Fahrmeir, Pritscher:

Regression analysis of forest damage by marginal models for correlated ordinal responses

Sonderforschungsbereich 386, Paper 9 (1995)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Regression analysis of forest damage by marginal models for correlated ordinal responses

By Ludwig Fahrmeir and Lisa Pritscher

Abstract

Studies on forest damage can generally not be carried out by common regression models, mainly for two reasons: Firstly, the response variable, damage state of trees, is usually observed in ordered categories. Secondly, responses are often correlated, either serially, as in a longitudinal study, or spatially, as in the application of this paper, where neighborhood interactions exist between damage states of spruces determined from aerial pictures. Thus so–called marginal regression models for ordinal responses, taking into account dependence among observations, are appropriate for correct inference. To this end we extend the binary models of Liang and Zeger (1986) and develop an ordinal GEE1 model, based on parametrizing association by global cross–ratios. The methods are applied to data from a survey conducted in Southern Germany. Due to the survey design, responses must be assumed to be spatially correlated. The results show that the proposed ordinal marginal regression models provide appropriate tools for analyzing the influence of covariates, that characterize the stand, on the damage state of spruce.

key words: aerial infra–red pictures, categorical data, correlated observations, cumulative logit model, damage of spruce, generalized estimating equations, global cross–ratio, multivariate regression, spatial correlation.

1 Introduction

In forest damage surveys the state of trees, e.g. the degree of defoliation, is usually measured in ordered categories ranging from 'no damage' to 'very strong damage'. To analyze the influence of covariates such as site, age or mixture of stand on damage state, regression models for ordinal response are a natural choice, see e.g. Kublin (1987). However, many forest damage surveys are conducted in such a way that direct application of common ordinal regression models, for example cumulative logit regression models, is questionable since they rely on the basic assumption of independent observations.

A typical survey of this kind arises in the specific application of this paper. First, a grid with rectangular meshes is placed over a map of the survey area, and then, for each grid point, damage states and covariates of a fixed number of trees next to the grid point are measured, using coloured infra–red aerial pictures taken from helicopters. Due to neighboring effects, observations among trees within each of these clusters around the grid points will often be spatially correlated and cannot be assumed to be independent.

Another typical situation with dependent data is a longitudinal study where observations on a sample of trees are made repeatedly over successive years. Again, repeated measurements for each tree form a cluster of observations, correlated over time.

For a correct analysis we have to take into account statistical dependence among observations within clusters. A variety of methods, mostly for regression with binary response, has been recently proposed and developed, ranging from the first generalized estimating approach of Liang and Zeger (1986) to full likelihood analyses (Molenberghs and Lesaffre, 1994). Reviews are given in Liang, Zeger and Qaqish (1992), Diggle, Liang and Zeger (1994), and Fahrmeir and Tutz (1994, ch.3 and ch.6).

As in the application of this paper, often the influence of covariates on the marginal probabilities of response categories, e.g. damage states, is of prime interest, whereas dependence is regarded only as a nuisance. Then the so–called GEE1 approach is a good choice, since only marginal regression relationships have to be specified correctly to obtain reasonable results. If

joint modelling of marginal probabilities and of association, e.g. neighboring effects, is of interest, then the GEE2 approach or full likelihood models are appropriate. However, these methods require that two–way and higher–order associations are correctly specified, otherwise inference on marginal effects can be severely biased. For the forest damage survey considered here, analysis of the influence of covariates on damage state is a primary goal. Also there is too little information in the data for correct specification of dependence. Therefore we rely on the GEE1 approach.

The purpose of our paper is two–fold. Firstly, for the survey described in more detail in Section 2, we aim at providing a reliable analysis of the influence of covariates, mainly characterizing the stand of trees, on damage state. For this aim, secondly, we extend in Section 3 Liang and Zeger's (1986) original proposal to ordinal responses and develop an ordinal marginal regression model based on parametrizing dependence by global cross–ratios (Dale, 1986). For binary responses the latter model reduces to the well-known odds ratio parametrization (Lipsitz, Laird and Harrington, 1991). The correlation parametrization for ordinal models (Miller, Davies and Landis, 1993) turned out to be less useful, at least for our particular data set, since it led to numerical problems, e.g. non-convergence. Section 4 describes the analysis of forest damage data carried out with the methods of Section 3. The results show that ordinal GEE1 methods provide reasonable and consistent estimates of covariate effects, while naive application of common ordinal response models based on independence can lead to erroneous inference.

