
# SAS/STAT® 15.1 User's Guide

# Introduction to Survival Analysis Procedures

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

**SAS/STAT® 15.1 User's Guide**



SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

# 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, a variable of interest might be the lifetime of diesel engines, the length of time a person stays at 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 that are 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 with most medical data) or physical (as with 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 of censoring of observations (that is, the actual time until the event is not observed). Such censoring can arise from withdrawal by a subject from the experiment or termination of the experiment. Because the response is usually a duration, some of the possible events might not yet have occurred when the period of data collection ends. 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 patient; consequently, the time to the event for all patients in the study might not be ascertained. In addition, some of the responses might be lost to follow-up

(for example, a participant might move or refuse to continue to participate) before data collection ends. In either case, only a lower bound on the failure time of the censored observations is known. Such 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 be known only to be smaller than a given value (*left-censored*) or known only to be within a given interval (*interval-censored*). Many possible censoring schemes arise in survival analysis. Maddala (1983) discusses several related types of censoring situations, and Kalbfleisch and Prentice (1980) also discuss several censoring schemes. Data that contains censored observations cannot be analyzed by ignoring the censored observations because, among other considerations, the longer-lived individuals are usually more likely to be right-censored. The method of analysis must take the censoring into account and correctly use both the censored observations and 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 of the error term. It is especially important to check any underlying assumptions as part of the analysis, because some of the models that are used are very sensitive to these assumptions.

## Survival Analysis Procedures

The following SAS/STAT procedures are specifically designed for analyzing survival data:

| | |
|---|---|
| ICLIFETEST | computes nonparametric estimates of survivor functions for interval-censored data. You can use this procedure to compare the underlying survival distributions of two or more samples of interval-censored data. |
| ICPHREG | fits proportional hazards regression models to interval-censored data. You can select a piecewise constant function as the baseline hazard function, or you can model the cumulative baseline hazard function by using a cubic spline or a discrete function. |
| LIFEREG | fits parametric models to failure time data that can be left-censored, right-censored, or interval-censored. The log of the survival time is modeled as a linear effect of covariates and a random disturbance term, the distribution of which includes the Weibull, log-normal, and log-logistic distributions. |
| LIFETEST | computes the Kaplan-Meier estimate of a survivor function and provides the log-rank test to compare the underlying hazards of two or more samples of right-censored data. You can also use this procedure to study the association between the failure time and a number of concomitant variables. |
| PHREG | fits the Cox proportional hazards model and its extensions, which include the multiplicative intensity model, the shared frailty model, and models for competing-risks data. |
| QUANTLIFE | performs quantile regression for survival data by modeling the quantiles of the lifetime variable as a function of the covariates. Because lifetime distributions are usually more skewed, the quantiles of the lifetime are more informative than the mean for summarizing the lifetime distribution. |

| | |
|---|---|
| RMSTREG | fits generalized linear models to the restricted mean survival time (RMST). The RMST, defined as the expected value of the time-to-event variable up to a prespecified time, is an interpretable summary measure of the time-to-event outcome. |
| SURVEYPHREG | is a Cox modeling procedure similar to PROC PHREG, appropriate for analyzing data that are collected from a survey sample. |

The SEVERITY procedure in SAS/ETS software is also a survival analysis procedure.

## Parametric Accelerated Failure Time Models: The LIFEREG Procedure

The LIFEREG procedure fits parametric accelerated failure time models to survival data that can 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, $\sigma$ is an unknown scale parameter, and $\epsilon$ 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. Texts that discuss these parametric models include Kalbfleisch and Prentice (1980); Lawless (1982); Nelson (1990); Meeker and Escobar (1998). For more information about PROC LIFEREG, see Chapter 73, "The LIFEREG Procedure."

## Nonparametric Methods for Right-Censored Data: 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. Cox and Oakes (1984) and Kalbfleisch and Prentice (1980) provide good discussions of the product-limit estimator, and 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 between the survival time variable and other concomitant variables, as given in Kalbfleisch and Prentice (1980, Chapter 6). For more information about PROC LIFETEST, see Chapter 74, "The LIFETEST Procedure."

## Nonparametric Methods for Interval-Censored Data: The ICLIFETEST Procedure

The ICLIFETEST procedure computes nonparametric estimates of the survival functions and examines the equality of the survival functions via statistical tests. PROC ICLIFETEST is intended primarily for handling interval-censored data, whereas the LIFETEST procedure deals exclusively with right-censored data. You can use PROC ICLIFETEST to analyze data that are left-censored, interval-censored, or right-censored. However, if the data to be analyzed contain only exact or right-censored observations, it is recommended that PROC LIFETEST be used because it provides specialized methods for dealing with right-censored data. For more information about PROC ICLIFETEST, see Chapter 66, "The ICLIFETEST Procedure."

## Proportional Hazards Regression for Interval-Censored Data: The ICPHREG Procedure

The ICPHREG procedure fits proportional hazards regression models to interval-censored data, including left-censored data and right-censored data as special cases. You can select a piecewise constant function as the baseline hazard, you can model the log of the cumulative baseline hazard function by using a cubic spline (Royston and Parmar 2002), or you can assume a discrete cumulative hazard function in which jumps are identified by the Turnbull method (Finkelstein 1986). PROC ICPHREG estimates the regression coefficients and the hazard parameters by maximizing the full likelihood function. For more information about PROC ICPHREG, see Chapter 67, "The ICPHREG Procedure."

