Journal of Statistical Software

July 2016, Volume 71, Book Review 3. *doi: 10.18637/jss.v071.b03*

Reviewer: Ulrike Grömping
Beuth University of Applied Sciences Berlin

Practical Guide to Logistic Regression

Joseph M. Hilbe
Chapman & Hall/CRC, Boca Raton, 2016.
ISBN 978-1-4987-0957-6. 174 pp. USD 49.95 (P).
https://www.crcpress.com/9781498709576

"Practical Guide to Logistic Regression" introduces the analysis of binary response data for working analysts and researchers who have an understanding of basic statistical principles and linear regression analysis – this is its intention as declared in the preface. The book is structured as follows: a brief section (Chapter 1, 12 pages) on statistical models (focusing on basic ideas of logistic regression, the Bernoulli distribution, and ML estimation) is followed by a detailed introduction (Chapter 2, 36 pages) to the logistic model with a single predictor, which starts with one binary predictor, presents all formal tools in this context, and moves on to a multi-category and a continuous predictor. Chapter 3 (22 pages) discusses logistic regression with multiple predictors, covering interpretation of coefficients, model fit statistics, information criteria (e.g., AIC, BIC), adjustment of standard errors (by scaling, through the sandwich estimator or via the bootstrap), and common expressions used in risk modeling, like "risk factor", "confounder" or "effect modifier". Chapter 4 ("Testing and Fitting a Logistic Model", 36 pages) covers goodness-of-fit tests, diagnostic plots, numeric instabilities resulting from (almost) perfect separation, exact logistic regression, and – perhaps somewhat out of place – analysis of data given as a table of counts. The next chapter, "Grouped Logistic Regression" (Chapter 5, 20 pages), would have been a more natural place for the analysis of data given as frequency tables; it starts with an introduction to the binomial distribution with more than one trial and moves on to overdispersion and beta binomial regression (for which the book's author is a renowned expert). The last chapter covers "Bayesian Logistic Regression" (Chapter 6, 24 pages). A reference section and an index complete the book.

The book presents many worked examples, and the choice of interesting data sets all of which are available to the reader is one of its greatest assets. Data availability makes it easy for readers to reproduce the examples from the book, and example code is available for R, SAS and Stata: R code is incorporated into the book chapters, and the end of each chapter gives SAS and Stata code. The data sets can be downloaded from the book author's website in all three formats, and it is very convenient for R users that the package **LOGIT** (on CRAN) contains all data sets and several functions that the book author provides. The SAS code for the examples can also be found on the book author's website, while I have not been able to

locate files with R code or Stata code for the book examples (in spite of a statement to that effect in the book's preface). It would be very helpful if author or publisher would also place R and Stata code online, ideally in an ASCII format that would make the book's reiterated warnings about copy-pasting code from PDF or Word documents superfluous. Regarding typesetting, it would also be helpful to make sure that future printings show code in proper ASCII format rather than allowing, e.g., quotes to be shown in fancy formats not suitable for code. Nevertheless, even as is, the example code in the book will be quite useful to the readers.

In the following, before sharing my personal conclusion about the role of this book in my teaching, I will discuss my perceptions on the book's general approach, coverage of the interpretation of coefficients and effect plots, quality of example code, examples of inaccurate or controversial statements, and the introduction to Bayesian logistic regression.

I quite like the book's slow approach of first presenting the most important aspects of logistic regression for a single predictor model – it is suitable for creating a good intuitive understanding of the inner workings of a logistic model. The exposition is not always as clear as it could have been with more careful choices of wording, but the slow initial speed is definitely useful. In the part on a single continuous predictor, some of the later subsections remain a bit unclear: for example, GAM modeling is proposed for diagnosing the functional form of a predictor, but the exposition could be more explicit about its role for this purpose.

Interpretation of coefficients is discussed in various places; the initial general discussion of odds ratios contains some quite sloppy wording, e.g., in the following sentence: "The odds of y given $x = 1$ is $\exp(-1.504077)$ or 0.22222 times greater than the odds of $x = 0$. This is the same as saying that the odds of $x = 0$ is $1/\exp(-1.504077)$ or 4.5 times greater than $x = 1$." Later specific interpretation examples are much clearer, e.g., when interpreting the effect of admission type on the probability of death in hospital in Section 2.4. For continuous predictors, I missed a hint on how to obtain an odds ratio for a specified difference other than one unit, e.g., for an age difference of 10 years. Unfortunately, the important topic of "Risk Factors, Confounders, Effect Modifiers, and Interactions" is treated superficially only; the book explains an effect modifier through an interaction, without ever properly explaining the concept of interactions (the book references Hilbe 2009, for that purpose). In my opinion, a discussion of interaction effects and their interpretation is crucial for an introduction to applied regression modeling. The example data set for the Bayesian chapter would have offered a good opportunity for showing the benefit of incorporating interactions: the inclusion of an interaction effect between having kids and being female on the probability of not working would have improved the model in a very plausible way. It would also have been helpful to demonstrate effect plots in the context of interpreting coefficients, especially in case of interactions and continuous predictors. The book touches upon effect plots once, but in a very dubious way: Figure 4.4 shows the effect of length of hospital stay (`los`) on the probability of death in hospital (`died`), given the admission status (factor `type` with levels 1 = elective, 2 = urgent or 3 = emergency); this plot is based on a very questionable calculation using the three levels (1, 2, 3) of `type` in a numeric regressor with one linear coefficient (instead of correctly keeping admission type as a factor the way it was included in the model `mymod` earlier in that chapter); this is seriously flawed and completely unnecessary. The R code for the creation of the questionable effect plot of Figure 4.4 (in Table 4.3) could be replaced by one line of code for creating a more appropriate effect plot with package **effects** (provided that the factor `type` is formatted as a factor within the data frame and not "on the

fly"): `plot(Effect(c("type", "los"), mymod), multiline = TRUE, type="response")` would do the trick. The SAS and Stata implementations could be fixed accordingly, e.g., using SAS `PROC LOGISTIC` or `PROC GENMOD` with the `EFFECTPLOT` statement.

