BIOMETRICS 57, 829–836
September 2001

Sequential Ordinal Modeling with Applications to Survival Data

James H. Albert

Department of Mathematics and Statistics, Bowling Green State University,
Bowling Green, Ohio 43403, U.S.A.
email: albert@bgnet.bgsu.edu

and

Siddhartha Chib

John M. Olin School of Business, Washington University,
One Brookings Drive, St. Louis, Missouri 63130, U.S.A.
email: chib@olin.wustl.edu

SUMMARY. This paper considers the class of sequential ordinal models in relation to other models for ordinal response data. Markov chain Monte Carlo (MCMC) algorithms, based on the approach of Albert and Chib (1993, *Journal of the American Statistical Association* **88,** 669–679), are developed for the fitting of these models. The ideas and methods are illustrated in detail with a real data example on the length of hospital stay for patients undergoing heart surgery. A notable aspect of this analysis is the comparison, based on marginal likelihoods and training sample priors, of several nonnested models, such as the sequential model, the cumulative ordinal model, and Weibull and log-logistic models.

KEY WORDS: Bayes factor; Cumulative ordinal probit and logit model; Discrete hazard function; Gibbs sampling; Marginal likelihood; Metropolis–Hastings algorithm; Model comparison; Nonnested models; Sequential ordinal probit and logit; Training sample prior.

1. Introduction

Ordinal data are often analyzed using the cumulative ordinal regression model (McCullagh, 1980). In this model, the observed ordinal response Y is derived from a continuous-valued latent variable z by a threshold specification where the latent data are modeled by a regression on covariates. This model is appropriate when one can assume that a single unobservable continuous variable underlies the ordinal response. For example, this representation can be used for modeling ordinal letter grades in a mathematics class by assuming that the grade is an indicator of the student's unobservable continuous-valued intelligence. Albert and Chib (1993) and Johnson and Albert (1999) describe Bayesian fitting algorithms and residual methods for the cumulative model by exploiting the latent variable representation.

In other circumstances, however, a different latent variable structure may be required to model the ordinal response. For example, McCullagh (1980) and Fahrmeir and Tutz (1994) consider a data set where the relative tonsil size of children is classified into the three states: present but not enlarged, enlarged, and greatly enlarged. The objective is to explain the ordinal response as a function of whether the child is a carrier of the *Streptococcus pyogenes.* In this setting, it may not be appropriate to model the ordinal tonsil size in terms of a single latent variable. Rather, it may be preferable to imagine that each response is determined by two continuous latent variables where the first latent variable measures the propensity of the tonsil to grow abnormally and pass from the state present but not enlarged to the state enlarged. The second latent variable measures the propensity of the tonsil to grow from the enlarged to the greatly enlarged states. Thus, the final greatly enlarged state is realized only when the tonsil has passed through the earlier levels. This latent variable representation leads to a sequential model because the levels of the response are achieved in a sequential manner.

As our primary example, we consider data on the length of hospital stay for a thousand patients undergoing heart surgery. The ordinal response variable is the number of days of hospital stay. A relatively small number of patients stay longer than 12 days, and for these patients, the response variable is defined, for convenience, to equal 12. Associated with the response variable are the following eight covariates that might influence the length of stay: AGE (in years), GENDER, RACE (white or nonwhite), PREVINS (if primary payer is private insurance), INDIG (if this is a charity care by the hospital), COMORB (number of patient comorbidities), CAMB (if the patient had a coronary artery bypass graft), and PTCA (if the patient had an angioplasty). In this particular example, the length of stay is censored if the patient dies in the hospital because then one does not know what the length of

the patient's stay would have been if he or she had lived. It may be possible to avoid the censoring problem by defining the response to be the length of time until death, but in that case, it would be more difficult to interpret the response as the recovery time from heart surgery. For these data, the sequential model is a plausible alternative to the cumulative model since a long observed hospital stay requires passage through shorter hospital stays. One latent variable may be used to represent the patient's propensity to pass through a short stay into a moderate stay, a second latent variable to represent the patient's propensity to pass from a moderate stay to a longer stay, and so on.

Suppose that one observes independent observations $Y_1, \ldots, Y_n$, where each Y_i is an ordinal categorical response variable with J possible values $\{1, \ldots, J\}$. Associated with the ith response Y_i, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ denote a set of p covariates. The cumulative model (McCullagh, 1980) is defined by the distribution

$$\Pr(Y \le j \mid \boldsymbol{\gamma}, \boldsymbol{\delta}) = F(\gamma_j^c - \mathbf{x}_i'\boldsymbol{\delta}), \qquad j = 1, \ldots, J, \quad (1)$$

where $F(\cdot)$ is a specified cumulative distribution function, $\boldsymbol{\delta}$ is the regression parameter vector, and $\boldsymbol{\gamma}^c = (\gamma_1^c, \ldots, \gamma_{J-1}^c)$ are the ordered category specific cutpoints or thresholds. This model can be derived by assuming the existence of a latent variable z_i generated according to the model $z_i = \mathbf{x}_i'\boldsymbol{\delta} + \epsilon_i$,, where ϵ_i is distributed according to F. The ordered cutpoints $\boldsymbol{\gamma}^c$ categorize the continuous variable z_i; $Y_i = j$ is observed if $\gamma_{j-1}^c < z_i < \gamma_j^c$. Albert and Chib (1993) use this latent variable representation to develop a Bayesian fitting approach for the ordinal model.

