*Harvard University*
Harvard University Biostatistics Working Paper Series

*Year* 2007 *Paper* 65

# Estimating Time-to-Event From Longitudinal Categorical Data Using Random Effects Markov Models: Application to Multiple Sclerosis Progression

Micha Mandel* Rebecca A. Betensky†

*The Hebrew University of Jerusalem, msmic@mscc.huji.ac.il

†Harvard School of Public Health, betensky@hsph.harvard.edu

# Estimating Time-to-Event From Longitudinal Categorical Data Using Random Effects Markov Models: Application to Multiple Sclerosis Progression *

Micha Mandel, The Hebrew University of Jerusalem

and

Rebecca A. Betensky, Harvard School of Public Health

June 27, 2007



*This research was supported in part by NIH CA075971, the Harvard Center for Neurodegeneration and Repair (HCNR), and the Partners Multiple Sclerosis Center. We thank Dr. Richard Rudick, Director of the Mellen Center for Multiple Sclerosis at the Cleveland Clinic Foundation, for providing us the Avonex trial data on behalf of the Multiple Sclerosis Collaborative Research Group (Principal Investigator, Lawrence Jacobs, MD, deceased). The original Avonex study was sponsored by NIH (NS26321) and Biogen, Inc (Cambridge Massachusetts).

**Abstract**

Longitudinal categorical data are common in many scientific studies, including those of multiple sclerosis (MS), and are frequently modeled using Markov dependency. Several authors have proposed random effects Markov models to account for heterogeneity in the population. In this paper, we go one step further and study prediction based on random effects Markov models. In particular, we show how to calculate the probabilities of future events and confidence intervals for those probabilities, given observed data on the categorical outcome and a set of covariates, and how to update them overtime. We discuss the usefulness of depicting these probabilities for visualization and interpretation of model results and illustrate our method using data from a phase III clinical trial that evaluated the utility of interferon beta-1a to MS patients of type relapsing-remitting.

KEYWORDS: Markov model, transition model, ordinal response, prediction.

# 1 Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of central nervous system myelin. In time, the disease causes disability which is measured on the ordinal expanded disability status scale (EDSS). In MS studies, disability is often evaluated semi-annually with the aim of estimating the probability of progression, as defined on the EDSS scale. The natural assumption of Markov dependency provides a convenient framework for the estimation of probabilities of various time-dependent events that are of biological interest (prediction). Examples of such events are: reaching a certain level, being in a certain level for two consecutive visits, reaching a group of levels and so forth. Mandel et.al. (2007) developed methodology for analyzing these kinds of events using fixed effects transition models and applied it to MS data. However their first-order models suffered from lack of fit, partially due to the heterogenous nature of the disease.

Transition models for longitudinal data (Diggle et.al., 2002) express the joint distribution of repeated measures as a product of conditional distributions, and are especially useful when the ultimate goal of the analysis is that of prediction. However, when important subject specific covariates are not measured, the model explains only part of the true heterogeneity in the data, and moreover, the Markov assumption may be violated. One remedy is to incorporate random effects into the Markov transition model. The basic assumption is that conditional on an observed set of covariates and an unobserved latent variable, the sequence of the categorical variables follows a Markov chain.

Two-state Markov transition models with random effects have been studied by several authors (Cook and Ng, 1997, Albert and Waclawiw, 1998, Albert and Follmann, 2003). These papers assume implicitly that the distribution of the first (baseline) state is indepen-

dent of the random effect given covariates. This strong and somewhat unnatural assumption is relaxed in a series of papers (i.e., Aitkin and Alfó, 1998, 2003 Alfó and Aitkin, 2000) that suggest models for the influence of the first state on the random effect's distribution. A related framework in which subjects transition between states according to a continuous time Markov process, but only are observed at $n$ time points was studied by Kalbfleisch and Lawless (1985), and random effects were introduced to this model by Cook (1999) and by Cook, et al. (2004).

The current paper focuses on the problem of prediction, or second stage estimation, from random effects Markov models. It is based on estimation results derived in earlier work to make inferences about the future of the process given its past. Specifically, let $Y_0, Y_1, \ldots, Y_n$ be the sequence of the categorical variables observed for a subject at the equally spanned time points $0, 1, \ldots, n$, and let $X$ be a vector of covariates. We are interested in estimating $P(Y_v = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$ or more generally, $P(A_v|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$, where the event $A_v$ is determined by $(Y_0, Y_1, \ldots, Y_v)$, and $v > n$. Only a few papers have dealt with second stage inference in the framework of transition models. Albert (1994) and Mandel et.al. (2007) studied parameters such as mean first passage time and probabilities of time to events, but only for fixed effects models. Albert and Waclawiw (1998) also estimated mean first passage time in random effects models, but only at the population level, and not based on subject specific history.

In this paper we extend the methodology of Mandel et.al. (2007) by developing methods for estimation of time-to-event probabilities and associated confidence intervals under random effects Markov models. Our estimates take into account subject specific history and can be updated over time when more data are collected. In the presence of random effects,

much more computational effort is required for deriving estimates and confidence intervals, and more importantly, careful interpretation of estimated coefficients and predicted values is necessary. However, proper interpretation of model results with the aid of graphical tools presented below enables important insights to the longitudinal process studied, e.g., to the natural history of MS.

It is important to distinguish between our usage of the term *prediction* to refer to $P(Y_v = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x)$, and its usage in the linear and generalized linear mixed models framework (Robinson, 1991, Jiang and Lahiri, 2006). The latter usage of prediction is in a subject-specific sense; either the random effect is estimated or a conditional distribution given the random effect is estimated. Thus, letting $U$ denote the latent random effect, the analog of prediction in the generalized linear models framework is estimation of $U|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x$ or $P(Y_t = y|Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n, X = x, U)$ and their associated mean squared errors. Because we do not have many observations on each subject, and subjects are quite heterogeneous, neither of these quantities is suitable in our setting. Instead, we integrate over the random effect to obtain an estimate for future events that is based on observed data alone. This very important distinction is illustrated on the MS data in Section 4.

The paper is organized as follows. Section 2 defines the model and reviews estimation based on previous papers mentioned above. Section 3 deals with prediction under the mixed effects model. It describes generation of probability estimators and their variances as a function of time given subject-specific covariates and history. Section 4 applies the method to data from a phase III clinical trial of MS patients. It discusses several important points regarding interpretation of the models and provides important insights into the natural

history of MS. Section 5 presents results of a simulation study and Section 6 completes the paper with a discussion.

# 2 Preliminaries

This section gives a brief review of the construction and estimation of the Markov transition mixed model. It also describes an identification problem that affects estimation of model parameters, but not prediction.

