Statistics for Social and Behavioral Sciences

Advisors:
S.E. Fienberg W.J. van der Linden

Statistics for Social and Behavioral Sciences

Brennan: Generalizability Theory
DeBoeck/Wilson: Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach
Devlin/Fienberg/Resnick/Roeder: Intelligence, Genes, and Success: Scientists Respond to The Bell Curve
Dorans/Pommerich/Holland: Linking and Aligning Scores and Scales
Finkelstein/Levin: Statistics for Lawyers, 2nd ed.
Gastwirth: Statistical Science in the Courtroom
Handcock/Morris: Relative Distribution Methods in the Social Sciences
Johnson/Albert: Ordinal Data Modeling
Kolen/Brennan: Test Equating, Scaling, and Linking: Methods and Practices, 2nd ed.
Longford: Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician
Lynch: Introduction to Applied Bayesian Statistics and Estimation for Social Scientists
Morton/Rolph: Public Policy and Statistics: Case Studies from RAND
van der Linden: Linear Models for Optimal Test Design
von Davier/Carstensen: Multivariate and Mixture Distribution Rasch Models
von Davier/Holland/Thayer: The Kernel Method of Test Equating
Zeisel/Kaye: Prove It with Figures: Empirical Methods in Law and Litigation

Scott M. Lynch

Introduction to Applied Bayesian Statistics and Estimation for Social Scientists

With 89 Figures

Springer

Scott M. Lynch
Department of Sociology and
Office of Population Research
Princeton University
Princeton, NJ 08544
slynch@princeton.edu

Series Editors:
Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213–3890
USA

Wim J. van der Linden
Department of Measurement
and Data Analysis
Faculty of Behavioral Sciences
University of Twente
7500 AE Enschede
The Netherlands

Library of Congress Control Number: 2007929729

ISBN 978-0-387-71264-2 e-ISBN 978-0-387-71265-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

For my Barbara

Preface

This book was written slowly over the course of the last five years. During that time, a number of advances have been made in Bayesian statistics and Markov chain Monte Carlo (MCMC) methods, but, in my opinion, the market still lacks a *truly* introductory book written explicitly for social scientists that thoroughly describes the actual process of Bayesian analysis using these methods. To be sure, a variety of introductory books are available that cover the basics of the Bayesian approach to statistics (e.g., Gill 2002 and Gelman et al. 1995) and several that cover the foundation of MCMC methods (e.g., beginning with Gilks et al. 1996). Yet, a highly applied book showing how to use MCMC methods to complete a Bayesian analysis involving typical social science models applied to typical social science data is still sorely lacking. The goal of this book is to fill this niche.

The Bayesian approach to statistics has a long history in the discipline of statistics, but prior to the 1990s, it held a marginal, almost cult-like status in the discipline and was almost unheard of in social science methodology. The primary reasons for the marginal status of the Bayesian approach include (1) philosophical opposition to the use of "prior distributions" in particular and the subjective approach to probability in general, and (2) the lack of computing power for completing realistic Bayesian analyses. In the 1990s, several events occurred simultaneously to overcome these concerns. First, the explosion in computing power nullified the second limitation of conducting Bayesian analyses, especially with the development of sampling based methods (e.g., MCMC methods) for estimating parameters of Bayesian models. Second, the growth in availability of longitudinal (panel) data and the rise in the use of hierarchical modeling made the Bayesian approach more appealing, because Bayesian statistics offers a natural approach to constructing hierarchical models. Third, there has been a growing recognition both that the enterprise of statistics is a subjective process in general and that the use of prior distributions need not influence results substantially. Additionally, in many problems, the use of a prior distribution turns out to be advantageous.

The publication of Gelfand and Smith's 1990 paper describing the use of MCMC simulation methods for summarizing Bayesian posterior distributions was the watershed event that launched MCMC methods into popularity in statistics. Following relatively closely on the heels of this article, Gelman et al.'s (1995) book, *Bayesian Data Analysis*, and Gilks et al.'s (1996) book, *Markov Chain Monte Carlo in Practice*, placed the Bayesian approach in general, and the application of MCMC methods to Bayesian statistical models, squarely in the mainstream of statistics. I consider these books to be classics in the field and rely heavily on them throughout this book.

