BIOMETRICS 44, 1049–1060
December 1988

Models for Longitudinal Data: A Generalized Estimating Equation Approach

Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert

Department of Biostatistics, The Johns Hopkins University, 615 North Wolfe Street,
Baltimore, Maryland 21205, U.S.A.

SUMMARY

This article discusses extensions of generalized linear models for the analysis of longitudinal data. Two approaches are considered: subject-specific (SS) models in which heterogeneity in regression parameters is explicitly modelled; and population-averaged (PA) models in which the aggregate response for the population is the focus. We use a generalized estimating equation approach to fit both classes of models for discrete and continuous outcomes. When the subject-specific parameters are assumed to follow a Gaussian distribution, simple relationships between the PA and SS parameters are available. The methods are illustrated with an analysis of data on mother's smoking and children's respiratory disease.

1. Introduction

This article considers statistical methods for longitudinal data where the broad scientific objective is to describe an outcome, y_{it}, for subject i at time t as a function of covariates, $\mathbf{x}_{it}$. Longitudinal data are characterized by the fact that repeated observations for a subject tend to be correlated. This correlation presents additional opportunities and challenges for analysis.

With independent observations, generalized linear models (GLMs) (McCullagh and Nelder, 1983) and quasi-likelihood (Wedderburn, 1974; McCullagh, 1983) have recently unified regression methods for a variety of discrete and continuous variables. Linear, logistic, and Poisson regression as well as some parametric survival analysis models are special cases. The objective of this article is to discuss approaches to the analysis of dependent, longitudinal data with similarly diverse types of outcome variables.

The GLM can be extended for time-dependent data in a variety of ways. Zeger and Qaqish (1988) and Kaufmann (1987) discuss generalized linear models for the conditional distribution of an outcome given its past. Alternatively, the parameters in a GLM can be assumed to vary across time as a stochastic process and/or across subjects according to a mixing distribution. We focus on the case where there is heterogeneity across subjects.

There are two distinct approaches to longitudinal data analysis in this case. First, the heterogeneity can be explicitly modelled; we will refer to this as the "subject-specific" (SS) approach. The mixed model is an example where the subject-specific effects are assumed to follow a parametric distribution across the population. Mixed linear models (Laird and Ware, 1982; Ware, 1985) for continuous longitudinal data are in common use. Mixed generalized linear models for non-Gaussian outcomes have recently become a research

Key words: Generalized estimating equations; Generalized linear models; Longitudinal data; Quasi-likelihood; Random effects.

focus. See Stiratelli, Laird, and Ware (1984), Anderson and Aitkin (1985), and Gilmour, Anderson, and Rae (1985) for applications to binomial data.

Second, the population-averaged response can be modelled as a function of covariates without explicitly accounting for subject to subject heterogeneity. The regression coefficients have interpretation for the population rather than for any individual and hence we will use the term "population-averaged" (PA) model in this case. Liang and Zeger (1986), Zeger and Liang (1986), Stram, Wei, and Ware (1988), and Moulton (unpublished Ph.D. dissertation, The Johns Hopkins University, 1986) have previously discussed examples of PA models.

This article first contrasts SS and PA models in more detail, indicating their respective domains of application. The mixed GLM is used as a basis for discussion. A generalized estimating equations approach (Liang and Zeger, 1986) useful for fitting both SS and PA models is then discussed in Section 3. This approach is an extension of quasi-likelihood to the analysis of dependent data. The methodology is illustrated with an analysis of respiratory infection data from the Harvard Study of Air Pollution and Health (Ware et al., 1984).

2. Subject-Specific and Population-Averaged Models

This section distinguishes between SS and PA models for longitudinal data, indicating their domains of application. While extra nomenclature can be a nuisance, we believe it is useful in this case to differentiate two distinct extensions of GLMs for longitudinal data analysis.

We begin with the mixed generalized linear model, an example of a subject-specific model. Let y_{it} be an outcome random variable and $\mathbf{x}_{it}$ a $p \times 1$ vector of fixed covariates at time t for subject i, where $t = 1, \ldots, n_i$ and $i = 1, \ldots, K$. Let $\mathbf{z}_{it}$ be a $q \times 1$ vector of covariates (typically a subset of $\mathbf{x}_{it}$) associated with a $q \times 1$ random effect, $\mathbf{b}_i$, and let $u_{it} = \mathrm{E}(y_{it} \mid \mathbf{b}_i)$. Under the mixed GLM, the responses for subject i are assumed to satisfy

$$h(u_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i, \qquad \mathrm{var}(y_{it} \mid \mathbf{b}_i) = g(u_{it}) \cdot \phi, \tag{2.1}$$

where $\mathbf{b}_i$ is an independent observation from a mixture distribution, F. The functions h and g are referred to as the "link" and "variance" functions, respectively. The objective of analysis is to estimate the fixed effects coefficients, $\boldsymbol{\beta}$, parameters of F, and possibly the scale parameter, ϕ. An example is the logistic Gaussian mixed model studied by Stiratelli et al. (1984), in which it is assumed that $\mathrm{logit}(u_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i$, $\mathrm{var}(y_{it} \mid \mathbf{b}_i) = u_{it}(1 - u_{it})$, and $\mathbf{b}_i$ is an independent Gaussian random vector with mean $\mathbf{0}$ and covariance $\mathbf{D}$, i.e., $\mathbf{b}_i \sim \mathrm{G}(\mathbf{0}, \mathbf{D})$.

