Mixed Model Equations and Best Prediction for Binary and Discrete Data

by

Charles E. McCulloch

BU-1158-M

May 1992

Abstract

The mixed model equations are used in linear, normal models to find predictors for the values of random effects and estimators for the fixed effects and variance components. The rationale behind the mixed model equations and best prediction are investigated and used to extend the ideas to the binary data situation as well as to review some currently used approaches.

1. Introduction

Henderson (in Henderson, Kempthorne, Searle and von Krosigk, 1959) proposed the use of the mixed model equations (MMEs) in a mixed linear model in order to simultaneously calculate estimates of the fixed effects and predictions of the random effects. He motivated the equations on computational grounds and also noted that they arise from maximizing the joint density of the data and the unobserved random effects. Henderson (1969, 1973) and Harville (1976) later showed that the predictions of the random effects are related to the best linear unbiased predictor (BLUP). Harville (1977) indicates the connection between the MMEs and maximum likelihood (ML) estimation of the variances of the random effects.

In this paper we explore the use of these ideas for a binary data model. We derive the joint equations for fixed and random effects and relate them to ML estimation. We compare these to currently used techniques based on maximization of the joint density.

2. The Mixed Model Equations

In the mixed linear model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} ,$$

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

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) ,$$

Henderson (in Henderson, Kempthorne, Searle, and von Krosigk, 1959) proposed the use of the MMEs for simultaneously calculating estimates of $\boldsymbol{\beta}$ and predictions of $\mathbf{u}$. The MMEs take the following form:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{1}$$

or

$$\mathbf{C} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \tag{2}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix}.$$

He proposed the use of (1) on the grounds that direct calculation of $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ required inversion of the $n \times n$ matrix $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$, while $\mathbf{C}$ of (2) is often much smaller. He also noted that the MMEs arise from maximizing the joint density of $\mathbf{y}$ and $\mathbf{u}$.

$\tilde{\mathbf{u}}$ from (1) takes the form

$$\begin{aligned} \tilde{\mathbf{u}} &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) . \end{aligned} \tag{3}$$

Henderson (1969, 1973) and Harville (1976) have shown that (3) is the BLUP of $\mathbf{u}$, assuming $\mathbf{D}$ and $\mathbf{V}$ are known. It is also the best predictor of $\mathbf{u}$, namely $\mathrm{E}[\mathbf{u} \mid \mathbf{y}] = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ replaced by $\tilde{\boldsymbol{\beta}}$.

Harville (1977) has also shown that the MMEs are related to an iterative scheme for solving the

ML equations of **D** and **R**. These, in turn, have been shown to be related to the EM algorithm (Laird, 1982). To be more specific, consider the model where

$$\mathbf{R} = \mathbf{I}\sigma_e^2 , \qquad \mathbf{D} = \text{diag}\{\mathbf{I}\theta_i\}, \qquad \text{and} \qquad \mathbf{u}' = [\mathbf{u}_1' \; \mathbf{u}_2' \cdots \mathbf{u}_r'] , \tag{4}$$

where **u** is partitioned to conform to the blocks of **D**. Then Harville (1977) derives the following iterative scheme for θ_i:

$$\theta_i^{(m+1)} = \left(\tilde{\mathbf{u}}_i^{(m)\prime} \tilde{\mathbf{u}}_i^{(m)} + \theta_i^{(m)} \text{tr}(\mathbf{W}_{ii}) \right) \Big/ q_i , \tag{5}$$

where the (m) in superscript denotes iteration m and $\mathbf{W}_{ii}$ = $i^{\underline{th}}$ diagonal block of $(\mathbf{I} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D})^{-1}$. This iterative scheme is also a version of the EM algorithm (see Laird, 1982).

The MMEs can thus be thought of in a variety of ways: as computing formulae for the MLE of $\boldsymbol{\beta}$, as equations for the maximizing values of the joint density of **y** and **u**, as a means of calculating best predictions of **u**, and as integral pieces of the EM algorithm and iterative ML procedures. In the next section we explore how these different viewpoints generalize for binary data.