2 Survey and data description

This section provides some details on the forest damage survey together with a description and a preliminary exploratory analysis of the data. Additional information is contained in Mössmer et al. (1991). The survey was conducted in the forest district of Flossenbürg in the north eastern part of Bavaria. A primary goal was to get information about damage state of spruce, the denominating tree species in this area, and about the influence of other variates on it.

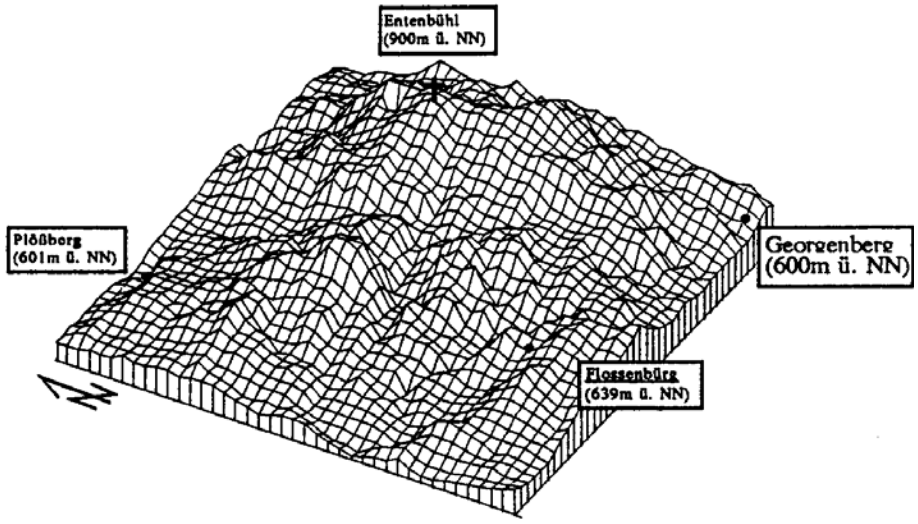

Figure 1: Digital elevation model of the survey area Flossenbürg, 250×250 m grid. Source: Bavarian State Institute of Forestry

To determine damage state, infra–red coloured aerial pictures of the area were taken from helicopters, and the degree of defoliation of tree–tops was used as an indicator of damage state. On infra–red coloured pictures healthy green tops without defoliation have an intensive red colour, while strongly damaged trees appear without any red colouring. In this way damage state was classified into five categories, indicating the degree of defoliation.

A sample of spruce trees was obtained in the following way: Aerial pictures were related to a digital terrain model with a 250 × 250 m grid, see Figure 1. Taking grid points as primary units, damage state of the eight spruce trees next to a grid point was determined. The whole data set of the survey consists of clusters with eight trees for each of 771 grid points, giving a total sample of 6168 trees.

For regression analysis the response variable damage (**D**) was further condensed into three ordered categories: strong, distinct and light. Figure 2 shows corresponding relative frequencies in the sample. All covariates are categorical and due to the survey plan constant within clusters, i.e. trees, belonging to

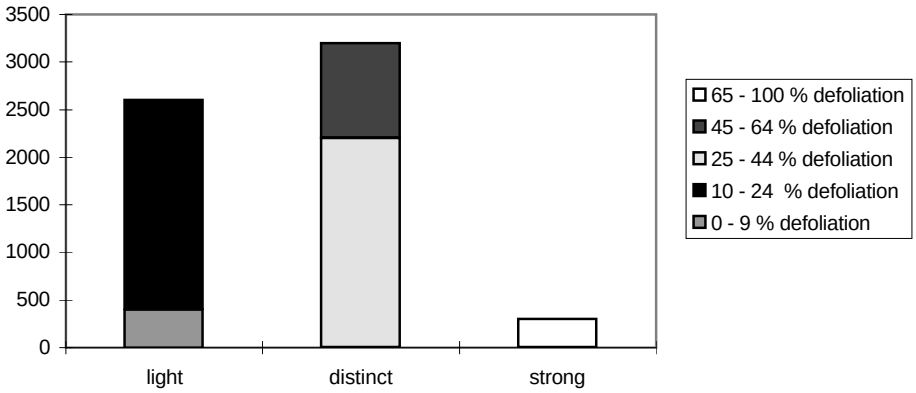

Figure 2: Frequency distribution of damage classes

the same cluster, have the same covariate values.