For discussion, suppose $\mathbf{x}_{it} = \mathbf{z}_{it} = (1, t)'$ and let $\mathbf{b}_i = (b_{0i}, b_{1i})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Then in the logistic mixed model, the log-odds of a positive response for subject i at time t is the linear function of time $(\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t$. Thus, $\exp(\beta_1 + b_{1i})$ is the odds ratio of a positive response at time $t + 1$ relative to time t for subject i. Since $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$, the parameter β_1 describes on average how an *individual's* probability of positive response depends on time.

In the population-averaged approach to longitudinal analysis, the marginal expectation, $\mu_{it} = \mathrm{E}(y_{it})$, is the focus. That is, we assume

$$h^*(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}^* \quad \text{and} \quad \mathrm{var}(y_{it}) = g^*(\mu_{it}) \cdot \phi$$

for some link function h^* and variance function g^*. Here, $\boldsymbol{\beta}^*$ describes how the population-averaged response rather than one subject's response depends on the covariates. In the logistic case with $\mathbf{x}_{it} = (1, t)'$, β_1^* is the change on a logit scale in the fraction of positive responses per unit time, rather than the typical change for an individual subject.

The principal distinction between SS and PA models is whether the regression coefficients describe an individual's or the average population response to changing $\mathbf{x}$. A secondary distinction is in the nature of the assumed time dependence. PA models only describe the covariance among repeated observations for a subject; SS models explain the source of this covariance. In PA models, the covariance matrix must be positive-definite but is otherwise unrestricted. In SS models, the time dependence arises solely from the shared subject effects, $\mathbf{b}_i$, in the conditional mean. The covariance matrix is thus fully determined by the choices of $g(u_{it})$ and F. For example, in a logistic model with Gaussian random intercept, only positive correlation is possible.

SS models are desirable when the response for an individual rather than for the population is the focus—for example, in studies of growth curves. Effective use of SS models is limited, however, by the information available per subject. In many longitudinal studies, each subject has few observations and it is not possible to estimate separate regression coefficients, $\boldsymbol{\beta} + \mathbf{b}_i$. Assuming the $\mathbf{b}_i$'s follow a particular distribution, as is done in the mixed model, is a vehicle for borrowing strength across subjects to estimate $\boldsymbol{\beta}$, the typical SS parameter. However, inferences about $\boldsymbol{\beta}$ may depend on the assumed form of the distribution of the $\mathbf{b}_i$'s, which cannot be checked without extensive data per subject.

PA models are most effectively used in population studies such as in epidemiology. Here the difference in the population-averaged response between two groups with different risk factors is more the focus than is the change in an individual's response. For example, if x_{it} indicates whether subject i smokes at time t, and y_{it} is the presence/absence of respiratory infection, the PA model estimates the difference in infection rates between smokers and nonsmokers; the SS model estimates the expected change in an individual's probability of infection given a change in smoking status.

An advantage of PA models is that the population-averaged response for a given covariate value, x_{it}, is directly estimable from observations without assumptions about the heterogeneity across individuals in the parameters. PA parameters are in this sense one step closer to the data than SS parameters. On the other hand, PA parameters depend on the degree of heterogeneity in the population (F). The same process in two populations with different degrees of heterogeneity will lead to different PA parameter values.

Under the mixed GLM in (2.1),

$$\mu_{it} = \mathrm{E}(y_{it}) = \mathrm{E}[\mathrm{E}(y_{it} \mid \mathbf{b}_i)] = \int h^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i)\, dF\,(\mathbf{b}_i). \tag{2.2}$$

Note that μ_{it} depends only on $\mathbf{x}_{it}$ and on F. When h is a nonlinear function and we assume $h(u_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i$, it is usually not true that $h(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$. Trivially, if there is no heterogeneity, i.e., $\mathbf{b}_i = \mathbf{0}$ for all i, then PA and SS models are the same and $h(u_{it}) = h(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$. Also, if h is the identity link, $h(\mu_{it}) = \mu_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}$. But in general, the link function that transforms u_{it} into a linear function of $\mathbf{x}_{it}$ does not also do the same for μ_{it}. This is illustrated in Figure 1, where we have assumed

$$\operatorname{logit}(u_{it}) = \eta_{it} + \mathbf{b}_i, \quad \mathbf{b}_i \sim \mathrm{G}(\mathbf{0}, \mathbf{D}),$$

and display the marginal mean, μ_{it}, as a function of η_{it} for several values of $\mathbf{D}$. Note the dependence of the marginal expectation on the random effects variance. There is attenuation of the effect of the covariates, as is well known in the context of errors-in-variables regression (e.g., Stefanski, 1985).

For mixed models with identity link, the distinction between subject-specific and population-averaged models is less important. In addition, inferences about the regression coefficients are robust to misspecification of the model for time dependence, a principal

