Chapter 10
Introduction to Survival Analysis Procedures

Chapter Table of Contents

Chapter 10
Introduction to Survival Analysis Procedures

Overview

Data that measure lifetime or the length of time until the occurrence of an event are called lifetime, failure time, or survival data. For example, variables of interest might be the lifetime of diesel engines, the length of time a person stayed on a job, or the survival time for heart transplant patients. Such data have special considerations that must be incorporated into any analysis.

Background

Survival data consist of a response variable that measures the duration of time until a specified event occurs (event time, failure time, or survival time) and possibly a set of independent variables thought to be associated with the failure time variable. These independent variables (concomitant variables, covariates, or prognostic factors) can be either discrete, such as sex or race, or continuous, such as age or temperature. The system that gives rise to the event of interest can be biological, as for most medical data, or physical, as for engineering data. The purpose of survival analysis is to model the underlying distribution of the failure time variable and to assess the dependence of the failure time variable on the independent variables.

An intrinsic characteristic of survival data is the possibility for censoring of observations, that is, the actual time until the event is not observed. Such censoring can arise from withdrawal from the experiment or termination of the experiment. Because the response is usually a duration, some of the possible events may not yet have occurred when the period for data collection has terminated. For example, clinical trials are conducted over a finite period of time with staggered entry of patients. That is, patients enter a clinical trial over time and thus the length of follow-up varies by individuals; consequently, the time to the event may not be ascertained on all patients in the study. Additionally, some of the responses may be lost to follow-up (for example, a participant may move or refuse to continue to participate) before termination of data collection. In either case, only a lower bound on the failure time of the censored observations is known. These observations are said to be *right censored*. Thus, an additional variable is incorporated into the analysis indicating which responses are observed event times and which are censored times. More generally, the failure time may only be known to be smaller than a given value (*left censored*) or known to be within a given interval (*interval censored*). There are numerous possible censoring schemes that arise in survival analyses. The monograph by Maddala (1983) discusses several related types of censoring situations, and the text by Kalbfleisch and Prentice (1980) also discusses several censoring schemes. Data with censored observations cannot be analyzed by ignoring the censored observations because, among other considerations, the longer-lived individuals are generally more likely to be censored. The method of analysis must take the censoring into account and correctly use the censored observations as well as the uncensored observations.

Another characteristic of survival data is that the response cannot be negative. This suggests that a transformation of the survival time such as a log transformation may be necessary or that specialized methods may be more appropriate than those that assume a normal distribution for the error term. It is especially important to check any underlying assumptions as a part of the analysis because some of the models used are very sensitive to these assumptions.

Survival Analysis Procedures

There are three SAS procedures for analyzing survival data: LIFEREG, LIFETEST and PHREG. PROC LIFETEST is a nonparametric procedure for estimating the distribution of survival time and testing the association of survival time with other variables. PROC LIFEREG and PROC PHREG are regression procedures for modeling the distribution of survival time with a set of concomitant variables.

The LIFEREG Procedure

The LIFEREG procedure fits parametric accelerated failure time models to survival data that may be left, right, or interval censored. The parametric model is of the form

$$y = \mathbf{x}'\beta + \sigma\epsilon$$

where y is usually the log of the failure time variable, $\mathbf{x}$ is a vector of covariate values, β is a vector of unknown regression parameters to be fit, σ is an unknown scale parameter, and ϵ is an error term. The baseline distribution of the error term can be specified as one of several possible distributions, including, but not limited to, the log normal, log logistic, and Weibull distributions. Several texts that discuss these parametric models are Nelson (1990), Lawless (1982), and Kalbfleish and Prentice (1980).

The LIFETEST Procedure

The LIFETEST procedure computes nonparametric estimates of the survival distribution function. You can request either the product-limit (Kaplan-Meier) or the life table (actuarial) estimate of the distribution. The texts by Cox and Oakes (1984) and Kalbfleisch and Prentice (1980) provide good discussions of the product-limit estimator, and the texts by Lee (1992) and Elandt-Johnson and Johnson (1980) include detailed discussions of the life table estimator. The procedure also computes rank tests of association of the survival time variable with other concomitant variables as given in Kalbfleish and Prentice (1980, Chapter 6).

