Food and Scientific Reports
ISSN 2582-5437

Introduction to Survival Analysis and Application in Agricultural Research

Santosha Rathod[1] and Amit Saha[2] and Kanchan Sinha[3]

[1]ICAR- Indian Institute of Rice Research, Hyderabad; [2]Central Sericultural Research & Training Institute (CSRTI), Mysuru; [3]ICAR-Indian Agricultural Statistics Research Institute, New Delhi

ABSTRACT

Survival analysis is a statistical methodology, which generally use to study time to event data. It is a methodology use to analyze the outcome variable time until the occurrence of event of interest, say for example, time to germination of seed, time to plant undergoes particular disease or stress, unemployment after education etc. Censored or time to event data are common in medical research, particularly when studying the survival time of patients. Survival data are also frequently occurring in agricultural research for example, seed germination times, insect survival in plant, disease occurrence, agricultural insurance etc. In agricultural research, however, survival analysis has been overlooked may be because of availability of analysis tools. However, in this short article information on details of Survival Analysis and illustration of clinical data has been presented.

Survival analysis is a branch of statistics, which deals with the expected duration of time until one or more events occurs, such as death in biological organisms and failure in mechanical systems. Survival Analysis is called as reliability theory/ analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event of interest can be can be death, occurrence of a disease, marriage, divorce, unemployment after education, etc. In logistic regression, we were interested in studying how independent variables were associated with the occurrence of categorical dependent event/ disease. Sometimes, though, we are interested in how an independent variable affects the dependent event to occur and data underlying process is assumed to be normal. But, in case of time to event occurred data, the underlying process may not follow the normal distribution it may falls under other than normal distributions. In these cases, logistic regression is not appropriate, then survival analysis come into the picture. In survival analysis, subjects are usually followed over a specified time and the focus is on the time at which the event of interest occurs. Linear regression cannot model the survival time as a function of predictor variables because, times are typically positive numbers; ordinary linear regression may not be the best choice Secondly, ordinary linear regression cannot effectively handle the censoring of observations.

Censored or time to event data are common in medical research, particularly when studying the survival time of patients. Survival data are also frequently occurring in agricultural research for example, seed germination times, insect survival in plant, disease occurrence, agricultural insurance etc. In agricultural research, however, survival analysis has been overlooked may be because of availability of analysis tools. Scherm and Ojiambo (2004) reviewed survival analysis and its application in plant pathology and further illustrated the methodology in longitudinal data on the timing of defoliation of blueberry leaves. Recently, Załuski *et al.* (2018), analyzed survival of willow plants in a long-term experiment. Onofri *et al.* (2019) reviewed about analyzing censored data in agricultural research with available software information as well. However, in this short article information on details of Survival Analysis and illustration of clinical data has been presented.

Terminologies:

Survival Time: Survival time refers to a variable which measures the time from a starting time (e.g., time initiated the treatment) to a endpoint of interest (e.g., attaining certain functional abilities).

Survival Rate: Survival rate is defined as the percent of observations/ peoples who survive a disease such as cancer for a specified amount of time.

Time-to-event: It is the time until the event occurs. Time-to-event is a positive random variable.

Hazard*:* The instantaneous rate at which a randomly-selected individual known to be alive at time (t - 1) will die at time t is called the conditional failure rate or instantaneous hazard.

Life table: Also called a mortality table or actuarial table, shows, for each age, what the probability is that a person of that age will die before his or her next birthday ("probability of death"). In other words, it represents the survivorship of objects from a certain population, also called as Kaplan–Meier survival life table.

The Survival Function:

It is a function describing the proportion of individuals surviving to or beyond a given time. Let T denote the survival time, then the survival function S (t) is defined as follows;

$$\hat{S}(t) = \frac{No. of\ observations\ survives\ longer\ than\ t}{Total\ number\ of\ obseravtions}$$

S(t) = P(surviving longer than time t)

= P(T > t)

The function S(t) is also known as the cumulative survival function, which lies between 0 to 1.

The Hazard Function:

An alternative characterization of the distribution of *T* is given by the hazard function, or instantaneous rate of occurrence of the event, `defined as follows;

$$\lambda(t) = \lim_{dt\to 0} \frac{\Pr\{t \le T < t + dt | T \ge t\}}{dt}$$

In other words, the hazards function is defined as the ratio of rate of occurrence of the event at duration *t* equals the density of events at *t*, to the probability of surviving to that duration without experiencing the event.

