GLMM Parameter Estimation

Mickey Warner

Introduction

In this paper we review "Approximate Inference in Generalized Linear Mixed Models" by Breslow and Clayton (1993), referred to hereafter as the authors. The generalized linear model (GLM) (McCullagh and Nelder, 1989) allows us to work with non-normal likelihoods, such as for binary, count, or non-nonegative data, by specifying a link function between the mean of a random variable and a systematic component.

For example, suppose Y has some distribution F with mean μ. We could specify a link function $g(\cdot)$ such that $g(\mu) = \mathbf{x}^\top \boldsymbol{\alpha}$. This form is now familiar to linear regression. The regression coefficients $\boldsymbol{\alpha}$ are often estimated through some iterative procedure (such as IWLS). These coefficients are then interpreted with respect to the function of the mean. Yet, suitable choices of $g(\cdot)$, with its inverse $h(\cdot) = g^{-1}(\cdot)$, may still produce easily interpretable results.

Linear mixed models (Henderson et al., 1959) assume there are additional sources of variation, beyond measurement error. Such models have the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

where typical assumptions are $\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ with $\mathbf{b}$ and $\mathbf{e}$ independent. These models are useful for repeated measures or blocked data or other complicated designs. The power of these models comes from specifying the covariances $\mathbf{G}$ and $\mathbf{R}$, either as unstructured or having some known form. Many choices of covariance structures are available for use. Methods exist for estimating the fixed and random effects and also the variance-covariance parameters.

Generalized linear mixed models (GLMMs) are a hybrid of the GLM and the linear mixed model. We have a need to use a non-normal likelihood, but also wish to model random effects and covariances. However, analytic solutions often require numerical integration, which becomes infeasible in high dimensions. The authors address the capability of Bayesian methods in avoiding the need for numerical integration. However, they criticize

the Bayesian approach due to its need for intense computational resources. Though significant advances in scientific computing and computational resources have made Bayesian methods far more accessible. At the time of their writing (some twenty years ago), their method of parameter estimation and inference was groundbreaking and is still a powerful tool.

The remainder of the paper will proceed as follows. In section 1 we describe in more detail the model that is considered by the authors. A summary of the model fitting procedure is given in section 2, and an example is presented in section 3. Section 4 provides concluding remarks.

1   Hierarchical model

For an observation i, a univariate response y_i is measured with covariates $\mathbf{x}_i$ and $\mathbf{z}_i$ associated with the fixed and random effects, respectively. Units may be blocked in some way. Given a q-vector $\mathbf{b}$ of random effects, the responses are assumed to be conditionally independent with means $\mathrm{E}(y_i|\mathbf{b}) = \mu_i^b$ and variances $\mathrm{Var}(y_i|\mathbf{b}) = \phi a_i v(\mu_i^b)$, where ϕ is a dispersion parameter, a_i is a known constant, and $v(\cdot)$ is a specified variance function. (Note: in μ_i^b, the b is not a power, but a superscript to denote that the mean depends on the random effects.) The conditional mean is related to the linear component $\eta_i^b = \mathbf{x}_i^\top \boldsymbol{\alpha} + \mathbf{z}_i^\top \mathbf{b}$ through the link function $g(\mu_i^b) = \eta_i^b$, with inverse $h(\cdot) = g^{-1}(\cdot)$, where $\boldsymbol{\alpha}$ is a p-vector of fixed effects.

Denote $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and let $\mathbf{X}$ and $\mathbf{Z}$ have rows comprised of $\mathbf{x}_i^\top$ and $\mathbf{z}_i^\top$. We then write the conditional means using vector notation as

$$\mathrm{E}(\mathbf{y}|\mathbf{b}) = \boldsymbol{\mu}^b = h(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b}). \tag{1}$$

Finally, we assume $\mathbf{b}$ is multivariate normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$ for an unknown vector $\boldsymbol{\theta}$ of covariance parameters.

We now give some remarks on this specification. The conditional variance has three terms, ϕ, a_i, and $v(\mu_i^b)$. The dispersion parameter ϕ may be estimated, but is occasionally fixed at unity. The authors used a_i as the "reciprocal of a binomial denominator." That is, for experiments involving binomial outcomes, a_i is the number of replicates on each observation. In the examples the authors provide, a_i is fixed at one. The choice of $v(\cdot)$ should be thought of as choosing the likelihood for the data. For example, for count data, $v(\mu_i^b) = \mu_i^b$ is comparable to the Poisson distribution, while $v(\mu_i^b) = \mu_i^b(1 - \mu_i^b)$ is similar to the binomial distribution. Note that we are

2

not necessarily making any distributional assumptions about the likelihood. This requires the use of a quasi-likelihood, as discussed in the next section. The link function $g(\cdot)$ must be monotonic and maps the real numbers onto the support for the mean. Choices include the logit or inverse cdf of the normal distribution for binomial data and the natural logarithm for Poisson data. Ease of interpretation and adequacy of model fit should be taken into account when deciding on a link function.

2 Penalized quasi-likelihood

As mentioned earlier, the assumptions made about the conditional mean and variance do not equate to choosing a likelihood. Rather, the link and variance functions amount to specifying first and second moments for the data. The authors propose a method of parameter estimation through various approximations of the penalized quasi-likelihood (PQL) and iterative updates.