Since the mid-1990s, there has been an explosion in advances in Bayesian statistics and especially MCMC methodology. Many improvements in the recent past have been in terms of (1) monitoring and improving the performance of MCMC algorithms and (2) the development of more refined and complex Bayesian models and MCMC algorithms tailored to specific problems. These advances have largely escaped mainstream social science.

In my view, these advances have gone largely unnoticed in social science, because purported introductory books on Bayesian statistics and MCMC methods are not truly introductory for this audience. First, the mathematics in introductory books is often too advanced for a mainstream social science audience, which begs the question: "introductory *for whom?*" Many social scientists do not have the probability theory and mathematical statistics background to follow many of these books beyond the first chapter. This is not to say that the material is impossible to follow, only that more detail may be needed to make the text and examples more readable for a mainstream social science audience.

Second, many examples in introductory-level Bayesian books are at best foreign and at worst irrelevant to social scientists. The probability distributions that are used in many examples are not typical probability distributions used by social scientists (e.g., Cauchy), and the data sets that are used in examples are often atypical of social science data. Specifically, many books use small data sets with a limited number of covariates, and many of the models are not typical of the regression-based approaches used in social science research. This fact may not seem problematic until, for example, one is faced with a research question requiring a multivariate regression model for 10,000 observations measured on 5 outcomes with 10 or more covariates. Nonetheless, research questions involving large-scale data sets are not uncommon in social science research, and methods shown that handle a sample of size 100 measured on one or two outcomes with a couple of covariates simply may not be directly transferrable to a larger data set context. In such cases, the analyst without a solid understanding of the linkage between the model and the estimation routine may be unable to complete the analysis. Thus, some discussion tailored to the practicalities of *real* social science data and computing is warranted.

Third, there seems to be a disjunction between introductory books on Bayesian theory and introductory books on applied Bayesian statistics. One of the greatest frustrations for me, while I was learning the basics of Bayesian statistics and MCMC estimation methods, was (and is) the lack of a book that links the theoretical aspects of Bayesian statistics and model development with the application of modern estimation methods. Some examples in extant books may be substantively interesting, but they are often incomplete in the sense that discussion is truncated after model development without adequate guidance regarding how to estimate parameters. Often, suggestions are made concerning how to go about implementing only certain aspects of an estimation routine, but for a person with no experience doing this, these suggestions are not enough.

In an attempt to remedy these issues, this book takes a step back from the most recent advances in Bayesian statistics and MCMC methods and tries to bridge the gap between Bayesian theory and modern Bayesian estimation methods, as well as to bridge the gap between Bayesian statistics books written as "introductory" texts for statisticians and the needs of a mainstream social science audience. To accomplish this goal, this book presents very little that is new. Indeed, most of the material in this book is now "old-hat" in statistics, and many references are a decade old (In fact, a second edition of Gelman et al.'s 1995 book is now available). However, the trade-off for not presenting much new material is that this book explains the process of Bayesian statistics and modern parameter estimation via MCMC simulation methods in great depth. Throughout the book, I painstakingly show the modeling process from model development, through development of an MCMC algorithm to estimate its parameters, through model evaluation, and through summarization and inference.

Although many introductory books begin with the assumption that the reader has a solid grasp of probability theory and mathematical statistics, I do not make that assumption. Instead, this book begins with an exposition of the probability theory needed to gain a solid understanding of the statistical analysis of data. In the early chapters, I use contrived examples applied to (sometimes) contrived data so that the forest is not lost for the trees: The goal is to provide an understanding of the issue at hand rather than to get lost in the idiosyncratic features of real data. In the latter chapters, I show a Bayesian approach (or approaches) to estimating some of the most common models in social science research, including the linear regression model, generalized linear models (specifically, dichotomous and ordinal probit models), hierarchical models, and multivariate models.

