A class of models for semicontinuous longitudinal data

Maren K. Olsen, Joseph L. Schafer, The Pennsylvania State University
Maren K. Olsen, 326 Thomas Building, University Park, PA 16801 (olsen@stat.psu.edu)

Key Words: growth curve models, EM algorithm, importance sampling

1. Introduction

Semicontinuous variables have a proportion of responses equal to a single value (often zero) and a continuous distribution among the remaining responses, as shown in Figure 1. Semicontinuous variables are common in many fields of research. Examples include: individual consumption of alcohol, tobacco, or other controlled substances; annual household expenditures on a class of durable goods (e.g. refrigerators); and annual income from specific sources (e.g. dividend income). Despite the prevalence of semicontinuous variables, current techniques do not handle this type of data well, particularly when change over time is to be assessed. Using methods not specifically tailored to semicontinuous variables may produce incorrect estimates or limit the types of hypotheses researchers would like to test.

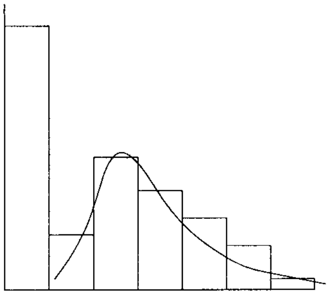

Figure 1: Histogram of a semicontinuous variable

Multilevel linear models, or general linear mixed models, (e.g. Bryk and Raudenbush, 1987; Lindstrom and Bates, 1988) may be used to study individual growth when it can be assumed that the response is continuous and normally distributed. Many software packages are available for fitting multilevel linear models, including MLn (Multilevel Models Project, 1996), HLM (Bryk, Raudenbush, and Congdon, 1996), and SAS PROC MIXED (Littell *et al.*, 1996). For convenience, researchers often apply these models even when the distributional assumptions are clearly violated, as when the response semicontinuous, leading to unreliable estimates and standard errors.

Another popular method for longitudinal data is to estimate marginal or population-averaged effects using generalized estimating equations (GEE) (Diggle, Liang, and Zeger, 1994). This method does not impose a full parametric distribution on the response, nor does it require correct specification of the covariance structure for repeated observations within units. GEE methods estimate population-average regression functions for mean response. When the response is semicontinuous, however, estimates of a single regression function may be of dubious value; dual regression functions—one describing the binary split, the other describing the continuous aspect—are probably more appropriate. Another limitation of GEE methods is that they do not handle missing values well; they may be appropriate when the missing values are missing completely at random (MCAR) but not missing at random (MAR) as defined by Rubin (1976), which tends to be quite restrictive.

Analyses of semicontinuous variables in cross-sectional data have appeared in the econometric literature. Manning *et al.*, (1987) and Duan *et al.*, (1983) address a semicontinuous distribution of medical expenses with a two-part model. The observed response is decomposed into two random variables and two regression equations. The first random variable represents whether or not a person has any medical expenses; this probability is modeled with a logit or probit dichotomous regression. The second random variable represents the amount of medical expenses given that the person had any. The parameters in the two regression equations are functionally independent and can be estimated separately using standard methods.

An extension of this two-part model forms the conceptual basis of our model for semicontinuous longitudinal data. The longitudinal semicontinuous

Maren K. Olsen is graduate student, Department of Statistics, The Pennsylvania State University, University Park, PA 16802. Joseph L. Schafer is Assistant Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802. This research was supported by grant 1-P50-DA10075-01 from the National Institute on Drug Abuse. Thanks to John Graham for providing data from the AAPT study.