3. Binary Data

The mixed, linear model of Section 2 is only appropriate for data well approximated by a continuous normal model. Yet much of the data gathered for variance components estimation is categorical or binary. An example is the estimation of genetic variances for breeding purposes (e.g., Im and Gianola, 1988). For our discussion we consider the threshold model for binary data, a flexible mixed model for binary or ordered categorical data. Let $\mathbf{y}$ represent a latent variable following the mixed model of Section 2:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \\ \mathbf{u}' &= [\mathbf{u}_1' \ \mathbf{u}_2' \cdots \mathbf{u}_r'] \\ \mathbf{u}_i &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\theta_i\right), \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) , \end{aligned} \tag{6}$$

but we observe only the binary response $W_i = \mathbf{I}_{\{y_i > 0\}}$. This is a well-used model (e.g., Gianola and Foulley, 1983) and it reduces to the usual probit analysis model if either $\mathbf{u} \equiv \mathbf{0}$ or if there is a single random effect and only one observation per level of the random effect. It is unimportant whether we actually believe in the latent variable $\mathbf{y}$ or merely use it as a device to build a flexible class of models. We now investigate the analogs of the MMEs for this model.

McCulloch (1992) has shown that $\mathrm{E}[\mathbf{u}_i \,|\, \mathbf{W}]$, the best predictor of $\mathbf{u}_i$, is given by

$$\tilde{\mathbf{u}}_i = \mathrm{E}[\mathbf{u}_i \,|\, \mathbf{W}] = \theta_i^2 \mathbf{Z}_i' \mathbf{V}^{-1} (\boldsymbol{\mu}_{y|W} - \mathbf{X}\boldsymbol{\beta}) ,$$

where

$$\boldsymbol{\mu}_{y|W} = \mathrm{E}[\mathbf{y} \,|\, \mathbf{W}] \quad \text{and} \quad \mathbf{V} = \mathbf{I} + \sum_{i=1}^{r} \theta_i \mathbf{Z}_i \mathbf{Z}_i' .$$

This is the same as for the linear, normal model (3) with $\mathbf{y}$ replaced by $\mathrm{E}[\mathbf{y} \,|\, \mathbf{W}]$. McCulloch (1992) also shows that the EM algorithm has iterates for $\boldsymbol{\beta}$ which are of the form

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1} \boldsymbol{\mu}_{y|W} ,$$

again of the same form as for the linear, normal model with $\mathrm{E}[\mathbf{y} \,|\, \mathbf{W}]$ replacing $\mathbf{y}$. From this we can back-solve for a set of MMEs which are nearly identical to (1), namely

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mu_{y|W} \\ \mathbf{Z}'\mu_{y|W} \end{bmatrix}. \tag{7}$$

This is less useful from a computational viewpoint than (1) since $\mu_{y|W}$ depends on the variance components and would need to be recomputed during any iterative procedure. We now look at the connection between an iterative algorithm for ML (EM) and its connection to the MMEs.

McCulloch (1992) shows that the EM algorithm for the model (6) takes the form

$$\begin{aligned} q_i\theta_i^{(m+1)} &= \theta_i^{(m)2}\operatorname{tr}\mathbf{V}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{V}_{y|W} \\ &\quad + \theta_i^{(m)2}(\mu_{y|W}-\mathbf{X}\beta)'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mu_{y|W}-\mathbf{X}\beta) \\ &\quad + \operatorname{tr}\left(\theta_i^{(m)}\mathbf{I}-\theta_i^{(m)2}\mathbf{Z}_i'\mathbf{V}\mathbf{Z}_i\right), \end{aligned}$$

where

$$\mathbf{V} = \sum_{i=1}^{r}\theta_i^{(m)}\mathbf{Z}_i\mathbf{Z}_i'+\mathbf{I} \quad \text{and} \quad \mathbf{V}_{y|W} = \operatorname{var}(\mathbf{y}\,|\,\mathbf{W}).$$

