

# THE STATA JOURNAL

## Editor

H. Joseph Newton  
Department of Statistics  
Texas A & M University  
College Station, Texas 77843  
979-845-3142  
979-845-3144 FAX  
jnewton@stata-journal.com

## Executive Editor

Nicholas J. Cox  
Department of Geography  
University of Durham  
South Road  
Durham City DH1 3LE  
United Kingdom  
n.j.cox@stata-journal.com

## Associate Editors

Christopher Baum  
Boston College

Rino Bellocco  
Karolinska Institutet

David Clayton  
Cambridge Inst. for Medical Research

Charles Franklin  
University of Wisconsin, Madison

Joanne M. Garrett  
University of North Carolina

Allan Gregory  
Queen's University

James Hardin  
Texas A&M University

Stephen Jenkins  
University of Essex

Jens Lauritsen  
Odense University Hospital

Stanley Lemeshow  
Ohio State University

J. Scott Long  
Indiana University

Thomas Lumley  
University of Washington, Seattle

Marcello Pagano  
Harvard School of Public Health

Sophia Rabe-Hesketh  
Inst. of Psychiatry, King's College London

J. Patrick Royston  
MRC Clinical Trials Unit, London

Philip Ryan  
University of Adelaide

Jeroen Weesie  
Utrecht University

Jeffrey Wooldridge  
Michigan State University

**Copyright Statement:** The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by Stata Corporation. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from Stata Corporation if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publically accessible web sites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or Stata Corporation. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The *Stata Technical Journal* (ISSN 1536-867X) is a publication of Stata Press, and Stata is a registered trademark of Stata Corporation.

# Review of An Introduction to Survival Analysis Using Stata

David W. Hosmer  
University of Massachusetts  
hosmer@schoolph.umass.edu

**Abstract.** The new book by Cleves et al. (2002) is reviewed.

**Keywords:** gn0000, survival time regression models, time to event analysis

## 1 Introduction

The Stata staff principally responsible for the `st` suite of programs in Stata, Mario A. Cleves, William W. Gould, and Roberto G. Gutierrez, have written a book that builds on a web course they offer on survival analysis. The preface states that the intended audience is a professional data analyst. As such, the goal is for this reader to emerge with a clear understanding of what a survival analysis is, the models, their estimators and what information they are based on. The book has 15 chapters covering essentially four broad areas: seven chapters discussing survival data and how to use the `stset` command to describe survival data to Stata; one chapter on classic nonparametric methods; three chapters dealing with the Cox proportional hazards model; and four chapters dealing with parametric regression models. My comments will address each of the four broad areas.

## 2 Survival analysis, the data, and using `stset`

The authors begin by using a modeling approach to compare and contrast analysis of survival time data and the usual normal errors linear regression model and to compare the three methods of analysis to be discussed in the book: nonparametric, semiparametric, and fully parametric models. This is followed by a careful presentation of the various functions of time used in survival analysis: survival, hazard, and cumulative hazard. The numerical examples on interpretation of these functions will be quite useful to readers, especially those new to the concept of hazard. This is followed by a careful discussion of analysis time, another topic that someone new to survival time will find most helpful.

The chapter on censoring and truncation was mixed in my opinion. The treatments of right censoring and left truncation (delayed entry) were clear and easy to follow. The discussions of left censoring and right truncation were too brief to be informative or useful. Granted these do not occur frequently in practice, but the authors should have provided some references. (In my opinion, one of the biggest weaknesses of the book is the lack of references for further study or alternative points of view.)

The next three chapters consider how survival data is likely to be recorded in practice, how to describe the data using `stset`, and how to interpret output from `stset`. Since `stset` may be Stata's most complicated command, the effort placed here to demystify it will be especially appreciated by users new to Stata's survival analysis commands.

At the conclusion of the first seven chapters, readers should have a good idea as to what constitutes a survival analysis and how to get their data into the format required by the `st` suite of commands.

### **3 Nonparametric survival time methods**

The authors present in Chapter 8 standard nonparametric methods such as the Kaplan–Meier estimator, the Nelson–Aalen estimator, log-rank tests etc. One topic not discussed is estimation, point and interval, of survival-time quantiles. Since these are routinely reported in the subject-matter literature in conjunction with Kaplan–Meier curves, the omission, in my opinion, is substantial. Again, this is a topic with no references.