## 2.1 The Model

Consider a discrete time Markov process over the states $\{1, 2, \ldots, J\}$. Let $Y_{iv}$ be the state of subject $i$ at visit (time) $v$, and let $Y_{iv}^{\bullet} = (Y_{iv}^2, \ldots, Y_{iv}^J)'$, where $Y_{iv}^j = I\{Y_{iv} = j\}$ and $I$ is the indicator function. Let $U_i \sim G$ (with density $g$) denote a subject specific latent variable and $X_i \sim H$ a vector of covariates. Using lower case letters for realizations of random variables, the data for subject $i$ having $n_i$ transitions are $(y_{i0}, y_{i1}, \ldots, y_{in_i}, x_i)$ and its contribution to the likelihood, $\mathcal{L}_i$, is given by (omitting the subscript $i$)

$$H(dx)P(Y_0 = y_0|X = x)\int_{-\infty}^{\infty}\prod_{v=1}^{n} P(Y_v = y_v|Y_{v-1} = y_{v-1}, X = x, U = u)g(u|Y_0 = y_0, X = x)du. \tag{2.1}$$

Albert and Waclawiw (1998) assume that each Markov process is in equilibrium, and Aitkin and Alfó (2003) specify a parametric model to the distribution of $Y_0|X, U$. Under these assumptions, $P(Y_0 = y_0|X = x)$ is informative and should be used for likelihood-based inference. While such additional assumptions enable exploitation of more information from the data, they cannot always be justified. Hence, our inference will be conditional on $(Y_0, X)$. Following Aitkin and Alfó (1998), we assume that $g(u|Y_0 = y_0, X = x) = \sigma^{-1}g_0([u-\eta' y_0^{\bullet}]/\sigma)$

for some known $g_0$, e.g., the standard normal density. Thus, given the initial state, the random effect is independent of the covariates, and the initial state affects $g$ only through its location. Then, by omitting $H(dx)P(Y_0 = y_0|X = x)$ and after changing variables, the likelihood contribuion reduces to

$$\int_{-\infty}^{\infty} \prod_{v=1}^{n} P(Y_v = y_v|Y_{v-1} = y_{v-1}, X = x, U = \sigma u + \eta' y_0^{\bullet}) g_0(u) du. \tag{2.2}$$

To model the transition probabilities, suppose that

$$P(Y_v = j|Y_{v-1} = k, X = x, U = u) = p_{k,j}(\beta' x + \gamma u). \tag{2.3}$$

There are two assumptions embedded in (2.3). First, the Markov model is time homogeneous, and second, the transition probabilities depend on $X$ and $U$ only through their linear combination. The second assumption reflects our view of the random effect as representing unmeasured covariates, that if observed, would be modelled as we do the observed covariates. The likelihood contribution becomes

$$\int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x + \gamma[\sigma u + \eta' y_0^{\bullet}]) g_0(u) du. \tag{2.4}$$

Finally, the transition probability depends on the transition $k \to j$ according to a vector $\alpha = (\alpha_{kj})$

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x + \gamma[\sigma u + \eta' y_0^{\bullet}]; \alpha) g_0(u) du. \tag{2.5}$$

Note that (i) the probability of the transition $k \to j$ may be a function of components of $\alpha$ other than $\alpha_{kj}$, and (ii) this vector is subject to several constraints to ensure that the rows of the transition matrix sum to one.

## 2.2 Identifiability

It is clear from (2.4) and (2.5) that $(\gamma, \sigma, \eta)$ is not identifiable. This, however, does not present a problem for estimation of $\beta$ or for prediction, hence $\gamma$ will be set to 1 in the sequel, giving the final working model

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x + \sigma u + \eta' y_0^\bullet; \alpha) g_0(u) du. \tag{2.6}$$

Suppose that $x$ contains the initial state $y_0^\bullet$. Then, it is clear from (2.6) that the effect of $y_0^\bullet$ as a covariate (i.e., the effect of $y_0^\bullet$ given $U = u$) is not identifiable. Thus, interpretation of $\eta$ should be made with care. It seems more reasonable to tie $\eta$ to the random effect's distribution rather than to the transition probabilities when the initial state is somewhat arbitrary, relating to the sampling time. In that case, interpretation of $\eta$ is as the center of the random effect's distribution and not in terms of an odds ratio. Moreover, there is nothing special about $y_0^\bullet$ in the discussion above, and the same reasoning applies for any other covariate; $\beta$ is not identifiable if the random effects' density takes the form $g(u|Y_0 = y_0, X = x) = \sigma^{-1} g_0([u - \eta' y_0^\bullet - \zeta' x]/\sigma)$. This is quite a reasonable form for $g(u|Y_0 = y_0, X = x)$ when viewing $U$ as an unmeasured covariate and assuming a multivariate normal distribution for $(X, U)$. Although interpretation of model results is problematic, for prediction purposes this identification issue raises no difficulties since prediction is based on $\beta' x + \gamma[\sigma u + \eta' y_0^\bullet]$, which is identifiable (for given $x, u$ and $y_0^\bullet$).

## 2.3 Estimation

The likelihood of $N$ independent subjects is given by

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}_i = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{v=1}^{n_i} p_{y_{iv-1},y_{iv}}(\beta' x_i + \sigma u + \eta' y_{i0}^\bullet; \alpha) g_0(u) du. \tag{2.7}$$

Estimation of nonlinear random effects model is done via maximization of (2.7). This is a standard, though difficult task. A convenient and flexible routine for maximization of the likelihood is the procedure NLMIXED in SAS, which contains several methods for integration and optimization when $g_0$ is normal (Littell et.al., 2006). The EM algorithm is an alternative way of maximizing $\mathcal{L}$ (Aitkin and Alfó, 1998) without requiring specificaion of the distribution of the random effects. However, Agresti found that the random effect distribution has to be extremely nonormal for the normal GLMM to suffer from bias or inefficiency (Agresti 2002, pp 547-548). Thus, the normal model for the random effects distribution seems reasonable in most circumstances in terms of simplicity and interpretability.

### 2.4 Notation

The following notation will be used in the sequel: $\theta \equiv (\beta', \eta')'$ is the vector of fixed effects, $\vartheta \equiv (\alpha', \theta', \sigma)'$ is a vector of length $m$ of all unknowns, $z \equiv (x', y_0^{\bullet\prime})'$ is the vector of observed predictors, and $w \equiv \theta' z + \sigma u$ is the linear predictor. Depending on the context, the transition probabilities will be denoted either by $p_{y_{v-1},y_v}(\theta' z + \sigma u; \alpha)$, $p_{y_{v-1},y_v}(x, u, y_0; \vartheta)$ or $p_{y_{v-1},y_v}(w; \alpha)$. The superscript $(s)$ will be added to denote transitions in $s$ steps: for example, $p^{(s)}_{k,j}(x, u, y_0; \vartheta) = P(Y_{v+s} = j | Y_v = k, X = x, U = u, Y_0 = y_0; \vartheta)$.

## 3 Prediction

The ultimate goal of our analysis is that of prediction of a future observation given the past observations and covariates:

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x). \tag{3.1}$$

Note that here, in contrast to the fixed effects case,

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) \neq P(Y_{v-n} = y_v | Y_0 = y_n, X = x),$$

because the distribution of $(Y_0, Y_1, \ldots, Y_v)$ has the Markov property only conditional on $U$. Thus, the process itself provides information on the latent $U$, which in turn, is used to predict future events.

