Chapter 8

Other estimation methods for GLMMs

8.1 Introduction

In this chapter I consider alternatives to maximum likelihood, which we saw are computationally demanding. Generalized estimating equations (GEEs) are a popular approach to fitting a related class of models. I also consider Penalized Quasi-Likelihood (PQL) and variants and the more general topic of estimating equations as a unifying theme. In a number of points in this chapter I fall back on the linear mixed model to give insight in a situation in which calculations are a bit more straightforward.

8.2 Generalized estimating equations

Generalized Estimating Equations (GEEs) is a computationally less demanding method than ML estimation. It has mainly been designed as a marginal modeling approach for longitudinal data, wherein data is collected on "subjects" on two or more occasions. It works best when the number of occasions is small compared to the number of subjects.

To motivate the ideas it is easiest to return to linear models. Consider the standard linear model with a full-rank $\mathbf{X}$ matrix:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

with well-known ordinary least squares estimator (OLSE)

$$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{8.1}$$

The OLSE is optimal (best linear unbiased or best unbiased) in the situation in which the data can be assumed to be homoscedastic and uncorrelated (with a further normality assumption to get best unbiased). However, it works well in many situations for which those assumptions are not met. In particular it is always unbiased when $\mathbf{X}$ can be regarded as fixed:

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}_{ols}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{8.2}$$

Furthermore, it is often relatively efficient. To see this we need to calculate the variance of the OLSE, which is straightforward. Let $\mathbf{V}$ denote the variance-covariance matrix of $\mathbf{Y}$. Then we have

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}) &= \mathrm{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{8.3}$$

If $\mathbf{V}$ were known we could use the weighted least squares estimator (WLSE), which is then more efficient than OLS

$$\hat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{8.4}$$

with variance

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{wls}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{V})^{-1}. \tag{8.5}$$

From this we can make some comparisons. Consider the simple linear regression model

$$\mathrm{E}[Y_{ij}] = \beta_0 + \beta_1 x_j, \tag{8.6}$$

for $i = 1, \ldots, k$ clusters and $x_j = -2, -1, 0, 1, 2$ as j goes from 1 to 5. Let $\mathbf{Y}_i$ represent the vector of five measurements within cluster i for $j = 1, \ldots, 5$ and consider various assumptions for $\mathbf{V}_i = \mathrm{Var}(\mathbf{Y}_i)$.

If $\mathbf{V}_i = \mathbf{I}_5$ then WLS reduces to OLS and, of course, OLS is fully efficient. If the observations within a cluster are equicorrelated, namely $\mathbf{V}_i = (1-\rho)\mathbf{I}_5 + \rho\mathbf{J}_5$ (where $\mathbf{J}_n$ denotes an $n \times n$ matrix of all ones) then OLS is again fully efficient. This is perhaps a bit surprising, but takes advantage of the fact that, with many forms of balanced data the OLSE is exactly equal to the WLSE (Zyskind, 1969). Finally, consider an example with an autoregressive structure with $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma^2\rho^{|j-k|}$. Table 8.1 gives the ratio of the variances of the OLSE and WLSE (assuming $\mathbf{V}_i$ known) of estimating β_1 in the model (8.6) for various values of ρ and the autoregressive structure and using $k = 10$. When $\rho = 0$ we are back to the situation in which the WLSE reduces to the OLSE and the ratio of variances is one. Perhaps surprisingly, the efficiency is always quite high and does not decrease monotonically as ρ increases.

While it is certainly possible to find situations in which the OLSE performs arbitrarily poorly compared to the WLSE, there are a wide variety of situations in which the OLSE is relatively efficient.

If the OLSE is relatively efficient, easy to calculate, and does not require estimation of the variance covariance structure, why not just perform OLS estimation

TABLE 8.1.
Efficiency of OLS to WLS in a simple linear regression model

	Autoregressive correlation ρ						
	0	0.1	0.3	0.5	0.7	0.9	0.99
$\text{Var}(\hat{\boldsymbol{\beta}}_{wls})/\text{Var}(\hat{\boldsymbol{\beta}}_{ols})$	1	0.997	0.980	0.962	0.952	0.955	0.961

on a regular basis? The answer is that, while the estimator is fine, estimation of its variance by the usual methods can be far from correct. It is not at all unusual for the estimated variance of an OLSE using the standard OLS calculations (which assumes all the observations are independent) to be off by a factor of two or more. And, unfortunately, it can be either too big or too small.

a. Non-normal data

In modeling non-normal data it is often common to use nonlinear models. This leads to a distinction between the marginal and conditional models, which can be quite different (see Section 1.2a). Examples of the marginal modeling approach include Zhao and Prentice (1990); Prentice and Zhao (1991); Liang et al. (1992); Zhao et al. (1992); Fitzmaurice and Laird (1993) and Fitzmaurice et al. (1994).