In the sequential model, the variable Y_i can take the value j only after the levels $1, \ldots, j-1$ are reached. In other words, to get to the outcome j, one must pass through levels $1, 2, \ldots, j-1$ and stop (or fail) in level j. The probability of stopping in level j $(1 \le j \le J-1)$, conditional on the event that the jth level is reached, is given by

$$\Pr(Y_i = j \mid Y_i \ge j, \boldsymbol{\gamma}, \boldsymbol{\delta}) = F(\gamma_j - \mathbf{x}_i'\boldsymbol{\delta}), \quad (2)$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})$ are unordered cutpoints and $\mathbf{x}_i'\boldsymbol{\delta}$ represents the effect of the covariates. This probability function is referred to as the discrete-time hazard function (Tutz, 1990, 1991; Fahrmeir and Tutz, 1994). It follows that the probability of stopping at level j is given by

$$\Pr(Y_i = j \mid \boldsymbol{\gamma}, \boldsymbol{\delta}) = F(\gamma_j - \mathbf{x}_i'\boldsymbol{\delta}) \prod_{k=1}^{j-1} \{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\}, \qquad j \le J-1, \quad (3)$$

whereas the probability that the final level J is reached is

$$\Pr(Y_i = J \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) = \prod_{k=1}^{J-1} \{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\} \quad (4)$$

since the event $Y_i = J$ occurs only if all previous $J-1$ levels are passed.

The sequential model can also be formulated in terms of latent variables. Corresponding to the ith observation, define latent variables $\{w_{ij}\}$, where $w_{ij} = \mathbf{x}_i'\boldsymbol{\delta} + e_{ij}$ and the e_{ij} are independently distributed from F. We observe $Y_i = 1$ if $w_{i1} \le \gamma_1$ and observe $Y_i = 2$ if the first latent variable $w_{i1} > \gamma_1$ and the second variable $w_{i2} \le \gamma_2$. In general, we observe $Y_i = j$ $(1 \le j \le J-1)$ if the first $j-1$ latent variables exceed their corresponding cutoffs and the jth variable does not: $Y_i = j$ if $w_{i1} > \gamma_1, \ldots, w_{ij-1} > \gamma_{j-1}, w_{ij} \le \gamma_j$. In this model, the latent variable w_{ij} represents one's propensity to continue to the $(j+1)$st level in the sequence, given that the individual has already attained level j.

The sequential model is formally equivalent to the continuation-ratio ordinal models, discussed by Agresti (1990, Chapter 9), Cox (1972), and Ten Have and Uttal (1994). It is useful here to view these models from the sequential perspective since, in some applications, observations move through the ordered outcomes sequentially and the sequential perspective motivates an attractive fitting procedure, as described below.

The sequential model is useful in the analysis of discrete-time survival data (Kalbfleish and Prentice, 1980; Fahrmeir and Tutz, 1994, Chapter 9). More generally, the sequential model can be used to model nonproportional and nonmonotone hazard functions and to incorporate the effect of time-dependent covariates. Suppose that one observes the time to failure for subjects in the sample, and let the time interval be subdivided (or grouped) into the J intervals $[a_0, a_1)$, $[a_1, a_2), \ldots, [a_{J-1}, \infty)$. We define the event $Y_i = j$ if a failure is observed in the interval $[a_{j-1}, a_j)$. The discrete hazard function (2) now represents the probability of failure in the time interval $[a_{j-1}, a_j)$ given survival until time a_{j-1}. The vector of cutpoint parameters $\boldsymbol{\gamma}$ represents the baseline hazard of the process.

This article presents a Bayesian analysis of the sequential probit model. Section 2 describes classical and Bayesian fitting methods for the sequential model. The Bayesian fitting is based on the latent variable/Markov chain Monte Carlo sampling approach of Albert and Chib (1993). Section 3 discusses generalizations of the sequential model, relevant for survival data, which differ by the specification of the baseline hazard function, the choice of covariates, and the effect of the covariates on the hazard function. The issue of model comparison is taken up in Section 4, where the approach of Chib (1995) and Chib and Jeliazkov (2001) is used to find the marginal likelihood of the sequential model. An extended analysis of the length of hospital stay dataset is presented in Section 5.

2. Fitting of the Sequential Model

2.1 *Likelihood Function and Classical Fitting*

Suppose that the hazard function has the form (2) and let $\mathbf{y} = (y_1, \ldots, y_n)$ represent the vector of observed ordinal responses. Then the likelihood function of the vector of cutpoints $\boldsymbol{\gamma}$ and the regression vector $\boldsymbol{\delta}$ is given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \prod_{i:y_i<J} \left[F(\gamma_{y_i} - \mathbf{x}_i'\boldsymbol{\delta}) \prod_{k=1}^{y_i-1} \left\{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\right\} \right] \times \prod_{i:y_i=J} \left[\prod_{k=1}^{J-1} \left\{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\right\} \right]. \quad (5)$$

In the context of failure-time data, it can happen that the survival times are not observed for all individuals. In the hos-

pital stay application, patients may die during the observation period and the actual length of hospital stay, if they had lived, is not known. In this case, these particular observations are censored from the right. If a particular patient, say i, dies in interval $[a_{j-1}, a_j)$, then (under the assumption that censoring occurs at the start of the interval) we only know that $Y_i \geq j-1$. To deal with this situation, let d_i be a noncensoring indicator that takes the value one if the ith patient's observation is not censored and is equal to zero otherwise. The probability of observing a censored outcome $(d_i = 0, Y_i = j)$ is given by

$$\Pr(Y_i = j, d_i = 0 \mid \boldsymbol{\delta}, \boldsymbol{\gamma}) = \prod_{k=1}^{j-1} \left\{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\right\}. \quad (6)$$

The likelihood function that incorporates the censoring resembles (5) with the exception that the ith component of the likelihood corresponding to a censored observation is replaced by $\Pi_{k=1}^{j-1}\{1 - F(\gamma_k - \mathbf{x}_i'\boldsymbol{\delta})\}$.

Fahrmeir and Tutz (1994) show that sequential ordinal models are a special case of multivariate generalized linear models (McCullagh and Nelder, 1989), and maximum likelihood (ML) estimates can be computed using an iterative reweighted least-squares algorithm. The goodness of fit of the model is judged by the computation of the Pearson or deviance statistic. A difference in deviances statistic is used to compare the fit of nested models.

