# An Introduction to Generalized Linear Mixed Models

by
Charles E. McCulloch
Biometrics Unit and Statistics Center
Cornell University

BU-1340-MA

January, 1997

# AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS

Charles E. McCulloch
Biometrics Unit and Statistics Center
Cornell University
Ithaca, New York 14853

## ABSTRACT

The generalized linear mixed model (GLMM) generalizes the standard linear model in three ways: accommodation of non-normally distributed responses, specification of a possibly non-linear link between the mean of the response and the predictors, and allowance for some forms of correlation in the data. As such, GLMMs have broad utility and are of great practical importance. Two special cases of the GLMM are the linear mixed model (LMM) and the generalized linear model (GLM). Despite the utility of such models, their use has been limited due to the lack of reliable, well-tested estimation and testing methods. I first describe and give examples of GLMMs and then discuss methods of estimation including maximum likelihood, generalized estimating equations, and penalized quasi-likelihood. Finally I briefly survey current research efforts in GLMMs.

Keywords and Phrases: Non-normal data, nonlinear models, EM algorithm, Newton-Raphson, maximum likelihood, generalized estimating equations and penalized quasi-likelihood.

## 1. Introduction

Generalized linear mixed models (GLMMs) are a natural outgrowth of both linear mixed models and generalized linear models. GLMMs can be developed for non-normally distributed responses, will allow nonlinear links between the mean of the response and the predictors, and can model overdispersion and correlation by incorporating random effects. As such, they are of wide utility (e.g., Breslow and Clayton, 1993).

While maximum likelihood and variants (e.g. restricted maximum likelihood or REML) are standard for both linear mixed models and generalized linear models (e.g., logistic regression), its use in GLMMs has been limited to simple models due to the need to numerically evaluate high dimensional integrals.

In this paper we first motivate and give examples of the GLMM approach and relate it to both generalized linear models (GLMs) and linear mixed models (LMMs). In Section 4 we then turn to the more technical issues of estimation and testing. Section 5 offers a brief review of the research literature and Section 6 gives conclusions.

## 2. Generalized linear mixed models

### *2.1 An example*

We begin by considering an example patterned after Stiratelli, Laird and Ware (1984) on the effect of air pollution on asthma attacks. Suppose that 200 schoolchildren are followed for a number of days. On each day we record a response: Asthma attack (yes/no) and several predictors: total suspended particulates (TSP - a measure of air pollution), sex of child (SEX), and history of hay fever (HAYF - measured as yes/no). If we were conducting a thorough data analysis it would also make sense to consider other predictors such as temperature, humidity, age of child, whether the mother or father is a smoker, day of week, and perhaps whether the child had an asthma attack on a previous day. However, for ease of exposition of the modelling issues we will restrict ourselves to the above three. We will focus on two questions of interest. Does polluted air increase the risk of an asthma attack? Are some children more sensitive than others to air pollution and, if so, which ones?

If we consider which features of this problem are relevant from a statistical modelling viewpoint, an immediate realization is that the data are binary and hence Bernoulli distributed. We also need to decide how to relate the response to the three predictors. A common approach is to define p=Pr{asthma attack} and model the log odds as a linear function of the predictors:

$$\ln(p/(1-p)) = \mu + \beta_1 SEX + \beta_2 HAYF + \gamma TSP. \qquad (1)$$

A typical way to fit such a model is via ordinary logistic regression. However, this approach has two serious drawbacks. First, since the data are gathered repeatedly on the same children, they are likely to be correlated and the model does not accommodate correlated data. Viewed another way, some children will undoubtedly be more likely to have asthma attacks and there is nothing in the model which reflects this fact. Secondly, the model is incapable of answering our second question which concerns the possibility of sensitive individuals.

Assuming independence when the data are, in fact, highly correlated is well-known to cause dramatically incorrect results (Cox and Snell, 1989, p.107). Estimates are usually little affected but standard errors, tests and confidence intervals are usually far from correct. Before discussing how to address such problems, we first introduce generalized linear models.

### *2.2 Generalized Linear Models*

A basic precept of generalized linear models (GLMs) is to dissect the modelling process into three distinct components or questions to be answered:

1. What is the distribution of the data?
2. What aspect of the problem will be modelled?
3. What are the predictors?

For our example, the data are binary so the distribution has to be Bernoulli. As mentioned above, a typical approach to question 2. is to model the log odds, ln(p/(1-p)), as a linear function of the predictors. Finally, for the third part, we have decided to use the predictors TSP, SEX, and HAYF.