response, Y_{ij}, can be recoded into two variables,

$$U_{ij} = \begin{cases} 1 & \text{if } Y_{ij} > 0, \\ 0 & \text{if } Y_{ij} = 0, \end{cases}$$

and

$$V_{ij} = \begin{cases} g(Y_{ij}) & \text{if } Y_{ij} > 0, \\ \text{missing} & \text{if } Y_{ij} = 0, \end{cases}$$

where $j = 1, \ldots, n_i$ indexes the time points for individual $i = 1, \ldots, m$ and g is a monotone increasing function (e.g. log) that will make V_{ij} approximately normally distributed. A multilevel logistic model for U_{ij}, and a multilevel linear model for V_{ij} using only the occasions where $Y_{ij} > 0$, can be fit separately with existing software (e.g. HLM). Fitting these models separately, however, does not allow for relationships between the two parts of the data, implying that U_{ij} is independent of $V_{ij'}$, $j \neq j'$. For example, in a study of adolescent alcohol use, this assumption means that whether or not a student uses alcohol in seventh grade does not influence his or her amount of use in eighth grade. Not allowing for relationships between the two parts of the data is equivalent to giving identical treatment to values of V_{ij} that are unseen because $Y_{ij} = 0$ and values of V_{ij} that are unseen because Y_{ij} is truly missing.

We propose to model incomplete longitudinal semicontinuous data by fitting correlated longitudinal models for the binary and continuous parts of the response. By working with U_{ij} and V_{ij}, we will be able to express a dual set of relationships among the Y_{ij}'s across time, and a dual set of relationships between the response variable and other covariates.

2. Proposed Model

As shown in Figure 2, each individual has two correlated growth curves — one for the logit probability of $U_{ij} = 1$ and one for the mean response, $E(V_{ij})$ for the occasions when $U_{ij} = 1$.

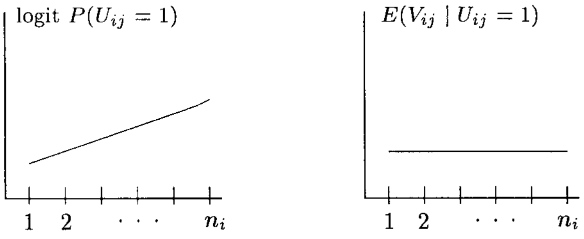

Figure 2: An individual's two growth curves

The logit model is

$$\eta_i = \text{logit}(\pi_i) = X_i^T \beta + Z_i^T c_i,$$

where $U_{ij} \sim \text{Bernoulli}(\pi_{ij})$, $\pi_{ij} = P(U_{ij} = 1)$, and $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots)^T$. The linear model is

$$V_i = X_i^{*T} \gamma + Z_i^{*T} d_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2 I)$ and V_i is the vector of V_{ij} for all j such that $U_{ij} = 1$. The random coefficients, c_i and d_i, are possibley correlated,

$$b_i = \begin{pmatrix} c_i \\ d_i \end{pmatrix} \sim N\left(0, \psi = \begin{pmatrix} \psi_{cc} & \psi_{cd} \\ \psi_{dc} & \psi_{dd} \end{pmatrix}\right).$$

This model has several desireable properties. The intercepts and slopes for each curve can be either fixed or random, and additional covariates (either static or time-varying) may be included in either curve. In addition, because the random effects of each curve are correlated, this model can describe possible relationships between the binary and continuous parts of the data.

In many situations, the hypotheses of primary interest will focus on β and γ, the fixed effects for the logit and linear models, respectively. It may also be useful, however, to examine ψ, the covariances of the subject-specific features, and to test hypotheses of interest — for example, the hypothesis that the models are separable ($\psi_{cd} = 0$).

3. Strategies for model fitting — the EM algorithm

Because our proposed model has logit and linear parts, obtaining parameter estimates and standard errors poses a unique challenge. Some preliminary approaches that we have investigated include Markov chain Monte Carlo (Gilks, *et al.*, 1994) and an approximate EM algorithm. In this paper we will present results from the approximate EM algorithm, in which the E-step is carried out by importance sampling (for an example of importance sampling, see Gelman *et al.*, 1995).

The EM algorithm (Dempster, Laird, and Rubin, 1977) is a common technique for parameter estimation in incomplete-data problems. Traditional approaches of EM to growth curve models treat the random effects as missing data. We maximize the expected loglikelihood which assumes that the random effects are observed (M-step). This expectation is taken with respect to the conditional distribution of the random effects given the observed data and the other model parameters (e.g. β or ψ) fixed at their most recent estimates (E-step). The EM algorithm iterates between the E-step and the M-step until the difference between the parameter estimates at each iteration is negligible.