There are questions about the interpretation of classical test statistics, particularly in small samples. Both the Pearson and deviance statistics have asymptotic chi-squared distributions only in the grouped data situation where the sizes of the groups tend to infinity at specific rates. The difference in deviance statistics is suitable only for comparing nested models of the same type. The Bayesian marginal likelihood/Bayes factor criterion developed in Section 4 provides a more unified perspective for comparing nonnested models.

2.2 *Latent Variable Representation and MCMC Fitting*

If $(\boldsymbol{\gamma},\boldsymbol{\delta})$ is assigned the prior $\pi(\boldsymbol{\gamma}, \boldsymbol{\delta})$, the posterior density is proportional to $\pi(\boldsymbol{\gamma}, \boldsymbol{\delta})L(\boldsymbol{\gamma}, \boldsymbol{\delta})$. To summarize this posterior density, we utilize the framework of Albert and Chib (1993) and incorporate the latent variables in the MCMC sampling.

Note that the latent variable representation outlined in Section 1 can be simplified by incorporating the cutpoints $\{\gamma_j\}$ into the mean function. Define the new latent variable $z_{ij} = w_{ij} - \gamma_j$. Then it follows that $z_{ij} \mid \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{x}_{ij}'\boldsymbol{\beta}, 1)$, where $\mathbf{x}_{ij}' = (0, 0, \ldots, -1, 0, 0, \mathbf{x}_i')$ with -1 in the jth column, and $\boldsymbol{\beta} = (\gamma_1, \ldots, \gamma_{J-1}, \boldsymbol{\delta}')'$. The observed data are then generated according to

$$Y_i = \begin{cases} 1 & \text{if } z_{i1} \leq 0 \\ 2 & \text{if } z_{i1} > 0, z_{i2} \leq 0 \\ \vdots & \vdots \\ J-1 & \text{if } z_{i1} > 0, \ldots, z_{iJ-2} > 0, z_{iJ-1} \leq 0 \\ J & \text{if } z_{i1} > 0, \ldots, z_{iJ-1} > 0. \end{cases} \quad (7)$$

To estimate this model by MCMC methods, we simulate the joint posterior distribution of $(\{z_{ij}\}, \boldsymbol{\beta})$ by sequentially sampling from the two conditional distributions $[\{z_{ij}\} \mid \mathbf{y}, \boldsymbol{\beta}]$, $[\boldsymbol{\beta} \mid \{z_{ij}\}]$. The latent data $\{z_{ij}\}$, conditional on $(\mathbf{y}, \boldsymbol{\beta})$, are independently distributed as truncated normal. For example, if $y_i = 3$ and $J = 4$, then $[z_{i1} \mid y_i = 3, \boldsymbol{\beta}] \sim \mathcal{TN}_{(0,\infty)}(\mathbf{x}_{i1}'\boldsymbol{\beta}, 1)$; $[z_{i2} \mid y_i = 3, \boldsymbol{\beta}] \sim \mathcal{TN}_{(0,\infty)}(\mathbf{x}_{i2}'\boldsymbol{\beta}, 1)$; $[z_{i3} \mid y_i = 3, \boldsymbol{\beta}] \sim \mathcal{TN}_{(-\infty,0)}(\mathbf{x}_{i3}'\boldsymbol{\beta}, 1)$, where $\mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes the normal(μ, σ^2) distribution truncated to the interval (a, b). In general, if $y_i = j$, then the latent data corresponding to this observation are represented as $\mathbf{z}_i = (z_{i1}, \ldots, z_{ij_i})$, where $j_i = \min\{j, J-1\}$ and the simulation of $\mathbf{z}_i$ is from a sequence of truncated normal distributions.

Similarly, the posterior distribution of the parameter vector $\boldsymbol{\beta}$, conditional on the latent data $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, has a simple form. Let $\mathbf{X}_i$ denote the covariate matrix corresponding to the ith subject consisting of the j_i rows $\mathbf{x}_{i1}', \ldots, \mathbf{x}_{ij_i}'$. If $\boldsymbol{\beta}$ is assigned a multivariate normal prior with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\mathbf{B}_0$, then it is not difficult to show that the posterior distribution of $\boldsymbol{\beta}$, conditional on $\mathbf{z}$, is $\mathcal{N}\{\hat{\boldsymbol{\beta}}, (\mathbf{B}_0^{-1} + \Sigma_{i=1}^n \mathbf{X}_i'\mathbf{X}_i)^{-1}\}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{B}_0^{-1} + \Sigma_{i=1}^n \mathbf{X}_i'\mathbf{X}_i)^{-1}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \Sigma_{i=1}^n \mathbf{X}_i'\mathbf{z}_i)$.

These facts lead to the following MCMC algorithm for fitting the basic sequential ordinal model:

(1) Sample $\mathbf{z} \mid \mathbf{y}, \boldsymbol{\beta}$ by sampling $\mathbf{z}_i \mid (y_i = j, \boldsymbol{\beta})$ for each i. If $y_i = 1$, generate z_{i1} from $\mathcal{TN}_{(-\infty,0)}(\mathbf{x}_{i1}'\boldsymbol{\beta}, 1)$. If $y_i = j$ $(2 \leq j \leq J-1)$, generate $\{z_{ik}\}_{k=1}^{j-1}$ independently from $\mathcal{TN}_{(0,\infty)}(\mathbf{x}_{ik}'\boldsymbol{\beta}, 1)$ and z_{ij} from $\mathcal{TN}_{(-\infty,0)}(\mathbf{x}_{ij}'\boldsymbol{\beta}, 1)$. If $y_i = J$, generate $\{z_{ik}\}_{k=1}^{J-1}$ independently from $\mathcal{TN}_{(0,\infty)}(\mathbf{x}_{ik}'\boldsymbol{\beta}, 1)$.

(2) Simulate $\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{z}$ by sampling from the distribution $\mathcal{N}_k\{\hat{\boldsymbol{\beta}}, (\mathbf{B}_0^{-1} + \Sigma_{i=1}^n \mathbf{X}_i'\mathbf{X}_i)^{-1}\}$.

(3) Go to step 1 and repeat the iterations using the most recent values of the conditioning variables.

