Statistical Science
2004, Vol. 19, No. 2, 219–238
DOI 10.1214/088342304000000305

Conditional and Marginal Models: Another View

Youngjo Lee and John A. Nelder

Abstract. There has existed controversy about the use of marginal and conditional models, particularly in the analysis of data from longitudinal studies. We show that alleged differences in the behavior of parameters in so-called marginal and conditional models are based on a failure to compare like with like. In particular, these seemingly apparent differences are meaningless because they are mainly caused by preimposed unidentifiable constraints on the random effects in models. We discuss the advantages of conditional models over marginal models. We regard the conditional model as fundamental, from which marginal predictions can be made.

Key words and phrases: Generalized linear model, hierarchical generalized linear model, joint modeling of mean and dispersion, spatial correlation, temporal correlation.

1. INTRODUCTION

In longitudinal studies, models for repeated measurements, constructed directly to describe marginal means and treating any covariance structure as nuisance parameters, have come to be widely used. These marginal (or so-called population-average) models are often contrasted with conditional (subject-specific or random-effect or multilevel) models. The principal distinction between marginal and conditional models has often been asserted to depend on whether the regression coefficients are to describe an individual's response or the marginal response to changing covariates, that is, one that does not attempt to control for unobserved subjects' random effects. For example, a marginal gender contrast compares the mean among men to that among women, while a conditional gender contrast compares the mean among men to that among women holding the same value of a random effect (a particular value that corresponds to each individual). Diggle, Liang and Zeger (1994) recommended the random-effect model for inferences about individual responses and the marginal model for inferences about margins, that is, the objectives (or the type of inferences) in a study should determine which suitable statistical model to use. By contrast, we see the analysis process as consisting of two main activities: the first is model selection, which aims to find parsimonious well-fitting models for the basic responses being measured; the second is model prediction, where estimates from selected models are used to predict quantities of interest and their uncertainties. In our view, inferences about both margins and individual subjects' responses belong to the prediction phase of the analysis. We show that alleged differences in the behavior of regression coefficients in so-called marginal and conditional models are based on a failure to compare like with like. We shall see that these differences are mainly caused by the choice of unidentifiable constraints on the random effects. To compare two different models, we must compare analogous quantities. We show that different constraints can lead to seemingly very different, but inferentially identical, models. We believe the conditional model is the basic model and that any conditional model leads to a specific marginal model. We work within a framework of conditional models derived from hierarchical generalized linear models (HGLMs; Lee and Nelder, 1996, 2001a), and marginal models derived in turn from these conditional models.

Youngjo Lee is Professor, Department of Statistics, Seoul National University, Seoul, Korea (e-mail: youngjo@plaza.snu.ac.kr). John A. Nelder is Visiting Professor, Department of Mathematics, Imperial College, London SW7 2BZ, UK (e-mail: j.nelder@ic.ac.uk).

Marginal models have often been fitted using generalized estimating equations (GEEs), whose drawbacks are also discussed.

2. CONDITIONAL VERSUS MARGINAL MODELS

In this section we discuss why random-effect models should be preferred to marginal models. Consider two normal models: one is a random-effect model

$$Y_{ij} = X_{ij}\beta + v_i + e_{ij}, \tag{1}$$

where $v_i \sim N(0, \lambda)$ is a random effect and $e_{ij} \sim N(0, \phi)$; the other is a marginal model

$$E(Y_{ij}) = X_{ij}\beta, \tag{2}$$

where the parameters in $\text{var}(Y) = \Sigma$ are nuisance parameters that have an arbitrary chosen pattern. Zeger, Liang and Albert (1988) pointed out that given only (2), their GEE solution is consistent. An obvious advantage of using random-effect models is that they allow conditional inferences in addition to marginal inferences (Robinson, 1991). With model (1) we can obtain not only a conditional mean

$$\mu^c_{ij} = E(Y_{ij}|v_i) = X_{ij}\beta + v_i,$$

but also the marginal mean

$$\mu_{ij} = E(\mu^c_{ij}) = E(Y_{ij}) = X_{ij}\beta,$$

while with the marginal model (2), we can obtain only the marginal mean μ_{ij}. The conditional model (1) is a basic model, which leads to a specific marginal model $Y \sim N(X\beta, \Sigma)$, that is, a multivariate normal model with a specific covariance structure. In this paper, we distinguish the marginal model (2), where the distribution of Y is given an arbitrary covariance structure, from a true multivariate model.

It may be reasonable to assume that an individual's unobservable trait (v_i) follows a certain distribution. However, the center of this distribution cannot be identified because it is confounded with the intercept term. Thus, in the random-effect model (1) we put the unidentifiable constraints $E(v_i) = 0$ and $E(e_{ij}) = 0$ as we do for error terms in linear models. This puts constraints $\sum_i \hat{v}_i = 0$ and $\sum_{ij} \hat{e}_{ij} = 0$ in any corresponding estimating procedure (e.g., that of Lee and Nelder, 1996). In this paper, we show that these constraints on random components are crucial if the βs in the models (1) and (2) are to be comparable in general. It is the constraints $E(v_i) = 0$ and $E(e_{ij}) = 0$ in the random-effect model (1) that lead to $E(Y_{ij}) = X_{ij}\beta$ in the marginal model (2), so that the βs in models (1) and (2) share a common meaning.

A marginal mean and a population average have often been assumed to mean the same thing. In this paper, we interpret the marginal mean to be the mean obtained by integrating out individuals' heterogeneities, and this will be a population average if and only if the individuals in the study can be regarded as a random sample from a population. Thus β cannot be interpreted as a population average unless the subjects can be considered as a representative random sample from a population: See the detailed discussion in Lindsey and Lambert (1998) about why the subjects in longitudinal studies often cannot be considered as a representative random sample (e.g., because they are volunteers). However, β can still be interpreted as the marginal effect for the covariate X, eliminating the heterogeneities of individual units.

In random-effect models, the difference between the conditional mean μ^c_{ij} and the marginal mean μ_{ij} of an individual is the random effect

$$v_i = E(Y_{ij}|v_i) - E(Y_{ij}).$$

There are two ways to define a marginal mean $E(Y_{ij})$ from the conditional mean $E(Y_{ij}|v_i)$: Either we take the expectation over v_i to give

$$\mu_{ij} = E(\mu^c_{ij}) \tag{3}$$

or take the value at $v_i = 0$ to give

$$\mu_{ij} = \mu^c_{ij}|_{v_i=0}. \tag{4}$$

Because individuals' deviations have been eliminated in (3) by integration, the marginal mean μ_{ij} does not involve any individual. However, in (4), μ_{ij} is the response of a notional individual with a null random effect, that is, one at the center of the distribution of random effects. With normal models the two marginal means in (3) and (4) are the same, but this is not generally true in nonlinear models (Crowder and Hand, 1990). In the next section, we show that for HGLMs both forms of marginal means are possible, but only on particular scales of the mean parameters.

We prefer the random-effect model. First, we can have a simple marginal interpretation using the conditional model not only in normal random-effect models, but also in a wider class. Second, ignoring important random effects may render invalid many traditional techniques of statistical analysis (Goldstein, 1995). Consider the two models

$$Y_{ijk} = \beta_0 + \beta_j + v_i + e_{ijk} \tag{C1}$$

and

$$Y_{ijk} = \beta_0 + \beta_j + v_i + v_{ij} + e_{ijk}, \tag{C2}$$

where β_0 is the intercept, β_j are fixed treatment effects, $v_i \sim N(0, \lambda_1)$ are random subject effects, $v_{ij} \sim N(0, \lambda_2)$ are random treatment–subject interactions and $e_{ijk} \sim N(0, \phi)$. The common marginal model M that corresponds to C1 and C2 has the form

$$\text{(M)} \qquad E(Y_{ijk}) = \beta_0 + \beta_j$$

with an arbitrary Σ. Users of such a marginal model cannot check assumptions about the covariance structure, so many advocates of marginal approaches treat covariance structures as nuisance parameters, claiming that their methods are insensitive to assumptions about Σ. However, C1 and C2 are qualitatively very different models, and ignoring differences between them could lead to wrong conclusions. When C1 is true, that is, there are no treatment–random-effect interactions ($v_{ij} = 0$), and if the subject effects v_i are not of inferential interest, we can use the marginal model M. However, in the presence of treatment–subject interactions (i.e., when C2 is true), marginal parameters in M may have a misleading interpretation in terms of treatment effects, exhibiting the so-called Simpson (1952) paradox. Lindsey and Lambert (1998) provided an example where a treatment can be superior on the average, while being poorer for every individual. They further listed no less than eight drawbacks of such a marginal approach and concluded that

> ... the "statistical" argument that we should directly model margins if scientific interest centres on them, is not acceptable on scientific grounds, for it implies that we are generally imposing more unrealistic physiological mechanisms on our data than by direct conditional modelling and that these are most likely rendering simple marginal models greatly biased.

For example, with a heterogeneous population the usual marginal model can show a long-term decreasing risk of adverse events under the treatment because that treatment has killed off the more frail subjects. Thus, the use of marginal models can be dangerous, even when marginal inferences are of interest. The usefulness of marginal inferences requires the absence of interactions, checkable only via conditional models.

The random-effect models leads to an equivalent multivariate model whose distribution of Y is obtained by integrating out the random effects v from the joint distribution of Y and v. However, the integration necessary to obtain the multivariate distribution inevitably uses model assumptions about the form of the random effects, which then become impossible to check given just that multivariate distribution. For example, in the random-effect model (1), the errors are decomposed into two independent components ($e_{ij} = Y_{ij} - X_{ij}\beta - v_i$ and v_i), so that model checking can be done separately for each component (Lee and Nelder, 2001a), while with the errors ($Y_{ij} - X_{ij}\beta$) of the corresponding multivariate model, such model checking is difficult, maybe impossible: See Lee and Nelder (2001b) for model checking with various spatial and temporal correlations.

3. POISSON HGLMs

In this section we study parametrizations of random effects and show that each parametrization leads to simple models about margins on a particular scale. In contrast to the parametrization of fixed effects, those for random effects do not seem to be well known. To clarify the argument, we start with a data set which can be modelled with a Poisson HGLM.