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Estimation of Survival Function:

A very widely used method of estimating survival function is by using Kaplan–Meier method and the resulting plot is called as **Kaplan–Meier curve** (Kaplan and Meier, 1958). This is a non-parametric method of estimating the survival function. Let $t_1 < t_2 < \ldots < t_2$ be the observed event times and $n=n_0$ the sample size. Let d_j be the number of individuals who have an event at time t_j, where j=1,…,k, and m_j the number of individuals censored in the interval $[t_j, t_j+1)$. Then $n_j=(m_{j+} d_j)+\ldots+(m_k+d_k)$ is the number of individuals at risk just prior to t_j.

$$\hat{S}(t) = \prod_{j:t_{j\le t}} \frac{n_j - d_j}{n_j}$$

Standard errors can be calculated using Greenwood's formula, which approximates the variance as

$$\widehat{Var}\{\hat{S}(t)\} = \{\hat{S}(t)\}^2 \prod_{j:t_{j\le t}} \frac{d_j}{n_j(n_j - d_j)}$$

Illustration:

Data from a clinical trial on colon cancer adjuvant therapy1 are used as an illustration. A group of colon cancer patients were followed up from diagnosis to death. That is, the time scale has origin the time of diagnosis of colon cancer and endpoint the time of death from colon cancer. The dataset is freely available in R Software R (dataset 'colon' in package 'survival'), contains 929 observations on colon cancer (Therneau, 2015).

The survival function: A very widely used method of doing that is calculating and plotting a Kaplan–Meier curve. In figure 1 Kaplan–Meier survival curve, is depicted calculated from the colon data by considering sex variable. Probability value depicts that the proportion surviving remains same for both the sex.

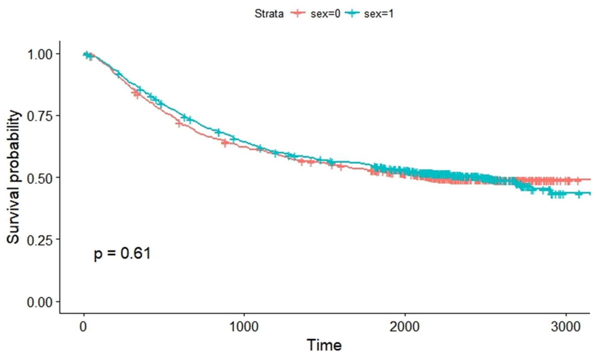

Fig.1. Kaplan–Meier survival curve for variable sex

Cox proportional hazards regression model: As part of the analysis Cox proportional hazards regression model was also fitted and results are depicted in Table 1.

Table 1: Fitting of Cox proportional hazards model

	coef e	xp(coef)	se(coef)	z	p
sex	-0.06607	0.93606	0.06828	-0.97	0.33323
age	0.00213	1.00213	0.00291	0.73	0.46476
rxLev	-0.03789	0.96282	0.07957	-0.48	0.63397
rxLev+5FU	-0.43677	0.64612	0.08622	-5.07	4.10E-07
obstruct	0.22062	1.24685	0.08471	2.6	0.00921
perfor	0.12898	1.13767	0.18578	0.69	0.48752
adhere	0.16846	1.18348	0.09254	1.82	0.06871
nodes	0.04075	1.04159	0.01078	3.78	0.00016
differ	0.13872	1.14881	0.07015	1.98	0.04797
extent	0.45389	1.57442	0.08401	5.4	6.60E-08
surg	0.24301	1.27508	0.07432	3.27	0.00108
node4	0.62764	1.87319	0.10059	6.24	4.40E-10
etype	-0.26914	0.76404	0.06778	-3.97	7.20E-05

Applications

Survival Analysis is an age old statistical technique which can be applied in many areas where our interest is model the time to event occurrence. Survival analysis has been mainly utilized in medical research, however it can be applied to many areas such as unemployment studies, crop insurance, insect pest modeling, seed germination (Hay et al 2014), plant growth in long term experiments (Zaluski et al 2014), germination test data (Hunter et al 1984), botanical epidemiology etc. In general survival Analysis can be applied to all the problems where our interest study is time to the happening of event.

References

Cox DR. Regression models and life-tables. J R Stat Soc B 1972; 34: 187-220.

Hay, F. R., Mead, A., & Bloomberg, M. (2014). Modelling seed germination in response to continuous variables: Use and limitations of probit analysis and alternative approaches. Seed Science Research, 24, 165– 186.

Hunter, E. A., Glasbey, C. A., & Naylor, R. E. (1984). The analysis of data from germination tests. Journal of Agricultural Science, 102, 207– 213.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 58: 457-481.

Onofri A, Piepho H-P, Kozak M. (2019). Analysing censored data in agricultural research: a review with examples and software tips. Ann Appl Biol. 2019; 174:3–13.

Scherm, H., & Ojiambo, P. S. (2004). Applications of survival analysis in botanical epidemiology. Phytopathology, 94, 1022– 1026.

Therneau T (2015). *A Package for Survival Analysis in S.* version .38, https://CRAN.R-project.org/package=survival.

Załuski, D., Mielniczuk, J., Bronowicka-Mielniczuk, U., Stolarski, M. J., Krzyżaniak, M., Szczukowski, S., & Tworkowski, J. (2018). Survival analysis of plants grown in long-term field experiments. Agronomy Journal, 110, 1791– 1798.