The above algorithm can be adjusted to account for censoring of the type described in Section 2.1. If the observation is censored at $y_i = j$ $(d_i = 0)$, then in step 1 of the algorithm, one generates $j_i = j - 1$ independent variates $z_{i1}, \ldots, z_{ij}$ from the truncated normal distributions $\mathcal{TN}_{(0,\infty)}(\mathbf{x}_{ik}'\boldsymbol{\beta}, 1)$, $k = 1, \ldots, j$. Step 2 of the algorithm is unchanged.

3. Discrete-Time Survival Analysis

In the survival analysis setting, the hazard of the basic model is given by (1), which allows for an arbitrary shape of the baseline. We refer to the covariates in this model as global covariates because the influence of these variables on the discrete hazard will not depend on the time interval.

In some cases, the effect of a particular covariate on the hazard function may be different across time. Suppose we subdivide the vector of covariates $\mathbf{x}_i$ into a vector $\mathbf{x}_{i(1)}$ of category-specific covariates, whose effect may depend on the particular response category, and a vector $\mathbf{x}_{i(2)}$, which consists of global covariates. We call these interaction models since there is interaction present between some covariates and the time variable. In this case, the hazard is represented as

$$\Pr(Y_i = j \mid Y_i \geq j, \boldsymbol{\gamma}, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{J-1}, \boldsymbol{\delta}) = F\left(\gamma_j - \mathbf{x}_{i(1)}'\boldsymbol{\delta}_j - \mathbf{x}_{i(2)}'\boldsymbol{\delta}\right). \quad (8)$$

The regression parameter $\boldsymbol{\delta}$ represents the effect of the global

covariates and the set $\{\boldsymbol{\delta}_j\}$ reflects the effects of the category-specific covariates. The fitting of these models is accomplished by letting the matrix $\mathbf{X}_i$ in algorithm 1 be given by the first $j_i = \min\{j, J-1\}$ rows of the matrix

$$\begin{pmatrix} -1 & 0 & \cdots & 0 & -\mathbf{x}'_{i(1)} & 0 & 0 & \cdots & \mathbf{x}'_{i(2)} \\ 0 & -1 & \cdots & 0 & 0 & -\mathbf{x}'_{i(1)} & 0 & \cdots & \mathbf{x}'_{i(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & 0 & \cdots & -\mathbf{x}'_{i(1)} & \mathbf{x}'_{i(2)} \end{pmatrix}$$

with $\mathbf{x}'_{ij}$ equal to the jth row of this matrix.

In the discrete-survival context, there may be a large number of time intervals and therefore a large number of possible response categories, J. In this case, it is desirable to model the baseline categories $\{\gamma_j\}$ and/or the $\{\boldsymbol{\delta}_j\}$ by a parsimonious function $f(j)$ containing a small number of parameters that reflect the general features of the discrete hazard (Mantel and Hankey, 1978; Efron, 1988). If the hazard is believed to be constant across time, this motivates the model of the form $\gamma_j = \phi_0$. An example of a more flexible function that could be used to model a constant, an increasing, or a decreasing hazard function is the quadratic form $\gamma_j = \phi_0 + \phi_1 j + \phi_2 j^2$, where the parameter values $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2)$ are unknown and must be estimated from the data. In the case where one is fitting an interaction model, then the category-specific coefficients $\{\boldsymbol{\delta}_j\}$ may also be hypothesized to lie on a low-order polynomial on the category index. In particular, suppose that δ_j is m dimensional; then one can let the lth component of $\boldsymbol{\delta}_j$ (i.e., δ_{jl}) lie on a polynomial $\delta_{jl} = \eta_{l0} + \eta_{l1} j + \eta_{l2} j^2$ with unknown parameters η_{k0}, η_{k1}, and η_{k2}. In vector notation, these constraints can be expressed as $\boldsymbol{\delta}_j = \mathbf{A}_j \boldsymbol{\eta}$, $j \leq J-1$, where $\mathbf{A}_j = I_m \otimes (1 \quad j \quad j^2)$ is a m times $3m$ matrix, $\boldsymbol{\eta} = (\boldsymbol{\eta}'_1, \ldots, \boldsymbol{\eta}'_m)'$ is a $3m$ vector and each $\boldsymbol{\eta}_l$ is a three-by-one vector consisting of $(\eta_{l0}, \eta_{l1}, \eta_{l2})$.

The sequential model under polynomial restrictions on both $\{\gamma_j\}$ and $\{\boldsymbol{\delta}_j\}$ can again be fit using algorithm 1 after defining $\mathbf{X}_i$ to be the first j_i rows of the matrix

$$\begin{pmatrix} -1 & -1 & -1 & -\mathbf{x}'_{i(1)} A_1 & \mathbf{x}'_{i(2)} \\ -1 & -2 & -2 & -\mathbf{x}'_{i(1)} A_2 & \mathbf{x}'_{i(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -(J-1) & -(J-1)^2 & -\mathbf{x}'_{i(1)} A_{J-1} & \mathbf{x}'_{i(2)}, \end{pmatrix}$$

with $\mathbf{x}'_{ij}$ equal to the jth row of this matrix. The parameters of the resulting model are given by $\boldsymbol{\beta} = (\boldsymbol{\phi}', \boldsymbol{\eta}', \boldsymbol{\delta}')'$.

