Generalized Linear Models

Charles E. McCulloch

August, 1999

Charles E. McCulloch is Professor, Departments of Statistical Science and Biometrics, Cornell University, Ithaca NY 14853. This is BU-1449-M in the Biometrics Unit Technical Report Series and was supported by NSF grant DMS-9625476.

Keywords: Probit, logistic, quasi-likelihood, nonlinear model

Abstract

The term "generalized linear models" encompasses both a class of models and a style of thinking about building models that is applicable to a wide variety of distributions and types of responses. Key elements are a distributional assumption for the data and a transformation of the mean which is assumed to create a linear model for the predictors. The history of generalized linear models is traced, current work is reviewed and some predictions are made.

1 Introduction and Some History

What is the difference, in absolute value, between logistic regression and discriminant analysis? I won't make you read this entire article to find the answer, which is 2. But you will have to read a bit further to find out why.

As most statisticians know, logistic regression and probit regression are commonly-used techniques for modelling a binary response variable as a function of one or more predictors. These techniques have a long history, with the word "probit" traced by David (1995) back to Bliss (1934) and Finney (1952) attributing the origin of the technique itself to psychologists in the late 1800's. In its earliest incarnations, probit analysis was little more than a transformation technique: scientists realized that the sigmoidal shape often observed in plots of observed proportions of successes versus a predictor x could be rendered a straight line by applying a transformation corresponding to the inverse of the normal c.d.f.

For example, Bliss (1934) describes an experiment in which nicotine is applied to aphids and the proportion killed is recorded (how is that for an early anti-smoking message?). Letting $\Phi^{-1}(\cdot)$ represent the inverse of the standard normal c.d.f., and $\hat{p}_i$ the observed proportion killed at dose d_i of the nicotine, Bliss exhibits a plot of $\Phi^{-1}(\hat{p}_i)$ versus $\log d_i$. The plot seems to indicate that a two segment linear regression model in $\log d_i$ is the appropriate model.

In an article a year later, Bliss (1935) explains the methodology in more detail as a weighted linear regression of $\Phi^{-1}(\hat{p}_i)$ on the predictor x_i using weights equal to $\frac{n_i \phi(p_i)^2}{\Phi(p_i)[1-\Phi(p_i)]}$, where $\phi(\cdot)$ represents the standard normal p.d.f. and n_i is the sample size for calculating $\hat{p}_i$. These weights can be easily derived as the inverse of the approximate variance found by applying the delta method to $\Phi^{-1}(\hat{p}_i)$.

This approach obviously has problems if an observed proportion is either zero or one. As a brief appendix to Bliss' paper, Fisher (1935) outlines the use of maximum likelihood to obtain estimates using data in which $\hat{p}_i$ is either zero or one. Herein lies a subtle change: Fisher is no longer describing a model for the transformed proportions but instead is directly modeling the mean of the binary response. Users of generalized linear models will recognize the distinction between a transformation and a link.

This technique of maximum likelihood is suggested only as a method of last resort when the observed proportions which are equal to zero or one must be incorporated in the analysis. The computational burdens were simply too high for it to be used on a regular basis in that era.

So, by the 1930's, models of the following form were being posited and fitting with the method of maximum likelihood was at least being entertained. With p_i denoting the probability of a success for the ith observation the model is given by

$$\begin{aligned} Y_i &\sim \text{indep. Bernoulli}(p_i) \\ p_i &= \Phi\left(\mathbf{x_i}'\boldsymbol{\beta}\right), \end{aligned} \tag{1}$$

where $\mathbf{x}_i'$ denotes the ith row of the matrix of predictors. With a slight abuse of notation, and in order to make this look similar to a linear model we can rewrite (1) as

$$\begin{aligned} \mathbf{Y} &\sim \text{indep. Bernoulli}(\mathbf{p}) \\ \mathbf{p} &= \Phi\left(\mathbf{X}\boldsymbol{\beta}\right) \end{aligned} \tag{2}$$

or equivalently

$$\Phi^{-1}\left(\mathbf{p}\right) = \mathbf{X}\boldsymbol{\beta}. \tag{3}$$

By 1952 this had changed little. In that year Finney more clearly describes the use of maximum likelihood for fitting these models in an appendix entitled "Mathematical basis of the probit method" and spends six pages in another appendix laying out the recommended computational method. This includes steps such as "34. Check steps 30-33" and the admonishment to the computer (a person!) that "A machine is not a complete safeguard against arithmetical errors, and carelessness will lead to wrong answers just as certainly as in non-mechanized calculations." This is clearly sage advice against overconfidence in output even from today's software.

