TOWARD A MULTILEVEL GENERALIZED LINEAR MODEL: THE CASE FOR POISSON DISTRIBUTED DATA

By

Wing Shing Chan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology
and Special Education

1995

ABSTRACT

TOWARD A MULTILEVEL GENERALIZED LINEAR MODEL: THE CASE FOR POISSON DISTRIBUTED DATA

by

Wing Shing Chan

This study presents a maximum likelihood approach for estimating a random coefficient generalized linear model via the Monte Carlo EM algorithm. Monte Carlo integration through simulating multivariate normal variates is used to integrate out the random effects in the E-step. The M-step is equivalent to maximizing a weighted sum of log likelihood for an exponential family and a multivariate normal distribution. The former weighted likelihood is maximized by one step of the Iteratively Reweighted Least Square algorithm while the latter is maximized in closed form by choosing appropriate weights. Standard errors for the fixed effect parameters and variance components are obtained through the tractable observed complete data information matrix. The Poisson distribution is used for illustration. A simulation study with a non-diagonal variance-covariance matrix shows better accuracy than an earlier comparable Gibbs sampling approach. Computing efficiency is demonstrated as the FORTRAN program converges in about

40 iterations within 30 minutes for a two-variance component model with 700 Poisson observations (100 groups each with 7 within-group units) on an IBM compatible 486 DX/33MHz personal computer.

Dedicated to

the land and people of Michigan

ACKNOWLEDGMENTS

I wish to express my thanks to Prof. Betsy J. Becker (Measurement and Quantitative Methods), Prof. Ralph L. Levine (Psychology), Prof. Stephen W. Raudenbush (Measurement and Quantitative Methods) and Prof. James H. Stapleton (Statistics and Probability) for serving on my dissertation committee and giving me many useful suggestions. I am grateful to Prof. Raudenbush for introducing me the concepts and principles of the Hierarchical Linear Models and to Prof. Stapleton for teaching me five graduate courses in mathematical statistics. Moreover, the large and beautiful campus of the Michigan State University has provided a favorable environment for building the foundation of my graduate education.

I also wish to express my gratitude for the hospitality from the Department of Maternal and Child Health at Harvard University where I visited as a research specialist in 1991-1992. I wrote my dissertation proposal while residing in Somerville of Massachusetts during this period.

I wish to thank my wife, Li Ming Wei, for her patience, encouragement and support as well as for her assistance in computer programming. She accompanied me to live in Forest Hills, New York where I conducted my independent dissertation research in 1992-1994. In 1993 the Queens College of CUNY had offered me some helpful part-time research and teaching opportunities and an Internet access to the rest of the world. I wish to thank my family members in Hong Kong for their needful financial support during the last phase of the dissertation research and writing.

Last but not least, I wish to thank my Shih-Fu, Ch'an Master Dr. Sheng-Yen in Elmhurst, New York. From 1991 to 1995, his teaching in Buddhist meditation is of great help to my personal life and academic research as I attain a calmer and clearer mind.

W. S. CHAN

Queens, NY

March 1995

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

Figure

CHAPTER I

INTRODUCTION

Educational data concerning students often arise within classrooms or schools. Since different classrooms have teachers of various experiences and qualifications, and different schools have various school climates, policies and facilities, educational results vary according to the quality of the educational context. One of the most important questions in educational research is what kind of teachers or schools will enhance learning most. To answer this question empirically, we must collect data about students' academic performance and characteristics of the teachers, classrooms and schools. For research results to be readily used for academic theory construction or public policy making, large samples of classrooms or schools must be studied to make broad generalizations.

However, traditional statistical research tools such as linear regression analysis fail to account for the variability due to the varying effects of each classroom or school. Ordinary random effects analysis of variance models also fail to allow for unbalanced data and flexible covariate adjustments. Because the scores of students within the same classroom or school are

correlated, the standard statistical independence assumption is violated when data from large numbers of classrooms or schools are analyzed. As a remedy to these deficiencies in traditional statistical tools, multilevel or hierarchical linear statistical models have been developed for educational and social research applications (Aitkin, Anderson, and Hinde, 1981, Goldstein, 1986, Longford, 1987 and Raudenbush and Bryk, 1986).

Multilevel linear models have two or more levels of regression equations. For example, in two-level models, a level-one equation relates students' outcomes (e.g., math achievement or verbal ability) to a set of independent variables (e.g., previous GPA, IQ, sex or socioeconomic status). However, each school is allowed to have its own regression equation. The slopes or coefficients of the level-one regression equation for each school are considered randomly distributed. The level-one coefficients are predicted by another set of independent variables related to the level-two units, that is, the school variables (e.g., private or public school, rural or urban school and racial composition of the school). The explicit modeling of level-one coefficients by level-two variables helps educational researchers study the intriguing relationships among schools, teachers and students. For example, Raudenbush and Bryk

(1986) were able to demonstrate through a multilevel model that the type of school (Catholic or public) has a differentiating effect on the influence of students' socioeconomic status (SES) on Math achievement. Specifically, they found that Catholic schools have a flatter slope in a regression of Math achievement on SES, which supports the contention that Catholic schools are more egalitarian than public schools.

Moreover, educational researchers making use of multilevel models need not be restricted by either choosing the student level or the aggregated school level for regression analysis. The estimated residual variances of the regression equations from both levels not only account for the components of variances at the student and school levels, they also give information about the relative amount of unexplained variation for the two levels.

When repeated measurements on students over time are taken as the level-one units and the students are correspondingly taken as level-two units, the multilevel model can be used as an appropriate model to study academic growth over time. Student variables such as sex, race or income can be used to predict the effect of time on academic achievement (i.e., the rate of growth or change). Therefore multilevel models are a means to solving two of the most persistent methodological problems in

educational research, namely the assessment of multilevel effects and the measurement of change or growth (Bryk and Raudenbush, 1992). Growing numbers of educational studies, for example, about vocabulary growth, school effectiveness and teaching styles have been conducted in a multilevel perspective (e.g., Aitkin, Anderson, and Hinde, 1981, Raudenbush and Bryk, 1986 and Huttenlocher et al., 1991). A review of educational applications of multilevel models is included in Raudenbush (1988).

Notwithstanding the growing popularity of multilevel models, most of the research applications are confined to continuous outcome variables with normal error distributions. Details about the methodology and applications of normal multilevel models in social and educational research can be found in the books written by Goldstein (1987), Bryk and Raudenbush (1992) and Longford (1993). However, a wide range of educational outcomes are not normally distributed, for example, pass or failure (dichotomous variable), absence from school (discrete counts) and time before graduation or dropout of school (censored survival time data). These non-normal outcomes can be dealt with, respectively, by statistical models based on the binomial, Poisson and exponential distributions.

These common distributions are subsets of the exponential family of distributions. A unified approach for maximum likelihood estimation of single-level regression models based on the exponential family of distributions is the generalized linear model (Nelder and Wedderburn, 1972). Computer software for such statistical models is now also available (Aitkin et al., 1989). Many attempts have been made to develop multilevel non-normal models and they are described in the next chapter. However, a unified approach for a multilevel generalized linear model with full maximum likelihood estimation has not been developed.

This dissertation is an attempt to extend the multilevel normal models to include other non-normal outcome variables so that a wider range of educational dependent variables can be investigated within the multilevel framework. In other words, the goal of the dissertation is to develop a unified maximum likelihood estimation approach for the multilevel generalized linear model. Since the generalized linear model covers a wide range of distributions, the dissertation will demonstrate statistical estimation of one member of the exponential family: the Poisson distribution. The same approach can be followed for the other distributions of the exponential family.

Though the Poisson distribution is not commonly used in educational research, its potential for application should not be underestimated. The Poisson distribution is a standard model for independent count data. Much educational data exists in terms of counts, for example, school absence (Aitkin et al., 1989, p.223), classroom behaviors such as speaking in class, altruistic behaviors, antisocial behaviors, vocabularies of children's utterances, number of peer-reviewed publications, teenage pregnancies, frequencies of using school facilities such as the library, gymnasium and counseling service, and numbers of times repeating a difficult required course or a certifying exam of a professional institution. The Poisson model is especially useful if the mean occurrence of counts per unit time is low. In these instances many people will have zero observed counts. Approximation of the empirical data by the normal distribution often fails to account for the positive skewness of the data.

Since the educational outcome variables in the above paragraph also arise within educational settings, for example, classrooms, high schools and tutorial schools, the multilevel approach will often need to be applied. The multilevel Poisson model has an additional advantage of being able to help resolve the over-dispersion problem (Cox, 1983) due to more than

expected variance. In theory, data of the Poisson distribution should have its mean approximately equal to its variance. However data arising from groups (e.g., educational institutions) are statistically dependent within a group, so the observed variance of the whole set of data can be much larger than the corresponding mean. A multilevel Poisson model accounts for the variation due to grouping by including a second or third level of variation. Thus it helps remove the over-dispersion due to the natural grouping of the data. Similar problems also exist for the binomial distribution and the multilevel approach can offer the same benefit.

In essence, this dissertation will undertake research that fulfills the objectives below:

1) To develop a multilevel statistical model for the exponential family of distributions.
2) To provide maximum likelihood estimation of the parameters and standard errors for the parameters of the model.
3) To use the Poisson distribution to demonstrate the details of the estimation method.

4) To write an iterative computer program to obtain statistical estimation for the multilevel Poisson model by iteration.

5) To use a small simulation study to demonstrate the validity of the computer program.

Before describing the state of previous research or the technical details for the formulation and estimation of the multilevel generalized linear model, an example on how the model can be applied could be illuminating.

An illustration for the application of a multilevel generalized linear model to educational research

To understand the formulation of the multilevel generalized linear model, it might be best to study a simple hypothetical example. A similar data analytic scenario from a real national school survey, with a normal outcome variable analyzed through the Hierarchical Linear Model, can be found in Raudenbush and Bryk (1986).

Suppose a survey on the number of altruistic behaviors of the recent month is made on students from a large number of schools. Some of the schools have an ethics education program. The socioeconomic status (SES) of each student was also

recorded. The research question is to study whether schools with ethics education programs affect the differentiation effect of SES on the exhibition of altruistic behavior in school. The analysis requires that we model the random effects due to each individual school, thus a multilevel model is formulated. Since the dependent variable takes on a random natural number greater or equal to zero, a Poisson model can be used to fit the data. As the Poisson model is a member of the exponential family, this is an example of the multilevel generalized linear model.

In level 1, the i^{th} student's altruistic behavior y_{ij} in school j is modeled by his/her SES level:

$$\log \lambda_{ij} = \beta_{0j} + \beta_{1j} SES_{ij}.$$ [1]

Here λ_{ij} refers to the mean of a student's counts of altruistic behavior. The response variable $y_{ij}|\lambda_{ij}$ is assumed to be Poisson distributed with parameter λ_{ij}. The logarithmic function is the 'link' function (Aitkin, Anderson, Francis and Hinde, 1989, p.76) that maps the natural numeric count onto the set of real numbers being predicted linearly by the student's SES variable. β_{0j} is the random intercept for school j. β_{1j} is the random slope for the effect of SES on the logarithm of the mean altruistic behavior for students in school j. A positive β_{1j} will mean that the higher

the SES of a student is, the more altruistic behavior the student will elicit. For illustrations, we now assume that empirically β_{1j} is positive. However, the implementation of an ethics education program can either suppress or elevate the influence of SES on altruistic behavior. To study the effect of higher level variables on the relationships between the outcome and the lower level variables, we need the level-2 equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(ETHICS\ PRGM)_j + u_{0j},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(ETHICS\ PRGM)_j + u_{1j}.$$

The random intercepts and slopes of the schools are being predicted by the dummy variable for the ethics education program which is zero for absence and one for presence. The residual errors or random effects u_j's at the school level are assumed to form a multivariate normal distribution with zero expectations.

As an example, if schools with ethics education are found to have higher β_{0j}'s on average (i.e., $\gamma_{01} > 0$), it means that ethics education can help students behave more altruistically. If schools with an ethics education program have smaller β_{1j}'s on average (i.e., $\gamma_{11} < 0$), then we might infer that ethics education can suppress the influence of one's social class background on

one's exhibition of altruistic behaviors in schools. In other words, ethics education enhances the egalitarian tendency of students toward acting altruistically.

This example has demonstrated the potential of applying multilevel generalized linear models for real world educational research involving student-school interaction effects with a discrete count outcome variable.

---

[1] To facilitate understanding for readers who are previously acquainted with the Hierarchical Linear Model (Raudenbush and Bryk, 1986), the mathematical symbols are chosen to be similar to those of HLM.

CHAPTER II

BACKGROUND

Much work has already been done on the multilevel normal models. A review of the methodology and applications of such models in educational research is given by Raudenbush (1988). Representative approaches that have been widely applied in educational research include the models and computer programs formulated by Goldstein (1986, 1987), Longford (1987), and Raudenbush and Bryk (1986). Since the present dissertation focuses on the multilevel extension for non-normal models, the review on multilevel normal models will not be repeated here.

Multilevel extensions of various particular non-normal models have also been published. For example, Anderson and Aitkin (1985), Conaway (1990), Stiratelli, Laird and Ware (1984) and Wong and Mason (1985) have published papers on the binomial models; Goldstein (1991) on the log-linear model; and Albert (1985, 1988, 1992), Morton (1987), Tsutakawa (1985, 1988) on the Poisson models. Albert's (1985, 1988, 1992) models are Bayesian models with a gamma prior distribution. Morton (1987) uses the quasi-likelihood approach in an analysis of variance framework. Tsutakawa (1985, 1988) employs

approximating techniques for both Bayesian and empirical Bayes estimation. All of these past Poisson models do not have the flexibility to allow for a full scale two-level model with multiple explanatory variables and variance-covariance components.

In terms of the whole exponential family of distributions, sometimes labeled as the generalized linear model with random effects, there are models developed by Anderson and Hinde (1988), Longford (1988), Zeger, Liang and Albert (1988), Schall (1991), Zeger and Karim (1991) and Breslow & Clayton (1993).

Various estimation approaches and algorithms have been adopted in the above past research. Statistical estimation approaches include the Bayesian (e.g., Zeger and Karim, 1991), maximum likelihood (e.g., Anderson and Hinde, 1988) and the maximum quasi-likelihood method (e.g., Longford, 1988). The Bayesian approach is especially relevant when the level-2[1] sample size is small and the asymptotic normal approximation of the posterior distribution using the maximum likelihood approach becomes inadequate. Contrarily, the maximum likelihood approach generally involves less computation and simpler analytic methods. It also gives results similar to those of the Bayesian approach when the sample size is large. The maximum quasi-likelihood approach is similar to the maximum

likelihood approach except that the quasi-likelihood function only specifies the relationship between the mean and the variance and does not contain the full parametric likelihood (For details, see Wedderburn, 1974 and McCullagh and Nelder, 1989, p.325). Although the quasi-likelihood approach requires fewer assumptions than the ordinary likelihood approach, it can fail to give reasonable results in some cases (Crowder, 1987). There is also some loss of efficiency when the data depart from a natural exponential family (Firth, 1987).