4. Comparison of Models by Bayes Factors

4.1 *Introduction*

A formal Bayesian approach for comparing nested or nonnested models is based on the notion of a Bayes factor. If $\mathbf{y}$ denotes the observation and $\boldsymbol{\theta}$ denotes the parameter, a Bayesian model $\mathcal{M}$ consists of a sampling density $f(\mathbf{y} \mid \mathcal{M}, \boldsymbol{\theta})$ and a prior density $\pi(\boldsymbol{\theta} \mid \mathcal{M})$. For any two models, $\mathcal{M}_r = \{f(\mathbf{y} \mid \mathcal{M}_r, \boldsymbol{\theta}_r), \pi(\boldsymbol{\theta}_r \mid \mathcal{M}_r)\}$ and $\mathcal{M}_s = \{f(\mathbf{y} \mid \mathcal{M}_s, \boldsymbol{\theta}_s), \pi(\boldsymbol{\theta}_s \mid \mathcal{M}_s)\}$, the Bayes factor of r versus s is defined as

$$B_{rs} = \frac{m(\mathbf{y} \mid \mathcal{M}_r)}{m(\mathbf{y} \mid \mathcal{M}_s)} = \frac{\int f(\mathbf{y} \mid \mathcal{M}_r, \boldsymbol{\theta}_r) \pi(\boldsymbol{\theta}_r \mid \mathcal{M}_r) d\boldsymbol{\theta}_r}{\int f(\mathbf{y} \mid \mathcal{M}_s, \boldsymbol{\theta}_s) \pi(\boldsymbol{\theta}_s \mid \mathcal{M}_s) d\boldsymbol{\theta}_s},$$

where $m(\mathbf{y} \mid \mathcal{M}_k)$ is the marginal likelihood of $\mathcal{M}_k$. The Bayes factor can be understood much like a likelihood ratio. For example, if $B_{rs} = 10$, then model $\mathcal{M}_r$ is supported 10 times more by the data then model $\mathcal{M}_s$. It is convenient, as in Good (1967), to express the Bayes factor on a log base 10 scale. If $\log B_{rs} = 0$, the models receive equal support from the data, and if $\log B_{rs} = -2$, model $\mathcal{M}_s$ has 100 times more support from the data.

To compute the Bayes factor, the marginal likelihood $m(\mathbf{y} \mid \mathcal{M}_r)$ needs to be evaluated for each model. In this article, we adopt the procedure of Chib (1995) to estimate the marginal likelihood from the MCMC output. The Chib method is based on the identity

$$m(\mathbf{y} \mid \mathcal{M}_s) = \frac{f(\mathbf{y} \mid \mathcal{M}_s, \boldsymbol{\theta}^*) \pi(\boldsymbol{\theta}^* \mid \mathcal{M}_s)}{\pi(\boldsymbol{\theta}^* \mid \mathcal{M}_s, \mathbf{y})},$$

where $\boldsymbol{\theta}^*$ is any point in the parameter space (usually taken to be a high density point). The likelihood ordinate $f(\mathbf{y} \mid \mathcal{M}_s, \boldsymbol{\theta}^*)$ is available for each model and the remaining issue is about the specification of the prior and the computation of the posterior ordinate.

4.2 *Priors*

The formulation of proper prior distributions is an important element when the goal of the analysis is to compare (nonnested) models using Bayes factors. For the cumulative model, it is convenient to assume that the parameters γ^c and $\boldsymbol{\delta}$ are *a priori* independent and to assign $\boldsymbol{\delta}$ a multivariate normal distribution. A normal prior distribution is not appropriate for γ^c due to the order restriction in the components. However, γ^c can be reexpressed to a real-valued vector $\alpha = (\alpha_1, \ldots, \alpha_{J-1})$ by means of the transformation $\alpha_1 = \log \gamma_1^c$, $\alpha_j = \log(\gamma_j^c - \gamma_{j-1}^c)$, $2 \leq j \leq J-1$. One can then assign α a multivariate normal prior distribution.

In the sequential ordinal case, the cutpoints $\gamma_1, \ldots, \gamma_{J-1}$ are not restricted and so it is convenient to assign a $\mathcal{N}_k(\boldsymbol{\beta}_0, \mathbf{B}_0)$ prior to $\boldsymbol{\beta}$. The assignment of the hyperparameters values $\{\boldsymbol{\beta}_0, \mathbf{B}_0)$ is generally difficult because $\boldsymbol{\beta}$ and the ordinal probabilities are nonlinearly related and little information is typically available about the location of the cutpoints. One method for assigning a proper diffuse prior relies on the notion of a training sample (Lempers, 1971, Chapter 5; Spiegelhalter and Smith, 1982). Suppose that sufficient data are available for one to use a portion of the data to assess the values of the multivariate normal hyperparameters and the remainder of the data for fitting and model comparison. Subdivide the data $\{(\mathbf{x}_i, y_i)\}$ at random into two parts, $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}$ and $\{(\mathbf{x}_i^{(e)}, y_i^{(e)})\}$. First, fit the ordinal model to the training sample $\{(\mathbf{x}_i^{(t)}, y_i^{(t)})\}$ using a flat noninformative prior on $\boldsymbol{\beta}$. From this fit, posterior means and variances of the parameters are computed; the posterior variance from this fit is then inflated (say be a factor of two) and the mean and the inflated variance are used to define the hyperparameters of the multivariate normal prior. Then the model is fit a second time to the remainder of the dataset $\{(\mathbf{x}_i^{(e)}, y_i^{(e)})\}$ using this informative prior. An alternative method of specifying a vague proper prior is based on statements about a set of cumulative probabilities. Beta priors are used to represent knowledge about specific cumulative probabilities, and then this distribution is transformed back

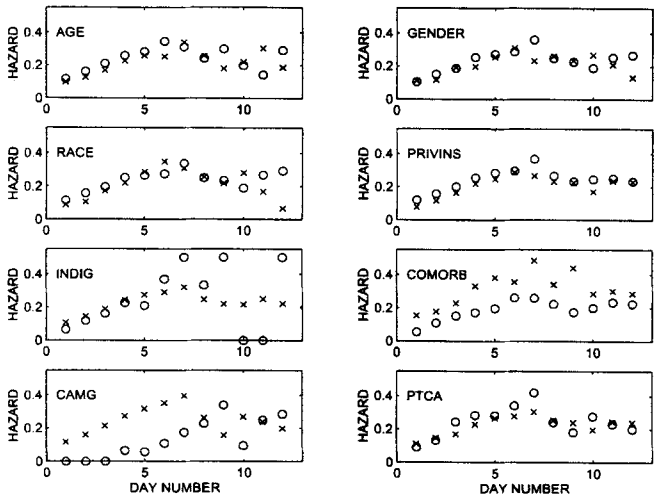

Figure 1. Empirical hazard function $\Pr(Y_i = j \mid Y_i \geq j)$ plotted for subgroups of the sample formed by different values of the eight covariates. In each graph, the symbol 'x' corresponds to the hazard for the covariate value of zero and a symbol 'o' corresponds to the hazard for a positive value of the covariate.

to obtain a prior distribution on $(\boldsymbol{\beta}, \gamma)$. See Albert and Chib (1997) for details on this prior construction in the ordinal data setting.