Table 1 shows a more general prescription of the structure of generalized linear models and illustrates the specific case of simple linear logistic regression. We decide on a distribution for the data (often from the exponential family and perhaps after transformation). The mean of that distribution is then modelled by selecting a link function (typically specified, not estimated from the data) and assuming that that function applied to the mean is a linear function of the predictors.

Estimation of the parameters for GLMs is typically accomplished by calculating maximum likelihood or maximum quasi-likelihood estimates. Hypothesis tests can then be based on analysis of deviance, where deviance is defined as = 2(max possible loglikelihood - loglikelihood of fitted model). Differences in the deviance of two models then give the likelihood ratio statistic for comparing the models: difference in deviance for models 1 and 2 = 2(loglik model 2 - loglik model 1). For details see McCullagh and Nelder (1989).

### *2.3 Generalized Linear Mixed Models*

We now return to the asthma example of Section 2.1. Write $Y_{ij}$ for the response for child i at time j, where $Y_{ij}$ equals 1 for an asthma attack and is zero otherwise. Then $Y_{ij} \sim \text{Bernoulli}(p_{ij})$, where $p_{ij}$ is the probability of an attack for child i at time j. Next assume that

$$\ln(p_{ij}/(1-p_{ij})) = \mu_i + \beta_1\text{SEX} + \beta_2\text{HAYF} + \gamma\text{TSP}, \qquad (2)$$

where $\mu_i \sim \text{Normal}(0,\tau_\mu)$.

The main difference here is that we have *assumed a distribution for $\mu_i$*. This induces a correlation between the logits of the probabilities of response for the ith child on occasions j and k: $\text{cov}(\ln(p_{ij}/(1-p_{ij})), \ln(p_{ik}/(1-p_{ik})) = \tau_\mu$. If the $p_{ij}$ are correlated then the data are likewise correlated for observations taken on the same child.

Alternatively or additionally we could assume a distribution for the parameter $\gamma$ by using the model:

$$\ln(p_{ij}/(1-p_{ij})) = \mu_i + \beta_1\text{SEX} + \beta_2\text{HAYF} + \gamma_i\text{TSP},$$

with $\gamma_i \sim \text{Normal}(\gamma, \tau_\gamma)$.

From model (2) we can see that $\gamma$ is the air pollution effect which is assumed constant across children. More precisely, it is the increase in the log odds of an asthma attack associated with an increase of one unit in TSP. Hence $\gamma_i$ is the air pollution effect for the *ith* child. We are now able to answer the second question of Section 2.1 in the following way. The hypothesis $\tau_\gamma > 0$ is equivalent to children having differential sensitivities to pollution (TSP). Furthermore, if we can predict the values of $\gamma_i$ then we can identify the sensitive individuals. That is, children with the highest values of $\gamma_i$ are the ones who are most sensitive to TSP.

Finally we can consider assuming a distribution on $\beta_2$:

$$\ln(p_{ij}/(1-p_{ij})) = \mu_i + \beta_1 SEX + \beta_{2i} HAYF + \gamma TSP,$$

$$\text{with } \beta_{2i} \sim \text{Normal}(\beta_2, \tau_\beta).$$

What effect does this have? If HAYF is coded 1 for yes and 0 for no, then for the non-hayfever group, the contribution of the $\beta_{2i}$HAYF term is zero, while for the hayfever group it is $\beta_{2i}$. If $\tau_\beta > 0$, then the hayfever group will have a larger variance.

We can see that the simple device of assuming a distribution on a parameter is capable of modelling correlation in the data, identifying differential sensitivity and predicting the most sensitive individuals, and modelling unequal variances. This is now a more adequate model for inference in the asthma example.

The steps on specifying a GLMM are almost the same as for a GLM. We must consider:

1. What is the distribution of the data?
2. What aspects will be modelled?
3. What are the factors?
4. Which factors will be assumed to have a distribution?

Table 2 illustrates the structure of GLMMs. The fourth decision in the list is the only new one but should be familiar from usage of LMMs. That is, which factors will be assumed to have a distribution and be declared random and which will be declared fixed?

## *2.4 Fixed versus Random Factors*

It has long been suggested (Eisenhart, 1947; Scheffe, 1959) that two main assumptions can be made about the parameters describing the parameters in a linear model. They can be assumed to be fixed, unknown constants or to them can be attributed a distribution. This is well accepted.