The algorithms relevant for the present research are: the EM algorithm (Dempster, Laird and Rubin, 1977), Fisher-scoring (Longford, 1987), iterative generalized least squares (Goldstein, 1986, 1989), data augmentation (Tanner and Wong, 1987) and Gibbs sampling (Gelfand, Hills, Racine-Poon and Smith, 1990). Tanner (1991) provides an excellent introduction to many data augmentation methods, broadly defined. It should be noted that a given algorithm might not be restricted to be used only for a particular estimation approach. For example, the EM algorithm can be used for maximizing a general likelihood (Dempster et al., 1977), for empirical Bayes estimation (Dempster, Rubin and Tsutakawa, 1981) or for Bayesian estimation (Racine-Poon, 1985).

Although several models already exist for formulating a multilevel generalized linear model, there has not been a full maximum likelihood (ML) model available for applications in educational and social studies. For example, although Anderson and Hinde (1988) formulate a maximum likelihood approach via the EM algorithm, their estimation method can only allow for a single random intercept model. The potential extension to high dimensions of random coefficients is also hindered by their use of Gaussian quadrature integration technique that works only for small dimensions (Rubinstein, 1981). Longford's (1988) non-normal extension to the multilevel normal model uses an approximation of the quasi-likelihood for estimation.

A recent simulation study by Rodriguez and Goldman (1993) concludes that the approximate quasi-likelihood method of Longford (1988) is equivalent to Goldstein's (1991) approximate generalized least square approach with regard to the multilevel logit model. The simulation results reveal substantial biases in the estimates of the fixed effects and/or the variance components whenever the random effects are large enough to be interesting. Moreover, the multilevel estimates of the fixed effects from the approximate quasi-likelihood method are virtually the same as those obtained using standard logit

models that ignore the hierarchical structure of the data (Rodriguez and Goldman, 1993, p.15).

Zeger, Liang and Albert (1988) use a generalized estimating equation approach (Liang and Zeger, 1986) which combines quasi-likelihood and robust variance estimation for the marginal or so called 'population average' model. The estimating equation (Zeger et al., 1988, p.1053) is essentially a variant of the quasi-likelihood estimation (c.f. McCullagh and Nelder, 1989, p.327). Schall's (1991) approach is an approximate maximum likelihood and quasi-likelihood estimation by means of a first order Taylor's expansion of the linked data. Zeger and Karim (1991) adopt the Bayesian approach via the Gibbs sampling method which is heavily computing intensive. Karim and Zeger (1992) have reported that a disadvantage of the Gibbs sampler is the computational burden. In their analysis of an ecological data set on Salamander mating with 360 binary responses, it takes about 5 hours of computer time on a 14 MIP DEC station 3100 microcomputer to generate 2000 simulated values from the posterior distribution, each obtained after 80 iterations of the Gibbs algorithm. They consider that the time required is sufficiently long as to possibly discourage the fitting of several different models (Karim et al., 1992, p.643). This

would not be feasible for educational applications with large data sets. In their paper on approximate inference for generalized linear mixed models, Breslow and Clayton (1993) use Laplace's method for integral approximation in conjunction with a quasi-likelihood approach.

Even for a particular distribution of the generalized linear model, there has not been a true maximum likelihood model. For example, Stiratelli, Laird and Ware (1984) estimate a multilevel binomial model via the EM algorithm. Because the joint posterior distribution of the fixed effect parameters and variance components is intractable, it is approximated by a multivariate normal approximation. The properties of their estimates will depart from those of ML estimates. Goldstein (1991) uses a first order Taylor's expansion for the nonlinear part of the likelihood and the resultant Iterative Reweighted Least Square estimates do not remain maximum likelihood estimates (See Rodriguez and Goldman, 1993).

Overall, Zeger and Karim (1991) have provided the most promising simulation results so far for a multilevel Bayesian logit model with two variance components. They have demonstrated a simulation study using a diagonal variance-covariance matirx. A random intercept model is used to analyze

a clinical data set for illustration. Breslow and Clayton (1993) present a comparative simulation study against Zeger and Karim (1991). Their results approach that of Zeger and Karim (1991) when the sample size is increased.

Implications for the present research

Vis-a-vis the present status of research in the field of multilevel generalized linear models, this dissertation will attempt to provide an alternative of a full maximum likelihood estimation approach with multiple random regression coefficients.

In order to reduce the expected intensive computation, I choose to adopt the maximum likelihood approach instead of the Bayesian approach because the latter generally requires solving high dimensional multiple integrals (Smith et al., 1985) or many repeated rounds of simulations (e.g., Zeger and Karim, 1991). The maximum likelihood approach is used in preference to the quasi-likelihood approach because of the former's well-behaved asymptotic properties of consistency, unbiasedness and efficiency (Rice, 1987, p.234-254).

As for the algorithm that maximizes the likelihood function, the EM algorithm (Dempster et al., 1977) is used for

the following reasons. Since much educational data involves national surveys with huge numbers of schools, the number of unobserved random effects due to schools could be large (e.g., in the thousands). A naive and direct maximization of the likelihood involves simultaneously estimating the fixed effects and all the random effects. These huge number of parameters are unstable to estimate for most common maximization routine. However, in maximizing the marginal likelihood, the EM algorithm treats the random effects as missing data and essentially estimates just the fixed effect and variance component parameters. The EM algorithm is preferable to the other Newton-type algorithms because it does not require computing the inverse of the information matrix. Generally, simple closed form solutions are attainable for the iterating steps of the EM algorithm. Iterates are also confined within the parameter space during the execution of the EM algorithm. It is also proved that the EM algorithm increases the marginal likelihood of the observations for every iteration (Dempster et al., 1977). A disadvantage of the EM algorithm is its relatively slow linear convergent rate. Because I have adopted Monte Carlo integration to use with the EM algorithm, a slower convergent rate could prove to be an advantage because the

iterates will not so easily jump out of the parameter space due to the randomness of the Monte Carlo simulations. More details about the EM algorithm will be provided in the next chapter.

Before giving the mathematical details in formulating and estimating the multilevel generalized linear model, the definitions and properties of the exponential family of distributions and the generalized linear model are described as follows.

Exponential family of distributions

The distribution of a random variable Y belongs to the exponential family if it can be expressed in the form (McCullagh and Nelder, 1989, p.28-29.):

$$f(y|\theta,\phi) = \exp[(y\theta - b(\theta))/a(\phi) + c(y,\phi)] \qquad (2.1)$$

where $a(.), b(.)$ and $c(.)$ are some known functions. The parameter θ is called the canonical parameter and ϕ is the dispersion parameter. The function $a(\phi) = \phi/m$, where m is the prior weight for the data. This function reduces to ϕ if the data is unweighted. It can be easily shown that the first two moments are (McCullagh and Nelder, 1989, p.29):

$$E(Y) = b'(\theta), Var(Y) = b''(\theta)a(\phi). \qquad (2.2)$$

Therefore, the mean of the observations is related to the canonical parameter θ. The variance depends on the canonical parameter (and hence on the mean) as well as on the dispersion parameter ϕ. Other representations of the exponential family of distributions can be found in Dempster, Laird and Rubin (1977) and Zeger and Karim (1991). For example, the Poisson, binomial, and normal distributions can be expressed in the form of the exponential family of distributions as shown below (McCullagh and Nelder, 1989, p.30):

Table 1. Poisson, binomial and normal distributions as members of the exponential family

Distribution	Notation	ϕ	θ	$b(\theta)$	$c(y,\phi)$
Poisson	$p(\lambda)$	1	$\log\lambda$	$\exp(\theta)$	$-\log y!$
Binomial	$B(n,\pi)/n$	$1/n$	$\log\left(\frac{\pi}{1-\pi}\right)$	$\log(1+e^{\theta})$	$\log\binom{n}{ny}$
Normal	$N(\mu,\sigma^2)$	σ^2	μ	$\theta^2/2$	$-\frac{1}{2}[\frac{y^2}{\phi}+\log(2\pi\phi)]$

Generalized linear models

The linear model based on the exponential family of distributions is introduced by Nelder and Wedderburn (1972) as the generalized linear model. The model is defined by N independent random variables $Y_1, Y_2, ..., Y_N$ sampled from the exponential family of distribution with corresponding canonical parameters $\theta_1, \theta_2, ..., \theta_N$ and a common dispersion parameter ϕ. The joint probability density function of the Y's (unweighted) is therefore

$$f(y_1, y_2, ..., y_N | \theta_1, \theta_2, ..., \theta_N, \phi) = \exp\{\sum_{i=1}^{N} [y_i \theta_i - b(\theta_i)] / \phi + \sum_{i=1}^{N} c(y_i, \phi)]\}. \quad (2.3)$$

In actual modeling, the expected value μ_i of Y_i is predicted by a linear combination of explanatory variables $x_1, x_2, ..., x_N$ as follows:

$$g(\mu_i) = \mathbf{x}_i' \beta, \quad (2.4)$$

where $g(.)$ is a monotone link function (Aitkin et al., 1989, p.76), β is the p × 1 vector of parameters and $\mathbf{x}_i$ is the vector of explanatory variables. The link function relates the expected values of Y to its linear predictors through a transformation function. The principal usage of the link function is to map the limited domain of the mean of the observations onto the real line. For example, the domain of the mean number of counts from a Poisson distribution is never negative, the logarithmic

function puts the mean onto the real number domain. The commonly used link functions for the Poisson, binomial and normal distributions are the logarithm, logit and identity functions respectively. These link functions have the property of $\theta_i = g(\mu_i) = \mathbf{x}_i'\beta$ and are named as the canonical links.

---

[1] Personal communication suggested by Dr. Stephen W. Raudenbush.

CHAPTER III

STATISTICAL MODEL AND ESTIMATION STRATEGY

A multilevel generalized linear model

In the multilevel framework, an observation is represented by y_{ij}, i=1,2,... n_j; j=1,2,...,J. The subscript i refers to the units in level-one, e.g., pupils, and j refers to the units in level-two, e.g., schools. P is the number of level-one predictors and Q_p is the number of level-two predictors for the pth random coefficient in level-one. Conditional on a P+1 × 1 random effect vector $\mathbf{u}_j$, the observations are random samples from the exponential family of distributions with density (McCullagh and Nelder, 1989):

$$f(y_{ij}|\mathbf{u}_j,\theta_{ij},\phi)=\exp[(y_{ij}\theta_{ij}-b(\theta_{ij}))/\phi+c(y_{ij},\phi)]. \qquad (3.1)$$

The conditional moments are

$$\mu_{ij}\equiv E(y_{ij}|\mathbf{u}_j)=b'(\theta_{ij}),\ Var(y_{ij}|\mathbf{u}_j)=b''(\theta_{ij})\phi. \qquad (3.2)$$

Let the level-1 equation be:

$$g(\mu_{ij})\equiv\eta_{ij}=\beta_{0j}+\beta_{1j}x_{1ij}+...+\beta_{Pj}x_{Pij} \qquad (3.3)$$

$$= \begin{pmatrix} 1 \\ x_{1ij} \\ x_{2ij} \\ \vdots \\ x_{Pij} \end{pmatrix}^T \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \vdots \\ \beta_{Pj} \end{pmatrix}.$$

The level-2 equations are

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} w_{01j} + \ldots + \gamma_{0Q_0} w_{0Q_0 j} + u_{0j}. \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} w_{11j} + \ldots + \gamma_{1Q_1} w_{1Q_1 j} + u_{1j}. \\ &\quad . \quad . \quad . \quad . \quad . \quad . \\ \beta_{Pj} &= \gamma_{P0} + \gamma_{P1} w_{P1j} + \ldots + \gamma_{PQ_p} w_{PQ_p j} + u_{Pj}. \end{aligned} \tag{3.4}$$

The above can be expressed as

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \vdots \\ \beta_{Pj} \end{pmatrix} = \begin{pmatrix} 1, w_{01j}, \ldots, w_{0Q_0 j} & 0 & \cdots & 0 \\ 0 & 1, w_{11j}, \ldots, w_{1Q_1 j} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1, w_{P1j}, \ldots, w_{PQ_p j} \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \vdots \\ \gamma_{0Q_0} \\ \hline \gamma_{10} \\ \gamma_{11} \\ \vdots \\ \gamma_{1Q_1} \\ \hline \vdots \\ \hline \gamma_{P0} \\ \gamma_{P1} \\ \vdots \\ \gamma_{PQ_p} \end{pmatrix} + \begin{pmatrix} u_{0j} \\ u_{1j} \\ \vdots \\ u_{Pj} \end{pmatrix}.$$

In matrix notation, the level-1 (3.3) and level-2 (3.4) equations can be rewritten compactly as

$$g(\mu_{ij}) \equiv \eta_{ij} = \mathbf{x}_{ij}' \beta_j, \tag{3.5}$$

$$\beta_j = \mathbf{w}_j\gamma + \mathbf{u}_j. \qquad (3.6)$$

The matrix $\mathbf{w}_j$ has dimension equal to (P+1) $\times$ $\sum_{p=0}^{P}(Q_p+1)$ and the vector γ's dimension is equal to $\sum_{p=0}^{P}(Q_p+1) \times 1$. If Q_p=Q for all p, then $\sum_{p=0}^{P}(Q_p+1)$=(P+1)(Q+1).

Combining the equations (3.5) and (3.6), we have

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}'\mathbf{w}_j\gamma + \mathbf{x}_{ij}'\mathbf{u}_j \qquad (3.7)$$

$$= \mathbf{A}_{ij}\gamma + \mathbf{B}_{ij}\mathbf{u}_j \quad \text{(say)}. \qquad (3.8)$$

The random effect vector $\mathbf{u}_j$ is assumed to vary with a multivariate normal distribution with zero expectations and variance-covariance matrix T with dimension P+1 $\times$ P+1.

The principal objective of the statistical estimation for this model is to estimate the P+1 fixed effects γ and the symmetric matrix T with (P+1)(P+2)/2 unique variance-covariance components.

