Stochastic Modeling and Applications
Vol. 25 No. 2 (July-December, 2021)
ISSN: 0972-3641

Received: 01st March 2021 *Revised: 06th April 2021* *Selected: 16th April 2021*

ESTIMATION OF SURVIVAL PROBABILITY FOR LUNG CANCER PATIENTS USING KAPLAN-MEIER APPROACH

JAYASANKAR RATHAKRISHNAN AND RAVANAN.R

Abstract. Amongst all the statistical procedures, Survival Analysis plays a vital role if the dependent (outcome) variable is time (time to an event). Survival Analysis is preferred over other regression analysis as it accommodates well the censoring concepts. Our study consists of 168 lung cancer patients with seven covariates. In this article the survival probability is estimated non-parametrically from the observed survival time using Kaplan-Meier method. Confidence interval for survival probability and median survival time are calculated. To compare the two survival groups based on the gender, survival comparison tests, log-rank test and Gehan-Wilcoxon tests are used. Study finds that there is a marked difference in survival pattern of patients based on gender

1. INTRODUCTION

In data analysis if the out-come variable of interest is time to an event then the most appropriate method is Survival Analysis. Survival Analysis accommodates censoring concept, which is its unique feature. The three main goals of survival analysis are i). To estimate and interpret the survivor function from the survival data ii). To compare the survivor functions iii). To assess the relationship of explanatory variable to survival time [5]. In this article we focus on the first two. In majority of the of the studies related to cancer, the major outcome under estimation is the time to an event of attention. In case if the event happened with every individuals, various methods of analysis would be suitable. Though, it is normal that at the end of study time (follow-up period) a few individuals have not had the event of interest, and thus their original time to event is not known. Additional, most of the survival data are not symmetrically distributed, but are mostly right skewed and at the beginning of the study the evets are more and towards the end of the study relatively less events. Such features of the data make the unique methods like survival analysis crucial.

2. MATERIALS AND METHODS

Our study datasets consists of 168 lung cancer patients in which 121 patients (72 %) attained the event (death) and the remaining 47 patients are censored. The covariates are Age, Gender. To estimate the survival probability, a non-parametric approach based on Kaplan-Meier method is used. To find the survival patterns of male and female patients, KM curves for two groups are drawn using

Key words and phrases. Kaplan-Meier, log-rank, Gehan-Wilcoxon, Lung cancer.

survminer and analysis is done using survival package in R and to verify whether any statistical difference exists between these groups, we used hypothesis tests, log-rank and Wilcoxon test.

2.1. Kaplan-Meier survival estimates. Estimation of survival probabilities can be done by non-parametric approach based on the observed survival times of both censored and un-censored events, by Kaplan-Meier method (Kaplan-Meier, 1958) also known as product-limit method. Assume that ' p' patients have attained the events during this period of different follow up times . Since the events are unspecified and expected to happen independently with each other, the survival probability from one interval to the successive interval should be multiplied in order to evaluate the cumulative survival probability. In general the probability of the individual being alive at any given time ti, is evaluated using the following formula.

$$S(ti) = S(t_{i-1})(1 - d_i/n_i) \tag{2.1}$$

When, t(0) = 0; then S(0) = 0

The evaluated probability is expected to be a step function as S(t) is constant between the successive time of events. This estimator permits every patient to provide information to the calculations until they are free from the event. In case if the individual meets the event (assume no censoring), then the estimator will be reduced to a simple ratio between the number of individual events free at a given time 't' and the number of patients who entered the study. The Log-rank test is a non-parametric way of testing hypothesis to compare the survival distribution of two groups. The test is also called "Mantel-Cox test"(1966) and it is necessary to find out significance difference exists or not between two survival distributions. Gehan (1965) and Breslow (1970) generalized the Wilcoxon rank sum test to permit for censored data. The generalized Wilcoxon test, because it assumes weights identical to the number of patients at risk, will put comparatively more weight on differences between the survival functions at smaller values of time. The log-rank test, because it assumes weights equal to 1, and will place more emphasis than the generalized Wilcoxon test on differences between the functions at larger values of time.

3. RESULTS AND DISCUSSIONS

Table 1 explains the important features of the KM survival probability. The estimator at some point in time is attained by multiplying a sequence of conditional survival probabilities, with the estimate being unchanged between subsequent event times.