Using

$$\begin{aligned} & P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) \\ &= \int_{-\infty}^{\infty} P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, u) g(u | \{Y_t = y_t\}_{0 \le t \le n}, X = x) du \quad (3.2) \\ &= \int_{-\infty}^{\infty} p_{y_n, y_v}^{(v-n)}(x, u, y_0; \vartheta) \frac{\prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u, y_0; \vartheta)}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u^*, y_0; \vartheta) g_0(u^*) du^*} g_0(u) du, \end{aligned}$$

(3.1) is estimated by

$$\hat{P}(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x) = \int_{-\infty}^{\infty} p_{y_n, y_v}^{(v-n)}(x, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u, y_0; \hat{\vartheta}) g_0(u)}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} du, \quad (3.3)$$

where $\hat{\vartheta}$ is an estimate of $\vartheta$. This last integral can be calculated using numerical methods. A natural simple way is to conduct Monte Carlo integration with respect to the assumed known density $g_0$. Alternatives are MCMC, which eliminates the burden of approximating the integral in the denominator, and general numerical integration methods.

Of special interest for us is prediction for a subject without any observed transitions. This represents a patient at diagnosis and is the analogous to prediction in a model without random effects. For such a case, (3.3) reduces to

$$P_{\hat{\vartheta}}(Y_v = y_v | Y_0 = y_0, X = x) = \int_{-\infty}^{\infty} p_{y_0, y_v}^{(v)}(x, u, y_0; \hat{\vartheta}) g_0(u) du. \quad (3.4)$$

As mentioned in Section 1, it is important to distinguish (3.1) from

$$P(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, X = x, U = u). \quad (3.5)$$

The problem of estimating quantities similar to (3.5) is referred to as *prediction* in the mixed model literature (Jiang and Lahiri, 2006). In using (3.5), one aims at estimating the *subject specific* transition probability which is a random variable, even in a frequentist's point of view. A point predictor for (3.5) is $P(Y_v = y_v|\{Y_t = y_t\}_{0\leq t\leq n}, X = x, U = \hat{u}_i)$, where $\hat{u}_i$ is the mean or the mode of $U|\{Y_t = y_t\}_{0\leq t\leq n}, X = x$ under $\hat{\vartheta}$ (Booth and Hobert, 1998). Jiang (2003) suggests the empirical best predictor $\mathbb{E}\{P(Y_v = y_v|\{Y_t = y_t\}_{0\leq t\leq n}, X = x)|U = u\}$ which is exactly (3.3). Thus, the point estimators of the two prediction problems are the same, but the estimands differ. With small numbers of observations on each subject, the utility of subject-specific parameters, such as those in (3.5) is questionable. This point will be illustrated further in the data analysis in Section 4.

## 3.1 Prediction Variance

Letting $\hat{\vartheta}$ be the estimator of $\vartheta$ and assuming that $\sqrt{N}(\hat{\vartheta} - \vartheta) \to N(0, \Sigma)$, we can calculate the asymptotic variance of (3.3) and (3.4) by the delta method. Let $P(x, u, y_0; \vartheta)$ be the transition matrix evaluated at $(x, u, Y_0 = y_0; \vartheta)$, and let $p^{*(v-n)}(y_n, y_v, P(x, u, y_0; \vartheta)) = p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta)$ be the $(v - n)$-step transition probability as a function of the one-step transition matrix. We have that

$$\frac{\partial}{\partial\vartheta}p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta) = \frac{\partial}{\partial \text{vec}(P)}p^{*(v-n)}(y_n, y_v, P(x, u, y_0; \vartheta))\frac{\partial}{\partial\vartheta}\text{vec}(P(x, u, y_0; \vartheta)), \quad (3.6)$$

where vec$(P)$ is the vector representation of the matrix $P$. The first term in the right hand side of (3.6) can be calculated by a simple matrix multiplication as shown by Mandel et.al. (2007), and the rows of the second term are

$$\left(\frac{\partial}{\partial\alpha}p(\cdot, \cdot, w; \alpha)\ ,\ \frac{\partial}{\partial w}p(\cdot, \cdot, w; \alpha)(z', u)\right)$$

where $p(\cdot, \cdot, w; \alpha)$ is the generic form of the transition probabilities evaluated at $w$ and $\alpha$. An illustration of a specific calculation of these derivatives for a partial proportional odds model (see (4.1) below) is given in Appendix A. To calculate the derivative of (3.4), one should average (3.6) with respect to $g_0$ which can be done again by numerical integration.

Differentiation of (3.3) is more complicated, but can still be carried out analytically as shown in Appendix B. An alternative approach for estimating the variance is based on a simulation that replaces the analytic differentiation, but makes use of the asymptotic properties of the parameters' estimators:

1. Calculate $\hat{\vartheta}$ and $\hat{\Sigma}$, the estimates of $\vartheta$ and $\Sigma$.

2. Sample $B$ vectors $\vartheta_1, \ldots, \vartheta_B$ from the normal distribution with parameters $\hat{\vartheta}$ and $\hat{\Sigma}/N$.

3. Calculate (3.3) with $\vartheta_b$ instead of $\hat{\vartheta}$ $(b = 1, \ldots, B)$.

4. Calculate the variance of the estimates in the previous step, or calculate confidence intervals using their distribution.

Note the complexity of this algorithm that requires numerical integration in step 3 for each of the simulated samples. The calculation of prediction variance for models having fixed effects only are considerably simpler (Mandel et al., 2007).

## 3.2 Choice of Parameters

Simple manipulations of the estimated transition matrix enable estimation of different parameters of interest. For example, one may be interested in estimating time until the process first visits a certain set of states $\mathcal{S}$ (hitting time), or time until the first two

consecutive visits to $\mathcal{S}$. The second parameter is of great interest in MS since EDSS in one visit may indicate a temporary progression that is much less important than sustained progression (see Section 4). As an example, consider time until the first two consecutive visits to $\mathcal{S} = \{k : k > j\}$, with the aim of ultimately estimating the probability of two consecutive visits to states larger than $j$ before time $v$. To estimate these probabilities, one should replace $P = (p_{ij})$ with the working $(J+1) \times (J+1)$ transition matrix

$$Q_{j+} = \begin{pmatrix} p_{11} & \cdots & p_{1j} & p_{1(j+1)} & \cdots & p_{1J} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{j1} & \cdots & p_{jj} & p_{j(j+1)} & \cdots & p_{jJ} & 0 \\ p_{(j+1)1} & \cdots & p_{(j+1)j} & 0 & \cdots & 0 & \sum_{k>j} p_{(j+1)k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{J1} & \cdots & p_{Jj} & 0 & \cdots & 0 & \sum_{k>j} p_{Jk} \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{3.7}$$

Thus, an additional absorbing state is added to indicate the event of interest (see Mandel et. al. (2007) for other modifications). The $(k, J+1)$'th cell of $Q_{j+}^v$ is the probability that two consecutive visits to $\mathcal{S}$ occurred during the first $v$ transitions when the process started at $k$.

Prediction and calculation of variances or confidence intervals are based on the modified transition matrix, $Q_{j+}$. Letting $q_{l,k}^{(v)}(x, u, y_0; \vartheta)$ denote the $(l, k)$ element of $Q_{j+}^v$ under $\vartheta$ for $(X = x, U = u, Y_0 = y_0)$, the probability of two consecutive visits in states $> j$ before time $v$, given the process' history up to time $n$, is estimated by

$$\int_{-\infty}^{\infty} q_{y_n, J+1}^{(v-n)}(x, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u, y_0; \hat{\vartheta})}{\int_{-\infty}^{\infty} \prod_{t=1}^{n} p_{y_{t-1}, y_t}(x, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} g_0(u) du. \tag{3.8}$$

Equation (3.8) assumes that the event $\{$*two consecutive visits in states* $> j\}$ has not oc-

curred before visit $n$ (otherwise the probability is 1), or alternatively, that we are interested in the probability of that event from the present time to time $v$. The variance is calculated as described in Section 3.1 (see Appendix B for some technical details).