However, conventional wisdom holds that a factor be treated as fixed if one is interested in drawing inferences about the specific levels included in the experiment (Searle, 1987, p.4; Snedecor and Cochran, 1989, p.320) or if, in repeated selection of the levels of that factor, the same levels are selected (Ott, 1984, p.638; Snedecor and Cochran, 1989, p.320). If inferences focus on the population from which the parameters are selected or if, on repeated selections of the parameters, the same levels are not used, then the factor is declared to be a random factor.

I argue that for both linear and generalized linear mixed models these criteria are incorrect. In practice we need to divorce the fixed versus random distinction from the scope of the inferences and instead base the decision on a criterion more closely related to the assumption of a distribution for the parameters. To make this point consider the idea of best prediction of the value of the level of a random effect. We are in this case willing to assume that the parameters follow a distribution (it is a random factor), but by calculating best predicted values we are making inferences about (and are "interested in") the specific levels included in the experiment.

In arguing that the conventional criteria are incorrect I consider two generic examples: a randomized blocks design and prediction of sire effects in animal breeding. I first consider the randomized blocks design and the criterion of whether or not we would get the same levels (blocks) of the random factor if the experiment were conducted again. In many experiments the following facts are all true:

a) The same blocks would be used if the experiment were repeated (which blocks are used is often determined by the availability of experimental material),
b) The investigator wants to draw inferences beyond the blocks on hand in the experiment,
c) The investigator is willing to make inferences to a population of blocks similar to the ones in the experiment, and
d) It is reasonable to assume the blocks in the experiment are an i.i.d.sample from the population described in c).

Points b), c), and d) mean that, by definition, blocks are a random factor. But a) argues that blocks should be declared fixed. Essentially the criterion fails because we do not have a physical sampling scheme which guarantees random sampling, but we are willing to assume the blocks form an i.i.d. sample from some distribution.

Next consider the second criterion: interest in the levels actually included in the experiment. The primary example is prediction of sire effects in animal breeding, but there are parallels in spatial prediction (kriging). Animal breeders often face the following problem. They wish to improve the genetic value of animals (e.g., the ability of cows to produce protein in milk) by selective breeding of the population. The data used for the analysis often includes the daughters of all the sires whose data are available through a registry. The goal is to estimate the ability of a specific sire to produce genetically superior offspring. On one hand it is easy to envision the sires included in the analysis as coming from a population (actual or conceptual) of sires; hence the argument for treating the effect as random. On the other hand, interest focusses specifically on the sires to be included in the analysis. Those are the only ones which could be considered for use in a breeding program. This, according to the second criterion, would argue for treating sire effects as fixed.

How does one reconcile assuming a distribution for parameters, but still being interested in them? This is now straightforward using the ideas of best prediction or best linear unbiased prediction (BLUP) ("prediction" rather than "estimation" since it is a random variable) as detailed in Robinson (1991) and Searle, Casella and McCulloch (1992) and implemented in software such as SAS PROC MIXED. With apologies to T.S. Geisel, I would summarize as follows:

## Up with BLUP

No ma'am, No ma'am. No one
knows ma'am. Whether FIXED
or whether RANDOM.

Should we this one, or do that one?
I.I.D. says that they're RANDOM.

But if we must say that's nixed,
then we'll say that they are FIXED.

What of those who won't predict?
To them I say interdict.
Up with BLUP,
BLUP's for you.
That's for when you're interested too.

### 3. Other Examples

To illustrate the versatility of GLMMs, I would like to briefly describe two other examples. The first involves Potomac River Fever (equine monocytic ehrlichiosis) in horses and is more carefully described in Atwill, et al (1996). Potomac River Fever is a blood-borne rickettsial disease whose transmission mechanism is unknown. Both arthropod (e.g. blackfly) and direct oral transmission have been suspected but not verified. Identification of risk factors of horses in New York State might give clues to the spread of this disease and help with reducing its frequency. The study involved 511 farms each with several social groups of horses, for a total of 2,587 horses. The response variable was seropositivity (yes/no) for the disease. Again, because the data are binary the distribution must be Bernoulli. We used the standard link for binary data, the logit link. A number of fixed predictors were used, examples of which are: frequency stall cleaned, frequency fly spray applied, breed, sex, etc. Two random factors were also used: farm and social group nested within farm. So there were a number of fixed factors and two nested random factors. If we let $Y_{ijk}$ denote seropositivity for horse k in social group j on farm i then the model was given by

