Further Details On The Analysis of Longitudinal Ordinal Response Data In Continuous Time

University of Wisconsin-Madison

Department of Biostatistics Technical Report 92

Michael R. Kosorok and Wei-Hsiung Chao[1]

ABSTRACT

A simple Markov model is developed for assessing the predictive effect of time-dependent covariates on an intermittently observed ordinal response in continuous time. This is accomplished by reparameterizing an ergodic intensity matrix in terms of its equilibrium distribution and a parametrically independent component which assesses the rate of movement between ordinal categories. The effect of covariates on the equilibrium distribution can then be modeled using any link appropriate for ordinal data. A robust maximum likelihood estimator based on this model which is consistent and asymptotically normal is constructed. Practical data analysis issues are discussed and a simple diagnostic tool for assessing model adequacy is developed. The utility of these methods is demonstrated with several analyses of visual acuity data, including a comparison analysis based on generalized estimating equation (GEE) methods.

KEY WORDS: Generalized estimating equation; Intensity matrix; Markov process; Maximum likelihood estimator; Time-dependent covariates; Transitional model.

[1]Michael R. Kosorok is Assistant Professor, Departments of Statistics and Biostatistics, University of Wisconsin, Madison, WI 53706. Wei-Hsiung Chao is a graduate student, Department of Statistics, University of Wisconsin, Madison, WI 53706. This research was supported in part by a University of Wisconsin Graduate School Award. The authors thank Douglas M. Bates, Richard J. Chappell, David L. DeMets, Thomas G. Kurtz, and Michael A. Newton for helpful discussions on the methods proposed here. They also thank Marian R. Fisher and Stephen J. Gange for both providing and discussing the Diabetic Retinopathy Study data. Finally, the authors thank the editors and referees of the *Journal of the American Statistical Association* for their helpful comments.

1. INTRODUCTION

Many longitudinal studies involve taking repeated measurements of an ordinal response and several covariates from a sample of individuals. Examples of ordinal responses include visual acuity (ranging from better than 20/20 to nearly blind), Apgar score (ranging from 0 to 10, depending on the health of the infant), or even a binary indicator of the presence or absence of pain (a two-category response). The search for meaningful and statistically valid models for analyzing this kind of longitudinal data has been fairly intense in the last several decades. The depth and diversity of this research is evident in several excellent surveys on the topic, including the work of Agresti (1989), Ashby et al. (1992), and Lindsey (1993).

Typically, there are two distinct approaches to modeling longitudinal data depending on whether scientific interest lies in the pattern of change in response over time or simply in the marginal dependence of a response on a set of covariates. The latter approach primarily focuses on marginal relationships between the covariates and the response while accounting for the correlation among the repeated observations for a given subject. Examples of this approach include the works of Wei and Johnson (1985), Liang and Zeger (1986), Stram, Wei, and Ware (1988), and Zeger, Liang, and Albert (1988). The former approach, however, focuses on predictive relationships between the current values of the response and past values of the covariates and response. In this setting, it is common to use Markov models for either discrete or continuous time. Examples of such models for discrete time include the logistic autoregressive regression model (Slud and Kedem 1994), the conditional log-linear model (Gilula and Haberman 1994), and the categorical time series regression model of Fahrmeir and Kaufmann (1987).

Several investigators have developed Markov models for continuous time. Kalbfleisch and Lawless (1985) described a general approach for applying Markov models to continuous time categorical data with time-independent covariates. This approach involves modeling the intensity matrix which generates the categorical response as a function of covariates with, for example, each intensity element as a log-linear model in the covariates. Anderson, Hansen, and Keiding (1991) proposed a nonhomogeneous Markov process for censored categorical longitudinal data in which proportional intensity regression is used to model each intensity element. The implicit assumption in this proportional intensity model is that the response is completely observed throughout the entire period of the study. Unfortunately, ordinal responses are frequently only assessed on an intermittent basis.

A common feature of both of these continuous time models is that each element of the intensity matrices is modeled as a separate function of the covariates. This can sometimes result in models which are complex and considerably more difficult to interpret than marginal models (Stram, Wei, and Ware 1988). This is particularly true in settings where the actual categories are somewhat arbitrary, as is the case with ordinal data, where transitions between specific categories are frequently not as interesting as the general direction and rate of movement.

In this paper, we present a continuous time Markov model for intermittently observed ordinal data which permits unified inference on the relationship between time-dependent covariates and the general direction and rate of movement between ordinal categories. This is accomplished by decomposing an ergodic intensity matrix into its equilibrium distribution and a parametrically independent component which assesses the transfer rate between categories. The formulation of this model—the "Local Equilibrium Distribution" (LED) model—is given in Section 2. In Section 3, we develop a robust maximum likelihood estimation approach for this model which produces consistent and asymptotically normal parameter estimates. In Section 4, we discuss a simple diagnostic tool for assessing model appropriateness. The utility of these methods is evaluated with several analyses of visual acuity data from patients with diabetic retinopathy—including a comparison analysis based on generalized estimating equation (GEE) methods—in Section 5. Finally, several summary and concluding remarks are given in section 6. This paper is essentially a correction and refinement of Kosorok and Chao (1994).

2. A LOCAL EQUILIBRIUM DISTRIBUTION (LED) MODEL

2.1 The Model Framework

Suppose individuals move independently among K states, denoted by $k = 1 \ldots K$. Let $\mathcal{Y}(\cdot)$ be a K-dimensional ordinal response process in continuous time—for a randomly chosen individual—intermittently observed at a sequence of time points, with a 1 in the appropriate ordinal category and zeros elsewhere. Suppose we also observe a right continuous covariate process $\mathbf{Z}(\cdot)$ at possibly different time points from the observation of $\mathcal{Y}(\cdot)$. We assume that $\mathbf{Z}(\cdot)$ is constant between the times we observe it. We will assume that the response process $\mathcal{Y}(\cdot)$ is a first order Markov process generated by a piecewise constant intensity matrix which is a function of the process $\mathbf{Z}(\cdot)$ and several parameters. The $K \times K$ intensity matrix at time t, $\mathbf{Q}(\boldsymbol{\theta}, t)$, has entries q_{ab} of the following form:

$$\begin{aligned} q_{ab} &= \pi_b \mu_{(a\wedge b)(a\vee b)},\ a \neq b, \\ q_{aa} &= -\sum_{b\neq a} q_{ab},\ a = 1 \ldots K, \end{aligned}$$

for some strictly positive $\boldsymbol{\pi} = \{\pi_1 \ldots \pi_K\}'$ such that $\pi_1 + \cdots + \pi_K = 1$ and some nonnegative $\boldsymbol{\mu} = \{\mu_{ab}, 1 \leq a < b \leq K\}'$ such that enough of the μ_{ab} values are strictly positive to ensure that $\mathbf{Q}(\boldsymbol{\theta}, t)$ is ergodic, and where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. The quantities $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ are modeled as distinct functions of $\mathbf{z} = \mathbf{Z}(t)$ and parameter vector $\boldsymbol{\theta}$.

It is not difficult to show that π_a is the equilibrium probability for category a, $a = 1 \ldots K$, since the equilibrium distribution for an ergodic first order Markov process, with intensity matrix $\mathbf{Q}^*$, is the unique solution of $\boldsymbol{\pi}'\mathbf{Q}^* = \mathbf{0}$ such that $\pi_1 + \cdots + \pi_K = 1$ (Kalbfleisch and Lawless 1985). The remaining quantities, μ_{ab}, $1 \leq a < b \leq K$, are nonnegative real values which reflect the rates of transfer between pairs of categories. It is not difficult to show that this parameterization fully specifies all time reversible ergodic intensity matrices. However, because $\mathbf{Q}(\boldsymbol{\theta}, \cdot)$ is allowed to change whenever the covariate process $\mathbf{Z}(\cdot)$ changes, both the time reversible property and the equilibrium distribution reparameterization are local phenomena: the actual resulting process $\mathcal{Y}(\cdot)$ is not necessarily time reversible and is not in general stationary.

For these reasons, we will refer to this model as the local equilibrium distribution (LED) model for longitudinal ordinal data in continuous time. A similar kind of reparameterization is discussed by Tuma, Hannan, and Groeneveld (1979) for stationary birth-death processes and by Kosorok (1993) for nonstationary birth-death processes which generate longitudinal ordinal data. As we will describe later, when $\mathbf{Q}(\boldsymbol{\theta}, t)$—for a given time t—is reparameterized into $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, the $\boldsymbol{\pi}$ term assesses the direction of movement (trend) while the $\boldsymbol{\mu}$ term assesses the rate of movement (diffusion)—which is related to the autocovariance—in what is essentially a nonstationary AR(1) process.

There exists a number of intuitive models for assessing the effect of covariates on a marginal ordinal response distribution (see Agresti 1989 for some examples). We can apply any one of these marginal models to describe the relationship between the covariate vector $\mathbf{z}$ and the local equilibrium distribution $\boldsymbol{\pi}$ at time t. One reasonable way of modeling the transfer rate vector $\boldsymbol{\mu}$ is to assume that

$$\mu_{ab} = \exp[\xi_{ab} + \boldsymbol{\gamma}'\mathbf{z}],$$

$1 \leq a < b \leq K$: thus transfer rates are permitted to vary from pair to pair yet the overall influence of the covariates on rate of movement can be assessed by the single parameter vector $\boldsymbol{\gamma}$.

The parsimony that can be achieved by this reparameterization, as an alternative to modeling each intensity element as a separate function of covariates, can be substantial. For example, the four category setting has 12 intensity elements q_{ab}, $a = 1 \ldots 4$ and $b = 1 \ldots 4$ such that $a \neq b$, which could be modeled individually as

$$q_{ab} = \exp[\alpha_{ab} + \boldsymbol{\beta}'_{ab}\mathbf{z}].$$

This requires fitting *and interpreting* 12 different covariate coefficient vectors and estimating 12 constant terms. In contrast, the LED model with a cumulative logit link for the local equilibrium distribution only requires fitting two covariate coefficient vectors (one for the trend and one for the diffusion) as well as estimating 3 cutpoint parameters (for the cumulative logit link) and 6 additional constant terms (for the 6 transfer rates). If there are 5 covariates, this is a reduction from 72 parameters $(12 \times 5 + 12)$ to 19 parameters $(2 \times 5 + 3 + 6)$.

Before completing the formulation of the LED model, we need to introduce some additional notation. Let $\mathcal{Y}_i(\cdot)$ be the ordinal response process and $\mathbf{Z}_i(\cdot)$ the corresponding covariate process for individual i, $i = 1 \ldots N$. Also for individual i, define

$t_{ij0} \equiv$ ordinal response observation times, $j = 0 \ldots n_i$,

$\mathbf{y}_{ij} \equiv \mathcal{Y}_i(t_{ij0})$,

$t_{ijl} \equiv$ covariate and response observation times, $l = 0 \ldots m_{ij}$, $j = 0 \ldots n_i$, and $m_{ij} \geq 1$ for $j < n_i$,

$\mathbf{z}_{ijl} \equiv \mathbf{Z}_i(t_{ijl})$, and

$\mathbf{Q}_{ijl}(\boldsymbol{\theta}) \equiv$ intensity matrix at time t_{ijl}, with local equilibrium distribution $\boldsymbol{\pi}^{ijl}$ and transfer rate vector $\boldsymbol{\mu}^{ijl}$, both of which are functions of $\mathbf{z}_{ijl}$ and $\boldsymbol{\theta}$ according to the LED model;

set $t_{ij(m_{ij})} \equiv t_{i(j+1)0}$ for $j = 0 \ldots n_i$ and $m_{i(n_i)} \equiv 0$; and assume $n_i \geq 1$. There is some redundancy here, in that $\mathbf{z}_{ij(m_{ij})} \equiv \mathbf{z}_{i(j+1)0}$, but this will simplify some of the notation used later.

Let $\mathbf{X}_i \equiv \{\mathbf{y}_{i0}, t_{ijl}, \mathbf{z}_{ijl}, l = 0 \ldots m_{ij}, j = 0 \ldots n_i\}$ and $\mathbf{Y}_i = \{\mathbf{y}_{ij}, j = 1 \ldots n_i\}$. Thus, if we condition on the initial ordinal response $\mathbf{y}_{i0}$, $\mathbf{X}_i$ is the observed design and covariate process while $\mathbf{Y}_i$ is the observed ordinal response process for individual i, $i = 1 \ldots N$. If we assume that $\mathcal{Y}_i(\cdot)$ is generated according to the LED model, then

$$\mathrm{E}[\mathbf{y}_{i(j+1)}|\mathbf{y}_{ij}, \mathbf{X}_i] \quad = \quad \left(\prod_{l=1}^{m_{ij}} \exp[\mathbf{Q}_{ij(l-1)}(\boldsymbol{\theta})\{t_{ijl} - t_{ij(l-1)}\}]\right)' \mathbf{y}_{ij} \tag{1}$$

completely specifies the distribution of $\mathbf{Y}_i$ conditional on $\mathbf{X}_i$ by the properties of first order Markov processes and multinomial distributions.

