Journal of Statistical Software

April 2009, Volume 30, Book Review 3. http://www.jstatsoft.org/

Reviewer: Joseph M. Hilbe
Jet Propulsion Laboratory, California Institute of Technology

---

Data Analysis Using Regression and Multilevel/Hierarchical Models

Andrew Gelman and Jennifer Hill
Cambridge University Press, Cambridge, UK, 2007.
ISBN 978-0-521-68689-1. 625 pp. USD 69.95 (P).
http://www.stat.columbia.edu/~gelman/arm/

Introductory comments

Gelman and Hill's text is one of the most outstanding statistical publications of which I am aware. I am pleased that I had the opportunity to read it. The book is a member of Cambridge University Press's *Analytical Methods for Social Research* series.

The text, which is how I shall henceforth refer to Gelman and Hill's text, is general work on linear models, with an emphasis on multilevel or hiearchical models. It uses R throughout the book for examples, and is an excellent teaching guide for learning how to employ R for statistical modeling.

This book is unlike many other texts on multilevel and mixed models. It guides the reader from basic linear modeling to complex hieracrchical models, including Bayesian approaches. The text discusses nearly every aspect of the modeling process, from basic data entry concerns and hypothesis testing, to the use of simulation, model checking techniques, and methods of handling missing data. It is a rather complete text, bridging frequency-based statistical modeling with Bayesian methodology.

The text is not a book to read casually. However, the level of discussion is not difficult, and assumes no more than basic calculus and multiple regression as requisite backgrounds. On the other hand, there is so much material, and insightful comments about the subjects being discussed, that the reader must pay close attention to the text in order to abstract all of the information presented.

I do have reservations about the discussion related to certain count models, and I will address those later in this review. However, other than these relatively minor points, the book is accurate and has a minimum of errata.

I should mention here that Appendix C, on software, presents a comparison of code for major software applications. Estimation code for six multilevel models of increasing complexity are shown for R, Stata, SAS, SPSS, and **AD Model Builder**. Unfortunately, the book was

prepared prior to the release of Stata 10 and its new random coefficient logistic and Poisson mixed models commands, `xtmelogit` and `xtmepoisson`. Instead, examples for such models are given using a user-authored command, `gllamm`. The code comparisons are, however, quite helpful for those who do not use R as their primary statistical language.

In order to do justice to the scope of the text, I shall divide the review into five sections, reflecting the five sections into which the book itself is partitioned. There are 25 chapters in the book, and an additional three appendices. Chapters are grouped into sections, which are themselves divided into parts. The five sections I refer to are labeled Parts 1A, 1B, 2A, 2B, and 3. There is a good logic to this division, which will soon become apparent.

Each chaper concludes with a bibliography related to the discussion, and exercise questions. There appear to be an average of approximately ten questions per chapter; however, each of these questions is at times separated into 4 or more ancillary questions. There are, therefore, many questions which instructors can choose from for examination, as well as homework, purposes.

Preliminary two chapters

Two initial chapters start the text by outlining the basic nature of multilevel modeling, and providing the reader with an overview of probability distributions, and the meaning of statistical inference, confidence intervals, and hypothesis testing. A real example is presented which demonstrates the major concepts discussed in these two preliminary chapters. The example was given to the first author by representatives of a large residential organization – with 15,372 households – who believed that the votes taken for election to the Board of Directors appeared to be doctored. They saw that the proportion of votes for each of the six candidates was nearly the same for each interval in voting tallies. That is, the first 600 votes had voting percentages that were the same as those for the next 1244 votes, and the next 1000, the next 1000, and final 1109 votes, for 5,553 votes in total. They believed that there was too little variation in voting from period to period, concluding that the votes were rigged. The authors show how they dealt with the problem, and the reasons why the voting could not be demonstrated to be rigged.

The text has a large number of similar interesting examples which draw the reader into the discussion. As the book unfolds, examples reflect the discussion of the current chapter, calling on previously learned statistical methods.

Part 1A: Single-level regression

Part A is over one hundred pages in length, and is devoted to reviewing basic linear models. The authors review least-squares and maximum-likelihood methods of estimation, and overview model fitting and evaluation. Chapters 3 and 4 emphasize Gaussian models, or simple linear regression. Chapters 5 and 6 deal respectively with logistic regression and generalized linear models. Generalized linear models, or GLM, is a key chapter since multilevel models are based on the concepts developed in generalized linear models. That is, GLM rests as a foundation to multilevel models.

Those with a solid foundation in GLM may skip this part without losing much of what will be expected to be understood in subsequent chapters.

Part 1B: Working with regression inferences

There are four chapters in this section, beginning with a discussion of the simulation of probability models and statistical inferences. Bayesian concepts are introduced at this point. It should be mentioned that Bayesian modeling is not an add-on to frequentist-based linear and multilevel modeling; rather it is integrated into the discussion from the outset.