He is practically apologetic about the effort required: "The chief hindrances to the more widespread adoption of the probit method ... (is) ... the

apparent laboriousness of the computations." He recognizes that his methods must be iterated until convergence to arrive at the maximum likelihood estimates but indicates that "With experience the first provisional line may often be drawn so accurately that only one cycle of the calculations is needed to give a satisfactory fit ..."

With computations so lengthy, what iterative method of fitting was employed? Finney recommended using "working probits," which he defined as (ignoring the shift of five units historically used to keep all the calculations positive):

$$z_i = \mathbf{x}_i'\boldsymbol{\beta} + \frac{y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})}{\phi(\mathbf{x}_i'\boldsymbol{\beta})}. \tag{4}$$

The working probits for a current value of $\boldsymbol{\beta}$ were regressed on the predictors using weights the same as suggested by Bliss, namely $\frac{\phi(p_i)^2}{\Phi(p_i)[1-\Phi(p_i)]}$, in order to get the new value of $\boldsymbol{\beta}$.

When I first learned about the EM algorithm (Dempster, Laird and Rubin, 1977), I was struck by its similarity to Finney's algorithm. A common representation of (1) is via a threshold model. That is, hypothesize a latent variable W_i such that

$$W_i \sim \text{indep. } N(\mathbf{x}_i'\boldsymbol{\beta}, 1). \tag{5}$$

Then, using $Y_i = I_{\{W_i>0\}}$ yields (1). To implement the EM algorithm, it is natural to regard the W_i as missing data and fill them in. Once the W_i are known, ordinary least squares can be used to get the new estimate of $\boldsymbol{\beta}$. The E-step fills in the W_i using the formula

$$E[W_i \mid Y_i] = \mathbf{x}_i'\boldsymbol{\beta} + \phi(\mathbf{x}_i'\boldsymbol{\beta})\frac{y_i - \Phi(\mathbf{x}_i'\boldsymbol{\beta})}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta})]}, \tag{6}$$

and the M-step estimates $\boldsymbol{\beta}$ as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$.

Thus the term added to $\mathbf{x}_i'\boldsymbol{\beta}$ in the EM algorithm is the same as the term added using working probits, once they are multiplied by the weight. Practical usage of EM and working probits, however, shows that working probits invariably converges much more quickly than does EM!

So as early as 1952 we see many of the key ingredients of generalized linear models: the use of "working variates" and link functions, fitting using a method of iteratively weighted fits, and the use of likelihood methods. But lack of computational resources simply did not allow widespread use of such techniques.

Logistic regression was similarly hampered. Over a decade later, Cox (1966) states, "Since the maximization of a function of many variables may not be straightforward, even with an electronic computer, it is worth having 'simple' methods for solving maximum likelihood equations, especially for use when there are single observations at each x value, so that the linearizing transformation is not applicable." Note the need for simple methods despite the fact that "computers" in 1966 are now machines.

For the logistic regression model akin to (2), namely:

$$\begin{aligned} \mathbf{Y} &\sim \text{indep. Bernoulli}(\mathbf{p}) \\ \mathbf{p} &= 1/(1+\exp[-\mathbf{X}\boldsymbol{\beta}]) \end{aligned} \tag{7}$$

or

$$\log[\mathbf{p}/(1-\mathbf{p})] = \mathbf{X}\boldsymbol{\beta},$$

it is straightforward to show that the maximum likelihood equations are given by

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{p}. \tag{8}$$

Since $\boldsymbol{\beta}$ enters $\mathbf{p}$ in a nonlinear fashion in (8) it is not possible to analytically solve this equation for $\boldsymbol{\beta}$. However, using the crude approximation (Cox, 1966), $1/(1+\exp[-t]) \approx \frac{1}{2} + \frac{t}{6}$, which is clearly only applicable for the mid-range of the curve, we can rewrite (8) approximately as:

$$\begin{aligned} \mathbf{X}'\mathbf{Y} &= \mathbf{X}'(\tfrac{1}{2}\mathbf{1} + \tfrac{1}{6}\mathbf{X}\boldsymbol{\beta}) \\ &= \tfrac{1}{2}\mathbf{X}'\mathbf{1} + \tfrac{1}{6}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{9}$$