2.2 The LED Model as a Nonstationary AR(1) Process

As mentioned earlier, the LED model is essentially a nonstationary continuous time autoregressive process of order 1. This follows from the fact that, for any ergodic and aperiodic intensity matrix $\mathbf{Q}$ with equilibrium distribution $\boldsymbol{\pi}$, any nonnegative K-vector $\mathbf{y}$ which sums to 1, and any $d \geq 0$,

$$\mathbf{y}' \exp[\mathbf{Q}d] \quad = \quad \exp[-\lambda d]\mathbf{y}'\mathbf{B}(d) + (1 - \exp[-\lambda d])\boldsymbol{\pi}', \tag{2}$$

where $\lambda > 0$ and $\mathbf{B}(d)$ is a bounded matrix with all rows summing to 1 and such that $\mathbf{B}(0) = \mathbf{I}$. The proof of this result is given in the appendix.

If, for example, $\mathbf{Q}$ is parameterized according to the LED model and the elements of $\boldsymbol{\mu}$ are restricted to be $\mu_{ab} = \mu$, for $1 \leq a < b \leq K$, then the decomposition given by (2) becomes simply

$$\mathbf{y}' \exp[\mathbf{Q}d] = \exp[-\mu d]\mathbf{y}' + (1 - \exp[-\mu d])\boldsymbol{\pi}',$$

which clearly has a nonstationary AR(1) structure.

In general, the result given in (2) implies that the conditional expectation (1) can be reexpressed as an exponentially weighted average of local equilibrium distributions and $\mathbf{y}_{ij}$. To see this, first let λ_{ijl}, $\mathbf{B}_{ijl}(\cdot)$, and $\boldsymbol{\pi}^{ijl}$ be the terms corresponding to the decomposition given in (2) for the intensity matrix $\mathbf{Q}_{ijl}(\boldsymbol{\theta})$, and define $d_{ijl} \equiv t_{ij(l+1)} - t_{ijl}$. Now, the conditional expectation (1) can be reexpressed as

$$
\begin{aligned}
&\mathrm{E}[\mathbf{y}_{i(j+1)}|\mathbf{y}_{ij}, \mathbf{X}_i] \\
&= \exp\left[-\sum_{l=1}^{m_{ij}} \lambda_{ij(l-1)} d_{ij(l-1)}\right] \left(\prod_{l=1}^{m_{ij}} \mathbf{B}_{ij(l-1)}(d_{ij(l-1)})\right)' \mathbf{y}_{ij} \\
&\quad + \left(1 - \exp\left[-\lambda_{ij(m_{ij}-1)} d_{ij(m_{ij}-1)}\right]\right) \boldsymbol{\pi}^{ij(m_{ij}-1)} \\
&\quad + I_{\{m_{ij}>1\}} \sum_{l=1}^{m_{ij}-1} \left(1 - \exp\left[-\lambda_{ij(l-1)} d_{ij(l-1)}\right]\right) \\
&\quad \times \exp\left[-\sum_{r=l+1}^{m_{ij}} \lambda_{ij(r-1)} d_{ij(r-1)}\right] \left(\prod_{r=l+1}^{m_{ij}} \mathbf{B}_{ij(r-1)}(d_{ij(r-1)})\right)' \boldsymbol{\pi}^{ij(l-1)}, \qquad (3)
\end{aligned}
$$

where

$$
I_{\{C\}} = \begin{cases} 1, & \text{if } C \text{ is true} \\ 0, & \text{otherwise.} \end{cases}
$$

Thus, the LED model can also be interpreted as a fairly rich class of continuous time nonstationary AR(1) models, with the trend terms being the local equilibrium distributions and the autocovariance terms reflecting the transfer rates.

2.3 A Special Class of LED Models

In some settings, it may be appropriate to restrict the structure of the intensity matrix $\mathbf{Q}$ so that all transitions between states greater than some $d \geq 0$ categories apart share a common transfer rate, i.e., $\mu_{ab} = \mu$ for $b - a > d$. For a given value of d, we will refer to this restricted model as a "d-level local equilibrium distribution" (LED-d) model. If the value of μ is assumed to be zero, then the associated LED model will be abbreviated LED-$d(0)$.

For example, an LED-1(0) model is a simple birth-death process with boundaries 1 and K. This model would be appropriate when individuals are thought to move a distance of only one category at a time. It turns out that the class of LED-1(0) models fully specifies all ergodic, aperiodic Markov processes with boundaries 1 and K defined by a tridiagonal intensity matrix. This follows from the fact that all ergodic birth-death processes are time reversible (Ross 1993). Another example is the LED-0 model described in section 2.2 with $\mu_{ab} = \mu$, for $1 \leq a < b \leq K$.

The binary, or two category, birth-death case is the simplest of all LED-1 models. In fact, because there is only one transfer rate, the binary LED-1 model is also an LED-0 model. Furthermore, the corresponding 2-vector $\mathbf{y}$ can be replaced with a dichotomous variable y^*, where

$$y^* = \begin{cases} 0, & \text{for } \mathbf{y} = \{1,0\}', \\ 1, & \text{for } \mathbf{y} = \{0,1\}'. \end{cases}$$

With this substitution, the conditional expectation (3) becomes

$$\begin{aligned} &\mathrm{E}[y^*_{i(j+1)} | y^*_{ij}, \mathbf{X}_i] \\ &= \exp\left[-\sum_{l=1}^{m_{ij}} \mu^{ij(l-1)} d_{ij(l-1)}\right] y^*_{ij} + \left(1 - \exp\left[-\mu^{ij(m_{ij}-1)} d_{ij(m_{ij}-1)}\right]\right) \pi^*_{ij(m_{ij}-1)} \\ &\quad + I_{\{m_{ij}>1\}} \sum_{l=1}^{m_{ij}-1} \left(1 - \exp\left[-\mu^{ij(l-1)} d_{ij(l-1)}\right]\right) \exp\left[-\sum_{r=l+1}^{m_{ij}} \mu^{ij(r-1)} d_{ij(r-1)}\right] \pi^*_{ij(l-1)}, \end{aligned}$$

where π^*_{ijl} is the local equilibrium probability of being in category 2 which corresponds to the 2×2 intensity matrix $Q_{ijl}(\boldsymbol{\theta})$, and where μ^{ijl} is the associated transfer rate between categories 1 and 2.

3. ESTIMATION AND LARGE SAMPLE THEORY

3.1 The Estimation Procedure

Before outlining the maximum likelihood estimation procedure and corresponding large sample characteristics, we need to define several quantities. Let

$$\ell_{ij}(\boldsymbol{\theta}) = \log\left\{\mathbf{y}'_{ij}\left(\prod_{l=1}^{m_{ij}} \exp[\mathbf{Q}_{ij(l-1)}(\boldsymbol{\theta})\{t_{ijl} - t_{ij(l-1)}\}]\right)\mathbf{y}_{i(j+1)}\right\}$$

be the i'th individual's contribution to the log likelihood on the interval $(t_{ij0}, t_{i(j+1)0}]$ defined by the conditional expectation (1); and let the associated contribution to the score vector be $\mathbf{s}_{ij}(\boldsymbol{\theta})$ with elements

$$s^u_{ij}(\boldsymbol{\theta}) = \frac{\partial \ell_{ij}(\boldsymbol{\theta})}{\partial \theta_u},$$

$u = 1 \ldots M$, and where θ_u is the u'th element of the parameter M-vector $\boldsymbol{\theta}$.

Our procedure for estimating the model parameter vector $\boldsymbol{\theta}$ is to solve the score equation $\mathbf{S}_N(\hat{\boldsymbol{\theta}}_N) = N^{-1} \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \mathbf{s}_{ij}(\hat{\boldsymbol{\theta}}_N) = \mathbf{0}$. As shown by Kalbfleisch and Lawless (1985), the required first derivatives can be obtained directly from the canonical decomposition and first derivatives of $\mathbf{Q}_{ijl}(\boldsymbol{\theta})$. In fact, as will be shown in section 3.2, second and higher derivatives can also be obtained from the canonical decomposition and derivatives of $\mathbf{Q}_{ijl}(\boldsymbol{\theta})$. The existence and behavior of these derivatives will be utilized in section 3.3 to obtain asymptotic consistency and normality for $\hat{\boldsymbol{\theta}}_N$ under fairly reasonable regularity conditions, even when the LED based likelihood is not correct for the data at hand but is reasonably close.

This estimation procedure begins with determining an initial estimate of $\boldsymbol{\theta}$ by assuming that all covariate effects are zero; that the cutpoint parameters for the local equilibrium distributions are such that $\boldsymbol{\pi}^{ijl} = \{1/K, 1/K, \cdots, 1/K\}'$ for all i, j, and l; and that $\boldsymbol{\mu}^{ijl} = \{\mu, \mu, \cdots, \mu\}'$ for all i, j, and l, where $\mu > 0$ is very small. Under these assumptions, an approximate estimate of μ is simply

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \sum_{j=0}^{n_i-1} x_{ij}}{(K-1) \sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \delta_{ij}},$$

where $\delta_{ij} = t_{i(j+1)0} - t_{ij0}$ and $x_{ij} = I_{\{\mathbf{y}_{ij} \neq \mathbf{y}_{i(j+1)}\}}$. This estimate is simply an average transition rate per unit time and is approximately unbiased for μ, provided μ is small enough so that at most one change can occur in the interval $(t_{ij0}, t_{i(j+1)0}]$ for all i and j.

Beginning with this initial estimate $\hat{\boldsymbol{\theta}}_N^{(0)}$, a Fisher-scoring algorithm with updates of the form

$$\hat{\boldsymbol{\theta}}_N^{(n+1)} = \hat{\boldsymbol{\theta}}_N^{(n)} + \left(\sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \mathbf{s}_{ij}(\hat{\boldsymbol{\theta}}_N^{(n)}) \left\{ \mathbf{s}_{ij}(\hat{\boldsymbol{\theta}}_N^{(n)}) \right\}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{j=0}^{n_i-1} \mathbf{s}_{ij}(\hat{\boldsymbol{\theta}}_N^{(n)}) \right)$$

is used several times to get into the neighborhood of $\hat{\boldsymbol{\theta}}_N$, after which the exact second derivative is employed in several Newton-Raphson iterations until convergence.

3.2 The Computation of Derivatives

The greatest difficulty in computing derivatives of the log-likelihood components ℓ_{ij}, $j = 0 \ldots n_i$, $i = 1 \ldots N$, is computing the derivatives of $\exp[Q_{ij(l-1)}(\boldsymbol{\theta})\{t_{ijl} - t_{ij(l-1)}\}]$, $l = 1 \ldots m_{ij}$. Kalbfleisch and Lawless (1985) showed that for a general intensity matrix $\mathbf{Q}$ with canonical decomposition $\mathbf{ADA}^{-1}$ and a time change t,

$$\frac{\partial \exp[\mathbf{Q}t]}{\partial \theta_u} = \mathbf{A}\dot{\mathbf{V}}_u\mathbf{A}^{-1},$$

for $u = 1 \ldots M$, where $\dot{\mathbf{V}}_u$ is a $K \times K$ matrix with (a, b)'th element

$$(\dot{g}^u_{ab})\frac{\exp[d_a t] - \exp[d_b t]}{d_a - d_b} \text{ for } d_a \neq d_b, \text{ or } (\dot{g}^u_{ab}t)\exp[d_a t] \text{ for } d_a = d_b,$$

where d_a is the a'th diagonal element of $\mathbf{D}$ and where $\dot{g}^u_{ab}$ is the (a, b)'th element of $\mathbf{A}^{-1}(\partial\mathbf{Q}/\partial\theta_u)\mathbf{A}$.

