Roger Newson (Department of Public Health Sciences, King's College, London, UK)

roger.newson@kcl.ac.uk

James Hardin and Joseph Hilbe. 2001. *Generalized Linear Models and Extensions.* College Station, TX: Stata Press. 245pp.

The authors, both well-known for their contributions to Stata's modelling software, have written a book which continues the good work done by McCullagh and Nelder (1989), and contains a lot of new developments. However, the two books are complementary, each containing things the other lacks. In an ideal world, all statisticians would have access to both books.

A generalized linear model (GLM) can be defined as any model where the conditional mean of the outcome variable Y is transformable to a linear combination of X-variables (using a link function), and where the variance of the outcome variable is proportional to some function of the mean (known as the variance function). A set of generalized linear models with the same variance function is called a family of models, and usually corresponds to a family of distributions, such as the Gaussian, Poisson or binomial. By choice of link and variance functions (and/or transformation of the outcome variable), the user can estimate parameters which may be proportions, rates, probabilities, odds, probits or arithmetic, geometric, harmonic or algebraic* means, and their differences or ratios. These differences or ratios can be either contrasts between groups defined by a categorical predictor variable, or changes in the Y-variable in response to a unit change in a quantitative predictor variable. Arguably, therefore, GLMs include most of the methods that most applied statisticians use, most of the time. GLM theory was developed by John Nelder and his colleagues at Rothamsted Experimental Station in the 1970s and extended by others elsewhere, including the authors of this book. The theory provides a single unified class of mathematical and computational methods which can be used to estimate parameters for all these models, and to calculate confidence limits. This book is a very useful definitive handbook for anybody who wants to understand these methods, and to choose one to apply to a particular data analysis. The book also includes detailed documentation of the `glm` command in Stata, explaining the rationale behind the options and formulae listed in the manual. However, the authors also explain in detail the fitting of GLMs using commands other than `glm`.

After an introductory section, the authors devote Section I to the mathematical theory behind GLMs, the possible algorithms used in fitting the parameters, the many methods available for calculating variances, standard errors and confidence limits, and many varieties of residuals and methods for assessing goodness of fit. Sections II, III, IV and V give the reader a panoramic tour of the many specific varieties of GLMs available for continuous, binary, count and multinomial outcomes. The authors present many useful combinations of link and variance function that are not widely known even amongst statisticians, such as using the gamma, inverse Gaussian and binomial variance functions with the identity link (to estimate differences between means) or the log link (to estimate ratios between means). Section VI introduces some further extensions to the already broad class of GLMs. These include quasi-likelihood, generalized additive models, and methods for clustered data, including generalized estimating equations (GEEs). Section VII gives a survey of available Stata software for the various GLMs. Finally, there is a list of Appendices for reference to the various components of commonly-used GLMs (such as link functions, variance functions and likelihood functions). Throughout the book, examples are demonstrated using Stata. The authors have helpfully placed on the Web the data sets and programs used in these examples, and these can be accessed from within Stata by typing `findit hardin hilbe`. The reader can therefore easily repeat the authors' analyses, and carry out similar analyses on other data, or alternative analyses on the same data.

---

* The pth-power algebraic mean of a variable Y is defined as $[E(Y^p)]^{\frac{1}{p}}$, where $E(\cdot)$ denotes expectation. Algebraic means can be estimated using `glm` with power-transformed data, together with their differences (using an inverse power link) or their ratios (using a log link, scaling the estimates and variances post-estimation by the inverse and squared inverse powers respectively, and displaying the result using the `eform` option).

How might this excellent book have been even better? This book is clearly intended mainly as a reference for statisticians, rather than to be used by non-statisticians as a substitute for statisticians. However, applied statisticians have to justify their methods, and to explain their parameters and confidence intervals in words, to non-statisticians. The authors cover parameter interpretation in Chapter 10, Subsection 10.6 in the case of binomial models, but I personally would have stressed this point more often with more classes of models. In particular, if I use Stata to fit models with a log or logit link (or to fit models with an identity link to log-transformed data), then, unlike the authors, I nearly always use the `eform` option (or its equivalent), so that the parameters in the output can be interpreted as arithmetic or geometric means (or rates or proportions or odds) and their ratios. In connection with this subject, the authors (in Subsection 5.5) discuss "GLM log-normal models", meaning GLMs fitted to untransformed data with a Gaussian variance function and a log link, and argue that this method is better than the more traditional practice of fitting a Gaussian model with an identity link to the log-transformed data, because the latter method yields parameters that are not easy to interpret. However, I find that medical colleagues can usually understand geometric means and their ratios, if statisticians explain them. Moreover, in the case of positively-skewed, positive-valued variables such as serum viral loads or triglyceride concentrations, the geometric mean (or even an algebraic mean) may be a better approximation to the median than the arithmetic mean. Also, in Chapter 17, where the authors discuss methods for clustered data, they explain clearly the formulae behind clustered sandwich variances and GEEs, and demonstrate the two methods using a data set with five clusters. It would be even better, however, if they had used the two methods on a data set with many more clusters, and demonstrated the reductions in the width of confidence intervals made possible by using GEEs instead of clustered sandwich variances for the ordinary GLM estimator. This would enable statisticians to explain to their non-statistical colleagues the reason for using a more complicated method, which requires much more computer time.

However, these are minor criticisms. The authors have written a very comprehensive account of a very large subject indeed, which is still very much under construction. I know of no other book where as much equally up-to-the-minute information on GLMs and their extensions can be found in one place. For readers wanting to explore the subject in even greater depth than is presented in this book, the References section of the book gives a wealth of places to start. Readers who might consider buying this book, but who still need to be convinced, can find further details of the contents at the Stata bookstore website under
`http://www.stata.com/bookstore/glmext.html`.

References

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models. 2nd edition.* London: Chapman and Hall.