### 3.3 Models Without Random Effects

A model without random effects is obtained as a special case by setting $u \equiv 0$ and considering $g_0$ as degenerate at zero. The averaging over the random effects is not needed, making the calculation much simpler. Also, (3.3) reduces to (3.4), where the last observed state carries all important information of the history of the process. Interpretation of $\eta$ is now as a coefficient of the covariate $y_0^\bullet$, and may be omitted according to the specification of the model.

## 4 Illustration - Progression of Multiple Sclerosis

The data set analyzed here is part of a double-blinded phase III clinical trial that evaluated the utility of interferon beta-1a (Avonex) for MS patients with relapsing-remitting disease (Jacobs et.al., 1996). It includes all patients who were accrued early enough to complete two years of follow-up by the end of the study and who had brain MRI scans at baseline and yearly thereafter. As seen in Table 3 of Jacobs et al. (1996), the time to sustained progression distribution of this subgroup was the same as that of all study subjects. Visits were scheduled to be every six months, however, actual visits deviated slightly from the schedule. We used all visits that had a maximum discrepancy of 30 days from the schedule. This resulted in only 16 missed visits (2.5% of the total scheduled visits), quite a small number for such a complex study. The outcome of interest was time to sustained progres-

Table 1: Observed transitions of MS patients between EDSS scores.

| | Placebo | | | Avonex | | |
|---|---|---|---|---|---|---|
| | EDSS≤1.5 | EDSS=2,2.5 | EDSS≥3 | EDSS≤1.5 | EDSS=2,2.5 | EDSS≥3 |
| EDSS≤1.5 | 49 | 23 | 6 | 50 | 26 | 3 |
| EDSS=2,2.5 | 15 | 45 | 30 | 33 | 52 | 25 |
| EDSS≥3 | 4 | 21 | 124 | 1 | 21 | 79 |

sion, defined as the time to two consecutive visits with EDSS of at least one point greater than baseline.

The data set contains 66 individuals with a total of 290 transitions in the Avonex group and 72 individuals with 317 transitions in the placebo group. Due to the small number of transitions, we collapsed the EDSS values into three categories: EDSS≤ 1.5 (no disability), EDSS of 2 or 2.5 (mild disability), and EDSS≥ 3 (moderate to severe disability). The total number of transitions is summarized in Table 1.

Model (2.6) was fitted to the data with $J = 3$, $g_0$ the standard normal density and

$$p_{k,j}(x, u, y_0^\bullet; \vartheta) = \frac{\exp(\alpha_{kj} + \beta' x + \eta' y_0^\bullet + \sigma u)}{1 + \exp(\alpha_{kj} + \beta' x + \eta' y_0^\bullet + \sigma u)} - \frac{\exp(\alpha_{k(j-1)} + \beta' x + \eta' y_0^\bullet + \sigma u)}{1 + \exp(\alpha_{k(j-1)} + \beta' x + \eta' y_0^\bullet + \sigma u)}, \tag{4.1}$$

where $\alpha_{k0} = -\infty$ and $\alpha_{kJ} = \infty$ $(k = 1, 2, 3)$. This is a partial proportional odds model that does not constrain the baseline transition matrix (assigns $J - 1$ parameters for each row), but assumes proportional odds of covariates among all transitions. Parameters were estimated by SAS procedure NLMIXED, using the Dual Quasi-Newton algorithm with integrals evaluated by the default adaptive Gaussian quadrature. Convergence problems

were solved by first fitting models without random effects, and using the resulting estimates as initial values for the mixed effects models.

We first estimated the model with a treatment indicator as the only component of $x$. The estimated coefficient for treatment was 1.19, with estimated standard error of 0.48. Thus, interferon beta-1a significantly decreases the probability of worsening in disability. This finding is consistent with the results of the original study. We then estimated the model for each arm separately, testing for the influence of patient specific covariates (i.e., age, sex, disease duration, brain lesion volume and brain parenchymal fraction). None of these covariates showed a significant effect. We thus continued our analysis fitting two separate models (one for each arm) without any covariates.

Using the estimated transition matrices, we calculated the probability of two consecutive visits with EDSS higher than the baseline value as a function of time. This was the definition of sustained progression used by Jacobs et.al. (1996). We used the modification of the transition matrix given in (3.7) for the analysis.

Figure 1 depicts the probability of progression for a subject in her first visit as a function of time (six months units) stratified by arm and baseline EDSS. After two years (the end point of the original study), the probability of sustained progression among those who had EDSS of one at baseline is estimated as 0.44 and 0.54 for the case and control arms, respectively. For those having EDSS of two at baseline, the difference is more pronounced, being 0.27 for the interferon and 0.59 for the placebo patients. It appears that interferon prevents progression for people with mild disability better than for those having no disability. However, this may be related to the nature of the scale; a change from EDSS of one to two is considered a smaller step than a change from two to three.

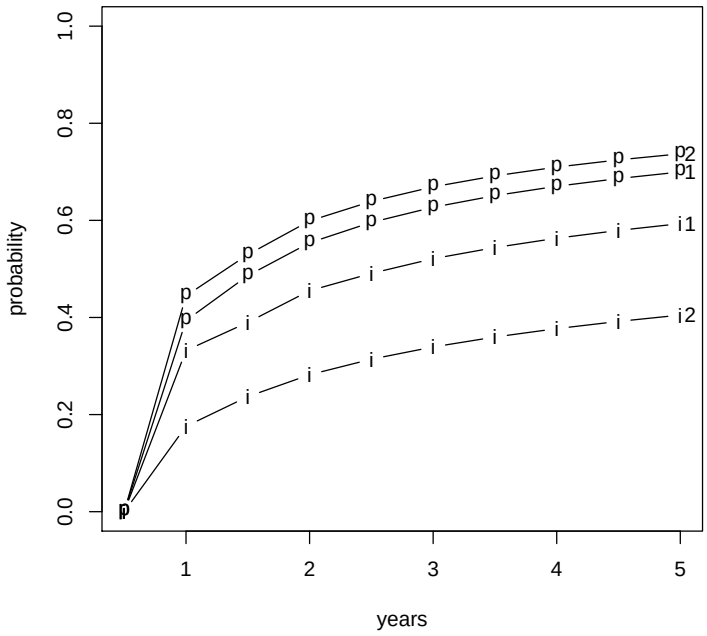


Figure 1: Probability of progression. Placebo arm denoted by p and interferon beta-1a arm denoted by i. The two curves for each arm show different level of baseline EDSS (1 and 2).

Jacobs et.al. (1996) estimated time to progression without conditioning on baseline EDSS. To generate a similar estimate, one can weigh the curves according to the probability of baseline EDSS. For example, in the interferon beta-1a arm there were 18 and 29 individuals with baseline EDSS of one and two, respectively, and the overall estimate of the probability of progression would use the weights 18/47 and 29/47. Estimating progression of individuals with EDSS of three or higher is impossible, because EDSS values greater than three were combined.