Galbraith and Laslett (1993) considered a set of data that comprise numbers of spontaneous and induced fission tracks (Y_{i1}, Y_{i2}) counted in matched areas (A_i) of crystal and mica for 27 (I) zircon crystals. Spontaneous tracks form over geological time by spontaneous fissions of trace ^{238}U. Induced tracks are created artificially by placing the sample in a nuclear reactor and bombarding it with thermal neutrons, a measured proportion of which collide with trace ^{235}U atoms, thereby causing fission. This indirectly measures the amount of uranium in the crystal (see Table 1).

TABLE 1
Number of spontaneous and induced fission tracks counted in matched areas for 27 zircon crystals

Crystal	Y_{i1}	Y_{i2}	A_i	Crystal	Y_{i1}	Y_{i2}	A_i
1	24	459	80	15	2	70	49
2	8	52	30	16	3	94	28
3	136	310	30	17	23	128	60
4	56	257	70	18	153	264	70
5	3	57	70	19	90	143	32
6	6	332	80	20	31	49	16
7	73	98	14	21	38	120	40
8	131	226	50	22	51	46	25
9	9	173	80	23	38	85	12
10	6	28	12	24	127	45	20
11	141	229	70	25	5	24	30
12	11	74	36	26	24	56	20
13	12	61	18	27	10	31	18
14	10	28	40				

A natural model for such data would be that the counts have Poisson distributions with means that vary both within and between pairs. We can model these data using two Poisson HGLMs. One is the Poisson–normal model (PN1)

$$Y_{ij}|v_i \sim P(\mu^n_{ij}),$$
$$\eta^n_{ij} = \log(\mu^n_{ij}) = \log A_i + \beta^n + v_i, \quad v_i \sim N(0, \lambda^n),$$

where $P(\mu^n_{ij})$ means the Poisson distribution with mean $\mu^n_{ij} = E(Y_{ij}|v_i)$. The other is the Poisson–gamma model (PG1)

$$Y_{ij}|u_i \sim P(\mu^g_{ij}),$$
$$\mu^g_{ij} = A_i \exp(\beta^g) u_i, \quad u_i \sim G(1, \lambda^g),$$

where $\mu^g_{ij} = E(Y_{ij}|u_i)$ and $G(1, \lambda^g)$ denotes the gamma distribution with mean 1 and variance λ^g. Here we use superscripts n and g to refer to normal and gamma distributions of random effects, respectively.

In the multiplicative model PG1, the log link leads to an additive model with linear predictor

$$\eta^g_{ij} = \log \mu^g_{ij} = \log A_i + \beta^g + \log u_i.$$

Each conditional model leads to a specific marginal model. Model PG1 leads to the marginal model

$$\log E(Y_{ij}) = \log E(\mu^g_{ij}) = \log A_i + \beta^g,$$

while PN1 leads to the marginal model

$$E\{\log(\mu^n_{ij})\} = \log A_i + \beta^n.$$

These models are different because the operators E and log do not commute. The definition of marginal means depends on the scale on which the margins are formed. For example, when $\mu_{ij} = E(Y_{ij})$, $g(\mu_{ij})$ is the marginal mean of $g(Y_{ij})$ only if $g(\cdot)$ is a linear transformation. Similarly, $\log(\mu_{ij})$ is not a marginal mean, but a nonlinear transformation of the marginal mean μ_{ij}, while $E(\log(\mu^n_{ij}))$ with $\mu^n_{ij} = E(Y_{ij}|v_i)$ could be a marginal mean of interest, after integrating out individual heterogeneities. Once the marginal model is formed, we may not be able to make inferences about the conditional means, so conditional models such as PN1 or PG1 are more basic.

Here β^n and β^g describe marginal means, but on different scales, so that they cannot be directly compared. Now suppose that we are interested in estimating the marginal mean $E(Y_{ij})$ of Y_{ij}. In PN1,

$$E(Y_{ij}) = A_i(\exp\beta^n)E(\exp v_i) = A_i \exp\beta^g_n,$$

say, where

$$\beta^g_n = \beta^n + \lambda^n/2. \tag{5}$$

In PG1, $E(Y_{ij}) = A_i(\exp\beta^g)$, so that β^g_n and β^g are comparable. Suppose we are interested in estimating the marginal mean $E(\log(\mu^n_{ij}))$. In PN1, $E(\log(\mu^n_{ij})) = \log A_i + \beta^n$, and in PG1,

$$E(\log(\mu^g_{ij})) = \log A_i + \beta^g + E(\log u_i)$$
$$= \log A_i + \beta^n_g,$$

say, where

$$\beta^n_g = \beta^g + \log\lambda^g + \psi(1/\lambda^g) \tag{6}$$

and $\psi(\cdot)$ is the digamma function. So β^n and β^n_g are comparable. For comparable quantities such as (β^n and β^n_g) or (β^g and β^g_n) the differences in our example are not very great (Table 2) compared with the difference between β^n and β^g. In Table 2, for gamma random effects, we use Lee and Nelder's (2001a) second-order Laplace method to estimate the dispersion components; for normal random effects, we use the first-order method, which works well with these models.

In an additive model such as PN1, the location of v_i is unidentifiable since $\beta + v_i \equiv (\beta + a) + (v_i - a)$ for $i = 1, \ldots, I$, while in a multiplicative model such as PG1, the scale of u_i is unidentifiable since $(\exp\beta)u_i \equiv (a\exp\beta)(u_i/a)$ for $a > 0$. When we form random-effect models we can impose a constraint either on the fixed effects or on the random effects. However, imposing constraints on random effects is more convenient when we move to models with more than one random component. Thus, in the additive model we may use the unidentifiable constraint $E(v_i) = 0$ and in the multiplicative model use $E(u_i) = 1$; this results in constraints on the estimators $\sum \hat{v}_i/I = 0$ and $\sum \hat{u}_i/I = 1$, respectively (Lee and Nelder, 1996). Remember that if, in linear models, we put a constraint on fixed effects such as $\sum \beta_i/I = 0$, least-squares estimates similarly satisfy $\sum \hat{\beta}_i/I = 0$.

In linear models we sometimes put constraints on parameters, but any relevant inferential quantities must

TABLE 2
Parameter estimates from two HGLMs

PN1	$\hat{\beta}^n = 0.629$	$\hat{\beta}^g_n = 0.905$	$\hat{\lambda}^n = 0.5515$
PG1	$\hat{\beta}^n_g = 0.609$	$\hat{\beta}^g = 0.870$	$\hat{\lambda}^g = 0.4847$

be independent of the constraints imposed. For example, we do not compare two estimates of $\hat{\beta}_i$ that arise from two different constraints (e.g., $\beta_1 = 0$ and $\sum \beta_i / I = 0$), because different constraints lead to different meanings for the parameter estimates; for example, $\hat{\beta}_i$ under $\beta_1 = 0$ estimates $\beta_i - \beta_1$, while under $\sum \beta_i / I = 0$ it estimates $\beta_i - \sum \beta_i / I$. Nelder (1994) discussed the unnecessary complexity caused by treating such constraints as an intrinsic property of linear models. Similarly, we should not treat constraints on random components as intrinsic properties of models. Consequently, parameters from random-effect models that have different constraints on the random components cannot be compared directly because they have different meanings. In PN1, it is the log scale on which the marginal mean $E(\log(\mu_{ij}^n))$ can be interpreted as the response of a notional individual having $\log \mu_{ij}^n|_{v_i=0} = \log A_i + \beta^n$, while in PG1, it is the original scale on which the marginal mean $E(\mu_{ij}^g)$ can be interpreted as that of a typical individual with an average random effect $\mu_{ij}^g|_{u_i=1} = A_i \exp(\beta^g)$.

Care is necessary with inferences about quantities with constraints imposed. For example, in making a profile likelihood for β_i, any constraint on the β_is should be kept. Otherwise, the resulting likelihood inferences do not compare the different values of parameter estimates from the same model. Similarly, in making valid profile-likelihood inferences for individual responses, constraints on random effects should also be kept (Lee and Nelder, 2002).

Note that the Poisson–normal model with a marginal parametrization (PNM),

$$Y_{ij}|v_i \sim P(\mu_{ij}^n),$$

$$\mu_{ij}^n = A_i \exp(\beta_n^g + v_i^*) \quad \text{with } v_i^* \sim N(-\lambda^n/2, \lambda^n),$$

provides the marginal mean $E(Y_{ij}) = \mu_{ij} = A_i \exp \beta_n^g$, which has the constraint $E(u_i^*) = 1$ with $u_i^* = \exp(v_i^*)$, while the Poisson–gamma model with a conditional parametrization (PGC),

$$Y_{ij}|u_i \sim P(\mu_{ij}^g), \quad \log(\mu_{ij}^g) = \log A_i + \beta_g^n + v_i^{**},$$

where $v_i^{**} = \log(u_i) - \log \lambda^g - \psi(1/\lambda^g)$ and $u_i \sim G(1, \lambda^g)$, provides the marginal mean $E(\log(\mu_{ij}^g)) = \log A_i + \beta_g^n$, which has the constraint $E(v_i^{**}) = 0$. Thus PNM gives parameters for $E(Y_{ij})$, while PGC gives those for $E(\log(\mu_{ij}^g))$. Thus, it is the constraint on the first moment of the random effects which determines whether the parameters contribute to $E(Y_{ij})$ or $E(\log(\mu_{ij}^g))$, not the shape of random-effect distribution. In all four models, there is a particular scale on which one obtains simple marginal interpretations: PG1 and PNM provide the marginal mean $E(Y_{ij})$, while PN1 and PGC provide $E(\log(\mu_{ij}^c))$.