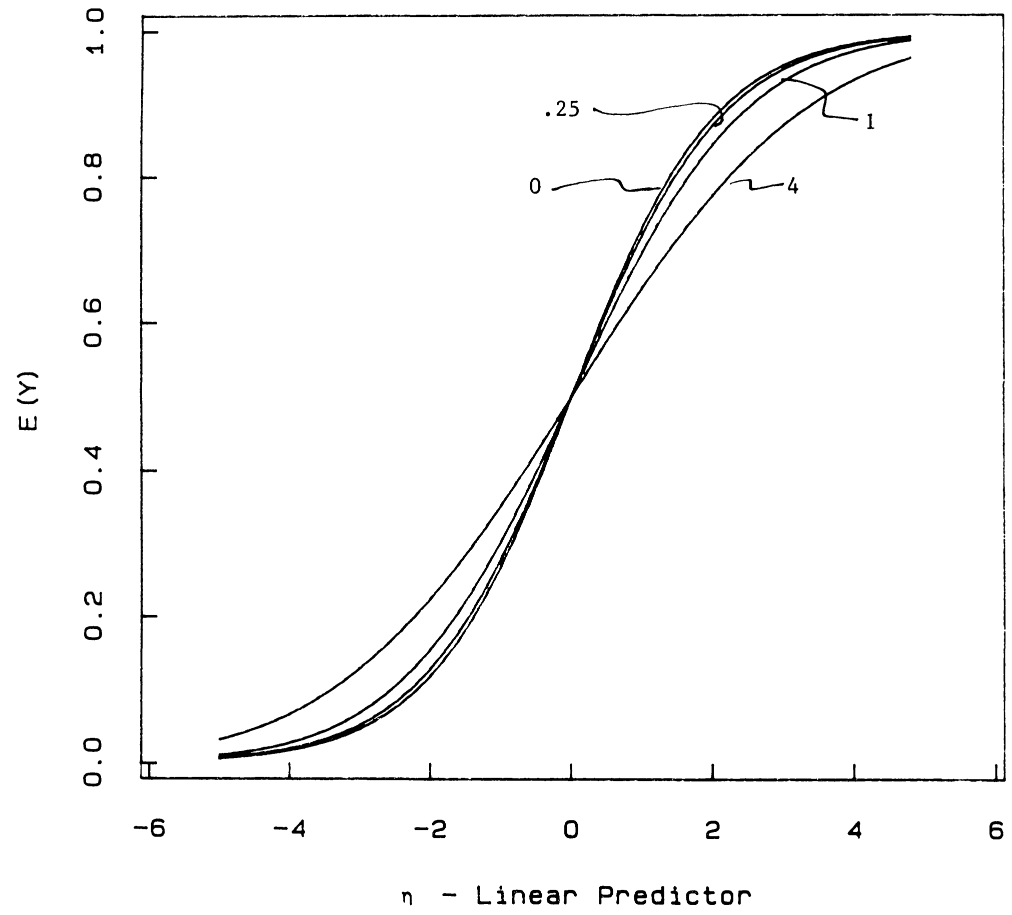

Figure 1. E(Y) vs η, where Y satisfies logit[E($Y\,|\,b$)] = $\eta + b$ and $b \sim G(0, D)$.

difference between the PA and SS approaches in linear models. To develop these points, briefly consider the linear mixed model (Laird and Ware, 1982), which in vector notation is given by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.3}$$

where

$$\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})', \quad \mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})', \quad \mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})', \quad \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})',$$

and where E($\mathbf{b}_i$) = $\mathbf{0}$, cov($\mathbf{b}_i$) = $\mathbf{D}$, cov(ε_i) = $\sigma^2\mathbf{I}$, and cov($\mathbf{b}_i$, ε_i) = $\mathbf{0}$. For simplicity of notation, we assume $\mathbf{X}_i = \mathbf{Z}_i$; extension to the general case is automatic. Note that the subject-specific coefficient for the ith individual is $\boldsymbol{\beta} + \mathbf{b}_i$. Since E($\mathbf{b}_i$) = $\mathbf{0}$, $\boldsymbol{\beta}$ has interpretation as the typical SS parameter. Alternatively, (2.3) can be expressed as

$$\mathrm{E}(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}, \quad \mathrm{cov}(\mathbf{Y}_i) = \sigma^2\mathbf{I} + \mathbf{X}_i\mathbf{D}\mathbf{X}_i' = \mathbf{V}_i. \tag{2.4}$$

Here, $\boldsymbol{\beta}$ has the interpretation as the rate of change in the population-averaged $\mathbf{Y}$ with $\mathbf{X}$. The random effects in the linear mixed model do not alter the marginal expectation of $\mathbf{Y}$, only the marginal covariance matrix. Hence, $\boldsymbol{\beta}$ has both a SS and PA interpretation.

The second point about the linear case is that consistent inferences about $\boldsymbol{\beta}$ can be obtained by least squares given only correct specification of the marginal expectation of $\mathbf{Y}$

and the usual regularity conditions. The least squares estimating equations for $\boldsymbol{\beta}$ have the form

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbf{X}_i' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) = \mathbf{0}. \tag{2.5}$$

Note that even when $\mathbf{V}_i$ is misspecified, $\mathrm{E}[\mathbf{U}(\boldsymbol{\beta})] = \mathbf{0}$ and hence the root of (2.5) is consistent. In addition, the robust variance estimate (White, 1982; Royall, 1986),

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{K} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \left[\sum_{i=1}^{K} \mathbf{X}_i' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{Y}_i - \hat{\mathbf{X}}_i\boldsymbol{\beta})' \mathbf{V}_i^{-1} \mathbf{X}_i\right] \left(\sum_{i=1}^{K} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1},$$

is also consistent given only that $\mathrm{E}(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. Thus, for large K, consistent inferences require correct specification of only the first moment. In the linear case, the specification of the first moment is the same in subject-specific and in population-averaged models. Hence, this distinction is less important.

3. Generalized Estimating Equations for PA and SS Models

In this section, we describe an estimating equation approach for fitting either PA or SS models to longitudinal data. To introduce the method, we briefly discuss PA models. See Liang and Zeger (1986) and Zeger and Liang (1986) for details. We then focus on the SS case.