In a preliminary variable selection (Pritscher, 1992) the following covariates turned out to be most influential:

C canopy density: very low(1), low(2), medium(3), high(4).

M mixture of stand: coniferous(1) or mixed(2).

U utilization method: second commercial thinning(1), first commercial thinning(2), precommercial thinning(3). This variable can be considered as a surrogate for tree stand age, since utilization methods change with age.

S site: bedrock or non–fertile granite weathering(1), fertile granite weathering(2), gneiss weathering(3), soil with water surplus(4).

A altitude : 500-600 m(1), 601-650 m(2), 651-700 m(3), 701-750m(4), 751 - 900 m(5).

In addition to main effects of the covariates, it is reasonable to consider possible interaction effects, for example between utilization method and canopy density or site and altitude.

To get some insight into the influence of covariates, conditional relative frequencies of damage classes given the categories of covariates are displayed

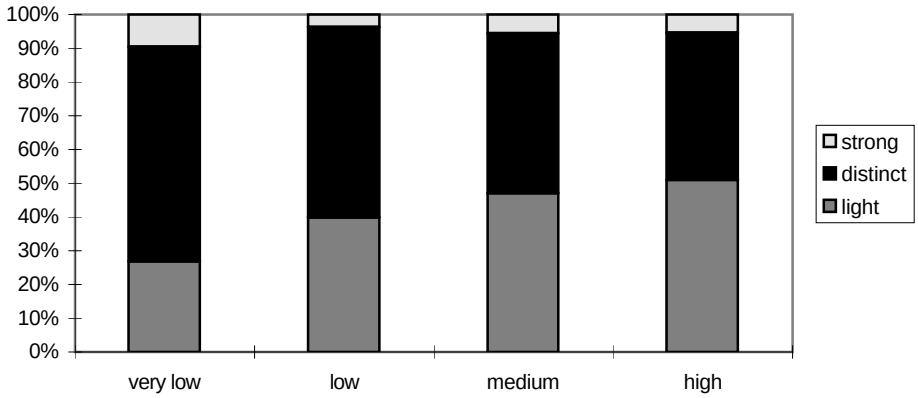

Figure 3: Damage class distribution by canopy density

in Figures 3 to 5. In Figure 3 frequencies of distinctly or strongly damaged trees decrease with increasing canopy density.

Figure 4 exhibits the expected dependence on utilization method as an indicator of age: Distinct and strong damage increases with age.

Figure 5 shows relative frequencies of damage classes conditioned by interaction categories of canopy density and utilization method, providing evidence that it is reasonable to include interaction effects in regression analyses. For example, while precommercial thinning in high canopy stands increases the tendency for only mild damage, there is an inverse tendency for first commercial thinning.

Further bivariate descriptive statistics describing the association between damage and covariates are found in Mössmer et al. (1991).

As already discussed in the introduction, common ordinal regression, based on the assumption of independent observations, is not appropriate for the survey at hand. Since covariates are constant within clusters and do not contain tree–specific information, correlations among trees of the same cluster are very likely. This has already be pointed out by Kublin (1987) and Quednau (1989). To account for that, we will apply marginal ordinal regression models developed in the next section for analyzing the data.