Models PN1 and PNM (and similarly PG1 and PGC) are equivalent, but with different parametrizations, providing identical inferences for the same thing. Estimates of β^n and β_g^n should be integrated over the random-effect distribution before any comparison is made with estimates of β_n^g and β^g. Thus, estimates of $\beta_n^g = \log\{E(Y_{ij})/A_i\}$ and $\beta^n = E[\log\{E(Y_{ij}|v_i)/A_i\}]$ (and also $\beta^g = \log\{E(Y_{ij})/A_i\}$ and $\beta_g^n = E[\log\{E(Y_{ij}|v_i)/A_i\}]$) cannot be directly compared. The difference between the estimates is not caused by differences between the two models (they are equivalent here), but because they predict two different quantities. The difference between $\hat{\beta}^n$ and $\hat{\beta}_n^g$ (or $\hat{\beta}^g$ and $\hat{\beta}_g^n$) is caused by putting different constraints in the equivalent normal (gamma) random-effect models, while the difference between $\hat{\beta}^n$ and $\hat{\beta}_g^n$ (or $\hat{\beta}^g$ and $\hat{\beta}_n^g$) is caused by assuming different random-effect distributions, keeping a common constraint $E(v_i) = 0$ [$E(u_i) = 1$]. Because arbitrary constraints can lead to seemingly very different models, especially in nonlinear cases, care should be exercised when comparing such models. When we compare two different models, we should compare them under common constraints. Models PN1 and PG1 (and similarly PNM and PGC) are similar when heterogeneities of individuals (variances of random effects, λ^g and λ^n) are small. However, when heterogeneities are large, they can be quite different because they assume different shapes for the random-effect distributions.

We (Lee and Nelder, 1996, 2000, 2001a, b) have developed various model-checking procedures. Our 1996 paper extends the scaled deviance test of GLMs, and in these tests the degrees of freedom for random effects are noninteger. Here, for PN1 the scaled deviance is 1994 with 27.49 degrees of freedom; for PG1 the scaled deviance is 1994 with 27.38 degrees of freedom. If the Poisson assumption were appropriate for the $Y_{ij}|v_i$ distribution, the scaled deviance would be close to its degrees of freedom. Thus, its large value implies that there exists extra-Poisson variation in the conditional distribution. Subsequent investigation showed that a term for pair effects should have been added to the fixed effects.

Our final Poisson-normal model (PNF) for Galbraith and Laslett's (1993) data is

$$Y_{ij}|(v_i, v_{ij}) \sim P(\mu_{ij}^n)$$

and

$$\log(\mu_{ij}^n) = \log A_i + \beta^n + \tau_j^n + v_i + v_{ij},$$

where τ_j^n, for $j = 1, 2$, is a pair effect, $v_i \sim N(0, \lambda_1^n)$ and $v_{ij} \sim N(0, \lambda_{2j}^n)$, that is, $\text{var}(v_{ij})$ is changing with j. Our final Poisson–gamma model (PGF) is

$$Y_{ij}|(u_i, u_{ij}) \sim P(\mu_{ij}^g)$$

and

$$\mu_{ij}^g = A_i \exp(\beta^g + \tau_j^g) u_i u_{ij},$$

where τ_j^g is a pair effect, $u_i \sim G(1, \lambda_1^g)$ and $u_{ij} \sim G(1, \lambda_{2j}^g)$. Differences in regression coefficient estimates for the two models are again caused mainly by the parametrizations rather than by the choice of random-effect distribution. For example, in Table 3 the intercept estimates in PNF and PGF have different signs. Note that we use the log scale for dispersion parameters because it often gives near-quadratic profile likelihoods (Lee and Nelder, 1996).

Now suppose that we are interested in inferences about $E(Y_{ij})$. In PNF,

$$\begin{aligned}\mu_{ij}^{*n} &= E(Y_{ij}/A_i) = \exp(\beta^n + \tau_j^n) E(\exp(v_i + v_{ij})) \\ &= \exp\{\beta^n + \tau_j^n + (\lambda_1^n + \lambda_{2j}^n)/2\},\end{aligned}$$

and in PGF,

$$\mu_{ij}^{*g} = E(Y_{ij}/A_i) = \exp(\beta^g + \tau_j^g).$$

We notice some differences in marginal mean predictions between the two models in Table 3. Note that in Poisson HGLMs with the log link, prediction from the normal random-effect model (such as PNF) gives similar results to the use of geometric means, and that from the gamma random-effect model (such as PGF) yields similar results to arithmetic means. Geometric means are preferred if the log scale provides symmetry of responses, while arithmetic means are preferred if the original scale does. Thus, on which scale we should take margins may depend on the shape of the distribution assumed for the responses. Similarly, in HGLMs the shape of the random-effect distribution may determine an appropriate scale of margins for analysis.

TABLE 3
Marginal mean prediction from two models for nuclear data

PNF		PGF	
Estimate	**s.e.**	**Estimate**	**s.e.**
$\hat\beta^n = -0.386$	0.257	$\hat\beta^g = 0.293$	0.242
$\hat\tau_2^n = 1.401$	0.228	$\hat\tau_2^g = 0.893$	0.214
$\hat\gamma_1^n = -0.899$	0.283	$\hat\gamma_1^g = -0.970$	0.287
$\hat\gamma_2^n = 0.266$	0.287	$\hat\gamma_2^g = 0.122$	0.285
$\hat\gamma_{22}^n = -5.734$	2.980	$\hat\gamma_{22}^g = -4.231$	1.261
$\hat\mu_{i1}^{*n} = 0.886$	GM = 0.642	$\hat\mu_{i1}^{*g} = 1.340$	AM = 1.467
$\hat\mu_{i2}^{*n} = 2.763$	GM = 2.727	$\hat\mu_{i2}^{*g} = 3.272$	AM = 3.318

$$\hat\tau_1^n = \hat\tau_1^g = 0 \quad \gamma_1^n = \log(\lambda_1^n) \quad \gamma_1^g = \log(\lambda_1^g)$$
$$\log(\lambda_{2j}^n) = \gamma_2^n + \gamma_{2j}^n \quad \hat\gamma_{21}^n = 0 \quad \log(\lambda_{2j}^g) = \gamma_2^g + \gamma_{2j}^g \quad \hat\gamma_{21}^g = 0$$

NOTE: GM stands for geometric means of y_{ij}/A_i for $j = 1, 2$; AM stands for arithmetic means of y_{ij}/A_i for $j = 1, 2$.

Certain parametrizations may achieve insensitivity of estimates to model assumptions. This is not a matter of different models, but a choice of parametrization within the same model. All the models we have considered for Poisson HGLMs are conditional models, but they provide both marginal and conditional interpretations of the regression coefficients. The conditional interpretation of β^n in PN1 (or of β^g in PG1) depends on a preimposed unidentifiable constraint on the random effects v_i, while the marginal interpretation does not, so that only the meaningful interpretation of β^n (β^g) is a marginal one. The principal distinction in the regression coefficients is caused by parametrization (constraints on random effects), and each parametrization allows a marginal interpretation, but only on a particular scale.

Our final question concerns which parametrizations should be preferred. For example, in choosing between the PN1 and PNM (or PG1 and PGC) parametrizations, we prefer PN1 (PG1) because β^n (β^g) are orthogonal to the dispersion parameters (Lee and Nelder, 1996), whereas β_n^g (β_g^n) are sums of components from orthogonal fixed effects and dispersion parameters; see (5) and (6). However, Heagerty and Zeger's (2000) study indicates that some people may prefer PNM (PG1) because the β_n^g (β^g) parametrization could yield estimates insensitive to covariance assumptions. Further study is required on parametrizations in random-effect models.

4. GEEs

For independent responses, Wedderburn (1974) showed that quasilikelihood (QL) estimating equations can be formed by using assumptions solely about the first two moments. It is sometimes said that use of QL depends on assumptions about the mean–variance relationship only, but this is not so. Lee and Nelder (1999) pointed out that QL estimating equations are the score equations derived from a QL and that the shape of

the distribution follows a pattern of higher-order cumulants similar to that predicted from a one-parameter exponential family if one existed. This makes the resulting QL estimator robust against misspecification of skewness. By replacing the variance with the covariance matrix of responses in the QL estimating equations, this approach can be extended to correlated responses (Liang and Zeger, 1986; Zeger and Liang, 1986). However, for correlated errors, there is in general no QL for which these QL-type estimating equations are score equations (McCullagh and Nelder, 1989). Because of the lack of a likelihood basis, we regard this approach as not being a proper extension of Wedderburn's (1974) QL approach to models with correlated errors. By contrast, random-effect models can give QL's for correlated errors (Lee and Nelder, 2001a).

The GEEs are QL-type estimating equations in which the pattern of the covariance matrix is to some extent arbitrary. Thus, they inherit the lack of a likelihood basis. Estimates of regression coefficients from GEEs have been claimed to be consistent under various model misspecifications as long as the regression equation for the mean is correctly specified (Zeger, Liang and Albert, 1988); however, this was shown by Crowder (1995) to be incompletely established. A referee pointed out that the more specific assumptions one makes about a model, the more likely it is that some of them fail in practice. We support the use of robust methods, but dislike the use of marginal models without probabilistic or likelihood basis, because use of such models makes checking almost impossible. Without proper model checking, there is no simple empirical means to discover whether the regression for the mean has been correctly or, more exactly, adequately specified. Estimates can of course be biased if important covariates are omitted. Without proper model checking, the validity of inference cannot be assured. Note also that an insensitivity of estimates to model assumptions does not necessarily imply that the conclusions are correct. Lee and Nelder (2001b) gave an example where estimates from three different models are similar, but model checking shows that none of them is the right one.

The use of robust procedures and the use of marginal models are separate issues. All robust procedures used in GEEs can also be used for random-effect models. Consider use of the model C2 in Section 2, which assumes that the v_{ij} are independent, when, in fact, they are serially correlated. Obviously in this case the conditional model fails. However, estimates $\hat{\beta}$ from the random-effect model C2 are as robust as the GEE estimator against such misspecification (Seely and Hogg, 1982). If the standard error estimates are unlikely to capture the true variation, we can use the sandwich standard error estimates (Kent, 1982); furthermore, another useful sandwich estimator (Lee, 2002) is available which cannot be derived from marginal models. There is no single robust procedure that protects against all circumstances, that is, robustness is not an absolute quality, so various robust methods need to be developed. For example, the nonparametric maximum likelihood estimator of Laird (1978) is robust against the wrong choice of distribution of random effects, the semiparametric approach against the wrong choice of baseline distribution of survival data (Ha, Lee and Song, 2001) and so forth. Thus, robustness is not necessarily a guarantee of the usefulness of the GEE approach. For a list of drawbacks to the GEE approach, see Lindsey and Lambert (1998).

A referee pointed out that given two approaches that provide similar inferences for equivalent elements, the choices of approach should be decided by the robustness of the method and the stability of the algorithm. We agree that the GEE approach is competitive in both respects. The use of numerically intractable marginal likelihoods causes difficulty in implementing random-effect methods for complicated correlation structures. The use of hierarchical likelihood allows both computationally simple and statistically efficient estimation algorithms (Lee and Nelder, 2001a) that are easily extendable to more general models that allow for various temporal and spatial correlations (Lee and Nelder, 2001b). Developments of various robust procedures and stable algorithms for various correlation structures in random-effect models are fertile areas for further research.