As noted earlier, direct maximization of the marginal likelihood function log $f(\mathbf{Y}|\gamma,\mathbf{T},\mathbf{u})$ is not feasible because the number of unknown vectors $\mathbf{u}_j$, which depends on the number of schools, can be very large. Alternatively, one can integrate out

the random effects **u** and maximize the log likelihood by partial differentiation. Let $\varphi=(\gamma,\mathbf{T})$, we have

$$\frac{\partial}{\partial\varphi}\log f(\mathbf{Y}|\varphi)=\frac{\partial}{\partial\varphi}\log\int f(\mathbf{Y}|\varphi,\mathbf{u})f(\mathbf{u}|\varphi)d\mathbf{u}\,.^{1} \qquad (3.9)$$

However, except for normal distributions, the above integral cannot be solved analytically. High dimensional numerical integration of the above requires the knowledge of φ the unknown parameter. Thus we need some kind of iterative procedure to maximize the log likelihood.

Maximum likelihood estimation via the Monte Carlo EM algorithm

Treating the random effects **u** as missing data, the EM algorithm (Dempster, Laird and Rubin, 1977) can be used to maximize the analytically intractable marginal log likelihood function $f(\mathbf{Y}|\gamma,\mathbf{T})$ indirectly through iterations derived from a combination of the E (expectation) step and the M (maximization) step.

The E step computes a Q function which is the conditional expectation of $\log f(\mathbf{Y},\mathbf{u}|\gamma,\mathbf{T})$, the 'complete data' log likelihood, with respect to the distribution of the 'missing data' **u**

conditioned on the observed data **Y** and the parameter values $\varphi^{(l)}$ at the l^{th} iteration.

$$Q(\varphi;\varphi^{(l)}) = \int \log f(\mathbf{Y},\mathbf{u}|\varphi) f(\mathbf{u}|\varphi^{(l)},\mathbf{Y}) d\mathbf{u}. \qquad (3.10)$$

The Q function is a function of φ and is maximized to obtain the parameter values $\varphi^{(l+1)}$ for the $l+1^{\text{th}}$ E step. That is, the M step solves the following equation:

$$\frac{\partial}{\partial\varphi} Q(\varphi;\varphi^{(l)}) = 0. \qquad (3.11)$$

Iteration between the E and M steps continues until the parameter values converge. Dempster, Laird and Rubin (1977) have shown that the EM algorithm increases the marginal likelihood $f(\mathbf{Y}|\gamma,\mathbf{T})$ for every iteration and the algorithm converges to at least a local maxima. Wu (1983) discovered an error in the proof for the convergence of the EM algorithm in Dempster et al. (1977). However, Wu (1983) shows that under mild regularity conditions the EM sequence converges to the maximum likelihood estimate. Wu (1983, p.95) contends that Dempster et al.'s (1977) results on the monotonicity of likelihood sequence and the convergence rate of the EM sequence remain valid. In other words, employment of the EM algorithm by itself guarantees maximum likelihood estimation if

the starting values are close to the maxima and convergence is achieved.

In practice, the integral for the expectation step may be hard to obtain. One can use the so called Monte Carlo EM algorithm (Wei and Tanner, 1990) to bypass the high dimensional multiple integral. The Monte Carlo EM algorithm is an EM algorithm with the expectation in the E-step approximated by a Monte Carlo integration.

The conditional expectation of the complete data log likelihood is now estimated by the arithmetic mean of M realizations of the complete data log likelihood with the unobserved random effects substituted by their simulated values:

$$Q(\varphi;\varphi^{(l)}) \cong \frac{1}{\mathbf{M}} \sum_{m=1}^{\mathbf{M}} \log f(\mathbf{Y}, \mathbf{u}_m | \varphi), \qquad (3.12)$$

where $\mathbf{u}_m$ is generated from the distribution $f(\mathbf{u}|\varphi^{(l)}, \mathbf{Y})$. It should be noted that the accuracy of the above approximation can be improved to any degree by increasing the number of random samples in the calculation.

---

[1] For simplicity of notation, the domain of the multiple integral in this dissertation is not printed. Interested readers can refer to the notation used in Appendix A1 for a more rigorous presentation.

CHAPTER IV

PARAMETER ESTIMATION

Convergence to maximum likelihood estimates is achieved by alternative implementations of the E-step and M-step of the EM algorithm. The E-step consists of computing the conditional expectation of the 'complete data' log likelihood function below:

The expectation step of the EM algorithm

$$Q(\varphi;\varphi^{(l)}) = \int \log f(\mathbf{Y},\mathbf{u}|\varphi) f(\mathbf{u}|\varphi^{(l)},\mathbf{Y}) d\mathbf{u} \qquad (4.1)$$

$$= \int [\sum_{j=1}^{J} \log f(\mathbf{Y}_j,\mathbf{u}_j|\varphi)] f(\mathbf{u}|\varphi^{(l)},\mathbf{Y}) d\mathbf{u}$$

$$= \sum_{j=1}^{J} \int \log f(\mathbf{Y}_j,\mathbf{u}_j|\varphi) f(\mathbf{u}|\varphi^{(l)},\mathbf{Y}) d\mathbf{u}. \qquad (4.2)$$

By applying the results from Appendix A1, Corollary 1, expression (4.2) becomes

$$\sum_{j=1}^{J} \int \log f(\mathbf{Y}_j,\mathbf{u}_j|\varphi) f(\mathbf{u}_j|\varphi^{(l)},\mathbf{Y}) d\mathbf{u}_j$$

$$= \sum_{j=1}^{J} \int \log f(\mathbf{Y}_j,\mathbf{u}_j|\varphi) f(\mathbf{u}_j|\varphi^{(l)},\mathbf{Y}_j) d\mathbf{u}_j, \qquad (4.3)$$

due to the independence between level-2 units.

Applying Bayes' theorem,

$$Q(\varphi;\varphi^{(l)}) = \sum_{j=1}^{J} \int \log f(\mathbf{Y}_j, \mathbf{u}_j|\varphi) \frac{f(\mathbf{Y}_j|\mathbf{u}_j,\varphi^{(l)}) f(\mathbf{u}_j|\varphi^{(l)})}{f(\mathbf{Y}_j|\varphi^{(l)})} d\mathbf{u}_j ,$$

$$= \sum_{j=1}^{J} \frac{1}{C_j} \int \log f(\mathbf{Y}_j, \mathbf{u}_j|\varphi) f(\mathbf{Y}_j|\mathbf{u}_j,\varphi^{(l)}) f(\mathbf{u}_j|\varphi^{(l)}) d\mathbf{u}_j . \qquad (4.4)$$

The constant $C_j \equiv f(\mathbf{Y}_j|\varphi^{(l)}) = \int f(Y_j|\mathbf{u}_j,\varphi^{(l)}) f(\mathbf{u}_j|\varphi^{(l)}) d\mathbf{u}_j$ is computed by the Monte Carlo integration method (Rubinstein, 1981).

$$C_j \approx \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)}), \qquad (4.5)$$

where $\mathbf{u}_{1j}, \mathbf{u}_{2j}, \ldots, \mathbf{u}_{Mj} \sim f(\mathbf{u}_j|\varphi^{(l)})$, which are M samples of simulations from the multivariate normal distribution.

Applying the Monte Carlo integration technique again to (4.4):

$$Q(\varphi;\varphi^{(l)}) \approx \sum_{j=1}^{J} \frac{1}{MC_j} \sum_{m=1}^{M} \log f(\mathbf{Y}_j, \mathbf{u}_{mj}|\varphi) f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)}) \qquad (4.6)$$

$$= \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{1}{MC_j} \log f(\mathbf{Y}_j, \mathbf{u}_{mj}|\varphi) f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})$$

$$\propto \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{1}{C_j} [\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi) + \log f(\mathbf{u}_{mj}|\varphi)] f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})$$

$$\propto \sum_{m=1}^{M} \sum_{j=1}^{J} [\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\gamma) + \log f(\mathbf{u}_{mj}|\mathrm{T})] \psi_{mj} , \text{ where}$$

$$\psi_{mj} = \frac{f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})}{C_j} \Bigg/ \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})}{C_j} . \qquad (4.7)$$

The M-step maximizes Q(.) to give the new parameter values $\varphi^{(l+1)}$. Parameters $\gamma^{(l+1)}$ are obtained by solving the equations

$$\sum_{m=1}^{M}\sum_{j=1}^{J}\partial\frac{\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\gamma)}{\partial\gamma}\psi_{mj}=0 \qquad (4.8)$$

and $\mathbf{T}^{(l+1)}$ by solving

$$\sum_{m=1}^{M}\sum_{j=1}^{J}\partial\frac{\log f(\mathbf{u}_{mj}|\mathrm{T})}{\partial Vech\mathrm{T}}\psi_{mj}=0. \qquad (4.9)$$

The maximization step of the EM algorithm

The solution for maximizing an unweighted single level analogue of the likelihood function in equation (4.8) can be found in Aitkin et al. (1989, p.322-325) for the generalized linear models. Anderson and Hinde (1988, p.3851-3852) have also derived the solution to a random intercept multilevel model with weighted likelihood using the numerical Gaussian quadrature method. In this dissertation, the solution to a random coefficient multilevel model with weighted likelihood function involving Monte Carlo integrations is derived below.

Estimation for the fixed-effect parameters

The maximization of Q(.) with respect to γ is equivalent to maximizing a weighted sum of log likelihoods of the conditional distribution of the outcome variable given the random effect variables $\mathbf{u}_{mj}$. The inclusion of the outer summation sign in subscript m is equivalent to expanding the original data set by M times with the simulated random effects $\mathbf{u}_{mj}$ appropriately substituted (c.f. Anderson and Hinde, 1988). We have to maximize the following weighted log likelihood function of the exponential family with respect to γ:

$$L(\gamma) = \sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\{[y_{ij}\theta_{mij} - b(\theta_{mij})]/\phi + c(y_{ij},\phi)\}\psi_{mj}, \qquad (4.10)$$

with $g(\mu_{mij}) = \theta_{mij} = \eta_{mij} = \mathbf{x}_{ij}'\mathbf{w}_j\gamma + \mathbf{x}_{ij}'\mathbf{u}_{mj} \equiv \mathbf{A}_{ij}\gamma + \mathbf{B}_{ij}\mathbf{u}_{mj}$ using the canonical link function. The derivative of (4.10) w.r.t. γ is

$$\frac{\partial L}{\partial \gamma} = \sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}[y_{ij} - b'(\theta_{mij})]\frac{\partial \theta_{mij}}{\partial \gamma}\psi_{mj}/\phi. \qquad (4.11)$$

For exponential families, $\mu = E(y) = b'(\theta)$ and $Var(y) = \phi b''(\theta)$, thus

$$\frac{\partial \theta_{mij}}{\partial \gamma} = \frac{\partial \theta_{mij}}{\partial \mu_{mij}} \cdot \frac{\partial \mu_{mij}}{\partial \eta_{mij}} \cdot \frac{\partial \eta_{mij}}{\partial \gamma}$$

$$= \frac{1}{b''(\theta_{mij})} \cdot \frac{1}{g'(\mu_{mij})} \cdot \mathbf{A}_{ij}. \qquad (4.12)$$

Therefore $\frac{\partial L}{\partial \gamma} = \sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{(y_{ij} - \mu_{mij})\mathbf{A}_{ij}\psi_{mj}}{V_{mij}g'(\mu_{mij})}$. $\qquad (4.13)$

To form a Fisher's scoring algorithm (c.f. Seber, 1989, p.685) as in the GLIM program (Aitkin et al., 1989), we need the expected value of the second derivative of the weighted log likelihood:

By results proved in Appendix A2,

$$E\left(\frac{\partial^2 L}{\partial\gamma\partial\gamma'}\right) = -\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{\psi_{mj}}E\left(\frac{\partial L_{mij}}{\partial\gamma}\frac{\partial L_{mij}}{\partial\gamma'}\right) \tag{4.14}$$

$$= -\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{\psi_{mj}}\frac{E(y_{ij}-\mu_{mij})^2\mathbf{A}_{ij}\mathbf{A}_{ij}'\psi_{mj}^2}{V_{mij}^2[g'(\mu_{mij})]^2}$$

$$= -\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\mathbf{A}_{ij}\mathbf{A}_{ij}'\psi_{mj}}{V_{mij}[g'(\mu_{mij})]^2}$$

$$= -\sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\omega_{mij}\mathbf{A}_{ij}\mathbf{A}_{ij}' \tag{4.15}$$

$$= -\mathbf{A}'\Omega\mathbf{A} \tag{4.16}$$

where $\mathbf{A} = \{[\mathbf{A}_1,\ldots,\mathbf{A}_J]_{m=1},\ldots,[\mathbf{A}_1,\ldots,\mathbf{A}_J]_{m=M}\}'$ with $\mathbf{A}_{ij} = \{\mathbf{A}_{1j},\ldots,\mathbf{A}_{n_jJ}\}'$ and Ω is a diagonal weight matrix with elements $\omega_{mij} = \frac{\psi_{mj}}{V_{mij}[g'(\mu_{mij})]^2}$ (4.17).

For the l^{th} iteration of the Fisher's scoring algorithm, the new estimate $\gamma^{(l+1)}$ is given by

$$\gamma^{(l+1)} = \gamma^{(l)} + (\mathbf{A}'\Omega^{(l)}\mathbf{A})^{-1}\frac{\partial L}{\partial\gamma}^{(l)}. \tag{4.18}$$

Now
$$\frac{\partial L}{\partial\gamma} = \sum_{m=1}^{M}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\mu_{mij})g'(\mu_{mij})\mathbf{A}_{ij}\omega_{mij} \tag{4.19}$$

$$= A'\Omega^{(l)}\delta \tag{4.20}$$

where $\delta = [(\delta'_{111}...\delta'_{1n_11})...(\delta'_{11J}...\delta'_{1n_JJ})],...,[(\delta'_{M11}...\delta'_{Mn_11})...(\delta'_{M1J}...\delta'_{Mn_JJ})]\}'$ with

$\delta_{mij} = (y_{ij} - \mu_{mij})g'(\mu_{mij})$ (4.21).

Thus
$$\gamma^{(l+1)} = \gamma^{(l)} + (\mathbf{A}'\Omega^{(l)}\mathbf{A})^{-1}\mathbf{A}'\Omega^{(l)}\delta^{(l)}$$
$$= (\mathbf{A}'\Omega^{(l)}\mathbf{A})^{-1}(\mathbf{A}'\Omega^{(l)}\mathbf{A}\gamma^{(l)} + \mathbf{A}'\Omega^{(l)}\delta^{(l)})$$
$$= (\mathbf{A}'\Omega^{(l)}\mathbf{A})^{-1}\mathbf{A}'\Omega^{(l)}\mathbf{Z}^{(l)} \qquad (4.22)$$

where $\mathbf{Z}^{(l)} = \mathbf{A}\gamma^{(l)} + \delta^{(l)}$ (4.23).