$$Y_{ijk} \sim \text{Bernoulli}(p_{ijk})$$

$$\text{logit}(p_{ijk}) = \mu + s_{j(i)} + f_i + \text{fixed effects},$$

$$s_{j(i)} \sim \text{i.i.d. } N(0, \sigma^2_{group(farm)}),$$

$$f_i \sim \text{i.i.d. } N(0, \sigma^2_{farm}),$$

where the $s_{j(i)}$ denote the social group effects and the $f_i$ represent the farm effects. We focus on inferences for the random factors. The estimated variances of the random effects were:

$$\hat{\sigma}^2_{farm} = 1.26$$

$$\hat{\sigma}^2_{group(farm)} = 0$$

So the difference in loglikelihood for testing $\sigma^2_{group(farm)} = 0$ is zero and hence not statistically significant when compared to a $\frac{1}{2}\chi^2_1$ (see Section 4). On the other hand, the farm variance component is statistically significant. This has the following implications. There is a significant correlation among horses within a farm on the logit scale (0.32), but no correlation within social groups. This suggests that the disease is not transmitted directly from horse to horse, but instead is related to environmental or management factors operating at a farm scale.

Another example in which GLMMs could be used is in analyzing data from the Breeding Bird Survey (Peterjohn, 1994). Counts of number of birds "sighted" has been made each June at thousands of locations across the U.S. and Canada. Many of the locations have been surveyed since the mid 1960s. Responses are a count of the number of birds of each species at each location. A possible distribution to try for such data would be Poisson. Fixed factors would include time (in order to gauge trends in population sizes) and possibly observer effects (Sauer, Peterjohn and Link, 1994) and location could serve as a random factor. This would serve to incorporate correlations for data taken repeatedly at the same location.

## 4. Inference for GLMMs

### *4.1 Maximum Likelihood Estimation*

Since maximum likelihood estimation is used for both LMMs and GLMs it is a logical place to start for estimation in GLMMs. To fix ideas, consider a very simple GLMM, a logit-normal model:

$$Y_{ij} \mid u \sim \text{Bernoulli}(p_{ij}),\ i=1,2,\ldots n;\ j=1,2,\ldots q.$$

$$\ln(p_{ij}/(1-p_{ij})) = \beta x_{ij} + u_j$$

$$u_j \sim \text{Normal}(0,\sigma^2).$$

This model has n observations in each of q clusters, within which the data are correlated. It uses a logit link and has one fixed and one random factor. The likelihood for this model would be calculated as follows

$$
\begin{aligned}
\text{likelihood} &= \mathrm{P}\{\mathbf{Y}=\mathbf{y}|\beta,\sigma^2\} \\
&= \int \mathrm{P}\{\mathbf{Y}=\mathbf{y}|\beta,\sigma^2,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \int \mathrm{P}\{\mathbf{Y}=\mathbf{y}|\beta,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \int \prod_{i,j} \mathrm{P}\{Y_{ij}=y_{ij}|\beta,\mathbf{u}\}f(\mathbf{u}|\sigma^2)d\mathbf{u} \\
&= \prod_j \int \prod_i \mathrm{P}\{Y_{ij}=y_{ij}|\beta,\mathbf{u}\}f(u_j|\sigma^2)du_j \\
&= \prod_j \int \exp\{\beta\Sigma_i y_{ij}x_{ij}+y_{+j}u_j\}\Pi_i(1+\exp\{\beta x_{ij}+u_j\})^{-1}\times \\
&\qquad\qquad \exp\{-u_j^2/2\sigma^2\}/(2\pi\sigma^2)^{1/2}du_j
\end{aligned}
$$

This cannot be evaluated in closed form, and, in general, the calculation of the likelihood can be quite difficult. For the general case,

$$
\text{likelihood} = \underbrace{\int\int\ldots\int}_{\text{dim of } \mathbf{u}} \exp(\Sigma_i \mathrm{Y}_i(\mathbf{x}_i'\beta+\mathbf{z}_i'\mathbf{u}))\Pi_i(1+\exp(\mathbf{x}_i'\beta+\mathbf{z}_i'\mathbf{u}))^{-1}d\mathrm{F}(\mathbf{u}).
$$

Unfortunately, the dimension of **u** and hence the order of integration can get large quickly. For example, in the salamander data set from McCullagh and Nelder (1989) with two crossed random factors (males and females) each with 6 levels, the above is a 12 dimensional integral. This makes numerical evaluation of the integral problematic.