5. BERNOULLI HGLMs

Heagerty and Zeger (2000) analyzed binary data from Weil (1970). The 21-day survival of pups from the litters of 16 exposed and 16 unexposed rats was compared. Let Y_{ij} denote the survival of pup j, $j = 1, \ldots, n_i$, born to animal i, $i = 1, \ldots, m$. The single covariate of interest is a between-subject binary indicator of the treatment assignment of the mother. Let $p_{ij}^c = E(Y_{ij}|v_{ij})$ and $p_{ij} = E(Y_{ij})$ be, respectively, the conditional and marginal probabilities. We consider the binomial-normal model (BN)

$$\begin{aligned} Y_{ij}|v_{ij} &\sim B(p_{ij}^c), \\ \operatorname{logit}(p_{ij}^c) &= \beta_0^c + x_{ij}\beta^c + v_{ij}, \quad v_{ij} \sim N(0, \lambda_j^c), \end{aligned} \tag{7}$$

TABLE 4
Parameter estimates from three models for Weil's data

	OBM		BM		BN		BN0	
Coefficient	**Estimate**	**s.e.**	**Estimate**	**s.e.**	**Estimate**	**s.e.**	**Estimate**	**s.e.**
β								
Intercept	2.183	0.315	2.175	0.286	2.191	0.294	2.340	0.446
Treatment	−0.961	0.518	−1.069	0.476	−0.728	0.608	−0.955	0.607
$\log\phi_{\bar{j}}$								
Intercept	0.356	0.377	−1.593[a]		−1.380	2.071	0.583[b]	0.304[b]
Treatment	1.107	0.525	2.188[c]		2.611	2.097		

[a]Heagerty and Zeger used the $\lambda_j^{0.5}$ scale, so we have transformed their estimate to our $\log\lambda_j$ scale.
[b]$\log\lambda$ scale.
[c]$\log\lambda_j$ scale.

where $B(p)$ means the Bernoulli distribution with a probability p, and also the overdispersed binomial model (OBM)

$$Y_{ij}|v_{ij} \sim \mathrm{OB}(p_{ij}), \quad \mathrm{logit}(p_{ij}) = \beta_0^o + x_{ij}\beta^o,$$

where $\mathrm{OB}(p_{ij})$ means the overdispersed Bernoulli with $p_{ij} = E(Y_{ij})$ and $\mathrm{var}(Y_{ij}) = \phi_j p_{ij}(1 - p_{ij})$. We also consider the binomial–normal model with common heterogeneity λ^c (BN0)

$$Y_{ij}|v_{ij} \sim B(p_{ij}^c),$$
$$\mathrm{logit}(p_{ij}^c) = \beta_0^c + x_{ij}\beta^c + v_{ij}, \quad v_{ij} \sim N(0, \lambda^c).$$

The results are shown in Table 4. In Heagerty and Zeger (2000), $|\hat{\beta}^c|$ in BN0 is twice that in BN, while our estimates (Lee and Nelder, 2001a) for the two models BN0 and BN are not very different.

5.1 A Marginalized Random-Effect Model

There may not exist a simple constraint on v_{ij} in BN that allows a logit model in the form $\mathrm{logit}(p_{ij}) = \beta_0^m + x_{ij}\beta^m$. For inferences about marginal probabilities, Heagerty and Zeger (2000) proposed to use the marginalized random-effect model (BM)

$$Y_{ij}|v_{ij} \sim B(p_{ij}^h), \quad \mathrm{logit}(p_{ij}^h) = \Delta + w_{ij},$$
$$\mathrm{logit}(p_{ij}) = \beta_0^m + x_{ij}\beta^m, \quad w_{ij} \sim N(0, \lambda_j^m),$$

where

$$p_{ij}^h = E(Y_{ij}|w_{ij}) = \frac{\exp(\Delta + w_{ij})}{1 + \exp(\Delta + w_{ij})}$$

and

$$p_{ij} = \frac{\exp(x_{ij}\beta^m)}{1 + \exp(x_{ij}\beta^m)} = \int \frac{\exp(\Delta + w_{ij})}{1 + \exp(\Delta + w_{ij})}\, dF(w_{ij}), \tag{8}$$

$F(w_{ij})$ being the cumulative distribution of w_{ij}. Given the values of $x_{ij}\beta^m$ and λ_j^m, the integral equation (8) can be solved numerically for Δ. In BM, the model assumption about the linear predictor of the ordinary BN model is split into two parts:

$$\mathrm{logit}(p_{ij}) = x_{ij}\beta^m \quad \text{and} \quad \mathrm{logit}(p_{ij}^h) = \Delta + w_{ij}. \tag{9}$$

However, the multivariate logit model

$$\mathrm{logit}(p_{ij}) = x_{ij}\beta^m$$

implicitly implies a strange random-effect model in which

$$p_{ij}^h = E(Y_{ij}|u_{ij}) = \frac{\exp(x_{ij}\beta^m)}{1 + \exp(x_{ij}\beta^m)} u_{ij} \quad \text{with } E(u_{ij}) = 1. \tag{10}$$

Appropriate distributional assumptions on u_{ij} lead to a multivariate model for the marginal parameters p_{ij}. However, we cannot find a distribution of u_{ij} which leads to Heagerty and Zeger's BM. This shows that there are two alternatives for generating multivariate logit models.

Regression coefficients from BN and BM are not directly comparable. Using the first equality in (8), we can directly predict the marginal probability from the BM. However, from BN models, we can also easily predict it by using the fact

$$p_{ij} = E(Y_{ij}) = \int \frac{\exp(\beta_0^c + x_{ij}\beta^c + v_{ij})}{1 + \exp(\beta_0^c + x_{ij}\beta^c + v_{ij})}\, dF(v_{ij}).$$

From Table 5 we see that there are only small differences in the predicted margins for the three models BN, BM and OBM. From $\hat{p}_{ij}$, we can compute marginal logistic regression estimates from BN, giving an

TABLE 5
Predicted margins from three models

OBM		BM		BN		BN0	
$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$
0.899	0.772	0.898	0.751	0.891	0.719	0.858	0.740

intercept $\hat{\beta}_0^m = \text{logit}(\hat{p}_{i1}) = 2.101$ and treatment effect $\hat{\beta}_1^m = \text{logit}(\hat{p}_{i2}) - \text{logit}(\hat{p}_{i1}) = -1.156$. Comparing these with those from BM in Table 4, we see that there is not much difference between the results from the so-called marginalized random-effect model BM and the ordinary BN. Furthermore, BN0 provides marginal estimators for the intercept $\hat{\beta}_0^m = 1.801$ and for the treatment effect $\hat{\beta}_1^m = \text{logit}(\hat{p}_{i2}) - \text{logit}(\hat{p}_{i1}) = -0.758$. A look at the difference in $\hat{\beta}_1^m$ between models with the common and separate heterogeneity suggests that the insensitivity of $\hat{\beta}_1^m$ (compared with $\hat{\beta}_1^c$) to dispersion misspecification may not hold, at least with this dataset. Furthermore, if we compare like with like, we see no dramatic differences in sensitivity.

In Poisson random-effect models the two seemingly different models, such as PN1 and PNM, are equivalent, but assume different constraints on the random effects. However, in binomial random-effect models, BN and BM are different and we may have to make a choice. If we compare like with like, these two models may not be very different. With BM inferences about conditional probability are difficult. By contrast, the integration required to obtain a marginal probability at the prediction stage from BN is not difficult; see also Goldstein and Rasbash (1996). Further studies, in particular about parametrizations related to the model prediction stage, would be interesting.

ACKNOWLEDGMENTS

The authors thank Sir David Cox, Norman Breslow, William Browne, Martin Crowder, Harvey Goldstein, Patrick Heagerty, Jerry Lawless, Jim Lindsey, Nick Longford, Yudi Pawitan, Stephen Senn, an anonymous editor and the referees for their helpful comments. This work was supported by Korean Research Foundation Grant KRF-2001-015-DP0069.

REFERENCES

CROWDER, M. J. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82** 407–410.

CROWDER, M. J. and HAND, D. J. (1990). *Analysis of Repeated Measures.* Chapman and Hall, London.

DIGGLE, P. J., LIANG, K. Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

GALBRAITH, R. F. and LASLETT, G. M. (1993). Statistical models for mixed fission track ages. *Nuclear Tracks and Radiation Measurements* **21** 459–470.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. Arnold, London.

GOLDSTEIN, H. and RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **159** 505–513.

HA, I. D., LEE, Y. and SONG, J. K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika* **88** 233–243.

HEAGERTY, P. J. and ZEGER, S. L. (2000). Marginalized multilevel models and likelihood inference (with discussion). *Statist. Sci.* **15** 1–26.

KENT, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69** 19–27.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc*. **73** 805–811.

LEE, Y. (2002). Robust variance estimators for fixed-effect estimates with hierarchical-likelihood. *Statist. Comput.* **12** 201–207.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B* **58** 619–678.

LEE, Y. and NELDER, J. A. (1999). Robustness of the quasilikelihood estimator. *Canad. J. Statist.* **27** 321–327.

LEE, Y. and NELDER, J. A. (2000). Two ways of modelling overdispersion in non-normal data. *Appl. Statist.* **49** 591–598.

LEE, Y. and NELDER, J. A. (2001a). Hierarchical generalised linear models: A synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika* **88** 987–1006.

LEE, Y. and NELDER, J. A. (2001b). Modelling and analysing correlated non-normal data. *Statist. Model.* **1** 3–16.

LEE, Y. and NELDER, J. A. (2002). Analysis of the ulcer data using hierarchical generalized linear models. *Statistics in Medicine* **21** 191–202.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

LINDSEY, J. K. and LAMBERT, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* **17** 447–469.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

NELDER, J. A. (1994). The statistics of linear models: Back to basics. *Statist. Comput.* **4** 221–234.

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statist. Sci.* **6** 15–51.

SEELY, J. and HOGG, R. V. (1982). Symmetrically distributed and unbiased estimators in linear models. *Comm. Statist. A—Theory Methods* **11** 721–729.