Under our proposed model, the likelihood function based on the assumption that the random effects are observed can be written as

$$
\begin{aligned}
L_A \propto \\
&\prod_{i=1}^{m} |\psi|^{-1/2} \exp \frac{-1}{2} b_i^T \psi^{-1} b_i \\
&\times \prod_{j=1}^{n_i} \left[\pi_{ij}^{U_{ij}} (1 - \pi_{ij})^{1-U_{ij}} \right] \\
&\times \prod_{j^*=1}^{n_i} \exp \left(\frac{1}{\sigma^2} (V_{ij} - A_{ij})^T (V_{ij} - A_{ij}) \right), \qquad (1)
\end{aligned}
$$

where $A_{ij} = X^{*T}_{ij}\gamma + Z^{*T}_{ij} d_i$ and j^* indicates only those time points j for which $U_{ij} = 1$. The expected loglikelihood can be written as the sum of three parts,

$$
\begin{aligned}
E(l_A) \propto \\
&E\left[-\frac{m}{2} \log |\psi| - \frac{1}{2} \sum_{i=1}^{m} b_i^T \psi^{-1} b_i \right] + \\
&E\left[\sum_{i,j} U_{ij}\eta_{ij} - \log(1 + \exp(\eta_{ij})) \right] - \\
&E\left[\frac{N^*}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i,j^*} (V_{ij} - A_{ij})^T (V_{ij} - A_{ij}) \right], (2)
\end{aligned}
$$

where $N^* = \sum_{i=1}^{m} n_i$ for those time points j for which $U_{ij} = 1$. Notice that $E(l_A)$ can be factored into three distinct functions of ψ, β, and (γ, σ^2), respectively; therefore, we can maximize each of the three parts of $E(l_A)$ separately.

3.1 The E-step

In contrast to a standard growth curve model, which assumes the errors are normally distributed, the multidimensional integral associated with $E(l_A)$ cannot be evaluated directly. Gelman *et al.* (1995) describe several computational techniques to approximate the integral expression, including *importance sampling.* The part of $E(l_A)$ which is a function of β can be written as

$$\sum_{i=1}^{m} \int h(c_i) q(c_i | \beta, \gamma, \sigma^2, \psi, \text{data}) dc_i \qquad (3)$$

where $h(c_i) = \sum_{j=1}^{n_i} (U_{ij}\eta_{ij} - \log(1 + \exp(\eta_{ij})))$. The marginal distribution of the random effects for the logit part of the model, $q(c_i|\beta, \gamma, \sigma^2, \psi, \text{data})$, is nonstandard, so we cannot obtain draws from it directly. Details of this distribution are given in a technical report (Olsen and Schafer, 1998). Instead, we can choose a standard distribution, $g(c_i)$ from which we can sample K random draws of c_i and approximate the integral given in (3) by

$$\sum_{i=1}^{m} \frac{\frac{1}{K} \sum_{k=1}^{K} h(c_i^{(k)}) w(c_i^{(k)})}{\frac{1}{K} \sum_{k=1}^{K} w(c_i^{(k)})}. \qquad (4)$$

The importance ratios, $w(c_i^{(k)})$ are defined to be

$$w(c_i^{(k)}) = \frac{q(c_i^{(k)} | \beta, \gamma, \sigma^2, \psi, \text{data})}{g(c_i^{(k)} | \beta, \gamma, \sigma^2, \psi, \text{data})}.$$

Gelman *et al.* (1995) note that importance sampling works best if the ratio hq/g is fairly constant.

We take $g(c_i)$ to be a multivariate t-distribution with df=4, mean equal to the maximum likelihood (ML) estimate, $\hat{c}_i$, and covariance matrix equal to the inverse Hessian of the loglikelihood. At each iteration of the EM algorithm, the mean and covariance matrices for g are calculated via the Newton-Raphson algorithm. Details of this algorithm are in Olsen and Schafer (1998).