For comparison, models without random effects were fitted to the same data. Figure 2 presents the estimated progression curves and 95% pointwise confidence intervals for patients in the interferon beta-1a arm who had baseline EDSS of two, and the i2 curve

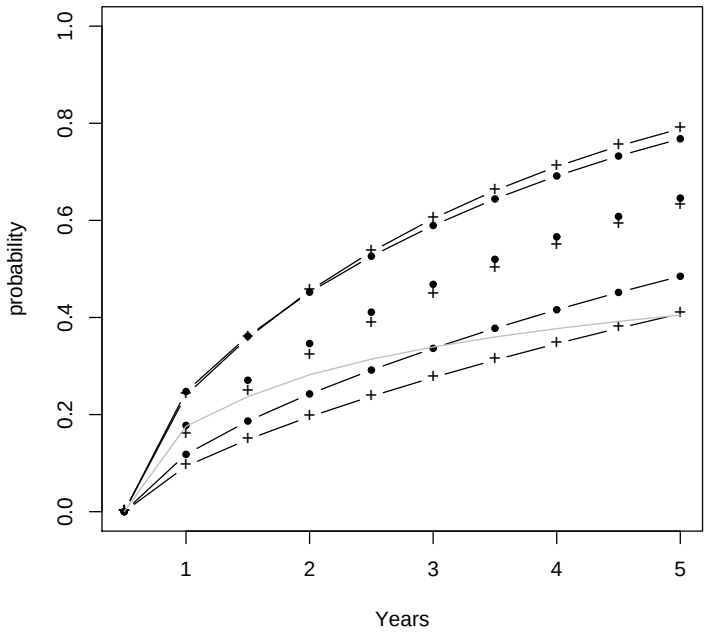


Figure 2: Probability of progression starting in state 2 of the interferon beta-1a arm under the fixed models. Pluses and circles denote estimates with and without conditioning on the baseline EDSS (state 2 in this graph). Lines are the corresponding estimates of the 95% pointwise confidence intervals. The gray line is the estimate under the mixed model (curve i2 of Figure 1).

of Figure 1 which is the analogous curve from the random effects model. The two curves represent models with (pluses) and without (circles) baseline EDSS as a covariate (see Section 3.3) and their estimated probabilities are quite similar. However, the probabilities are considerably larger than those predicted by the random effects model (gray curve). Under the random effects model, the estimated $\sigma^2$ values are 4.86 and 5.85 for the placebo and interferon beta-1a arms, respectively. This indicates that the heterogeneity in the data is large and supports the choice of the random effects model. Various other publications have found that progression is slower than that predicted by the models without random effects (e.g., Jacobs et.al., 1996, Weinshenker et.al., 1989).

To estimate prediction curves for patients who have history of EDSS, realizations of curves can be generated using the posterior distribution of the random effect (given history and covariates), as seen in (3.2). Such realizations represent the hypothetical population of curves that the patient specific curve comes from, and their mean is the probability of progression given the data, i.e., unconditional on the random effect. Depicting curves from the posterior distribution is a useful descriptive tool that helps to understand and to interpret the results. To illustrate that, Figure 3 depicts 100 curves for two hypothetical subjects in the interferon beta-1a group. The left panel represents a subject with baseline EDSS of two and without follow-up data (i.e., at the first visit). The curve on the right represents a subject after ten visits with EDSS history (2,2,3,2,3,2,1,2,2,2), and can be considered as a five year update of the progression curve for the subject on the left. The mean curves and 95% pointwise confidence intervals are depicted too. The variability of the curves on the right is much smaller than that on the left as a result of the additional information. The confidence intervals are not much smaller since they contain the sampling variability of the coefficient estimators. With increasing follow-up data on the same individual, the variability of the gray curves disappears and the graph shows the predicted (or estimated) probability of progression of a specific subject.

Figure 3 illustrates the heterogeneity in the course of the disease and indicates that subject-specific prediction is very difficult, almost impossible. It provides a nice platform for understanding the distinction between (3.1) and (3.5). The quantity (3.5) is essentially one of the gray curves appearing in the figure, while (3.1) is the average of the curves. Thus, (3.1) is a functional of the distribution of (3.5), and can be estimated consistently from the data.

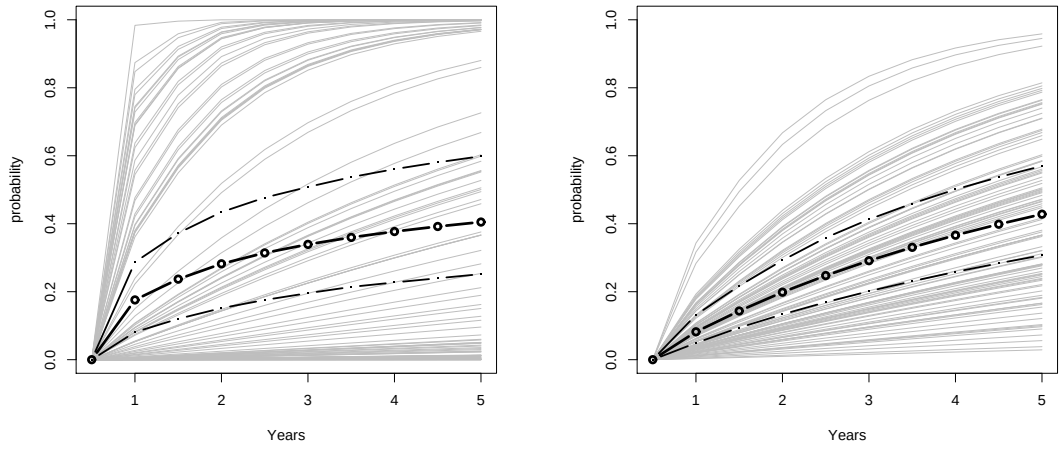


Figure 3: Distribution of curves of the probability of progression. One hundred realizations of the estimated model for the interferon beta-1a arm (gray lines) with the estimated mean curve and 95% confidence intervals based on the simulation method. Left - baseline EDSS of 2 no follow-up data; right - ten visits with observed data (2,2,3,2,3,2,1,2,2,2).

To generate realizations of progression curves for Figure 3, a sampling algorithm from the posterior $g(u|\{Y_t = y_t\}_{0\le t\le n}, X = x)$ is required. A natural choice is the MCMC algorithm. However, for the current purpose, a rather small number of independent realizations is needed and a simple alternative is to use a direct sampling. Let $c > \sup_u P(\{Y_t = y_t\}_{0\le t\le n}|X = x, u)$ (for practical purposes, it is enough to approximate $c$ by calculating $P(\{Y_t = y_t\}_{0\le t\le n}|X = x, u)$ on a fixed grid). The rejection/acceptance algorithm (e.g., Evans and Swartz, 2000) for generating one realization is then

1. Generate independently $u$ from $g_0$ and $v$ from $\mathrm{U}(0, 1)$.

2. If $P(\{Y_t = y_t\}_{0\le t\le n}|X = x, u) > cv$ stop and return $u$. Otherwise go back to step 1.

| visit | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_0 = 1$ | 0 | .422 | .586 | .692 | .759 | .805 | .838 | .863 | .882 | .897 |
| $y_0 = 2$ | 0 | .270 | .407 | .501 | .569 | .621 | .662 | .696 | .723 | .746 |

Table 2: Probabilities of time to two consecutive visits in a larger state.