The book is not meant to be a manual for R, Stata or SAS, but the code is meant to support the understanding of statistical concepts; nevertheless, the R code interspersed in between the text is given a lot of attention, and rightly so, because it would otherwise not be useful. The code works, but is not always ideal; for example, since R does not have a built-in function for residual standard errors, the book provides several lines of R code for obtaining standardized residuals, instead of simply using function `rstandard` for that purpose. Or it is stated that standard errors for the coefficients can only be obtained through taking square roots of the diagonal elements of `vcov` output, while they are also available in the `coefficients` slot of a `summary.glm` object. Or points are added to a figure (Figure 2.2.) using the command `lines` with option `type = "p"` instead of using the command `points`. Or readers are referred to the `modelfit` function of Hilbe's package **COUNT** because it is alleged that R does not have a function for the Schwarz BIC criterion, although there is a function `BIC` in the core package **stats**. The creation of effect plots discussed in the previous paragraph is another case for which useful established capabilities of the software are ignored. These examples demonstrate that the book's focus is not to be a guide for efficient use of the software products discussed. Nevertheless, the code snippets throughout the text are in most cases quite helpful for coming to grips with the statistical content.

I have been educated strictly in the frequentist world. In this thinking, Bayesian methods are either too subjective or – with a flat prior – a complicated way to do the same thing one could have done through maximum likelihood. While reading the other chapters of Hilbe's book as a critical expert, I have read the Bayesian chapter as a learner. As such, I was first of all grateful for pointers how to do a Bayesian logistic regression in R. For this, the book shows two ways, one completely within R and one using the external software **JAGS**, accessed through package **R2jags**. Getting the code of the first example to run was complicated by the fact that the book uses difficult-to-access custom functions for printing the output; for my taste, it would have been sufficient to use the built-in R function `print` with a digits option; this prints a little more information, but not unreasonably so. Besides this small technical point, the example data for this chapter is a bit unfortunate, since even the standard logistic regression without overdispersion is suitable for these data, so that the Bayesian model cannot demonstrate its benefits to a skeptic like me. On the other hand, trust can be built by seeing that the two ways of running the Bayesian logistic regression yield similar results which are also very close to the standard results.

While the book provides many useful thoughts and examples on fitting good logistic models to data, there are also various misleading statements. Right in the beginning, the book claims that it is assumed that "The predictors are uncorrelated with each other." In many fields of application, a necessity for this assumption would render logistic regression unusable. Some qualifying words would have been very desirable. Section 2.3.3 on p-values has some quite controversial wording, e.g., "the smaller the p-value, the more likely $\beta \neq 0$". Furthermore, I disagree with "A coefficient of 0 indicates no effect, and contributes nothing to understanding the response variable of interest"; indeed, the zero coefficient *does* contribute to the understanding that *the regressor variable in question*, given the other variables in the model, does not contribute anything to understanding the response variable of interest. There is also an occasional wrong claim, like "Odds ratio for beta binomial are inflated compared to the

grouped logit, but the p-values are closely the same" in Chapter 5: while the first part of this sentence correctly reflects the printed odds ratios, the versions of p-values obtained from the binomial model are quite different to each other and none of them very close to those of the beta binomial.

Joe Hilbe is a very experienced modeler, statistician and book author. I therefore had high expectations for this book, hoping that I could use it for teaching applied logistic regression to undergraduate or graduate students with training in calculus, linear algebra and the linear regression model, but otherwise relatively little theoretical background. Unfortunately, the book is unsuitable for that purpose; it can serve as supporting material in a course based on more conventionally structured material on logistic regression, like, e.g., Fox and Weisberg (2011) or Hosmer and Lemeshow (2013). The reason for rejecting the book as the base material is two-fold: for the mathematical audience and anyone who wants to work with the formulae, the book has too many formal issues: formulae are sometimes quite sloppy (using or not using indices in sums or expressions appears almost haphazard in places), the letter y has too many roles, and there is no proper notational distinction between a parameter and its estimate (the use of b and β remains a mystery to me) or between the response, its prediction or its expectation; while I appreciate that too much notation can put off readers with little mathematical background, the level of ignorance regarding formalism in this book is too high for using it in teaching mathematically-minded students. In addition, the book not only benefits but also suffers from its author's vast experience: while there are various useful passing comments, such comments are not always deferred to the adequate place for the learner. Nevertheless, the book has its positive aspects: in particular, the examples and the slow introduction into the topic will be quite useful to some readers. Thus, you are invited to make up your own mind. Make sure to get the second printing with errata corrected – the first printing suffered from many errata, including serious ones in formulae.

References

Fox J, Weisberg S (2011). *An R Companion for Applied Regression.* 2nd edition. Sage Publications, Thousand Oaks.

Hilbe J (2009). *Logistic Regression Models.* Chapman & Hall/CRC, Boca Raton.

Hosmer DW, Lemeshow S (2013). *Applied Logistic Regression.* John Wiley & Sons, New York. doi:10.1002/9781118548387.

Reviewer:

Ulrike Grömping
Beuth University of Applied Sciences Berlin

Department II
D-13353 Berlin, Germany
E-mail: groemping@bht-berlin.de
URL: http://prof.beuth-hochschule.de/groemping/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2016, Volume 71, Book Review 3 *Published:* 2016-07-26
doi:10.18637/jss.v071.b03