# The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

**Subscriptions** are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

| U.S. and Canada | | Elsewhere | |
|---|---|---|---|
| **Printed & electronic** | | **Printed & electronic** | |
| 1-year subscription | $ 98 | 1-year subscription | $138 |
| 2-year subscription | $165 | 2-year subscription | $245 |
| 3-year subscription | $225 | 3-year subscription | $345 |
| 1-year student subscription | $ 75 | 1-year student subscription | $ 99 |
| 1-year university library subscription | $125 | 1-year university library subscription | $165 |
| 2-year university library subscription | $215 | 2-year university library subscription | $295 |
| 3-year university library subscription | $315 | 3-year university library subscription | $435 |
| 1-year institutional subscription | $245 | 1-year institutional subscription | $285 |
| 2-year institutional subscription | $445 | 2-year institutional subscription | $525 |
| 3-year institutional subscription | $645 | 3-year institutional subscription | $765 |
| **Electronic only** | | **Electronic only** | |
| 1-year subscription | $ 75 | 1-year subscription | $ 75 |
| 2-year subscription | $125 | 2-year subscription | $125 |
| 3-year subscription | $165 | 3-year subscription | $165 |
| 1-year student subscription | $ 45 | 1-year student subscription | $ 45 |

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The Stata Journal (2013)
**13**, Number 1, pp. 212–216

# Review of Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model by Patrick Royston and Paul C. Lambert

Nicola Orsini
Unit of Nutritional Epidemiology
and
Unit of Biostatistics
Institute of Environmental Medicine, Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

**Abstract.** In this article, I review *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*, by Patrick Royston and Paul C. Lambert (2011 [Stata Press]).

**Keywords:** gn0057, flexible parametric survival models, survival analysis

## 1 Introduction

*Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*, by Patrick Royston and Paul C. Lambert, is a welcome complement to the existing documentation on survival analysis using Stata ([ST] and Cleves et al. [2010]). The authors clearly define the goal of the book—to describe and illustrate the use and applications of flexible parametric survival models—and most certainly succeed in achieving it. The commands used in this book have been written by the authors and their collaborators, and it is clear that the authors bring considerable experience in the development, application, and teaching of statistical methods to this book. With a clear pedagogical and practical approach, the key features of flexible parametric survival models are explained, and the context of application, with its challenges and problems, is carefully described. Implementation of the user-friendly estimation and postestimation commands is clearly illustrated. Throughout the book, a rich set of worked examples and graphical representations of different quantities of interest in survival analysis are used to aid readers. At every stage, the similarities and differences between flexible parametric models and more traditional parametric (Poisson, Weibull), semiparametric (Cox), and nonparametric (Kaplan–Meier) models are highlighted. This book provides a thorough introduction to flexible parametric survival models and is suitable for data analysts regardless of their background or experience with Stata. It is also suitable for use as a textbook in a course on survival analysis.

## 2 Overview of the book

Chapter 1 establishes the goals of the book, describes what the authors mean by going beyond the Cox model, and explains why one should be interested in knowing how to do that. After a brief review of the Cox model and its proportionality assumption, the authors illustrate how to obtain, represent graphically, and interpret the baseline hazard. The advantages listed in a concise paragraph are smoothed baseline hazard and survival functions, time-dependent associations, modeling on different scales, relative survival, out-of-sample prediction, and modeling on multiple time scales.

Chapter 2 presents two important and helpful Stata commands, `stset` and `stsplit`. This chapter can be very useful for readers not yet familiar with the suite of `st` commands. The authors explain some `stset` options commonly used in survival analysis, such as defining the units of time (months, years), restricting the follow-up time (a maximum of five years), and defining the time scale (time since entry, time since diagnosis, time from birth). The `stsplit` command is demonstrated with associations (time-dependent regression coefficients) or predictors (time-varying covariates) that change with the time scale. Several worked examples illustrating the use of the `stsplit` command, followed by estimation of Poisson, Cox, and Royston–Parmar models, can be found in chapters 4 and 7.