The scoring algorithm can now be viewed as an iterative reweighted least square algorithm (IRLS). The 'adjusted dependent variable' **Z** is regressed on explanatory variables **A** with diagonal weight matrix Ω. The new estimate of $\gamma^{(l+1)}$ can be found by solving the following system of equations using ordinary Gaussian elimination (Johnson and Riess, 1988):

$$(\mathbf{A}'\Omega^{(l)}\mathbf{A})\gamma^{(l+1)} = \mathbf{A}'\Omega^{(l)}\mathbf{Z}^{(l)}. \qquad (4.24)$$

Estimation for the variance component parameters

The new variance component parameters $\mathbf{T}^{(l+1)}$ belong to the multivariate normal distribution and are estimated by the sum of the weighted squares and cross-products of the simulated random effects. This is based on an extension of the standard maximum likelihood estimate (Seber, 1984) for the dispersion

matrix of the multivariate normal distribution. The solution to equation (4.9) becomes

$$T^{(l+1)} = \sum_{m=1}^{M}\sum_{j=1}^{J} \psi_{mj}\mathbf{u}_{mj}\mathbf{u}_{mj}' . \qquad (4.25)$$

The validity of the above equation is supported by the following theorem:

Theorem 1. Let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_J$ be J independent p-dimensional vectors of random variables and each $\mathbf{u}_j$ have a non-singular MVN distribution with zero expectation and a positive definite variance-covariance matrix T. Let $\psi_j \geq 0 \ \forall \ j$ and r be the count of $\psi_j > 0$. For $r \geq p$ and subject to $\sum_{j=1}^{J}\psi_j = 1$, the weighted log likelihood $\sum_{j=1}^{J}\psi_j \log f(\mathbf{u}_j|T)$ is maximized uniquely at $T = \sum_{j=1}^{J}\psi_j \mathbf{u}_j\mathbf{u}_j'$.

Proof: The following two lemmas are required for the proof of this theorem.

Lemma 1: Consider the matrix function f, where $f(T) = \log|T| + tr[T^{-1}\Phi]$. If Φ is positive definite, subject to T being positive definite, $f(T)$ is minimized uniquely at $T = \Phi$ (Watson, 1964).

Lemma 2: Let $\mathbf{U}'=(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_J)$, where the $\mathbf{u}_j$ are J independent p-dimensional vectors of random variables, and let Ψ be a positive semidefinite $J\times J$ matrix of rank r ($r\geq p$). Suppose that for each j and all $\mathbf{b}$ ($\mathbf{b}\neq 0$) and c, $\mathrm{pr}(\mathbf{b}'\mathbf{u}_j=c)=0$. Then $\mathbf{U}'\Psi\mathbf{U}$ is positive definite with probability 1 (Das Gupta, 1971).

Let $c=-\frac{p}{2}\sum_j \psi_j \log 2\pi$, then

$$\sum_j \psi_j \log f(\mathbf{u}_j|\mathrm{T}) = c-\frac{1}{2}\sum_j \psi_j \log|\mathrm{T}|-\frac{1}{2}\sum_j \psi_j \mathbf{u}_j'\mathrm{T}^{-1}\mathbf{u}_j$$

$$= c-\frac{1}{2}\log|\mathrm{T}|\sum_j \psi_j-\frac{1}{2}\sum_j \psi_j tr[\mathbf{u}_j'\mathrm{T}^{-1}\mathbf{u}_j]$$

$$= c-\frac{1}{2}\log|\mathrm{T}|-\frac{1}{2}tr\left[\mathrm{T}^{-1}\sum_j \psi_j \mathbf{u}_j\mathbf{u}_j'\right]$$

$$= c-\frac{1}{2}\left(\log|\mathrm{T}|+tr\left[\mathrm{T}^{-1}\sum_j \psi_j \mathbf{u}_j\mathbf{u}_j'\right]\right). \qquad (4.26)$$

Let $\mathbf{U}'=(\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_J)$ and $\Psi=\mathrm{diag}(\psi_1,\psi_2,\ldots,\psi_J)$, then $\sum_{j=1}^{J}\psi_j\mathbf{u}_j\mathbf{u}_j'$ = $\mathbf{U}'\Psi\mathbf{U}$. Ψ is positive semidefinite of rank r because r is the count of $\psi_j>0$. Since $\mathbf{u}_j \sim \mathrm{N}_p(0,\mathrm{T})$ and T is positive definite, then for all $\mathbf{b}$ ($\mathbf{b}\neq 0$) and c, $\mathbf{b}'\mathbf{u}_j \sim \mathrm{N}_p(0, \mathbf{b}'\mathrm{T}\mathbf{b})$. This distribution

is nondegenerate as $\mathbf{b}'\mathbf{T}\mathbf{b}>0$. Hence $pr(\mathbf{b}'\mathbf{u}_j=c)=0$. Therefore for $r \geq p$, $\sum_{j=1}^{J} \psi_j \mathbf{u}_j \mathbf{u}_j'$ is positive definite with probability 1 by lemma 2.

Directly applying lemma 1 to (4.26), $\sum_{j=1}^{J} \psi_j \log f(\mathbf{u}_j|\mathrm{T})$ is maximized uniquely at $\mathrm{T} = \sum_{j=1}^{J} \psi_j \mathbf{u}_j \mathbf{u}_j'$. □

Applying the above theorem, the weighted likelihood of the multivariate normal distribution $\sum_m \sum_j \log f(\mathbf{u}_{mj}|\mathrm{T}) \psi_{mj}$ in equation (4.9) is maximized by $\mathrm{T}^{(l+1)} = \sum_m \sum_j \psi_{mj} \mathbf{u}_{mj} \mathbf{u}_{mj}'$ because it could be easily verified that $\sum_m \sum_j \psi_{mj} = 1$.

In practice, because the M-step needs only one iteration of an IRLS, this EM algorithm becomes a generalized EM (GEM) algorithm (Dempster et al., 1977). The fixed parameters are thus expressed as a weighted least squares expression (4.22). Applying the theorem above, a simple closed form solution exists for the variance component parameters. Hence, the advantage of using the EM algorithm, which often involves closed form solutions in the M-step, is maintained.

The convergence to the maximum likelihood estimate can be monitored by tracing through the expected complete data log likelihood function, $Q(.)$. The reason is that every step of the EM algorithm will increase the log likelihood function through the increase of $Q(.)$ (Dempster, Laird and Rubin, 1977). With Monte Carlo EM algorithm, the complete data likelihood or any other estimated iterates converge in a stochastic fluctuating manner. By plotting the iterative values of $Q(.)$, the algorithm is stopped after the $Q(.)$ function has risen to a maximum plateau. In an alternative parallel way, one can also monitor convergence by noting the increment of the model parameters by plotting parameter values against the number of iteration. After convergence, the algorithm can be allowed to run for more steps with an increased number of Monte Carlo samples so as to decrease the stochastic variation of the estimated parameters.

CHAPTER V

VARIANCE OF PARAMETER ESTIMATES

The variance for the estimates of the fixed parameters and variance components can be estimated using the inverse of the information matrix (see Aitkin et al., 1989, p.81):

$$V(\hat{\varphi}) \approx \mathbf{I}^{-1} \equiv \left(-\frac{\partial^2 \log f(\mathbf{Y}|\varphi)}{\partial \varphi^2} \right)^{-1} \Bigg|_{\hat{\varphi}} . \qquad (5.1)$$

The diagonal elements of V correspond to the large sample variances of the maximum likelihood estimator for the parameters φ. Louis (1982) shows that the Fisher information matrix of the incomplete data likelihood can be expressed as a function of the conditional expectations of the complete data information matrix and the cross-products of the complete data score function. In the terminology of this study, Louis's (1982) result can be re-expressed as:

$$-\frac{\partial^2 \log f(\mathbf{Y}|\varphi)}{\partial \varphi \partial \varphi'} = -E[\frac{\partial^2 \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \varphi \partial \varphi'}] - E[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \varphi} \frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \varphi'}]$$

$$+\frac{\partial \log f(\mathbf{Y}|\varphi)}{\partial \varphi} \frac{\partial \log f(\mathbf{Y}|\varphi)}{\partial \varphi'}, \qquad (5.2)$$

with the expectation taken with respect to $f(\mathbf{u}|\mathbf{Y},\varphi)$. The observed information matrix is thus

$$-\frac{\partial^2 \log f(\mathbf{Y}|\varphi)}{\partial\varphi\partial\varphi'}\bigg|_{\hat{\varphi}} = -E[\frac{\partial^2 \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\varphi\partial\varphi'}]\bigg|_{\hat{\varphi}} - E[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\varphi}\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\varphi'}]\bigg|_{\hat{\varphi}} \quad (5.3)$$

for the last term in equation (5.2) above vanishes at the maximum likelihood estimator $\hat{\varphi}$.

This method makes use of the complete data distribution $f(\mathbf{Y},\mathbf{u}|\varphi)$ rather than the marginal distribution $f(\mathbf{Y}|\varphi)$ which does not have an analytic derivation. Similar to its application in the EM algorithm, the method of Monte Carlo can be applied to (5.3) to ease the integration problem (Tanner, 1991, p.39). Since maximum likelihood estimators are asymptotically normally distributed, the estimated standard errors can be easily used to create confidence intervals and hypothesis tests.

Estimation for the variance of the fixed effect estimates

The components of equation (5.3) required for the evaluation the variance of the fixed effect estimates are derived as follows:

$$E[\frac{\partial^2 \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\gamma\partial\gamma'}] = \int \sum_{j=1}^{J} \frac{\partial^2 \log f(\mathbf{Y}_j,\mathbf{u}_j|\varphi)}{\partial\gamma\partial\gamma'} f(\mathbf{u}|\mathbf{Y},\varphi)d\mathbf{u}$$

$$\approx \sum_{m=1}^{M}\sum_{j=1}^{J} \frac{1}{\mathbf{M}C_j} \frac{\partial^2 \log f(\mathbf{Y}_j,\mathbf{u}_{mj}|\varphi)}{\partial\gamma\partial\gamma'} f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi}),$$

(see Appendix A3)

$$= \sum_{m=1}^{M}\sum_{j=1}^{J} \frac{1}{MC_j}\left[\frac{\partial^2 \log f(\mathbf{Y_j}|\mathbf{u_{mj}},\varphi)}{\partial\gamma\partial\gamma'} + \frac{\partial^2 \log f(\mathbf{u_{mj}}|\varphi)}{\partial\gamma\partial\gamma'}\right] f(\mathbf{Y_j}|\mathbf{u_{mj}},\hat{\varphi})$$

$$= \sum_{m=1}^{M}\sum_{j=1}^{J} \frac{1}{MC_j}\left[\frac{\partial^2 \log f(\mathbf{Y_j}|\mathbf{u_{mj}},\varphi)}{\partial\gamma\partial\gamma'}\right] f(\mathbf{Y_j}|\mathbf{u_{mj}},\hat{\varphi}) \qquad (5.4)$$

with $\dfrac{\partial^2 \log f(\mathbf{Y_j}|\mathbf{u_{mj}},\varphi)}{\partial\gamma\partial\gamma'} = \sum_{i=1}^{n_j} -b''(\theta_{mij})\dfrac{\partial\theta_{mij}}{\partial\gamma}\dfrac{\partial\theta_{mij}}{\partial\gamma'}/\phi = -\sum_{i=1}^{n_j}\dfrac{\mathbf{A_{ij}}\mathbf{A_{ij}}'}{V_{mij}[g'(\mu_{mij})]^2}.$ (5.5)

Moreover,

$$E\left[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\gamma'}\right]$$

$$= E\left[\left(\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial\gamma} + \frac{\partial \log f(\mathbf{u}|\varphi)}{\partial\gamma}\right)\left(\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial\gamma'} + \frac{\partial \log f(\mathbf{u}|\varphi)}{\partial\gamma'}\right)\right]$$

$$= E\left[\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial\gamma'}\right]$$

$$= E\left[\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma}\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma'}\right]$$

$$= E\left[\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma'}\right]$$

$$+E\left[\sum_{j,j'=1;j\neq j'}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y_{j'}}|\mathbf{u_{j'}},\varphi)}{\partial\gamma'}\right]$$

$$= \int\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma'} f(\mathbf{u_j}|\mathbf{Y_j},\varphi)d\mathbf{u_j}$$

$$+\int\sum_{j,j'=1;j\neq j'}^{J}\frac{\partial \log f(\mathbf{Y_j}|\mathbf{u_j},\varphi)}{\partial\gamma}\frac{\partial \log f(\mathbf{Y_{j'}}|\mathbf{u_{j'}},\varphi)}{\partial\gamma'} f(\mathbf{u_j}|\mathbf{Y_j},\varphi)f(\mathbf{u_{j'}}|\mathbf{Y_{j'}},\varphi)d\mathbf{u_j}d\mathbf{u_{j'}}$$

with the latter term equal to (c.f. Appendix A1)

$$\sum_{j,j'=1;j\neq j'}^{J}\int\frac{\partial\log f(\mathbf{Y}_j|\mathbf{u}_j,\varphi)}{\partial\gamma}f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j\int\frac{\partial\log f(\mathbf{Y}_{j'}|\mathbf{u}_{j'},\varphi)}{\partial\gamma'}f(\mathbf{u}_{j'}|\mathbf{Y}_{j'},\varphi)d\mathbf{u}_{j'} \qquad (5.6)$$

$$=\sum_{j,j'=1;j\neq j'}^{J}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma}\frac{\partial\log f(\mathbf{Y}_{j'}|\varphi)}{\partial\gamma'},\ \text{(Tanner, 1991)}^{1} \qquad (5.7)$$

$$=\sum_{j=1}^{J}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma}\sum_{j'=1}^{J}\frac{\partial\log f(\mathbf{Y}_{j'}|\varphi)}{\partial\gamma'}-\sum_{j=1}^{J}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma'}\ ^{2} \qquad (5.8)$$

with $\sum_{j}^{J}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma}\Big|_{\varphi=\hat{\varphi}}\sum_{j}^{J}\frac{\partial\log f(\mathbf{Y}_j|\varphi)}{\partial\gamma}\Big|_{\varphi=\hat{\varphi}}=0\cdot 0=0$ at MLE of φ. Hence

at MLE of φ,

$$E\left[\frac{\partial\log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\gamma}\frac{\partial\log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial\gamma'}\right]_{\hat{\varphi}}$$

$$\approx\sum_{j=1}^{J}\sum_{m=1}^{M}\frac{1}{MC_j}\frac{\partial\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi)}{\partial\gamma}\frac{\partial\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi)}{\partial\gamma'}f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})$$

$$-\sum_{j=1}^{J}\left\{\left[\sum_{m=1}^{M}\frac{1}{MC_j}\frac{\partial\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi)}{\partial\gamma}f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right]\cdot\right.$$

$$\left.\left[\sum_{m=1}^{M}\frac{1}{MC_j}\frac{\partial\log f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi)}{\partial\gamma'}f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right]\right\}, \qquad (5.9)$$

(using Appendix A3 and footnote 2)

$$=\sum_{m=1}^{M}\sum_{j=1}^{J}\frac{1}{MC_j}\left(\sum_{i=1}^{n_j}\frac{(y_{ij}-\mu_{mij})\mathbf{A}_{ij}}{V_{mij}g'(\mu_{mij})}\right)\left(\sum_{i=1}^{n_j}\frac{(y_{ij}-\mu_{mij})\mathbf{A}_{ij}'}{V_{mij}g'(\mu_{mij})}\right)f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})$$

$$-\sum_{j=1}^{J}\left\{\left[\sum_{m=1}^{M}\frac{1}{MC_j}\left(\sum_{i=1}^{n_j}\frac{(y_{ij}-\mu_{mij})\mathbf{A}_{ij}}{V_{mij}g'(\mu_{mij})}\right)f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right]\cdot\right.$$

$$\left[\sum_{m=1}^{M}\frac{1}{MC_j}\left(\sum_{i=1}^{n_j}\frac{(y_{ij}-\mu_{mij})\mathbf{A}_{ij}'}{V_{mij}g'(\mu_{mij})}\right)f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right]\Bigg\}. \qquad (5.10)$$

Estimation for the variance of the variance component estimates

The variances of the covariance estimates are estimated in an analogous way as that for the fixed effects estimates in the preceeding section. Results follow by replacing the partial deferential of γ by that of τ which stands for any arbitrary element of the covariance matrix T.