We thus have

$$\mathbf{X}'(\mathbf{Y} - \tfrac{1}{2}\mathbf{1}) = \tfrac{1}{6}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{10}$$

which we can solve as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'6(\mathbf{Y} - \tfrac{1}{2}\mathbf{1}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*, \tag{11}$$

where Y_i^* is equal to 3 for a success and -3 for a failure. That is, we can approximate the logistic regression coefficients in a crude way by an ordinary least squares regression on a coded Y.

Logistic regression is often used as an alternate method for two-group discriminant analysis (Cox and Snell, 1989), by using the (binary) group identifier as the "response" and the multivariate vectors as the "predictors". This is a useful alternative when the usual multivariate normality

assumption for the multivariate vectors is questionable, e.g., when one or more variables are binary or categorical.

When it *is* reasonable to assume multivariate normality, the usual Fisher discriminant function is given by $\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$, where $\bar{\mathbf{X}}_i$ is the mean of the vectors for the ith group. If we code the successes and failures as 1 and -1 then $\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 = \mathbf{X}'\mathbf{Y}$. Thus we see that the difference between logistic regression and discriminant function analysis is 2, in absolute value.

2 Origins

Generalized linear models appeared on the statistical scene in the path-breaking paper of Nelder and Wedderburn (1972). Even though virtually all the pieces had previously existed, they were the first to put forth a unified framework which showed the similarities between seemingly disparate methods, such as probit regression, linear models, and contingency tables. They recognized that fitting a probit regression by iterative fits using the "working probits", namely (4), could be generalized in a straightforward way to unify a whole collection of maximum likelihood problems. Replacing $\Phi^{-1}(\cdot)$ with a general "link" function, $g(\cdot)$ and defining a "working variate" via

$$z \equiv g(\mu) + (y - \mu)g'(\mu) \tag{12}$$

gave, via iterative weighted least squares, a computational method for finding the maximum likelihood estimates. More formally we can write the model as follows:

$$\begin{aligned} Y_i &\sim \text{indep. } f_{Y_i}(y_i) \\ f_{Y_i}(y_i) &= \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) - c(y_i, \phi)\} \\ E[Y_i] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{13}$$

where $\boldsymbol{\theta}_i$ is a known function of $\boldsymbol{\beta}$ and $g(\cdot)$ is a known function which transforms (or links) the mean of y_i (not y_i itself!) to the linear predictor. The iterative algorithm is used to give maximum likelihood estimates of $\boldsymbol{\beta}$.

More importantly, it made possible a style of thinking which freed the data analyst from necessarily looking for a transformation which simultaneously achieved linearity in the predictors and normality of the distribution (as in Box and Cox, 1964).

I think of building generalized linear models by making three decisions:

1. What is the distribution of the data (for fixed values of the predictors and possibly after a transformation)?

2. What function of the mean will be modeled as linear in the predictors?

3. What will the predictors be?

What advantages does this have? First, it unifies what appear to be very different methodologies, which helps to understand, use and teach the techniques. Second, since the right-hand-side of the equation is a linear model after applying the link, many of the concepts of linear models carry over to GLMs. For example, the issues of full-rank versus overparameterized models are similar.

The application of generalized linear models became a reality in the mid 1970's when GLMs were incorporated into the statistics package GENSTAT and made available interactively in the GLIM software. Users of these packages could then handle linear regression, logistic and probit regression, Poisson regression, log-linear models, and regression with skewed continuous distributions, all in a consistent manner. Both packages are still widely used and are currently distributed by the Numerical Algorithms Group (www.nag.com). Of course, by now, most major statistical packages have facilities for generalized linear models, e.g., SAS Proc GENMOD.

GLM's received a tremendous boost with the development of quasi-likelihood by Wedderburn in 1974. Using only the mean to variance relationship, Wedderburn showed how statistical inference could still be conducted. Perhaps surprisingly, given the paucity of assumptions, these techniques often retain full or nearly full efficiency (Firth, 1987). Further, the important modification of *overdispersion* is allowed, that is, models with variance proportionally larger than predicted by the nominal distribution, say, a Poisson distribution. Such situations arise commonly in practice. Quasi-likelihood was put on a firmer theoretical basis in McCullagh (1983).