GEEs work most naturally for models specified marginally. In contrast, GLMMs are specified conditional on the random effects. We now elaborate on this difference in the context of an example.

Let $Y_{ij} = 1$ if the ith woman miscarries during her jth pregnancy and is 0 otherwise. We hypothesize a probit model,

$$\begin{aligned} Y_{ij}|\mathbf{u} &\sim \text{indep. Bernoulli}(p_{ij}), \\ p_{ij} &= \text{E}[Y_{ij}|\mathbf{u}] = \Phi(\mu + \beta x_{ij} + u_i), \end{aligned}$$

which yields

$$\begin{aligned} \text{E}[Y_{ij}|\mathbf{u}] &= \Phi\left(\frac{\mu + \beta x_{ij}}{\sqrt{1+\sigma_u^2}}\right) \\ &\equiv \Phi(\mu^* + \beta^* x_{ij}), \end{aligned} \tag{8.7}$$

with $\beta^* = \beta/\sqrt{1+\sigma_u^2}$. Thus, β represents the conditional effect of x_{ij} on a probit scale and β^* represents the marginal effect of x_{ij}. What are the interpretations of β and β^*?

The interpretation of β is the increase in $\Phi^{-1}(\cdot)$ of the probability of a miscarriage for each woman associated with each increase in the order of childbirth. The interpretation of β^* is the increase in $\Phi^{-1}(\cdot)$ of the probability of miscarriage *averaged over all women* associated with each increase in the order of childbirth. Because of the nonlinear model, averaging before versus after applying the function $\Phi^{-1}(\cdot)$ gives different answers.

The use of GEEs typically proceeds by specifying a marginal model. For example, for binary data, we could specify

$$\begin{aligned} \mathrm{E}[Y_{ij}] &= p_{ij}, \\ \operatorname{logit}(\mathbf{p}_i) &= \mathbf{X}_i\boldsymbol{\beta}. \end{aligned} \tag{8.8}$$

Estimates are obtained by solving the generalized estimating equations for $\hat{\boldsymbol{\beta}}$:

$$\sum_{i=1}^{n} \left(\frac{\partial \mathbf{p}_i}{\partial \boldsymbol{\beta}}\right)' [\mathrm{WVar}(\mathbf{Y}_i)]^{-1}(\mathbf{Y}_i - \mathbf{p}_i) = 0 \tag{8.9}$$

where WVar(·) indicates a "working" or assumed covariance structure, possibly dependent on unknown parameters.

This has properties similar to the estimating equations for the linear mixed model, which is given by

$$\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) = 0. \tag{8.10}$$

b. Comparison of marginal versus conditional modeling

In some cases, the inferential goals of a problem lead one clearly to marginal or conditional specification of the model. In other cases it is less clear and hence it is useful to contrast the approaches.

Marginal models have the following advantages:

1. Marginal models avoid the specification of the conditional structure; when only marginal questions are of interest, misspecification of this portion of the model can be avoided.

2. When paired with a GEE approach to estimation, estimates of the marginal parameter estimates are consistent, even under misspecification of the association structure.

3. When the underlying random effects distribution is heteroscedastic, assuming it is homoscedastic and using a conditional approach can lead to biased estimators (Heagerty and Zeger, 2000; Heagerty and Kurland, 2001).

However the marginal approach also has drawbacks. These include:

1. The form of a conditional mixing distribution required for typically assumed marginal models may be unusual (e.g., Wang and Louis, 2002).

2. The form of the marginal distribution may not be the same as any of the conditional models. In extreme circumstances, features of scientific interest present in every conditional model may not be present in the marginal model.

3. If the question is of marginal interest, a longitudinal design may not be the most appropriate.

4. Marginal analyses are subject to the Ecological Fallacy and Simpson's paradox, potentially giving misleading results.

5. Marginal quantities can be calculated from a conditional model but the converse is not typically true.

A more detailed discussion of some of the critiques of marginal modeling can be found in Lindsey and Lambert (1998).

c. Comparison of GEE and random effects estimation methods

A big advantage of the GEE approach is the ability to use a "robust" variance estimate. In such a case the inferences about the mean structure are asymptotically valid, even when the working variance is incorrect. This offers a useful tool for inference or, at least, model checking.

Drawbacks to standard application of GEEs include:

- Difficulty in adapting GEEs for GLMMs. GEEs are most naturally adapted to marginal models, not the conditional random effects models of GLMMs. But see Zeger et al. (1988) and Heagerty (1999) for some results in this direction.