# 5 Simulation

A small simulation study was conducted in order to examine the performance of the confidence intervals. Transitions between three states were generated according to the Markov model (4.1) with parameters $\alpha_{11} = 0$, $\alpha_{12} = 2.2$, $\alpha_{21} = -1.4$, $\alpha_{22} = 1.4$, $\alpha_{31} = -2.2$, $\alpha_{32} = -.85$, $\eta_2 = -.5$, $\eta_3 = -1$ and $\sigma = 1$ (no covariates). The derived probabilities of time to two consecutive visits in a larger state are listed in Table 2.

We considered two settings, both of which result in 300 transitions. The first was of 100 subjects each with 3 transitions and the second was of 20 subjects each with 15 transitions. In both settings, we sampled $Y_0$, with probability 1/3 assigned to each state. One thousand data sets were generated for each setting, and for each, we calculated confidence intervals for the probabilities of time to two consecutive visits in state 3 starting in state 2. We compared the confidence intervals based on the analytical delta method to those obtained by the simulation method described at the end of Section 3.1. For the latter approach, we used 1000 replications and calculated both normal based intervals using the variance, and percentile confidence intervals. We also calculated intervals for models without random effects to illustrate the consequences of model misspecification.

Since the parameters of interest are probabilities, we used the log(-log) transformation before generating the confidence intervals. This method is commonly used in survival

analysis (Kalbfleisch and Prentice, 2002) and has improved performance in the fixed effects model (Mandel, et al., 2007). Thus, the limits of our level $\alpha$ pointwise confidence intervals are $\hat{p}^{\exp\{\pm z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{p})}/[\hat{p}\log(\hat{p})]\}}$, where $\hat{p}$ and $\hat{\mathrm{Var}}(\hat{p})$ are the estimates of the parameter and its variance, and $z_\alpha$ is the $\alpha$ percentile of the standard normal distribution.

The maximum likelihood estimate of $\sigma^2$ was 0 in 120 and 18 data sets for the setting of 100 and 20 subjects, respectively. In these cases, the estimated model is exactly the fixed effects model. In evaluating the confidence intervals, we report the results for those data sets with $\hat{\sigma}^2 > 0$. In addition, the information matrix of $\vartheta$ was not positive definite in one of the data sets with 20 subjects. This data set was excluded from the evaluation of the confidence intervals. The proportion of intervals covering the true parameters for data sets with estimated $\sigma^2 > 0$ are listed in Table 3. The confidence intervals with 100 subjects perform well, whereas those for 20 subjects are anti-conservative. This is probably a result of the poor normal approximation for the distribution of the fixed effect estimators in data with few subjects. Another feature of the confidence intervals is their uneven distribution on the left and right of the true value, where most intervals that do not include the true parameter assign values that are too small. This is less pronounced in the percentile method, and suggests that it is mostly related to the linear approximation of the delta method. The similarity between the analytical derivative and the simulation based approach for the 100 subjects setting is remarkable. The simulation approach demands more computer time, but saves the burden of calculating and programming complicated derivatives. It also has the additional merit of eliminating the linear approximation of the delta method.

Table 4 presents the results of fitting models without random effects for the same

simulations reported in Table 3. As in Figure 2, both models, with and without $Y_0$ as a covariate, were fitted. Overall the confidence intervals perform poorly. In the setting of 100 subjects, there is a tendency of overestimation which is consistent with the finding of the data analysis in Section 4 (see Figure 2).

# 6 Discussion

We have presented prediction and confidence intervals estimation in the mixed effects Markov model framework, and have provided several graphical tools that help to interpret model results. These graphs, and especially Figure 3, indicate how heterogeneous is MS and provide a useful tool for better understanding the course of the disease. Although we have introduced our models for time independent covariates, they can be extended to time varying covariates. Letting $x_v$ be the values of the covariates as measured at visit $v$, (2.6) is replaced by

$$\mathcal{L}_i = \int_{-\infty}^{\infty} \prod_{v=1}^{n} p_{y_{v-1},y_v}(\beta' x_{v-1} + \sigma u + \eta' y_0^{\bullet}; \alpha) g_0(u) du, \tag{6.1}$$

and estimation of $\vartheta$ follows, similar to this extension in models without random effects (Mandel, et al., 2007). If the covariate process is external to the Markov process and its future is known at time $v$, then prediction can be done as in (3.3) by

$$\begin{aligned}
&\hat{P}(Y_v = y_v | \{Y_t = y_t\}_{0 \le t \le n}, \{x_t\}_{t=0}^{v-1}) \\
&= \int_{-\infty}^{\infty} p_{y_n,y_v}^{(v-n)}(\{x_t\}_{t=n}^{v-1}, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^{n} p_{y_{t-1},y_t}(x_{t-1}, u, y_0; \hat{\vartheta}) g_0(u)}{\int_{u^*} \prod_{t=1}^{n} p_{y_{t-1},y_t}(x_{t-1}, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*} du,
\end{aligned}$$

where $p_{y_n,y_v}^{(v-n)}(\{x_t\}_{t=n}^{v-1}, u, y_0; \hat{\vartheta})$ is the $(y_n, y_v)$ element of the transition matrix $P(x_n, u, y_0; \vartheta) \times P(x_{n+1}, u, y_0; \vartheta) \times \cdots \times P(x_{v-1}, u, y_0; \vartheta)$. The variance can be estimated by the simulation algorithm discussed in Section 3.1.

Table 3: Results of simulations fitting the random effects model. Entries are the proportions of replications in which the true parameter was to the left and to the right of the calculated confidence interval. The methods compared are the delta method (delta), the simulation method using the variance (simulation V) and the simulation method using the percentiles (simulation P)

| visit | 100 subjects with 3 transitions | | | | | | 20 subjects with 15 transitions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | delta | | simulation V | | simulation P | | delta | | simulation V | | simulation P | |
| | left | right | left | right | left | right | left | right | left | right | left | right |
| 2 | .017 | .030 | .018 | .028 | .016 | .018 | .019 | .064 | .022 | .062 | .023 | .058 |
| 3 | .017 | .036 | .018 | .039 | .016 | .025 | .018 | .063 | .021 | .068 | .029 | .053 |
| 4 | .016 | .036 | .017 | .035 | .018 | .023 | .021 | .062 | .019 | .066 | .032 | .049 |
| 5 | .010 | .038 | .010 | .038 | .020 | .020 | .024 | .066 | .015 | .070 | .034 | .047 |
| 6 | .005 | .040 | .007 | .041 | .022 | .016 | .021 | .066 | .012 | .070 | .038 | .043 |
| 7 | .002 | .043 | .006 | .043 | .018 | .017 | .020 | .065 | .011 | .068 | .044 | .038 |
| 8 | .000 | .041 | .001 | .043 | .017 | .016 | .020 | .063 | .008 | .064 | .044 | .036 |
| 9 | .000 | .042 | .001 | .044 | .014 | .016 | .018 | .054 | .007 | .065 | .045 | .032 |
| 10 | .000 | .045 | .000 | .048 | .013 | .014 | .019 | .054 | .004 | .065 | .054 | .030 |

Table 4: Results of simulations fitting the fixed effects model for replications having $\hat{\sigma}^2 > 0$. Entries are the proportions of replications in which the true parameter was to the left and to the right of the calculated confidence interval. Calculation of confidence intervals using the delta method.

