A SURVEY OF VARIANCE COMPONENTS ESTIMATION FROM BINARY DATA

by

Charles E. McCulloch
Biometrics Unit and Statistics Center
Cornell University

BU-1211-M May 1993

ABSTRACT

The basic problem of variance components estimation from binary data is described and motivated with an example. Marginal and conditional models which have been used are discussed, including random effect, beta-binomial and quadratic exponential family models. Methods of estimation are also addressed, including ML, various definitions of REML, estimating equations methods and methods based on the analogs of the mixed model equations.

1. INTRODUCTION

The response variable in many statistical analyses is binary which may necessitate the use of techniques specifically designed for such data. For the normal, linear, mixed model maximum likelihood (ML) and restricted maximum likelihood (REML) are well accepted techniques. However, in variance components estimation from binary data, computation of ML estimates can be quite difficult and even the definition of REML is subject to debate. Consequently a number of alternate techniques have been proposed. We review the models suggested for such situations and methods of estimation proposed for those models.

In the rest of this section we discuss some examples of variance components estimation from binary data and delineate some of the challenges in specifying models and fitting them to data. In Section 2 we discuss models which have been proposed; in Section 3 we discuss methods of estimation. Section 4 concludes with discussion.

To motivate these methods and for later discussion we now consider some examples. The primary example we will consider is the salamander mating experiment presented in McCullagh and Nelder (1989) which has been widely used as an illustrative example. In this experiment, 20 male and 20 female salamanders (ten of each of two species) were mated. The response variable is mating outcome (success/failure). We give some numerical details for the three stages of the experiment in each of which 120 matings were observed. One of the goals of the experiment was to quantify the variation in the probability of mating due to animal to animal differences in the males and females.

Other examples abound. Searle, Casella and McCulloch (1992) describe a study of reproductive success (survive to reproduce or do not survive) of aphids where the goal was to estimate the variation from aphid clone to aphid clone. Stiratelli, Laird and Ware (1984) describe a study of asthma attacks (yes/no) in which a goal was to predict the susceptibility of individuals to pollution and weather conditions. Thompson (1990) reviews some applications in animal breeding.

For several reasons variance components estimation for binary data is not as straightforward as for the "usual" model, i.e., the normal, linear, mixed model. Modelling E[y] in the binary data case means modelling the probability of a success. Since probabilities are bounded between 0 and 1, it is natural to consider nonlinear models. In the "usual" mixed model, the random effects contribute a fixed variance to the response. With binary data this is no longer possible since the variability must be less when probabilities are near 0 or 1 and hence another reason to consider nonlinear models. With binary (Bernoulli) data there is a tie between the mean and variance and this is further justification for a using random effects models whose the variance must be nonconstant. All of the above argue strongly for using techniques specifically for handling the binary data situation.

Given that a nonlinear model is to be used, a natural candidate would be to generalize the standard methods used in analyzing binary data, namely the logistic or probit models. Unfortunately,

the usual estimation methods of ML and REML are computationally quite difficult for these models. This has lead to the use of alternate models and alternate methods of estimation, which we now consider.

The focus here will be on flexible models and methods which can be used in a wide variety of situations. Some topics will therefore be dismissed with only the briefest of coverage.

2. MODELS

2.1 Marginal versus Conditional Models

A fundamental distinction for nonlinear models is whether they are models for the marginal distribution of the binary variable y or whether they are conditionally specified models (Zeger, Liang and Albert, 1988). This distinction is important because the interpretation of the parameters in the two types of models is quite different. Zeger, Liang and Albert (1988), Liang, Zeger, and Qaqish (1992) and Prentice and Zhao (1991) give guidelines as to the advantages of each method when interest centers on the mean of y and its behavior as a function of covariates or factors. If instead the focus is on variance components estimation, then the marginal approach is unsatisfactory.

Variance components estimation, by its very nature, attempts to divide up the variability and attribute it to various sources. As such, one must have a structure or model for how the variability arises and one is naturally led to conditionally specified models. In the presence of covariates which have (possibly) different values for each observation, a conditionally specified model is practically necessary to separate the effects of covariates and variability attributable to a factor.