SIMPSON, E. H. (1952). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.
WEDDERBURN, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61** 439–447.
WEIL, C. S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology* **8** 177–182.
ZEGER, S. L. and LIANG, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.
ZEGER, S. L., LIANG, K. Y. and ALBERT, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44** 1049–1060.

Comment

Stephen Senn

I welcome this paper by Youngjo Lee and John Nelder (L&N), which considerably advances our understanding of random-effect modeling. There is much that the authors state with which I am in complete agreement, in particular, their demonstration that the claim is false that marginal models must be used where inferences about populations are desired. As John Nelder pointed out many years ago in a paper with Peter Lane, prediction is a different purpose to estimation and you do not have to *estimate* directly analogous quantities to those which you wish to *predict* (Lane and Nelder, 1982). Estimates form the building blocks of predictions and often much work is required before you can construct the latter from the former. In any case, in the field in which I work, that of clinical trials, it is a pernicious but all too widespread delusion that the patients are a representative sample of those for whom the treatments might be indicated: In reality we have precise control of the allocation algorithm, but very little of the presenting process (Senn, 2000a). In fact, the results for the patients in a clinical trial *as a population* are frequently not of interest (Lindsey and Lambert, 1998). What one is trying to establish is the causal effects of treatments on individuals: After all, if the treatment cannot affect individuals, it has no effect on populations, and it is individuals we treat. Establishing such effects can, of course, yield predictions for populations. In short, I have a great deal of sympathy with the authors' creed, which can be summed up succinctly using the last sentence in their abstract, "We regard the conditional model as fundamental, from which marginal predictions can be made." If, in the rest of this note, I raise a few quibbles with this point of view, this is simply for the sake of discussion; basically, I think that their thesis is correct.

However, before proceeding to see if anything can be said in favor of marginal models, I wish to amplify some points of agreement. For example, in my opinion, and as already stated, estimation and prediction are not the same except by accident. It is misleading that a standard statistical paradigm, to which textbooks often return, is that of estimating a population mean using simple random sampling. For this purpose, the parameter estimate of the simple model is, indeed, the same as the prediction. However, as soon as we turn to more complex sampling schemes, this is not so. Stratified random sampling, for example, yields estimates of stratum means from which the population mean can be predicted using the sampling fractions *if one wishes*, but there is no *immediate* connection between any of the parameters estimated and the target quantity. There is also a very common confusion between samples and experiments: The latter carry with them no necessary implication about any population quantity whatsoever. Consider, for example, random-effect meta-analysis. The usual random-effect estimate down-weights the influence of larger trials compared to fixed-effect analysis in producing an "overall mean effect." However, this overall mean effect is not, although it is a point regularly overlooked, a prediction in itself of anything useful whatsoever. For example, if one believed that larger trials were more likely to have recruited "typical" patients and that a prediction for typical patients was needed, one might wish to give larger trials more weight after all (Senn, 2000b).

These distinctions sometimes come to a head when causal analyses are carried out on sample surveys. For example, one might have carried out a stratified survey on dietary habits with very different sampling fractions per stratum. To the extent that one wishes to establish the dietary habits of the population, these

Stephen Senn is Professor, Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland (e-mail: stephen@stats.gla.ac.uk).

fractions are relevant to inferences, but if a separate object of the study is the effect of diet on health, then for this purpose the study is quasiexperimental and the fractions are not relevant for causal inferences. It is perhaps interesting to note that the statistical package SUDAAN is particularly strong with regard to the former purpose and that great play is made of its ability in estimating effects for clustered data to implement the GEE approaches criticized by L&N.

A related issue concerns certain types of measurement. We are now also seeing a plethora of recommendations regarding, for example, sigma-divided measures of treatment effects. For instance, instead of measuring the difference between two treatments on some natural scale, such as liters of forced expiratory volume in one second in asthma or millimeters of mercury diastolic blood pressure in hypertension, we do so in terms of the degree of overlap between the two treatment groups (Rom and Hwang, 1996; Stine and Heyse, 2001). Such measures, depending as they do on the variability observed in the sample, which may be quite different from that in some target population, are almost uninterpretable (Senn, 1997) and have no application at an individual level. Again, a reification of the population is involved.

Thus, I agree wholeheartedly with the authors' claim that marginal predictions do not require marginal estimation and indeed that such predictions will be better served (usually) by conditional models. Once the parameters of these models have been estimated, they may then, together with further quantities that may be necessary, yield marginal predictions for a given population of interest.

What causes me to hesitate in signing up 100% to the thesis of L&N is the thought that because our knowledge is limited, we may be forced from time to time to be marginal. There has been considerable work over the last two decades on the effect of missing covariates or model-misspecification in nonlinear models. For example, Gail, Wieand and Piantadosi (1984) examined exactly what type of model randomization would ensure asymptotically unbiased estimates despite a missed prognostic covariate. They showed that whereas many such models existed, for "certain important nonlinear regression models," biased estimates were produced. (See in particular their extremely useful Table I, Gail, Wieand and Piantadosi, 1984, page 437.) In labeling the estimates as biased, they implicitly assumed that the model conditioning on the covariate was the correct one. Ford, Norrie and Ahmadi (1995) in the specific context of proportional hazards models, took the more cautious view that different quantities were being measured. Similar points have been made by others (Beach and Meier, 1989; Robinson and Jewell, 1991). It seems plausible that those cases where missing covariates lead to biased estimates are the cases for which, in a random-effect framework, conditional and marginal models appear to yield different results. (This is probably also relevant to the discussion in L&N toward the end of Section 3 regarding choice of parametrization.) The connection between the effect of stratification and fitting covariates on parameter estimation and random effects modeling must be particularly close when one considers that, from one point of view, a fixed-effect model is simply a more complex version of the random-effect model, involving separate distributions for each stratum or covariate rather than regarding them as realizations of some parent distribution. Furthermore, since one can always imagine covariates that could have been measured but were not, there is always a possible model to which the model actually used is marginal.

Suppose we are faced with performing a meta-analysis of a mixture of parallel and crossover trials for the sort of models for which the marginal and conditional model parameters represent different quantities. How should we proceed? One argument is that since we can fit a marginal model in both cases, this is the way that we should analyze individual trials. Presumably the L&N philosophy would be that we fit conditional models where we can do so and then combine them at the level of marginal predictions.

I am puzzled, however, by L&N's opening example. Consider a concrete instance. When modeling an AB/BA crossover trial using random effects and a Normal model [fixed patient effects are also common (Senn, 2002)], the two usual approaches are the conditional approach using between- and within-subject errors corresponding to L&N model (1) and the marginal approach using a block diagonal variance–covariance matrix. However, I do not see what is wrong with the latter approach. Indeed, it seems slightly more general than the former, allowing as it does for negative correlations. Such negative correlations are implausible, but are occasionally encountered, as in Grizzle's (1965) famous paper (where, however, the analysis was of differences from baseline, which therefore has the patient effect eliminated). In fact, I would say that the two models are equivalent, apart from the implied constraint $\rho \geq 0$ in the former, which constraint might or might not be appropriate, depending on the degree of

prior knowledge, but which in any case could be incorporated in estimation for the latter. I also believe that the reference to Simpson's paradox here is a red herring. This *pace* L&N has no essential connection to treatment–subject interaction and no such interaction is present in the example of Lindsey and Lambert (1998; at least on the probability scale) nor do they claim that there is. Indeed, the pure paradox exhibits itself most forcibly when there is no interaction at all and the treatment effect is constant in every stratum, but different in terms of marginal contrasts. For excellent discussions of confounding and Simpson's paradox, see Pearl's (2000) important book, Garrett (2003) or Greenland, Robins and Pearl (1999).

Another issue is that of judgement in modeling. All statisticians choose to be marginal at some level or another in their choice of measure on some occasion or another. Consider a simple AB/BA crossover trial in bioequivalence. The object is to estimate relative bioavailability and show that this is close to unity. Conventionally this is done by comparing the area under the concentration time curve (AUC) using a log transformation. It is sometimes mistakenly claimed that, due to properties of the log-Normal, this implies something about comparison of population medians rather than means. In fact, if there are no missing observations, efficient analysis may be reduced to an analysis of the log ratios for individual subjects. Essentially one conditions on the subject effects and it make no difference whether the subject effects themselves are Normal or not. The subjects in the trial could have been recruited in equal numbers from sumo wrestlers and jockeys, yielding a curious bivariate bimodal distribution of the original measurements, but leaving the distribution of log ratios unaffected. The inference is about the causal effect of the treatment and not about any populations as such, and these population parameters are of no interest whatsoever apart from the relative bioavailability itself, each subject providing a means of estimating this. This is consonant with the philosophy of L&N.

However, an analysis of AUCs from such a trial will not permit identification of patient by formulation interactive effects: For that a crossover trial with three or more periods would be needed (Hauck, Hyslop, Chen, Patnaik and Williams, 2000; Senn, 2002). However, since blood samples will be taken at frequent intervals, by using a sufficiently parsimonious model for within-subject errors together with a suitable pharmacokinetic model, individual subject effects could be estimated using the individual concentrations at each time point. If the purpose of the exercise were to estimate pharmacokinetic parameters such as clearance, volume of distribution and so forth, this might, indeed, be a good thing to do. There is a long and impressive tradition of this approach, stretching back at least to the work of Sheiner and colleagues in the early 1970s (Sheiner, Rosenberg and Melmon, 1972). However, in a bioequivalence trial, this is not the object of the exercise, but simply to prove that the two concentration-time profiles are equivalent. In my view, there really would be little point to doing anything other than comparing AUC's (Senn, 2001). Many similar examples can be found where the statistician makes a judgement as to how far it is worth going into the business of analyzing lower levels of the data. A famous example that involves such a choice of level, although this has not always been appreciated, is the epilepsy data of Thall and Vail (1990), which also was used by Diggle, Liang and Zeger (1994). The data comprise an 8-week trial divided, in my opinion to no useful purpose, into four 2-week periods: why not eight 1-week periods or 56 1-day periods or, more naturally, one 8-week period? Probably the most logical approach would be to look at it in terms of repeated interseizure intervals in terms of some complex survival analysis model. In short, sometimes enthusiasm for driving the model downward is taken too far. This was certainly Yates' (1982) view of approaches to analyzing repeated measures designs as if they were split-plot experiments, which they are not; this habit is widespread in the psychometric literature.

