1 Introduction

Contents

In updating this text, our primary goal is to convey the practice of analyzing data via generalized linear models to researchers across a broad spectrum of scientific fields. To accomplish this goal, we lay out the framework used for describing various aspects of data and for communicating tools for data analysis. This part of the text contains no examples. Rather, we focus on the lexicon that we will use in later chapters. These later chapters use examples from such fields of interest as biostatistics, economics, and survival analysis.

In developing analysis tools, we illustrate techniques via their genesis in estimation algorithms. We believe that motivating the discussion through the estimation algorithms clarifies the origin and usefulness of the techniques. Instead of detailed theoretical exposition, we refer to texts and papers that present this type of material so that we may focus our detailed presentations on the algorithms and their justification. Our detailed presentations are mostly algebraic; we have minimized matrix notation whenever possible. We also provide a list of the statistics and tests associated with this area along with their formulas and utility.

1.1 Origins and motivation

We wrote this text for researchers who want to understand the theory, scope, and application of generalized linear models. For brevity's sake, we use the acronym GLM to refer to the *generalized linear model*, but we acknowledge that GLM has been used as an acronym for the *general linear model*. The latter usage, of course, refers to the area of statistical modeling that is based solely on the normal or Gaussian probability distribution. GLZ is another term referring to generalized linear models found in some software packages.

We take GLM to be the generalization of the general, for that is precisely what GLMs are; they are the result of extending ordinary least squares (OLS) regression, or the normal model, to a model that is appropriate for a variety of response distributions. We examine exactly how this extension is accomplished. We also aim to provide the reader with a firm understanding of how GLMs are to be evaluated and when their use is appropriate. We even advance a bit beyond the traditional GLM and give the reader a look at how GLMs can be extended to model certain types of data that do not fit exactly within the GLM framework.

Nearly every text that addresses a statistical topic uses one or more statistical computing packages to calculate and display results. We use Stata exclusively, though we do refer to other software packages.

The individual statistical models that make up GLMs are often found as standalone software modules, typically fitted using maximum likelihood methods based on quantities from model-specific derivations. Stata has several such commands for specific GLMs such as `poisson`, `logistic`, and `regress`. Some of these procedures were included in the Stata package from its first version. More models have been addressed through commands written by users employing Stata's programming language that has led to the creation of highly complex statistical models. Most of these user-written commands have since been incorporated into the official Stata package. Stata's `glm` command was itself created as a user program, first in 1993, and then 3 years later as part of Stata 6.0. The `glm` command, which we wrote to augment the first version of this text, was adopted and is now the official Stata `glm` command distributed with the commercial software; the software company maintains and supports the command. Examples in this edition reflect StataCorp's updates to the command.

Stata already offers many of the statistical models that are discussed in this text. Moreover, Stata allows the user to write complex statistical algorithms for themselves. You will see many examples throughout this text, including the `glm` command.

Readers of technical books often need to know about prerequisites, especially how much math and statistics background is required. To gain full advantage from the text and follow its every statement and algorithm, you should have an understanding equal to a two-semester calculus-based course on statistical theory. Without a background in statistical theory, the reader can accept the presentation of the theoretical underpinnings and follow the (mostly) algebraic derivations that do not require more than a mastery of simple derivatives. We assume prior knowledge of multiple regression but no other specialized knowledge is required.

We believe that you can best understand GLMs if their computational basis has been made clear. Hence, we begin our exposition with an explanation of the foundations and computation of GLMs; there are two major methodologies for developing algorithms. We then show how simple changes to the base algorithms lead to different GLM families, links, and even further extensions. In short, we attempt to lay the GLM open to inspection and to make every part of it as clear as possible. In this fashion, the reader

can understand exactly how and why GLM algorithms can be used, as well as altered, to better model a desired dataset.

Perhaps more than any other text in this area, we examine alternatively two major computational GLM algorithms:

1. Iteratively reweighted least squares (and modifications)
2. Newton–Raphson (and modifications)

Interestingly, some of the models we present can be calculated only by using one of the above methods. Iteratively reweighted least squares (IRLS) is the more specialized technique and is applied less often. Yet it is typically the algorithm of choice for quasi-likelihood models such as generalized estimating equations (GEEs). On the other hand, truncated models that do not fit neatly into the exponential family of distributions are modeled using Newton–Raphson methods—and for this, too, we show why. Once again, focusing on the details of calculation should help the reader understand both the scope and the limits of a particular model.

Whenever possible, we present the log likelihood for the model under discussion. In writing the log likelihood, we include offsets so that interested programmers can see how those elements enter estimation. In fact, we attempt to offer programmers the ability to understand and write their own working GLMs, plus many useful extensions. As programmers ourselves, we believe that there is value in such a presentation; we would have much enjoyed having it at our fingertips when we first entered this statistical domain.

1.2 Notational conventions

We use L to denote the likelihood and the script $\mathcal{L}$ to denote the log likelihood. We use X to denote the design matrix of independent (explanatory) variables. When appropriate, we use boldface type $\mathbf{X}$ to emphasize that we are referring to a matrix; a lowercase letter with a subscript $\mathbf{x}_i$ will refer to the ith row from this matrix.

We use Y to denote the dependent (response) variable and refer to the vector $\boldsymbol{\beta}$ as the coefficients of the design matrix. We will use $\widehat{\boldsymbol{\beta}}$ when we wish to discuss or emphasize the fitted coefficients. Throughout the text, we will discuss the role of the (vector) linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$. In generalizing this concept, we will also refer to the augmented (by an offset) version of the linear predictor $\xi = \eta + \text{offset}$.