The PHREG Procedure

The PHREG procedure fits the proportional hazards model of Cox (1972, 1975) to survival data that may be right censored. The Cox model is a semiparametric model in which the hazard function of the survival time is given by

$$h(t|x) = h_0(t)\exp(\beta'\mathbf{x}(t))$$

where $h_0(t)$ is an unspecified baseline hazard function, $\mathbf{x}(t)$ is a vectors of covariate values, possibly time-dependent, and β is a vector of unknown regression parameters. The model is referred to as a semiparametric model since part of the model involves the unspecified baseline function over time (which is infinite dimensional) and the other part involves a finite number of regression parameters. Several texts that discuss the Cox regression models are Collett (1994), Cox and Oaks (1984), Lawless (1982), Kalbfleish and Prentice (1980).

Survival Analysis with SAS/STAT Procedures

The typical goal in survival analysis is to characterize the distribution of the survival time for a given population, to compare this survival time among different groups, or to study the relationship between the survival time and some concomitant variables.

A first step in the analysis of a set of survival data is to use PROC LIFETEST to compute and plot the estimate of the distribution of the survival time. The association between covariates and the survival time variable can be investigated by computing estimates of the survival distribution function within strata defined by the covariates. In particular, if the proportional hazards model is appropriate, the estimates of the log(-log(SURVIVAL)) plotted against the log(TIME) variable should give approximately parallel lines, where SURVIVAL is the survival distribution estimate and TIME is the failure time variable. Additionally, these lines should be approximately straight if the Weibull model is appropriate.

Statistics that test for association between failure time and covariates can be used to select covariates for further investigation. The LIFETEST procedure computes linear rank statistics using either Wilcoxon or log-rank scores. These statistics and their estimated covariance matrix can be used with the REG procedure with the option METHOD=RSQUARE to find the subset of variables that produce the largest joint test statistic for association. An example of this method of variable selection is given in the “Examples” section of Chapter 37, “The LIFETEST Procedure.”

Another approach to examine the relationship between the concomitant variables and survival time is through a regression model in which the survival time has a distribution that depends on the concomitant variables. The regression coefficients may be interpreted as describing the direction and strength of the relationship of each explanatory variable on the effect of the survival time.

In many biological systems, the Cox model may be a reasonable description of the relationship between the distribution of the survival time and the prognostic factors. You use PROC PHREG to fit the Cox regression model. The regression coefficient is interpreted as the increase of the log hazard ratio resulting in the increase of one unit in the covariate. However, the underlying hazard function is left unspecified and, as in any other model, the results can be misleading if the proportional hazards assumptions do not hold.

Accelerated failure time models are popular for survival data of physical systems. In many cases, the underlying survival distribution is known empirically. You use PROC LIFEREG to fit these parametric models. Also, PROC LIFEREG can accommodate data with interval-censored observations, which are not allowed in PROC PHREG.

A common technique for checking the validity of a regression model is to embed it in a larger model and use the likelihood ratio test to check whether the reduction to the actual model is valid. Other techniques include examining the residuals. Both PROC LIFEREG and PROC PHREG produce predicted values, residuals, and other computed values that can be used to assess the model adequacy.

References

Collett, D. (1994), *Modelling Survival Data in Medical Research*, London: Chapman and Hall.

Cox, D.R. (1972), “Regression Models and Life-Tables (with Discussions),” *Journal of Royal Statistical Society. Series B*, 34, 187–220.

Cox, D.R. (1975), “Partial Likelihoods,” *Biometrika*, 62, 269–276.

Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.

Elandt-Johnson, R.C. and Johnson, N.L. (1980), *Survival Models and Data Analysis*, New York: John Wiley & Sons, Inc.

Gross, A.J. and Clark, V.A. (1975), *Survival Distributions: Reliability Applications in the Biomedical Sciences*, New York: John Wiley & Sons, Inc.

Kalbfleisch, J.D. and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons, Inc.

Lawless, J.E. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons, Inc.

Lee, E.T. (1992), *Statistical Methods for Survival Data Analysis*, Second Edition, New York: John Wiley & Sons, Inc..

Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press.

Nelson, W. (1990) *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, New York: John Wiley & Sons, Inc.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.