Using the notation $\tilde{\mathbf{u}}_i^{(m)} = \theta_i^{(m)}\mathbf{Z}_i'\mathbf{V}^{-1}(\mu_{y|W}-\mathbf{X}\beta)$ and the result $\operatorname{tr}(\theta_i\mathbf{I}-\theta_i^2\mathbf{Z}_i'\mathbf{V}\mathbf{Z}_i) = \theta_i\operatorname{tr}(\mathbf{W}_{ii})$ (Searle, Casella and McCulloch, 1992, p. 279) where $\mathbf{W}_{ii}$ is defined in Section 2, we have

$$\theta_i^{(m+1)} = \frac{\tilde{\mathbf{u}}_i^{(m)'}\tilde{\mathbf{u}}_i^{(m)}+\theta_i^{(m)}\operatorname{tr}(\mathbf{W}_{ii})+\theta_i^{(m)2}\operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{V}_{y|W})}{q_i}. \tag{8}$$

So for binary data we see that the EM algorithm is very similar to the linear, normal case (5), differing only in the presence of an additional term, $\theta_i^{(m)2}\operatorname{tr}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{V}_{y|W})$. Again, this is computationally less attractive than (5) because of the need to update the $n\times n$ matrix $\mathbf{V}_{y|W}$ at each iteration.

Another approach to deriving the MMEs is to maximize the joint density of $\mathbf{y}$ and $\mathbf{u}$. The joint density is given by

$$\begin{aligned} f_{\mathbf{y},\mathbf{u}}(\mathbf{y},\mathbf{u}) &= f_{\mathbf{y},\mathbf{u}}(\mathbf{y}|\mathbf{u})f_{\mathbf{u}}(\mathbf{u}) \\ &\propto \prod_{i=1}^{n}\Phi(\mathbf{x}_i\beta+\mathbf{z}_i\mathbf{u}+\epsilon_i)^{y_i}[1-\Phi(\mathbf{x}_i\beta+\mathbf{z}_i\mathbf{u}+\epsilon_i)^{1-y_i} \\ &\quad \times |\mathbf{D}|^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}}, \end{aligned}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i^{th} rows of $\mathbf{X}$ and $\mathbf{Z}$.

Maximizing this with respect to β and $\mathbf{u}$ leads to a nonlinear system of equations. A number of authors (Gianola, Foulley and Fernando, 1986; Harville and Mee, 1984; Gilmour, Anderson and Rae, 1985) have employed various approximations and iterative procedures (for a review see Foulley, Gianola and Im, 1990) to solve for the joint maxima. These give rise to equations of the form

$$\begin{bmatrix} \mathbf{X}'\mathbf{Q}\mathbf{X} & \mathbf{X}'\mathbf{Q}\mathbf{X} \\ \mathbf{Z}'\mathbf{Q}\mathbf{X} & \mathbf{Z}'\mathbf{Q}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Q}\hat{\mathbf{y}} \\ \mathbf{Z}'\mathbf{Q}\hat{\mathbf{y}} \end{bmatrix}, \tag{9}$$

where $\mathbf{Q}$ is a weight matrix and $\hat{\mathbf{y}}$ is an estimate of the underlying latent variable. These are similar to the equations (7) with $\boldsymbol{\mu}_{\mathrm{y}|\mathrm{W}}$ replaced by $\hat{\mathbf{y}}$ and with $\mathbf{Q}$ where an $\mathbf{R}^{-1}$ would be, if not assumed to be the identity as in (6). A drawback of (9) is that it is not clear what sort of frequentist properties the solution $\tilde{\mathbf{u}}$ will possess. See Robinson (1991) for a brief discussion.

Another drawback of (9) is that it does not directly give estimates of the variance components. Authors (e.g., Harville and Mee, 1984; Stiratelli, Laird and Ware, 1984) typically assume that the joint density can be approximated by a multivariate normal density. This then gives rise to EM iteration equations of the form (5) rather than (8).