| visit | 100 subjects | | | | 20 subjects | | | |
|---|---|---|---|---|---|---|---|---|
| | w/o $y_0$ | | with $y_0$ | | w/o $y_0$ | | with $y_0$ | |
| | left | right | left | right | left | right | left | right |
| 2 | .000 | .089 | .008 | .073 | .027 | .444 | .020 | .233 |
| 3 | .001 | .051 | .017 | .049 | .033 | .337 | .034 | .189 |
| 4 | .011 | .030 | .024 | .032 | .032 | .240 | .038 | .162 |
| 5 | .027 | .015 | .035 | .023 | .028 | .165 | .042 | .128 |
| 6 | .058 | .005 | .061 | .015 | .010 | .099 | .053 | .098 |
| 7 | .113 | .000 | .092 | .008 | .000 | .062 | .060 | .085 |
| 8 | .178 | .000 | .123 | .005 | .000 | .040 | .076 | .077 |
| 9 | .269 | .000 | .149 | .003 | .000 | .027 | .093 | .062 |
| 10 | .355 | .000 | .184 | .003 | .000 | .015 | .112 | .047 |

MS is known to be a heterogenous disease. Some patients experience no progression for ten or more years (benign MS) and others experience a fast and continuous progression from onset (primary progressive MS). The patients used in this paper are relatively homogenous, since enrollment was subjected to the strict criteria of a phase III clinical trial. Nonetheless, our results indicate that heterogeneity is present and it is significant. Prediction of the random effect $U$ would be too ambitious with the typical short follow-up data on each subject, as well illustrated by Figure 3. In other contexts, however, predicting $U$ and calculating the mean squared error of prediction, using the approach of Booth and Hobert (1998), is of both theoretical and applied interest and is a topic of current research.

# References


[1] AGRESTI, A. (2002). *Categorical Data Analysis* (second edition), Wiley & Sons (New Jersey).

[2] AITKIN, M., and ALFÓ, M. (1998). Regression models for binary longitudinal responses, *Statistics and Computing*, **8**, 289-307.

[3] AITKIN, M., and ALFÓ, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modelling, *Statistical Modelling*, **3**, 291-303.

[4] ALBERT, P. S. (1994). A Markov model for sequences of ordinal data from a relapsing-remitting disease, *Biometrics*, **50**, 51-60.

[5] ALBERT, P. S., and FOLLMANN, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica*, **57**, 100-111.

[6] ALBERT, P. S., and WACLAWIW, M. A. (1998). A two-state markov chain for heterogeneous transitional data: a quasi-likelihood approach, *Statistics in Medicine*, **17**, 1481-1493.

[7] ALFÓ, M., and AITKIN, M. (2000). Random coefficient models for binary longitudinal responses with attrition, *Statistics and Computing*, **10**, 279-287.

[8] BOOTH, J. G., and HOBERT, J. P. (1998). Standard errors of prediction in generalized linear mixed models, *Journal of the American Statistical Association*, **93**, 262-272.

[9] COOK, R. J. (1999). A mixed model for two-state Markov processes under panel observation, *Biometrics*, **55**, 915-920.

[10] COOK, R. J., and NG, E. T. M. (1997). A logistic-bivariate normal model for overdispersed two-state Markov processes, *Biometrics*, **53**, 358-364.

[11] COOK, R. J., YI, G. Y., LEE, K. A., and GLADMAN, D. D. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation, *Biometrics*, **60**, 436-443.

[12] DIGGLE, P., HEAGERTY, P., LIANG, K. Y., and ZEGER, S. L. (2002). *Analysis of longitudinal data* (second edition), Oxford University Press, (Oxford).

[13] Evans, M., and Swartz, T. (2000). *Approximating Integrals Via Monte Carlo and Deterministic Methods*, Oxford University Press, (Oxford).

[14] JACOBS, L. D., COOKFAIR, D. L., RUDICK, R. A., HERNDON, R. M., RICHERT, J. R., SALAZAR, A. M., FISCHER, J. S., GOODKIN, D. E., GRANGER, C. V.,

SIMON J. H., and others (1996). Intramuscular interferon beta-1 alpha for disease progression in relapsing multiple sclerosis, *Annals of Neurology*, **39**, 285-294.

[15] JIANG, J. M. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models, *Journal of Statistical Planning and Inference*, **111**, 117-127.

[16] JIANG, J., and LAHIRI, P. (2006). Mixed model prediction and small area estimation (with discussion), *Test*, **15**, 1-96.

[17] 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.

[18] KALBFLEISCH, J. D., and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data* (second edition), Wiley & Sons (New Jersey).

[19] LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D., and SCHABENBERBER, O. (2006). *SAS for Mixed Models* (second edition), SAS Publishing.

[20] MANDEL, M., GAUTHIER, S. A., GUTTMANN. C. R. G., WEINER, H. L., and BETENSKY, R. A. (2007). Estimating time to event from longitudinal categorical data: an analysis of multiple sclerosis progression *Journal of the American Statistical Association*, forthcoming.

[21] ROBINSON, G. K. (1991). That BLUP is a good thing: the estimation of random efffects (with discussion), *Statistical Science*, **6**, 15-32.

[22] WEINSHENKER, B. G., BASS, B., RICE, G. P. A., NOSEWORTHY, J., CARRIERE, W., BASKERVILLE, J., and EBERS, G. C. (1989). The natural history of multiple sclerosis: a geographically based study 2. Predictive value of the early clinical course, *Brain*, **112**, 1419-1928.

# A Calculation of the Variance of Model (4.1)

To calculate the first part on the right hand side of (3.6), consider the more general map $M_v(\text{vec}(P)) = \text{vec}(P^v)$ from $\mathbb{R}^{J^2}$ to $\mathbb{R}^{J^2}$. Let $\Delta_{kj}$ by a $J \times J$ matrix whose entries are all zeros except the $(k,j)$'th which is one. Then, the $(k-1)J+j$ column of $\frac{\partial}{\partial \text{vec}(P)} M_v(\text{vec}(P))$ is given by

$$\text{vec}\Big(\sum_{\ell=0}^{v-1} P^{\ell} \Delta_{kj} P^{v-1-\ell}\Big), \tag{A.1}$$

where $P^0 = \mathbf{I}_J$ is the identity matrix of order $J$ (Mandel et.al).

For the second part on the right hand side of (3.6), let

$$F(j,k,x,u,y_0^{\bullet};\vartheta) = \sum_{j' \le j} p_{k,j'}(x,u,y_0;\vartheta) = \frac{\exp(\alpha_{kj} + \beta' x + \eta' y_0^{\bullet} + \sigma u)}{1 + \exp(\alpha_{kj} + \beta' x + \eta' y_0^{\bullet} + \sigma u)}$$

for $j = 1, \ldots, J-1$ , $k = 1, \ldots, J$ and denote by $V = diag(F(j,k,\ldots)[1 - F(j,k,\ldots)])$ the $J(J-1) \times J(J-1)$ matrix containing the binomial variances in its diagonal (in the alphabetical order of $kj$, i,e, $11, 12, \ldots, 1J-1, 21, 22, \ldots, 2J-1, \ldots$). Let $A$ be a $J \times (J-1)$ matrix with $(i,i)$ elements equal to 1, $(i, i-1)$ elements equal -1 and all other entries 0. For example, for $J = 3$

$$A = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Let $B = \mathbf{I}_J \otimes A$, where $\otimes$ denotes the Kronecker product, and denote by $\mathbf{1}_a$ the vector of ones of length $a$, then

$$\frac{\partial}{\partial \vartheta}\text{vec}(P(x,u;\vartheta)) = \Big(BV, (BV\mathbf{1}_{J(J-1)}) \otimes (z',u)\Big), \tag{A.2}$$

which is the second part of the derivative as appeared on the right hand side of (3.6).