- GEEs by themselves do not help to separate out different sources of variation. It is often an advantage to be able to attribute variation as being associated with different factors.

- GEEs are not directly a technology for best prediction of random effects. Besides not having the random effects naturally imbedded, best prediction requires further knowledge of the distribution than is typically provided for GEEs. But see Waclawiw and Liang (1993).

- GEEs are not the best technique for other-than-longitudinal data, either crossed or nested random factors.

- GEEs may not perform well in situations where the number of time points is large in comparison to the number of "subjects" or when there is much missing or unequally spaced data.

- GEEs may be inefficient when the goal is estimation of the variance covariance structure.

To address this last concern, an improvement over the standard GEE methods was developed, called GEE2 (Prentice, 1988; Zhao and Prentice, 1990; Prentice and Zhao, 1991). GEE2 uses sums of squares and cross-products of the data vector in a secondary set of estimating equations to supplement the equations for $\boldsymbol{\beta}$. This comes at a cost. First there are a large number of squares and cross-products. Second, we now need to specify a working correlation structure for the squares and cross-products. Finally, we lose one of the attractive properties of GEEs, the consistency under misspecification of the association structure. Further improvements to overcome this last deficiency are addressed in Qu et al. (2000).

8.3 Dispersion–mean model

GEE2 is closely related to an earlier idea, due to Pukelsheim and developed by Anderson et al. (1984) for approximately normally distributed data. The description below borrows from Searle et al. (1992). Consider the variance components linear mixed model (2.9),

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i, \mathbf{R}\right), \\ \mathbf{u}_i &\sim \text{ indep. } \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_i^2), \end{aligned} \tag{8.11}$$

and define $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ along with $\mathcal{Y} \equiv \mathbf{MY} \otimes \mathbf{MY}$. In words, $\mathbf{MY}$ are the residuals after an OLS fit on the column space of $\mathbf{X}$ and the Kronecker product creates the sums of squares and cross-products. It is not surprising that estimation of the variance components is based on $\mathcal{Y}$.

To deduce how to use $\mathcal{Y}$ to estimate the variance components we start by calculating its expected value

$$\begin{aligned} \mathrm{E}[\mathcal{Y}] &= \mathrm{E}[\mathbf{MY} \otimes \mathbf{MY}] \\ &= \mathrm{E}[\mathbf{M}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \otimes \mathbf{M}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})] \\ &= \mathrm{E}[(\mathbf{M} \otimes \mathbf{M})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \otimes (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})] \\ &= (\mathbf{M} \otimes \mathbf{M})\mathrm{E}[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \otimes (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})] \\ &= (\mathbf{M} \otimes \mathbf{M})\mathrm{E}[\mathrm{vec}\{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\}] \\ &= (\mathbf{M} \otimes \mathbf{M})\mathrm{vec}\{\mathrm{E}[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})']\} \\ &= (\mathbf{M} \otimes \mathbf{M})\mathrm{vec}(\mathrm{Var}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})] \\ &= (\mathbf{M} \otimes \mathbf{M})\mathrm{vec}(\mathbf{V}), \end{aligned} \tag{8.12}$$

where $\mathbf{V}$=Var($\mathbf{Y}$),vec($\cdot$) is the matrix operator that stacks the columns of a matrix into a large column vector, $\mathrm{vec}(\mathbf{V}) = \{\mathrm{vec}(\mathbf{Z}_i\mathbf{Z}_i')\sigma_i^2\}$ and $\{\cdot\}$ indicates a block "column" matrix with blocks given inside of the parentheses.

On defining $\mathcal{X} = \mathbf{M} \otimes \mathbf{M}\{\mathrm{vec}(\mathbf{Z}_i\mathbf{Z}_i')\}$ we can rewrite (8.12) as

$$\mathrm{E}[\mathcal{Y}] = \mathcal{X}\sigma^2. \tag{8.13}$$

In words, (8.13) is a linear model for the variance components with a data vector given by the sums of squares and cross-products of the residuals. OLS applied to (8.13) yields an old variance components estimation technique called MINQUE0 (Hartley et al., 1978). Generalized least squares applied to this model gives the REML equations.