From Longford (1987), we have the following two useful derivative formulas:

$$\frac{\partial}{\partial\tau}\log|\mathrm{T}| = tr\left(\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\right) \qquad (5.11)$$

and

$$\frac{\partial\mathrm{T}^{-1}}{\partial\tau} = -\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\mathrm{T}^{-1}. \qquad (5.12)$$

Let τ' be any arbitrary element of the covariance matrix T other than τ. Thus,

$$\frac{\partial^2}{\partial\tau\partial\tau'}\log|\mathrm{T}| = tr\left(\frac{\partial\mathrm{T}^{-1}}{\partial\tau'}\frac{\partial\mathrm{T}}{\partial\tau}\right) = tr\left(-\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau'}\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\right)$$

(5.13) and

$$\frac{\partial^2\mathrm{T}^{-1}}{\partial\tau\partial\tau'} = -\left[\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\left(-\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau'}\mathrm{T}^{-1}\right)+\left(-\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau'}\mathrm{T}^{-1}\right)\frac{\partial\mathrm{T}}{\partial\tau}\mathrm{T}^{-1}\right]$$

$$= \mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau'}\mathrm{T}^{-1}+\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau'}\mathrm{T}^{-1}\frac{\partial\mathrm{T}}{\partial\tau}\mathrm{T}^{-1}. \qquad (5.14)$$

Since $\log f(u_{mj}|\varphi) \propto -\frac{1}{2}\log|\mathrm{T}| - \frac{1}{2}u_{mj}'\mathrm{T}^{-1}u_{mj}$, therefore,

$$\frac{\partial \log f(u_{mj}|\varphi)}{\partial \tau} = -\frac{1}{2}\left[tr\left(\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau}\right) - u_{mj}'\left(\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau}\mathrm{T}^{-1}\right)u_{mj}\right], \qquad (5.15)$$

and
$$\frac{\partial^2 \log f(u_{mj}|\varphi)}{\partial \tau \partial \tau'} = -\frac{1}{2}\left[tr\left(-\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau'}\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau}\right)\right.$$

$$\left. + u_{mj}'\left(\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau}\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau'}\mathrm{T}^{-1} + \mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau'}\mathrm{T}^{-1}\frac{\partial \mathrm{T}}{\partial \tau}\mathrm{T}^{-1}\right)u_{mj}\right]. \qquad (5.16)$$

As in the variance estimation of the fixed effects, we need to calculate $E\left[\frac{\partial^2 \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau \partial \tau'}\right]$ and $E\left[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau'}\right]$.

Now,

$$E\left[\frac{\partial^2 \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau \partial \tau'}\right]$$

$$\approx \sum_{\mathbf{m}=1}^{\mathbf{M}}\sum_{\mathbf{j}=1}^{\mathbf{J}} \frac{1}{\mathbf{M}C_j}\left[\frac{\partial^2 \log f(\mathbf{Y_j}|\mathbf{u_{mj}},\varphi)}{\partial \tau \partial \tau'} + \frac{\partial^2 \log f(\mathbf{u_{mj}}|\varphi)}{\partial \tau \partial \tau'}\right] f(\mathbf{Y_j}|\mathbf{u_{mj}},\hat{\varphi})$$

$$= \sum_{\mathbf{m}=1}^{\mathbf{M}}\sum_{\mathbf{j}=1}^{\mathbf{J}} \frac{1}{\mathbf{M}C_j}\left[\frac{\partial^2 \log f(\mathbf{Y_j}|\mathbf{u_{mj}},\varphi)}{\partial \tau \partial \tau'}\right] f(\mathbf{Y_j}|\mathbf{u_{mj}},\hat{\varphi}). \qquad (5.17)$$

And,

$$E\left[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau'}\right]$$

$$= E\left[\left(\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial \tau} + \frac{\partial \log f(\mathbf{u}|\varphi)}{\partial \tau}\right)\left(\frac{\partial \log f(\mathbf{Y}|\mathbf{u},\varphi)}{\partial \tau'} + \frac{\partial \log f(\mathbf{u}|\varphi)}{\partial \tau'}\right)\right]$$

$$= E\left[\frac{\partial \log f(\mathbf{u}|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}|\varphi)}{\partial \tau'}\right]$$

$$= E\left[\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau'}\right]$$

$$= E\left[\sum_{j=1}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau'}\right] + E\left[\sum_{j,j'=1;j\neq j'}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}_{j'}|\varphi)}{\partial \tau'}\right]$$

$$= \int \sum_{j=1}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau'} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j$$

$$+\int \sum_{j,j'=1;j\neq j'}^{J}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}_{j'}|\varphi)}{\partial \tau'} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi) f(\mathbf{u}_{j'}|\mathbf{Y}_{j'},\varphi)d\mathbf{u}_j d\mathbf{u}_{j'}$$

with the latter term equal to (c.f. Appendix A1)

$$\sum_{j,j'=1;j\neq j'}^{J}\int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j \int \frac{\partial \log f(\mathbf{u}_{j'}|\varphi)}{\partial \tau'} f(\mathbf{u}_{j'}|\mathbf{Y}_{j'},\varphi)d\mathbf{u}_{j'} \qquad (5.18)$$

$$= \sum_{j,j'=1;j\neq j'}^{J}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{Y}_{j'}|\varphi)}{\partial \tau'}, \text{ (see Appendix A4)} \qquad (5.19)$$

$$= \sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}\sum_{j'=1}^{J}\frac{\partial \log f(\mathbf{Y}_{j'}|\varphi)}{\partial \tau'} - \sum_{j=1}^{J}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau'} \qquad (5.20)$$

with $\sum_{j}^{J}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}\Big|_{\varphi=\hat{\varphi}} \sum_{j}^{J}\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}\Big|_{\varphi=\hat{\varphi}} = 0\cdot 0 = 0$ at MLE of φ. Hence

at MLE of φ,

$$E\left[\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{Y},\mathbf{u}|\varphi)}{\partial \tau'}\right]\Big|_{\hat{\varphi}}$$

$$\approx \sum_{j=1}^{J}\sum_{m=1}^{M}\frac{1}{\mathbf{M}C_j}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau'} f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})$$

$$-\sum_{j=1}^{J}\left[\sum_{m=1}^{M}\frac{1}{\mathbf{M}C_j}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right]\left[\sum_{m=1}^{M}\frac{1}{\mathbf{M}C_j}\frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau'} f(\mathbf{Y}_j|\mathbf{u}_{mj},\hat{\varphi})\right] \qquad (5.21)$$

(using Appendex A3 and A4)

---

[1] Tanner (1991, p.37) shows an equivalent of
$$\int \frac{\partial \log f(\mathbf{Y}_j|\mathbf{u}_j,\varphi)}{\partial \gamma} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi) d\mathbf{u}_j = \frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \gamma}.$$

[2] Result follows from a simple identity:
$$\sum_{j=1}^{J} a_j \sum_{j'=1}^{J} b_{j'} = \sum_{j=1}^{J} a_j b_j + \sum_{j,j'=1, j\neq j'}^{J} a_j b_{j'}$$
for any real number $a_j, b_{j'}$ with $j, j' \in \{1,2,...,J\}$.

CHAPTER VI

EMPIRICAL PROPERTIES OF THE STATISTICAL MODEL

The formulas for the estimates of the parameters and the associated standard errors of the Multilevel Generalized Linear Model (MGLM) in the preceeding chapters were used in building up a computer program. Microsoft FORTRAN Version 5.1 was the software used in writing the program in FORTRAN. Sample FORTRAN subroutines for mathematical and matrix operations and random Poisson variate generations were obtained from Press (1989). The program was built, tested and implemented on a Gateway2000 IBM compatible personal computer at clock speed 33 MHz with an Intel 486 DX microprocessor. Random numbers are generated by the default random function from the above FORTRAN software. Multivariate normal variates are generated by the LU decomposition method given by Rubinstein (1981).

In the following sections, we take a closer look at the performance of the estimators as programmed. The purpose is not so much to prove absolute validity of the program as to draw insight from several program runs. Simulation tests for validity are left for the next chapter.

To facilitate the analysis, I limit the model to the One-Way Poisson model with a single random component, also known as the random intercept model. In other words, the parameter set will contain just a fixed effect γ (Gamma) and a univariate variance-covariance matrix T (Tau). Because Monte Carlo integration is used in the algorithms, the stochastic nature of the random variate generations will be revealed in the subsequent figures as minor oscillations along the path to convergence. The parameter values for γ and T are respectively -0.7 and 0.16. The data set generated from these parameters has level-2 units $J = 50$ and level-1 units $n_j = 20$ for all j. The number of samples used in the Monte Carlo integrations is 200. Two starting positions are attempted. One set of initial γ and T pair is (-1.5, 0.3) and the other set is (-0.1, 0.05).

Maximization of the likelihood function

Under normal conditions, the EM algorithm will increase the likelihood function to at least a local maximum through the iterations. It is instructive though to watch how the likelihood function approaches the maximum for some sample runs. As shown in Figure 1, the log likelihood functions from two different sets of starting values increase and merge to a plateau

in around 15 iterations. After 15 iterations, the log likelihood function rises from -1015.6 to -970.6 when the starting values for γ and T are respectively -1.5 and 0.3. For the other set of starting values for γ and T, being -0.1 and 0.05, the rise is from -1031.9 to

-970.8. The results in Figure 1 give evidence supporting the achievement of maximum likelihood. The shape of the two rising likelihoods also gives support that the likelihood function is increased with the number of EM iterations. The strict monotonic increase of the likelihood function is masked by the random fluctuations due to the Monte Carlo samples.

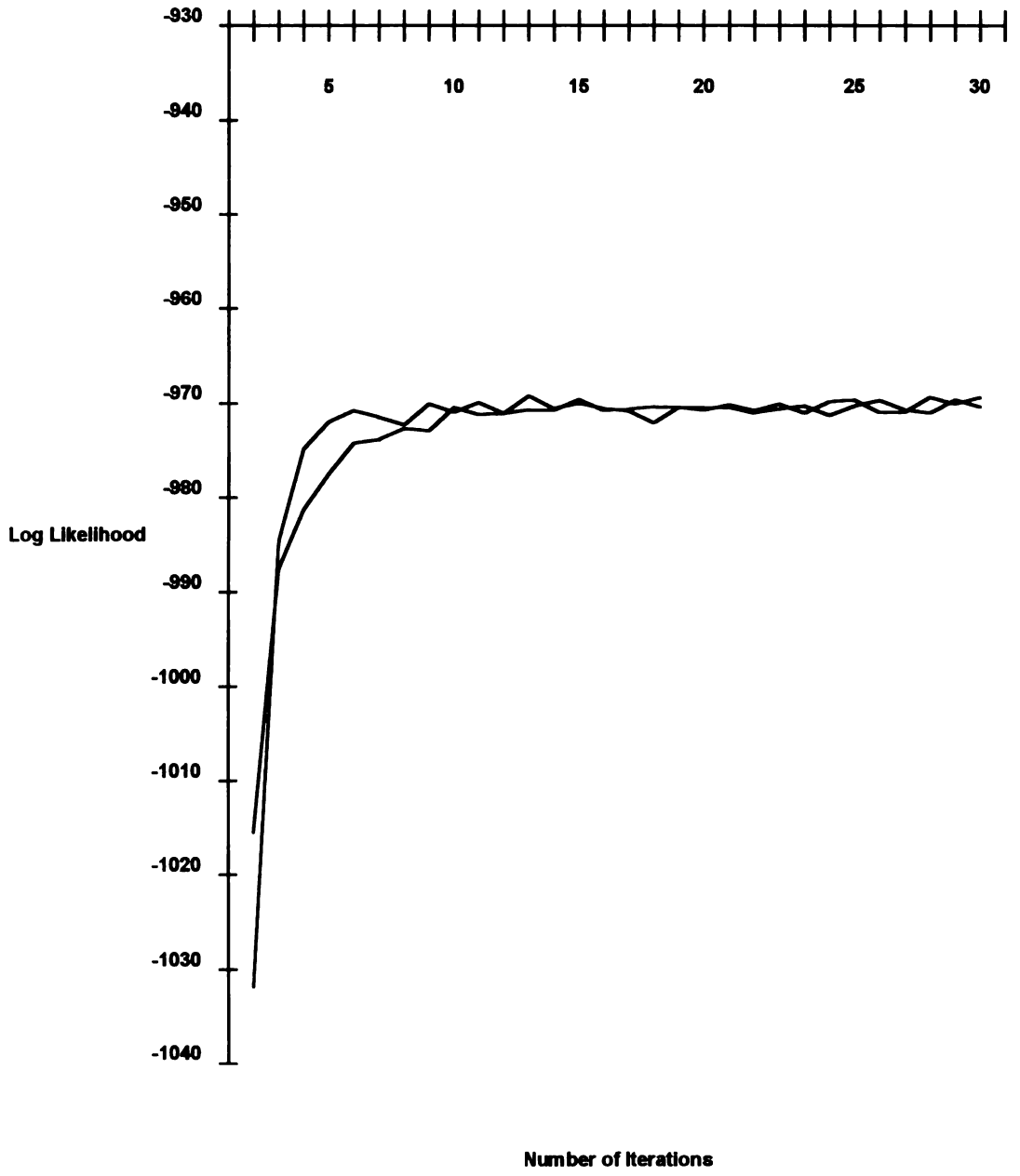
The log likelihood is calculated as follows:

$$\log f(\mathbf{Y}|\varphi^{(l)}) = \sum_{j=1}^{J} \log f(\mathbf{Y}_j|\varphi^{(l)})$$

$$= \sum_{j=1}^{J} \log C_j$$

$$\approx \sum_{j=1}^{J} \log\left(\frac{1}{M}\sum_{m=1}^{M} f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})\right). \qquad (6.1)$$