3.1 *Estimating Regression Coefficients*

To model the marginal expectation, μ_{it}, we assume $h^*(\mu_{it}) = x_{it}'\boldsymbol{\beta}^*$ and $\mathrm{var}(y_{it}) = g^*(\mu_{it})\phi$. Let $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{Y}_i) = \{h^{*-1}(\mathbf{x}_{i1}'\boldsymbol{\beta}^*), \ldots, h^{*-1}(\mathbf{x}_{in_i}'\boldsymbol{\beta}^*)\}'$ and $\mathbf{A}_i = \mathrm{diag}\{g^*(\mu_{i1}), \ldots, g^*(\mu_{in_i})\}$. For independent observations, $\mathrm{cov}(\mathbf{Y}_i) = \mathbf{A}_i \cdot \phi$. As we expect correlation among repeated observations for a subject, let $\mathbf{R}_i(\boldsymbol{\alpha})$ be a "working" correlation matrix perhaps depending on an $s \times 1$ vector of unknown parameters, $\boldsymbol{\alpha}$. We estimate $\boldsymbol{\beta}^*$ by solving the "generalized estimating equation" (GEE)

$$\mathbf{U}(\boldsymbol{\beta}^*) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}^*} \mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{V}_i(\boldsymbol{\alpha}) = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$. Liang and Zeger (1986) show that $\hat{\boldsymbol{\beta}}^*$, the solution of (3.1), is consistent and asymptotically ($K \rightarrow \infty$) Gaussian given only correct specification of the mean and the usual regularity conditions. A robust variance estimate

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}^*} = \mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1}, \tag{3.2}$$

where

$$\mathbf{M}_0 = \sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^*} \hat{\mathbf{V}}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}^*}$$

and

$$\mathbf{M}_1 = \sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^*} \hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{\mathbf{V}}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}^*}$$

is also consistent even when $\mathrm{cov}(\mathbf{Y}_i) \neq \mathbf{V}_i$.

The GEE can also be used to fit the mixed generalized linear model, as has previously been discussed by Gilmour et al. (1985) for a probit model of binomial responses in the animal breeding context. Below, we consider the GLM class of outcomes, give consistent variance estimates, and establish connections between PA and SS parameters when the random effects distribution is Gaussian.

To use the GEE approach for the mixed GLM, we calculate the marginal moments, $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$, from the conditional moments and the random effects distribution, F. We then solve the GEE (3.1) as discussed in Liang and Zeger (1986). Given the conditional moments in (2.1) and a distribution, F, for the random effects, the marginal expectation, $\boldsymbol{\mu}_i$, has the form of (2.2). The marginal covariance matrix is

$$\mathbf{V}_i = \operatorname{cov}[\mathrm{E}(\mathbf{Y}_i \mid \mathbf{b}_i)] + \mathrm{E}[\operatorname{cov}(\mathbf{Y}_i \mid \mathbf{b}_i)] \tag{3.3}$$

with s, t element

$$[\mathbf{V}_i]_{s,t} = \int (u_{is} - \mu_{is})(u_{it} - \mu_{it})\, dF(\mathbf{b}_i) + \phi I(s = t) \int g(u_{is})\, dF(\mathbf{b}_i),$$

where $u_{it} = h^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i)$ and $I(s = t)$ is the indicator function with value 1 if $s = t$ and 0 otherwise. Having evaluated $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ for each subject, we solve the GEE given in (3.1) for $\boldsymbol{\beta}$. Note that the GEE is a function of F, which is assumed to be known at a given iteration.

If F is the Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$, the expression for the marginal mean simplifies or is easily approximated for the standard link functions. For the identity link $[h(u) = u]$, we have trivially $\mu_{it} = \mathrm{E}(y_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$. For the log link $[h(u) = \log(u)]$, $\mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{D}\mathbf{z}_{it}/2)$. That is, the random effect leads to a simple offset, $\mathbf{z}_{it}'\mathbf{D}\mathbf{z}_{it}/2$, in the marginal mean. When $h(u) = \Phi^{-1}(u)$, the probit link, $\mu_{it} = \Phi(a_p(\mathbf{D}) \cdot \mathbf{x}_{it}'\boldsymbol{\beta})$, where $a_p(\mathbf{D}) = |\mathbf{D}\mathbf{z}_{it}\mathbf{z}_{it}' + \mathbf{I}|^{-q/2}$ and q is the dimension of $\mathbf{b}_i$. This expression for $a_p(\mathbf{D})$ is a generalized form of the parameter s in Gilmour et al. (1985). For the logit link, an exact closed-form expression for the marginal mean is unavailable. However, using a cumulative Gaussian approximation to the logistic function (Johnson and Kotz, 1970, p. 6) leads to the expression $\operatorname{logit}(\mu_{it}) \approx a_l(\mathbf{D}) \cdot \mathbf{x}_{it}\boldsymbol{\beta}$, where $a_l(\mathbf{D}) = |c^2\mathbf{D}\mathbf{z}_{it}\mathbf{z}_{it}' + \mathbf{I}|^{-q/2}$ and $c = 16\sqrt{3}/(15\pi)$. Figure 2 shows the quality of this approximation for different values of $\mathbf{D}$ when $q = 1$. Note that the mixed models on the logit and probit scales lead to a rescaling of the linear predictor in the expression for the marginal mean; however, the same link function can be used for both the conditional and marginal expectations.