## Quantile Regression: The QUANTLIFE Procedure

The QUANTLIFE procedure explores how the conditional quantile of the failure time variable depends on covariates. Quantile regression provides a flexible way to capture heterogeneous effects, in the sense that the tails and the central location of the conditional distributions can vary differently with the covariates. Thus, quantile regression offers a powerful tool in survival analysis, where the lifetimes are skewed and extreme survival times can be of special interest (Koenker and Geling 2001; Huang 2010). For more information about PROC QUANTLIFE, see Chapter 99, "The QUANTLIFE Procedure."

## Cox Regression and Extensions: 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

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

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. Texts that discuss the Cox regression models include Collett (1994); Cox and Oakes (1984); Kalbfleisch and Prentice (1980); Lawless (1982). Extensions of the Cox model are discussed in Therneau and Grambsch (2000); Andersen et al. (1992); Fleming and Harrington (1991); Fine and Gray (1999). For more information about PROC PHREG, see Chapter 89, "The PHREG Procedure."

## Restricted Mean Survival Time Regression: The RMSTREG Procedure

The RSMTREG procedure analyzes time-to-event data by performing a regression analysis on the restricted mean survival time (RMST), which is the expected value of the time-to-event variable up to a prespecified time point. The regression model is a generalized linear model that models the RMST directly and supports classification and interaction effects, along with the linear and log link functions. Models are fitted by using

methods based on the estimating equations, particularly the pseudovalue (PV) method (Andersen, Hansen, and Klein 2004) and the inverse probability censoring weighting (IPCW) method (Tian, Zhao, and Wei 2014).

## Cox Regression for Survey Data: The SURVEYPHREG Procedure

The SURVEYPHREG procedure fits the Cox proportional hazards model to sample survey data. The procedure is similar to the PHREG procedure, except that it incorporates complex sample design information in the analysis. The proportional hazards regression coefficients are estimated by maximizing a partial pseudo-log-likelihood function that incorporates the sampling weights. PROC SURVEYPHREG provides design-based variance estimates, confidence intervals, and tests for the estimated regression coefficients. For more information about PROC SURVEYPHREG, see Chapter 119, "The SURVEYPHREG 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 analyzing of a set of survival data is to use the LIFETEST or ICLIFETEST procedure to compute and plot the estimate of the distribution of the survival time. In many applications, you often have several survival curves to compare. For example, you might want to compare the survival experiences of patients who receive different treatments for their disease. You can investigate the relationship between covariates and the survival time by computing estimates of the survival distribution function within strata that are 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) should give approximately parallel lines, where SURVIVAL is the survival distribution estimate and TIME is the failure time variable. In addition, these lines should be approximately straight if the Weibull model is appropriate.

You can use knowledge of the association between failure time and the concomitant variables to select covariates for further investigation. The LIFETEST procedure computes linear rank statistics by using Wilcoxon or log-rank scores. These statistics and their estimated covariance matrix can be used with the REG procedure and the METHOD=RSQUARE option to find the subset of variables that produce the largest joint test statistic for association. An illustration of this methodology is given in Example 74.1 of Chapter 74, "The LIFETEST Procedure."

Another approach to examining the relationship between survival time and the concomitant variables 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 effect of each explanatory variable on 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 that results 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.

When you have interval-censored data, it is difficult to fit the semiparametric Cox regression models. But you can use the ICPHREG procedure to fit a different proportional hazards model, and the regression coefficients of such a model can still be interpreted as log-hazard ratios.

Accelerated failure time models are popular for fitting survival data from 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 that contain left-censored or interval-censored observations, which PROC PHREG does not allow.

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's adequacy.

## Bayesian Survival Analysis with SAS/STAT Procedures

You can request Bayesian analysis of survival models in the LIFEREG and PHREG procedures. In addition to enabling you to fit the Cox model, PROC PHREG also enables 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 by using the Bayes theorem. The prior distribution enables you to incorporate into your 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 the 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 hazard 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 are very similar to those of 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 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.

For an introduction to the basic concepts of Bayesian statistics, see Chapter 7, "Introduction to Bayesian Analysis Procedures." For a discussion of the advantages and disadvantages of Bayesian analysis, see "Bayesian Analysis: Advantages and Disadvantages" on page 136 in Chapter 7, "Introduction to Bayesian Analysis Procedures." For more information about Bayesian analysis, including guidance about choosing prior distributions, see Ibrahim, Chen, and Sinha (2001); Gelman et al. (2004); Gilks, Richardson, and Spiegelhalter (1996).

# References


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

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). "Regression Analysis of Restricted Mean Survival Time Based on Pseudo-observations." *Lifetime Data Analysis* 10:335–350.

Collett, D. (1994). *Modelling 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* 34:187–220. 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.

Fine, J. P., and Gray, R. J. (1999). "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94:496–509.

Finkelstein, D. M. (1986). "A Proportional Hazards Model for Interval-Censored Failure Time Data." *Biometrics* 42:845–854.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. 2nd ed. London: Chapman & Hall.

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

Huang, Y. (2010). "Quantile Calculus and Censored Regression." *Annals of Statistics* 38:1607–1637.

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.

Koenker, R., and Geling, O. (2001). "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis." *Journal of the American Statistical Association* 96:458–468.

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*. 2nd ed. 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.

Royston, P., and Parmar, M. K. B. (2002). "Flexible Parametric Proportional-Hazards and Proportional-Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects." *Statistics in Medicine* 21:2175–2197.

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

Tian, L., Zhao, L., and Wei, L. J. (2014). "Predicting the Restricted Mean Event Time with the Subject's Baseline Covariates in Survival Analysis." *Biostatistics* 15:222–233.