What then to do for ML estimation? For simple problems we can use numerical integration. When the model has a single random effect or two nested random effects, it is relatively easy to evaluate the integrals in the likelihood. For example, with a single random factor we have seen that the likelihood is a product of one-dimensional integrals. One can then maximize the likelihood numerically to find ML estimates and to perform likelihood ratio tests.

To evaluate the likelihood numerically, with a single, normally distributed random effect, the likelihood can be written as a product of integrals of the form: $\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx$. These can be accurately evaluated using Gauss-Hermite quadrature:

$$
\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx \approx \sum w_i g(x_i)
$$

The weights, $w_i$, and the evaluation points, $x_i$, can be found in books describing numerical integration, e.g., Abramowitz and Stegun (1964). There are other approaches to ML estimation (see Section 5) for more complicated models, but the computations are much more difficult.

If one can calculate the ML estimates and the maximized value of the likelihood, then likelihood ratio tests are a possibility. Inference using ML would then proceed using the usual asymptotic approximations: ML estimates are asymptotically normal, with SEs coming from second derivatives of the log likelihood. Tests would be based on the likelihood ratio test, comparing -2loglikelihood for nested models. Best predicted values would be estimated by calculating E[random effect|data] and plugging in ML or REML estimates. In general, the conditional expected values cannot be evaluated in closed form just as the likelihood cannot.

One point bears emphasis in using the likelihood ratio test for variance components. A common hypothesis of interest is whether a variance component is zero. This hypothesis lies on the boundary of the parameter space and the usual asymptotic theory breaks down. The intuition is seen easily by considering ANOVA estimators in a one-way random effects layout for a linear mixed model. When the variance of the random effect is zero and the sample sizes are large, the ANOVA estimator is negative about half the time. The ML estimator cannot be negative and so it is zero about half the time. The likelihood ratio test statistic which is formed as

$$-2\log\Lambda = -2(\log L(\sigma^2{=}0) - \log L(\sigma^2{=}\hat{\sigma}^2))$$

would be zero about half the time. The likelihood ratio theory breaks down because the estimate gets "stuck" on the boundary. The actual large-sample distribution under $H_o$: $\sigma^2=0$ is a 50:50 mixture of a $\chi^2_1$ and 0. Operationally, we would calculate the p-value under $\chi^2_1$ and cut the p-value in half, creating, in essence, a one-tailed test. See Stram and Lee (1994) and Self and Liang (1987) for proofs and further details.

In summary, ML estimation for GLMMs has known large sample properties and likelihood ratio tests can be based on them. Unfortunately, estimates are hard to compute for many GLMMs and their small sample performance needs to be assessed for any particular model.

### *4.2 Generalized Estimating Equations*

The computational difficulty of ML estimation has made approaches based on general estimating equations (GEEs) attractive. GEEs are a computationally less demanding method than ML estimation. They are applicable (mainly) to longitudinal data, where we define longitudinal data as data collected on a subject on two or more occasions with the number of occasions being small compared to the number of subjects.

To set the basic ideas, we first consider a longitudinal data modelling approach using a linear mixed model. It proceeds in three steps:

(1) Separate effects which are constant across subjects ($\beta$) from those which vary across subjects ($\mathbf{u}_j$).

(2) For the data of the *jth* subject, $\mathbf{Y}_j$, write a linear model conditional on the value of $\mathbf{u}_j$:

$$\mathbf{Y}_j = \mathbf{X}_j\beta + \mathbf{Z}_j\mathbf{u}_j + \varepsilon_j,$$

$$\varepsilon_j \sim N(\mathbf{0},\mathbf{R}_j)$$

(3) Incorporate subject-to-subject variability by assigning a distribution to $\mathbf{u}_j$:

$$\mathbf{u}_j \sim N(\mathbf{0},\mathbf{D}).$$

The resulting distribution is $\mathbf{Y}_j \sim \text{indep } N(\mathbf{X}_j\beta,\ \mathbf{V}_j = \mathbf{Z}_j\mathbf{D}\mathbf{Z}_j' + \mathbf{R}_j)$.

