Preface

Beauty will result from the form and correspondence of the whole, with respect to the several parts, of the parts with regard to each other, and of these again to the whole; that the structure may appear an entire and compleat body, wherein each member agrees with the other, and all necessary to compose what you intend to form.
—The First Book of Andrea Palladio's Architecture, 1570

In the early 1970s, the *generalized linear model* (GLM) class of statistical models was proposed by Nelder and Wedderburn (1972), providing a unified framework for several important regression models. They showed that the linear model, Poisson regression, logistic regression and probit analysis, and others could be treated as special cases of GLMs, and that one algorithm could be used to estimate them all. The unified GLM framework also provides an elegant overriding theoretical structure resulting in inference, diagnostics, software interface, etc. that applies to all of them. Prior to GLMs, these methods were largely treated as unrelated. Since then, GLMs have gained universal acceptance in the statistical world, and as Senn (2004, p. 7) writes, "Nelder and Wedderburn was a paper that changed the statistical landscape for ever and it is simply impossible now to envisage the modelling world without it."

Although GLMs were a great advance, they are largely confined to one-parameter distributions from an exponential family. There are many situations where practical data analysis and regression modelling demands much greater flexibility than this. To this end, this book describes a much larger and more flexible statistical framework for fixed-effects regression modelling that greatly extends GLMs. It comprises about half-a-dozen major classes of statistical models, and having at its heart two classes, called *vector generalized linear models* (VGLMs) and *vector generalized additive models* (VGAMs). (Other classes are listed below.) Each class is related to each other in a natural way within this framework, e.g., VGAMs are a smooth or data-driven version of VGLMs.

The purpose of this book is to introduce the framework and each major subclass of models. VGLMs might be thought of loosely as multivariate GLMs, which are not confined to the exponential family. VGAMs replace the linear functions of VGLMs by smooths. The remaining classes can be thought of as extensions of VGLMs and VGAMs. As a software implementation of our approach, the R package VGAM (and its companion VGAMdata) are used. Following the grand tradition of GLMs, we demonstrate its usefulness for data analysis over a wide range of problems and application areas. It is hoped that, in a way similar to what was pioneered by GLMs, the overall framework described in this book will provide a natural vehicle for thinking about regression modelling, and performing much more of applied statistics as a coherent whole.

What advantages does the VGLM/VGAM framework confer? It provides the same benefits that GLMs gave, but on a much larger scale. For example, the theory and methodology of this book unifies areas such as univariate distributions, categorical data analysis, aspects of quantile regression, and extremes. And the advantages which generalized additive models add to GLM-type analyses in terms of smoothing, VGAMs do for VGLMs. It has long been my belief that what should be 'ordinary' and 'routine' regression modelling in applied statistics has been hampered by a lack of common framework. For VGLMs/VGAMs, the user has the freedom to easily vary model elements within a large flexible framework, e.g., currently VGAM implements over 150 family functions.

The underlying conception is to treat almost all distributions and classical models as generalized regression models. By 'classical', we mean models and distributions which are amenable to estimation by first and second derivative methods, particularly Fisher scoring. For years there has been a need to broaden the scope of GLMs and GAMs, and to fortify it with some necessary infrastructure to make them more fully operable. The end is that the framework instils structure in many classical regression models.

Audience

The book was written with three types of readers in mind. The first are existing users of the VGAM R package needing a primary source for the methodology, and having sufficient examples and details to be useful. The second are people interested in a new and very general modelling framework. The third are students/teachers in courses on general regression following a basic regression course on GLMs. For these, some notes for instructors are given below.

Some assumptions have been made about the background of the reader. Firstly is a basic working knowledge of R. There are now many books that provide this, e.g., Venables and Ripley (2002), Dalgaard (2008), Maindonald and Braun (2010), Cohen and Cohen (2008), Zuur et al. (2009), de Vries and Meys (2012). Further references are listed at the end of Chap. 8. Second is a mid-undergraduate level knowledge of statistical theory and practice. Thirdly, some chapters assume a basic familiarity with linear algebra and calculus up to mid-undergraduate level.

Some of the applied chapters (in Part II) would benefit from some prior exposure to those subject areas, e.g., econometrics, plant ecology, etc. because only a minimal attempt can be made to establish the background, motivation, and notation there.

How This Book is Organized

There are two parts. Part I describes the general theory and computational details behind each major class of models. These classes are displayed in the flowchart of Fig. 1.2. Readers familiar with LMs, GLMs, and GAMs might jump directly into the VGLM and VGAM chapters (Chap. 2 is a summary of these topics), otherwise they could go through sequentially from the beginning. However, Chap. 1 (especially Sect. 1.3) is crucial, because it sets the scene and gives a brief overview of the entire framework. The other major classes of models are various generalizations of reduced-rank methods. The main motivation for these is dimension reduction, and they consequently operate on latent variables. They are:

- RR-VGLMs: these *reduced-rank VGLMs* (Chap. 5) are based on linear combinations of the explanatory variables, and should be useful for many readers.
- QRR-VGLMs: these *quadratic RR-VGLMs* (Chap. 6) being relevant to mainly ecologists could be generally skipped on first reading.
- RR-VGAMs: these allow for *constrained additive ordination* (CAO; Chap. 7), which is a smooth version of RR-VGLMs.

Chapter 8 on the VGAM package should be read by all software users.

Part II explores some major application areas. Practitioners of certain topics in Part II will usually not need to concern themselves with the other parts, because they are mainly separate. However, readers will be able to at least browse through the range of other applications there. The breadth of coverage should demonstrate the versatility of the framework developed in Part I; I feel this is a major attractive feature of the book. Chapter 18 is a specialist chapter addressed more to R programmers wishing to extend the capabilities of the current software, e.g., by writing new VGAM family functions.