Ideally, simple formulae would exist for $\mathbf{V}_i = \operatorname{cov}(\mathbf{Y}_i)$ as well. This is not the case except for the linear link. However, only an approximation for $\mathbf{V}_i$ is necessary to obtain consistent and nearly efficient inferences for $\boldsymbol{\beta}$ using the GEE approach when the number of subjects, K, is large relative to the number of observations per subject, n_i, and F is given (Liang and Zeger, 1986). Expanding the link function in a Taylor series about $\mathbf{b}_i = \mathbf{0}$ gives the approximation

$$\begin{aligned}\operatorname{cov}(\mathbf{Y}_i) &\approx \operatorname{cov}\left[h^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta}) + \frac{\partial h^{-1}}{\partial \mathbf{b}_i}(\mathbf{x}_{it}'\boldsymbol{\beta})\mathbf{b}_i\right] \\ &\quad + \phi \mathrm{E}\left\{g\left[h^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta}) + \frac{\partial h^{-1}}{\partial \mathbf{b}_i}(\mathbf{x}_{it}'\boldsymbol{\beta})\mathbf{b}_i\right]\right\} \\ &\approx \mathbf{L}_i\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{L}_i + \phi\mathbf{A}_i = \tilde{\mathbf{V}}_i,\end{aligned} \tag{3.4}$$

where $\mathbf{L}_i = \operatorname{diag}\{\partial h^{-1}(u)/\partial u, u = \mathbf{x}_{it}'\boldsymbol{\beta}, t = 1, \ldots, n_i\}$. The quality of a similar approximation has been studied for the probit link by Gilmour et al. (1985).

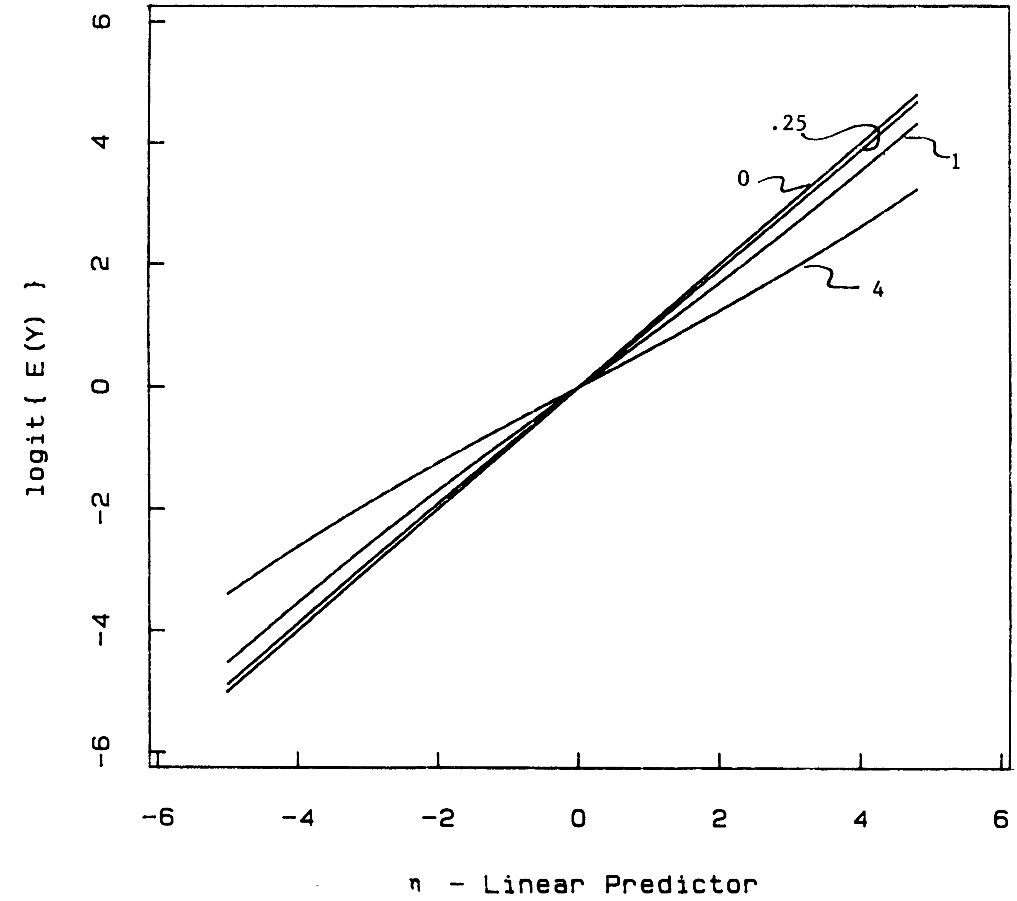

Figure 2. logit[E(Y)] vs η, where Y satisfies logit[E($Y \mid b$)] $= \eta + b$ and $b \sim$ G(0, D). Notice that the logit of the marginal mean is nearly linear.

Using $\tilde{\mathbf{V}}_i$ as an approximation to $\mathbf{V}_i$, we have that conditional on F, $\sqrt{K}(\hat{\boldsymbol\beta} - \boldsymbol\beta)$ is asymptotically multivariate Gaussian with mean $\mathbf{0}$ (if an exact expression for μ_{it} is used and small bias otherwise) and with variance that can be consistently estimated by

$$\hat{\mathbf{V}}_{\hat{\boldsymbol\beta}} = \left(\sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol\mu}_i'}{\partial \boldsymbol\beta} \hat{\tilde{\mathbf{V}}}_i^{-1} \frac{\partial \hat{\boldsymbol\mu}_i}{\partial \boldsymbol\beta}\right)^{-1} \left(\sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol\mu}_i'}{\partial \boldsymbol\beta} \hat{\tilde{\mathbf{V}}}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol\mu}_i)(\mathbf{Y}_i - \hat{\boldsymbol\mu}_i)' \hat{\tilde{\mathbf{V}}}_i^{-1} \frac{\partial \hat{\boldsymbol\mu}}{\partial \boldsymbol\beta}\right) \left(\sum_{i=1}^{K} \frac{\partial \hat{\boldsymbol\mu}_i'}{\partial \boldsymbol\beta} \hat{\tilde{\mathbf{V}}}_i^{-1} \frac{\partial \hat{\boldsymbol\mu}_i}{\partial \boldsymbol\beta}\right)^{-1}.$$