As shown in the appendix, a similar result can be obtained for the second derivative:

$$\frac{\partial^2 \exp[\mathbf{Q}t]}{\partial \theta_u \partial \theta_v} = \mathbf{A}\ddot{\mathbf{V}}_{uv}\mathbf{A}^{-1} + \mathbf{A}\mathbf{H}_{uv}\mathbf{A}^{-1}, \tag{4}$$

for $u = 1 \ldots M$ and $v = 1 \ldots M$, where $\ddot{\mathbf{V}}_{uv}$ is a $K \times K$ matrix with (a, b)'th element

$$(\ddot{g}^{uv}_{ab})\frac{\exp[d_a t] - \exp[d_b t]}{d_a - d_b} \text{ for } d_a \neq d_b, \text{ or } (\ddot{g}^{uv}_{ab}t)\exp[d_a t] \text{ for } d_a = d_b,$$

where $\ddot{g}^{uv}_{ab}$ is the (a, b)'th element of

$$\mathbf{A}^{-1}\left(\frac{\partial^2\mathbf{Q}}{\partial\theta_u\partial\theta_v}\right)\mathbf{A};$$

and where $\mathbf{H}_{uv}$ is a $K \times K$ matrix with (a, b)'th element

$$\sum_{r=1}^{K}[\dot{g}^u_{ar}\dot{g}^v_{rb} + \dot{g}^v_{ar}\dot{g}^u_{rb}]\,\Gamma_{arb},$$

where

$$\Gamma_{arb} = \begin{cases} \frac{\exp(d_a t)-d_a t-1}{(d_a-d_b)(d_a-d_r)} + \frac{\exp(d_b t)-d_b t-1}{(d_b-d_a)(d_b-d_r)} + \frac{\exp(d_r t)-d_r t-1}{(d_r-d_a)(d_r-d_b)}, & \text{if } d_a \neq d_b,\ d_a \neq d_r,\ d_b \neq d_r, \\ \frac{t[\exp(d_a t)-1]}{d_a-d_r} + \frac{\exp(d_r t)-d_r t-\exp(d_a t)+d_a t}{(d_a-d_r)^2}, & \text{if } d_a = d_b \neq d_r, \\ \frac{t[\exp(d_a t)-1]}{d_a-d_b} + \frac{\exp(d_b t)-d_b t-\exp(d_a t)+d_a t}{(d_a-d_b)^2}, & \text{if } d_a = d_r \neq d_b, \\ \frac{t[\exp(d_b t)-1]}{d_b-d_a} + \frac{\exp(d_a t)-d_a t-\exp(d_b t)+d_b t}{(d_b-d_a)^2}, & \text{if } d_b = d_r \neq d_a, \\ \frac{t^2}{2}\exp(d_a t), & \text{if } d_a = d_b = d_r. \end{cases}$$

Although complex, this result is easily computed; furthermore, close inspection of the terms of these derivatives reveals that they are bounded whenever the eigenvalues, t, and the derivatives of $\mathbf{Q}$ are bounded. It can be shown that higher derivatives have similar structures and are similarly bounded.

An essential part of obtaining these derivatives involves obtaining the canonical decomposition of $\mathbf{Q}$. This canonical decomposition is also useful for computing model based predictions. If we assume that $\mathbf{Q}$ is parameterized according to the LED model with local equilibrium distribution $\boldsymbol{\pi}$, then certain simplifications in the canonical decomposition are possible. Let $\mathbf{c} = \{\pi_1^{1/2}, \pi_2^{1/2}, \cdots, \pi_K^{1/2}\}'$ and let $\mathbf{C}$ be a $K \times (K-1)$ matrix with (a,b)'th element w_a/w_{a-1} if $a = b$, $-(\pi_a \pi_b)^{1/2}/(w_b w_{b-1})$ if $a > b$, and 0 if $a < b$, where $w_0 = 1$ and, for $a \geq 1$, $w_a = (1 - \sum_{r=1}^{a} \pi_r)^{1/2}$. If we let $\mathbf{D}_c = \text{diag}(\mathbf{c})$, it is not difficult to see that $\mathbf{D}_c \mathbf{Q} \mathbf{D}_c^{-1}$ is symmetric and therefore has real eigenvalues and orthonormal eigenvectors. Further, since $\mathbf{Q}$ is time reversible, we know that one eigenvalue is 0 and the remaining eigenvalues are strictly negative.

By computing several inner products, it is evident that $\mathbf{c}$ is the eigenvector of $\mathbf{D}_c \mathbf{Q} \mathbf{D}_c^{-1}$ corresponding to the 0 eigenvalue and that the remaining eigenvectors are orthonormal linear combinations of the columns of $\mathbf{C}$, since $\mathbf{C}'\mathbf{c} = 0$ and $\mathbf{C}'\mathbf{C}$ is the identity matrix. Thus $\mathbf{M} = \mathbf{C}'\mathbf{D}_c \mathbf{Q} \mathbf{D}_c^{-1} \mathbf{C}$ is a symmetric matrix of order $K-1$ with eigenvalues equal to the nonzero eigenvalues of $\mathbf{Q}$. Furthermore, it can be shown that the structure of $\mathbf{C}$ ensures that if $\mathbf{Q}$ arises from an LED-d model then $\mathbf{M}$ is $\{(2d+1) \wedge K\}$-diagonal. This means that computationally efficient algorithms for symmetric banded matrices can be fully utilized (see Golub and Van Loan, 1989). Now, if we let $\mathbf{M} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}'$ be the canonical decomposition of $\mathbf{M}$, then the full canonical decomposition of $\mathbf{Q}$ can be written $\mathbf{Q} = \mathbf{W}\mathbf{F}\mathbf{W}^{-1}$, where

$$\mathbf{W} = \mathbf{D}_c^{-1}[\mathbf{c} \vdots \mathbf{C}\mathbf{R}],\ \mathbf{F} = \begin{bmatrix} 0 & \vdots & 0 \\ \ldots & \vdots & \ldots \\ 0 & \vdots & \Lambda \end{bmatrix},\ \text{and } \mathbf{W}^{-1} = \begin{bmatrix} \mathbf{c}' \\ \cdots\cdots \\ \mathbf{R}'\mathbf{C}' \end{bmatrix} \mathbf{D}_c.$$

This device allows us to take advantage of our knowledge of the zero eigenvalue in a computationally efficient manner.

3.3 Large Sample Results

In this section, we present robust large sample results which take advantage of the existence of the derivatives described in the previous section. We will begin with a theorem and corollary giving the main results followed by a brief discussion of how these regularity conditions should be satisfied in practical data analysis settings and under what circumstances the resulting model estimates are robust:

Theorem 1 *Let $\boldsymbol{\Theta}$ be an open convex set and suppose the following conditions hold:*

1. *The random pairs $\{\mathbf{X}_i, \mathbf{Y}_i\}$, $i = 1 \ldots N$, are independent and identically distributed.*
2. *For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\ell_{i\cdot}(\boldsymbol{\theta}) = \sum_{j=0}^{n_i-1} \ell_{ij}(\boldsymbol{\theta})$ is three times differentiable with respect to $\boldsymbol{\theta}$.*
3. *For every compact set $T \subset \boldsymbol{\Theta}$, there exists a function $g(\cdot,\cdot)$ such that $E[g(\mathbf{X}_i, \mathbf{Y}_i)] < \infty$ and, for every $\boldsymbol{\theta} \in T$,*
   - *(a)* $|\ell_{i\cdot}(\boldsymbol{\theta})| \le g(\mathbf{X}_i, \mathbf{Y}_i)$,
   - *(b)* $\left|\frac{\partial s_{i\cdot}^u(\boldsymbol{\theta})}{\partial \theta_v}\right| \le g(\mathbf{X}_i, \mathbf{Y}_i)$,
   - *(c)* $\left|\frac{\partial^2 s_{i\cdot}^u(\boldsymbol{\theta})}{\partial \theta_w \partial \theta_v}\right| \le g(\mathbf{X}_i, \mathbf{Y}_i)$,
   - *(d)* $|s_{i\cdot}^u(\boldsymbol{\theta}) s_{i\cdot}^v(\boldsymbol{\theta})| \le g(\mathbf{X}_i, \mathbf{Y}_i)$, *and*

(e) $\left|\frac{\partial s_{i\cdot}^u(\boldsymbol{\theta})}{\partial\theta_v} s_{i\cdot}^w(\boldsymbol{\theta})\right| \leq g(\mathbf{X}_i, \mathbf{Y}_i)$,

where $s_{i\cdot}^u(\boldsymbol{\theta}) = \sum_{j=0}^{n_i-1} s_{ij}^u(\boldsymbol{\theta})$, for $u = 1 \ldots M$, $v = 1 \ldots M$, and $w = 1 \ldots M$.

4. $\mathbf{S}(\boldsymbol{\theta}) = E[\mathbf{s}_{i\cdot}(\boldsymbol{\theta})]$ has a unique zero in $\boldsymbol{\Theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

5. There exists an $N^* < \infty$ such that for every $N \geq N^*$,

$$\mathbf{S}_N(\boldsymbol{\theta}) = N^{-1}\sum_{i=1}^{N} \mathbf{s}_{i\cdot}(\boldsymbol{\theta})$$

has a zero in $\boldsymbol{\Theta}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_N$ almost surely.

6. For every $\epsilon > 0$, there exists a convex compact set $T^* \subset \boldsymbol{\Theta}$ and an $N^{**} < \infty$ such that $\boldsymbol{\theta}_0 \in T^*$ and, for every $N \geq N^{**}$, $\Pr\{\hat{\boldsymbol{\theta}}_N \in T^*\} > 1 - \epsilon$.

7. There exists a neighborhood of $\boldsymbol{\theta}_0$, $\mathcal{B}(\boldsymbol{\theta}_0)$, such that the matrix $\mathbf{I}(\boldsymbol{\theta})$, with (u,v)'th element

$$I^{uv}(\boldsymbol{\theta}) = E[-\frac{\partial s_{i\cdot}^u(\boldsymbol{\theta})}{\partial\theta_v}],$$

is nonsingular for all $\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}_0)$.

Then the following are true:

1. $\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_0$ in probability, as $N \to \infty$.

2. $\left[\mathbf{I}_N(\hat{\boldsymbol{\theta}}_N)\right]^{-1} \mathbf{V}_N(\hat{\boldsymbol{\theta}}_N) \left[\mathbf{I}_N(\hat{\boldsymbol{\theta}}_N)\right]^{-1} \to [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\, \mathbf{V}(\boldsymbol{\theta}_0)\, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ in probability, as $N \to \infty$, where $\mathbf{I}_N(\boldsymbol{\theta})$, $\mathbf{V}_N(\boldsymbol{\theta})$, and $\mathbf{V}(\boldsymbol{\theta})$ are matrices with (u,v)'th elements

$$\begin{aligned}
I_N^{uv}(\boldsymbol{\theta}) &= -N^{-1}\sum_{i=1}^{N} \frac{\partial s_{i\cdot}^u(\boldsymbol{\theta})}{\partial\theta_v}, \\
V_N^{uv}(\boldsymbol{\theta}) &= N^{-1}\sum_{i=1}^{N} s_{i\cdot}^u(\boldsymbol{\theta}) s_{i\cdot}^v(\boldsymbol{\theta}), \text{ and} \\
V^{uv}(\boldsymbol{\theta}) &= E[s_{i\cdot}^u(\boldsymbol{\theta}) s_{i\cdot}^v(\boldsymbol{\theta})],
\end{aligned}$$

respectively, for $u = 1 \ldots M$ and $v = 1 \ldots M$.

3. $N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \to \mathbf{Z}$ *in distribution, as* $N \to \infty$*, where*

$$\mathbf{Z} \;\sim\; N\Big(\mathbf{0}, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\,\mathbf{V}(\boldsymbol{\theta}_0)\,[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\Big).$$

Corollary 1 *If, in addition to the conditions specified in theorem 1, we have that* $\ell_{i\cdot}(\boldsymbol{\theta}_0)$ *is the true log conditional likelihood of* $\mathbf{Y}_i$ *given* $\mathbf{X}_i$*, for* $i = 1 \ldots N$*, then the following are true:*

1. $\hat{\boldsymbol{\theta}}_N \to \boldsymbol{\theta}_0$ *in probability as* $N \to \infty$*.*

2. $[\mathbf{I}_N(\hat{\boldsymbol{\theta}}_N)]^{-1} \to [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ *in probability as* $N \to \infty$*.*

3. $\mathbf{V}_N(\hat{\boldsymbol{\theta}}_N) \to \mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0)$ *in probability as* $N \to \infty$*.*

4. $N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \to \mathbf{Z}^*$ *in distribution, as* $N \to \infty$*, where* $\mathbf{Z}^* \sim N\Big(\mathbf{0}, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\Big)$*.*

Condition 6 serves a purpose similar to Lemma 2 of Huber (1967), but Huber's approach otherwise differs substantially from ours.

Under a number of practical scenarios, the conditions of theorem 1 should be satisfied. For example, if we are willing to assume that for each intensity matrix $\mathbf{Q}_{ijl}$ the cumulative logit of $\boldsymbol{\pi}^{ijl}$ is $-\boldsymbol{\beta}'\mathbf{z}_{ijl}$ (with the necessary cutpoint parameters); each (a,b)'th element of $\boldsymbol{\mu}^{ijl}$, μ^{ijl}_{ab}, is $\exp[\xi_{ab} + \boldsymbol{\gamma}'\mathbf{z}_{ijl}]$; the moment generating function of the time-dependent covariates $\mathbf{z}_{ijl}$ exists; and $\sum_{j=0}^{n_i-1} m_{ij}$ is bounded, then for any compact $T \subset \boldsymbol{\Theta}$, $\mathbf{Q}_{ijl}$ and the first three derivatives of $\mathbf{Q}_{ijl}$ are bounded in expectation on T. From the results of section 3.2, this means that $\exp[\mathbf{Q}_{ijl}t]$ and its first three derivatives are also bounded, satisfying conditions 2 and 3. If in addition we are willing to assume that individuals are independent and identically distributed and that $\mathbf{S}_N(\boldsymbol{\theta})$ is sufficiently well behaved so that

- its expectation has a unique zero in the interior of $\boldsymbol{\Theta}$ at some finite $\boldsymbol{\theta}_0$,

- it has a zero $\hat{\boldsymbol{\theta}}_N$ in the interior of $\boldsymbol{\Theta}$ which does not drift off to infinity as $N \to \infty$, and

- its derivative is full rank when $\boldsymbol{\theta}$ is sufficiently close to $\boldsymbol{\theta}_0$,

then conditions 1, 4, 5, 6, and 7 are also satisfied and the results of theorem 1 will follow.