An example of this is given in Diggle, Liang and Zeger (1994). Milk was collected from 79 cows on one of three diets: barley, lupins, and a mixture of both. Protein content of the milk was recorded weekly for 19 weeks after the earliest calving. Effects which are constant are diet and time and those which vary across subjects (animals) are the intercepts. That gives the model for the *jth* cow on diet i, at time t as

$$Y_{ijt} = \mu_i + a_j + f(t) + e_{ijt}$$

$$\mathbf{a} \sim N(\mathbf{0},\mathbf{I}\sigma^2) \quad e_{ij} \sim N(\mathbf{0},\mathbf{R}_j)$$

$$\mathbf{R}_j\text{: } \operatorname{corr}(e_{ijt}, e_{ijt'}) = \exp(-\phi|t-t'|),$$

where f(t) is a nonlinear function of time. This model incorporates both random effects for the animals ($a_j$) and a residual correlation governed by the parameter $\phi$.

What would the consequences be of using ordinary least squares (OLS) to estimate such a model? If we write

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{D}),\ \mathbf{e} \sim N(\mathbf{0}, \mathbf{R}),$$

so that $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V}=\mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R})$, then $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is well known that $\hat{\beta}_{OLS}$ is unbiased:

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

It is also well known (Diggle, Liang and Zeger, 1994) that $\hat{\beta}_{OLS}$ usually has high efficiency. In fact, with balanced designs, $\hat{\beta}_{OLS} = \hat{\beta}_{GLS}$, where $\hat{\beta}_{GLS}$ is the, fully efficient, generalized least squares estimator. What, then, is wrong with using $\hat{\beta}_{OLS}$ and standard software? $\operatorname{Var}(\hat{\beta}_{OLS})$ is actually $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ but, using standard software, it is estimated to be $(\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2$, which will often be very wrong. That is, the OLS estimate is not so bad, but the usual variance estimate is way off.

The basic idea behind GEEs is, with $\mathbf{Y}_j \sim$ independently, to use the "replication" across subjects to get an empirical (or so-called "robust") estimate of the variance. For the longitudinal data setting,

$$\hat{\beta}_{OLS} = (\sum_j \mathbf{X}_j'\mathbf{X}_j)^{-1}(\sum_j \mathbf{X}_j'\mathbf{Y}) \quad \text{and}$$

$$\mathrm{Var}(\hat{\beta}_{OLS}) = (\sum_j \mathbf{X}_j'\mathbf{X}_j)^{-1}(\sum_j \mathbf{X}_j'\mathbf{V}_j\mathbf{X}_j)(\sum_j \mathbf{X}_j'\mathbf{X}_j)^{-1}$$

which can be estimated by $(\sum_j \mathbf{X}_j'\mathbf{X}_j)^{-1}(\sum_j \mathbf{X}_j'(\mathbf{Y}_j - \hat{\mu}_j)(\mathbf{Y}_j - \hat{\mu}_j)'\mathbf{X}_j)(\sum_j \mathbf{X}_j'\mathbf{X}_j)^{-1}$. Intuitively, for the milk protein data from Diggle, Liang and Zeger (1994), if all the animals had all 19 weeks of data we could just get empirical estimates of the correlation within animal from the multivariate observations. With some missing data the above formula can be used.

GEEs work most easily for models specified on the unconditional distribution. In contrast, we have been specifying models which are conditional on the random effects, $\mathbf{u}$. An unconditional or marginal specification for binary data would be as follows:

$$E[Y_{ij}] = p_{ij}$$

$$\mathrm{logit}(p_{ij}) = \mathbf{x}_{ij}'\beta .$$

This hypothesizes a logistic relationship on the marginal distribution of $Y_{ij}$ rather than the conditional distribution. For these marginal distributions, we obtain $\hat{\beta}$ by solving the GEE:

$$\sum_{j=1}^{n} \left(\frac{\partial \mathbf{p}_j}{\partial \beta}\right) Var(\mathbf{Y}_j)^{-1}(\mathbf{Y}_j - \mathbf{p}_j) = 0$$

This has properties similar to the equations for the LMM:

$$\sum_j \mathbf{X}_j'\mathbf{V}_j^{-1}(\mathbf{Y}_j - \mathbf{X}_j\beta) = \mathbf{0} .$$

Again, the basic idea is to use the robust variance estimates (robust because they are not model dependent) to get proper estimates of $\mathrm{Var}(\hat{\beta})$. For the conditionally specified random effects models we have been using, one has to work a bit harder (see Zeger, Liang and Albert, 1988).