### **4 The Cox proportional hazards model**

In the next three chapters, the authors discuss the Cox proportional hazards model, estimation of regression coefficients, and post-estimation of the baseline and covariate-adjusted survival function. In my opinion, a clear strength of the chapter is the discussion of estimation in the presence of ties. To my knowledge, Stata is the only package that provides two “exact” methods. I assume that the authors do not want to acknowledge their competition by name, but an oblique reference as to which exact method is used in (say) TBT would be useful and informative. In addition, more explicit guidance about which method to use to correct for ties, and when, would have been helpful, or at least some references discussing this topic.

The authors' chapter on model building (Chapter 10) is, in my opinion, the weakest chapter in the book. They really do not provide a discussion of model building, only relatively superficial discussions of inclusion of categorical and continuous covariates and interactions. Given that Stata is the only package with the method of fractional polynomials, I was surprised that they did not use the opportunity to toot their horn. One positive aspect is the enhanced discussion of the various ways to include time-varying covariates.

The final chapter dealing with the Cox model considers methods for assessing model adequacy and fit. Some of the methods discussed, for example, determining functional form, to my way of thinking are basic steps in a careful model building and should have been discussed in the previous chapter. The authors suggest using the `linktest` command to assess adequacy in specification of the linear predictor. I have never found this test to be especially useful. The more focused tests and methods found in fractional polynomials combined with appropriate graphical displays are most certainly going to find inadequacies in the linear predictor with greater power than the omnibus link test. The authors carefully present Stata's test for proportional hazards based on the Schoen-

feld residuals. One omission is any guidance on what to do when the methods indicate that the hazard may be nonproportional in one or more covariates. They consider this in the context of parametric models later in the book but leave the reader with little help in this chapter. The authors' discussion of goodness-of-fit is limited to graphical displays based on Cox–Snell residuals. Again, this is a topic where considerably more work has been done, much of which is easily done in Stata, and yet no references are provided. The discussion and presentation of diagnostics for influence is quite brief. Since I have the floor, I will ask why these authors of the `st` commands did not include an option for obtaining the influence measures similar to the options for obtaining the score and Schoenfeld residuals. If these measures are useful and important, shouldn't Stata do all the calculations for us?

In summary, the three chapters dealing with the Cox proportional hazards model provide a reasonable introduction to modeling survival data with this much used and important model. However, considerably more could have been done, or at the very least a more inclusive set of references given for additional information.

## 5 Parametric survival-time models

In my opinion, the chapters on parametric models are the best in the book. This may be due to the fact that being more familiar with the Cox model, I learned the most from them. I found especially informative the explanations in Chapter 12 on the differences in the data and likelihood functions used by the semiparametric Cox model and parametric models likelihood and the comparison of the proportional hazards versus accelerated failure-time parameterizations. In Chapter 13 the authors compare the different parametric survival-time models available in Stata: exponential, Weibull, Gompertz, log-normal, gamma, and log-logistic. They have put together a nice collection of numerical examples that use parametric models to explore and then model a variety of different-shaped hazard functions. In Chapter 14 they discuss the various functions and statistics that can be obtained after a model has been fit. Goodness of fit and diagnostics are not specifically discussed, and the reader is referred to the chapter on the Cox model, which, as noted above, is not especially well done. Chapter 15 considers extensions such as stratified models and random effects–frailty parametric models. The discussion of random effects–frailty parametric models is well done. Given current interest in these models, it is not clear to me why frailty model extensions were not provided in Stata for the Cox model. Soon to come?

## 6 Summary and conclusions

In summary, I think the authors provide a good overview of survival analysis with Stata. The strengths of the book are the discussions of the `stset` command, methods for time-varying covariates, and the comparisons and modeling tricks shown in the chapters on parametric models. While not comprehensive enough to use as a stand-alone reference, I think the book would be a valuable addition to the library of analysts who use Stata to perform survival analyses.

## 7 References

Cleves, M., W. Gould, and R. Gutierrez. 2002. *An Introduction to Survival Analysis Using Stata*. College Station, TX: Stata Press.

### **About the Author**

David Hosmer is a professor of biostatistics in the Department of Biostatistics and Epidemiology of the University of Massachusetts School of Public Health and Health Sciences in Amherst, Massachusetts.