In addition, theorem 1 does not require that the working LED likelihood be correct for the data at hand. However, the actual likelihood for the data should be close enough to the working likelihood so that the resulting estimators are consistent for meaningful parameters. Whether or not the local equilibrium model is "close enough" for a given setting is a complex question. As pointed out earlier, the LED model generates a nonstationary AR(1) process. If the structure of the trend and diffusion terms for the resulting nonstationary AR(1) process seem appropriate for a given analysis setting, then the estimation procedure given here should provide estimates of meaningful parameters whether or not the LED model is precisely correct.

4. A SIMPLE DIAGNOSTIC FOR MODEL APPROPRIATENESS

A simple diagnostic can be generated from the conditional expectation based on the LED model. First let $\mathbf{p}_{i(j+1)}(\boldsymbol{\theta})$ be the right hand side of (1), then if the LED model is correct, $\mathrm{E}[\mathbf{y}_{ij}|\mathbf{y}_{i(j-1)}, \mathbf{X}_i] = \mathbf{p}_{ij}(\boldsymbol{\theta}_0)$, and $\mathrm{var}[\mathbf{y}_{ij}|\mathbf{y}_{i(j-1)}, \mathbf{X}_i] = \mathrm{diag}\left\{\mathbf{p}_{ij}(\boldsymbol{\theta}_0)\right\} - \mathbf{p}_{ij}(\boldsymbol{\theta}_0)\left\{\mathbf{p}_{ij}(\boldsymbol{\theta}_0)\right\}'$. If we now define $\mathbf{v} = \{1, 2, \ldots, K\}'$, then $\mathbf{v}'\mathbf{y}_{ij}$ is the ordinal category which individual i is in at time t_{ij0}. Furthermore, the quantity

$$r_{ij}(\boldsymbol{\theta}) = \frac{\mathbf{v}'\left\{\mathbf{y}_{ij} - \mathbf{p}_{ij}(\boldsymbol{\theta})\right\}}{\sigma_{ij}(\boldsymbol{\theta})},$$

where

$$\sigma_{ij}^2(\boldsymbol{\theta}) = \mathbf{v}'\left[\mathrm{diag}\left\{\mathbf{p}_{ij}(\boldsymbol{\theta})\right\}\right]\mathbf{v} - \left\{\mathbf{v}'\mathbf{p}_{ij}(\boldsymbol{\theta})\right\}^2,$$

has mean 0, variance 1, and is uncorrelated with each $r_{im}(\boldsymbol{\theta})$, $m \neq j$, when the LED model is correct and $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The same properties should be approximately true when $\boldsymbol{\theta}$ is replaced with $\hat{\boldsymbol{\theta}}_N$. We will hereafter refer to $r_{ij}(\hat{\boldsymbol{\theta}}_N)$ as the "summary residual." Scatterplots of this diagnostic against covariates can be used to assess a number of assumptions, including the correctness of the functional form of the covariates.

5. SEVERAL ANALYSES OF VISUAL ACUITY DATA

5.1 The Visual Acuity Data

The Diabetic Retinopathy Study (DRS) was a randomized controlled clinical trial initiated in 1972 to evaluate the effectiveness of photocoagulation treatment in preventing severe visual loss in patients with diabetic retinopathy. Eligible patients were randomized so that one eye received treatment and one did not. The patients were scheduled to return every 4 months to assess visual acuity and take several relevant measurements. On the basis of statistically significant evidence that photocoagulation therapy was effective in reducing the risk of severe visual loss in patients with advanced diabetic retinopathy, the protocol changed in 1976 to permit the treatment of initially untreated eyes which had "high risk of severe visual loss". For this initial assessment, severe visual loss was defined as having visual acuity worse than 5/200 for at least two consecutively completed four-month follow-up visits. Follow-up continued into 1979 (Diabetic Retinopathy Study Research Group 1976 and 1981; Rand et al. 1985).

In the analyses presented here, we will use data only from the right eye. Visits where either the covariates or response values were unavailable will be considered missing and the covariate value at the missing value will be assumed to be unchanged from the previous observed value (missing was actually extremely rare). This is not unreasonable if we consider the time-dependent covariates to be as-observed values. Since covariates and responses were observed at the same time, $m_{ij} = 1$ for all i and j. For this reason, the l subscript will not be needed and will be dropped from the analyses presented here. There are more than 1600 subjects included in this data set with an average of over 10 follow-up visits per subject. The analyses presented here were implemented in Fortran 77 on a Sun SPARCstation 20.

The covariates for this analysis consisted of four baseline covariates and six time-dependent covariates in addition to the intercept term. The baseline covariates (abbreviations given in parentheses) were age at enrollment in years, rounded to the nearest integer (age); patient gender, 0 if male, 1 if female (sex); baseline retinal group, taking on the value 1 if at least some new blood vessels were present on the optic disc of the right eye at enrollment, 0 otherwise (BRG); and treatment assignment, 1 if the right eye initially received treatment, 0 if not (Treatment). The six time-dependent covariates (abbreviations again given in parentheses) were time in the study since enrollment (time); protocol change status, taking on the value 1 for the period beginning with the visit in which the protocol changed, 0 otherwise (change); time by change interaction (t x c); treatment by time interaction (T x t); treatment by change interaction (T x c); and treatment by time by change interaction (T x t x c).

5.2 A Binary LED-1 Analysis

For the binary analyses presented here, we used the response

$$y_{ij}^* = \begin{cases} 1 \text{ if visual acuity at time } t_{ij} \text{ is } > 5/200, \\ 0 \text{ otherwise,} \end{cases}$$

for $j = 0 \ldots n_i$, $i = 1 \ldots N$, since this was the definition of severe visual loss as given in the original protocol. The logit link was used for the local equilibrium probabilities π_{ij}^*, ie., $\pi_{ij}^* = \exp(\alpha + \boldsymbol{\beta}'\mathbf{z}_{ij})/[1 + \exp(\alpha + \boldsymbol{\beta}'\mathbf{z}_{ij})]$, while the transfer rates μ^{ij} were assumed to be all equal, $\mu^{ij} = \mu = \exp[\xi]$ (and therefore not dependent on covariates). We also need to define $\phi \equiv \exp[-\mu]$.

Since the chi-squared full-versus-reduced test for the presence of the three-way interaction was profoundly nonsignificant (the p-value was $> .9$) it was dropped from the model. However, the full versus reduced test for the two-way interactions, after removing the three-way interaction, was highly significant (the p-value was .0002): all two-way interactions were therefore left in the model, along with the main effects. Parameter estimates and 95% confidence intervals based on the robust covariance estimator for this final model are presented in table 1. The μ value of .4572 indicates that the correlation between two visits four months apart—which is $\approx \phi^{\frac{1}{3}}$—is about .86. The statistical significance and direction of the BRG and time terms is in general agreement with results obtained in the previously mentioned Diabetic Retinopathy Study reports. The effects of treatment and protocol change are a little more difficult to assess.

Figure 1 shows the change in risk score $\hat{\boldsymbol{\beta}}'\mathbf{z}_{ij}$, over time, based on LED-1 model predictions (labeled "LED-1") for a subject with an initially treated right eye versus another subject with the same baseline characteristics but with an initially untreated right eye. Both subjects are assumed to have experienced a change in protocol two years after entering the study and then to have remained in the study for two years following the protocol change. The higher the value of this change in risk score, the greater the model-predicted "local" equilibrium probability of severe visual loss. This plot seems to indicate that both the treatment and the protocol change eventually have an overall beneficial effect, after an initially negative impact; although table 1 indicates that this net beneficial effect of treatment over time, as reflected in the treatment by time interaction (T x t), is not quite statistically significant. This is in qualitative agreement with results of a report from the Diabetic Retinopathy Study Research Group (1981) which concluded that, although beneficial, the treatment did cause initial eye damage.

Figure 2 shows the summary residuals against time in study, age at enrollment, and calendar time (in years). A horizontal straight line is drawn at zero for reference. In all cases, the residuals appear to not appreciably deviate from a zero expectation. This was examined more closely by computing a running mean smoothed version of the summary residuals, using a window width of 1001 observations (about 4% of the data). The smoothed mean was so close to the zero horizontal reference line that it was visually indistinguishable on these small plots and was omited. These plots seem to indicate that the model adequately captures the effects of time in the study, age at enrollment, and calendar time on the response. Similar plots could be used to explore the adequacy of the model for other covariates.

5.3 Several Binary GEE Analyses for Comparison

Since the parameters in the local equilibrium probability π^*_{ij} have an interpretation similar to the parameters in a standard (marginal) logistic regression, we will consider several other marginal models for comparison. As pointed out earlier, the LED model is essentially a nonstationary AR(1) process. This leads us to consider marginal regression models which have an AR(1) correlation structure for comparison, either regressing y^*_{ij} on the current covariate value $\mathbf{z}_{ij}$ (a strictly marginal model) or on the preceding covariate value $\mathbf{z}_{i(j-1)}$ (a "predictive" marginal model). We will include the same covariates in these marginal models as those included in table 1 for the LED-1 model.

Since our Markov process might not be stationary, and our primary interest of comparison lies in the parameters modeling the mean, we need to carry out the marginal analysis with an estimation procedure which produces consistent estimates and is robust to misspecification of the correlation structure. One such procedure is Liang & Zeger's (1986) generalized estimating equation (GEE) approach which models the marginal mean as a function of covariates, through a link function, and assumes a functional relationship between the marginal mean and marginal variance. That is, if $\eta_{ij} = \mathrm{E}[y^*_{ij}|\mathbf{z}_{ij}]$, then $h(\eta_{ij}) = \boldsymbol{\beta}'\mathbf{z}_{ij}$, and $\mathrm{var}[y^*_{ij}|\mathbf{z}_{ij}] \propto g(\eta_{ij})$, where h is the link function and g the variance function. By treating the correlation structure as a working entity, the GEE method produces a consistent estimate of $\boldsymbol{\beta}$, whether or not the correlation structure is correct. These marginal analyses were implemented in Fortran 77 for nonequally spaced continuous time data with a logit link and an AR(1) working correlation structure. To see how the result depends on the working correlation structure, we also fit a marginal model using the independence working correlation. The four models for our comparison are as follows, where "marginal model" refers to regressing on the current covariate value, and "marginal predictive model" refers to regressing on the preceding covariate value (abbreviations given in parentheses):

- marginal model with independence working correlation (M.I),
- marginal predictive model with independence working correlation (MP.I),
- marginal model with AR(1) working correlation (M.AR(1)), and
- marginal predictive model with AR(1) working correlation (MP.AR(1)).

This ordering of the four marginal models approximately reflects increasing degrees of agreement with the LED-1 model.

Tables 2 and 3 present the parameter estimates and their 95% confidence interval based on robust covariance estimates for the above four marginal models. There are no confidence intervals for the autocorrelation parameter ϕ since the GEE approach uses an estimating procedure which treats ϕ as nuisance parameter. Note that $\hat{\phi}^{\frac{1}{3}}$ is .8154 and .8108 for the M.AR(1) and MP.AR(1) models, respectively, which indicates a high correlation between responses four months apart and is fairly close to the corresponding $\hat{\phi}^{\frac{1}{3}} = .8586$ for the LED-1 model.

Comparing these results to the LED-1 results given in table 1, we find that the direction of effects is the same in all of the LED-1 and marginal models, except for the main effect of time and the interaction term of treatment and protocol change (T x c). The statistical significance of the main effects BRG and change as well as of the two-way interactions also agrees with the result from the LED-1 model. On the other hand, the time effect is significant in the marginal models but not significant in the LED-1 model.

The change in risk score over time for initially treated versus untreated subjects for the four marginal models is given in figure 1 with labels for reference. Both the treated and untreated subjects are presumed to have experienced the protocol change two years after entering the study. As with the LED-1 model, the higher the value of the risk score, the greater the model-predicted marginal probability of severe visual loss. This figure seems to indicate that both the treatment and the protocol change have a detrimental effect, though the detrimental effect is weakened after the protocol change. Qualitatively, the marginal models seem to indicate that the treatment has no long term or short term beneficial effect. This result is inconsistent with both the LED-1 results and the previously mentioned report from the Diabetic Retinopathy Study Research Group (1981) which concluded that the treatment is eventually beneficial. Figure 1 also demonstrates the interesting trend that the change in risk score for the marginal models is heading towards that for the LED-1 model in the order M.I, MP.I, M.AR(1), and MP.AR(1). It seems that none of the marginal models adequately assess the long term treatment effect.

Figure 3 shows standardized residuals (observed minus model-predicted response values adjusted to approximately have mean zero and variance one) against time in study for the marginal models. A running mean smoothed version of the standardized residuals, using the same window width as used in the LED-1 model, was examined but not presented here because the smoothed version was too close to the horizontal line at zero to be visually distinct. Although these plots show a trend in the skewness of the residuals, the running mean smooth did not differ substantially from zero; and it is not clear whether this skewness reflects any inadequacies in the marginal models.