A consequence of this choice of models is that the parameter estimates obtained via the Bayesian approach are often very consistent with those that could be obtained via a classical approach. This may make a reader ask, "then what's the point?" First, there are many cases in which a Bayesian approach and a classical approach will not coincide, but from my perspective, an introductory text should establish a foundation that can be built upon, rather than beginning in unfamiliar territory. Second, there are additional benefits to taking a Bayesian approach beyond the simple estimation of model parameters. Specificially, the Bayesian approach allows for greater flexibility in evaluating model fit, comparing models, producing samples of parameters that are not directly estimated within a model, handling missing data, "tweaking" a model in ways that cannot be done using canned routines in existing software (e.g., freeing or imposing constraints), and making predictions/forecasts that capture greater uncertainty than classical methods. I discuss each of these benefits in the examples throughout the latter chapters.

Throughout the book I thoroughly flesh out each example, beginning with the development of the model and continuing through to developing an MCMC algorithm (generally in R) to estimate it, estimating it using the algorithm, and presenting and summarizing the results. These programs should be straightforward, albeit perhaps tedious, to replicate, but some programming is inherently required to conduct Bayesian analyses. However, once such programming skills are learned, they are incredibly freeing to the researcher and thus well worth the investment to acquire them. Ultimately, the point is that the examples are thoroughly detailed; nothing is left to the imagination or to guesswork, including the mathematical contortions of simplifying posterior distributions to make them recognizable as known distributions.

A key feature of Bayesian statistics, and a point of contention for opponents, is the use of a prior distribution. Indeed, one of the most complex things about Bayesian statistics is the development of a model that includes a prior and yields a "proper" posterior distribution. In this book, I do not concentrate much effort on developing priors. Often, I use uniform priors on most parameters in a model, or I use "reference" priors. Both types of priors generally have the effect of producing results roughly comparable with those obtained via maximum likelihood estimation (although not in interpretation!). My goal is not to minimize the importance of choosing appropriate priors, but instead it is not to overcomplicate an introductory exposition of Bayesian statistics and model estimation. The fact is that most Bayesian analyses explicitly attempt to minimize the effect of the prior. Most published applications to date have involved using uniform, reference, or otherwise "noninformative" priors in an effort to avoid the "subjectivity" criticism that historically has been levied against Bayesians by classical statisticians. Thus, in most Bayesian social science research, the prior has faded in its importance in differentiating the classical and Bayesian paradigms. This is not to say that prior distributions are unimportant—for some problems they may be very important or useful—but it is to say that it is not necessary to dwell on them.

The book consists of a total of 11 chapters plus two appendices covering (1) calculus and matrix algebra and (2) the basic concepts of the Central Limit Theorem. The book is suited for a highly applied one-semester graduate level social science course. Each chapter, including the appendix but excluding the introduction, contains a handful of exercises at the end that test the understanding of the material in the chapter at both theoretical and applied levels. In the exercises, I have traded quantity for quality: There are relatively few exercises, but each one was chosen to address the essential material in the chapter. The first half of the book (Chapters 1-6) is primarily theoretical and provides a generic introduction to the theory and methods of Bayesian statistics. These methods are then applied to common social science models and data in the latter half of the book (Chapters 7-11). Chapters 2-4 can each be covered in a week of classes, and much of this material, especially in Chapters 2 and 3, should be review material for most students. Chapters 5 and 6 will most likely each require more than a week to cover, as they form the nuts and bolts of MCMC methods and evaluation. Subsequent chapters should each take 1-2 weeks of class time. The models themselves should be familiar, but the estimation of them via MCMC methods will not be and may be difficult for students without some programming and applied data analysis experience. The programming language used throughout the book is R, a freely available and common package used in applied statistics, but I introduce the program WinBugs in the chapter on hierarchical modeling. Overall, R and WinBugs are syntactically similar, and so the introduction of WinBugs is not problematic. From my perspective, the main benefit of WinBugs is that some derivations of conditional distributions that would need to be done in order to write an R program are handled automatically by WinBugs. This feature is especially useful in hierarchical models. All programs used in this book, as well as most data, and hints and/or solutions to the exercises can be found on my Princeton University website at: www.princeton.edu/∼slynch.