2.2 Models

As mentioned previously a natural approach would be to use standard methods for binary data analysis, e.g., logistic and probit regression analyses and extend them to mixed models. After all, it is often difficult to decide (in the normal, linear, mixed model) if a factor should be treated as fixed or random. Frequently, the model is formulated first and the decision as to fixed or random is made afterwards.

With nonlinear models it is natural to think about the factors acting in a similar fashion and again deciding "fixed" or "random" at a later stage. This suggests the use of models which can be viewed within a framework of generalized, linear, mixed models (GLMM) as recently described by

Schall (1991) and Breslow and Clayton (1993). The GLMM for binary data is specified in the following conditional manner:

$$\begin{aligned} &y_i \sim \text{Bernoulli}(p_i), \\ &\mathbf{p} = h(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}), \\ &\mathbf{u} \text{ has distribution with} \\ &E[\mathbf{u}] = \mathbf{0} \text{ and } \text{Var}(\mathbf{u}) = \mathbf{D}\,. \end{aligned} \tag{1}$$

In this formulation $\mathbf{u}$ represents the random effects with variance components $\mathbf{D}$ and $h(\cdot)$ is usually either the logistic cdf (for a logistic mixed model) or the standard normal cdf (for a probit mixed model). A common assumption would be that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$. The nonlinear function $h(\cdot)$ is the key to adapting many of the pieces of the normal, linear, mixed model for use with binary data.

The model (1) would probably be widely used were it not for the difficulty of ML and REML estimation (see Section 3). Because of this computational problem, a number of other approaches have been suggested. One which has been used for many years and is natural for simple problems is the beta-binomial (e.g., Crowder, 1978 or Searle, Casella, and McCulloch, 1992). This does not generalize easily to arbitrary numbers of factors and random effects. Rosner and co-workers (e.g., Rosner, 1984; Tosteson, Rosner and Redline, 1991) have made some extensions of this model, however.

Another class of models which have enjoyed recent interest are quadratic exponential models (Zhao and Prentice, 1990; Prentice and Zhao, 1991). The model has the following form:

$$\begin{aligned} &\mathbf{y}_i \sim \text{indep } f(\mathbf{y}_i;\, \Theta_i,\, \boldsymbol{\lambda}_i) \\ &f(\mathbf{y}_i;\, \Theta,\, \boldsymbol{\lambda}) = \Delta^{-1} \exp\{\mathbf{y}'\Theta + \mathbf{w}'\boldsymbol{\lambda} + c(\mathbf{y})\}\,, \end{aligned} \tag{2}$$

where $\mathbf{w}' = (y_1y_2,\, y_1y_3,\, \ldots\,,\, y_2y_3,\, \ldots)$, Δ^{-1} is a normalizing constant and $c(\cdot)$ is a "shape" function. This model has several drawbacks for variance components estimation: it requires independent blocks of observations (not true, for example, in the salamander data), it is a marginal model and ML estimation is computationally difficult when the size of $\mathbf{y}_i$ gets large.

3. METHODS OF ESTIMATION

3.1 Introduction

We now consider estimation methods for the models of Section 2, especially (1). As alluded to, ML and REML estimation are problematic so we consider alternatives based on estimating equation methods and analogs of the mixed model equations from normal, linear, mixed models (Searle, Casella, and McCulloch, 1992).

3.2 Maximum Likelihood

The discussion in Section 2 indicates that the model given by (1) is preferable. With the assumption that $\mathbf{u} \sim \mathbf{N(0, D)}$ the likelihood for (1) is proportional to

$$|\mathbf{D}|^{-\frac{1}{2}} \int \prod_{i=1}^{n} h(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})^{y_i} [1 - h(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})]^{1-y_i} \exp\{-\mathbf{u}'\mathbf{D}^{-1}\mathbf{u}\}d\mathbf{u} \qquad (3)$$

where $\mathbf{x}_i'$ and $\mathbf{z}_i'$ are the ith rows of $\mathbf{X}$ and $\mathbf{Z}$. The dimension of the integral in (3) is equal to the numbers of levels of random effects in $\mathbf{u}$ and can be quite large. In general (3) does not simplify and its evaluation is quite difficult. A computationally intensive approach for calculating the ML estimates without directly evaluating (3) has been suggested in McCulloch (1994).

