
Table 4.1 (Survival Probability Obtained By Kaplan Meier Estimator)

Time	n.risk	n.event	Survival probabilities	std.err	lower 95% CI	upper 95% CI
2	628	1	0.998	0.00159	0.995	1
3	627	3	0.994	0.00317	0.987	1
4	624	4	0.987	0.00448	0.979	0.996
5	620	4	0.981	0.00546	0.97	0.992
6	616	3	0.976	0.00609	0.964	0.988
7	613	4	0.97	0.00684	0.956	0.983
8	609	2	0.967	0.00717	0.953	0.981
9	607	2	0.963	0.0075	0.949	0.978
10	605	4	0.957	0.00809	0.941	0.973
11	601	3	0.952	0.00851	0.936	0.969
12	598	2	0.949	0.00878	0.932	0.966
13	596	1	0.947	0.0089	0.93	0.965
14	595	4	0.941	0.0094	0.923	0.96
15	591	3	0.936	0.00974	0.917	0.956
17	588	1	0.935	0.00986	0.916	0.954
18	587	1	0.933	0.00997	0.914	0.953
19	586	2	0.93	0.01019	0.91	0.95
20	584	2	0.927	0.0104	0.907	0.947
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
491	129	1	0.204	0.01607	0.175	0.238
494	128	1	0.202	0.01603	0.173	0.236
499	127	1	0.201	0.01598	0.172	0.235
502	125	1	0.199	0.01593	0.17	0.233
516	120	1	0.197	0.01589	0.169	0.231
519	119	1	0.196	0.01584	0.167	0.229
559	54	1	0.192	0.01596	0.163	0.226
568	38	1	0.187	0.01632	0.158	0.222
659	9	1	0.166	0.02438	0.125	0.222

The survival fit obtained using the following R code for UIS data.

```
us=read.table("d:rprog/usd.csv",header=T,sep=",")
kmpl=survfit(Surv(TIME,CENSOR)~1,data=us)
summary(kmpl)
plot(kmpl)
```

fig. 4-1 Kaplan Meier survival curve



The survival probabilities are obtained using the KM fit the curve obtained as a smooth curve for only 18% of censoring has occurred and also the event time are immediate [7]. The median survival time estimated Non-parametrically as 208 day.

Table 4.2(Parametric estimation of UIS data)

	(Intercept)	Scale	Loglik	AIC
Weibull	5.654021	1.138789	-3382	6768.022
Exponential	5.670383	1	-3388.6	6770.28
Lognormal	5.096428	1.419206	-3368	6739.905

```
weibull .null<- survreg(data = us, Surv(TIME,CENSOR) ~ 1, dist = "weibull")
plot(x = predict(weibull, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
y = rev(seq(0.01, 0.99, by = 0.01)), col = "red")

Expo.null<- survreg(data = us, Surv(TIME,CENSOR) ~ 1, dist = "weibull")
plot (x = predict(Expo.null, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
    y = rev(seq(0.01, 0.99, by = 0.01)),      col = "green")

lognormal <- survreg(data = us, Surv(TIME,CENSOR) ~ 1, dist = "lognormal")
lines(x = predict(lognormal, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
    y = rev(seq(0.01, 0.99, by = 0.01)),      col = "blue")
```

A null model for the Exponential weibull, and log normal distribution is obtained by the above R code and scale obtained by weibull fit(table.1.2) is 1.138789 nearly equals 1 and is a special case of exponential and a exponential can also a better fit for this data. The intercept of exponential fit is 5.670, λ = exp (-intercept) = 0.003448, μ = 1/.003448 = 290 days. The mean survival time obtained by exponential fit is 290 days and the median survival time $t_{0.5}$ = -log0.5/λ = 201 days. The two parameter of the weibull distribution are obtained from the weibull fit are λ= exp(-intercept)=.00350, and μ = 1/scale=0.878. The estimated parameters of log normal distribution are μ = intercept = 5.0964 and σ^2 = scale^2= 1.4192 and the mean survival time obtained from the log normal fit is 296 days. The survival probabilities for the above mention distribution are obtained from the estimated parameter of the survival fit using the R codes are given in the table (1.3). The log likelihood and the AIC obtained under the fit table (1.2) exhibits log normal is a suitable model for the above mentioned data.

Table 4.3. Estimated survival probabilities from parametric fit.

Time	Censor	Estimated Survival Probabilities		
		Exponential	Weibull	Lognormal
2	1	**0.993128**	**0.988431**	0.999996
3	1	**0.98971**	**0.983389**	0.999973
4	1	**0.986303**	**0.978542**	0.999913
5	1	**0.982908**	**0.973841**	0.999795
6	1	**0.979525**	**0.969257**	0.999601
7	1	**0.976154**	**0.964772**	0.999316
8	1	**0.972794**	**0.960371**	0.998928
9	1	**0.969446**	**0.956046**	0.99843
10	1	**0.966109**	**0.951789**	0.997814
11	1	**0.962784**	**0.947593**	0.997077
12	1	**0.95947**	**0.943455**	0.996216
13	1	**0.956167**	**0.939369**	0.995232
14	1	**0.952876**	**0.935332**	0.994125
15	1	**0.949597**	**0.931342**	0.992897
17	1	**0.943071**	**0.92349**	0.990085
18	1	**0.939825**	**0.919625**	0.988508
19	1	**0.93659**	**0.915797**	0.986821
20	1	**0.933367**	**0.912006**	0.985029
21	1	**0.930154**	**0.90825**	0.983135
22	1	**0.926952**	**0.904526**	0.981145
23	1	**0.923762**	**0.900836**	0.979061
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
655	0	**0.104523**	**0.120419**	0.22986
658	0	**0.103447**	**0.119374**	0.229278
659	0	**0.103091**	**0.119028**	0.229086
708	0	**0.087066**	**0.103306**	0.220447
720	0	**0.083537**	**0.099799**	0.218549
734	0	**0.079601**	**0.095864**	0.216431
762	0	**0.072275**	**0.088475**	0.212487
763	0	**0.072026**	**0.088222**	0.212353
805	0	**0.062316**	**0.078266**	0.207101
1172	0	**0.017582**	**0.028147**	0.181076