From equation (2), we can see that the remaining two parts of $E(l_A)$ (the functions of ψ and γ, σ^2) require the calculation of $E(b_i b_i^T)$ and $E(d_i)$. It can be shown that

$$d_i | c_i \sim N\left(\hat{d}_i, \sigma^2 (Z^*{}_i Z^{*T}_i + \sigma^2 H^{-1})^{-1} \right),$$

where

$$
\begin{aligned}
\hat{d}_i \quad = \quad & \psi_{dc}\psi_{cc}^{-1} c_i + \left(\sigma^2 H^{-1} + Z_i^* Z_i^{*T} \right)^{-1} \\
& \times \left(Z_i^* (V_i - X_i^{*T}\gamma) - Z_i^* Z_i^{*T} \psi_{dc}\psi_{cc}^{-1} c_i \right)
\end{aligned}
$$

and $H = \psi_{dd} - \psi_{dc}\psi_{cc}^{-1}\psi_{cd}$. Therefore, using the importance sampling results,

$$
\begin{aligned}
E(d_i) \quad &= \quad E(E(d_i | c_i)) \\
&= \quad \frac{\frac{1}{K} \sum_{k=1}^{K} \hat{d}_i^{(k)} w(c_i^{(k)})}{\frac{1}{K} \sum_{k=1}^{K} w(c_i^{(k)})}. \qquad (5)
\end{aligned}
$$

If we write out the expression for $E(b_i b_i^T)$ in terms of c_i and d_i, it is clear that we just need to find $E(c_i c_i^T), E(c_i d_i^T)$, and $E(d_i d_i^T)$. Applying the rules of taking the expectation of a conditional expectation and importance sampling, expressions for these can be found in the same way as for $E(d_i)$. For details, see Olsen and Schafer (1998).

3.2 The M-step

As stated previously, we can maximize the three parts of $E(l_A)$ separately. Expressions for finding the maximum values for ψ, γ, and σ^2 are attained by taking the derivative of $E(l_A)$, setting it equal to 0, and solving for the parameter of interest. The part involving ψ is maximized at $\hat{\psi} = m^{-1} \sum_{i=1}^{m} E(b_i b_i^T)$. The maximum for γ has the standard least-squares form,

$$\hat{\gamma} = \left(\sum_{i,j*} X^*{}_{ij} X^*{}_{ij}^T \right)^{-1} \sum_{i,j*} X^*{}_{ij} (V_{ij} - Z^*{}_{ij}^T E(d_i)).$$

The maximum for σ^2 is given by

$$\begin{aligned} \hat{\sigma}^2 \quad = \quad & \frac{1}{N^*} \sum_{i=1}^{m} \sum_{j^*=1}^{n_i} (V_{ij} - (X^*{}_{ij}^T \gamma^{(t)})^2 - \\ & 2(V_{ij} - (X^*{}_{ij}^T \gamma^{(t)}) Z^*{}_{ij}^T E(d_i) \mathrm{tr} Z^*{}_{ij} Z^*{}_{ij}^T E(d_i d_i^T) \end{aligned}$$

where $\gamma^{(t)}$ is the most recent value of γ in the $(t+1)^{st}$ iteration of the EM algorithm.

Maximizing $E(l_A)$ with respect to β requires an application of the Newton-Raphson algorithm. The maximum, $\hat{\beta}$, can be found numerically by repeated application of

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial E(l_A)^{(t)}}{\partial \beta \partial \beta^T} \right]^{-1} \frac{\delta E(l_A)^{(t)}}{\partial \beta}.$$

where

$$\frac{\partial E(l_A)}{\partial \beta \partial \beta^T} = \frac{1}{K} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \bar{w_i}^{-1} w(c_i^{(k)}) \pi_{ij}^{(k)} (1 - \pi_{ij}^{(k)})$$

and

$$\begin{aligned} \frac{\delta E(l_A)}{\partial \beta} \quad = \quad & \left[\left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} U_{ij} X_{ij}^T \right)^T - \right. \\ & \left. \frac{1}{K} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \bar{w_i}^{-1} w(c_i^{(k)}) X_{ij} \pi_{ij}^{(k)} \right]. \end{aligned}$$

When the Newton-Raphson algorithm converges, $\hat{\beta} = \beta^{(t+1)} \approx \beta^{(t)}$.