Acknowledgements

I have a number of people to thank for their help during the writing of this book. First, I want to thank German Rodriguez and Bruce Western (both at Princeton) for sharing their advice, guidance, and statistical knowledge with me as I worked through several sections of the book. Second, I thank my friend and colleague J. Scott Brown for reading through virtually all chapters and providing much-needed feedback over the course of the last several years. Along these same lines, I thank Chris Wildeman and Steven Shafer for reading through a number of chapters and suggesting ways to improve examples and the general presentation of material. Third, I thank my statistics thesis advisor, Valen Johnson, and my mentor and friend, Ken Bollen, for all that they have taught me about statistics. (They cannot be held responsible for the fact that I may not have learned well, however). For their constant help and tolerance, I thank Wayne Appleton and Bob Jackson, the senior computer folks at Princeton University and Duke University, without whose support this book could not have been possible. For their general support and friendship over a period including, but not limited to, the writing of this book, I thank Linda George, Angie O'Rand, Phil Morgan, Tom Espenshade, Debby Gold, Mark Hayward, Eileen Crimmins, Ken Land, Dan Beirute, Tom Rice, and John Moore. I also thank my son, Tyler, and my wife, Barbara, for listening to me ramble incessantly about statistics and acting as a sounding board during the writing of the book. Certainly not least, I thank Bill McCabe for helping to identify an egregious error on page 364. Finally, I want to thank my editor at Springer, John Kimmel, for his patience and advice, and I acknowledge support from NICHD grant R03HD050374-01 for much of the work in Chapter 10 on multivariate models.

Despite having all of these sources of guidance and support, all the errors in the book remain my own.

Princeton University

Scott M. Lynch

April 2007

Contents

List of Figures

List of Tables

1

Introduction

The fundamental goal of statistics is to summarize large amounts of data with a few numbers that provide us with some sort of insight into the process that generated the data we observed. For example, if we were interested in learning about the income of individuals in American society, and we asked 1,000 individuals "What is your income?," we would probably not be interested in reporting the income of all 1,000 persons. Instead, we would more likely be interested in a few numbers that summarized this information—like the mean, median, and variance of income in the sample—and we would want to be able to use these sample summaries to say something about income in the population. In a nutshell, "statistics" is the process of constructing these sample summaries and using them to infer something about the population, and it is the inverse of probabilistic reasoning. Whereas determining probabilities or frequencies of events—like particular incomes—is a *deductive* process of computing probabilities given certain parameters of probability distributions (like the mean and variance of a normal distribution), statistical reasoning is an *inductive* process of "guessing" best choices for parameters, given the data that have been observed, and making some statement about how close our "guess" is to the real population parameters of interest. Bayesian statistics and classical statistics involving maximum likelihood estimation constitute two different approaches to obtaining "guesses" for parameters and for making inferences about them. This book provides a detailed introduction to the Bayesian approach to statistics and compares and contrasts it with the classical approach under a variety of statistical models commonly used in social science research.

Regardless of the approach one takes to statistics, the process of statistics involves (1) formulating a research question, (2) collecting data, (3) developing a probability model for the data, (4) estimating the model, and (5) summarizing the results in an appropriate fashion in order to answer the research question—a process often called "statistical inference." This book generally assumes that a research question has been formulated and that a random sample of data has already been obtained. Therefore, this book focuses on model development, estimation, and summarization/inference. Under a classical approach to statistics, model estimation is often performed using canned procedures within statistical software packages like SAS®, STATA®, and SPSS®. Under the Bayesian approach, on the other hand, model estimation is often performed using software/programs that the researcher has developed using more general programming languages like R, C, or C++. Therefore, a substantial portion of this book is devoted to explaining the mechanics of model estimation in a Bayesian context. Although I often use the term "estimation" throughout the book, the modern Bayesian approach to statistics typically involves simulation of model parameters from their "posterior distributions," and so "model estimation" is actually a misnomer.

In brief, the modern Bayesian approach to model development, estimation, and inference involves the following steps:

1. Specification of a "likelihood function" (or "sampling density") for the data, given the model parameters.
2. Specification of a "prior distribution" for the model parameters.
3. Derivation of the "posterior distribution" for the model parameters, given the likelihood function and prior distribution.
4. Simulation of parameters to obtain a sample from the "posterior distribution" of the parameters.
5. Summarization of these parameter samples using basic descriptive statistical calculations.