The point I am making here is that all statisticians decide on a certain degree of aggregation. Therefore, if the drive for conditional models is interpreted as implying that no level of aggregation is ever acceptable, in my opinion this goes too far. However, it seems to me that the general message of L&N is sound and I offer my summary of some of the lessons of this paper as follows.

1. The desire to issue marginal predictions is not in itself a reason for not using conditional models.
2. Inferences will usually be superior if conditional models are employed.
3. We should be careful when comparing parameter estimates from different models; it may require much thought to compare like with like.
4. In particular, certain explicit or implicit side conditions on models may have important consequences.

Comment

Naisyin Wang

I would like to begin by thanking Professor Lee and Professor Nelder for an interesting and well written article. The comparison between conditional and marginal modeling of longitudinal data has generated great interest in recent years. As pointed out by Lee and Nelder, certain controversies were simply raised due to the failure to "compare like with like." Even though this issue may be well known among statisticians whose research interests include longitudinal data analysis, the clearly presented examples in this article certainly help to make the issue transparent to general users of these methods. The current article discusses another interesting and related question: the choice between conditional and marginal modeling. The authors tend to focus on the notion that conditional modeling is absolutely superior to the marginal modeling. Beside the scenarios that the scientific aims dictate the choice of modeling approach, I feel there are various issues that are worthwhile to consider and have not been discussed fully in the current article.

A POPULATION-AVERAGING INTERPRETATION FOR MARGINAL MODELING

As the authors have noted, when the subjects in the study can be regarded as random samples from a population, a marginal mean and a population mean are often taken as the same thing. It is known that a weighted marginal mean with the weights inversely proportional to the sampling probabilities of the samples still possesses the same interpretation. Precisely, when a marginal estimation equation is solved using data from a study, its solution is consistent with the parameter that solves the corresponding population estimating equation. This can be established using general M-estimator theory under mild conditions, provided that the subjects under study are properly sampled from the target population. That is, the estimator based on marginal modeling is meaningful in the population-averaging sense. The issue that needs to be carefully addressed now is whether this population parameter can be used to answer the scientific question of interest. Obviously, the construction of the marginal estimating equations should have this goal as the top priority.

The authors argue that this population-averaging interpretation often does not hold for marginal modeling because subjects in longitudinal studies may not represent the population of interest (e.g., they are volunteers). They further state that, nevertheless, even under this situation, the estimates obtained from the conditional modeling are still meaningful. I am puzzled by this argument. Consider the situation that all subjects under study are random samples from one subpopulation and the parameter of interest has values that differ between this subpopulation and the rest of the target population. In this situation, regardless of which modeling approach is taken, the conclusion can only be made for the subpopulation where the samples are taken and not for the entire target population. One might argue that the assumption that the parameter of interest is the same across all subjects in the population can be used to rule out this problem. It is worth noting that such an assumption is uncheckable because no subjects are taken from the complement of the subpopulation, and thus no observations from there can be used for assumption checking.

STABILITY AND ROBUSTNESS OF MARGINAL ESTIMATION

For any assumed structure, there is always a possibility that the structure is misspecified. When the model is misspecified, a method is preferred if either it is less susceptible to the potential biases or it is easy to diagnose such a misspecification. Marginal modeling is strong in both aspects. One known advantage of using a marginal approach is that it is less susceptible than the conditional approach to biases induced by the misspecification of random-effect models. Such biases were discussed in detail by Neuhaus, Hauck and Kalbfleisch (1992) and Heagerty and Kurland (2001). Since the main concentration of the marginal approach is to properly model the first moment marginally, there are many traditional graphical diagnostic tools that are applicable here. Furthermore, because the marginal approach fits directly under the simplest M-estimation framework in which many robustness procedures are available, it is easy to adopt existing robustness methods (e.g., outliers down-weighting) into the marginal

Naisyin Wang is Professor, Department of Statistics and Faculty of Toxicology, Texas A&M University, College Station, Texas 77843-3143, USA.

approach. On this front, more research is warranted for conditional methods under mixed effect models.

Finally, I wish to point out that both approaches have their strengths and applications. I emphasized certain advantages of the marginal approach above simply because the authors have provided strong supportive arguments for conditional methods. Among them, the conditional prediction (e.g., providing a prediction confidence interval for a potential outcome given a subject's risk factors) is one aspect that is unique to conditional modeling and has broad aspects of application. With more new developments on topics such as outlier detection or model checking, both approaches could be more accessible and useful to practitioners.

ACKNOWLEDGMENT

This work was supported by the National Cancer Institute Grant CA74552 and the Texas Advanced Research Program.

Comment

Jiming Jiang

Professor Lee and Professor Nelder have presented us with a well written article regarding a controversy about the use of conditional and marginal models, particularly in the analysis of longitudinal data. In this field, there have been two main approaches: one using the GEE method based on marginal models and the other using methods of mixed model analysis based on conditional models. My comments focus on three important issues: the use of conditional and marginal models, differences between linear and generalized linear mixed models, and consistency.

1. CONDITIONAL OR MARGINAL?

The choice between the two models—conditional or marginal—should depend on the kind of inference needed in practice. Note that there are some similarities between analysis of longitudinal data and small-area estimation (SAE; e.g., Datta and Lahiri, 2000; for a review on SAE, see Ghosh and Rao, 1994), so I use an example from SAE for illustration. Consider the model

$$Y_{ij} = x_{ij}'\beta + u_i + e_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n_i,$$

where Y_{ij} corresponds to the j response from the ith small area, x_{ij} is a vector of known covariates, β is a vector of unknown regression coefficients, u_i is a small-area-specific random effect and e_{ij} is an error. This model is often called a *nested error regression* model (e.g., Ghosh and Rao, 1994) and it is a conditional model. A conditional model is needed here, because in SAE the small-area means are often of main interest, which for the ith small area is associated with the random effect u_i. More specifically, the prediction of a mixed effect of the form $x'\beta + u_i$ is of primary interest, where x is a known vector. No marginal model can provide inference about such a quantity, because the random effects do not appear in a marginal model.

On the other hand, in analysis of longitudinal data the main interest is often about the (fixed) regression coefficients (e.g., Diggle, Liang and Zeger, 1994). In such cases, I do not see why it is always necessary to assume a conditional model, because, as Lee and Nelder mentioned, the more specific the assumptions one makes, the more likely it is that some of the assumptions will fail in practice. A conditional model assumes the appearance of random effects in a specific manner (e.g., linear or generalized linear). Such assumptions are delicate, and one certainly risks the possibility of failure. By the way, unlike standard regression diagnostics, methods of diagnosing mixed effect models are not fully developed (e.g., Jiang, 2001), especially in the generalized linear case. Furthermore, the random effects themselves are not of interest in this case, unlike in SAE. Of course, this is not to say that conditional models should never be used unless the random effects are of direct interest, but the idea is certainly questionable that the conditional model is fundamental and, therefore, should be preferred over the marginal model.

Jiming Jiang is Professor, Department of Statistics, University of California, Davis, California 95616, USA.

2. LINEAR OR GENERALIZED LINEAR?

Lee and Nelder used an example of a linear mixed model to show that two conditional models can lead to the same marginal model. What they did not say is that, for the most part, this is true only in the linear case. For example, consider the following mixed logistic models, which are, in a way, similar to models C1 and C2 considered by Lee and Nelder,

(D11) $\text{logit}\{p(Y_{ijk}=1|v_i)\}=\beta_0+\beta_j+v_i$

and

(D12) $\text{logit}\{p(Y_{ijk}=1|v_i,v_{ij})\}=\beta_0+\beta_j+v_i+v_{ij}$,

where the βs are fixed effects (there should be a constraint $\sum_j \beta_j = 0$ to ensure identifiability) and the v's are random effects which have mean zero. In this case, the marginal model is defined in terms of the (unconditional) probability $p(Y_{ijk}=1)$. Under D1, we have

(1) $$p(Y_{ijk}=1)=E\{h(\beta_0+\beta_j+\xi)\},$$

where $h(x)=e^x/(1+e^x)$ and ξ has the same distribution as v_i, while under D2,

(2) $$p(Y_{ijk}=1)=E\{h(\beta_0+\beta_j+\eta)\},$$

where η has the same distribution as v_i+v_{ij}. If v_i and v_{ij} are independent and both normally distributed, models (1) and (2) are considered to be the same; otherwise, the two marginal models may be different. Note that models C1 and C2 considered by Lee and Nelder have the same marginal model as long as the random effects have mean zero (regardless of normality). To see an example in which the marginal models are different even under normality, consider

(D3) $\text{logit}\{p(Y_{ijk}=1|v_i,s_i)\}=\beta_0+\beta_j+v_i+s_i\beta_j$,

where s_i is another random effect with mean zero. Note that here the model has a *random slope* s_i in addition to the *random intercept* v_i. A linear analogue of C3, $Y_{ijk}=\beta_0+\beta_j+v_i+s_i\beta_j+e_{ijk}$ (e_{ijk} is the same as in Lee and Nelder's paper), results in the same marginal model as C1 and C2. However, under D3 we have

(3) $$p(Y_{ijk}=1)=E\{h(\beta_0+\beta_j+\zeta)\},$$

where ζ has the same distribution as $v_i+s_i\beta_j$. For example, under normality and independence of v and s, ξ has distribution $N(0,\sigma_v^2)$, while ζ has distribution $N(0,\sigma_v^2+\sigma_s^2\beta_j^2)$, so the marginal models (1) and (3) are different unless the β_j's are all zero (note the sum constraint about the β_j's), which is a meaningless case.

When a conditional model is assumed, I am in favor of using robust methods of inference, an approach that Lee and Nelder also seem to support (see Section 4 of their paper). However, they dislike the use of marginal models without a likelihood basis. Can a likelihood-based method also be robust? There are some well-known examples in linear (mixed) models. For example, the weighted least squares (WLS) estimator, which is the maximum likelihood estimator under the assumption $Y\sim N(X\beta, W^{-1})$, where W is the weight matrix (e.g., Diggle, Liang and Zeger, 1994, page 58), is known to be consistent under failure of normality and misspecification of the covariance matrix; the restricted maximum likelihood (REML) estimator, which is derived under normality, is known to be consistent, even if normality fails (Richardson and Welsh, 1994; Jiang 1996, 1997). Note that WLS and REML work under both conditional and marginal models. However, similar properties do not seem to be shared by existing methods, likelihood-based or not, in generalized linear mixed models (GLMM). For example, the likelihood equation under GLMM and normality of the random effects will be biased if normality fails. This suggests that, unlike GEE, likelihood-based inference in GLMM is more sensitive to failure of distributional assumptions, such as normality, than in linear mixed models.

