Journal of Statistical Software

April 2016, Volume 70, Book Review 2. doi: 10.18637/jss.v070.b02

Reviewer: James E. Helmreich
Marist College

Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression and Survival Analysis (2nd Edition)

Frank E. Harrell, Jr.
Springer-Verlag, Cham, 2015.
ISBN 978-3-319-19424-0. 582 pp. USD 89.99 (H).
http://biostat.mc.vanderbilt.edu/rms

Regression Modeling Strategies is an advanced text, aimed at graduate students and researchers with a solid, comprehensive background in regression modeling. Readers should have a good working knowledge of regression analysis as well as R as all code is written for that software. The extensive graphics are in color and make use of the package **ggplot2**. There is a companion package **rms**; **Hmisc** (Harrell's well-known miscellaneous package) is used frequently as well. All R commands for analyses are in the text and available at the companion website. There are extensive notes on further reading as well as exercises for students at the end of each chapter.

Harrell's text is a massively detailed work; an impressive compendium of current knowledge in the field with nearly 700 cited references, a third published since the first edition appeared. The text is worth the remarkably reasonable price for this bibliography alone. Yet the exposition and case study analyses are exceptional and are recommended reading for all those interested in the field. There are general discussions of various topics (missing data, modeling strategies, data reduction, validation, resampling, maximum likelihood estimation, binary and ordinal logistic regression, survival analysis etc.), with each thoroughly presented, referenced, and then used in the beautiful, chapter length case studies.

The intent of the book is, as the title suggests, overall strategies for regression modeling, as opposed to isolated techniques and examples. Harrell's main emphasis is on model building for predictive purposes, but he takes pains to point out how effect estimation and various tests are done within the framework of a model. The emphasis is on the process – what the steps are, the choices available at each stage, evaluation of those choices and models as well as implications of those models. As an example, missing data and imputation issues seem rarely to be incorporated into a text on regression, and books that discuss these issues rarely include much about modeling. Here, however, Harrell provides a comprehensive early consideration of them all (Chapter 3). The discussion is careful to put the actions taken in context: what one does is contingent on why and how the data values are missing. Similarly, data reduction (remove unimportant, uninformative, highly missing variates, use PSA techniques) is introduced early,

in Chapter 4, to reflect the number of parameters that can be reasonably estimated for a model given the size of the dataset. Strong emphasis is put upon degrees of freedom, that they depend on the the total number of predictors and covariates modeled, not just the number decided upon after model building. These early chapters are not so much about the *how*, but the *what to consider* and *why*. Immediately following these first chapters is a case study on data reduction which includes sections on imputation. Other case studies generally begin with these topics as well.

Case studies use several large and rich datasets, primarily from the health care field (prostate cancer, meningitis, survival and costs of in-hospital patients etc.). These extensive case studies will be extremely useful for students of that field, but the methods of analysis are certainly generally applicable to other fields. The R code for analyses is shown, as well as extensive color graphics used for the analyses.

Chapters are intensively annotated, footnoted and referenced to the literature. Book references frequently include specific sections and page numbers. If there are differences of opinion Harrell carefully notes them with a balanced discussion. For instance, there is a presentation of different definitions of 'effect size' that have been proposed by various authors. Frequent marginal references to the end of chapter provide suggestions for further reading – these can number in the dozens for some chapters. Overall, the book is an impressive description and synthesis of current research.

How case studies proceed depends on the dataset and goals, but a broad description of the processes presented is informative. Basic exploratory data analysis is conducted, and a discussion of the goals of the analysis is given. Appropriate models and the number of parameters that can be reasonably modeled are estimated. There is always a discussion of the degrees of freedom that are available and 'spent' in various ways. This may lead to issues with imputation, and extensive consideration of the reduction in the number of variables. In fact, the degrees available drive the need for data reduction and inform the choice of models first inspected. Models are suggested and assessed, then revised using multiple approaches. Usually these are non-additive, and use carefully chosen levels of splines and complex interactions. Assumptions are always explicitly stated and examined for validity. Various validation techniques are used; the bootstrap especially is employed to good effect. Warnings against overfitting of models are prominent. Graphical techniques are stressed at every step and are incorporated in an integral way. Assumptions are assessed, and if necessary or pedagogically appropriate entirely new approaches will be tried. The implications of the model are interpreted, various tests of hypotheses may be performed, and predictions are found and compared, especially graphically. This description really does not do full justice to the careful and thorough presentations as well as the nuanced discussions Harrell provides.

Despite this breadth of analysis the writing is terse and very precise. Especially towards the end of an analysis I found myself wishing for more elaboration, though this spurred me to more careful thought. There is an immense amount of information, experience and expertise packed into the 500 odd pages of this text that demands careful reading.

The packages used with the text are excellent. They contain a wealth of functions, not least of which are some that are 'wrappers' for functions found elsewhere. For example, quantile regression can be performed with `Rq` in **rms**, a nice extension of `rq` in Koenker's **quantreg**. These provide extensive additional analytic tools for confidence intervals, validations, and graphical analyses of a model. Harrell does a good job showing examples of their use throughout the

book and especially in case studies. In addition, the companion website is quite nice; it contains all R code given in the book, as well as a link to a course taught by the author using his text. The site is unusually detailed as well as carefully laid out; it is likely to be informative for instructors in the classroom.

Regression Modeling Strategies is a monumental scholarly work of the highest order. The literature discussions on every topic offer an amazing breadth and depth of scope. I would begin any serious investigation of a technique new to me with this text, especially as every technique is integrally embedded in the context of a broader analysis. In the age of Wikipedia, blogs, and online forums, one often wonders if the need for hardback books is coming to an end. This is an excellent counterexample to the trend, a book we should all have on our bookshelves and use frequently.

Reviewer:

James E. Helmreich
Marist College
Department of Mathematics
3399 North Road
Poughkeepsie, NY 12601, United States of America
E-mail: James.Helmreich@Marist.edu
URL: http://foxweb.marist.edu/users/james.helmreich/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
April 2016, Volume 70, Book Review 2 *Published:* 2016-04-01
doi:10.18637/jss.v070.b02