Most acronyms are listed on pp. xxiii–xxiv, and the appendices, notation, and glossary are located at the back of the book. R packages are denoted in sans-serif fonts, e.g., stats, splines; and R commands in typewriter font and functions ending in parentheses, e.g., `coef()`. All logarithms are to base e unless otherwise specified.

The scope of this book is broad. To keep things to a manageable size, many topics which I would like to have included have had to be omitted. Rather than citing hundreds of journal articles, I have largely cited books and review papers. Other references and sources of data, etc. can be found in the packages' online help files. I apologize beforehand for many peoples' work that should probably have been cited but are not, due to space limitations or ignorance. A slightly unfortunate consequence of the book's breadth is that some of the notation had to be recycled, however, despite some symbols having multiple meanings, this should not raise excessive confusion because the topics where they are used are far enough apart.

Topics considered too specialized or technical for most readers have been marked with a dagger (†); these may be safely skipped without losing the overall discourse.

Book Website and Software

Resources are available at the book's webpage, starting at

`http://www.stat.auckland.ac.nz/~yee`

Amongst other things, these should include R scripts, complements, errata, and any latest information. Readers wishing to run the software examples will need access to R (see `http://www.R-project.org`), and the VGAM and VGAMdata packages installed along side. The latter comprises some data used here. Version 1.0-0 or later for both packages are needed for compatibility with this book.

Note for Instructors

Most of the contents are pitched at the level of a senior undergraduate or first-year postgraduate in statistics. It might be considered a follow-on course after a first course in GLMs. Students are assumed to have a reasonable working R background already, because only a few important R topics that are crucial are covered (at a cursory level too) in the first chapter. The first few chapters of Fox and Weisberg (2011) would be useful preparation, as well as those listed above.

For a course based on this book, it is recommended that about half of the time be devoted to Part I, so that the overall framework can be seen. The remaining time might be used on a selection of topics drawn from Part II depending on interest and need. Each Part II chapter is at an introductory level, and the references at the end of each chapter can be pursued for a deeper coverage.

Some of the exercises might be suitable for homework or tutorial purposes. Overall, they are a blend of short analyses involving real and simulated data, mathematical statistics problems, and statistical computing (R programming) tasks. While most problems are short, some are less straightforward and require more advanced R programming to obtain an elegant solution. The most demanding or time-consuming problems are marked with a dagger (†) and should be considered optional or for a sizeable project.

The book is primarily focused on estimation. Theoretical rigour that was unnecessary to the main messages has been omitted, and the presentation is chiefly driven by a pragmatic approach. Most of the material is self-contained, and I have tried to strike a balance between theory, computational details and practice for an applied statistician.

Acknowledgements

This is a synthesis of work that started while I was a PhD student at the University of Auckland, and it has spanned the two decades since. Not surprisingly, there are many people to thank, both directly and indirectly. I am indebted to my PhD supervisors Chris Wild and Neil Mitchell who had a very positive effect during the formative years of my academic life. Ross Ihaka and Trevor Hastie pointed me in the right direction soon after in those very early days.

Many graduate students have contributed to this body of work in one way or another over the years, especially James Gray, Kai Huang, Víctor Miranda, and Alvin Sou. Successive heads of departments have helped arrange my teaching to allow for a faster progression of this work. Additionally, I wish to acknowledge the hospitality of many (mostly statistics) departments all across this terrestrial ball, which I have had the good fortune to visit over the years.

Many thanks are extended to Garry Tee for proofreading the book in its entirety. Also to many others who have given valuable feedback on portions thereof; those are Michel Dedual, Stéphane Guindon, Lian Heng, Richard Huggins, Murray Jorgensen, Ioannis Kosmidis, Louise McMillan, Alastair Scott, Alec Stephenson, Jakub Stoklosa, Berwin Turlach, Yong Wang, Alan Welsh, and Chris Wild. All errors, however, are my own. Thanks to John Young for his translation of several quotes from the ancients used at the beginning and ending of several chapters. For help with the typesetting (especially LaTeX and knitr), my thanks to Trevor Hastie, Alan Izenman, Andrew Robinson, and Yihui Xie. Also, the R-core members and the R community at large have been of great direct and indirect help; these include

the countless users who have sent in bug reports and suggestions over the years. Feedback from workshop attendees has been helpful too. I would also like to thank Hannah Bracken at Springer.

Last but not least, thanks to my wife Selina who has been very understanding during the intensive preparations.

Future Directions

This book is a snapshot from a project that is a work in progress, and it will probably never be finished. Large tracts of theory and methodology presently lie undeveloped, and many existing parts require refinement. It is hoped that, over time, bugs will be attended to and further additions made to the software. Consequently, a warning to the reader: low-level internal changes to the software are ongoing, however high-level usage will change less. (Good coding practice such as the use of the generic functions listed in Chap. 8, as well as the suggestions of Sect. 8.5.1, should help minimize such problems). The `NEWS` file in the packages lists successive changes between versions.

Over time, I hope to further develop the theory and continue to add to the functionality of VGAM. Due to the volume of emails received (which exceeds the number I can reply to), please send bug reports only. And my apologies if I am unable to reply at all!

Auckland, New Zealand Thomas W. Yee
May 2015

It is a good thing for an uneducated man to read books of quotations.
—Winston Churchill, *My Early Life*

Springer

http://www.springer.com/978-1-4939-2817-0

Vector Generalized Linear and Additive Models
With an Implementation in R
Yee, Th.W.
2015, XXIV, 589 p. 103 illus., 99 illus. in color.,
Hardcover
ISBN: 978-1-4939-2817-0