It is important to emphasize that the distributional result assumes F is given. In mixed models with nonlinear links, estimates for $\boldsymbol\beta$ and for parameters of F ($\mathbf{D}$ in the Gaussian case) are not even asymptotically orthogonal as they are in linear models. In general, inferences about $\boldsymbol\beta$ depend in a complicated way on those for F. There are important exceptions, however, when $\mathbf{b}_i$ is Gaussian. For the log link, changing $\mathbf{D}$ changes only an offset in the marginal mean. For $\mathbf{X}$'s approximately orthogonal to the intercept, changing $\mathbf{D}$ has little effect on their coefficients and the conditional inferences are satisfactory. For the probit and logit links, $\mathbf{D}$ need not be known for testing the null hypothesis $\beta_j = 0$. To see this, note that for any $\mathbf{D}$, $h[\mathrm{E}(y_{it})] \approx \mathbf{x}_{it}'\boldsymbol\beta a_h(\mathbf{D})$, where h is the logit or probit function. The standard error of $\hat\beta_j$ is also proportional to $a_h(\mathbf{D})$. Hence, the ratio of $\hat\beta_j$ to its standard error is approximately independent of $\mathbf{D}$.

3.2 *Estimating* D and ϕ

The approximation (3.4) can be used to obtain a rough estimate of the random effects variance, **D**. We have

$$\mathbf{V}_i \approx \mathbf{L}_i \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \mathbf{L}_i + \phi \mathbf{A}_i$$

so that

$$\mathbf{D} \approx (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i' \mathbf{L}_i (\mathbf{V}_i - \phi \mathbf{A}_i) \mathbf{L}_i^{-1} \mathbf{Z}_i (\mathbf{Z}_i' \mathbf{Z}_i)^{-1}.$$

We use the moment estimator

$$\hat{\mathbf{D}} = \frac{1}{K} \sum_{i=1}^{K} (\mathbf{Z}_i' \mathbf{Z}_i)^{-1} \mathbf{Z}_i \hat{\mathbf{L}}_i^{-1} [(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)' - \hat{\phi} \hat{\mathbf{A}}_i] \hat{\mathbf{L}}_i^{-1} \mathbf{Z}_i' (\mathbf{Z}_i' \mathbf{Z}_i)^{-1}. \qquad (3.5a)$$

The scale parameter is estimated from the diagonal terms of the covariance matrix. Note that

$$\mathrm{E}(y_{it} - \mu_{it})^2 \approx \phi g(\mu_{it}) + (\mathbf{L}_{i_{tt}})^2 \mathbf{z}_{it}' \mathbf{D} \mathbf{z}_{it},$$

which leads to the moment estimator

$$\hat{\phi} = \sum_{i=1}^{K} \sum_{t=1}^{n_i} \frac{(y_{it} - \hat{\mu}_{it})^2 - (\hat{L}_{i_{tt}})^2 \mathbf{z}_{it}' \hat{\mathbf{D}} \mathbf{z}_{it}}{g(\hat{\mu}_{it})}. \qquad (3.5b)$$

To calculate $\hat{\boldsymbol{\beta}}$ and $(\hat{\mathbf{D}}, \hat{\phi})$ simultaneously, we iterate solving equations (3.1) and (3.5). This algorithm has been found to converge except when the linear approximation becomes inaccurate, for example, when the probability of response in logistic regression becomes too large or small and/or when **D** becomes large. Because of the approximations involved for nonlinear links, we currently use $\hat{\mathbf{D}}$ and $\hat{\phi}$ as rough estimators and examine the sensitivity of $\hat{\boldsymbol{\beta}}$ to changes in **D** and ϕ in the neighborhood of $\hat{\mathbf{D}}$, $\hat{\phi}$. Further work on estimating **D** and ϕ for specific link functions is required.

4. Example: Children's Respiratory Disease and Mother's Smoking

We illustrate the GEE approach with an analysis of data from the Harvard Study of Air Pollution and Health (Ware et al., 1984). Data are available for 537 children from Steubenville, Ohio, each of whom was examined annually from age 7 to age 10. Whether the child had respiratory infection in the year prior to each exam was reported by the mother. Mother's smoking status [regular smoker (1) or not (0)], a time-independent variable, was determined at the first interview. The subset of data used here was obtained from and previously analyzed by Laird, Beck, and Ware (unpublished technical report, Department of Biostatistics, Harvard University, 1986). Only subjects with complete records were available; however, the previous analysis found little difference when all subjects were included. Our objective is to illustrate the GEE method, demonstrating the connections between PA and SS models.

To fit a PA model, the marginal probability of respiratory infection, μ_{it}, is assumed to satisfy

$$\mathrm{logit}(\mu_{it}) = \beta_0^* + MS\, \beta_1^* + AGE\, \beta_2^* + (AGE \cdot MS)\beta_3^*, \qquad (4.1)$$

where $MS = 1$ if mother smoked and 0 if not, and *AGE* is in years since the 9th birthday. Note that whether a child had infection the previous year is not explicitly included in the model. The purpose is to compare the rate of respiratory disease for children whose mothers smoke to the rate for children whose mothers do not smoke.