4. Summary

The calculation of BLUPs, efficient calculation of the MLE of $\boldsymbol{\beta}$, maximization of the joint density of $\mathbf{y}$ and $\mathbf{u}$, and close ties between BLUP and EM all come together when using the linear, normal, mixed model in the form of the MMEs. When these separate ideas are considered for binary data, the resulting techniques do not coincide.

Best prediction of the random effects in the binary data, threshold model through the use of $E[\mathbf{u}\,|\,\mathbf{y}]$ does not result in a linear function of the data, since the distribution is non-normal. However, the best predictor of $\mathbf{u}$ and the MLE of $\boldsymbol{\beta}$ both are familiar functions of $\boldsymbol{\mu}_{y|W}$ and a simple set of MMEs can be formed to calculate $\tilde{\mathbf{u}}$ and $\tilde{\boldsymbol{\beta}}$.

Turning to the EM algorithm and its relation to the MMEs, we see that the EM algorithm for binary data has a form very similar to that for normal data, differing only in the presence of an additional term in the iteration equations. Thus the computational advantage of the MMEs can be exploited for the pieces which are the same. Unfortunately, the additional piece is computationally intensive (McCulloch, 1992) and no similar simplified calculation of it is apparent from the MMEs.

Other authors have attempted the maximization of the joint distribution of $\mathbf{u}$ and $\mathbf{y}$ with respect to $\mathbf{u}$ and $\boldsymbol{\beta}$. In connection with iterative schemes to find the maximum and certain approximations, these lead to approximate MMEs of a form similar to that of the linear, normal, mixed model. These are problematic in two regards: 1) It is not clear what frequentist properties such a $\tilde{\mathbf{u}}$ will possess and 2) Assumptions of approximate normality have been needed to form iterative equations for estimation of the variance components.

The fact that these different approaches to estimation in the case of binary data lead to different techniques yields both problem and opportunity. It yields the problem of different solutions where before we had but a single one. It yields the opportunities to choose from a variety of techniques and to judge the performance of different criterion for distributional models other than the linear, normal, mixed model.

5. References

Foulley, J.L., Gianola, D. and Im, S. (1990). Genetic evaluation for discrete polygenic traits in animal breeding. In *Advances in Statistical Methods for Genetic Improvement of Livestock,* Gianola, D. and Hammond, K. (Eds.). Springer-Verlag, Berlin, pp. 361-409.

Gianola, D. and Foulley, J.L. (1983). Sire evaluation for ordered categorical data with a threshold model. *Genet. Sel. Evol.* **15**, 201-224.

Gianola, D., Foulley, J.L., and Fernando, R.L. (1986). Prediction of breeding values when variances are not known. *Genet. Sel. Evol.* **18**, 485-498.

Gilmour, A.R., Anderson, R.D., and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593-599.

Harville, D.A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Ann. Statist.* **2**, 384-395.

Harville, D.A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-340.

Harville, D.A. and Mee, R.W. (1984). A mixed model procedure for analyzing ordered categorical data. *Biometrics* **40**, 393-408.

Henderson, C.R. (1969). Design and analysis of animal husbandry experiments. In *Techniques and Procedures in Animal Science Research*, 2nd ed., Chapter 1. American Society of Animal Science Monographs, Quality Corporation, Albany, New York.

Henderson, C.R. (1973). Sire evaluation and genetic trends. In *Proc. Anim. Breed. Genet. Symp. in Honor of Dr. J.L. Lush.* ASAS and ADSA, Champaign, Illinois, pp. 10-41.

Henderson, C.R., Kempthorne, O., Searle, S.R., and von Krosigk, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **1**: 192-218.

Im, S. and Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Appl. Statist.* **37**, 196-204.

Laird, N.M. (1982). Computation of variance components using the EM algorithm. *J. Statist. Comp. & Simul.* **14**, 295-303.

McCulloch, C.E. (1992). Maximum likelihood variance components estimation for binary data. Technical Report BU-1037-MB, Biometrics Unit and Statistics Center, Cornell University, Ithaca, New York.

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

Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance Components.* (In the Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics Section.) John Wiley and Sons, Inc., New York.

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