8.4 Penalized quasi-likelihood

Estimation in GLMs proceeds by developing a "working variate" through a Taylor series expansion of the link function. Will this work for GLMMs? I start with the

GLMM of (4.5) written somewhat more informally:

$$\begin{aligned} \mathbf{Y}|\mathbf{u} &\sim \text{ exponential family with mean } \boldsymbol{\mu}, \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \\ \mathbf{u} &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}). \end{aligned}$$

Next I expand $g(\cdot)$:

$$\begin{aligned} g(\mathbf{y}) &\approx g(\boldsymbol{\mu}) + (\mathbf{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + (\mathbf{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon g'(\boldsymbol{\mu}). \end{aligned}$$

The idea is then to treat $\mathbf{w} \equiv g(\boldsymbol{\mu}) + (\mathbf{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu})$ as a LMM with

$$\text{(8.14)} \qquad \text{var}(\mathbf{w}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{D}_g\mathbf{R}\mathbf{D}_g,$$

where $\mathbf{D}_g = \text{diag}\{g'(\boldsymbol{\mu})_i\}$ and $\mathbf{R} = \text{var}(\mathbf{y} - \boldsymbol{\mu})$.

One way to use this working variate approximation is to solve the mixed model equations (2.22) for $\hat{\boldsymbol{\beta}}$ and $\tilde{\mathbf{u}}_{blup}$. The $\tilde{\mathbf{u}}_{blup}$ is then used to obtain an estimate of $\mathbf{D}$ and the approximation is repeated, using the updated values of $\boldsymbol{\beta}$ and $\mathbf{u}$. Schall (1991) proposed a method along these lines as well as ways to get approximate standard errors.

The development so far does not explain the name *penalized quasi-likelihood.* For that we give some details of a Poisson-normal model. The development is patterned after Breslow and Clayton (1993).

Consider the following simple model with a single random effect and variance proportional to the mean:

$$\text{(8.15)} \qquad \begin{aligned} \text{E}[Y_{ij}|u_i] &= \mu_{ij}, \qquad i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i, \\ \log \mu_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i, \\ \text{var}(Y_{ij}|u_i) &= \tau^2\mu_{ij}, \\ u_i &\sim \text{ i.i.d. } \mathcal{N}(0, \sigma_u^2). \end{aligned}$$

In this model, the data come in m correlated clusters, indexed by i. The log quasi-likelihood corresponding to (8.15) would be

$$\text{(8.16)} \qquad \log QL(\boldsymbol{\beta}, \mathbf{u}) = \left(\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \sum_i y_{i\cdot}u_i - \sum_i e^{u_i} \sum_j e^{\mathbf{x}_{ij}'\boldsymbol{\beta}}\right) /\tau^2.$$

Since this includes the latent and unknown $\mathbf{u}$ it is not clear exactly how to use this log quasi-likelihood. One way would be to consider integrating the quasi-likelihood against the distribution of $\mathbf{u}$ giving

$$\text{(8.17)} \qquad \begin{aligned} &\text{Integrated } QL \\ &= \int \cdots \int \exp\{\log QL(\boldsymbol{\beta}, \mathbf{u}) - \mathbf{u}'\mathbf{u}/2\}(2\pi\sigma_u^2)^{-m/2}d\mathbf{u}. \end{aligned}$$

This is no easier to evaluate than the likelihood (it *is* the likelihood for a Poisson-normal model excepting the additional τ^2). However, the form of (8.17) suggests the use of a Laplace approximation.

That is, to approximate an integral of the form $\int_{-\infty}^{+\infty} e^{-\kappa(u)}du$, we utilize a Taylor series expansion on κ to obtain

$$\kappa(u) \approx \kappa(u_0) + \kappa'(u_0)(u-u_0) + \tfrac{1}{2}\kappa''(u_0)(u-u_0)^2. \tag{8.18}$$

If we chose u_0 so that $\kappa'(u_0) = 0$ (i.e., a mode of κ) then the expansion simplifies to

$$\kappa(u) \approx \kappa(u_0) + \tfrac{1}{2}\kappa''(u_0)(u-u_0)^2. \tag{8.19}$$

Using the approximation (8.19) in place of κ in the integral gives

$$\begin{aligned}\int_{-\infty}^{+\infty} e^{-\kappa(u)}du &\approx \int_{-\infty}^{+\infty} e^{\kappa(u_0)+\,\frac{1}{2}\kappa''(u_0)(u-u_0)^2}du \\ &= e^{-\kappa(u_0)}\sqrt{2\pi/\kappa''(u_0)}.\end{aligned} \tag{8.20}$$