Quasi-likelihoods provide an alternative to full-likelihoods. The authors omit the details on quasi-likelihoods, but we will provide some explanation here. Wedderburn (1974) defines the quasi-likelihood of a single observation as

$$\frac{\partial K(y_i|\mu_i^b)}{\partial \mu_i^b} = \frac{y_i - \mu_i^b}{a_i v(\mu_i^b)} \tag{2}$$

$$\implies K(y_i|\mu_i^b) = \int_{y_i}^{\mu_i^b} \frac{y_i - u_i}{a_i v(u_i)} du_i, \tag{3}$$

using the notation from section 1. The left side of the equal sign in (2) shares similar properties to the derivative of a log-likelihood (the score). For instance, after straightforward calculations, we have

$$\mathrm{E}_Y\left(\frac{\partial K(y_i|\mu_i^b)}{\partial \mu_i^b}\right) = 0 \tag{4}$$

$$\mathrm{Var}_Y\left(\frac{\partial K(y_i|\mu_i^b)}{\partial \mu_i^b}\right) = \frac{1}{a_i v(\mu_i^b)}. \tag{5}$$

For simple choices of $v(\mu_i^b)$ the integral in (3) is tractable. Certain choices also correspond to the log-likelihoods of particular distributions: $v(\mu_i^b) = 1$ results in a Gaussian likelihood, $v(\mu_i^b) = \mu_i^b$ a Poisson, $v(\mu_i^b) = (\mu_i^b)^2$ a Gamma, and $v(\mu_i^b) = \mu_i^b(1 - \mu_i^b)$ a Binomial, to name a few.

The quasi-likelihood for the set of n observations is the following sum

$$Q(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{b}) = \sum_{i=1}^{n} K(y_i|\mu_i^b). \tag{6}$$

The left side of the equation indicates that the quasi-likelihood is conditional on the parameters and random effects (recall that μ_i^b is a function of $\mathbf{b}$). Using Bayes theorem, we have the joint distribution

$$ql(\mathbf{y}, \mathbf{b}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = Q(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{b})p(\mathbf{b}). \tag{7}$$

Since $Q(\cdot)$ resembles a log-likelihood, we exponentiate it and then integrate out $\mathbf{b}$ from the joint in (7) to obtain the authors' integrated quasi-likelihood function, their equation (2),

$$e^{ql(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta})} = \int \exp\left[Q(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}, \mathbf{b})\right] p(\mathbf{b})d\mathbf{b} \tag{8}$$

$$\propto |\mathbf{D}|^{-1/2} \int \exp\left[\sum_{i=1}^{n} K(y_i|\mu_i^b) - \mathbf{b}^{\top}\mathbf{D}^{-1}\mathbf{b}\right] d\mathbf{b}, \tag{9}$$

where $p(\mathbf{b})$ is the multivariate normal density with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$. This is the function used to obtain estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. However, there is a clear difficulty in integrating the function. The authors use Laplace's method for integral approximation to obtain a second order approximation of (9).

Laplace's integral approximation is given by

$$\int e^{Mf(x)}dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} e^{Mf(x_0)} \tag{10}$$

for large M, where $f(x_0)$ is the global maximum of the function f. The approximation is generalized when integrating with respect to a vector (see Wikipedia). Let $\kappa(\mathbf{b}) = \sum_{i=1}^{n} K(y_i|\mu_i^b) - \mathbf{b}^{\top}\mathbf{D}\mathbf{b}$. Let $\kappa'(\mathbf{b})$ and $\kappa''(\mathbf{b})$ be the q-vector and $q \times q$ matrix of first- and second-order partial derivatives with respect to $\mathbf{b}$. After applying Laplace's method and taking the log of (9) and ignoring constants, we have

$$ql(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\log|\kappa''(\tilde{\mathbf{b}})| - \kappa(\tilde{\mathbf{b}}), \tag{11}$$

4

where $\tilde{\mathbf{b}}$ is the solution to $\kappa'(\mathbf{b}) = \mathbf{0}$. From here, the authors solve for the estimating equations for $\boldsymbol{\alpha}$ and $\mathbf{b}$. An iterated weighted least squares (IWLS) algorithm may be used to obtain estimates for these parameters.

When estimates for $\boldsymbol{\alpha}$ and $\mathbf{b}$ are calculated, the variance parameters $\boldsymbol{\theta}$ are estimated through an approximate profile quasi-likelihood. For the sake of space, we omit the technical details and refer the reader to the paper we have been discussing.

2.1 Fitting the model with available software

There are several software packages available that will fit generalized linear mixed models using the PQL approach. The `MASS` packages containts the `glmmPQL()` function for fitting models via PQL. Packages such as `lme4` and `MCMCglmm` contain functions that fit GLMMs, but not necessarily through PQL.

In SAS, `PROC GLIMMIX` has options to fit models with the PQL method. The benefit to using PQL is that the first- and second-order approximations guarantees that parameters for most models will be estimable, a significant advantage over methods which are unable to produce estimates due to overly-complex covariance specifications.

3 Example: epilepsy study

epilepsy data Thall and Vail (1990)

4 Conclusion

bayesian analysis on the epilepsy data Fong et al. (2009)

References

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

Fong, Y., Rue, H., and Wakefield, J. (2009), "Bayesian inference for generalized linear mixed models," *Biostatistics*, kxp053.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959), "The estimation of environmental and genetic trends from records subject to culling," *Biometrics*, 15, 192–218.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 2, Chapman and Hall London.

Thall, P. F. and Vail, S. C. (1990), "Some covariance models for longitudinal count data with overdispersion," *Biometrics*, 657–671.

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