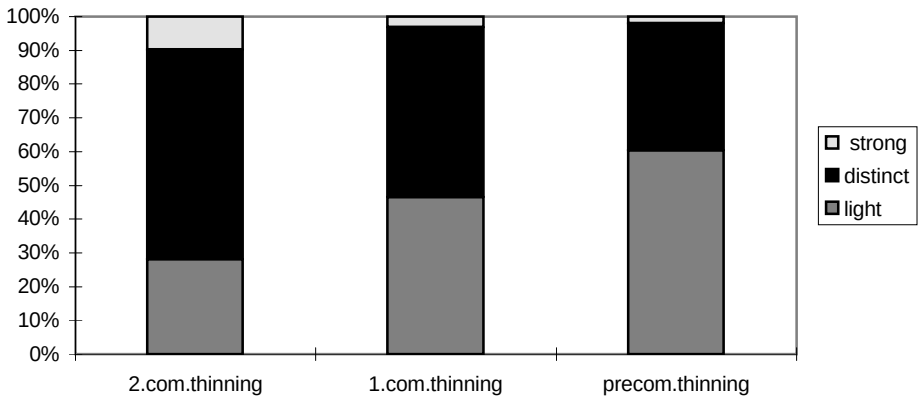

Figure 4: Damage class distribution by utilization method

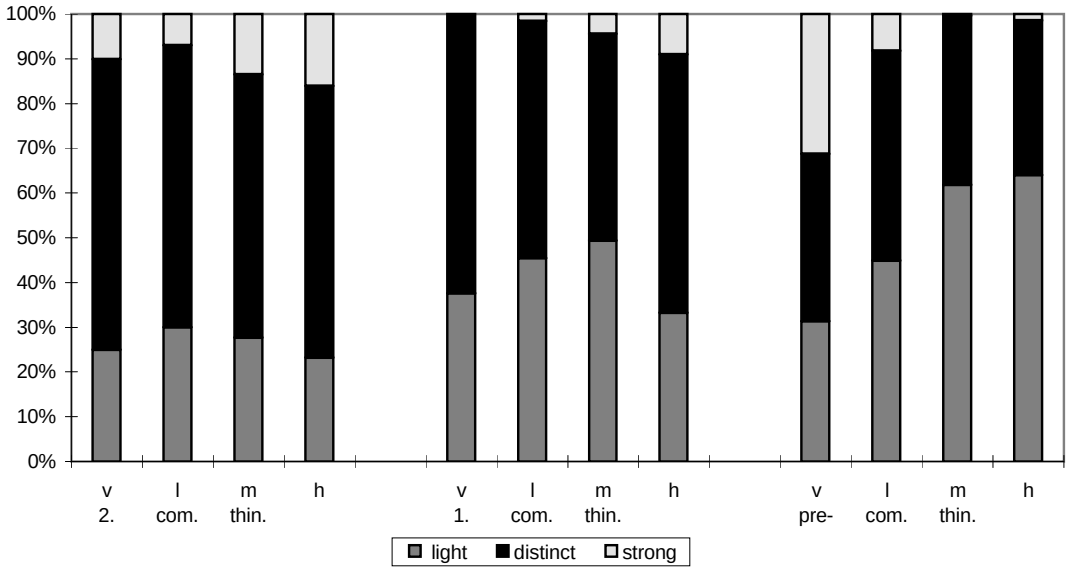

Figure 5: Damage class distribution by utilization method and canopy density

3 Marginal cumulative models for correlated ordinal response

In the following we consider a data situation as in the forest damage survey. Suppose a study has been conducted with I clusters as primary units. In each cluster $i = 1, \ldots, I$ ordinal responses Y_{ij} with $q+1$ categories together with discrete or continuous covariates $\mathbf{x}_{ij}$ are observed for n_i subjects within cluster i. For simplicity we will assume equal cluster sizes $n_i = n$. Covariates may be subject–specific or may be constant within clusters, i.e. $\mathbf{x}_{i1} = \ldots = \mathbf{x}_{in}$. Thus the data are given by $(Y_{ij}, \mathbf{x}_{ij}), i = 1, \ldots, I, j = 1, \ldots, n$.