Applying the multidimensional analog of this to our integral (8.17) requires $\mathbf{u}_0$ as the solution to

$$\tau^2\mathbf{D}^{-1}\mathbf{u}_0 = \mathbf{Z}'(\mathbf{y} - \mathrm{E}[\mathbf{y}|\mathbf{u}_0]), \tag{8.21}$$

with an approximation to the log of the integrated quasi-likelihood of

$$\begin{aligned}&-\tfrac{1}{2}\log|\mathbf{I} + \mathbf{Z}'\mathbf{W}\mathbf{Z}\mathbf{D}/\tau^2| \\ &\quad + \left[\mathbf{y}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}_0) - \sum_{i,j} e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i}\right]/\tau^2 - \tfrac{1}{2}\mathbf{u}_0\mathbf{D}^{-1}\mathbf{u}_0,\end{aligned} \tag{8.22}$$

where $\mathbf{W}$ is $diag\{\boldsymbol{\mu}\}/\tau^2$ for this Poisson-like model. Ignoring the fact that the term involving $\mathbf{W}$ depends on the unknown parameters, the remainder of (8.22) is the log likelihood or log quasi-likelihood with the additional "penalty" term of $\mathbf{u}_0'\mathbf{D}^{-1}\mathbf{u}_0$ added on. We can view the penalty function either as arising from a normal distribution for the random effects $\mathbf{u}$ or merely as a penalty function that prevents $\mathbf{u}$ from getting too "big."

This suggests an algorithm as follows:

1. Obtain starting values for $\boldsymbol{\beta}, \tau$, $\mathbf{D}$ and $\mathbf{u}$.

2. Solve for $\mathbf{u}_0$ using (8.21).

3. Use $\mathbf{u}_0$ to get an estimate of $\mathbf{D}$.

4. Maximize (8.22) as a function of $\boldsymbol{\beta}$ and τ for given values of $\mathbf{u}_0$ and $\mathbf{D}$.

5. Iterate to convergence.

The big question is whether, after all these approximations, the method works well? The answer is not when it is needed most. That is, when the conditional distribution of the data given the random effects is approximately normal (i.e., Poisson with mean counts above 7 or so) then the approximations and methods work well. But for distributions far from normal, that is, binary data with observation specific covariates, the method does not work well. Extensive investigations (Breslow and Clayton, 1993; Breslow and Lin, 1995; Lin and Breslow, 1996) show that the method can fail badly for such distributions.

Why does the method fail to work well? The short answer is that there are too many approximations. But here is a more complicated answer that explains why this method works for the normal, linear, mixed model, but not non-normal models

Whenever the marginal density of $\mathbf{Y}$ is formed as a mixture, with separate parameters for $\mathbf{Y}|\mathbf{u}$ and $\mathbf{u}$, we saw that the ML equations could be formed as

$$\text{(8.23)} \qquad \mathrm{E}\left[\frac{\partial \ln f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{U},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\middle| \mathbf{y}\right] = 0,$$

$$\text{(8.24)} \qquad \mathrm{E}\left[\frac{\partial \ln f_{\mathbf{u}}(\mathbf{U}|\mathbf{D})}{\partial \mathbf{D}}\middle| \mathbf{y}\right] = 0.$$

To see how these relate to BLUP and PQL and a different technique called maximum hierarchical likelihood or MHL (see Lee and Nelder, 1996) it is instructive to start with the linear mixed model

$$\text{(8.25)} \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

with the usual normality assumptions. For this model, (8.23) for $\boldsymbol{\beta}$ is

$$\text{(8.26)} \qquad \mathrm{E}[\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\sigma^2] = \mathbf{0},$$

or, since $\tilde{\mathbf{u}} = E[\mathbf{u}|\mathbf{Y}]$,

$$\text{(8.27)} \qquad \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{Z}\tilde{\mathbf{u}} = \mathbf{0},$$

which is the equation for $\boldsymbol{\beta}$ from the mixed model equations, (2.22).

Since $\partial \log f_{\mathbf{Y}|\mathbf{u}}/\partial \mathbf{u} = \partial \log f_{\mathbf{u}|\mathbf{Y}}/\partial \mathbf{u}$, setting this derivative equal to zero and solving gives the mode of $f_{\mathbf{u}|\mathbf{Y}}$. This is the same as the first step in the Laplace approximation: finding the mode of the joint distribution of $\mathbf{Y}$ and $\mathbf{u}$.

In the case of the linear mixed model, since the distribution of $\mathbf{u}$ given $\mathbf{Y}$ is normal, the mode is the mean, which is $\tilde{\mathbf{u}}$. Lee and Nelder (1996, 2001), in their MHL technique, use this approach directly: form the joint distribution $f_{\mathbf{Y},\mathbf{u}}$ and maximize simultaneously with respect to $\mathbf{u}$ and $\boldsymbol{\beta}$. This "joint maximization" method thus finds $\tilde{\mathbf{u}}$ that is needed to solve for the MLE of $\boldsymbol{\beta}$. It therefore works quite well for the normal-normal linear model.