4.3 *Marginal Likelihood of the Sequential Model*

In the sequential model, the parameter vector is $\boldsymbol{\theta} = (\gamma, \boldsymbol{\delta}) = \boldsymbol{\beta}$ and the posterior ordinate is $\pi(\boldsymbol{\beta} \mid \mathcal{M}, \mathbf{y}) = \int \pi(\boldsymbol{\beta} \mid \mathcal{M}, \mathbf{y}, \mathbf{z}) \pi(\mathbf{z} \mid \mathcal{M}, \mathbf{y}) d\mathbf{z}$, where $\pi(\boldsymbol{\beta} \mid \mathcal{M}, \mathbf{y}, \mathbf{z})$ is the multivariate normal density that appears in the sequential model fitting algorithm. Let $\{\mathbf{z}^{(g)}\}$ be draws from the MCMC run and

$$\hat{\boldsymbol{\beta}}^{(g)} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left(\mathbf{B}_0^{-1} \boldsymbol{\beta}_0 + \sum_{i=1}^{n} \mathbf{X}_i' z_i^{(g)} \right).$$

Then a simulation-consistent estimate of the posterior ordinate is given by $\hat{\pi}(\boldsymbol{\beta}^* \mid \mathcal{M}, \mathbf{y}) = G^{-1} \sum_{g=1}^{G} \phi_k(\boldsymbol{\beta}^* \mid \hat{\boldsymbol{\beta}}^{(g)}, \mathbf{B}^{-1})$, where $\phi_k(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the k-variate multivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The marginal likelihood of the sequential ordinal model on the log scale is now immediately available as

$$\ln m(\mathbf{y} \mid \mathcal{M}) = \ln f(\mathbf{y} \mid \mathcal{M}, \boldsymbol{\beta}^*) + \ln \phi_k(\boldsymbol{\beta}^* \mid \mathcal{M}, \boldsymbol{\beta}_0, \mathbf{B}_0) - \ln \left\{ G^{-1} \sum_{g=1}^{G} \phi_k \left(\boldsymbol{\beta}^* \mid \hat{\boldsymbol{\beta}}^{(g)}, \mathbf{B}^{-1} \right) \right\}, \quad (9)$$

where $f(\mathbf{y} \mid \mathcal{M}, \boldsymbol{\beta}^*)$ is the density of the data in (5) evaluated at $\boldsymbol{\beta}^*$.

5. Example

We return to the hospital stay example described in the Introduction, where Y_i denotes the number of days of hospital stay for the ith patient undergoing heart surgery. As a preliminary look at this data, Figure 1 plots values of the observed hazard function $\Pr(Y_i = j \mid Y_i \geq j)$ against the day number j $(1 \leq j \leq 12)$ for subgroups of the sample formed by different values of each of the eight covariates. For the binary covariates, values of the observed hazard for the subgroup coded 1 are plotted using the symbol 'o' and the hazard for the subgroup coded 0 are displayed using the symbol 'x'. For this graph, the AGE variable was categorized into two groups (1 = young, 0 = old) and the 'o' ('x') symbol corresponds to the young (old) patients. Likewise, the COMORB variable was divided into two levels (0–2 = low, 3 or more = high) and the 'o' in the plot corresponds to the high value of COMORB.

Some general features of the data are evident from Figure 1. Here the hazard can be interpreted as the probability of going home on a particular day, given that the patient is still in the hospital. The hazard generally increases from a value of 0.1 to a value of 0.4 about day 7 and then slowly decreases. The two covariates COMORB and CAMG appear to influence the hazard—for each variable, the hazard function is significantly higher for patients where the variable is absent (covariate value equal to zero) relative to patients where the variable is present (value of one). This means that patients with these characteristics will have longer hospital stays. The patterns for the remaining covariates are much less clear, and it is hard to tell from Figure 1 which variables may be statistically significant.

This preliminary analysis motivates the consideration of a series of ordinal response regression models, fit with the probit link where $F = \Phi$. For each ordinal model, we use an independent training sample of size 200 to estimate the hyperparameters of a multivariate normal prior on the unknown regression parameters and cutpoints. This prior, specified in this way for each model, is then used to find the marginal likelihood from the MCMC output as described in Section 4. We use these marginal likelihoods as a device for searching for good-fitting parsimonious models.

Initially, we consider the following four models:

(1) $\mathcal{M}_1$: The basic cumulative ordinal response model (1). This model was fit by the modified Albert and Chib (1993) algorithm described in Albert and Chib (1997).

(2) $\mathcal{M}_2$: The sequential model (2). As there are $J = 12$ possible values of the ordinal response, this model includes $J - 1 = 11$ distinct parameters $\{\gamma_j\}$ to model the baseline hazard function and a vector $\boldsymbol{\delta}$ of eight parameters to model the effects of the covariates, which will influence the baseline hazard in a global fashion.

(3) $\mathcal{M}_3$: The sequential model (8) using COMORB as a category-specific covariate. This model states that the effect of the variable COMORB differs across time. This particular interaction model was chosen based on the inspection of the empirical hazard functions for the two levels of COMORB in Figure 1. For small numbers of days (where most of the data fall), the hazard function appears to increase more sharply for a positive value of the covariate than for a negative value of the covariate. Note that, in the sequential model, only one parameter is used to model the effect of COMORB—here $J - 1 = 11$ parameters are used to model the effect of COMORB for each day.