4. Example

The data in this example come from a panel of the Adolescent Alcohol Prevention Trial (AAPT) (Hansen and Graham, 1991). In one wave of AAPT, measurements were collected from 3,581 fifth graders in public schools of Los Angeles county. The students were re-surveyed each year in grades 6-10. The response variable of interest is a composite measure of reported recent alcohol use. In this measure, 0 represents no recent alcohol use or sips for religious purposes only. The histograms in Figure 3 show the distribution of the response at each grade. The missing values are denoted NA. There is a high concentration of values at zero and a positive continuous right skewed distribution. Notice, also, that at each successive year the proportion of missing responses increases due to attrition.

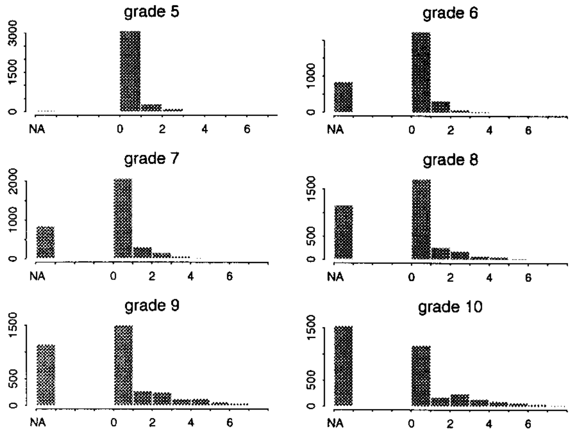

Figure 3: Histograms of reported recent alcohol use

In our linear and logit models, girls are coded as 0 and boys as 1. Time is coded 0 to 5, where 0 represents grade 5 and 5 represents grade 10. As a result, the intercepts in the both the linear and logit models represent the average level at grade 5.

Our design matrices for the logit model have four fixed effects and a random intercept. The fixed effects are an intercept, time, sex, and sex by time. The linear model has exactly the same form, except that there is an additional random effect allowing the slopes to vary by individual. For this model, the approximate EM algorithm took 341 iterations to converge using a maximum relative parameter change of 0.001. Histograms of the importance ratios indicate that importance sampling provided a good approximation.

5. Results

Parameter estimates for the fixed effects in the logit ($\hat{\beta}$) and linear ($\hat{\gamma}$) models are shown in Table 1. Recall that the logit model is for the probability of any recent alcohol use (except sips for religious purposes). The linear part models the expected amount of recent alcohol use among those who reported use. Figure 4 displays the average trends over time for boys and girls for the logit-probability of any recent alcohol use (shown on the probability scale) and the expected amount of recent alcohol use among those who reported use. Notice that boys show a higher probability of use initially, but the girls' probability increases faster so that beyond seventh grade girls have a higher probability of reporting any recent al-

	$\hat{\beta}$	$\hat{\gamma}$
intercept	-3.22	-.370
sex	0.408	0.021
time	0.604	0.194
sex*time	-0.161	0.0085

Table 1: Parameter estimates of the fixed effects

cohol use. Girls and boys have similar increasing trends of the mean amount of recent alcohol use, but the boys' use is consistently higher and increases at a slightly faster rate than the girls' use over time.