Chapter 3 describes the main motivating examples used in the book. Briefly, the outcomes are time from surgery to recurrence of or death from breast cancer (Rotterdam breast cancer data), time from breast cancer diagnosis to death (England and Wales breast cancer data), and time from birth to hip fractures (orchiectomy data). All of them are relatively large observational studies and are used to explain the advantages offered by flexible parametric models and their predictions in the analysis of such outcomes.

Chapter 4 shows how to use Poisson regression to flexibly model rates of an event. This is achieved by splitting the time scale in different ways. The authors start with a simple model, assuming that the baseline rate is constant throughout the time scale (exponential model). Next they allow the baseline rate to vary over the time scale by splitting the time scale into a number of intervals and then modeling changes of the rate over time with indicator variables (piecewise exponential model). By finely splitting the time scale, they can model the baseline rate like any other quantitative covariate by using fractional polynomials or splines. The relation between Poisson and Cox regression is illustrated using both examples and formulas. The chapter concludes with a useful discussion of possible advantages and disadvantages of using Poisson regression to model survival data.

Chapter 5 introduces Royston–Parmar regression, which generalizes the standard parametric regression models (Weibull, loglogistic, lognormal). These models implemented in the `stpm2` command play a central role in this book. The main advantage of this approach over Poisson regression is that it uses one row for each individual (which avoids data augmentation implied by `stsplit`) and therefore makes postestimation commands easier to use. Royston–Parmar models are motivated by showing the lack of fit of standard parametric models (exponential, Weibull) in an attempt to model years from surgery in the Rotterdam breast cancer data. The flexibility is obtained by modeling the log cumulative-hazard function as a smooth function of the log of time. A set of covariates is then added to the linear predictor for the log cumulative hazard, so the proportionality assumption for the covariates (regression coefficients that do not change with the time scale) continues to hold with Royston–Parmar regression. Nonproportionality, however, can be modeled, as shown later in chapter 7. Because flexibility is achieved by using spline transformations, the authors describe in detail how those spline transformations—more specifically, restricted cubic splines—are generated and how the choice of the spline model may affect the results. It is noteworthy that the findings based on Royston–Parmar models appear to be fairly insensitive to the number and, particularly, the location of the knots. The chapter continues with a more concise generalization of the loglogistic and lognormal (probit) models, demonstrating that results from standard parametric models can be obtained by choosing 1 degree of freedom for spline transformations. The Aranda–Ordaz family of link functions for survival models, of which cumulative hazards and cumulative odds are special cases, is described in paragraph 5.6 with some useful advice on how to choose between them.

Chapter 6 focuses on prognostic models, essentially multivariable regression models used to predict the occurrence of future outcomes. It discusses the different aspects involved in developing and reporting a prognostic model. The various steps required to build a multivariable model, such as the choice of a suitable scale (cumulative hazard versus cumulative odds) and spline transformations, the selection of covariates, and the functional form for quantitative predictors, are illustrated with the Rotterdam breast cancer data as the motivating example. Once a final model is obtained, the authors emphasize the benefits provided by using flexible parametric Royston–Parmar models. Those benefits include being able to easily estimate a variety of interpretable quantities—unconditional and conditional survival probabilities, survival probabilities at specified centiles of the prognostic index, survival probabilities at given covariate values, differences between survival probabilities, centiles of the predicted survival distribution—by using the helpful `predict` postestimation command. The chapter then describes how to assess the goodness of fit of the multivariable model based on residuals and some summary measure of discrimination and explained variation. The final two paragraphs are dedicated to out-of-sample predictions (to interpolate and extrapolate beyond the observed time points or to validate a prognostic model) and imputations of censored survival times (to visualize the relation between survival time and the prognostic index).