The first observed survival time is 5 days. During 5 days observation, there are 168 subjects at risk. The estimate of the survival function in 5 days is 0.994 and its confidence interval for 95% is 1.000. The next observed survival time is 11 days and thus the estimate of the survival function is obtained by multiplying the value of the survival function just prior to 11 days and the conditional survival probability for 11 days is 0.988 and its confidence interval for 95% is 1.000. The Median survival time from the above table is 310 days (approximately). The function is

TABLE 1. Kaplan-Meier Survival Probabilities of Lung Cancer Data

Time	Risk	Events	Survival Probability	Std Error	Lower CI	Upper CI
5	168	1	0.994	0.006	0.983	1
11	167	1	0.988	0.008	0.972	1
12	166	1	0.982	0.01	0.962	1
13	165	1	0.976	0.012	0.953	1
15	164	1	0.97	0.013	0.945	0.996
305	63	1	0.511	0.041	0.437	0.598
310	62	1	0.503	0.041	0.428	0.591
320	61	1	0.495	0.041	0.42	0.583
765	8	1	0.089	0.03	0.047	0.171
791	7	1	0.077	0.028	0.038	0.157
814	5	1	0.061	0.026	0.027	0.142

TABLE 2. Log-rank Test for Male and Female patients

Variable	N	Observed value	Expected value	(O-E)2/E	(O-E)2/V
Male	104	83	69.5	2.6	6.16
Female	64	38	51.5	3.52	6.16

TABLE 3. Wilcoxon Test for Male and Female patients

Variable	N	Observed value	Expected value	(O-E)2 / E	(O-E)2 / V
Male	104	51.5	42.3	2.00	6.78
Female	64	21.1	30.3	2.79	6.78

not defined beyond 814 days which is denoted in Table 1. As the time increases the survival probability decreases in this Kaplan-Meier table.

Figure 1: The Kaplan-Meier Survival Curve for lung Cancer Patients with Confidence Intervals and Risk Table. The median survival time is 310 days.

The estimates obtained are always depicted in graphical form. The graph showing approximation survival probabilities (on the Y axis) and time past after entry into the study in days (on X the axis) consists of horizontal and vertical lines. The survival curve is drawn as a step function. The figure above shows that KM curve of point estimation for lung Cancer Death.

Figure 2 represents the Non-parametric Kaplan-Meier survival curves of Lung Cancer data for the variable sex. Visual check recommends that there is a substantial difference between the gender survivals. From the above graph, the median survival time for the female patients is 426 days whereas for male patients it is 284 days. The two survival curves are compared statistically by testing the Null hypothesis highlighting that there is no a major difference between the two survival patterns. Log rank test (is a non-parametric test), makes no assumptions about the survival distributions. Log rank statistic is approximately equivalent to a Chi-Square test statistic.