Although this process and its associated terminology may seem foreign at the moment, the goal of this book is to thoroughly describe and illustrate these steps. The first step—as well as the associated parameter estimation method of maximum likelihood—is perhaps well understood by most quantitative researchers in the social sciences. The subsequent steps, on the other hand, are not, especially Step 4. Yet advances in Step 4 have led to the recent explosion in the use of Bayesian methods. Specifically, the development of Markov chain Monte Carlo (MCMC) sampling methods, coupled with exponential growth in computing capabilities, has made the use of Bayesian statistics more feasible because of their relative simplicity compared with traditional numerical methods. When approximation methods of estimation were more common, such methods generally relied on normality assumptions and asymptotic arguments for which Bayesians often criticize classical statistics. With the advent of MCMC sampling methods, however, more complicated and realistic applications can be undertaken, and there is no inherent reliance on asymptotic arguments and assumptions. This has allowed the benefits of taking a Bayesian approach over a classical approach to be realized.

1.1 Outline

For this book, I assume only a familiarity with (1) classical social science statistics and (2) matrix algebra and basic calculus. For those without such a background, or for whom basic concepts from these subjects are not fresh in memory, there are two appendices at the end of the book. Appendix A covers the basic ideas of calculus and matrix algebra needed to understand the concepts of, and notation for, mathematical statistics. Appendix B briefly reviews the Central Limit Theorem and its importance for classical hypothesis testing using a simulation study.

The first several chapters of this book lay a foundation for understanding the Bayesian paradigm of statistics and some basic modern methods of estimating Bayesian models. Chapter 2 provides a review of (or introduction to) probability theory and probability distributions (see DeGroot 1986, for an excellent background in probability theory; see Billingsley 1995 and Chung and AitSahlia 2003, for more advanced discussion, including coverage of Measure theory). Within this chapter, I develop several simple probability distributions that are used in subsequent chapters as examples before jumping into more complex real-world models. I also discuss a number of real univariate and multivariate distributions that are commonly used in social science research.

Chapter 2 also reviews the classical approach to statistical inference from the development of a likelihood function through the steps of estimating the parameters involved in it. Classical statistics is actually a combination of at least two different historical strains in statistics: one involving Fisherian maximum likelihood estimation and the other involving Fisherian and Neyman and Pearsonian hypothesis testing and confidence interval construction (DeGroot 1986; Edwards 1992; see Hubbard and Bayarri 2003 for a discussion of the confusion regarding the two approaches). The approach commonly followed today is a hybrid of these traditions, and I lump them both under the term "classical statistics." This chapter spells out the usual approach to deriving parameter estimates and conducting hypothesis tests under this paradigm.

Chapter 3 develops Bayes' Theorem and discusses the Bayesian paradigm of statistics in depth. Specifically, I spend considerable time discussing the concept of prior distributions, the classical statistical critique of their use, and the Bayesian responses. I begin the chapter with examples that use a point-estimate approach to applying Bayes' Theorem. Next, I turn to more realistic examples involving probability distributions rather than points estimates. For these examples, I use real distributions (binomial, poisson, and normal for sampling distributions and beta, gamma, and inverse gamma for prior distributions). Finally, in this chapter, I discuss several additional probability distributions that are not commonly used in social science research but are commonly used as prior distributions by Bayesians.

Chapter 4 introduces the rationale for MCMC methods, namely that sampling quantities from distributions can help us produce summaries of them that allow us to answer our research questions. The chapter then describes some basic methods of sampling from arbitrary distributions and then develops the Gibbs sampler as a fundamental method for sampling from high-dimensional distributions that are common in social science research.

Chapter 5 introduces an alternative MCMC sampling method that can be used when Gibbs sampling cannot be easily employed: the Metropolis-Hastings algorithm. In both Chapters 4 and 5, I apply these sampling methods to distributions and problems that were used in Chapters 2 and 3 in order to exemplify the complete process of performing a Bayesian analysis *up to, but not including,* assessing MCMC algorithm performance and evaluating model fit.