1983 also saw the publication of the first edition of the now-classic book, Generalized Linear Models (McCullagh and Nelder, 1983). With a nice blend of theory, practice and applications it made GLM's more widely used and appreciated. A colleague once asked me what I thought of the book Generalized Linear Models. I replied that it was absolutely wonderful and that the modeling and data analytic philosophy that it espoused was visionary. After going on for several minutes I noticed that he looked perplexed. When I inquired why he replied, "I think it is terrible - it has no theorems."

Perhaps that was the point.

3 Major Developments

Generalized linear models are now a mature data analytic methodology (e.g., Lindsey, 1996) and have been developed in numerous directions. There are techniques for choosing link functions and diagnosing link failures (e.g., Pregibon, 1980; Mallick and Gelfand, 1994) as well as research on the consequences of link misspecification (e.g., Li and Duan, 1989; Weisberg and Welsh, 1994). There are techniques for outlier detection and assessment of case influence for model checking (e.g., Pregibon, 1981; Cook and Croos-Dabrera, 1998). There are methods of modeling the dispersion parameters as a function of covariates (e.g., Efron, 1986) and for accommodating measurement error in the covariates (e.g., Stefanski and Carroll, 1990; Buzas and Stefanski, 1993). And there are ways to handle generalized additive models (Hastie and Tibshirani, 1990).

An extremely important extension of generalized linear models is the approach pioneered by Liang and Zeger (Liang and Zeger, 1986; Zeger and Liang, 1986) that is known as generalized estimating equations (GEEs). GEEs made, in my opinion, two valuable contributions: the accommodation of a wide array of correlated data structures and the popularization of the "robust sandwich estimator" of the variance-covariance structure. Current software implementations of GEEs are mostly designed to accommodate longitudinal data structures, i.e., ones in which the data can be arranged as repeat measurements on a series of independent "subjects" (broadly interpreted, of course). The use of the "robust sandwich estimator" which basically goes back to Royall (1986) and Huber (1967) allows the specification of a "working" covariance structure. That is, the data analyst must specify a guess as to the correct covariance structure, but inferences remain asymptotically valid even if this structure is incorrectly specified (as it always is to some degree). Not surprisingly, the efficiency of inferences can be affected if the "working" structure is far from truth (e.g., Fitzmaurice, 1995).

Distribution theory for modifications of exponential families for use in generalized linear models has been developed further in, for example, Jorgensen (1997) and the theory of quasi-likelihood is detailed in the book-length treatment of Heyde (1997).

4 Looking Forward

Anyone making predictions runs the risk of someone actually checking later to see if the predictions are correct. So I am counting on the "Jean Dixon effect," defined by the Skeptic's Dictionary (http://skeptic.com) as "the tendency of the mass media to hype or exaggerate a few correct predictions by a psychic, guaranteeing that they will be remembered, while forgetting or ignoring the much more numerous incorrect predictions."

Since likelihood and quasi-likelihood methods are based on large sample approximations, an important area of development will be the construction of tests and confidence intervals which are accurate in small and moderate sized samples. This may be through "small sample asymptotics" (e.g., Skovgaard, 1996; Jorgensen, 1997) or via computationally intensive methods like the bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997).

The extension of generalized linear models to more complex correlation structures has been an area of active research and will see more developments. Models for time series (e.g., Chan and Ledolter, 1995), random effects models (e.g., Stiratelli, Laird and Ware, 1984; Breslow and Clayton, 1993) and spatial models (e.g., Heagerty and Lele, 1999) have all been proposed. Unfortunately, likelihood analysis of many of the models lead to intractable, high-dimensional integrals. So likewise, computing methods for these models will continue to be an ongoing area of development. McCulloch (1997) and Booth and Hobert (1999) use a Monte Carlo EM approach, Quintana, Liu and del Pino (1999) use a stochastic approximation algorithm and Heagerty and Lele (1999) take a composite likelihood tack.