Finally, we will use the $\mathrm{E}(\cdot)$ notation to refer to the expectation of a random variable and the $\mathrm{V}(\cdot)$ notation to refer to the variance of a random variable. We will describe other notational conventions at the time of their first use.

1.3 Applied or theoretical?

A common question regarding texts concerns their focus. Is the text applied or theoretical? Our text is both. However, we would argue that it is basically applied. We show enough technical details for the theoretician to understand the underlying basis of GLMs. However, we believe that understanding the use and limitations of a GLM includes understanding its estimation algorithm. For some, dealing with formulas and algorithms appears thoroughly theoretical. We believe that it aids understanding the scope and limits of proper application. Perhaps we can call the text a bit of both and not worry about classification. In any case, for those who fear formulas, each formula and algorithm is thoroughly explained and that by book's end the formulas and algorithms will seem simple and meaningful. For completeness, we give the reader references to texts that discuss more advanced topics and theory.

1.4 Road map

Part I of the text deals with the basic foundations of GLM. We detail the various components of GLM, including various family, link, variance, deviance, and log-likelihood functions. We also provide a thorough background and detailed particulars of both the Newton–Raphson and IRLS algorithms. The chapters that follow highlight this discussion, which describes the framework through which the models of interest arise.

We also give the reader an overview of GLM residuals, introducing some that are not widely known, but that nevertheless can be extremely useful for analyzing a given model's worth. Finally, in part I we discuss the general notion of goodness of fit and provide a framework through which you can derive more extensions to GLM.

Part II concerns the continuous family of distributions, including the Gaussian, gamma, inverse Gaussian, and power families. We derive the related formulas and relevant algorithms for each family and then discuss the ancillary or scale parameters appropriate to each model. We also examine noncanonical links and generalizations to the basic model. Finally, we give examples, showing how a given dataset may be analyzed using each family and link. We give examples dealing with model application, including discussion of the appropriate criteria for the analysis of fit. We have expanded the number of examples in this new edition to highlight both model fitting and assessment.

Part III deals with binomial response models. It includes exposition of the general binomial model and of the various links. Major links described include the canonical logit, as well as the noncanonical links probit, log-log, and complementary log-log. We also cover other links. We present examples and criteria for analysis of fit throughout. This new edition includes extensions to generalized binomial regression resulting from a special case of building a regression model from the generalized negative binomial probability function.

We also give considerable space to overdispersion. We discuss the problem's nature, how it is identified, and how it can be dealt with in the context of discovery and analysis.

Pursuant to the latter, we explain how to adjust the binomial model to accommodate overdispersion. You can accomplish this task by internal adjustment to the base model, or you may need to reformulate the base model itself. We also introduce methods of adjusting of the variance–covariance matrix of the model to produce robust standard errors. The problem of dealing with overdispersion continues in the chapters on data.

Part IV concerns count response data. We include examinations of the Poisson, the geometric, and the negative binomial models. With respect to the negative binomial, we show how the standard models can be further extended to derive a class called heterogeneous negative binomial models. There are several "brands" of negative binomial, and it is wise for the researcher to know how each is best used. The distinction of these models is typically denoted NB-1 and NB-2 and relates to the variance-to-mean ratio of the resulting derivation of the model. We have updated this discussion to include the generalized Poisson regression model, which is similar to NB-1.

Part V deals with the categorical response regression models. Typically considered extensions to the basic GLM algorithm, categorical response models are divided into two general varieties: unordered responses, also known as multinomial models, and ordered responses models. We begin by considering ordered-response models. In such models, the discrete number of outcomes are ordered, but the integer labels applied to the ordered levels of outcome are not necessarily equally spaced. A simple example is the set of outcomes "bad", "average", and "good". We also cover unordered multinomial responses, whose outcomes are given no order. For an example of an unordered outcome, consider choosing the type of entertainment that is available for an evening. The following choices may be given as "movie", "restaurant", "dancing", or "reading". Ordered response models are themselves divisible into two varieties: 1) ordered binomial including ordered logit, ordered probit, ordered complementary log-log, or ordered log-log and 2) the generalized ordered binomial model with the same links as the nongeneralized parameterization. We have expanded our discussion to include more ordered-outcome models, generalized ordered-outcome models, and continuation ratio models.

Finally, part VI is about extensions to the GLM family of models. In particular, we examine the following models:

1. Fixed-effects models
2. Random-effects models
3. Quasilikelihood models
4. GEEs
5. Generalized additive models

We attempt throughout to give the reader a thorough outline or overview of GLMs. We have attempted to cover nearly every major development in the area. We have also tried to show the direction in which statistical modeling is moving, hence laying a foundation for future research and for ever-more-appropriate GLMs. Moreover, we have expanded each section of the original version of this text to bring new and expanded

regression models into focus. Our attempt, as always, is to illustrate these new models within the context of the GLM.

1.5 Installing the support materials

All the data used in this book are freely available for you to download from the Stata Press web site, http://www.stata-press.com. In fact, when we introduce new datasets, we merely load them into Stata the same way that you would. For example,

```
. use http://www.stata-press.com/data/hh2/medpar
```

To download the datasets, do-files, and user-written commands for this book, type

```
. net from http://www.stata-press.com/data/hh2/
. net install glme2-ado1
. net install glme2-ado2
. net get glme2-data
```

The user-written commands will be automatically installed for your copy of Stata. The datasets and do-files will be downloaded to your current working directory. We suggest that you create a new directory into which the materials will be downloaded.