Figure 4 shows the standardized residuals against age at enrollment for marginal models. A running mean smoothed version is again not presented here because the smoothed version was too close to the horizontal line at zero to be visible. All of the four models show the same residual pattern of symmetry across age at enrollment, indicating that marginal models adequately capture the effect of age at enrollment on the binary response.

Figure 5 shows the standardized residuals against calendar time for the marginal models. The running mean smoothed versions are not presented for the same reasons as before. All of the marginal models present the same residual pattern of asymmetry across calendar time, and there is a sharp change in the residual pattern at the year of protocol change. This is in contrast to the relatively homogeneous behavior of the LED-1 residuals over calendar time seen in figure 2. Note that calendar time is not included as a covariate in the model. Clearly, these marginal models assess properties of the data which are distinctly different from the effects of covariates on future response as assessed by the LED-1 model.

5.4 A 4-Category LED-1 Analysis

The 4-category response $\mathbf{y}_{ij}$ was determined by visual acuity at visit time t_{ij} as follows:

$\mathbf{y}_{ij}$	visual acuity
$\{1,0,0,0\}'$	20/15 and 20/20 (normal vision)
$\{0,1,0,0\}'$	20/30 through 20/70 (mildly impaired)
$\{0,0,1,0\}'$	20/100 through 5/200 (moderately impaired)
$\{0,0,0,1\}'$	worse than 5/200 (severely impaired)

For comparison with the previous analyses, we considered only the covariates that were included in the final model of the binary LED-1 analysis.

We used a cumulative logit link for the local equilibrium distribution and assumed the transfer rates were constants which did not depend on the covariates. That is,

$$\begin{aligned}
\pi_1^{ij} &= \frac{\exp\left[\alpha_1 - \boldsymbol{\beta}'\mathbf{z}_{ij}\right]}{1+\exp\left[\alpha_1 - \boldsymbol{\beta}'\mathbf{z}_{ij}\right]}, \\
\sum_{r=1}^{k} \pi_r^{ij} &= \frac{\exp\left[\left(\alpha_1 + \sum_{r=2}^{k}\exp[\alpha_r]\right) - \boldsymbol{\beta}'\mathbf{z}_{ij}\right]}{1+\exp\left[\left(\alpha_1 + \sum_{r=2}^{k}\exp[\alpha_r]\right) - \boldsymbol{\beta}'\mathbf{z}_{ij}\right]}, \text{ for } k = 2,3, \\
\pi_4^{ij} &= 1 - \sum_{k=1}^{3} \pi_k^{ij}, \\
\mu_{a(a+1)}^{ij} &= \mu_{a(a+1)} = \exp[\xi_{a(a+1)}], \text{ for } a = 1\ldots 3, \text{ and} \\
\mu^{ij} &= \mu = \exp[\xi], \text{ for all remaining transfer rates.}
\end{aligned}$$

Table 4 presents the parameter estimates and their 95% robust confidence intervals. The statistical significance and direction of the BRG and time terms is in general agreement with results obtained by the Diabetic Retinopathy Study Research Group (1981) and the binary LED-1 analysis, except that the time effect is not significant in the binary LED-1 analysis. The effects of treatment and protocol change are more complex.

Figure 6, upper left plot, presents the change in risk score over time for initially treated versus untreated subjects for the 4-category analysis. As before, both the treated and untreated subjects are presumed to have experienced the protocol change two years after entering the study. The higher the value of the risk score, the greater the model-predicted trend towards more severe visual loss. This figure seems to indicate that both the treatment and the protocol change eventually have an overall beneficial effect, after an initially negative impact. This result is consistent with both the binary LED-1 analysis and the previously described report from the Diabetic Retinopathy Study Research Group (1981).

Figure 6 also shows the standardized summary residuals against time in study, age at enrollment, and calendar time (in years) for the 4-category LED-1 model. A running mean smoothed version of these residuals, using the same window width as used in the binary analysis, was examined but not presented here for the same reasons given earlier (being visually indistinguishable from zero). These plots seem to indicate that the LED-1 model captures the effects of time in the study, age at enrollment, and calendar time on the 4-category response. It appears as though the local equilibrium distribution models are qualitatively able to capture the long term treatment effect.

6. SUMMARY AND CONCLUDING REMARKS

The essential feature of the local equilibrium distribution model is the way it decomposes a Markov process into two intuitive and parametrically independent components: one corresponding to the trend and the other to the rate of transfer between categories. The parameters in the trend component can be interpreted in much the same way that parameters in a marginal regression model can be; however, the LED model trend assesses how covariates effect the transitional behavior of the response. The transfer rate component is simply a measure of dependence between adjacent observations, closely related to the autocovariance. The difference between the local equilibrium distribution model and marginal models in terms of statistical inference has been demonstrated in this paper to be potentially large and important, and decisions as to which model to use should be made carefully.

A time reversible reparameterization of the intensity matrix was chosen because of its relative simplicity. A similar reparameterization of all ergodic and aperiodic matrices with a given equilibrium distribution is possible, but it is astronomically more complex to work with because it requires the use of a number of difficult linear constraints. However, it should be pointed out that the class of time reversible intensity matrices with a given equilibrium distribution $\boldsymbol{\pi}$ is parametrically rich, in the sense that time reversible intensity matrices utilize $K(K-1)/2$ of the $(K-1)^2$ degrees of freedom—over one half—available for all ergodic and aperiodic intensity matrices with equilibrium distribution $\boldsymbol{\pi}$. In many settings this ratio is much larger than one half. For example, this ratio is one for birth-death processes since, as pointed out earlier, all birth-death intensity matrices are time reversible. Furthermore, the LED models easily give rise to a rich class of non-time reversible models through incorporating time-dependent covariates. This means that natural biological processes, such as those related to aging, can be approximated by LED models through appropriate use of frailty measures such as age.

Clearly, the potential of this model has not been fully explored here. More work is needed to explore other useful parameterizations of the trend and transfer rate components; the use of this model for prediction; assessing the effect of absorbing phenomena—such as death—which may be related to the ordinal response; as well as other issues. However, the LED model in its current form offers potentially useful inference on the predictive relationship between time-dependent covariates and ordinal outcomes in continuous time longitudinal studies.

APPENDIX: PROOFS

A.1 Derivations of (2) and (4)

Derivation of (2). If $\mathbf{Q}$ is ergodic and aperiodic with equilibrium probability $\boldsymbol{\pi}$, then, for any $d > 0$, $\exp[\mathbf{Q}d]$ has one eigenvalue equal to 1 while the remaining eigenvalues are strictly between -1 and 1. Thus the second largest real component of the eigenvalues of $\mathbf{Q}$, $-\lambda$, is strictly negative. Now, let

$$\mathbf{B}(d) = \exp[\lambda d]\exp[\mathbf{Q}d] - (\exp[\lambda d] - 1)\,\mathbf{j}\boldsymbol{\pi}',$$

where $\mathbf{j} = \{1, 1, \ldots, 1\}'$ is a K-vector. Simple multiplication verifies (2). Because $\mathbf{j}$ is the right eigenvector of $\mathbf{Q}$ corresponding to the eigenvalue 0,

$$\mathbf{B}(d)\mathbf{j} = \exp[\lambda d]\mathbf{j} - (\exp[\lambda d] - 1)\,\mathbf{j} = \mathbf{j}$$

and the rows of $\mathbf{B}(d)$ add up to 1. It is easy to see that $\mathbf{B}(0)$ is the identity matrix. Furthermore, the structure of $\mathbf{B}(d)$ ensures that its eigenvalues range from 1 to just larger than -1, and this implies that $\mathbf{B}(d)$ is bounded.$\square$

Derivation of (4). For time change $t \geq 0$,

$$\frac{\partial^2 \exp[\mathbf{Q}t]}{\partial\theta_u \partial\theta_v} = \sum_{s=1}^{\infty} \frac{\partial^2}{\partial\theta_u \partial\theta_v} \frac{\mathbf{Q}^s t^s}{s!}$$

$$= \sum_{s=2}^{\infty}\sum_{q=0}^{s-2}\sum_{m=q+1}^{s-1} \left(\mathbf{Q}^q \frac{\partial \mathbf{Q}}{\partial\theta_u}\mathbf{Q}^{m-q-1}\frac{\partial \mathbf{Q}}{\partial\theta_v}\mathbf{Q}^{s-m-1}\frac{t^s}{s!} + \mathbf{Q}^q \frac{\partial \mathbf{Q}}{\partial\theta_v}\mathbf{Q}^{m-q-1}\frac{\partial \mathbf{Q}}{\partial\theta_u}\mathbf{Q}^{s-m-1}\frac{t^s}{s!}\right) \quad (5)$$

$$+\sum_{s=1}^{\infty}\sum_{r=0}^{s-1} \mathbf{Q}^r \frac{\partial^2 \mathbf{Q}}{\partial\theta_u \partial\theta_v}\mathbf{Q}^{s-r-1}\frac{t^s}{s!}. \quad (6)$$

Now, let the canonical decomposition of $\mathbf{Q}$ be $\mathbf{ADA}^{-1}$, where $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_K\}'$; and define $\dot{\mathbf{G}}_u \equiv \mathbf{A}^{-1}(\partial\mathbf{Q}/\partial\theta_u)\mathbf{A}$ and $\ddot{\mathbf{G}}_{uv} \equiv \mathbf{A}^{-1}(\partial^2\mathbf{Q}/\partial\theta_u\partial\theta_v)\mathbf{A}$, with $\dot{g}^u_{ab}$ as the (a, b)'th element of $\dot{\mathbf{G}}_u$ and $\ddot{g}_{ab}^{uv}$ as the (a,b)'th element of $\ddot{\mathbf{G}}_{uv}$. Then, similar to what was done by Kalbfleisch and Lawless (1985), term (6) can be written as $\mathbf{A}\ddot{\mathbf{V}}_{uv}\mathbf{A}^{-1}$, where the (a,b)'th element of $\ddot{\mathbf{V}}_{uv}$ is

$$(\ddot{g}_{ab}^{uv})\frac{\exp[d_a t]-\exp[d_b t]}{d_a-d_b} \text{ for } d_a \neq d_b \text{ and } (\ddot{g}_{ab}^{uv}t)\exp[d_a t] \text{ for } d_a = d_b.$$

In addition, term (5) can be written as $\mathbf{A}\mathbf{F}_{uv}\mathbf{A}^{-1}+\mathbf{A}\mathbf{F}_{vu}\mathbf{A}^{-1}$, where

$$\mathbf{F}_{uv}=\sum_{s=2}^{\infty}\sum_{q=0}^{s-2}\sum_{m=q+1}^{s-1}\mathbf{D}^q\dot{\mathbf{G}}_u\mathbf{D}^{m-q-1}\dot{\mathbf{G}}_v\mathbf{D}^{s-m-1}\frac{t^s}{s!}$$

which has (a,b)'th element

$$\sum_{r=1}^{K}\sum_{s=2}^{\infty}\sum_{q=0}^{s-2}\sum_{m=q+1}^{s-1} d_a^q(\dot{g}_{ar}^u)d_r^{m-q-1}(\dot{g}_{rb}^v)d_b^{s-m-1}\frac{t^s}{s!}.$$

After substantial additional derivation, it can be shown that the (a,b)'th element of $\mathbf{F}_{uv}$ is then $\sum_{r=1}^{K}\dot{g}_{ar}^u\dot{g}_{rb}^v\Gamma_{arb}$, where Γ_{arb} is as given in section 3.2.□

A.2 Proofs of Theorem 1 and Corollary 1

Before proceeding with the proof of theorem 1, we will need the following lemma:

Lemma 1 *Suppose conditions 1 through 3 of theorem 1 hold. Then, for every convex compact* $T \subset \mathbf{\Theta}$*,*

$$\lim_{N\to\infty}\sup_{\boldsymbol{\theta}\in T}|\mathbf{S}_N(\boldsymbol{\theta})-\mathbf{S}(\boldsymbol{\theta})|=0$$

in probability.