Figure 1. Log likelihood function as a function of the number of iterations for two sets of starting values

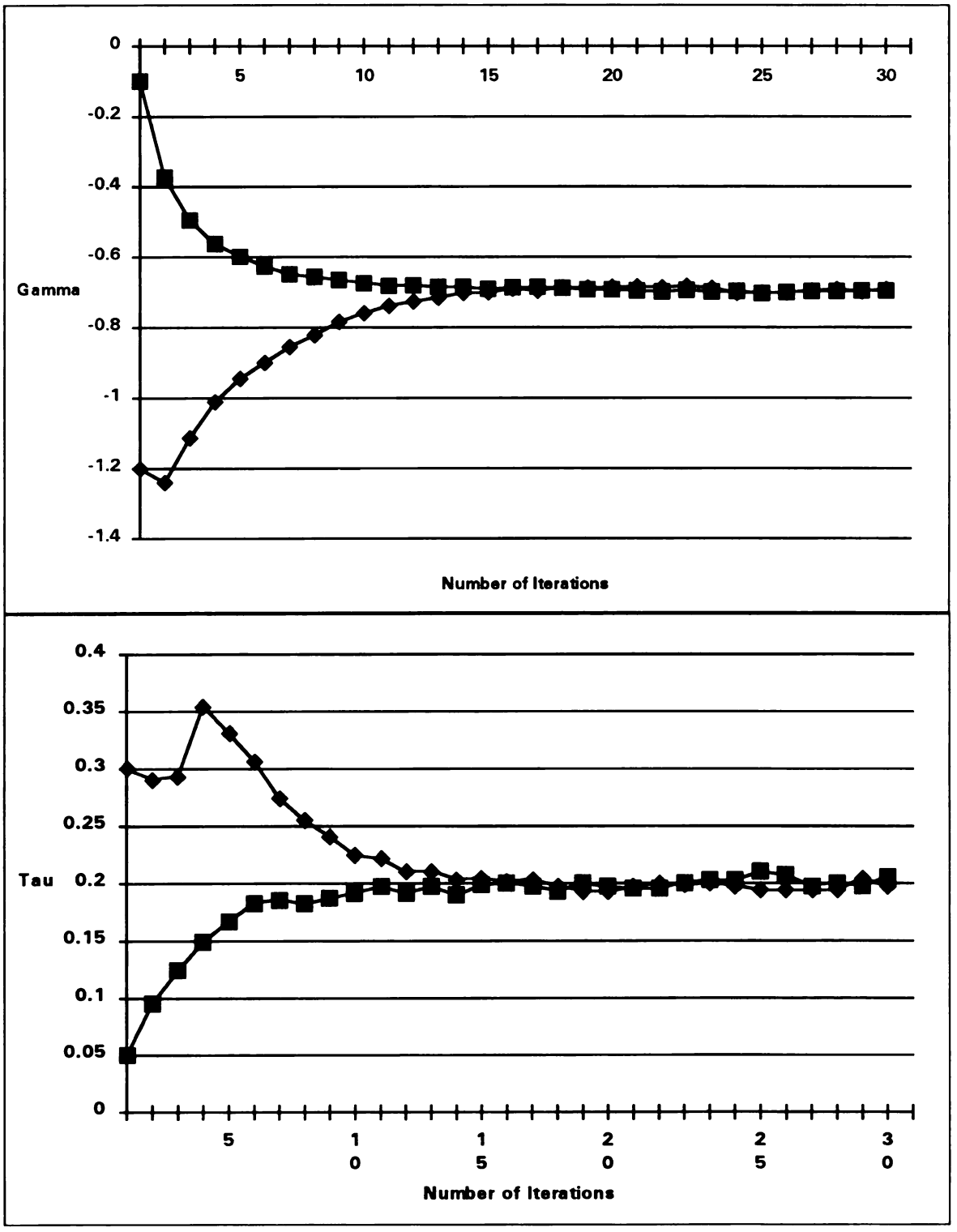


Convergence of estimated parameters

As shown from Figure 2, the estimates of γ and T both converge rather smoothly to stable values after some 15 iterations. The slight discontinuity for the converging T sequence with starting value 0.3 is caused by my fixing the value of T for the first three iterations in order to improve stability.

Compared to the parameter value of γ: -0.7, the iterations from two different initial positions obtain the values of -0.700 and -0.690 at the 15^{th} iteration. Similarly, estimations of the variance component T obtain the values of 0.205 and 0.199, which are not far from the theoretical parameter value of 0.16. These specific deviations are due to the random sampling errors in creating the data sets.

Figure 2. Convergent paths of model parameters from two initial positions



Convergence from different starting values

Given the parameter values of γ and T as respectively -0.7 and 0.16, the program was tested for two sets of starting values. The actual iterating paths are shown in the previous three figures. It is evident that the convergent parameter values estimated by the two iterating sequences from different starting positions are identical. The difference in precision obtained can be accounted by the Monte Carlo nature of the iterations. Results are summarized in the following table:

Table 2. Comparing parameter and likelihood values obtained by two starting values (Parameter values for γ and T are respectively -0.7 and 0.16)

Starting values		Estimated values at 15^{th} iterations		
γ	T	γ	T	Log Likelihood
-1.2	0.3	-0.700	0.205	-970.62
-0.1	0.05	-0.690	0.199	-970.77

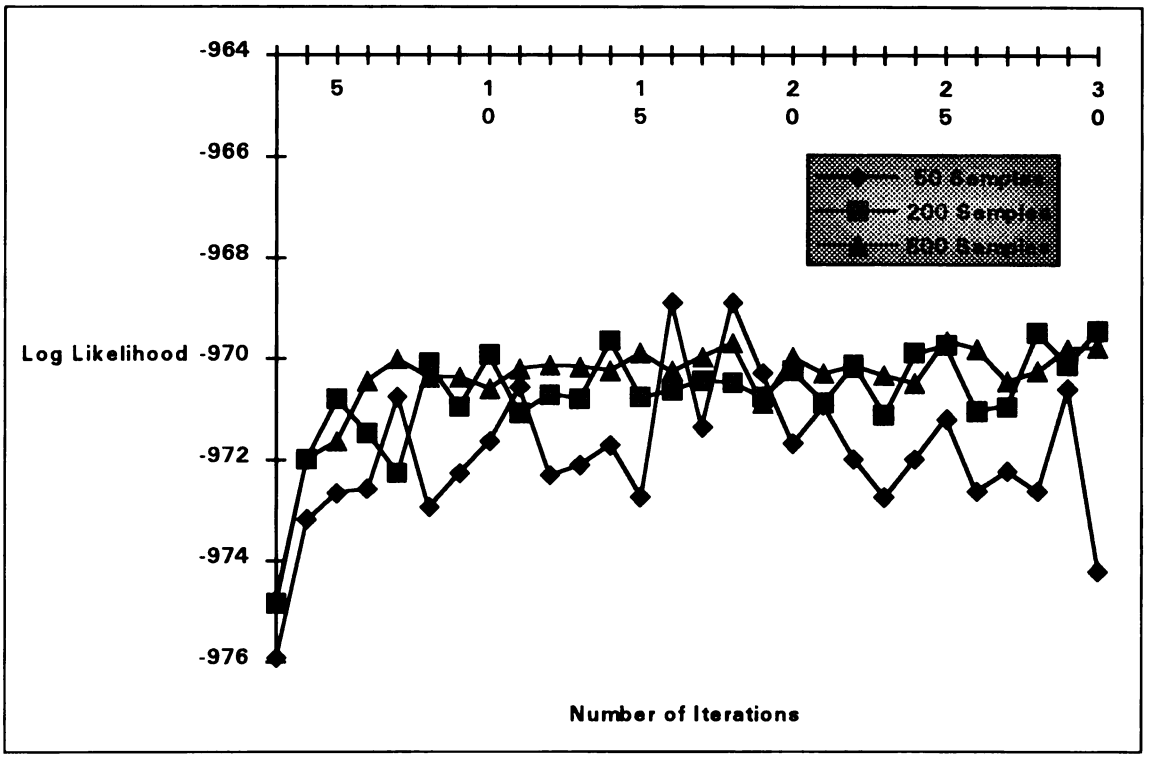
Influence of the size of Monte Carlo samples

This program employs Monte Carlo EM algorithm and computes Monte Carlo integration through the generation of multivariate normal variates. The higher the number of Monte

Carlo samples, the better the approximation to the theoretical integral by the Monte Carlo integral. Figure 3 depicts the influence of the number of Monte Carlo samples on the estimated log likelihood for a random intercept Poisson model. The same data set generated from the parameter set (γ,T) = (-0.1, 0.05) is being modeled on three different occasions with the size of Monte Carlo samples vary over 50, 200 and 800 per each integral. As expected, the run with the least Monte Carlo samples (M=50) produces the largest fluctuations on the estimated log likelihood values and vice versa. It should also be noted that the run with 800 samples seem to achieve the largest log likelihood on average than the other runs with fewer samples. It is because parameter estimates are more precise and get closer to their ML values as the Monte Carlo integration becomes more accurate with larger generated samples. The likelihood value computed from these parameter values which are closer to the ML values will approach nearer to the maximum likelihood value.

The effect of the size of Monte Carlo samples on parameter estimation is similar. The convergent paths of the estimated parameters fluctuate less when the number of Monte Carlo samples increases.

Figure 3. Influence of the size of Monte Carlo samples on the estimated log likelihood



In this chapter, illustrative program runs provide some confirmations on the validity of the estimation procedure and the computer program. The program does appear to be working reliably to produce maximum likelihood estimation for at least a random intercept Poisson model. Different starting values lead to the same results at convergence. The size of Monte Carlo samples is found to affect the smoothness of the convergent path and the proximity of the statistical estimation to the true ML

estimate. More formal simulation tests will be presented in the next chapter.

CHAPTER VII

SIMULATION STUDY

Simulation test for a random intercept model

Despite its costliness, the simulation study is one of the best strategies to access the quality of one's statistical estimation. Using selected parameter values, random variates can be generated from their respective density functions to form a large number of independently simulated data sets. Parameter values are then estimated from each of these data sets. The means of the estimated parameters averaged over all the data sets are compared with the respective chosen parameter values to determine the bias, if any, of the estimators. The empirical variance of the estimators might also be used to compare with the variance of the other estimators to determine their relative efficiency. The following paragraphs describe a simulation test for the simplest multilevel generalized linear model: a random intercept model with no predictor variables. This model is analogous to the one-way random effect ANOVA model.

The observation y_{ij} has a Poisson distribution with mean parameter μ_{ij}. Using a logarithmic link function, the model is

$$\log \mu_{ij} = \gamma + u_j,$$

where γ is the fixed effect and u_j is the random effect for school j. The random effect u_j is modeled by a normal distribution:

$$u_j \sim N(0,\tau)$$

For the simulation test, the parameters γ and τ are set to be -0.7 and 0.16. Thus the school mean is expected to be $e^{-0.7} = 0.497 \approx 0.5$. The mean 0.5 is chosen to simulate the problematic scenario in data analysis where many zero occurrences are expected from a Poisson distribution and the distribution will not look normal and symmetrical. Parameters for the standard error of $\hat{\gamma}$ and $\hat{\tau}$ are set by the empirical standard deviation of $\hat{\gamma}$ and $\hat{\tau}$ (For more detail, see Press (1986), p.529-532). Each simulated data set contains 50 schools each with 20 students. One hundred independent data sets are created. They are fitted by the random intercept Poisson model and the estimated parameters are recorded. To be economical in time, the beginning iterations of the program employ fewer Monte Carlo (MC) samples. The later iterations use larger samples as results approach convergence so as to reduce stochastic variation in the final estimates. Thus, 50 MC samples are used in the first 15 iterations, followed by 200 MC samples for the next 10 iterations and followed by 800 MC samples for

the last 7 iterations. Experience shows that convergence to about two significant figures of accuracy is achieved with this scheme. To reduce randomness, parameter estimates from the last 4 iterations are averaged to produce the final estimate. However, standard errors of the parameter estimates are only calculated at the last iteration. They are not being averaged because they usually fluctuate less and require a large additional computation resource per iteration. Results of the simulation test is tabulated as the case 1 study in Table 3.

The results show that the fixed effect parameter and the variance component are both estimated without bias with error less than 1.5%. The standard error of the fixed effect has a slight negative bias below 3%. The standard error of the variance component is negatively biased by about 18%. Coverage by 95% confidence intervals range from 88% to 94%. Overall, the simulation study shows that the statistical estimation of the model is quite satisfactory.

Table 3. Results of 2 simulation studies. Case 1: Random intercept model. Case 2: Random coefficient model.