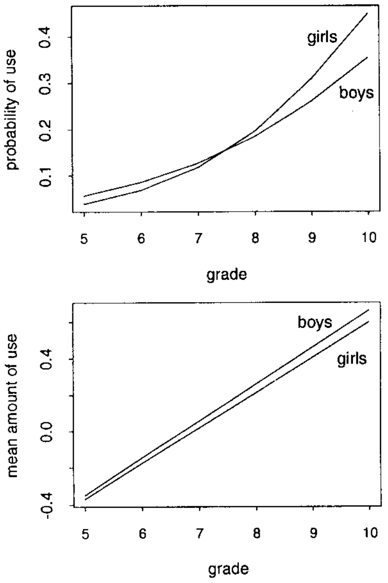

Figure 4: Average growth curves for boys and girls

To gain insight into the relationship between the two parts of the model, we can look at the variance components of the random effects.

$$\hat{\psi} = \begin{pmatrix} 2.32 & 0.316 & 0.056 \\ 0.316 & 0.203 & -0.024 \\ 0.056 & -0.024 & 0.010 \end{pmatrix}$$

Recall that the logit part of the model only has a random intercept, while the linear part has both a random intercept and slope. The relatively small variance component for the random slope of the linear model (0.010) indicates that including a random slope may not be necessary. So, we restrict our attention of the ψ_{cd} matrix to just the first element — the covariance between the random intercepts of the linear and logit model (estimated at 0.316). The correlation between the random intercepts for the linear and logit models is 0.46, providing us with strong evidence that modeling the two parts of the data separately is not appropriate.

6. Future Work

We regard this EM algorithm as a preliminary approach. Simulation work has shown that it is computationally accurate but too slow for practical use. Our future efforts will focus upon developing faster algorithms for parameter estimation.

Researchers have addressed analyzing binary longitudinal data under the generalized linear mixed model using a variety of methods, including penalized quasi-likelihood (e.g. Goldstein and Rabash, 1996; Lin and Breslow, 1996; Wolfinger and O'Connell, 1993) and MCMC methods (e.g. McCulloch, 1997; Zeger and Karim, 1991). Recently, Raudenbush and Yang (under review) have developed an algorithm which implements a Taylor series and Laplace approximation to evaluate the likelihood before maximizing it using Fisher scoring; they have found this approach to be both accurate and computationally fast. Accurate and efficient algorithms for the linear part of the model are considerably more developed. Schafer (under review) reviews standard methods for general linear mixed models and derives a new set of procedures (a combination of EM and scoring) which significantly speed up conventional algorithms for ML estimation. We plan to modify these current likelihood-based methods to incorporate the non-standard type of missing data in V_{ij} and the assumption that the random effects for the logit and linear parts of the model are correlated. The implementation of these full-likelihood methods will enable us to test hypotheses about the covariance parameters of the random effects and other types of goodness of fit measures in addition to hypotheses about the fixed effects. Finally, we plan to incorporate these algorithms into a software that researchers will be able to use to analyze semicontinuous longitudinal data.

7. References

Bryk, A. S. and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101, 147-158.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1996). *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*, Scientific Software International, Inc., Chicago.

Diggle, P., Liang, K. and Zeger, S. (1994). *Analysis of longitudinal data.* New York: Oxford University Press.

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 Series B*, 39, 1-38.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, 1, 115-126.

Gelfland, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, New York.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). Introducing Markov chain Monte Carlo. *Markov Chain Monte Carlo in Practice*, (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), Chapman & Hall, New York.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society* A, 159, 505-513.

Hansen, W.B. and Graham, J.W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, 20, 414-430.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91, 1007-1016.

Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014-1022.

Littell, R. C. Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models.* SAS Institute, Inc., Cary, NC.

Manning, W.G., Duan, N., and Rogers, W.H. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, 35, 59-82.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170.

Multilevel Models Project (1996). *Multilevel Modeling Applications - a Guide for Users of MLn.* (ed. Geoff Woodhouse) Institute of Education, University of London.

Olsen, M.K. and Schafer, J.L. Parameter estimates for semicontinuous longitudinal data using an approximate EM algorithm. Technical Report 98-31, The Methodology Center, The Pennsylvania State University, October 1998.

Raudenbush, S.W. and Yang, M. (under review). Maximum likelihood for hierarchical models via high-order, multivariate Laplace approximation.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Schafer, J.L. (under review). Some improved procedures for linear mixed models.

Wolfinger, R.D. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulations*, 48, 233-243.

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.