Table 1 presents the coefficients and robust Z-statistics for the PA model based on the following three working assumptions about the correlation: $\mathbf{R} = \mathbf{I}$ (repeated observations uncorrelated); $R_{jk} = \alpha, j \neq k$ (exchangeable correlation); $R_{jk} = \alpha(|j - k|)$ (stationary correlation). In practice, we choose $\mathbf{R}$ based on empirical estimates of the correlation. We use three different correlation assumptions here only to demonstrate that both the estimates and Z-statistics show little dependence on the choice of $\mathbf{R}$, despite the presence of substantial correlation among these data. In the exchangeable working model ($R_{jk} = \alpha, j \neq k$), $\hat{\alpha} = .346$. In the stationary case, the correlations for lags of 1, 2, and 3 years were estimated to be .40, .31, and .31, respectively. Either of these alternatives appears reasonable.

The yearly rate of respiratory infection for 9-year-olds ($AGE = 0$) with nonsmoking mothers is approximately 15%. The PA model indicates that the rate decreases with age by about 2% per year. The rate of illness for children of smoking mothers is about 1.35 as high as in those with nonsmoking mothers with approximate 95% interval (.92, 1.96). As the Z-statistic for β_1^* is 1.58, the evidence from these data only moderately supports the smoking–respiratory illness relationship.

Table 2 compares the robust Z-statistics with those obtained if we assume repeated observations are uncorrelated both in estimating $\boldsymbol{\beta}$ *and* in calculating its variance—that is, if we assume the independence working model is correct. The consistent statistic, Z_R, is smaller for both time-independent covariates (intercept, mother's smoking) and larger for the time-dependent covariates (age, age–smoking interaction). Positive correlation within

Table 1

Coefficients and robust Z-statistics ($\hat{\beta}^*/s.e._{\hat{\beta}^*}$) *for the population-averaged model of equation* (4.1) *for three different assumptions about the correlation among repeated observations for a subject*

	Uncorrelated		Exchangeable[a]		Stationary[b]	
Coefficient	$\hat{\beta}^*$	Z	$\hat{\beta}^*$	Z	$\hat{\beta}^*$	Z
Intercept	−1.90	−16.0	−1.90	−15.8	−1.90	−15.8
Mother's smoking ($MS = 0$ if no)	.313	1.7	.303	1.6	.298	1.6
AGE	−.140	−2.4	−.137	−2.4	−.139	−2.4
$MS \cdot AGE$	.0699	.79	.0657	.75	.0704	.80

[a] $\hat{\alpha} = .346$.

[b] $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3) = (.40, .31, .30)$.

Table 2

Comparison of robust Z-statistics from independence working model and naive Z-statistics in which the independence was assumed in calculating both the coefficients and their standard errors

	Naive Z-statistic	Robust Z-statistic	
Coefficient	Z_N	Z_R	$(Z_N - Z_R)/Z_n$
Intercept	−21.4	−15.8	.26
Mother's smoking (MS)	2.3	1.6	.30
AGE	−2.0	−2.4	−.20
$MS \cdot AGE$	.64	.79	−.23

subjects makes estimates of differences among subjects less precise than they would be with independent observations. In contrast, within-subject changes can be estimated more precisely, as is indicated in Table 2. Note that ignoring the correlation leads to incorrectly interpreting the data as strong evidence for a mother's smoking effect.

Now consider the subject-specific model. We describe the probability of respiratory infection for an individual, $u_{it} = \mathrm{E}(y_{it} \mid \mathbf{b}_i)$, as a function of the covariates assuming

$$\operatorname{logit}(u_{it}) = \beta_0 + MS\,\beta_1 + AGE\,\beta_2 + (AGE \cdot MS)\beta_3 + b_{0i}, \tag{4.2}$$

where we assume $b_{0i} \sim \mathrm{G}(0, D)$. Here, the coefficients are log-odds ratios for a single child. That is, the *MS* coefficient, β_1, indicates how one child's risk would change if his mother changed smoking status rather than how the average risk over the population differs, as is the case for PA coefficients. For illustration, only the intercept in the linear predictor is assumed to vary across subjects. Inferences about mother's smoking when both the intercept and age coefficients are random are qualitatively similar. Table 3 presents coefficients and robust *Z*-statistics for a range of values of *D*. The approximate variance, given in equation (3.5), was used in the estimating equation. Note that the subject-specific parameters are greater in absolute value than the population-averaged analogues ($D = 0$) and increase with the variance, *D*, of the random effect. The random effects variability shrinks the fixed effects parameters toward 0 in the logistic model. This is well known in the context of errors-in-variables (e.g., Stefanski, 1985). The degree of shrinkage depends on the study design ($\mathbf{z}_{it}$'s), but in this simple case can be approximated by $a_l(D)$. For example, the *MS* coefficients for various *D*'s can be compared with $a_l(D) \cdot \hat{\beta}_1(0)$, where $\hat{\beta}_1(0)$ is the *MS* coefficient when $D = 0$ (see Table 3). Note that this approximation is within 2% for all values of *D*. Hence, a reasonable estimate of SS parameters can in this case be obtained from the PA model results.

Tests for the null hypothesis that mother's smoking does not affect children's respiratory disease can be based on the *MS* *Z*-statistic in Table 3. It does not change even as *D* changes from 0 to 4.0. Hence, this inference is approximately independent of the random effects variance. On the other hand, the absolute magnitude of the SS coefficient changes substantially as a function of *D*.