Attempts to avoid likelihood analysis via techniques such as penalized quasi-likelihood (for a description see Breslow and Clayton, 1993) have not been entirely successful. Approaches based on working variates (e.g., Schall, 1991) and Laplace approximations (e.g., Wolfinger, 1994) generate inconsistent estimates (Breslow and Lin, 1994) and can be badly biased for distributions far from normal (i.e., the important case of Bernoulli-distributed data). Clearly, reliable and well-tested, general-purpose fitting algorithms need to be developed before these models will see regular use in practice.

The inclusion of random effects in generalized linear models raises several additional questions: What is the effect of misspecification of the random effects distribution (e.g. Neuhaus, Hauck, and Kalbfleisch, 1992) and how can it be diagnosed? What is the best way to predict the random effects and how can prediction limits be set, especially in small and moderate sized samples (e.g. Booth and Hobert, 1998)? How can outlying random effects be

identified or downweighted? All of these are important practical questions which must be thoroughly investigated for regular data analysis.

The whole idea behind generalized linear models is the development of a strategy and philosophy for approaching statistical problems, especially those involving non-normally distributed data, in a way that retains much of the simplicity of linear models. Areas in which linear models have been heavily used (e.g. simultaneous equation modeling in econometrics) have and will see adaptations for generalized linear models. As such, generalized linear models will continue in broad use and development for some time to come.

References

Bliss, C. (1934). The method of probits. *Science*, **79**, 38-39.

Bliss, C. (1935). The calculation of the dose-mortality curve. *Annals of Applied Biology*, **22**, 134–167.

Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American. Statistical Association*, **93**, 262–272.

Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**, 265–285.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. (With discussion) *Journal of the Royal Statistical Society, Series B*, **26**, 211-252.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N.E. and Lin, X. (1994). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Buzas, J.S. and Stefanski, L.A. (1996). Instrumental variable estimation in generalized linear measurement error models. *Journal of the American Statistical Association*, **91**, 999–1006.

Chan, K.S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, **90**, 242–252.

Cook, R.D. and Croos-Dabrera, R. (1998). Partial residual plots in generalized linear models. *Journal of the American Statistical Association*, **93**, 730–739.

Cox, D.R. (1966). Some procedures connected with the logistic qualitative response curve. In *Research Papers in Statistics* (F.N. David, Ed.), 55–72.

Cox, D.R., and Snell, E.J. (1989). *Analysis of Binary Data.* London: Chapman and Hall.

David, H.A. (1995). First (?) occurrence of common terms in probability and statistics. *The American Statistician*, **49**, 121–133.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press: Cambridge.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm.(With discussion) *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 709–721.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall: New York.

Finney, D.J. (1952). *Probit Analysis.* Cambridge University Press: Cambridge.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika*, **74**, 233–245.

Fisher, R.A. (1935). Appendix to the calculation of the dose-mortality curve (by Bliss). *Annals of Applied Biology*, **22**, 164–165.

Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, **51**, 309–317.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman and Hall: London.

Heagerty P. and Lele, S. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**, 1099–1111.

Heyde, C.C. (1997). *Quasi-likelihood and its Application*. Springer: New York.

Huber, P.J. (1967). The behaviour of maximum likelihood estimators under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, LeCam, L.M., and Neyman, J. (eds.). pp 221–233.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman and Hall: London.

Li, K.-C., and Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics*, **17**, 1009–1052.

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lindsey, J.K. (1996). *Applying Generalized Linear Models*. New York: Springer.

Mallick, B.K. and Gelfand, A.E. (1994). Generalized linear models with unknown link functions. *Biometrika*, **81**, 237–245.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, **11**, 59–67.

McCullagh, P., and Nelder, J. (1983). *Generalized Linear Models*. Chapman and Hall: London.

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

Nelder, J.A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*, **29**, 15–24.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–724.

Quintana, R., Liu, J., and del Pino, G. (1999). Monte Carlo EM with importance reweighting and it application in random effects models. *Computational Statistics and Data Analysis*, **29**, 429–444.

Royall, R. (1986). Model robust inference using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*. **78**, 719–727.

Skovgaard, I.M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, **2**, 145–165.

Stefanski, L.A. and Carroll, R.J. (1990). Score tests in generalized linear measurement error modelsa. *Journal of the Royal Statistical Society, Series B*, **52**, 345–359.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Weisberg, S., and Welsh, A.H. (1994). Adapting for the missing link. *Annals of Statistics*, **22**, 1674–1700.

Wolfinger, R.W. (1994). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791–795.

Zeger, S., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.