Estimation of survival times for lung cancer patients via Kaplan-Meier

Michael Holdgreve

Department of Mathematical Sciences - Purdue University Fort Wayne
Supervisor: Alessandro Maria Selvitella

Abstract

When analyzing a small population, it is important to include every data point possible. However, this becomes difficult when the study is constrained by the condition that data points will drop out of the study throughout the data collection time frame, such as in the case of survival analysis. Whether the study is analysing the life-cycle of manufacturing parts, aquatic biology, or medical cancer studies, certain methods must be used to obtain accurate conclusions regarding the population.

Introduction

The main problem with survival analysis in general is that the analyzed data set will never be 100% complete. This phenomenon is known as “censoring” or “censored data.” What this means is that the collection of the data has been prematurely terminated for some reason or another, leaving a gap in the data collection.

A few examples of censoring are: when a subject withdraws from the study due to the revocation of consent, from the subject prematurely passes away, or from another variable interfering with the subject completing the study such as if a study occurs for a fixed period, such as 10 years, and the event that is being studied, such as the development of cancer, does not occur. While this is ultimately a good thing for the patient, it poses a problem because it is unknown if the development of cancer occurred in 11 years, 20 years, or if it never occurred at all. .

It is this lack of information which causes a problem in being able to compile a complete and robust prediction model. Censoring is necessary because the alternative would be to keep the subjects in the study, even after the event of interest occurs. The result of this is that the subjects which continue to be in the study after the event of interest occurs inaccurately increases the predicted survival probability.

Problem

The ultimate goal for this study is to determine the survival probability of a patient at any particular time. There are two main problems when studying such a population, however. The first problem is that patients may either die or drop out of the study, but we’re not exactly sure when. If the patients check in at regular intervals, it can be determined when the last known survival point was. It is assumed that if the patient doesn’t check in at a scheduled check-in point, that the patient died somewhere between check-in A and check-in B. The second problem is that the overall population size is very small. To help boost the population size, the study allows patients to drop in the study no matter what time has elapsed since their diagnosis.

Kaplan-Meier

The Kaplan-Meier method takes into account the first and second problems by segmenting the study into several small intervals. This means that each interval has its own survival percentage with its own number of living subjects and number of subjects who died. The only caveat is that a patient cannot drop in during the middle of an interval. The patient is added to the study during a regular interval check-in, at which point the interval statistics are calculated as normal. The formula for finding the survival percentage for a given interval is

$$\textit{Survival Percentage } (S_t) = \frac{\textit{Number of subjects living} - \textit{Number of subjects died}}{\textit{Number of subjects living}} \text{ [5]}$$

These odds are then multiplied out in a cumulative fashion to find the survival probability for a specific interval after the first. For example:

$$\textit{Odds of Surviving to Day N} = \prod_{i=1}^{n-1} DailySurvivalOdds_i$$

This is done no matter how many patients are in the study, and continues until the study is concluded. The survival chart is then compiled in a step format to get an overall view of survival odds. The chart is typically annotated when the data is censored.

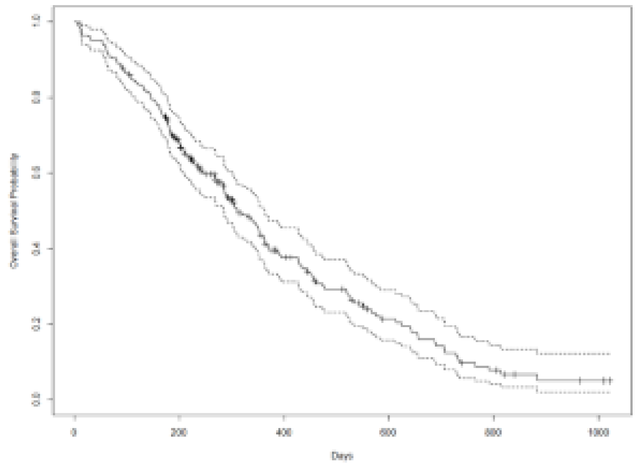

Fig. 1: Annotated Kaplan-Meier chart showing censoring

Figure 1 is a Kaplan-Meier plot for the probability of survival as it applies to the lung cancer survival data included in the attached libraries vs the amount of time the given person survives. The solid line represents the best-fit step function, with the dotted line indicating the 95% confidence intervals. The hash-marks along the step function indicate that there has been a censoring in the data. One of the main benefits for using the Kaplan-Meier method is that it is known whether the data is censored or not, and if so, where. Although it is difficult to tell how many people are in the study at any given interval, you can tell by the hash-marks where the data is censored the most. In this example, most of the censoring occurs between days 200 and 400, which corresponds to a survival probability of 40% to 60%.

What Happens If Data Is Not Censored

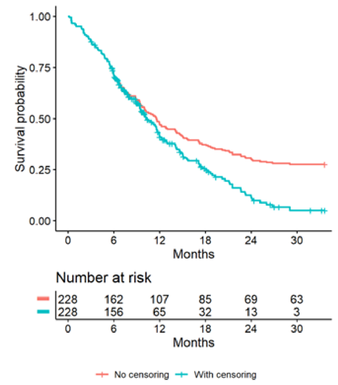

Fig. 2: Regression Line With Censoring vs Without Censoring

Figure 2 shows the effect of censoring on survival probability. Without any data being censored at all and keeping all data points in the study, even after the event of interest, it artificially inflates the survival probability. Without censoring, according to the model, the survival probability past 30 months is approximately 30%. However, by censoring the data and dropping the samples as soon as the event of interest occurs, we can see the results that the odds of survival past 30 months is approximately 10%. This discrepancy is one example of the necessity for censoring the data appropriately, even if it means a lower sample size at the end of the study.

Acknowledgements

Special thanks to Clark, Bradburn, Love, and Altman for their invaluable documentation and reference materials in understanding Survival Analysis. [4] [2] [1] [3]

References

[1] M Bradburn et al. “Survival analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit”. In: *British Journal of Cancer* 89 (2003), pp. 605–611.

[2] M J Bradburn et al. “Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods”. In: *British Journal of Cancer* 89 (2003), pp. 461–436.

[3] T Clark et al. “Survival analysis part IV: Further concepts and methods in survival analysis”. In: *British Journal of Cancer* 89 (2003), pp. 781–786.

[4] T.G. Clark et al. “Survival Analysis Part I: Basic concepts and first analyses”. In: *British Journal of Cancer* 89 (2003), pp. 232–238.

[5] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. “Understanding survival analysis: Kaplan-Meier estimate”. In: *International Journal of Ayurveda Research* 4 (2010), pp. 274–278.