Marginal probabilities for Y_{ij} are related to covariates $\mathbf{x}_{ij}$ by a cumulative model

$$\mathrm{g}\{\mathrm{pr}(Y_{ij} \leq r|\mathbf{x}_{ij})\} = \theta_r + \mathbf{x}_{ij}'\gamma\,, \quad r = 1, \ldots, q\,,$$

for some suitable link function g, ordered threshold parameters $\theta_1 < \ldots < \theta_q$ and a vector γ of covariate effects. In our application we will use a logit link, implying proportional odds

$$\frac{\mathrm{pr}(Y_{ij} \leq r|\mathbf{x}_{ij})}{\mathrm{pr}(Y_{ij} > r|\mathbf{x}_{ij})} = \exp(\theta_r + \mathbf{x}_{ij}'\gamma) = \exp(\theta_r)\exp(\mathbf{x}_{ij}'\gamma)\,. \tag{1}$$

This and other ordinal response models are discussed in more detail e.g. in Fahrmeir and Tutz (1994, ch.3). For the following it is convenient to represent Y_{ij} as a vector $\mathbf{y}_{ij} = (y_{j1}^{(i)}, \ldots, y_{jr}^{(i)}, \ldots, y_{jp}^{(i)})'$ of q dummies with $y_{jr}^{(i)} = 1$ if $Y_{ij} = r$, $y_{jr}^{(i)} = 0$ if $Y_{ij} \neq r$. We also gather the linear predictors $\eta_{jr}^{(i)} = \theta_r + \mathbf{x}_{ij}'\gamma$ in the q-dimensional predictor $\eta_{ij} = (\eta_{j1}^{(i)}, \ldots, \eta_{jq}^{(i)})'$. Defining the parameter vector $\beta = (\theta_1, \ldots, \theta_q, \gamma')'$ and the design matrix

$$\mathbf{X}_{ij}' = \begin{bmatrix} 1 & & & \mathbf{x}_{ij}' \\ & \ddots & & \vdots \\ & & 1 & \mathbf{x}_{ij}' \end{bmatrix}\,,$$

we have $\eta_{ij} = \mathbf{X}_{ij}'\beta$. Then the vector of marginal response probabilities $\pi_{ij} = (\pi_{j1}^{(i)}, \ldots, \pi_{jq}^{(i)})'$ with $\pi_{jr}^{(i)} = \mathrm{pr}(Y_{ij} = r|\mathbf{x}_{ij}) = \mathrm{pr}(y_{jr}^{(i)} = 1|\mathbf{x}_{ij})$ is related to the linear predictor by

$$\pi_{ij} = \pi_{ij}(\eta_{ij}) = \mathbf{h}(\mathbf{X}_{ij}'\beta) \tag{2}$$

with q-dimensional link $\mathbf{h}$ defined by (1). Gathering responses, response probabilities and design matrices of cluster i, we get

$\mathbf{y}_i = (\mathbf{y}_{i1}', \ldots, \mathbf{y}_{in}')'$, $\pi_i = (\pi_{i1}', \ldots, \pi_{in}')'$ and $\mathbf{X}_i = (\mathbf{X}_{i1}', \ldots, \mathbf{X}_{in}')'$. Then

$$\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \pi_i) = 0\,. \tag{3}$$

is a multivariate generalized estimating equation (GEE) for consistent estimation of β, where $\mathbf{D}_i = \text{blockdiag}(\partial \pi_{ij} / \partial \eta_{ij})$ and $\mathbf{V}_i$ is a 'working' covariance matrix. The matrix $\mathbf{V}_i$ need not be equal to the true covariance matrix of the responses. The estimator $\hat{\beta}$ that solves (3) is still consistent for β and asymptotically normal under mild regularity conditions if only marginal probabilities are correctly specified. However, the asymptotic covariance matrix of $\hat{\beta}$ has to be modified by a robust 'sandwich' estimator, see below, and there will be some loss of efficiency in general.

The working covariance matrix $\mathbf{V}_i$ can be modelled in a variety of ways. Our analyses will be based on the working assumptions of independence or exchangeable correlations, both direct extensions of the Liang and Zeger (1986) approach, already used in Pritscher, Bäumler and Fahrmeir (1994), and on a new GEE1 method, where global cross–ratios as in Dale (1986) together with a second estimating equation are used to model $\mathbf{V}_i$ and to estimate it simultaneously with β.