Proof. For every $\epsilon,\eta>0$, proceed as follows. The absolute continuity of $\mathbf{S}(\boldsymbol{\theta})$ follows directly from conditions 1 through 3 by dominated convergence; and, by theorem 4.19 of Rudin (1976), this implies that there exists a $\delta_1>0$ such that

$$\sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in T:|\boldsymbol{\theta}-\boldsymbol{\phi}|<\delta_1}|\mathbf{S}(\boldsymbol{\theta})-\mathbf{S}(\boldsymbol{\phi})|<\epsilon/3.$$

Conditions 2 and 3 imply that there exists a δ_2 such that $0 < \delta_2 < \epsilon\eta\left(3\mathrm{E}[g(\mathbf{X}_i, \mathbf{Y}_i)]\right)^{-1}$ and

$$\begin{aligned}
&\Pr\left[\sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in T:|\boldsymbol{\theta}-\boldsymbol{\phi}|<\delta_2} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}_N(\boldsymbol{\phi})| > \epsilon/3\right] \\
&\quad\leq \Pr\left[\sup_{\boldsymbol{\theta}\in T} |\mathbf{I}_N(\boldsymbol{\theta})|\delta_2 > \epsilon/3\right] \\
&\quad\leq \Pr\left[g(\mathbf{X}_i, \mathbf{Y}_i)\delta_2 > \epsilon/3\right] \\
&\quad\leq \frac{3\delta_2\mathrm{E}[g(\mathbf{X}_i, \mathbf{Y}_i)]}{\epsilon} \\
&\quad< \eta.
\end{aligned}$$

Now, let $C \subset T$ be a finite set of points such that

$$\sup_{\boldsymbol{\theta}\in T}\inf_{\boldsymbol{\phi}\in C} |\boldsymbol{\theta} - \boldsymbol{\phi}| < \delta^* = \min(\delta_1, \delta_2),$$

then

$$\begin{aligned}
&\lim_{N\to\infty} \Pr\left[\sup_{\boldsymbol{\theta}\in T} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta})| > \epsilon\right] \\
&\leq \lim_{N\to\infty} \Pr\left[\sup_{\boldsymbol{\theta}\in T,\boldsymbol{\phi}\in C:|\boldsymbol{\theta}-\boldsymbol{\phi}|<\delta^*} \left\{|\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}_N(\boldsymbol{\phi})| + |\mathbf{S}_N(\boldsymbol{\phi}) - \mathbf{S}(\boldsymbol{\phi})| + |\mathbf{S}(\boldsymbol{\phi}) - \mathbf{S}(\boldsymbol{\theta})|\right\} > \epsilon\right] \\
&\leq \lim_{N\to\infty} \Pr\left[\sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in T:|\boldsymbol{\theta}-\boldsymbol{\phi}|<\delta_2} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}_N(\boldsymbol{\phi})| > \epsilon/3\right] \\
&\quad + \lim_{N\to\infty} \Pr\left[\sup_{\boldsymbol{\phi}\in C} |\mathbf{S}_N(\boldsymbol{\phi}) - \mathbf{S}(\boldsymbol{\phi})| > \epsilon/3\right] \\
&\quad + \lim_{N\to\infty} \Pr\left[\sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in T:|\boldsymbol{\theta}-\boldsymbol{\phi}|<\delta_1} |\mathbf{S}(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\phi})| > \epsilon/3\right] \\
&< \eta + 0 + 0. \square
\end{aligned}$$

Proof of theorem 1. To obtain asymptotic consistency of $\hat{\boldsymbol{\theta}}_N$, proceed as follows for every $\epsilon, \eta > 0$. As in lemma 1, conditions 1 through 3 establish the absolute continuity of $\mathbf{S}(\boldsymbol{\theta})$ on every compact $T \in \boldsymbol{\Theta}$. Combining this with condition 4 implies that there exists $\epsilon_0 > 0$, $\epsilon_1 > 0$, and $\delta > 0$ such that $\epsilon > \epsilon_1 > \epsilon_0$ and, for any convex compact $T \subset \boldsymbol{\Theta}$ such that

$\boldsymbol{\theta}_0 \in T$ and

$$\sup_{\boldsymbol{\theta} \in T} |\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \epsilon_1, \tag{7}$$

then

$$\sup_{\boldsymbol{\theta} \in T : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \epsilon_0} |\mathbf{S}(\boldsymbol{\theta})| \quad < \quad \inf_{\boldsymbol{\theta} \in T : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \epsilon_1} |\mathbf{S}(\boldsymbol{\theta})| - \delta.$$

Conditions 5 and 6 now imply that there exists an N_1 and a convex compact set T^* such that $\boldsymbol{\theta}_0 \in T^* \subset \boldsymbol{\Theta}$, T^* satisfies (7), and, for every $N \geq N_1$, $\Pr\left[\hat{\boldsymbol{\theta}}_N \in T^*\right] > 1 - \eta/2$. Lemma 1 implies that there also exists an $N_2 \geq N_1$ such that for every $N \geq N_2$, $\Pr\left[\sup_{\boldsymbol{\theta} \in T^*} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta})| > \delta/3\right] < \eta/2$ and therefore, with probability greater than $1 - \eta/2$,

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in T^* : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \epsilon_0} |\mathbf{S}_N(\boldsymbol{\theta})| &\leq \sup_{\boldsymbol{\theta} \in T^* : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \epsilon_0} |\mathbf{S}(\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta} \in T^*} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta})| \\
&< \inf_{\boldsymbol{\theta} \in T^* : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \epsilon_1} |\mathbf{S}(\boldsymbol{\theta})| - \sup_{\boldsymbol{\theta} \in T^*} |\mathbf{S}_N(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta})| \\
&\leq \inf_{\boldsymbol{\theta} \in T^* : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| > \epsilon_1} |\mathbf{S}_N(\boldsymbol{\theta})|.
\end{aligned}$$

Thus $\Pr\left[|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0| > \epsilon\right] < \eta$.

Conditions 1 through 3 ensure that, for every compact $T \subset \boldsymbol{\Theta}$, both $\mathbf{I}_N(\boldsymbol{\theta}) \to \mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{V}_N(\boldsymbol{\theta}) \to \mathbf{V}(\boldsymbol{\theta})$ in probability, as $N \to \infty$, for all $\boldsymbol{\theta} \in T$. Conditions 3(c) and 3(e), along with the consistency of $\hat{\boldsymbol{\theta}}_N$, now establishes that both $\mathbf{I}_N(\hat{\boldsymbol{\theta}}_N) \to \mathbf{I}(\boldsymbol{\theta}_0)$ and $\mathbf{V}_N(\hat{\boldsymbol{\theta}}_N) \to \mathbf{V}(\boldsymbol{\theta}_0)$ in probability, as $N \to \infty$. Consistency of the "sandwich" estimator of the variance now follows from condition 7.

For each M-vector $\mathbf{u}$, proceed as follows. By a simple Taylor's expansion, $N^{1/2}\mathbf{u}'\mathbf{S}_N(\boldsymbol{\theta}_0) = N^{1/2}\mathbf{u}'\mathbf{I}_N(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}^*$ on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_N$. Conditions 1 through 4 establish that $\mathrm{E}[\mathbf{S}_N(\boldsymbol{\theta}_0)] = 0$ and that $\mathrm{var}[N^{1/2}\mathbf{S}_N(\boldsymbol{\theta}_0)] = \mathbf{V}(\boldsymbol{\theta}_0)$ is bounded. The central limit theorem now implies that $N^{1/2}\mathbf{S}_N(\boldsymbol{\theta}_0)$ is asymptotically multivariate normal with mean 0 and variance $V(\boldsymbol{\theta}_0)$. The asymptotically normality of $N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0)$ then follows from the Cramer-Wold device, the previously established consistency of $\mathbf{I}_N(\hat{\boldsymbol{\theta}}_N)$, and condition 7.□

Proof of corollary 1. Corollary 1 now follows directly from theorem 1 by verifying that

$$\mathrm{E}[\mathbf{s}_{i\cdot}^u(\boldsymbol{\theta}_0)\,\{\mathbf{s}_{i\cdot}^v(\boldsymbol{\theta}_0)\}'] = \mathrm{E}[-\frac{\partial \mathbf{s}_{i\cdot}^u(\boldsymbol{\theta}_0)}{\partial \theta_v}]$$

when the working conditional log likelihood is correct.□

REFERENCES

Agresti, A. (1989), "A Survey of Models for Repeated Ordered Categorical Response Data," *Statistics in Medicine,* 8, 1209–1224.

Anderson, P. K., Hansen, L. S., and Keiding, N. (1991), "Non- and Semi- parametric Estimation of Transition Probabilities From Censored Observation of A Non-homogeneous Markov Process," *Scandinavian Journal of Statistics,* 18, 153–167.

Ashby, M., Neuhaus, J. M., Hauck, W. W., Bacchetti, P., Heilbron, D. C., Jewell, N. P., Segal, M. R., and Fusaro, R. E. (1992), "An Annotated Bibliography of Methods for Analysing Correlated Categorical Data," *Statistics in Medicine,* 11, 67–99.

Diabetic Retinopathy Study Research Group. (1976), "Preliminary Report on Effects of Photocoagulation Therapy," *American Journal of Opthalmology,* 81, 383–396.

——— (1981), "Clinical Applications of Diabetic Retinopathy Study Findings, DRS Report Number 8," *Ophthalmology,* 88, 583–600.

Fahrmeir, L., and Kaufmann, H. (1987), "Regression Models for Non-stationary Categorical Time Series," *Journal of Time Series Analysis,* 8,147–160.

Gilula, Z., and Haberman, S. J. (1994), "Conditional Log-linear Models for Analyzing Categorical Panel Data," *Journal of the American Statistical Association,* 89, 645–656.

Golub, G. H., and Van Loan, C. F. (1989), *Matrix Computations* (2nd ed.), Baltimore, MD: The Johns Hopkins University Press.

Huber, P. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), pp. 221–234.

Kalbfleisch, J. D., and Lawless, J. F. (1985), "The Analysis of Panel Data Under A Markov Assumption," *Journal of the American Statistical Association,* 80, 863–871.

Kosorok, M. R. (1993), "A Longitudinal Model for Binary Response Data With Time-dependent Covariates," *Technical Report 84*, Department of Biostatistics, University of Wisconsin, Madison, Wisconsin.

Kosorok, M. R., and Chao, W.-H. (1994), "The Analysis of Longitudinal Ordinal Response Data In Continuous Time," *Technical Report 87*, Department of Biostatistics, University of Wisconsin, Madison, Wisconsin.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

Lindsey, J. K. (1993), *Models for Repeated Measurements,* New York: Oxford University Press.

Rand, L. I., Prud'homme, G. J., Ederer, F., Canner, P. L., and The Diabetic Retinopathy Study Research Group. (1985), "Factors Influencing The Development of Visual Loss in Advanced Diabetic Retinopathy," *Investigative Ophthalmology and Visual Science,* 26, 983–991.

Ross, S. M. (1993), *Introduction to Probability Models* (5th ed.), San Diego, CA: Academic Press.

Rudin, W. (1976), *Principles of Mathematical Analysis* (3rd ed.), New York: McGraw-Hill.

Slud, E., and Kedem, B. (1994), "Partial Likelihood Analysis of Logistic Regression and Autoregression," *Statistica Sinica,* 4, 89–106.

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.

Tuma, N. B., Hannan, M. T., and Groeneveld, L. P. (1979), "Dynamic Analysis of Event Histories," *American Journal of Sociology,* 84, 820–854.

Wei, L. J. and Johnson, W. E. (1985), "Combining Dependent Tests With Incomplete Repeated Measurements," *Biometrika,* 72, 359–364.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics,* 44, 1049–1060.

Table 1: Parameter estimates and 95% confidence intervals (in parentheses) for the final LED-1 model (omitting the three way interaction), including estimates of $\mu = \exp[\xi]$ and $\phi^{\frac{1}{3}} = \exp[-\mu/3]$. Confidence intervals for μ and $\phi^{\frac{1}{3}}$ are based on the delta method. An "$*$" denotes statistical significance at the .05 level.

Term	Estimate	Confidence Interval
intercept	-2.2355	$(-2.8941, -1.5769)^*$
age	$.0038$	$(-.0051, .0128)$
sex	$-.0234$	$(-.2580, .2112)$
BRG	1.0758	$(.8210, 1.3306)^*$
Treatment (T)	$.2417$	$(-.1552, .6387)$
time (t)	$.4459$	$(-.0130, .9049)$
change (c)	$.7552$	$(.1202, 1.3901)^*$
t x c	$-.4540$	$(-.8730, -.0349)^*$
T x t	$-.1911$	$(-.4576, .0755)$
T x c	$.3207$	$(-.3181, .9595)$
μ	$.4572$	$(.4077, .5069)^*$
$\phi^{\frac{1}{3}}$	$.8586$	$(.8444, .8728)^*$

Table 2: Coefficient estimates and 95% confidence intervals (in parentheses) for the marginal models M.I and MP.I. An "$*$" denotes statistical significance at the .05 level.

Term	M.I		MP.I	
intercept	-4.6830	$(-5.1194, -4.2466)^*$	-4.0103	$(-4.4318, -3.5889)^*$
age	.0003	$(-.0071, .0076)$	.0003	$(-.0070, .0076)$
sex	$-.0112$	$(-.2246, .2023)$	$-.0121$	$(-.2249, .2007)$
BRG	.9070	$(.6898, 1.1241)^*$	.9063	$(.6898, 1.1228)^*$
Treatment (T)	.0702	$(-.2419, .3824)$	.0627	$(-.2221, .3475)$
time (t)	1.1590	$(1.0409, 1.2770)^*$	.9468	$(.8346, 1.0591)^*$
change (c)	1.8842	$(1.6054, 2.1630)^*$	1.4001	$(1.1331, 1.6670)^*$
t x c	$-.8793$	$(-.9756, -.7831)^*$	$-.6919$	$(-.7788, -.6051)^*$
T x t	$-.0263$	$(-.1227, .0702)$	$-.0368$	$(-.1440, .0704)$
T x c	.0487	$(-.3011, .3985)$	.0935	$(-.2449, .4319)$

Table 3: Coefficient estimates, autocorrelation coefficients $\phi^{\frac{1}{3}}$, and 95% confidence intervals (in parentheses) for the marginal models M.AR(1) and MP.AR(1). An "$*$" denotes statistical significance at the .05 level.