Equations (A.1) and (A.2) give the derivatives in terms of the elements of $P$ while the working transition matrix is usually a modification of $P$ as discussed in Section 3.2. The application (3.6) of the derivative of the transition probabilities can be used with all references to $P$ replaced with the modified transition matrix, say $Q$, and in particular, (A.1) needs no other changes. The matrix $\frac{\partial}{\partial \vartheta}\text{vec}(Q)$ is obtained from the $J^2 \times m$ matrix $\frac{\partial}{\partial \vartheta}\text{vec}(P)$ by summing the appropriate terms using rows of zeros for the designed 0 and 1's cells. As an example, consider the matrix $Q_{j+}$ defined in (3.7). Let

$$C_1 = \begin{pmatrix} \mathbf{I}_J \\ \mathbf{0}_{1,J} \end{pmatrix}, \quad C_2 = \begin{pmatrix} \mathbf{I}_j & \mathbf{0}_{j,J-j} \\ \mathbf{0}_{J-j,j} & \mathbf{0}_{J-j,J-j} \\ \mathbf{0}_{1,j} & \mathbf{1}'_{J-j} \end{pmatrix},$$

be $(J+1) \times J$ matrices, where $\mathbf{0}_{r,c}$ is a zero matrix with $r$ rows and $c$ columns, and define the $(J+1)^2 \times J^2$ matrix

$$C = \begin{pmatrix} \mathbf{I}_j \otimes C_1 & \mathbf{0}_{j(J+1),J(J-j)} \\ \mathbf{0}_{(J+1)(J-j),jJ} & \mathbf{I}_{J-j} \otimes C_2 \\ \mathbf{0}_{(J+1),jJ} & \mathbf{0}_{(J+1),J(J-j)} \end{pmatrix}$$

then

$$\frac{\partial}{\partial \vartheta}\text{vec}(Q_{j+}) = C\frac{\partial}{\partial \vartheta}\text{vec}(P),$$

which for the partial proportional odds model reduces to (see (A.2))

$$\frac{\partial}{\partial \vartheta}\text{vec}(Q_{j+}) = C\Big(BV, (BV\mathbf{1}_{J(J-1)}) \otimes (z',u)\Big).$$

# B Derivative of (3.3)

Letting

$$h(y_v, \{y_t\}_{0\le t\le n}, x, u, \vartheta) = p^{(v-n)}_{y_n,y_v}(x, u, y_0; \hat{\vartheta}) \frac{\prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \hat{\vartheta})}{\int_{u^*} \prod_{t=1}^n p_{y_{t-1},y_t}(x, u^*, y_0; \hat{\vartheta}) g_0(u^*) du^*},$$

and assuming differentiation under the integral is permitted, we have

$$\begin{aligned}
&\frac{\partial}{\partial\vartheta} P_\vartheta(Y_v = y_v | \{Y_t = y_t\}_{0\le t\le n}, X = x) \\
&\quad = \int_{-\infty}^{\infty} \frac{\partial}{\partial\vartheta}\, h(y_v, \{y_t\}_{0\le t\le n}, x, u, \vartheta) g_0(u) du \\
&\quad = \int_{-\infty}^{\infty} h(y_v, \{y_t\}_{0\le t\le n}, x, u, \vartheta) \frac{\partial}{\partial\vartheta} \log\{h(y_v, \{y_t\}_{0\le t\le n}, x, u, \vartheta)\} g_0(u) du.
\end{aligned}$$

Next,

$$\begin{aligned}
\frac{\partial}{\partial\vartheta} \log\{h(y_v, \{y_t\}_{0\le t\le n}, x, u, \vartheta)\} \;=\; & \frac{\partial}{\partial\vartheta} \left[ \log\{p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta)\} + \sum_{t=1}^n \log\{p_{y_{t-1},y_t}(x, u, y_0; \vartheta)\} \right] \\
& + \frac{\partial}{\partial\vartheta} \log\{ \int_{u^*} \prod_{t=1}^n p_{y_{t-1},y_t}(x, u^*, y_0; \vartheta) g_0(u^*) du^* \}. \qquad \text{(B.1)}
\end{aligned}$$

Each of the components of the first two terms on the right hand side of (B.1) has the form

$$\frac{\partial}{\partial\vartheta} \log\{p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta)\} = [p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta)]^{-1} \frac{\partial}{\partial\vartheta} p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta).$$

To calculate the third term on the right hand side of (B.1), differentiate again under the integral sign using

$$\frac{\partial}{\partial\vartheta} \log\{ \int_u \prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \vartheta) g_0(u) du \} = \frac{\int_u \frac{\partial}{\partial\vartheta} \prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \vartheta) g_0(u) du}{\int_u \prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \vartheta) g_0(u) du}$$

and

$$\frac{\partial}{\partial\vartheta} \prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \vartheta) = \prod_{t=1}^n p_{y_{t-1},y_t}(x, u, y_0; \vartheta) \times \sum_{t=1}^n [p_{y_{t-1},y_t}(x, u, y_0; \vartheta)]^{-1} \frac{\partial}{\partial\vartheta} p_{y_{t-1},y_t}(x, u, y_0; \vartheta).$$

Thus, the basic term needed for the calculation is

$$\frac{\partial}{\partial\vartheta} p^{(v-n)}_{y_n,y_v}(x, u, y_0; \vartheta) = \frac{\partial}{\partial \mathrm{vec}(P)} p^{*(v-n)}(y_n, y_v, P) \frac{\partial}{\partial\vartheta} \mathrm{vec}(P(x, u, y_0; \vartheta))$$

(see (3.6)), and can be calculated as illustrated in Appendix A.

A great simplification is obtained by noticing that $p^*(k,j,P)$ is just the $(k,j)$'th cell of $P$, thus, $\frac{\partial}{\partial \text{vec}(P)} p^*(k,j,P) = [\text{vec}(\Delta_{kj})]'$, where $\Delta_{kj}$ is a $J \times J$ matrix whose elements are zeros except the $(k,j)$ cell which is 1.

To calculate the variance via the delta method, the derivative should be evaluated at the MLE $\hat{\vartheta}$. This is done by repeatedly generating $u$ from $g_0$ and calculating the derivative or by any other numerical integration methods.

To calculate the derivative of (3.8), the same steps as described above are applied, but with $q(\cdot)$ instead of $p(\cdot)$ in the first term on the right hand side of (B.1). Calculation of $\frac{\partial}{\partial \vartheta} q^{(v-n)}_{y_n,y_v}(x,u,y_0;\vartheta)$ is a transformation of $\frac{\partial}{\partial \vartheta} p^{(v-n)}_{y_n,y_v}(x,u,y_0;\vartheta)$ and described in Appendix A.