3. CONSISTENT OR OTHERWISE?

Since I mentioned consistency, I had better continue, because there is some serious issue of consistency regarding methods based on Laplace approximation (to integrals). Such a method was used by Lee and Nelder to make inference about the hierarchical generalized linear model; they claim that the method "works well." The Laplace-approximation-based method is known to work well when the variances of the random effects are close to zero, and the second-order Laplace approximation is more accurate than the first-order one (e.g., Lin and Breslow, 1996). However, these estimators are not consistent unless the number of observations that correspond to each random effect goes to infinity (e.g., Lee and Nelder, 1996; Jiang 1999). While the latter assumption may be realistic in some cases (of longitudinal data), it is certainly not in SAE, where the sample size for each small area is usually small.

Rejoinder

Youngjo Lee and John A. Nelder

We thank the discussants for their comments. The title of our paper was carefully chosen and was meant to imply that we were making the case for conditional models, rather than trying to survey all possible approaches to this kind of modeling. We are glad to have Senn's support for our distinction between estimation and prediction, the case for which Lane and Nelder made more than 20 years ago, but which seems still to be little appreciated. Senn's arguments deserve to be widely studied.

Senn's crossover example raises the interesting point about what should be done with negative correlations (within litters, say). Formally these can be expressed by negative variance components, and in some software this is the route taken. However if, as in our HGLM software, the variance components are modelled on the log scale (which has advantages), the correlations must be left as correlations and estimated accordingly. We accept his point about the irrelevance of Simpson's paradox, having learnt much from the recent discussion about it on the Genstat bulletin board.

Senn's general point about decisions on levels of aggregation is well taken. An analysis with no aggregation may be practically impossible because it would involve impossibly large datasets. What we suspect the statistician is doing, or should be doing, is aggregating over lower-level classifications which he or she believes have little effect on the inferences made.

RESPONSE TO WANG

When the subjects in a study can be regarded as random samples from a population, a marginal model is useful for estimation of population parameters. For such cases, however, there will almost always be a conditional model, leading to that marginal model, to allow inferences about population parameters, unless the marginal model does not have a likelihood or probabilistic base. Thus, inferences for the population can be made from the conditional model as well. When the subjects in the study cannot be regarded as random samples from a population, the marginal means from the conditional model can still be useful for inference about causal effects of treatments and so forth on individuals' margins. As Senn noted in his discussion, it also indicates the change of population, because the treatments affect individuals that compose the population, which would be informative on the population change.

RESPONSE TO JIANG

All three models, D1, D2 and D3 of Jiang, lead to the common marginal model

$$E(\text{logit}\{P(Y_{ijk}=1|v_i)\})=\beta_0+\beta_1,$$

with arbitrary covariance structure for $\text{logit}(P(Y_{ijk}=1|v_i))$. So each leads to a marginal model on a particular scale. As we said, these three models are qualitatively very different, and ignoring differences between them could lead to wrong conclusions.

Jiang raised doubts about the performance of the h-likelihood method, and both Wang and Jiang raised the robustness issue. The rest of our discussion is devoted to these points.

PERFORMANCE OF THE *h*-LIKELIHOOD METHOD

There have been various criticisms of the h-likelihood method, deriving from a belief that h-likelihood provides qualitatively different (i.e., noninvariant) inferences for trivial reexpressions of the underlying model and that the h-likelihood estimator does not work well, especially for binary data; see Jiang's last comment. This we believe is mainly due to a misunderstanding of the h-likelihood procedure. Lee and Nelder (2003a) discussed invariance associated with h-likelihood, and Noh and Lee (2003) gave numerical evidence to show that it outperforms all the other Laplace-approximation-based alternatives in estimation of GLMMs for binary data. We developed the h-likelihood method and read a paper to the Royal Statistical Society about it in 1996. The discussion was a disaster because everybody took the worst possible case of binary data and described difficulties with it. Nobody said it worked in other cases. We are glad to have an opportunity to clarify these matters. There have been many criticisms about the performance of the h-likelihood estimator, but surprisingly we never see any actual numerical studies to verify such criticism. We have not seen any method which outperforms the h-likelihood estimator. We do not say that the current h-likelihood method will always perform the best, but we believe that it can always be modified to give

an improvement, as has been done with Fisher's likelihood method.

We explained why Lee and Nelder's proposal generally "works well," while the other Laplace-approximation-based methods, such as those of Schall (1991), Breslow and Clayton (1993), Breslow and Lin (1995), Shun and McCullagh (1995), Lin and Breslow (1996) and Shun (1997), do not. All of them, except for the h-likelihood method, are limited (i.e., restricted to GLMMs and/or to some particular design structures) and miss some terms (Noh and Lee, 2003).

h-LIKELIHOODS

Ever since Fisher (1921) introduced the concept of likelihood, the likelihood function has played an important part in the development of both the theory and the practice of statistics. There have been several attempts to extend likelihood beyond its use in parametric inference to inference from models of a more general nature that may include fixed parameters, random parameters and unobserved variables. Special cases are subject-specific inference, prediction of unobserved future observations and missing data problems. For the use of h-likelihood as a predictive likelihood, see Pawitan (2001); for missing data, see Lee, Noh and Ryu (2003).

Consider HGLMs

$$\mu = E(y|u) \quad \text{and} \quad \text{var}(y|u) = \phi V(\mu)$$

with a linear predictor

$$\eta = g(\mu) = X\beta + Zv,$$

where $g(\cdot)$ is a generalized linear model (GLM) link function, X and Z are model matrices for fixed and random parameters (effects), respectively, and $v_i = v(u_i)$ are random effects after some transformation $v(\cdot)$.

The joint density of the responses y and the random effects v can be written

$$L(v(u), y|\beta, \phi, \lambda) = f_{\beta,\phi}(y|v(u))\, f_\lambda(v(u)),$$

where $f_{\beta,\phi}(y|v(u))$ is a density with a distribution from a one-parameter exponential family for GLMs and the second term $f_\lambda(v)$ is the density function of the random effects v with parameter λ. Note that the function $v(u)$ defines the scale on which the random effects are assumed to combine additively with the fixed effects β in the linear predictor. Lee and Nelder (1996) defined the h (log-)likelihood as

$$h = \log\{L(v(u), y|\beta, \phi, \lambda)\}.$$

In this definition we use a particular scale $v = v(u)$ for the h-likelihood, which gets rid of alleged counterexamples related to noninvariance (Lee and Nelder, 2003a).

USE OF *h*-LIKELIHOOD

The h-likelihood is not a likelihood in the Fisherian sense because of the presence of unobservables, namely random effects. Lee and Nelder (1996) claimed that a systematic likelihood inference is possible for HGLMs by using the h-likelihood. In this discussion, we concentrated on estimation of the fixed parameters (β, ϕ, λ). From the h-likelihood we have the following two profile likelihoods: (1) The marginal log-likelihood m can be obtained from the h-likelihood by integrating out the random parameters, that is,

$$m = l(y|\beta, \phi, \lambda) = \log \int \exp(h)\, dv.$$

(2) In mixed linear models the restricted (or residual) likelihood r of Patterson and Thompson (1971), that is,

$$r = l(y|\tilde{\beta}, \phi, \lambda) = \log f_{\phi,\lambda}(y|\tilde{\beta})$$

$$\text{where } \tilde{\beta} \text{ are ML estimators given } (\phi, \lambda),$$

has been proposed for inference about the dispersion parameters (ϕ, λ) to reduce bias, especially in finite samples. Under the h-likelihood framework, the marginal likelihood is a profile likelihood for the fixed parameters (β, ϕ, λ), after eliminating random parameters v by integration from the h-likelihood; the restricted likelihood is that for the dispersion parameters (ϕ, λ), after eliminating fixed effects β by conditioning on the marginal likelihood.

Lee and Nelder (2001a) considered a function defined as

$$p_\alpha(l) = \left[l - \tfrac{1}{2}\log\det\{D(l, \alpha)/(2\pi)\}\right]\Big|_{\alpha=\tilde{\alpha}},$$

where $D(l, \alpha) = -\partial^2 l/\partial\alpha^2$ and $\tilde{\alpha}$ solves $\partial l/\partial\alpha = 0$. For fixed effects β, the use of $p_\beta(m)$ is equivalent to conditioning on $\tilde{\beta}$ [i.e., $p_\beta(m) \simeq r = l(y|\tilde{\beta}, \phi, \lambda)$ to first order (Cox and Reid, 1987)], while for random effects v, the use of $p_v(h)$ is equivalent to integrating them out using the first-order Laplace approximation, [i.e., $p_v(h) \simeq m = \log \int L(y, v|\beta, \phi, \lambda)\, dv$ (Lee and Nelder, 2001a)]. In mixed linear models, Lee and Nelder (2001a) noted that

$$m \equiv p_v(h) \quad \text{and} \quad p_\beta(m) \equiv p_{\beta,v}(h).$$

The use of $p_{\beta,v}(h)$ for estimating the dispersion parameters (ϕ, λ) means that we eliminate both random and fixed effects simultaneously from the h-likelihood. Lee and Nelder (2001a) showed that, in general, $p_{\beta,v}(h)$ is approximately $p_\beta(p_v(h))$ and that numerically $p_{\beta,v}(h)$ provides good dispersion estimators for HGLMs. In principle we should use the h-likelihood h

for inferences about v, the marginal likelihood m for β and the restricted likelihood $p_\beta(m)$ for the dispersion parameters. When m is numerically hard to obtain, we can use $p_v(h)$ and $p_{\beta,v}(h)$ as approximations to m and $p_\beta(m)$; $p_{\beta,v}(h)$ gives approximate restricted MLEs and $p_v(h)$ gives approximate MLEs.