Term	M.AR(1)		MP.AR(1)	
intercept	−4.1221	(−4.5426, −3.7017)*	−3.8409	(−4.2567, −3.4251)*
age	.0014	(−.0058, .0087)	.0027	(−.0044, .0099)
sex	−.0494	(−.2569, .1582)	−.0266	(−.2332, .1799)
BRG	.9257	(.7132, 1.1382)*	.9118	(.7008, 1.1229)*
Treatment (T)	.0891	(−.1754, .3535)	.0215	(−.2409, .2840)
time (t)	.8213	(.7380, .9047)*	.7053	(.6223, .7884)*
change (c)	1.0317	(.7982, 1.2652)*	.8877	(.6316, 1.1439)*
t x c	−.4866	(−.5742, −.3990)*	−.3860	(−.4772, −.2948)*
T x t	−.0071	(−.0719, .0578)	−.0350	(−.1030, .0330)
T x c	−.0568	(−.2627, .1490)	.1362	(−.1035, .3759)
$\phi^{\frac{1}{3}}$	.8154	–	.8108	–

Table 4: Parameter estimates and 95% confidence intervals (in parentheses) for the 4-category LED-1 model. Confidence intervals for the μ's are based on the delta method. An "$*$" denotes statistical significance at the .05 level.

Term	Estimate	Confidence Interval
α_1	.3346	$(-.4086,\ 1.0778)$
α_2	.0700	$(-.0421,\ .1821)$
α_3	$-.4678$	$(-.5837,\ -.3519)^*$
age	.0139	$(.0053,\ .0224)^*$
sex	.1005	$(-.0888,\ .2899)$
BRG	.4061	$(.2121,\ .6000)^*$
Treatment (T)	.0812	$(-.2310,\ .3934)$
time (t)	.5084	$(.2915,\ .7253)^*$
change (c)	.8043	$(.3507,\ 1.2580)^*$
t x c	$-.4226$	$(-.6146,\ -.2306)^*$
T x t	$-.0918$	$(-.2501,\ .0665)$
T x c	$-.0709$	$(-.4932,\ .3514)$
μ_{12}	2.8665	$(2.4833,\ 3.2497)^*$
μ_{23}	2.7721	$(2.2146,\ 3.3295)^*$
μ_{34}	1.6615	$(1.4058,\ 1.9171)^*$
μ	.1036	$(.0715,\ .1357)^*$

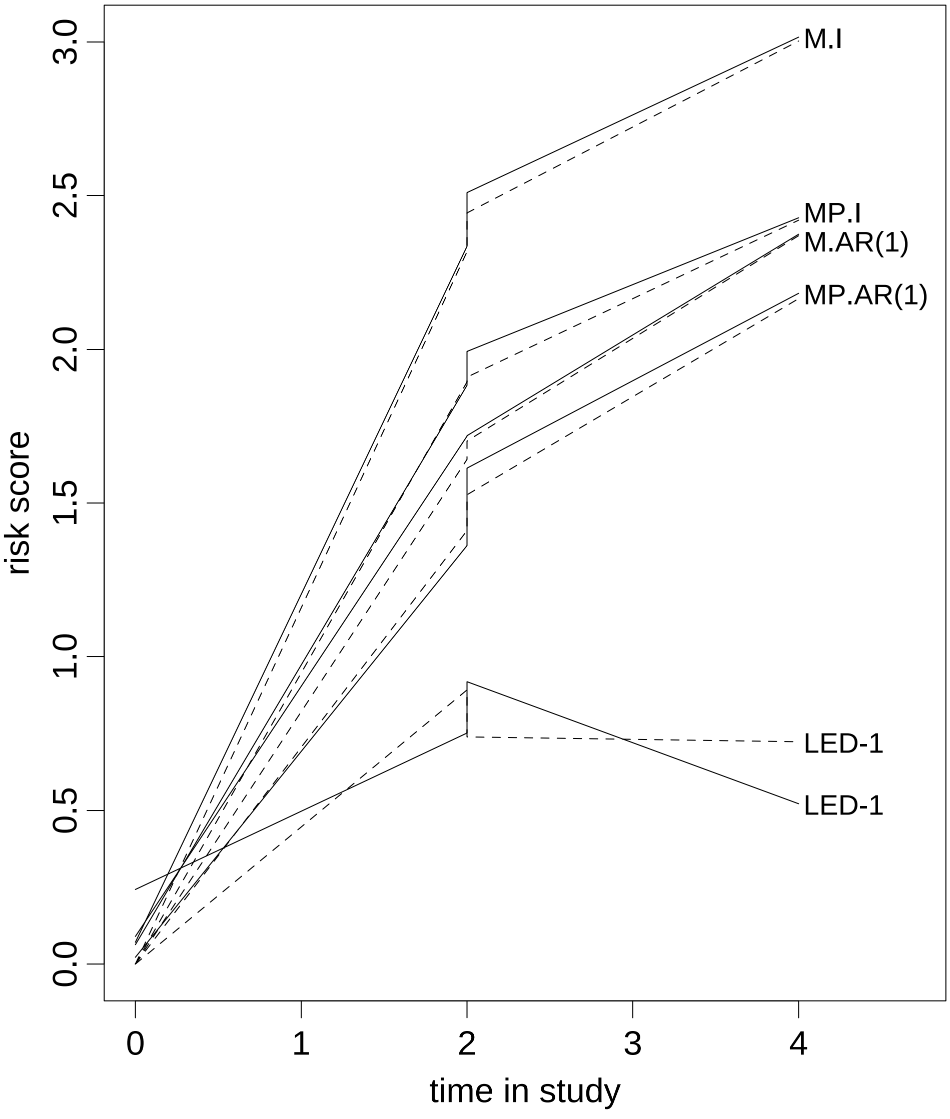

Figure 1: Plot of change in risk score over time for a subject whose right eye initially received treatment (solid line) versus a subject with the same baseline characteristics whose right eye was initially untreated (dashed line), for the LED-1 and four marginal binary models (labelled on the right).

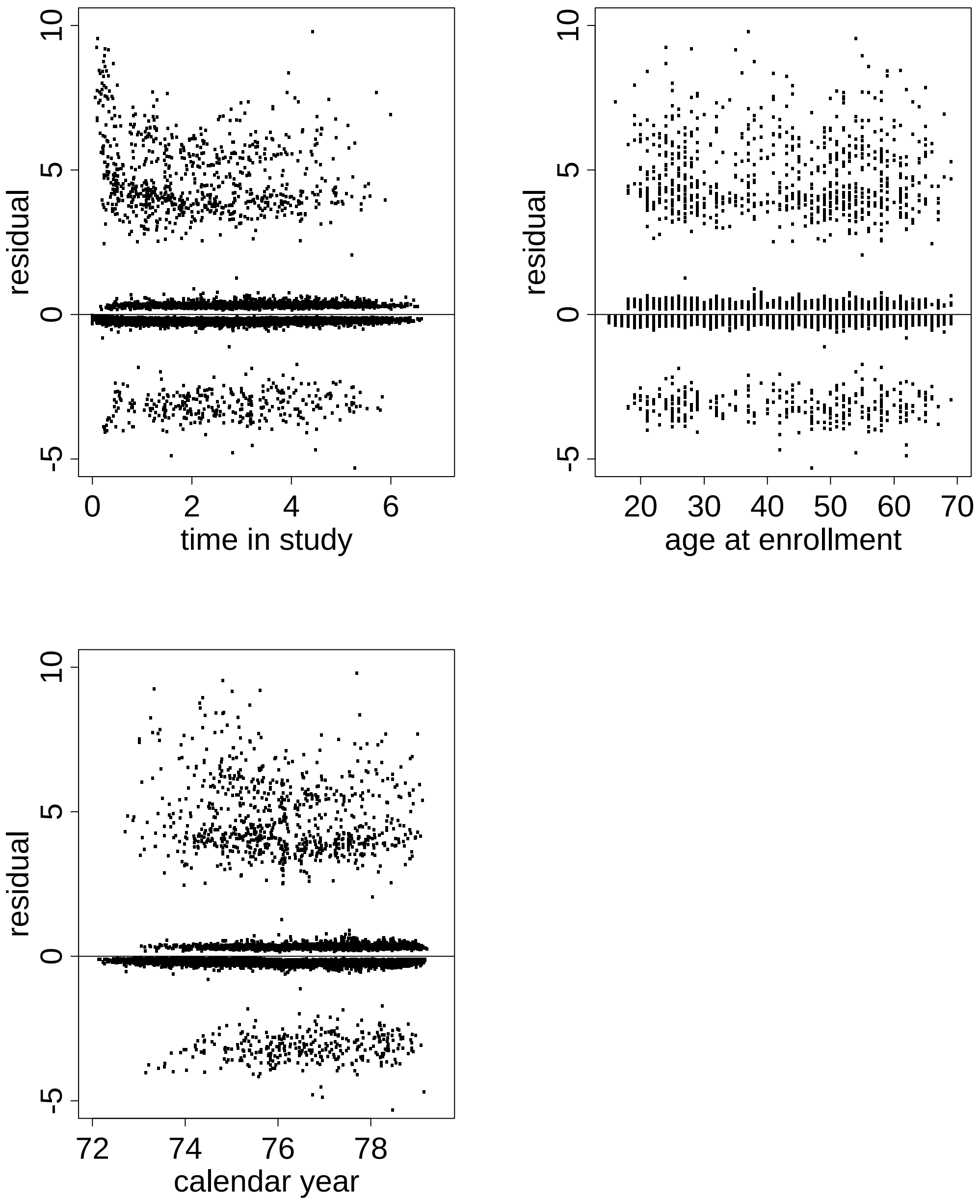

Figure 2: Plots of summary residuals for the LED-1 binary model. A horizontal line at zero is given for reference.

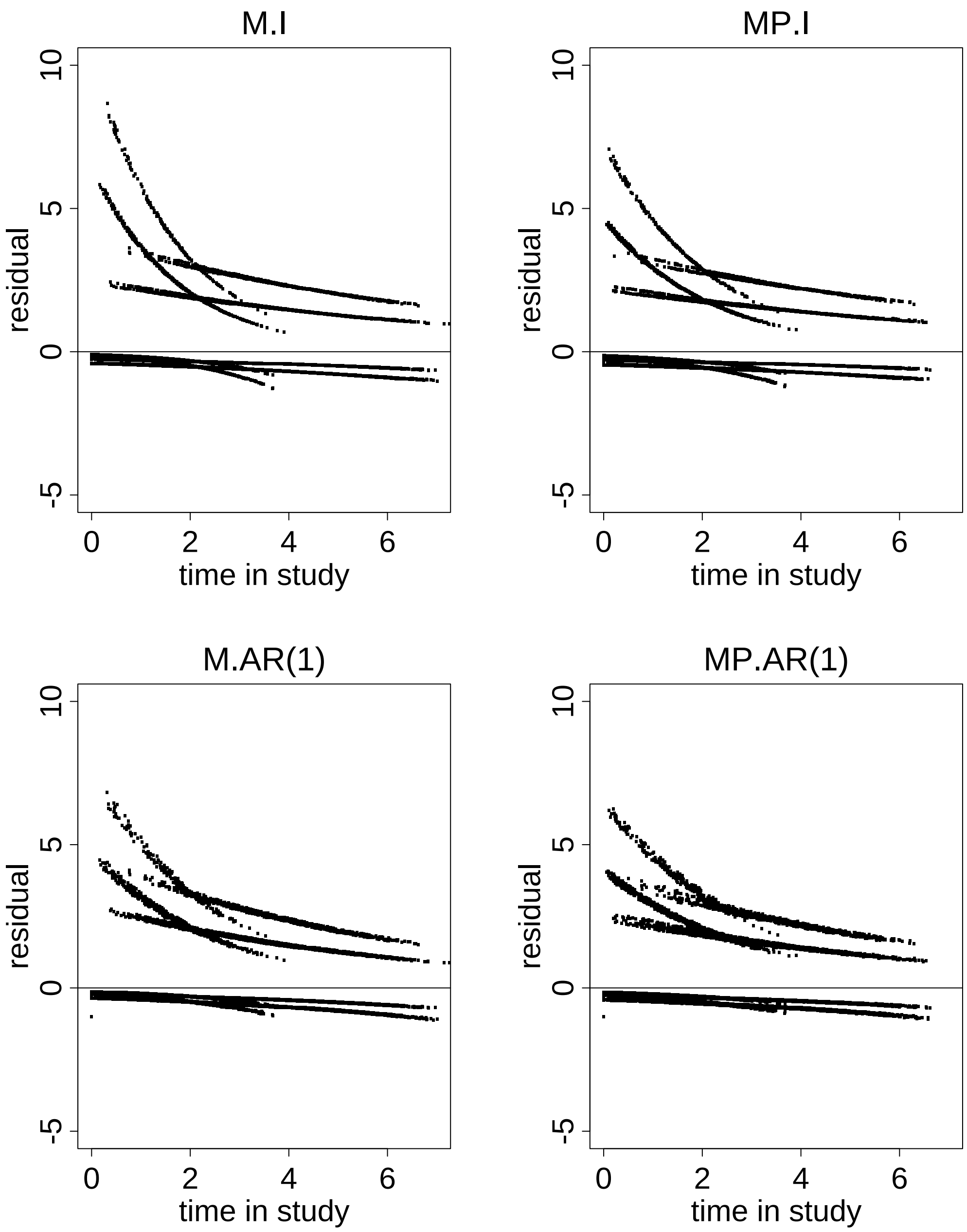

Figure 3: Plots of standardized residuals versus duration time in study for marginal and marginal predictive binary models with independence and AR(1) working correlation structures.