Chapter 6 completes the exemplification of a Bayesian analysis by showing (1) how to monitor and assess MCMC algorithm performance and (2) how to evaluate model fit and compare models. The first part of the chapter is almost entirely devoted to technical issues concerning MCMC implementation. A researcher must know that his/her estimation method is performing acceptably, and s/he must know how to use the output to produce appropriate estimates. These issues are generally nonissues for most classical statistical analyses, because generic software exists for most applications. However, they are important issues for Bayesian analyses, which typically involve software that is developed by the researcher him/herself. A benefit to this additional step in the process of analysis—evaluating algorithm performance—is that it requires a much more intimate relationship with the data and model assumptions than a classical analysis, which may have the potential to lull researchers into a false sense of security about the validity of parameter estimates and model assumptions.

The second part of the chapter is largely substantive. All researchers, classical or Bayesian, need to determine whether their models fit the data at hand and whether one model is better than another. I attempt to demonstrate that the Bayesian paradigm offers considerably more information and flexibility than a classical approach in making these determinations. Although I cannot and do not cover all the possibilities, in this part of the chapter, I introduce a number of approaches to consider.

The focus of the remaining chapters (7-10) is substantive and applied. These chapters are geared to developing and demonstrating MCMC algorithms for specific models that are common in social science research. Chapter 7 shows a Bayesian approach to the linear regression model. Chapter 8 shows a Bayesian approach to generalized linear models, specifically the dichotomous and ordinal probit models. Chapter 9 introduces a Bayesian approach to hierarchical models. Finally, Chapter 10 introduces a Bayesian approach to multivariate models. The algorithms developed in these chapters, although fairly generic, should not be considered endpoints for use by researchers. Instead, they should be considered as starting points for the development of algorithms tailored to user-specific problems.

In contrast to the use of sometimes contrived examples in the first six chapters, almost all examples in the latter chapters concern real probability distributions, real research questions, and real data. To that end, some additional beneficial aspects of Bayesian analysis are introduced, including the ability to obtain posterior distributions for parameters that are not directly estimated as part of a model, and the ease with which missing data can be handled.

1.2 A note on programming

Throughout the text, I present R programs for virtually all MCMC algorithms in order to demystify the linkage between model development and estimation. R is a freely available, downloadable programming package and is extremely well suited to Bayesian analyses (www.r-project.org). However, R is only one possible programming language in which MCMC algorithms can be written. Another package I use in the chapter on hierarchical modeling is WinBugs. WinBugs is a freely available, downloadable software package that performs Gibbs sampling with relative ease (www.mrc-bsu.cam.ac.uk/bugs). I strongly suggest learning how to use WinBugs if you expect to routinely conduct Bayesian analyses. The syntax of WinBugs is very similar to R, and so the learning curve is not steep once R is familiar. The key advantage to WinBugs over R is that WinBugs derives conditional distributions for Gibbs sampling for you; the user simply has to specify the model. In R, on the other hand, the conditional distributions must be derived mathematically by the user and then programmed. The key advantage of R over WinBugs, however, is that R—as a generic programming language—affords the user greater flexibility in reading data from files, modeling data, and writing output to files. For learning how to program in R, I recommend downloading the various documentation available when you download the software. I also recommend Venables and Ripley's books for S and S-Plus® programming (1999, 2000). The S and S-Plus languages are virtually identical to R, but they are not freely available.

I even more strongly recommend learning a generic programming language like C or C++. Although I show R programs throughout the text, I have used UNIX-based C extensively in my own work, because programs tend to run *much* faster in UNIX-based C than in any other language. First, UNIX systems are generally faster than other systems. Second, C++ is the language in which many software packages are written. Thus, writing a program in a software package's language when that language itself rests on a foundation in C/C++ makes any algorithm in that language inherently slower than it would be if it were written directly in C/C++.

C and C++ are not difficult languages to learn. In fact, if you can program in R, you can program in C, because the syntax for many commands is close to identical. Furthermore, if you can program in SAS or STATA, you can learn C very easily. The key differences between database programming languages like SAS and generic programming languages like C are in terms