(4) $\mathcal{M}_4$: The sequential model where the baseline hazard function, as modeled by the cutpoints $\{\gamma_j\}$, is restricted to fall on a quadratic polynomial with unknown coefficients.

Table 1

*Log marginal likelihoods (by the Chib method) and BIC values of alternative models (see text for model definitions). A few of the BIC values were not computable due to the difficulty of finding the mode of the likelihood, and these are indicated by *.*

	Covariates			
	Complete		Reduced	
$\mathcal{M}_1$	−2146.7	(−2168.9)	−2144.0	(−2162.5)
$\mathcal{M}_2$	−2098.6	(−2125.2)	−2102.9	(−2118.7)
$\mathcal{M}_3$	−2117.0	*	−2113.7	*
$\mathcal{M}_4$	−2094.0	(−2107.0)	−2092.1	(−2101.4)
$\mathcal{M}_5$	−2191.2	(−2198.0)	−2188.5	(−2191.8)
$\mathcal{M}_6$	−2223.2	(−2235.6)	−2220.9	(−2229.5)

Table 1 (the "complete" column) gives values of the logarithm of the marginal likelihood, $\ln m(\mathbf{y} \mid \mathcal{M})$, for each of the four models $\mathcal{M}_1 - \mathcal{M}_4$. These marginal likelihoods can be used to compute Bayes factors to compare models. For example, the Bayes factor in support of the sequential model $\mathcal{M}_2$ compared with the basic ordinal model $\mathcal{M}_1$ is given by $BF_{21} = m(\mathbf{y} \mid \mathcal{M}_2)/m(\mathbf{y} \mid \mathcal{M}_1) = \exp\{-2098.6 + 2146.7\} = 7.7 \times 10^{20}$, giving support to the sequential model. Comparing all four models, the largest value of the log marginal likelihood (−2094.0) is attained at model $\mathcal{M}_4$, which implies that there is support for restricting the baseline hazard function to a quadratic form. There is little evidence for the inclusion of COMORB as a category interaction variable—the value of the log marginal likelihood for model $\mathcal{M}_3$ is significantly higher than the value for the basic sequential model $\mathcal{M}_2$.

Table 1 also includes values of the Bayesian Information Criterion (BIC), which is the value of the maximized log likelihood penalized by a term that is a function of the number of parameters in the model. Looking at the "complete" column, we see that model $\mathcal{M}_4$ is also preferred from the BIC criterion.

The posterior means and standard deviations for all eight covariates (the "complete" model) for model $\mathcal{M}_4$ are displayed in Table 2. By looking at the sizes of the posterior means relative to the sizes of the standard deviations, it appears that variables AGE, GENDER, and INDIG are insignificant and can be removed from the model. All of the models are refit with the reduced set of covariates; the log marginal likelihoods for these models are listed under the "reduced" column of Table 1. On the basis of the resulting marginal likelihoods, we can conclude that the models $\mathcal{M}_1 - \mathcal{M}_4$ are each better fit with the reduced set of covariates. This analysis also reveals that the best model is $\mathcal{M}_4$ containing the five covariates RACE, PRIVINS, COMORB, CAMG, and PTCA.

For comparison purposes, we also model our response Y_i as continuous rather than as discrete and fit two parametric survival models, the Weibull ($\mathcal{M}_5$) and log logistic ($\mathcal{M}_6$) by a single-block Metropolis–Hastings sampler (Chib and Greenberg, 1995). As in the previous ordinal analyses, our prior is based on a training sample of 200 observations. In Table 1, we report the log marginal likelihood for these two parametric survival models under both the complete and reduced set of covariates. We see that the parametric models are not supported by the data because the values of the marginal likelihoods of models $\mathcal{M}_5$ and $\mathcal{M}_6$ are substantially smaller than those for the sequential models.

The goodness of fit of the sequential model $\mathcal{M}_4$ can be assessed by comparing the fitted multinomial probabilities with the observed ordinal values for the patients who did not die while they were in the hospital. For each patient, we observe the ordinal value $\mathbf{y}$ with associated observed fitted probabilities $\{\Pr(\widehat{Y = j}), j \leq J\}$. To assess the consistency of the observed value with the multinomial fitted distribution, one can compute the value $\eta = \min\{\Sigma_{j \leq y} \Pr(\widehat{Y = j}), \Sigma_{j \geq y} \Pr(\widehat{Y = j})\}$. If this value for a particular observation is small, then that observation is inconsistent with the fitted model. For the dataset of noncensored patients, 32% of the η values are smaller than 0.05 and 22% are smaller than 0.01. On closer examination, it can be seen that this lack of fit is due to length of stays that are longer than those predicted using the model. This lack of fit for long lengths of stay raises the possibility that our relatively rich data set is incomplete in some respects and that additional covariates may be helpful. For example, it is possible that some patients develop surgical-site infections or

Table 2

Posterior means and standard deviations of covariates of model $\mathcal{M}_4$ (complete and reduced forms) for the hospital stay example. Inefficiency factors from the autocorrelations of the sampled output are also reported.