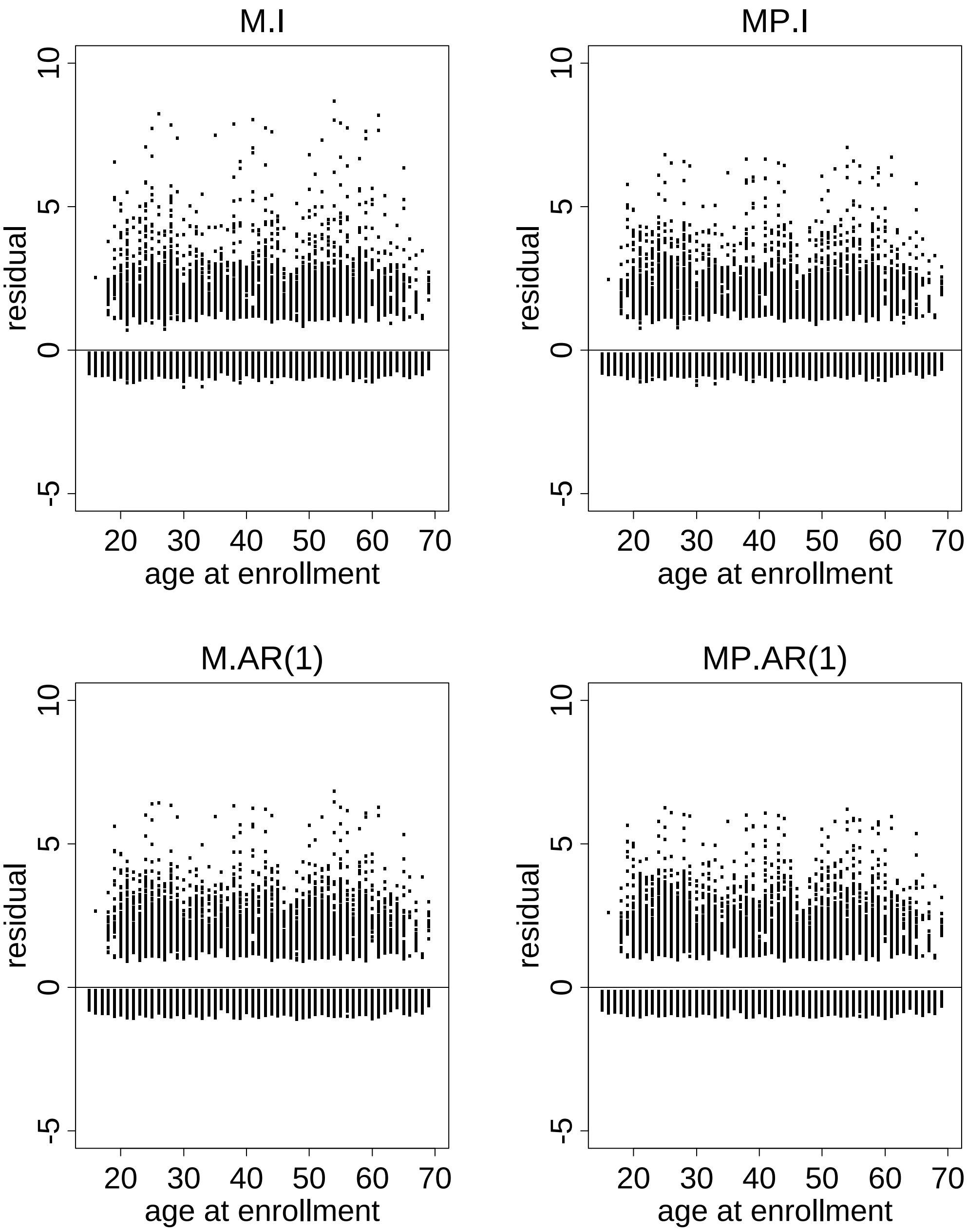

Figure 4: Plots of standardized residuals versus age at enrollment for marginal and marginal predictive binary models with independence and AR(1) working correlation structures.

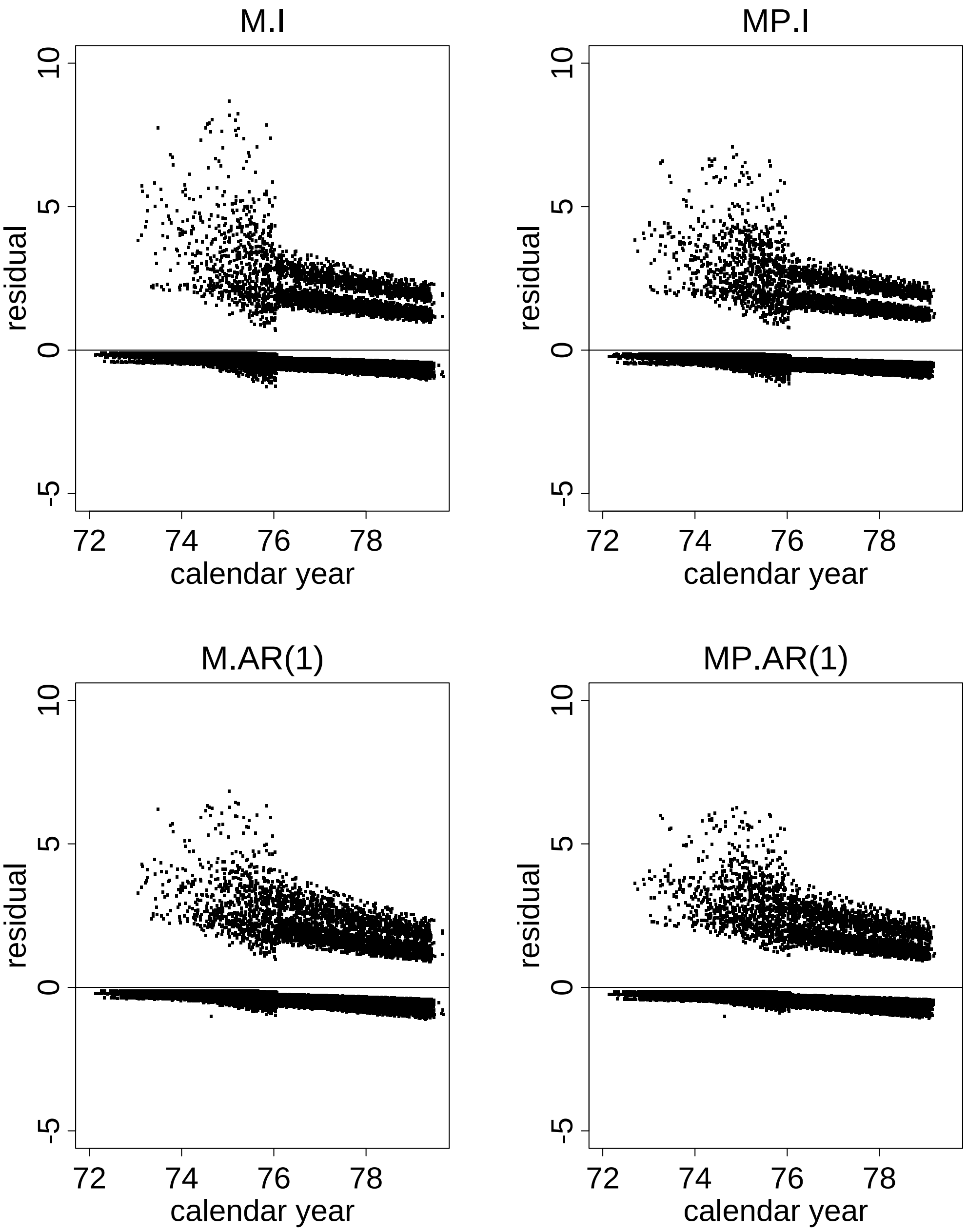

Figure 5: Plots of standardized residuals versus calendar time for marginal and marginal predictive binary models with independence and AR(1) working correlation structures.

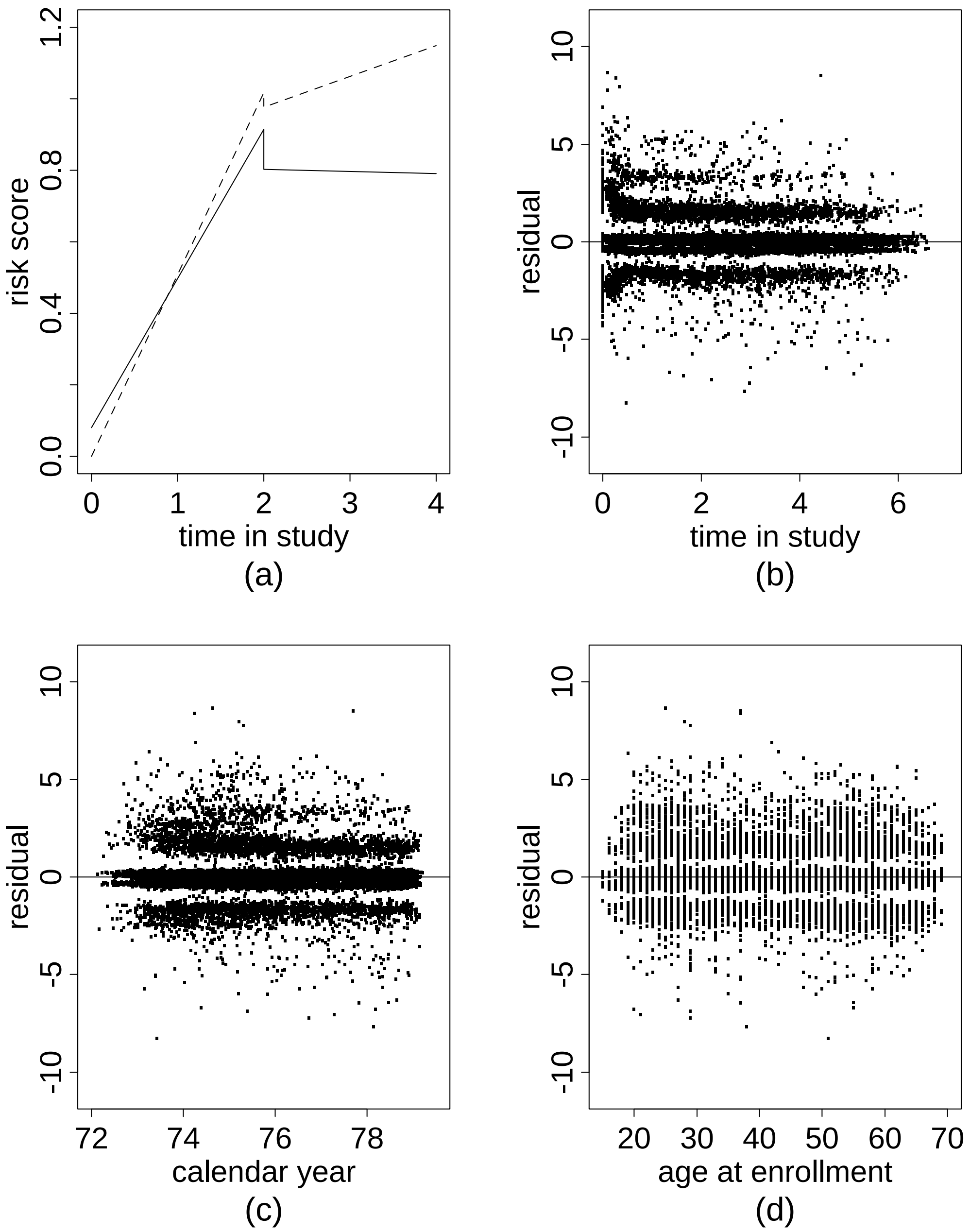

Figure 6: The 4-category LED-1 analysis: (a) plot of change in risk score for a subject whose right eye initially received treatment (solid line) versus a subject with the same baseline characteristics whose right eye was initially untreated (dashed line), as well as plots (b), (c), and (d) of the summary residuals versus time in study, calendar years, and age at enrollment, respectively.