h-LIKELIHOOD VERSUS PENALIZED-QUASILIKELIHOOD ESTIMATION

Lee and Nelder (1996) observed that although, in general, a joint maximization of h-likelihood does not provide marginal MLEs for β, the deviance differences constructed from h and $p_v(h)$ are often very similar, so they proposed to use h for estimating β. In these models joint optimization of the h-likelihood offers a numerically and statistically efficient fitting algorithm (Lee and Nelder, 2001a). This algorithm can be expressed as the fitting of a set of interlinked GLMs; it requires neither prior distributions of parameters nor multidimensional quadratures. Except for binary data, the resulting estimators generally work well; see the simulation studies of Poisson and binomial models (Lee and Nelder, 2001a), of frailty models (Ha, Lee and Song, 2001) and of mixed linear models with censoring (Ha, Lee and Song, 2002). Schall's (1991) method is the same as Breslow and Clayton's (1993) penalized-quasilikelihood (PQL) method for GLMMs. They are the same as the h-likelihood method, but ignore $\partial\hat{v}/\partial\phi$ and $\partial\hat{v}/\partial\lambda$ in the dispersion estimation (Lee and Nelder, 2001a), which results in severe bias, especially in binary data (Noh and Lee, 2003). So Breslow and Lin (1995) and Lin and Breslow (1996) proposed a bias correction for the PQL estimator; however, this cannot overcome the difficulty caused by ignoring important terms (Noh and Lee, 2003). Bellamy et al. (2000) and Ten Have and Localio (1999) observed empirical results through simulation studies that the PQL estimator performs well in situations involving small numbers of large clusters. For good performance, the h-likelihood estimator does not require large clusters, but works well with a small number of clusters (Yun and Lee, 2004; Kang, Lee and Lee, 2003). Shun and McCullagh (1995) and Shun (1997) omitted terms related to profiling of β in $p_\beta(m)$, and this results in a nonignorable bias in finite samples such as the salamander data (Noh and Lee, 2003).

USE OF PROFILE LIKELIHOOD FOR BIAS REDUCTION

In binary data the h-likelihood method can have nonignorable bias. However, this reflects a general difficulty with likelihood inference, namely how to deal with a subset of parameters in the presence of many nuisance parameters. It should be noted that MLEs for the dispersion parameters also have severe biases when the number of fixed effects increases with the sample size. Such biases can be avoided by introducing restricted likelihood (i.e., profile likelihood), which, however, results in larger variance. With binary data such biases of the h-likelihood estimator β can also be avoided by introducing the profile h-likelihood, either m or $p_v(h)$, when m is hard to obtain. The bias of the h-likelihood estimator in binary data should be treated as that for the MLE for dispersion parameters; use of a profile h-likelihood $p_v(h)$, based on the Laplace approximation, eliminates such undesirable bias (Yun and Lee, 2004) in the same way as the restricted likelihood does for dispersion parameters. Noh and Lee (2003) found that when the cluster size exceeds 3 there is no recognizable asymptotic bias at all. In their study for binary data, the computed asymptotic bias is within the third decimal point in dispersion parameter estimation and within the fourth decimal point in fixed-effects estimation. Furthermore, higher-order approximation such as the second order is useful for improved approximation. So for the analysis of small-area estimation the h-likelihood procedure will work very well if the proper procedure is used. In summary, we believe that the complaints about the h-likelihood method are caused by a confusion with other methods which miss some important terms.

ROBUSTNESS AND HGLMs

Diagnostic tools for HGLMs are well developed (Lee and Nelder, 2001a, b, Lee, Yun and Lee, 2003), while those for marginal models are not, due to lack of a probabilistic basis. It has been claimed that marginal models provide robust estimators. However, we have demonstrated that robustness is not a matter of models, but rather of parameterizations, that is, a margin on one particular scale may be less sensitive than one on another scale. Without a probabilistic base (and therefore without the possibility of proper model checking), there may be no way to guarantee robustness. Recently, Lee and Nelder (2003b) introduced double HGLMs in which random effects are allowed not only in the means, but also in the dispersions. Heavy tailed distributions are available for $y|v$ and v by introducing random effects in ϕ and λ, respectively. This is very promising, because it will allow robust estimation for random-effect models. This class will, among other things, also enable models of types widely used in the

analysis of financial data to be explored, and should give rise to new extended classes of models within that framework.

h-LIKELIHOOD: LIKELIHOOD FOR UNOBSERVABLES

Under the *h*-likelihood framework, the marginal likelihood appears as a profile likelihood similar to the restricted likelihood, which has been recommended to reduce the bias. With the use of the *h*-likelihood, inference for random or combined fixed and random parameters is possible. It is perhaps unfortunate that Bayesians, from Lindley and Smith (1972) onward, seem to have made a takeover bid for all hierarchical models, implying that one has to be Bayesian to deal with them. The availability of Markov chain Monte Carlo methods, which make all problems seem more easily solvable via Bayesian computations, has appeared to justify this. However, by using *h*-likelihood, we can deal with such models directly in a likelihood framework because there is an explicit analytic form for that kind of likelihood. Furthermore, inferences for unobservables are possible without resorting to an empirical Bayesian framework. *h*-likelihood gives a powerful and practical tool for statistical inference; being a natural extension of Fisher likelihood to models with unobservables, it will become, we believe, widely used for inference from hierarchical models.

ADDITIONAL REFERENCES

BEACH, M. L. and MEIER, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials* **10** 161S–175S.

BELLAMY, S. L., GIBBERD, R., HANCOCK, L., HOWLEY, P., KENNEDY, B., KLAR, N., LIPSITZ, S. and RYAN, L. (2000). Analysis of dichotomous outcome data for community intervention studies. *Statistical Methods in Medical Research* **9** 135–159.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

BRESLOW, N. E. and LIN, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82** 81–91.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* **10** 613–627.

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

FORD, I., NORRIE, J. and AHMADI, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* **14** 735–746.

GAIL, M. H., WIEAND, S. and PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71** 431–444.

GARRETT, A. D. (2003). Therapeutic equivalence: Fallacies and falsification. *Statistics in Medicine* **22** 741–762.

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statist. Sci.* **9** 55–93.

GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.

GRIZZLE, J. E. (1965). The two-period change-over design and its use in clinical trials. *Biometrics* **21** 467–480.

HA, I. D., LEE, Y. and SONG, J. K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis* **8** 163–176.

HAUCK, W. W., HYSLOP, T., CHEN, M. L., PATNAIK, R. and WILLIAMS, R. L. (2000). Subject-by-formulation interaction in bioequivalence: Conceptual and statistical issues. *Pharmaceutical Research* **17** 375–380.

HEAGERTY, P. J. and KURLAND, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88** 973–985.

JIANG, J. (1996). REML estimation: Asymptotic behavior and related topics. *Ann. Statist.* **24** 255–286.

JIANG, J. (1997). Wald consistency and the method of sieves in REML estimation. *Ann. Statist.* **25** 1781–1803.

JIANG, J. (1999). On maximum hierarchical likelihood estimators. *Comm. Statist. Theory Methods* **28** 1769–1775.

JIANG, J. (2001). Goodness-of-fit tests for mixed model diagnostics. *Ann. Statist.* **29** 1137–1164.

KANG, W., LEE, M. and LEE, Y. (2003). HGLM versus conditional estimators for the analysis of clustered binary data. *Statistics in Medicine*. To appear.

LANE, P. W. and NELDER, J. A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics* **38** 613–621.

LEE, Y. and NELDER, J. A. (2003a). Likelihood for random-effect models. Unpublished manuscript.

LEE, Y. and NELDER, J. A. (2003b). Double hierarchical generalized linear models. *Comput. Statist.* To appear.

LEE, Y., NOH, M. and RYU, K. (2003). HGLM modelling of dropout process using frailty model. Unpublished manuscript.

LEE, Y., YUN, S. and LEE, Y. (2003). Analyzing weather effects on airborne particulate matter with HGLM. *Environmetrics* **14** 687–697.

LIN, X. and BRESLOW, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Amer. Statist. Assoc.* **91** 1007–1016.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayesian estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 1–41.

NEUHAUS, J. M., HAUCK, W. W. and KALBFLEISCH, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79** 755–762.

NOH, M. and LEE, Y. (2003). Review of estimating methods for binary data in generalised linear mixed models. Unpublished manuscript.

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58** 545–554.

PAWITAN, Y. (2001). *In All Likelihood*: *Statistical Modelling and Inference Using Likelihood*. Clarendon Press, Oxford.

PEARL, J. (2000). *Causality*: *Models*, *Reasoning and Inference*. Cambridge Univ. Press.

RICHARDSON, A. M. and WELSH, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Austral. J. Statist.* **36** 31–43.

ROBINSON, L. D. and JEWELL, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *Internat. Statist. Rev.* **59** 227–240.

ROM, D. M. and HWANG, E. (1996). Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine* **15** 1489–1505.

SCHALL, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78** 719–727.

SENN, S. J. (1997). Comment on "Testing for individual and population equivalence based on the proportion of similar responses," by D. M. Rom and E. Hwang (letter; comment). *Statistics in Medicine* **16** 1303–1306.

SENN, S. J. (2000a). Consensus and controversy in pharmaceutical statistics (with discussion). *The Statistician* **49** 135–176.

SENN, S. J. (2000b). The many modes of meta. *Drug Information J.* **34** 535–549.

SENN, S. J. (2001). Statistical issues in bioequivalance. *Statistics in Medicine* **20** 2785–2799.

SENN, S. J. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. Wiley, Chichester.

SHEINER, L. B., ROSENBERG, B. and MELMON, K. L. (1972). Modelling of individual pharmacokinetics for computer-aided drug dosage. *Computers and Biomedical Research* **5** 411–459.

SHUN, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. *J. Amer. Statist. Assoc.* **92** 341–349.

SHUN, Z. and MCCULLAGH, P. (1995). Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B* **57** 749–760.

STINE, R. A. and HEYSE, J. F. (2001). Non-parametric estimates of overlap. *Statistics in Medicine* **20** 215–236.

TEN HAVE, T. and LOCALIO, A. R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* **55** 1022–1029.

THALL, P. F. and VAIL, S. C. (1990). Some covariance-models for longitudinal count data with overdispersion. *Biometrics* **46** 657–671.

YATES, F. (1982). Reader reaction—regression-models for repeated measurements. *Biometrics* **38** 850–853.

YUN, S. and LEE, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Comp. Statist. Data Anal.* **45** 639–650.