3.3 Restricted Maximum Likelihood

Restricted maximum likelihood (REML) was developed specifically for the normal, linear, mixed model to produce estimates of the variance components invariant to the fixed effects. There are a number of ways to justify REML on more general grounds which are applicable to the binary data situation. These do not lead to the same method as in the normal, linear, mixed model. Stiratelli, Laird and Ware (1984) use a Bayesian approach where the variance of the fixed effects becomes diffuse. Reichert (1992), generalizing the dispersion-mean model of Pukelsheim (see Searle, Casella, and McCulloch, 1992 for an accessible discussion), gives another version of REML. Her approach is similar in spirit to Prentice (1988) and the model (2) which incorporates quadratic functions of the ys. Unfortunately, in either case the exact REML computations are no more feasible than the ML ones.

3.4 Mixed Model Analogs

Given the difficulty of ML and REML for (1), other approaches have been suggested for computing variance components estimates. Stiratelli, Laird and Ware (1984), Schall (1991) and Breslow and Clayton (1993) have justified, each on different grounds, the same analog of the mixed model equations. This analog is based on approximating the binary variates with a normal distribution and using the mixed model equations for ML or REML. Gilmour, Anderson and Rae (1984) and Foulley, Gianola and Im (1990) derive similar equations.

3.5 Generalized Estimating Equations

A very popular approach to the analysis of correlated binary data has been the use of generalized estimating equations (GEE). This approach can be used for conditionally specified models (Zeger, Liang and Albert, 1988) like (1). However, the GEE method focuses mainly on the estimation of fixed effects and, in its original form, is both ad hoc and inefficient for estimating variance components

(Zeger, Liang, and Qaqish, 1992). They detail extensions of the GEEs for models like (2). However, these extensions are subject to the same drawbacks as (2).

4. DISCUSSION

We have argued that models of the form (1) are most natural for variance components estimation. Unfortunately they pose computational problems in forming ML or REML estimates. Alternate methods of estimation have been proposed based on normal approximations and the mixed model equations. These are ad hoc in nature and their performance as estimators needs to be evaluated. They give equations which are fundamentally different in form from ML equations for binary data (McCulloch, 1992). GEE approaches are limited in the requirement that the data fall into independent blocks and by their ad hoc approach to variance components estimation.

Arguments about which methods are best are not academic. On a particular data set, they can give quite different estimates. Table 1 below gives the estimated variance components using four different methods for a portion of the salamander data (McCullagh and Nelder, 1989).

Table 1. Variance components estimates for the salamander data set by four methods

		Data Set					
		1		2		3	
Method	Variance	σ^2_F	σ^2_M	σ^2_F	σ^2_M	σ^2_F	σ^2_M
Moment[a]		1.37	.70	.98	.60	.40	1.34
Bayes[b]		2.35	.14	2.99	1.42	.33	2.89
Mixed model[c]		1.41	.09	1.26	.62	.26	1.50
ML[d]		1.73	.17	1.42	1.30	.29	1.27

[a]McCullagh and Nelder (1989) table 14.10

[b]Karim and Zeger (1992) table 3, medians

[c]Breslow and Clayton (1993) table 8

[d]McCulloch (1994) table 1

REFERENCES

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

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

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.

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

Liang, K.-Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**: 3-40.

McCulloch, C.E. (1992). Mixed model equations and best prediction for binary and discrete data. Biometrics Unit Technical Report BU-1158-M, Biometrics Unit, Cornell University.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. To appear in *Journal of the American Statistical Association.*

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (Second Edition). Chapman and Hall. London.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**: 1033-1048.

Prentice, R.L., and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses.

Rosner, B. (1984). Multivariate methods in ophthalmology with application to other paired data situations. *Biometrics* **40**: 1025- 1035.

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

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

Thompson, R. (1990). Generalized linear models and applications in animal breeding. In *Advances in Statistical Methods for Genetic Improvement of Livestock.* Gianola, D. and Hammond, K. (eds.). Springer-Verlag, Berlin.

Tosteson, T.D., Rosner, B. and Redline, S. (1991). Logistic regression for cluster binary data in proband studies with application to familial aggregation of sleep disorders. *Biometrics* **47**: 1257-1265.

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

Zhao, L.P., and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**: 642-648.