4.1 Graphical representation of survival function using R

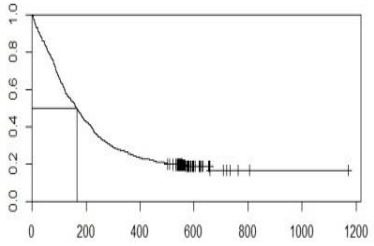


Fig 4 - 1 Kaplan-Meier Estimator

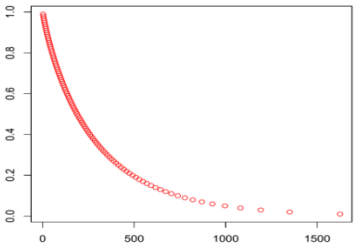


Fig 4 - 2 Weibull

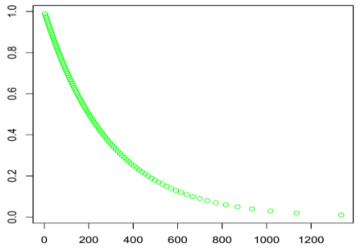


Fig 4 - 3 Expoenetial

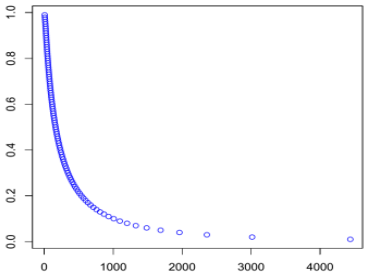


Fig 4 - 4 Lognormal

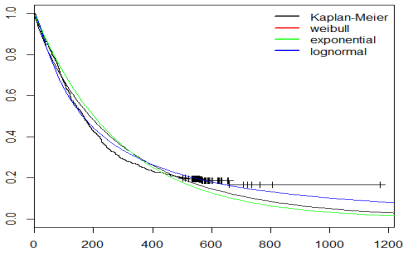


Fig 4 - 5 Non-parametric and parametric survival function

V. CONCLUSIONS

On the analysis of UIS data the estimated survival probabilities were given in the table (4.1) & (4.3) and the median survival time was calculated graphically as 166 days under Non parametric approach. The survivorship curves obtained by Kaplan Meier method under Non-Parametric approach and also by parametric method Exponential ,Weibull and Log-normal can be visualised from fig(4.5) there is no significance difference between the estimates of survivorship function obtained by parametric and non parametric method. On a comparison with the parametric distributions log normal is found to be a better fit in accordance with the AIC value for the UIS data, moreover the scale parameter obtained by weibull is nearly 1 which can be reduced to exponential survival function so obtained for the UIS data.

REFERENCES

[1] Altman DG, Bland JM.(1998)," Time to event (survival) data". British Medical Journal (BMJ); vol. 317:pp. 468-469.

[2] Altman DG, Bland JM.(1998), "Survival probabilities (The Kaplan-Meier method)". BMJ; vol. 317; pp.1572 - 1580.

[3] Collett D. (2003), "Modeling of Survival Data in Medical Research". Chapman Hall, London, U.K.

[4] Elisa T. Lee. (1992), "Statistical methods for Survival Data Analysis". Second Edition. A Wiley-Inter science publication, United States of America.

[5] German Rodrıguez, grodri@princeton.edu, Parametric Survival Models, Spring, 2001; revised Spring 2005, Summer 2010

[6] Kaplan, E. L., and Paul Meier (Jun. 1958)," Non parametric Estimation from Incomplete Observations", Journal of the American Statistical Association. Vol. 53, pp457 – 481.

[7] Kleinbaum, D.G. (1996). Survival analysis. New York: Springer-Verlag.

[8] Maryam Siddiqa, Mueen - ud -Din Azad, Muhammad Khalid Pervaiz,Muhammad Ghias, Gulzar H. Shah, Uzma Hafeez Survival analysis of dialysis patients under parametric and non-parametric approaches, Electronic Journal of Applied Statistical Analysis, EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 2, 271 – 288

[9] Maarten L. Buis, Department of Social Research Methodology, Vrije Universiteit Amsterdam,m.buis@fsw.vu.nl,April 2, 2006, An introduction to Survival Analysis

[10] Rupert G. Miller J R (1981), "Survival Analysis", John Wiley & Sons, United States of America.

[11] Ramakrishnan, M. Ravanan, R. Estimation of Survival Distribution Using R Software International Journal of Scientific and Research Publications, Volume 3, Issue 4, April 2013 1, ISSN 2250-3153

AUTHORS

First Author – S. Deepapriya, M.Sc.(Maths), M.Phill. (Maths), M.Sc. (Stats),Research Scholar, Presidency college, Chennai. Email – sbdpriya@gmail.com

Second Author – Dr.. R Ravanan, M.Sc., M.Phil., Ph.D., Head, Department of statistics, Presidency college, Chennai. Email – ravananstat@gmail.com