In summary, GEEs are mainly for longitudinal data and are easiest for marginal models, not random effects models. Primary advantages are that they are computationally easier than ML estimators and use robust standard errors. They work best when the number of time points is relatively small compared to the number of subjects and when the data consist mainly of essentially complete vectors. Otherwise a parametric modelling approach may be more attractive (Diggle, Liang and Zeger, 1994).

### *4.3 Penalized quasi-likelihood*

Another approach which has been suggested is that of penalized quasi-likelihood (Breslow and Clayton, 1994). The same approach has been arrived at by using Laplace approximations (Wolfinger, 1994) and the "joint-maximization" point of view (Gilmour, Anderson and Rae, 1984; Schall, 1991). It can be derived by the following argument. Let

$$\mathrm{Y} \sim \text{exponential family with mean } \mu$$

$$\mathrm{g}(\mu) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \qquad \mathbf{u} \sim \mathrm{N}(\mathbf{0},\mathbf{D}).$$

First we approximate g by linearization:

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

Next we treat **z** as a LMM with

$$\mathrm{Var}(\mathbf{z}) = \mathbf{ZDZ}' + \mathbf{R}(\mathrm{g}'(\mu))^2$$

The basic idea is then to use the Mixed Model Equations (Searle, Casella, and McCulloch, 1992) to iteratively to find both $\hat{\beta}$ and the BLUP of **u**. Schall (1991) also suggests ways to get approximate standard errors.

This approach has several advantages. It is computationally fairly easy and it works well when the data are approximately normal to start with. Unfortunately it does not work well (Breslow and Lin, 1995; McCulloch, 1997) for highly non-normal data (e.g. binary data). It is also tied to random effects distributions which are normal (McCulloch, 1997).

### *4.4 Other approaches*

Other approaches are to derive models for specific situations. Examples can be found in Crowder (1978) for the beta-binomial model, in Abu-Libdeh, et al (1990) for the Poisson-gamma, and in Conaway (1990). Conditional approaches have been used as exemplified in Conaway (1989) and Cox and Snell (1989). Marginal models have been explored in Liang, Zeger and Qaqish (1992).

## 5. Some Current Research Topics

The research literature for GLMMs is growing quickly. Because maximum likelihood estimation is computationally difficult, a number of authors have tried simulation based approaches to ML estimation. McCulloch (1994, 1997) uses a Gibbs sampler to find ML estimates in a probit-normal model and a Metropolis algorithm to suggest Monte Carlo EM and Monte Carlo Newton-Raphson approaches to calculating ML estimates for general GLMMs. Alternate approaches to ML estimation include those of Geyer (1994), Geyer and Thompson (1992), and Casella and Berger (1995). The econometrics literature also contains simulation based approaches to ML estimation, e.g., Borsch-Supan and Hajivassiliou (1993).

Examples of the penalized quasi-likelihood approaches can be found in Gilmour, Anderson and Rae (1984), Schall (1991), Breslow and Clayton (1994), Breslow and Lin (1995), Lin and Breslow (1997), and Wolfinger (1994). Another philosophy is Bayesian estimation which is natural since it does not distinguish fixed and random factors and so the "mixed" model presents no extra complications. Bayes techniques have been propounded in Gilks, et al (1993), and Zeger and Karim (1991) among others, but Natarajan and McCulloch (1995) caution against using flat priors for variance components.

Generalized estimating equations have rightly become popular. The original articles are Zeger and Liang (1986), and Liang and Zeger (1986). Recently, some authors have found situations where GEEs may not be very efficient (Fitzmaurice, 1995; Lipsitz, et al, 1994), so care needs to be taken in their use.

Other research has tried to directly look at analogs of the Mixed Model Equations, e.g., Engel and Keen (1994), and McGilchrist (1994) and Kuk (1995) offers a general simulation based approach for improving estimators.

## 5. Summary

GLMMs are very versatile in that they can handle non-normal data, nonlinear models, and a random effects covariance structure. They can be used to incorporate correlations in models, model the correlation structure, identify sensitive subjects and can be used to handle heterogeneous variances. The modelling process is relatively straightforward, requiring the following decisions:

1. What is the distribution of the data?
2. What is to be modelled?
3. What are the factors? and
4. Are the factors fixed or random?

This all makes GLMMs attractive for use in modelling.