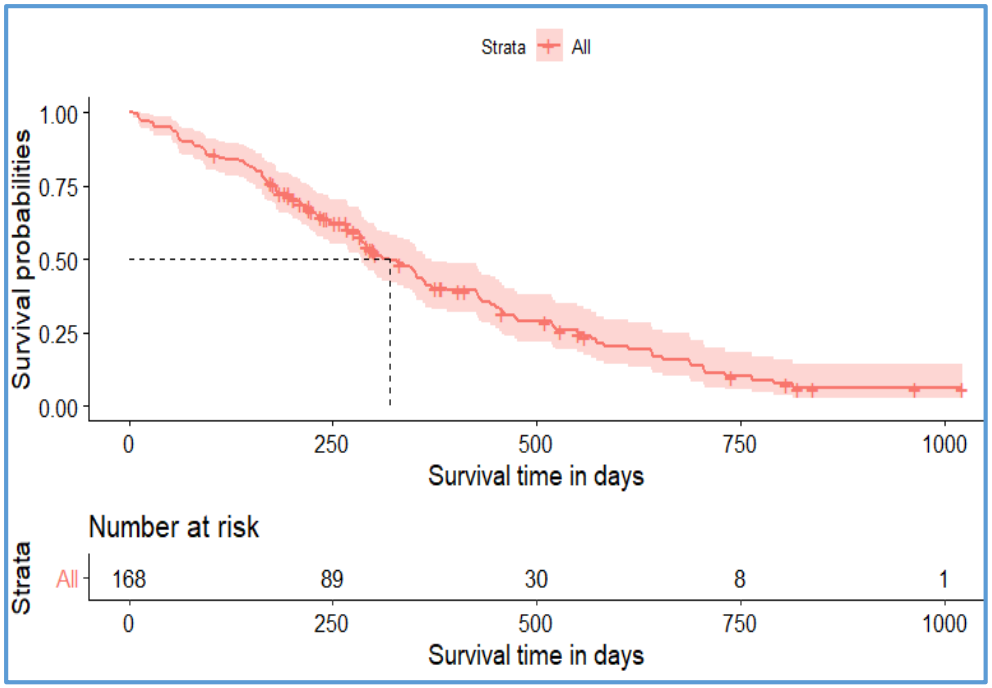

Figure 1: The Kaplan-Meier Survival Curve

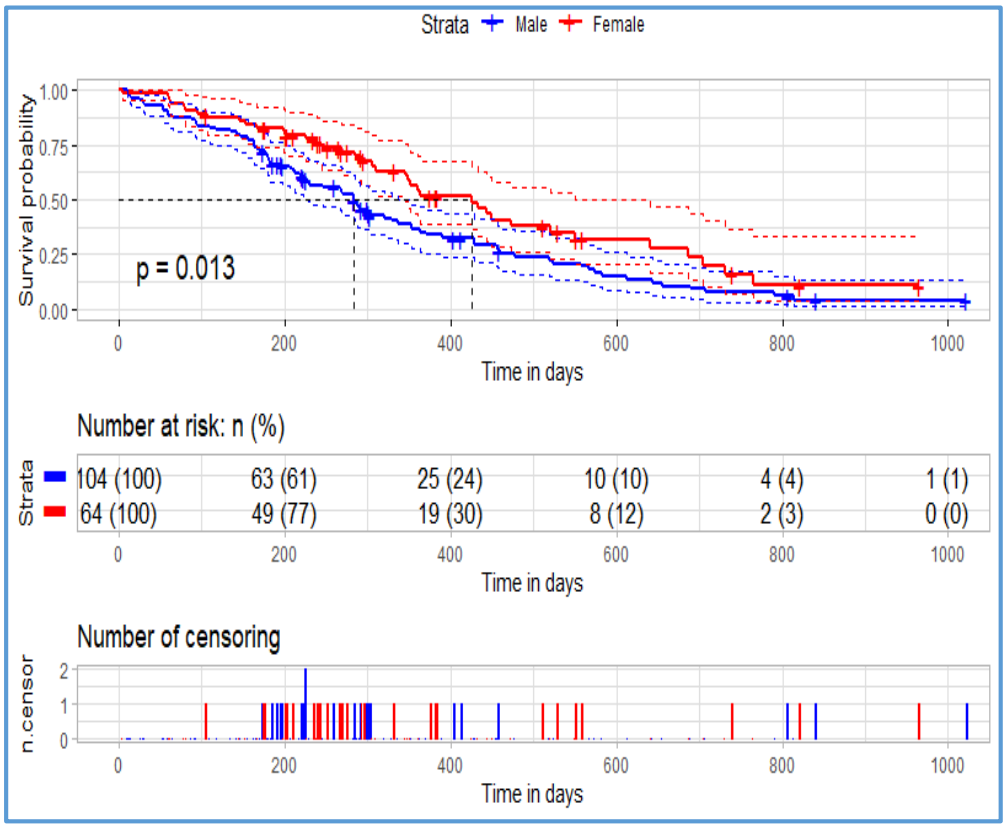

Figure 2: Survival curves for male and female

Outcome of table 2 represents the log-rank test for comparison of survival probabilities yielded, p-value is 0.013. So, there is a marked difference between the male and female patients. The log-rank test assumes equal weight to all survival times and emphasizes the tail of the survival curve.

The above table 3 outcome represents the Wilcoxon test p-value is 0.04 and chi-square value is 6.8. So, there is a marked difference between male and female patients in mortality. The Wilcoxon give more weight to earlier survival times and emphasizes beginning of the survival curve. The larger p-value in Wilcoxon test compared to the log-rank test is due to the differences between the two survivals occurs mainly at later failure times.

4. CONCLUSION

The basic concepts of Non-parametric model the Kaplan-Meier outcome represents that as the time increases the survival probability decreases in this Kaplan-Meier table and curve. Log-rank test table demonstrates that there is a marked difference between the male and female patients and the curve also represents the same p-value is 0.013. The Wilcoxon test also significant at 5% level of significant and p-value is 0.009 shows in the above table. From this we can understand the gender plays a vital role in mortality of the lung cancer patients.

5. LIMITATIONS

Kaplan-Meier is not a regression model. Since it is non-parametric, we cannot summarize the relationship between the survival time and other explanatory variables and also it can only incorporate on categorical variables.

References

1. Altman, D.G and Bland J.M: Statistics Notes: Interaction revised: The difference between two estimates: *British Medical Journal* (2003) 326 –329.
2. Clark. T.G, Altman D.G and De Stavola B.L: Quantifying the completeness of follow-up: *Lancet* (2002) 1309–1310.
3. Clark.T.G, Stewart M.E, Gabra H and Smyth.J: A Prognostic model for Ovarian Cancer: *British Journal Cancer* (2001) 85 944–952.
4. Collett, D. Modelling Survival Data in Medical Research: *London, Chapman and Hall* (1994)
5. Kleinbaum. D.G and Klein. M: Survival Analysis, A Self-Learning Text. Second Edition, Springer (1996)
6. Hosmer. D.W and Lemeshow. S. Applied Survival Analysis: Regression Modelling of Time to event Data *New York: Wiley* 1999
7. Kaplan. E.L and Meier. P Non-Parametric estimation from incomplete observations *J Am Stat. Assoc* 53, 457–481.

Research Scholar, Department of Statistics, Presidency College, Chennai, Tamilnadu, 600025, India
Email address: jaipb19@gmail.com

Joint Director of Collegiate Education, Chennai Region, Chennai-600018, India
Email address: ravananstat@gmail.com