φ:	γ_{00}	γ_{01}	γ_{10}	γ_{11}	τ_{00}	τ_{01}	τ_{11}
Study 1							
φ	-0.7	—	—	—	0.16	—	—
$\hat{\varphi}$	-0.71				0.16		
$S_{\hat{\varphi}}$	0.074				0.051		
$\hat{S}_{\hat{\varphi}}$	0.072				0.042		
Coverage of 95% CI	95				88		
Study 2							
φ	-0.5	0.4	0.3	0.2	0.083	0.017	0.0092
$\hat{\varphi}$	-0.50	0.39	0.30	0.20	0.086	0.011	0.012
$S_{\hat{\varphi}}$	0.089	0.12	0.036	0.044	0.031	0.0068	0.0046
$\hat{S}_{\hat{\varphi}}$	0.084	0.11	0.037	0.050	0.015	0.0050	0.0024
Coverage of 95% CI	94	94	95	97	67	69	73

Simulation test for a random coefficient model

A linear growth model predicted by the sex of students is used in a simulation study for a random coefficient model. The time data x_{ij} is coded as -3, -2, -1, 0, 1, 2, 3 and sex ω_j is coded as 0 (girls) and 1 (boys). There are 7 repeated time observations nested within 100 students equally divided by the two sexes.

The level-one model is

$$\log \mu_{ij} = \beta_{0j} + \beta_{1j} x_{ij}.$$

The level-two model consists of

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\omega_j + u_{oj},$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\omega_j + u_{1j}.$$

The fixed parameters $\gamma_{00}, \gamma_{01}, \gamma_{10}$ and γ_{11} are respectively set to be: -0.5, 0.4, 0.3 and 0.2. The dispersion matrix is $T = \begin{pmatrix} 0.0827 & 0.0165 \\ 0.0165 & 0.00919 \end{pmatrix}$. Thus the correlation between u_{0j} and u_{1j} is set to be

$$\frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}} = \frac{0.0165}{\sqrt{0.0827 \times 0.00919}} = 0.6.$$

The log expected mean and expected mean (in parenthesis) of a student are tabulated below by sex and by the beginning and ending time points:

Table 4. Log expected mean and expected mean (in parenthesis) by time and sex

	Time x_{ij}	
Sex ω_j	-3.0	3.0
Girls 0	-1.40 (0.25)	0.40 (1.50)
Boys 1	-1.60 (0.20)	1.40 (4.10)

The iteration scheme is as follows: Iteration begins with 50 MC samples. At the 20^{th} iteration, it increases to 200 MC samples for 10 iterations. At the 30^{th} iteration, 800 MC samples are used for the last 8 iterations. Experience with some trial runs shows that this iteration scheme is generally more than sufficient for attaining the desired accuracy of the estimates. The final estimates for the parameters and the parameter variances are calculated by the same method as described in the previous random intercept case. Again, 100 data sets are generated from the known set of parameters.

Results of the simulation study are displayed in Table 3 along side with the results for the random intercept case. Estimation for the fixed effects reach 2 significant figures of accuracy in general except for γ_{01}, which has a slight negative bias by about 3%. Coverage of the 95% confidence intervals are very good with results ranging from 94% to 97%. Standard error estimates for the fixed effects γ_{00} and γ_{01} come very close to their 'true' estimate with only 6% and 8% negative bias. The standard error estimates for γ_{10} and γ_{11} deviates positively from their 'true' values by 3% and 14% respectively. The 'true' values for the standard errors of the fixed effects and the variance components are not pre-selected parameters. They are estimated

by the empirical standard deviations of the fixed effects and the variance components obtained from the simulation runs.

Estimations of the variance components tend to be less accurate than those for the fixed effects. Nevertheless, τ_{00} is estimated accurately with only about 4% positive bias. Variance component τ_{11} is estimated with 30% positive bias, while component τ_{01} is esimated with a 35% negative bias. Standard errors of the variance components tend to be larger than their conterparts for the fixed effects, with biases equal to +52%, -26%, -48%. Coverage probabilities by the confidence intervals of the variance components are 67%, 69% and 73%. These larger errors of the variance components estimates may be due to chance differences because the ML estimators are only asymptotically unbiased. The large sample assumption for the normality of the MLE of T may be violated due to the small within-group sample size of 7.

The quality of my simulation results is comparable to those of a random effect binomial model using a similar design by Zeger and Karim (1991, p.83). Although they use the binomial distribution instead of the Poisson distribution, it is still interesting to compare the degree of accuracy of their results against mine because both studies have identical structural

models, variable values and number of cases. For easy comparison, only absolute value of the biases are shown below.

In their simulation study with 2 variance components, fixed effect estimates are biased by 10%, 14%, 10% and 14% (Mine: 0%, 3%, 0% and 0%) with their corresponding standard errors biased by 3%, 6%, 12% and 10% (Mine: 6%, 8%, 3% and 14%). The nominal 90% intervals by these 4 standard errors have coverage probabilities of 80%, 89%, 87% and 87% (Mine: 94%, 94%, 95% and 97% for 95% confidence interval). Their variance components are biased by 69% and 48% (Mine: 4%, 30%) respectively. Their covariance component (parameter=0.0) is only biased by 8% (Mine: 35% for a non-zero parameter=0.0165 with 0.6 correlation between the 2 random effects). Their standard errors of the variance components are biased by 18%, 75% and 57% (Mine: 52%, 26% and 48%). However, the coverage probabilities of their nominal 90% intervals for the variance components provided by their standard errors achieve high values of 88%, 95% and 100% (Mine: 67%, 69% and 72% for 95% confidence interval). Apparently, the standard errors of their variance components are quite over-estimated. Thus their confidence intervals are longer and will often include the parameter values for more than 90% of the samples. On the contrary, the standard errors of the variance components of my

estimation method tend to be under-estimated. Because of the incorporation of vague prior distributions, Bayesian approaches tend to produce larger standard errors than the non-Bayesian approaches (c.f. Tsuatakawa, 1985)

CHAPTER VIII

CONCLUSIONS

This dissertation has demonstrated a maximum likelihood (ML) estimation approach for multilevel generalized linear model via the Monte Carlo EM algorithm. Although only the Poisson distribution is used for illustration, the same approach can be used to cover other distributions in the exponential family such as the binomial, exponential and normal distributions. Anderson and Hinde (1988) have published the first paper on the random effects generalized linear model using a true ML approach also with EM algorithm. Unfortunately, their model is limited to a single variance component and they have not given any method to find out the standard errors for the fixed effects and the variance component. Their adoption of a numerical quadrature technique for solving multiple integrals also limited their potentials for extension to multiple variance component model because quadrature techniques are good for low dimensions and their required computations accelerate with increasing dimensions (Rubinstein, 1981).

Built on Anderson and Hinde's (1988) pioneering work, I have been able to extend their random intercept model to a fully

random coefficient model because of the successful implementation of the Monte Carlo methods and the discovery and derivation for the proof of Theorem 1. With this theorem, we can easily estimate in closed form the variance components of any dimensions by equation (4.25). Maximum likelihood estimation is achieved because the EM algorithm is a smooth likelihood maximizer through indirectly maximizing the expectation of the 'complete data' likelihood. Using Monte Carlo integrations by generating multivariate normal distribution, the present approach is able to work at higher dimensions of variance components. The growth rate of computer resources for Monte Carlo integration is linear with the increase of integral dimensions. Employing Louis's (1982) method, I derived the formulas for computing the standard errors for the fixed effects and the variance components using tractable complete data information and avoided the intractable incomplete data information matrix. Simulation studies on a random intercept and a random coefficient model have shown promising results of the accuracies of my program. It is in general superior to the best existing approach (Zeger and Karim, 1991) with regard to the degree of estimation accuracy. Besides, computation time of my program is comparatively much shorter.

Besides its statistical contribution, the present model can help bring some of the challenging data analytic scenarios such as multilevel models, longitudinal studies (e.g. Raudenbush and Chan, 1992, 1993) and meta-analysis (e.g. Becker, 1988) into a new realm involving non-normal random effects models. Research which involves discrete count, dichotomous and survival time data is often difficult for an applied researcher. Coupled with a complicated multilevel, longitudinal or meta-analytic scenario, suitable statistical models that solve these analytic problems are in great need. This approach can provide an additional option for doing research using random effects models. Practical educational studies include a large number of level-one variables and the demand for high dimension random slopes can now be solved through the Monte Carlo approach. The present approach will also serve as a standard to evaluate the existing approximate maximum or quasilikelihood approaches.

Further research

Notwithstanding the above mentioned success, a number of works that remedy some existing shortcomings or advance the

usefulness of the current model and program are awaiting. They are listed as follows:

1) Application to real data analysis

Due to the problems in finding a suitable educational data set to demonstrate the multilevel Poisson model, only simulation studies are presented. Upon the access of suitable data sets, applicational studies should be performed to demonstrate the usefulness of my program to solve real data analytic challenges in multilevel studies.

2) Extension to other members of the exponential family

Random effects binomial models may be more common in applied educational and social research than the corresponding Poisson models. Binomial distribution should be substituted for Poisson distributions in this model for deriving a random effect binomial model. Simulation studies should be performed to access the accuracy of the resulting model. Extension to exponential, normal or other member of the exponential family should also be persued.

3) Improving the speed of the program

Due to the large computer resource requirement by generating multivariate normal distributions for doing Monte Carlo integrations, run time for the simulation models can reach 20-30 minutes for about 2 significant figures of accuracy. To be applicable to larger real data sets, program speed has to be increased. Possible choices are choosing a faster programming language and computer system, employing more efficient random variable generating methods, using suitable variance reduction techniques (Rubinstein, 1981) or accelerators of the EM algorithm (e.g. Louis, 1982, Jamshidian and Jennrich, 1993).

4) Search for a good criterion of convergence

Difficulties have been reported about judging when to stop a stochastic convergent sequence (Gelfand and Smith, 1990). Ploting the sequence of parameters against number of iterations could tell us when the converging parameter path are stable against the background of Monte Carlo fluctuations. Further research should be done either to automate the ploting technique or to compute a criterion of convergence based on some suitable standard.

5) Improving the estimates of the standard errors

In my simulation study with two variance components, the biases of the standard errors of the variance components range from 26% to 52%. The reason could be due to the low sample size (n_j= 7) within each level-2 unit. Lack of information to estimate the variance components could give rise to a larger error in their standard error estimates. Rodriguez and Goldman (1993) report in their simulation study of a multilevel logit model that parameter estimates are more biased when the number of observations within group is small. Moreover, variance estimates tend to have a skewed asymmetric distributions when sample sizes are not large enough. With appropriate choice of prior distributions for the parameters, estimation by Bayesian techniques could improve over ML approaches especially when the available sample size is small. The use of the mean as a point estimate to summarize the sampling distribution of the variance components may not be the best choice. For example, Zeger and Karim (1991) suggest that the use of mode or geometric mean might alleviate some of the biases.

Data augmentation techniques (Tanner, 1987) could be used to obtain the whole posterior distributions of the standard error sampling distributions and the use of the posterior mode in

place of the MLE might lead to improvement of the standard error estimates.

APPENDIX A

Mathematical notes

A1 Expectation of the function of a random variable by a joint distribution

Let $A \subset \mathbf{R}^{(P+1)\times J}$, $B \subset \mathbf{R}^{P+1}$, $f: A \to \mathbf{R}$, and $g: \mathbf{R}^{P+1} \to \mathbf{R}$ and let $f(\mathbf{u}) = f(u_1)f(u_2),...,f(u_J)$ be a joint distribution function of J multivariate random variables $u_1,...,u_J$ each with dimension P+1. Then the expectation

$$E[g(u_j)] = \int_A g(u_j)f(\mathbf{u})d\mathbf{u} = \int_B g(u_j)f(u_j)du_j .$$

Proof:

$$\int_A g(u_j)f(\mathbf{u})d\mathbf{u} = \int_A g(u_j)f(u_1),...,f(u_J)du_1,...,du_J$$

Evaluating the integrand for some $u_{j'}$ axis with $j' \neq j$, $j' \in \{1,...,j-1,j+1,...,J\}$, it becomes

$$\int_A g(u_j)f(u_1),...,f(u_{j-1}),f(u_{j+1}),...,f(u_J)du_1,...,du_{j-1},du_{j+1},...,du_J .$$

Performing the above partial integration for all $u_{j'}$ axis, $j' \neq j$, the integrand becomes $\int_B g(u_j)f(u_j)du_j$. □

Corollary 1: $\sum_{j=1}^{J}\int_A g(u_j)f(\mathbf{u})d\mathbf{u} = \sum_{j=1}^{J}\int_B g(u_j)f(u_j)du_j$.

Proof: It follows directly from summing both sides of the results from Appendix A1. □

Corollary 2: Additionally, let $h:\mathbf{R}^{P+1} \to \mathbf{R}$, then

$$\int_A g(u_j)h(u_{j'})f(\mathbf{u})d\mathbf{u} = \int_B g(u_j)f(u_j)du_j \int_B g(u_{j'})f(u_{j'})du_{j'}$$

Proof: Following the proof in Appendix A1, it can be readily shown that for $C \subset \mathbf{R}^{2(P+1)}$,

$$\text{L.H.S.} = \int_C g(u_j)h(u_{j'})f(u_j)f(u_{j'})du_j du_{j'}$$

$$= \int_B g(u_j)f(u_j)du_j \int_B h(u_{j'})f(u_{j'})du_{j'} \text{ . } \square$$

Corollary 2a:

$$\sum_{j,j'=1;j\neq j'}^{J}\int_A g(u_j)h(u_{j'})f(\mathbf{u})d\mathbf{u} = \sum_{j,j'=1;j\neq j'}^{J}\int_B g(u_j)f(u_j)du_j \int_B g(u_{j'})f(u_{j'})du_{j'}$$

Proof: It follows directly from summing both sides of the results from Appendix A1, Corollary 2. □

A2 Expected information matrix with weights

Let $L(\varphi)=\sum_{i=1}^{I}\psi_i \log f(y_i|\varphi)$ be a weighted log likelihood function with I independent random samples and $L_i(\varphi)=\psi_i \log f(y_i|\varphi)$. $f(y_i|\varphi)$ is a probability density function with parameter φ and ψ_i is some weight of given value . Then

$$E\left(\frac{\partial^2 L(\varphi)}{\partial\varphi\partial\varphi'}\right)=-\sum_{i=1}^{I}\frac{1}{\psi_i}E\left(\frac{\partial L_i(\varphi)}{\partial\varphi}\frac{\partial L_i(\varphi)}{\partial\varphi'}\right).$$

Proof:

$$E\left(\frac{\partial^2 L(\varphi)}{\partial\varphi\partial\varphi'}\right)=E\left(\frac{\partial^2}{\partial\varphi\partial\varphi'}\sum_{i=1}^{I}\psi_i \log f(y_i|\varphi)\right)$$

$$=\sum_{i=1}^{I}\psi_i E\left(\frac{\partial^2}{\partial\varphi\partial\varphi'}\log f(y_i|\varphi)\right)$$

$$=-\sum_{i=1}^{I}\psi_i E\left(\frac{\partial \log f(y_i|\varphi)}{\partial\varphi}\frac{\partial \log f(y_i|\varphi)}{\partial\varphi'}\right)$$

(Seber, 1989, p.685),

$$=-\sum_{i=1}^{I}\frac{1}{\psi_i}E\left(\frac{\partial}{\partial\varphi}[\psi_i \log f(y_i|\varphi)]\frac{\partial}{\partial\varphi'}[\psi_i \log f(y_i|\varphi)]\right)$$

$$=-\sum_{i=1}^{I}\frac{1}{\psi_i}E\left(\frac{\partial L_i(\varphi)}{\partial\varphi}\frac{\partial L_i(\varphi)}{\partial\varphi'}\right). \square \qquad \text{(A2.1)}$$

A3 Conditional expectation by the missing data distribution via Monte Carlo method

Let $E[g(\mathbf{u})]$ be the expectation of a function $g(.)$ of $\mathbf{u}$, consisted of independent variates $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_J$, conditioned on the data $\mathbf{Y}$ with independent components $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_J$ and the current parameter value $\varphi^{(l)}$ at iteration l. Then

$$E[g(\mathbf{u})] = \int g(\mathbf{u}) f(\mathbf{u}|\mathbf{Y}, \varphi^{(l)}) d\mathbf{u}$$

$$\approx \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{1}{MC_j} g(\mathbf{u}_{mj}) f(\mathbf{Y}_j | \mathbf{u}_{mj}, \varphi^{(l)})$$

where

$$C_j \approx \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{Y}_j | \mathbf{u}_{mj}, \varphi^{(l)})$$

and

$$\mathbf{u}_{1j}, \mathbf{u}_{2j}, \ldots, \mathbf{u}_{Mj} \sim f(\mathbf{u}_j | \varphi^{(l)})$$

are M samples of simulations from the multivariate distribution $f(\mathbf{u}_j | \varphi^{(l)})$.

