Cornell University
Cornell Statistical Consulting Unit

StatNews #78

What is Survival Analysis?

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 can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack.

In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Why not use linear regression to model the survival time as a function of a set of predictor variables? First, survival times are typically positive numbers; ordinary linear regression may not be the best choice unless these times are first transformed in a way that removes this restriction. Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations.

Observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring. Suppose patients are followed in a study for 20 weeks. A patient who does not experience the event of interest for the duration of the study is said to be right censored. The survival time for this person is considered to be at least as long as the duration of the study. Another example of right censoring is when a person drops out of the study before the end of the study observation time and did not experience the event. This person's survival time is said to be censored, since we know that the event of interest did not happen while this person was under observation. Censoring is an important issue in survival analysis, representing a particular type of missing data. Censoring that is random and non informative is usually required in order to avoid bias in a survival analysis. For more information on censoring, including types of censoring, see our Statnews #67 at: http://www.cscu.cornell.edu/news/statnews/stnews67.pdf.

Unlike ordinary regression models, survival methods correctly incorporate information from both censored and uncensored observations in estimating important model parameters. The dependent variable in survival analysis is composed of two parts: one is the time to event and the other is the event status, which records if the event of interest occurred or not. One can then estimate two functions that are dependent on time, the survival and hazard functions. The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The survival function gives, for every time, the probability of surviving (or not experiencing the event) up to that time. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. While these are often of direct interest, many other quantities of interest (e.g., median survival) may subsequently be estimated from knowing either the hazard or survival function. It is generally of interest in survival studies to describe the relationship of a factor of interest (e.g. treatment) to the time to event, in the presence of several covariates, such as age, gender, race, etc. A number of models are available to analyze the relationship of a set of predictor variables with the survival time. Methods include parametric, nonparametric and semiparametric approaches.

Parametric methods assume that the underlying distribution of the survival times follows certain known probability distributions. Popular ones include the exponential, Weibull, and lognormal distributions. The description of the distribution of the survival times and the change in their distribution as a function of predictors is of interest. Model parameters in these settings are usually estimated using an appropriate modification of maximum likelihood.

A nonparametric estimator of the survival function, the Kaplan Meier method is widely used to estimate and graph survival probabilities as a function of time. It can be used to obtain univariate descriptive statistics for survival data, including the median survival time, and compare the survival experience for two or more groups of subjects. To test for overall differences between estimated survival curves of two or more groups of subjects, such as males versus females, or treated versus untreated (control) groups, several tests are available, including the log-rank test. This can be motivated as a type of chi-square test, a widely used test in practice, and in reality is a method for comparing the Kaplan-Meier curves estimated for each group of subjects.

A popular regression model for the analysis of survival data is the Cox proportional hazards regression model. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The Cox regression model is a semiparametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods described above. In particular, and in contrast with parametric models, it makes no assumptions about the shape of the so-called baseline hazard function.

The Cox regression model provides useful and easy to interpret information regarding the relationship of the hazard function to predictors. While a nonlinear relationship between the hazard function and the predictors is assumed, the hazard ratio comparing any two observations is in fact constant over time in the setting where the predictor variables do not vary over time. This assumption is called the proportional hazards assumption and checking if this assumption is met is an important part of a Cox regression analysis. It is by far the most popular model for survival data analysis and is implemented in a large number of statistical software packages, including SAS, STATA, SPSS, and JMP and R.

The above covers the simplest and most common methods for analyzing right-censored survival data. Other methods of analysis exist, as well as methods that handle different types of censoring and time-varying predictor variables. These are covered in the literature, with an overview of some of these methods being found in Hosmer and Lemeshow (2).

If you need assistance with a survival analysis problem, do not hesitate to contact a statistical consultant at the Cornell Statistical Consulting Unit.

References:
1. Kleinbaum D.G., Survival *Analysis, a self learning text, Springer-Verlag*, 1996.
2. Hosmer D.W., Lemeshow S., and May S., *Applied Survival Analysis: Regression Modeling of Time-to- Event Data*, Wiley, 2008.

Author: Simona Despa

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).