In contrast, for non-normal distributions, the conditional distribution of $\mathbf{u}$ given $\mathbf{Y}$ is not normal and hence maximizing $f_{\mathbf{Y},\mathbf{u}}$ with respect to $\mathbf{u}$ locates the mode which may not be $\tilde{\mathbf{u}}$. Furthermore, the likelihood equation may involve functions more complicated than $\mathbf{u}$ alone.

Hence conditional expected values of other functions of $\mathbf{u}$ are required to calculate the MLE of $\boldsymbol{\beta}$. Engel and Keen (1996), in their discussion of the Lee and Nelder paper, note this discrepancy, but seem to downplay its importance.

a. Higher-order Laplace approximations

In addition to investigating the performance of PQL, Breslow and Lin (1995) and Lin and Breslow (1996) looked at higher-order Laplace approximations, extending (8.18) to an extra third-order term in an attempt to improve the approximation. This worked well for small variance components, but not for larger ones and led to some instability in the estimation procedures. More recently, Raudenbush et al. (2000) have investigated extending the approximation to yet another term (the fourth order) and have obtained promising results. Their investigations are limited to a two-level nested structure and it would be nice to see if the excellent performance holds up in more difficult situations, for example, crossed random factors.

8.5 Choosing good estimating equations

The general theory of estimating equations (EE) offers some guidance to compare various methods of estimation, such as PQL and ML. A frequent place to start is to try to form an *unbiased estimating equation*, that is, to find a function $G(\mathbf{Y}, \boldsymbol{\theta})$ such that

$$\text{(8.28)} \qquad \mathrm{E}[G(\mathbf{Y}, \boldsymbol{\theta})] = \mathbf{0} \qquad \forall \boldsymbol{\theta}.$$

Solutions to $G(\mathbf{Y}, \boldsymbol{\theta}) = 0$ in $\boldsymbol{\theta}$ for given $\mathbf{Y}$ are, under regularity conditions, consistent and asymptotically normal. Familiar examples include

$$\begin{aligned}
\text{ML:}\quad & \frac{\partial \log L}{\partial \boldsymbol{\theta}} = \mathbf{0}, \\
\text{QL:}\quad & \frac{\partial \mu_i}{\partial \theta_j}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}, \\
\text{GEE:}\quad & \sum_k \frac{\partial \mu_{ki}}{\partial \theta_j}\mathbf{V}_k^{-1}(\mathbf{Y}_k - \boldsymbol{\mu}_k) = \mathbf{0}.
\end{aligned}$$

Potential advantages (Heyde, 1997) to approaching the problem from the estimating equations point of view rather than the estimator itself are:

- Automatic invariance to $1-1$ transformations of the parameter.
- Information and information-like quantities are functions of the EE, not the estimator.
- Asymptotic properties of the estimator are typically derived first for the EE and then transferred to the estimator.
- It is easy to combine EEs.
- There is flexibility in the choice of a family of estimating equations (perhaps followed by optimization within that family).

An obvious disadvantage is that it does not directly give properties of the estimator, which is what is used in practice.

Since it is advantageous to consider optimal EEs and since they are invariant to scalar multiplication, it makes sense to standardize them. A common standardization is to modify the basic estimating function G by

$$G^s \equiv -\mathrm{E}[\partial G/\partial \boldsymbol{\theta}]'(\mathrm{E}[GG'])^{-1}G. \tag{8.29}$$

This standardized form has two nice features:

- The large sample variance of $\hat{\boldsymbol{\theta}}$ is $[\mathrm{var}(G^s)]^{-1}$, so we can compare estimators on the basis of the variance of their EEs.

- $\mathrm{var}(G^s) = -\mathrm{E}[\partial G^s/\partial \boldsymbol{\theta}]$, that is, it satisfies a score equation.

a. Illustration: REML

To illustrate the calculation of optimal estimating equations we consider REML estimation for the linear mixed model. Consider the ANOVA linear mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_i \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}, \tag{8.30}$$

where $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_i^2)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, all mutually independent. The basic idea behind REML is to estimate the variances in a way that "takes account of" and is invariant to the choice of values of $\mathbf{X}\boldsymbol{\beta}$.

One way to do this is to restrict attention to residuals after removing the fixed effects, either by using $\mathbf{MY}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ or, more flexibly, by considering $\mathbf{K}'\mathbf{Y}$, where $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and $\mathbf{K}'$ is $(N - \mathrm{rank}[\mathrm{X}]) \times N$ and of full row rank.