The remainder of Part 1B relates to using simulation as a check for both statistical procedures and model fit, and a discussion on causal inference. Causal inference is explained first in terms of a treatment variable, and then expanded to more complex models, including imbalanced models, models with instrumental variables, matched models, and so forth. There is also a nice discussion, not often found in texts on this subject, of the assumptions and goals of observational studies in distinction to experimental studies.

The goal of Part 1 is to give the reader a solid background of statistical methodology before engaging in multilevel modeling. I believe the authors do a comendable job in that respect.

One point where I believe the author's presentation can lead to confusion relates to how they dicuss the overdispersed Poisson and negative binomial models. They claim on page 115 that "overdispersed Poisson" strictly speaking means any count model for which the variance is w times the mean, reducing to the Poisson when w equals one. They then assert that the negative binomial is a model of this type, but that it is not usually expressed in terms of the mean and overdispersion, but rather in terms of parameters a and b where the mean of the distribution is a/b and the overdispersion is $1 + 1/b$. This is simply not the case. Nearly all parameterizations of the negative binomial, referring here to the traditional NB2 model, are understood in terms of the mean and heterogeneity or overdispersion parameter α. The variance function is typically defined as $\mu + \alpha \cdot \mu^2$ and mean μ. The negative binomial can be considered as a Poisson-gamma mixture model, or as a logged linked negative binomial family member of generalized linear models. It is not a Poisson model. If anything, one may consider the Poisson as a negative binomial model when the value of α is zero. Poisson models that are parameterized such that their variance is multiplied by a constant are traditonally referred to as quasi-likelihood Poisson models, and are not generally used to model Poisson overdispersion.

Part 2A: Multilevel regression

Parts 2A and 2B are the meat of the text. Part 2A details the structure and scope of multilevel models. The five chapters address, respectively, multilevel structures, multilevel linear models, multilevel models with varying intercepts and slopes – i.e., random intercept and random coefficient or slope models, multilevel logistic models, and multilevel generalized linear models. The only discussion related to multilevel GLMs are ordered categorical slope models, overdispersed Poisson and negative binomial models. The only time continuous response models are discussed is when they relate to the negative binomial, e.g., the negative binomial as a Poisson-gamma mixture distribution.

Part 2B: Fitting multilevel models

Gelman and Hill discuss the estimation of multilevel models using both R software and BUGS. They recommend from the outset that when faced with a modeling situation, one should start

with a basic multiple regression using `lm` or in the case of binary and binomial responses or counts, using `glm`. If intercepts and slopes are to vary, then the modeling is advanced to linear mixed models, or multilevel models, using `lmre`. If we need to understand the uncertainty about the coefficients, the predictions, and other model statistics, they then recommend employing Bayesian methods using BUGS software. BUGS is used to calculate simulations that represent the inferential uncertainty of the model parameters. Much of this part is devoted to using BUGS for this type of modeling.

It should be mentioned that the authors also recommend that statisticians may need to do some programing with R when models become complex. Issues such as speed of estimation, and accuracy of estimates is also disussed at length in this section of the text.

When one has completed Part 2B, they will be able to use BUGS and R together to engage rather sophisticated data situations.

Part 3: From data collection to model understanding to model checking

This final part may be considered as ancillary discussion, but nevertheless essential material to the actual modeling process. Issues related to sample size and power, summarizing model conclusions, dealing with missing data and employing imputation methods, and model checking are all the subject of Part 3. A full 110 pages are taken up with this material, so the authors do not treat the subject lightly.

Of interest in Chapter 24 is a brief discussion of the deviance information criterion, or DIC, which is the multilevel equivalent of the AIC statistic which is a common measure of comparative fit for single level models. The authors correctly point out that the statistic, which is a measure of out-of-sample predictive error, is unstable to estimate for more complex multilevel models. They provisionally use it to compare the models discussed, but do so with caveats. They conclude that more work needs to be done to derive a statistic which adequately compares the fit of competing multilevel models.

Appendices

The text has three appendices, the first being a section on tips to improve the readers modeling efforts. The second appendix deals with the construction and presentation of statistical graphics, and the third which we have previously alluded to relates to current software used for multilevel modeling.

Concluding comments

Gelman and Hill have produced an outstanding text. It is not a book to be read casually, but is rather a text to study, to seriously consider, and to work with as one develops an understanding of the subjects being discusssed. For those who are already familiar with multilevel modeling, there are many useful hints and comments interspersed throughout the text, and of considerable value are the examples of R code that can be used to simulate and model data as required. I recommend it for use in any graduate level course on multilevel models, in particular ones that chose to use R for the modeling software. For those who are professional statisticians and researchers, and who either do or intend to engage in multilevel

modeling, I recommend Gelman and Hill's text as an useful reference text.

Reviewer:

Joseph Hilbe
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, California 91109, United States of America
E-mail: jhilbe@aol.com
URL: http://en.wikipedia.org/wiki/Joseph_Hilbe

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 30, Book Review 3 *Published:* 2009-04-27
April 2009