Proof:

$$E[g(\mathbf{u})] = \int g(\mathbf{u}) f(\mathbf{u}|\mathbf{Y}, \varphi^{(l)}) d\mathbf{u}$$

By Appendix A1, Corollary 1, the above is equal to

$$\sum_{j=1}^{J}\int g(\mathbf{u}_j)f(\mathbf{u}_j|\varphi^{(l)},\mathbf{Y}_j)d\mathbf{u}_j$$

$$=\sum_{j=1}^{J}\int g(\mathbf{u}_j)\frac{f(\mathbf{Y}_j|\mathbf{u}_j,\varphi^{(l)})f(\mathbf{u}_j|\varphi^{(l)})}{f(\mathbf{Y}_j|\varphi^{(l)})}d\mathbf{u}_j\text{, by Bayes' law}$$

$$\approx\sum_{j=1}^{J}\frac{1}{C_j}\int g(\mathbf{u}_j)f(\mathbf{Y}_j|\mathbf{u}_j,\varphi^{(l)})f(\mathbf{u}_j|\varphi^{(l)})d\mathbf{u}_j \qquad \text{(A3.1)}$$

where the constant $C_j \equiv f(\mathbf{Y}_j|\varphi^{(l)})=\int f(Y_j|\mathbf{u}_j,\varphi^{(l)})f(\mathbf{u}_j|\varphi^{(l)})d\mathbf{u}_j$ is computed by the Monte Carlo integration method (Rubinstein, 1981):

$$C_j \approx \frac{1}{\mathbf{M}}\sum_{m=1}^{\mathbf{M}} f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)}),$$

where

$$\mathbf{u}_{1j},\mathbf{u}_{2j},\ldots,\mathbf{u}_{Mj}\sim f(\mathbf{u}_j|\varphi^{(l)}),$$

are M samples of simulations from the multivariate distribution $f(\mathbf{u}_j|\varphi^{(l)})$.

Applying the Monte Carlo integration technique again, equation (A3.1) is approximated by

$$\sum_{j=1}^{J}\frac{1}{\mathbf{M}C_j}\sum_{m=1}^{\mathbf{M}}g(\mathbf{u}_{mj})f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)})$$

$$=\sum_{m=1}^{\mathbf{M}}\sum_{j=1}^{J}\frac{1}{\mathbf{M}C_j}g(\mathbf{u}_{mj})f(\mathbf{Y}_j|\mathbf{u}_{mj},\varphi^{(l)}).\ \square \qquad \text{(A3.2)}$$

A4: Conditional expectation of the score function

Lemma (c.f. a similar argument from Tanner, 1991, p.37):

$$\int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j = \frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau}.$$

Proof:

$$f(\mathbf{Y}_j|\varphi) = \frac{f(\mathbf{Y}_j,\mathbf{u}_j|\varphi)}{f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}$$

$$f(\mathbf{Y}_j|\varphi) = \frac{f(\mathbf{Y}_j|\mathbf{u}_j,\varphi)f(\mathbf{u}_j|\varphi)}{f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}$$

$$\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau} = \frac{\partial \log f(\mathbf{Y}_j|\mathbf{u}_j,\varphi)}{\partial \tau} + \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} - \frac{\partial \log f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}{\partial \tau}$$

$$= 0 + \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} - \frac{\partial \log f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}{\partial \tau}.$$

Taking the expected value for both sides w.r.t. $f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)$:

$$\frac{\partial \log f(\mathbf{Y}_j|\varphi)}{\partial \tau} = \int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j - \int \frac{\partial \log f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j$$

$$= \int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j - \int \frac{\partial f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)}{\partial \tau} d\mathbf{u}_j$$

$$= \int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j - \frac{\partial}{\partial \tau}\int f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j$$

$$= \int \frac{\partial \log f(\mathbf{u}_j|\varphi)}{\partial \tau} f(\mathbf{u}_j|\mathbf{Y}_j,\varphi)d\mathbf{u}_j. \quad \square$$

APPENDIX B

Estimation of Starting Values

Starting values are vital to the efficient and successful implementation of any iterative algorithm. In our case, if the starting values are too distant from the final ML estimates, it will take a long time for the program to converge. In the worse case, the program will fail. It is because the starting values may not be within a quadratic likelihood region of the maxima which is a requirement for Newton-type maximization algorithm to work (Seber and Wild, 1989). Precise estimation of starting values may sometimes require iterations and becomes time consuming. Hence a good choice of starting values are those which are easy to compute and close to the final ML estimates.

Starting values for the fixed effects γ are calculated by regressing the logarithm of the dependent variable on the independent variables related to the fixed part of the model. The random part of the model is ignored for simplicity:

$$\log(\mu_{ij}) = \mathbf{A}_{ij}\gamma + \mathbf{B}_{ij}u_j \approx \mathbf{A}_{ij}\gamma . \qquad (B.1)$$

By taking the datum Y_{ij} as an approximation for mean μ_{ij}, we can obtain starting values $\gamma^{(0)}$ as the least square solutions of the model:

$$\log(Y_{ij}) = \mathbf{A}_{ij}\gamma^{(0)} + e_{ij}, \qquad \text{(B.2)}$$

and

$$\gamma^{(0)} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Z}, \qquad \text{(B.3)}$$

where $\mathbf{A} = (\mathbf{A}_1', \mathbf{A}_2', ..., \mathbf{A}_J')'$ and $\mathbf{Z} = [(\log Y_{11}, ..., \log Y_{n_1 1}), ..., (\log Y_{1J}, ..., \log Y_{n_J J})]'$ are, respectively, the stacked matrix of the independent variables for the fixed effects and the stacked vector of the logarithm of the dependent variable across all the level-2 units. An arbitrary small negative value in place of zero is used to avoid undefined value for the logarithmic function: For example, if $Y_{ij} = 0$, Y_{ij} is set to be 10^{-5}.

The above computed starting values for fixed effects are used to calculate the starting values for the variance-covariance matrix of the random effects:

Let

$$Z_{ij}^{*} = \log(Y_{ij}) - \mathbf{A}_{ij}\gamma^{(0)} \approx \mathbf{B}_{ij}\mathbf{u}_j. \qquad \text{(B.4)}$$

Random effects of each level-2 units can be approximated by regressing the difference between the logarithm of the dependent variables and the predicted fixed part of the systematic

component on the independent variables related to the random effects:

$$\mathbf{u}_j^{(0)} = (\mathbf{B}_j'\mathbf{B}_j)^{-1}\mathbf{B}_j'\mathbf{Z}_j^*. \qquad (B.5)$$

Starting values for the variance-covariance matrix are estimated by

$$T^{(0)} = \frac{1}{J}\sum_{j=1}^{J}\mathbf{u}_j^{(0)}\mathbf{u}_j^{(0)\prime}, \text{ (Seber, 1984).} \qquad (B.6)$$

The above estimation has the advantage of being positive definite with probability 1, which is a requirement for generating the multivariate normal random variables in perfoming Monte Carlo integrations.

LIST OF REFERENCES

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modeling of data on teaching styles. Journal of the Royal Statistical Society, Series A, 144(4), 419-461.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). Statistical modeling in GLIM. NY: Oxford University Press.

Albert, J.H. (1985). Simultaneous estimation of Poisson means under exchangeable and independence models. Journal of Statistical Computation and Simulation, 23, 1-14.

Albert, J.H. (1988). Bayesian estimation of Poisson means using a Hierarchical log-linear model. In J.M. Bernado, M.H. Degroot, D.V. Lindley and A.F.M. Smith (Eds.), Bayesian statistics 3, (519-531). NY : Clarendon Press.

Albert, J. (1992). A Bayesian analysis of a Poisson random effects model for home run hitters. The American Statistician, 46, 246-253.

Anderson, D.A. and Hinde, J.P. (1988). Random effects in generalized linear models and the EM algorithm.

Communications in Statistics Theory and Methods, 17(11), 3847-3856.

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. Journal of the Royal Statistical Society, Ser. B, 47, 203-210.

Becker, B.J. (1988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 41, 257-278.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian inference in statistical analysis. Reading, MA: Addison-Wesley.

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

Bryk, A.S. and Raudenbush, S.W. (1992). Hierarchical linear models for social and behavioral research: Applications and data analysis methods. Beverly Hills, CA: Sage.

Crowder, M. (1987). On linear and quadratic estimating functions. Biometrika, 74, 591-597.

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

Cox, D.R. (1983). Some remarks on over-dispersion. Biometrika, 70, 269-274.

Das Gupta, S. (1971). Nonsingularity of the sample covariance matrix. Sankhya A, 33, 475-478.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Ser. B, 39, 1-38.

Firth, D. (1987). On the efficiency of quasi-likelihood estimation. Biometrika, 74, 233-245.

Gelfand, A.E., Hills, S.I., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in Normal data methods using Gibbs sampling. Journal of the American Statistical Association, 90, 972-985.

Goldstein, H.I. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika, 73, 43-56.

Goldstein, H.I. (1987). Multilevel models in educational and social research. London: Oxford University Press.

Goldstein, H.I. (1989). Restricted unbiased iterative generalised least squares estimation. Biometrika, 76, 622-623.

Goldstein, H.I. (1991). Nonlinear multilevel models, with an application to discrete response data. Biometrika, 78, 45-51.

Huttenlocher, J.E., Haight, W., Bryk, A.S. and Seltzer, M. (1991). Early vocabulary growth: Relationship to language input and gender. Developmental Psychology, 27(2), 236-249.

Jamshidian, M. and Jennrich, R. (1993). Conjugate gradient acceleration of the EM algorithm. Journal of the American Statistical Association, 88, 221-228.

Johnson, L.W. and Riess, R.D. (1982). Numerical Analysis, Philippines: Addison-Wesley

Karim, M.R. and Zeger, S.L. (1992) Generalized linear models with random effects; Salamander mating revisited. Biometrics, 48, 631-644.

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

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika, 74, 817-827.

Longford, N.T. (1988). A quasi-likelihood adaption for variance component analysis. Proc. Sect. Comp. Statist., Am. Statist. Assoc., 137-142.

Longford, N.T. (1993). Random coefficient models. NY: Oxford university press.

Louis, T.A. (1982). Finding observed information using the EM algorithm. Journal of the Royal Statistical Society, Ser. B, 44, 98-130.

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

Morton, R. (1987). A generalized linear model with nested strata of extra-Poisson variation. Biometrika, 74, 247-257.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. Journal of the Royal Statistical Society, Ser. A, 135, 370-384.

Press, W.H. (1986). Numerical recipes: The art of scientific computing. NY: Cambridge University Press.

Raudenbush, S.W. and Bryk, A.S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1-17.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: a review. Journal of Educational Statistics, 13, 85-116.

Raudenbush, S.W. and Chan, W.S. (1992). Growth curve analysis in accelerated longitudinal design with application to the National Youth Survey. Journal of Crime and Delinquency, 29, 387-411.

Raudenbush, S.W. and Chan, W.S. (1993). Application of a Hierarchical Linear Model to the study of adolescent deviance in an overlapping cohort design. Journal of Consulting and Clinical Psychology, 61, 941-951.

Racine-Poon, A. (1985). A Bayesian approach to Non-linear random effects models.Biometrics, 41, 1015-1023.

Rice, J.A. (1987). Mathematical statistics and data analysis. California: Wadsworth Inc.

Rodriguez, G. and Goldman, N. (1993). An assessment of estimation procedures for multilevel models with binary responses. Paper presented at a workshop on multilevel analysis organized by the Rand Corporation, Santa Monica, California, July, 1993.

Rubinstein, R.Y. (1981). Simulation and the Monte Carlo method. NY: Wiley.

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

Seber, G.A.F. (1984). Multivariate observations.NY: Wiley.

Seber, G.A.F. (1989). Nonlinear regression. NY: Wiley.

Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C. and Dransdield, M. (1985). The implementation of the Bayesian paradigm. Communications in Statistics: Theory

and Methods, 14(5), 1079-1102.

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

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-550.

Tanner, M.A. (1991). Tools for statistical inference: Observed data and data augmentation methods. Berlin: Springer-Verlag.

Tsutakawa, R.K. (1985). Estimation of cancer mortality rates: A Bayesian analysis of small frequencies. Biometrics, 41, 69-79.

Tsutakawa, R.K. (1988). Mixed model for analyzing geographic variablity in mortality rates. Journal of the American Statistical Association, 83, 37-42.

Watson, G.S. (1964). A note on maximum likelihood. Sankhya A, 26, 303-304.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms. Journal of the American Statistical Association, 85, 699-704.

Wong, G.Y. and Mason, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. Journal of the American Statistical Association, 80, 513-524.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. Annals of Statistics, 31, 144-148.

Zeger, S.L., Liang, K.Y. and 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.