It is easy to show that $\mathbf{K}'\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}'\mathbf{VK})$ so the log likelihood is (up to a constant)

$$-\tfrac{1}{2}\log|\mathbf{K}'\mathbf{VK}| - \tfrac{1}{2}\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{VK})^{-1}\mathbf{K}'\mathbf{y}. \tag{8.31}$$

With some tedious algebra (Searle et al., 1992) the equations that result after differentiating (8.31) and equating to zero are

$$\mathrm{tr}(\mathbf{PZ}_i\mathbf{Z}_i') = \mathbf{y}'\mathbf{PZ}_i\mathbf{Z}_i'\mathbf{Py}, \tag{8.32}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}$.

This same result can be derived by alternate means (Heyde, 1997) via optimal EEs with a slight gain in generality. Suppose $\mathbf{Y}$ is distributed with mean $\mathbf{X}\boldsymbol{\beta}$ and with variance $\sum_i \mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 + \mathbf{I}\sigma^2$, with third and fourth moments to match the normal distribution (so not quite assuming normality).

Then, within the family of EEs of the form

$$G_i(\mathbf{Y}, \sigma^2) = \mathbf{Y}'\mathbf{KA}_i\mathbf{K}'\mathbf{Y} - \mathrm{E}[\mathbf{Y}'\mathbf{KA}_i\mathbf{K}'\mathbf{Y}] \tag{8.33}$$

(i.e., $\mathbf{A}_i$ completely arbitrary) the REML EEs are optimal.

This idea, of using the REML EEs for non-normal data, was applied to binary data by Drum and McCullagh (1993). Other derivations of REML rely on marginal likelihood: the REML equations can be derived by integrating $\boldsymbol{\beta}$ out of the likelihood using a flat distribution (Searle et al., 1992). A third approach to REML is via quotient spaces (McCullagh, 1999).

Although each of these different approaches to REML give the same answer for the linear mixed model, they can give different answers for non-normal and nonlinear models and so leaves ambiguous the "proper" definition of REML for those cases.

Yet another justification of REML (Speed, 1991) is to regard REML as equating observed and expected values of a quadratic form in the best linear predicted values. Returning to model (8.30) it is straightforward to show that the best linear unbiased predicted value is given by (Searle et al., 1992)

$$\tilde{\mathbf{u}}_i = \sigma_i^2 \mathbf{Z}_i' \mathbf{P} \mathbf{y}, \tag{8.34}$$

so that the sum of squares of the BLUPs is given by

$$\tilde{\mathbf{u}}_i' \tilde{\mathbf{u}}_i = \sigma_i^4 \mathbf{y}' \mathbf{P} \mathbf{Z}_i \mathbf{Z}_i' \mathbf{P} \mathbf{y}. \tag{8.35}$$

By the usual expected value calculations for quadratic forms it is easy to show that

$$\mathrm{E}[\tilde{\mathbf{u}}_i' \tilde{\mathbf{u}}_i] = \sigma_i^4 \mathrm{tr}(\mathbf{P} \mathbf{Z}_i \mathbf{Z}_i'), \tag{8.36}$$

so that equating the observed and expected sums of squares of the BLUPs [equate (8.36) to (8.35)] is equivalent to the REML equation (8.32).

b. BLUP methods

If we wish to pursue the idea of EEs, we must decide on a family of EEs. One idea is to directly use the family that reproduces REML for the normal linear model, or similarly, quadratic functions of the data. But this leads to intractable calculations and necessitates the same sort of approximations as with PQL. This has been developed in a series of papers by McGilchrist and co-workers (McGilchrist, 1993, 1994; McGilchrist and Yau, 1995) and is essentially the idea behind GEE2 and the discussion in Chapter 10 of McCullagh and Nelder (1989).

I think a more profitable approach is not to directly work with quadratic forms but instead to consider the analog of equating observed and expected predicted values, as was illustrated for REML above. By construction, this gives an unbiased estimating equation.

For a special class of models, this idea has been partially developed in a PhD thesis by Renjun Ma under the direction of Bent Jorgensen at the University of British Columbia. Their class of distributional models is called the Tweedie class and it includes normal, Poisson, gamma and inverse gamma, among others. For this class they consider nested random effects (and arbitrary fixed effects). They actually use EEs based on the BPs to estimate the parameters in the mean structure and take a more ad hoc approach to estimating the variance parameters.