	Covariates					
	Complete			Reduced		
Covariate	Mean	SD	Ineff.	Mean	SD	Ineff.
AGE	0.003	0.003	2.77	*	*	*
GENDER	−0.011	0.050	3.15	*	*	*
RACE	−0.088	0.049	3.51	−0.091	0.049	2.30
PRIVINS	−0.119	0.047	2.43	−0.133	0.046	2.04
INDIG	0.092	0.101	2.54	*	*	*
COMORB	0.121	0.009	4.18	0.121	0.009	3.11
CAMG	0.772	0.074	2.39	0.768	0.076	2.16
PTCA	0.110	0.049	2.48	0.110	0.050	2.22

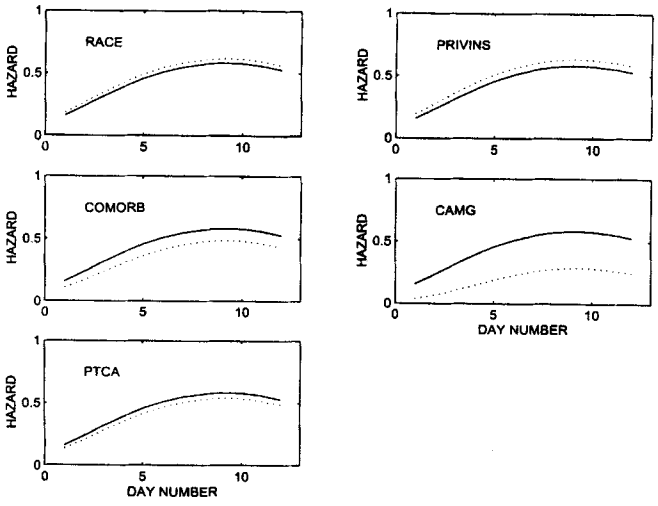

Figure 2. Posterior means of the hazard function for different covariate values. The solid line in each graph represents the baseline hazard for an individual where all covariates are equal to zero and the dotted line represents the hazard for an individual where the particular covariate has been changed to one (two for the COMORB covariate).

other complications that could lengthen the time at the hospital, regardless of the characteristics measured at baseline. This information is not available in our data set. Another nonmeasured but unobservable variable may be hospital-specific quality that may induce differences in the length of stay of patients across hospitals. The modeling of this factor would require matching patients with hospitals (which is not possible with our data) followed by a random effects formulation of the hospital-specific variable within the context of a clustered data sequential ordinal model. This modeling extension can be the focus of further study.

Table 2 gives the posterior means and standard deviations of the five covariates for the best model $\mathcal{M}_4$. The coefficients on RACE and PRIVINS are negative, which means that white patients and patients with private insurance had shorter hospital stays. In contrast, the COMBORB, CAMG, and PTCA variables have positive coefficients, indicating that these conditions cause extended hospital stays. The table also contains the inefficiency factors, computed as $\{1 + 2\Sigma_{k=1}^{\infty} \rho_j(k)\}$, where $\rho_j(k)$ is the sample autocorrelation at lag k from the MCMC output on the jth parameter. The inefficiency factor may be interpreted as the ratio of the numerical variance of the posterior mean under Markov chain sampling to the variance of the posterior mean from hypothetical independent sampling. We see that each of the inefficiency factors is small, indicating that our proposed sampler is mixing well.

The fitted hazard functions are displayed in Figure 2. Consider a patient who is nonwhite (WHITE = 0), does not hold private insurance (PRIVINS = 0), had zero comorbidities (COMORB = 0), did not have a coronary artery bypass graft (CAMG = 0), and did not have an angioplasty (PTCA = 0). The posterior mean of the hazard function $\mathrm{E}\{F(\gamma_k - \mathbf{x}'\boldsymbol{\delta})\}$ is graphed against the day number k and displayed as a solid line in each of the graphs in Figure 2. These expectations were computed by finding the mean of simulated values of the parameters over a grid of values of the covariate. The dotted line in each graph represents the posterior mean of the hazard function when the corresponding covariate value for this patient is changed. So, e.g., the dotted line in the upper left figure corresponds to a patient where PRIVINS = COMORB = CAMG = PTCA = 0 and WHITE = 1. These graphs clearly show how the baseline hazard changes in a global fashion as a function of the five significant covariates.

6. Conclusion

In this article, the sequential model has been analyzed in relation to other models for ordinal data. For such problems, the model was extended to include covariates that affect the hazard rate in a general or in a category-specific manner. There are several attractive aspects to the Bayesian approach discussed here. The latent data representation of the model leads to simple MCMC fitting procedures. In addition, the Bayes marginal likelihood approach provides an effective mechanism for evaluating the significance of covariates and comparing nonnested models, such as the cumulative ordinal, the sequential ordinal, and continuous survival models with Weibull and log-logistic error structures.

Acknowledgements

The authors are grateful to the referees and the editors for their helpful comments.

Résumé

Dans ce papier, on considère la classe des modèles ordinaux séquentiels, en relation avec d'autres modèles pour des données de réponse ordinales. Des algorithmes de Monte Carlo par Chaînes de Markov (MCMC) sont développés pour l'ajustement de ces modèles. Les idées et les méthodes sont illustrées en détail avec des données réelles de durée d'hospitalisation pour des patients admis en chirurgie cardiaque. Un aspect notable de cette analyse est la comparaison de plusieurs modèles non emboîtés, tels que le modèle séquentiel, le modèle ordinal cumulatif, et les modèles de Weibull et log-logistique, se basant sur les vraisemblances marginales et des échantillons tests.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88,** 669–679.

Albert, J. and Chib, S. (1997). *Bayesian methods for cumulative, sequential and two-step ordinal data regression models.* Technical Report. Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90,** 1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statistician* **49,** 327–335.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* **96,** 270–281.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association* **83,** 414–425.

Farhmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer Verlag.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society, Series B* **29,** 399–431.

Johnson, V. and Albert, J. (1999). *Ordinal Data Modeling.* New York: Springer Verlag.

Kalbfleish, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* Rotterdam: Rotterdam University Press.

Mantel, N. and Hankey, B. F. (1978). A logistic regression analysis of response time data where the hazard function is time dependent. *Communications in Statistics, Series A* **7,** 333–347.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42,** 109–127.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* New York: Chapman and Hall.

Spiegelhalter, D. L. and Smith, A. F. M. (1982). Bayes factors for linear and loglinear models with vague prior information. *Journal of the Royal Statistical Society, Series B* **44,** 377–387.

Ten Have, T. R. and Uttal, D. H. (1994). Subject-specific and population-averaged continuation ratio logit models for multiple discrete-time survival profiles. *Applied Statistics* **43,** 371–384.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology* **43,** 39–55.

Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics and Data Analysis* **11,** 275–295.

Received July 2000. *Revised January* 2001.
Accepted January 2001.