Chapter 7 illustrates how to estimate and present survival models when one or more regression coefficients vary with the time scale; this is known as time-dependent effects or nonproportionality. The nonproportionality is modeled within different frameworks: Cox, Poisson, and Royston–Parmar regression. Each regression model can be extended to model the fact that a regression coefficient is not constant over the time scale. It is clear that the differences across approaches are in terms of flexibility when modeling the interaction between the covariate and time, the amount of data handling, and the possibility to easily predict different quantities of interest such as survival probabilities, hazard rates, hazard ratios, and hazard differences. Thanks to powerful estimation and postestimation commands, the flexible parametric Royston–Parmar models perform the best in all aspects. The possibility to control separately the complexity of spline transformations for the baseline distribution function and each of the time-dependent regression coefficients could not be simpler with the `stpm2` command. The family of Royston–Parmar models can offer an additional way of handling nonproportionality, that is, of changing the scale. Several times, the chapter emphasizes that proportionality of the regression coefficient is scale dependent. For instance, if cumulative odds are proportional, the hazard rates are unlikely to be proportional, and the hazard ratio will converge toward the null, that is, 1, as the follow-up time increases. The analysis of the orchiectomy data provides an excellent example of how multiple time scales can be defined (attained age and time since diagnosis).

Chapter 8 describes relative survival for the analysis of population-based cancer studies. The idea is to compare the observed survival experience of patients with the expected survival of the general populations as provided by national registries. The related concepts of excess mortality and relative survival are explained and illustrated with the England and Wales breast cancer data. The chapter begins with a review of the traditional life-table approach for estimating relative survival based on the user-written `strs` command. Next the Poisson regression model is extended to include expected mortality. This is accomplished in the generalized linear model framework, which allows user-written link functions. The Royston–Parmar models are also extended to incorporate information on expected mortality, providing a unique and easy-to-use framework that works seamlessly for all-cause, cause-specific, and relative survival. Modeling issues in relative survival are similar to the standard survival analysis illustrated in previous chapters. As nonproportional excess hazard is common in cancer studies, the advantages offered by Royston–Parmar models in terms of reduced data handling, easy model specification, and a rich set of predictions (excess mortality rates, excess mortality rate ratios or differences, relative survival, and relative survival difference) are once again evident.

Chapter 9 describes how flexible parametric models can be useful to estimate relevant quantities in a variety of contexts. In clinical trials, the number needed to treat as a function of time can be estimated by the inverse of the difference in predicted survival probabilities (the option `stdiff` of `predict`). When one presents survival curves after multivariable models, the predicted survival probabilities as a function of time can be obtained with the mean covariate method (the options `survival` and `at()` of `predict`) or with the direct method (the options `meansurv` and `at()` of `predict`). When one models a continuous outcome with no censoring and with a distribution that varies with a continuous covariate, an outcome-dependent association can be handled with an outcome-varying regression coefficient (the options `tvc()` and `dftvc()` of `stpm2`). In the case of multiple events, different kinds of marginal models that take into account dependent outcome data are estimated within the framework of Royston–Parmar models. These allow for delayed entry, robust cluster standard errors, and event-specific baseline hazards. To help those considering a Bayesian approach in survival analysis, the chapter illustrates step by step how to fit a Royston–Parmar model in WinBUGS from Stata. In the presence of competing risks, modeling a cause-specific hazard is done with Royston–Parmar models, from which one can easily predict hazard and survival and obtain the cumulative incidence function by numerical integration. In the analysis of population-based cancer studies, current period analysis estimates of patient survival and crude probability of death can be computed once again by using a flexible parametric approach.

## 3 Conclusion

This insightful book clearly explains what flexible parametric survival models are and what they can offer compared with traditional methods. The main strength of this book is that it gently introduces the reader to the significant advantages of a flexible parametric approach in survival analysis. These advantages can be summarized in one word: predictions. The authors' deep knowledge of flexible parametric survival models is evident in the clear way that model specification, prediction, and presentation are described. The Stata code and reproducible examples available on the book's website, http://www.stata-press.com/data/fpsaus.html, can greatly facilitate the application of flexible parametric models in different areas, especially medical research. I highly recommend the methods presented in the book to any professional data analyst dealing with time-to-event, eventually censored, outcomes. Because this is an area of active research, I am confident that the authors will continue expanding the set of tools and range of applications presented in this book.

## 4 References

Cleves, M., W. Gould, R. G. Gutierrez, and Y. V. Marchenko. 2010. *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Royston, P., and P. C. Lambert. 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press.

### About the author

Nicola Orsini is an associate professor of medical statistics and an assistant professor of epidemiology in the Unit of Biostatistics and Unit of Nutritional Epidemiology at the Institute of Environmental Medicine, Karolinska Institutet, Sweden.