I think there would be profit to looking at classes of EEs based either quadratic forms in the best predictor or the best linear predictor. For example, equations of the form

$$\tilde{\mathbf{u}}' \tilde{\mathbf{u}} = \mathrm{E}[\tilde{\mathbf{u}}' \tilde{\mathbf{u}}]. \tag{8.37}$$

TABLE 8.2.
Parameter estimates and SEs (subscripts) for the Progabide data

	Estimation Method		
Parameter	**MLE**[1]	**PQL**[2]	**GEE**[3]
Intercept	$0.96_{0.08}$	$1.00_{0.14}$	$1.35_{0.16}$
TRT	$-0.07_{0.07}$	$-0.01_{0.19}$	$-0.11_{0.19}$
POST	$0.11_{0.05}$	$0.11_{0.05}$	$0.11_{0.12}$
POST×TRT	$-0.30_{0.07}$	$-0.30_{0.07}$	$-0.30_{0.17}$
σ_s^2	$\hat{\sigma}_s^2 = 0.52_{0.10}$	$\hat{\sigma}_s^2 = 0.53_{0.10}$	$\hat{\rho} = 0.60$

[1] SAS GENMOD.
[2] SAS NLMIXED.
[3] Diggle et al. (1994).

Of course, calculation of $\tilde{\mathbf{u}}$ is often problematic so it might be advantageous to work only with the best linear predictor:

$$\tilde{\mathbf{u}}_{BLP} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \tag{8.38}$$

where $\mathbf{C}$ is the covariance between $\mathbf{u}$ and $\mathbf{Y}$ and $\mathbf{V}$=Var($\mathbf{Y}$). Even this does not exist in closed form but involves at most a two-dimensional integral.

c. Composite and working likelihoods

Another suggestion involves only using the distribution (and actually only their variances and covariances) of the Y_i taken two at a time (in $\mathbf{V}$) or the Y_i and the u_j taken in pairs in $\mathbf{C}$.

The ease with which EEs can be combined (e.g., weighted average) suggests that likelihoods for only portions of the data or simpler, "working" likelihoods might be a fertile ground for finding relatively efficient EEs. A common example of this is the use of GEEs with a working independence variance structure.

Working with the true likelihood of only two observations at a time gives tractable computations and combining them into a composite likelihood can be relatively efficient (e.g., Heagerty and Lele, 1998).

8.6 Progabide and seizures revisited

I now return to the Progabide and seizures dataset previously discussed in Section 5.3 to contrast the estimation methods. Table 8.2 displays the estimates using ML, PQL and GEE. The maximum likelihood and penalized quasi-likelihood estimates are quite close, undoubtedly because the average counts were around 7, giving approximately normally distributed data, for which PQL would be expected to perform like ML. The estimates for GEE are similar except for the intercept. This similarity is perhaps surprising since GEE is fitting a marginal model, while the ML and PQL approaches fit a conditional model. For example, with the probit model we saw attenuation of all the coefficients in (1.8). However, with a log link and random intercepts, the only coefficient that will differ between the marginal and conditional approaches is the intercept. The intercept would be expected to be larger by $\sigma_s^2/2$, which makes them much more comparable, though still somewhat different.

What *does* appear to be different are the standard errors, in particular for the coefficient of primary interest, that associated with the pre- versus post-baseline by treatment interaction. ML and PQL give the standard error as 0.07, while GEE gives 0.17. Two logical explanations present themselves. First, GEE is known to estimate standard errors with higher variability (Kauermann and Carroll, 2001) so perhaps it is just an inaccurate standard error estimate. Second, the robust estimation in GEE may be picking up a model specification error since its variance estimates are less affected by specification errors.

In this case it appears to be the latter. Some further analysis using SAS PROC NLMIXED indicates that there is a random POST effect. Inclusion of this extra source of randomness yields the following estimates and standard errors: Intercept = $1.07_{0.13}$, POST × TRT $= -.35_{0.15}, \sigma_s^2 = 0.45, \sigma_p^2 = 0.22, \sigma_{sp} = 0.02$, where σ_p^2 is the additional subject-specific variance component in the post-baseline period and σ_{sp} is the covariance between it and the subject intercepts. Now the standard error for POST × TRT is much more comparable and the intercept is larger by $\frac{1}{2}0.45 = 0.23$ in the pre-baseline group and $\frac{1}{2}(0.45 + 2(0.02) + 0.22) = 0.35$ in the post-baseline group, or, averaging the two, about 0.29 larger. The difference between the GEE estimated intercept and the intercept in this new ML fit is 0.28, quite close to that predicted by the conditional model.

8.7 Further notes

Non- and semi-parametric regression methods and accommodation of measurement error for generalized linear models with correlated data are developed in Carroll et al. (1997); Wang et al. (1998); Lin and Carroll (1999); Wang et al. (1999); Lin and Carroll (2000, 2001) and Welsh et al. (2002). Heagerty (1999) develops an interesting melding of the marginal and conditional approaches. Jiang (1998) uses an alternate approach, "simulated moments" to fit GLMMs.