Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Jon Wakefield

Bayesian and Frequentist Regression Methods

Springer

Jon Wakefield
Departments of Statistics and Biostatistics
University of Washington
Seattle, Washington
USA

ISSN 0172-7397
ISBN 978-1-4419-0924-4 ISBN 978-1-4419-0925-1 (eBook)
DOI 10.1007/978-1-4419-0925-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012952935

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

In the order of my meeting them, this book is dedicated to:

Norma Maureen Wakefield
Eric Louis Wakefield
Samantha Louise Wakefield
Felicity Zoe Moodie
Eleanor Anna Wakefield
Eric Stephen Wakefield

Preface

The past 25 years have seen great advances in both Bayesian and frequentist methods for data analysis. The most significant advance for the Bayesian approach has been the development of Markov chain Monte Carlo methods for estimating expectations with respect to the posterior, hence allowing flexible inference and routine implementation for a wide range of models. In particular, this development has led to the more widespread use of hierarchical models for dependent data. With respect to frequentist methods, estimating functions have emerged as a unifying approach for determining the properties of estimators. Generalized estimating equations provide a particularly important example of this methodology that allows inference for dependent data.

The aim of this book is to provide a modern description of Bayesian and frequentist methods of regression analysis and to illustrate the use of these methods on real data. Many books describe one or the other of the Bayesian or frequentist approaches to regression modeling in different contexts, and many mathematical statistics texts describe the theory behind Bayesian and frequentist approaches without providing a detailed description of specific methods. References to such texts are given at the end of Chaps. 2 and 3. Bayesian and frequentist methods are not viewed here as competitive, but rather as complementary techniques, and in this respect this book has some uniqueness.

In embarking on the writing of this book, I have been influenced by many current and former colleagues. My early training was in the Mathematics Department at the University of Nottingham and my first permanent academic teaching position was in the Mathematics Department at Imperial College of Science, Technology and Medicine in London. During this period I was introduced to the Bayesian paradigm and was greatly influenced by Adrian Smith, both as a lecturer and as a Ph.D. adviser. I have also benefited, and continue to benefit, from numerous conversations with Dave Stephens who I have known for over 25 years. Following my move to the University of Washington in Seattle I was exposed to a very modern view of frequentist methods in the Department of Biostatistics. In particular, Scott Emerson, Patrick Heagerty and Thomas Lumley have provided constant stimulation. These interactions, among many others, have influenced the way I now think about statistics, and it is this exposure which I hope has allowed me to write a balanced account of Bayesian and frequentist methods. There is some theory in this book and some data analysis, but the focus is on material that lies between these endeavors and concerns methods. At the University of Washington there is an advanced three-course regression methods sequence and this book arose out of my teaching of the three courses in the sequence.

If modern computers had been available a 100 years ago, the discipline of statistics would have developed in a dramatically different fashion to the way in which it actually evolved. In particular, there would probably be less dependence on linear and generalized linear models, which are mathematically and computationally convenient. While these model classes are still useful and do possess a number of convenient mathematical and computational properties, I believe they should be viewed as just two choices within a far wider range of models that are now available. The approach to modeling that is encouraged in this book is to first specify the model suggested by the background science and to then proceed to examining the mathematical and computational aspects of the model.

As a preparation for this book, the reader is assumed to have a grasp of calculus and linear algebra and have taken first courses in probability and statistical theory. The content of this book is as follows. An introductory chapter describes a number of motivating examples and discusses general issues that need consideration before a regression analysis is carried out. This book is then broken into five parts: I, Inferential Approaches; II, Independent Data; III, Dependent Data; IV, Nonparametric Modeling; V, Appendices. The first two chapters of Part I provide descriptions of the frequentist and Bayesian approaches to inference, with a particular emphasis on the rationale of each approach and a delineation of situations in which one or the other approach is preferable. The third chapter in Part I discusses model selection and hypothesis testing. Part II considers independent data and contains three chapters on the linear model, general regression models (including generalized linear models), and binary data models. The two chapters of Part III consider dependent data with linear models and general regression models. Mixed models and generalized estimating equations are the approaches to inference that are emphasized. Part IV contains three chapters on nonparametric modeling with an emphasis on spline and kernel methods. The examples and simulation studies of this book were almost exclusively carried out within the freely available `R` programming environment. The code for the examples and figures may be found at:

http://faculty.washington.edu/jonno/regression-methods.html

along with the inevitable errata and links to datasets. Exercises are included at the end of all chapters but the first. Many of these exercises concern analyses of real data. In my own experience, a full understanding of methods requires their implementation and application to data.

In my own teaching I have based three one-quarter courses on the following. *Regression Methods for Independent Data* is based on Part II, dipping into topics in Part I as needed and using motivating examples from Chap. 1. *Regression Methods* *for Dependent Data* centers on Part II, again using examples from Chap. 1, and building on the independent data material. Finally, *Nonparametric Regression and Classification* is based on the material in Part IV. The latter course is stand-alone in the sense of not requiring the independent and dependent data courses though extra material on a number of topics, including linear and generalized linear models and mixed models, will need to be included if not previously encountered.

In the 2003–2004 academic year I was the Genentech Professor and received funding specifically to work on this book. The staff at Springer have been very helpful at all stages. John Kimmel was the editor during most of the writing of this book and I am appreciative of his gentle prodding and advice. About 18 months from the completion of this book, Marc Strauss stepped in and has also been very supportive. Many of my colleagues have given comments on various chapters, but I would like to specifically thank Lurdes Inoue, Katie Kerr, Erica Moodie, Zoe Moodie, Ken Rice, Dave Stephens, Jon Wellner, Daniela Witten, and Simon Wood for feedback on different parts of this book. Finally, lest we forget, I would like to thank all of those students who suffered through initial presentations of this material—I hope your sacrifices were not in vain…

Seattle, WA
June 2012

Jon Wakefield

Contents