Unfortunately, computing methods for much of the class of GLMMs is an area of active research. Advances are being made in ML estimation, PQL, GEEs and Bayes methods. No general purpose software exists and tests and confidence intervals are asymptotic and approximate. Nevertheless, I foresee heavy usage of GLMMs in the future as the computational issues are resolved and the validity of tests is established and/or improved.

## REFERENCES

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions. National Bureau of Standards, Washington, D.C.

Abu-Libdeh, H., Turnbull, B. and Clark, L.C. (1990) Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. *Biometrics* **46**: 1017-1034.

Atwill, E.R., H.O. Mohammed, J.W. Lopez, C.E. McCulloch, and E.J. Dubovi. Cross-sectional evaluation of environmental, host, and management factors associated with the risk of seropositivity to *Ehrlichia risticii* in horses of New York State. *American Journal of Veterinary Research*, 57: 278-285, 1996.

Borsch-Supan, A. and Hajivassiliou, V. (1993). Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variables. *Journal of Econometrics* **58**: 347-368.

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. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**: 81-91.

Casella, G.C. and Berger, R.L. (1994). Estimation with selected binomial information, or Do you really believe Dave Winfield is batting .471? *Journal of the American Statistical Association* **89**: 1080-1090.

Conaway, M.R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association* **89**: 53-62.

Conaway, M.R. (1990). A random effects model for binary data. *Biometrics* **46**: 317-328.

Cox, D.R. and Snell E.J. (1989). Analysis of Binary Data, 2*nd* Edition. Chapman and Hall, London.

Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27**: 34-37.

Diggle, P., Liang, K.-Y., and Zeger, S.L. (1994). Longitudinal Data Analysis. Oxford University Press, Oxford.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3**: 1-21.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica* **48**:1-22.

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

Geyer, C.J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B,* **54**: 657-699.

Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics* **49**: 441-453.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1984). The analysis of binary data by a generalized linear mixed model. *Biometrika* **72**: 593-599.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society, Series B* **57**:395-407.

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

Liang, K.-Y., Zeger, S.L., and Qaqish, B.H. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B,* **54**: 673-687.

Lin, X. and Breslow, N.E. (1997). Bias correction in generalized linear mixed models with a multiple components of dispersion. *Journal of the American Statistical Association* **91**: 1007-1016.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J., and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics* **50**: 270-278.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, 2nd Ed. Chapman and Hall, London.

McCulloch, C.E. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association* **89**: 330-335.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. To appear in *Journal of the American Statistical Association.*

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society, Series B* **56**: 61-69.

Natarajan, R. and McCulloch, C.E. (1995). A note on existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**:639-643.

Ott, L. (1984). An introduction to statistical methods and data analysis, 2*nd* Edition. Duxbury, Boston.

Peterjohn, B.G. (1994). The North American Breeding Bird Survey. Birding **26**: 386-398.

Robinson, G.K. (1991). That BLUP is a good thing - the estimation of random effects. *Statistical Sciences* **6**:15-51.

Sauer, J.R., Peterjohn, B.G., and Link, W.A. (1994). Observer differences in the North American Breeding Bird Survey. Auk **111**: 50-62.

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

Scheffe, H. (1959). *The Analysis of Variance*. Wiley, New York.

Searle, S.R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York.

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

Self and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**: 605-610.

Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods, 8*th* Edition. Iowa State University Press, Ames, Iowa.

Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**:1171-1177.

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

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

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

Zeger, S.L., Liang, K.-Y., Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**: 1049-1060.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**: 79-86.

Table 1: The structure of generalized linear models.

| General case | Simple linear logistic regression |
|---|---|
| Y ~ distribution | Y ~ Bernoulli |
| $\mu$ = mean of Y | p = mean of Y |
| $g(\mu) = X\beta$ | $\ln(p/(1-p)) = \alpha + \beta x$ |
| link function $g(\cdot)$ | logit link |
| covariates $X\beta$ | one predictor x |

Table 2: The structure of generalized linear mixed models.

| general case | logit-normal |
|---|---|
| Y ~ distribution | Y ~ Bernoulli |
| $\mu$ = mean of Y | p = mean of Y |
| $g(\mu) = X\beta + Zu$ | $\ln(p/(1-p)) = \beta x + u_i$ |
| link function $g(\cdot)$ | logit link |
| fixed factors $X\beta$ | fixed factor x |
| random factors Zu | random intercepts $u_i$ |
| u ~ distribution | $u_i \sim \text{Normal}(\mu_u, \tau_u)$ |