Unfortunately, there is little information about *D* in these data. The algorithm suggested above fails to converge because the linear approximation to the logit link function is not adequate for small respiratory disease propensity and large *D*. In their unpublished technical report, Laird et al. have found the profile likelihood for *D* to be flat when a Gaussian random effect is assumed. They give a point estimate of 4.4 and an approximate interval of (3, 6.5). In this range, the mother's smoking coefficient varies from .45 to .56.

Table 3

Coefficients and robust Z-statistics from subject-specific model with a Gaussian random intercept with variance D

	D							
	0		.5		1		4.0	
Coefficient	$\hat{\beta}$	*Z*	$\hat{\beta}$	*Z*	$\hat{\beta}$	*Z*	$\hat{\beta}$	*Z*
Intercept (β_0)	−1.91	−16	−2.1	−16	−2.2	−17	−2.9	−16
MS (β_1)	.313	1.6	.344	1.7	.368	1.7	.488	1.6
AGE (β_2)	−.14	−2.4	−.15	−2.5	−.16	−2.5	−.22	−2.5
MS · *AGE* (β_3)	.070	.79	.074	.78	.079	.77	.11	.78
$a_l(D)$	1.00		1.08		1.16		1.54	
$\hat{\beta}_1(0) \cdot a_l(D)$	.313		.338		.363		.488	
$\log\left[\dfrac{\hat{\beta}_1(0) \cdot a_l(D)}{\hat{\beta}_1(D)}\right]$	0.0		−.02		−.01		−.01	

In summary, the PA model indicates that the rate of children's respiratory disease is approximately 35% greater for children of smoking mothers. The SS modelling indicates that a child's risk if his or her mother stopped smoking would decrease by between 35% and 63% as the random effect variance ranges from 0 to 4.0. Note that the assumption $D = 4.0$ corresponds to a substantial amount of heterogeneity in children's propensity for respiratory disease. Under a Gaussian assumption, it implies that 95% of children have a probability of infection in a given year in the interval .0001 to .75. Finally, either the PA or SS model leads to the interpretation of these data as mild evidence for the relationship of mother's smoking and children's respiratory infection.

5. Discussion

In this paper, we have distinguished between population-averaged and subject-specific models for longitudinal data. The GEE approach can be used for both types of models. PA models describe how the average response across subjects changes with the covariates. Only the link function need be correctly specified to make consistent inferences about PA coefficients. The SS mixed models use the information contained in the population-averaged response as well as a distributional assumption about the heterogeneity among subjects to estimate subject-specific coefficients. Both the link function and the random effects distribution must be correctly specified for consistent inferences in this case. Estimates of SS parameters and of the random effects distribution are asymptotically correlated except for the linear link. Hence, the precision of $\hat{\boldsymbol{\beta}}$ depends on that of $\hat{F}$ ($\hat{\mathbf{D}}$ in the Gaussian case), which is more difficult to estimate from longitudinal data especially with nonlinear links and few observations per subject. When SS parameters are of primary interest, we believe care must be exercised in their interpretation.

The methods described here are closely related to previous work in two other contexts. Gail, Wieand, and Piantadosi (1984) examined the bias in estimates of treatment effect when a balanced covariate is omitted in the generalized linear model setting. They found that the bias is zero only for the identity and log links. The notion of omitting balanced covariates is conceptually similar to that of ignoring random effects. Our results agree with theirs regarding the identity and log links. In addition, we have established simple relationships between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ for other link functions. Second, the random effects problem is related to errors-in-variables regression as discussed for binary regression by Carroll et al. (1984) and for generalized linear models by Stefanski (1985). In the mixed model, the coefficients are random; in errors-in-variables the covariates are random. While this distinction leads to different results, there is substantial overlap.

Acknowledgements

We gratefully acknowledge partial support from NIMH Grant R01 MH40859-01, and NIH Grants 1-R29-AI25529-01 and 1-R29-GM39261-01. We thank the associate editor and referees, whose comments substantially improved an earlier draft.

Résumé

On discute, dans cet article, d'extensions des modèles linéaires généralisés, pour l'analyse de données longitudinales. On considère deux approches: les modèles spécifiques au sujet (SS), dans lesquels l'hétérogénéité dans les paramètres de régression, est modélisée explicitement; et les modèles moyennant sur une population (PA) dans lesquels, on s'intéresse à la réponse globale dans la population. On utilise, une approche généralisée des équations d'estimation, pour ajuster les deux classes de modèles pour des résultats discrets ou continus. Quand on suppose, que les paramètres spécifiques au sujet, suivent une distribution Gaussienne, on dispose de relations simples, entre les paramètres PA et SS. On illustre, les méthodes avec une analyse de données sur le tabagisme des mères et les maladies respiratoires des enfants.

REFERENCES

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

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71,** 19–25.

Gail, M. A., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. *Biometrika* **71,** 431–444.

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

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions, Vol.* 2. Boston: Houghton-Mifflin.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Annals of Statistics* **15,** 79–98.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

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

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics* **11,** 59–67.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Royall, R. M. (1986). Model robust inference using maximum likelihood estimators. *International Statistical Review* **54,** 221–226.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72,** 583–592.

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

Stram, D. O., Wei, L. J., and Ware, J. H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association* **83,** 631–637.

Ware, J. H. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *American Statistician* **39,** 95–101.

Ware, J. H., Dockery, D. W., Spiro, A., III, Speizer, F. E., and Ferris, B. G., Jr. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *American Review of Respiratory Disease* **129,** 366–374.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika* **50,** 1–25.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44,** 1019–1031.

Received November 1986; *revised July* 1987 *and March* 1988.