SAS/STAT® 9.2 User's Guide

Introduction to Survival Analysis Procedures
(Book Excerpt)

Chapter 13

Introduction to Survival Analysis Procedures

Contents

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.

Survival data consist of a response (event time, failure time, or survival time) variable that measures the duration of time until a specified event occurs 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 to indicate which failure times are observed event times and which are censored times. More generally, the failure time might 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 analysis. 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 right 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 might be necessary or that specialized methods might 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 LIFEREG is a parametric regression procedure for modeling the distribution of survival time with a set of concomitant variables. PROC LIFETEST is a nonparametric procedure for estimating the survivor function, comparing the underlying survival curves of two or more samples, and testing the association of survival time with other variables. PROC PHREG is a semiparametric procedure that fits the Cox proportional hazards model and its extensions.

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

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

where y is usually the log of the failure time variable, $\mathbf{x}$ is a vector of covariate values, $\boldsymbol{\beta}$ is a vector of unknown regression parameters, 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 Kalbfleisch and Prentice (1980), Lawless (1982), Nelson (1990), and Meeker and Escobar (1998). For more information about PROC LIFEREG, see Chapter 48, "The LIFEREG Procedure."

The LIFETEST Procedure

The LIFETEST procedure computes nonparametric estimates of the survival distribution function. You can request either the product-limit (Kaplan and Meier 1958) 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. PROC LIFETEST computes nonparametric tests to compare the survival curves of two or more groups. The procedure also computes rank tests of association of the survival time variable with other concomitant variables as given in Kalbfleisch and Prentice (1980, Chapter 6). For more information about PROC LIFETEST, see Chapter 49, "The LIFETEST Procedure."

The PHREG Procedure

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

$$\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\boldsymbol{\beta}' \mathbf{x}(t)}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\mathbf{x}(t)$ is a vector of covariate values (possibly time-dependent), and $\boldsymbol{\beta}$ is a vector of unknown regression parameters. The model is referred to as a semiparametric model because part of the model involves the unspecified baseline function over time (which has an infinite dimension) and the other part involves a finite number of regression parameters. Several texts that discuss the Cox regression models are Collett (1994), Cox and Oakes (1984), Kalbfleisch and Prentice (1980), and Lawless (1982). Extensions of the Cox model are discussed in Therneau and Grambsch (2000), Andersen et al. (1992), and Fleming and Harrington (1991). For more information about PROC PHREG, see Chapter 64, "The PHREG Procedure."

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 the survival distributions 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. In many applications, there are often several survival curves to compare. For example, you want to compare the survival experiences of patients who receive different treatments for their disease. 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 illustration of this methodology is given in Example 49.1 of Chapter 49, "The LIFETEST Procedure."

Another approach to examining 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 can 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 might 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 left-censored or 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.

Bayesian Survival Analysis with SAS/STAT Procedures

Bayesian analysis of survival models can be requested in the LIFEREG and PHREG procedures. In addition to the Cox model, PROC PHREG also allow you to fit a piecewise exponential model. In Bayesian analysis, the model parameters are treated as random variables, and inference about parameters is based on the posterior distribution of the parameters. A posterior distribution is a weighted likelihood function of the data with a prior distribution of the parameters using the Bayes theorem. The prior distribution enables you to incorporate into the analysis knowledge or experience of the likely range of values of the parameters of interest. You can specify normal or uniform prior distributions for the model regression coefficients in both LIFEREG and PHREG procedures. In addition, you can specify a gamma or improper prior distribution for the scale or variance parameter

in PROC LIFEREG. For the piecewise exponential model in PROC PHREG, you can specify normal or uniform prior distributions for the log-hazard parameters; alternatively, you can specify gamma or improper prior distributions for the hazards parameters. If you have no prior knowledge of the parameter values, you can use a noninformative prior distribution, and the results of a Bayesian analysis will be very similar to a classical analysis based on maximum likelihood.

A closed form of the posterior distribution is often not feasible, and a Markov chain Monte Carlo method by Gibbs sampling is used to simulate samples from the posterior distribution. You can perform inference by using the simulated samples, for example, to estimate the probability that a function of the parameters of interest lies within a specified range of values.

See Chapter 7, "Introduction to Bayesian Analysis Procedures," for an introduction to the basic concepts of Bayesian statistics. Also see "Bayesian Analysis: Advantages and Disadvantages" on page 149 for a discussion of the advantages and disadvantages of Bayesian analysis. See Ibrahim, Chen, and Sinha (2001), Gelman et al. (2004), and Gilks, Richardson, and Spiegelhalter (1996) for more information about Bayesian analysis, including guidance about choosing prior distributions.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1992), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

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

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 20, 187–220, with discussion.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

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

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

Fleming, T. R. and Harrington, D. (1991), *Counting Processes and Survival Analysis*, New York: John Wiley & Sons.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Second Edition, London: Chapman & Hall.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.

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

Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

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

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

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

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

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

Therneau, T. M. and Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, New York: Springer-Verlag.

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:
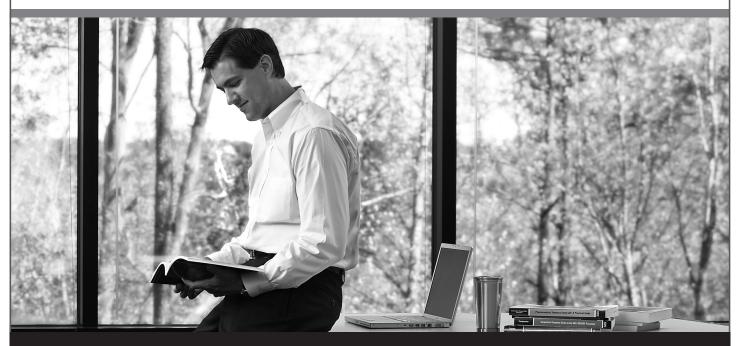
- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

§sas | THE POWER TO KNOW®