The simplest choice of $\mathbf{V}_i$ results from the working assumption of independent components $\mathbf{y}_{ij}$ of $\mathbf{y}_i$. Let

$$\mathbf{\Sigma}_{ij} = \text{cov}(\mathbf{y}_{ij}) = \text{diag}(\pi_{ij}) - \pi_{ij}\pi_{ij}'\,.$$

denote the covariance matrix of the multinomial distribution $M(1, \pi_{ij})$ for $\mathbf{y}_{ij}$. Then the independence working assumption corresponds to the choice of a block–diagonal independence working covariance matrix

$$\mathbf{V}_i = \mathbf{\Sigma}_i = \text{diag}(\mathbf{\Sigma}_{i1}, \ldots, \mathbf{\Sigma}_{in})\,. \tag{4}$$

in (3). The estimate $\hat{\beta}_{IEE}$ can be computed from the resulting independence estimating equation (IEE) in the same way as for independent ordinal responses. Also, as for binary response, $\hat{\beta}_{IEE}$ is consistent and asymptotically

normal under regularity assumptions:

$$\hat{\beta}_{IEE} \overset{a}{\sim} N(\beta, \mathbf{A}_{IEE}) ,$$

with the 'sandwich' matrix $\mathbf{A}_{IEE} = \mathbf{F}^{-1}\mathbf{H}\mathbf{F}^{-1}$, where

$$\mathbf{F} = \sum_{i=1}^{I} \mathbf{X}_i'\mathbf{D}_i\mathbf{V}_i^{-1}\mathbf{D}_i'\mathbf{X}_i \ , \ \mathbf{H} = \sum_{i=1}^{I} \mathbf{X}_i'\mathbf{D}_i\mathbf{V}_i^{-1}\,\mathrm{cov}(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i'\mathbf{X}_i \, . \tag{5}$$

An estimate $\hat{\mathbf{A}}_{IEE}$ for $\mathbf{A}_{IEE}$ is obtained by replacing the true but unknown covariance matrix $\mathrm{cov}(\mathbf{y}_i)$ by $(\mathbf{y}_i - \hat{\pi}_i)(\mathbf{y}_i - \hat{\pi}_i)'$ and $\hat{\pi}_i = \pi(\hat{\beta}_{IEE})$. If the components $\mathbf{y}_{ij}$ of $\mathbf{y}_i$ are indeed independent, then $\mathbf{V}_i = \mathrm{cov}(\mathbf{y}_i)$, and $\mathbf{A}_{IEE} = \mathbf{F}^{-1}$ is the usual 'naive' asymptotic covariance matrix of the maximum likelihood estimate. Generally however, $\mathbf{V}_i \neq \mathrm{cov}(\mathbf{y}_i)$, resulting in some loss of efficiency. Therefore, Liang and Zeger (1986) propose to set

$$\mathbf{V}_i = \mathbf{\Sigma}_i^{1/2}\mathbf{R}\mathbf{\Sigma}_i^{1/2} \, .$$

with various choices for the working correlation matrix, $\mathbf{R} = \mathbf{I}$ corresponding to the independence assumption. In our application, we will choose a working matrix for exchangeable correlations,

$$\mathbf{R} = \begin{bmatrix} \mathbf{I} & \mathbf{Q} & \cdots & \mathbf{Q} \\ \mathbf{Q}' & \mathbf{I} & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathbf{Q}' & \cdots & \cdots & \mathbf{I} \end{bmatrix} , \tag{6}$$

but other choices are also possible. The $q \times q$- matrix $\mathbf{Q}$ is generally unknown and is estimated by a method of moments, involving Pearson residuals, simultaneously with β as in the binary case. Details are given in Pritscher (1992). The estimate $\hat{\beta}_{GEE}$ is again consistent and asymptotically normal as in (5), however with $\mathbf{V}_i$ replaced by

$$\hat{\mathbf{V}}_i = \hat{\mathbf{\Sigma}}_i^{1/2}\hat{\mathbf{R}}\hat{\mathbf{\Sigma}}_i^{1/2} \, .$$

where $\hat{\mathbf{R}}$ is obtained by replacing $\mathbf{Q}$ by the estimate $\hat{\mathbf{Q}}$.

Instead of using Pearson residuals to estimate $\mathbf{Q}$, equation (3) can be augmented by a second GEE involving second–order moments of the observations, see e.g. Liang et al. (1992). However, if second order moments are directly parametrized, this approach implies undesirable restrictions on correlations. Therefore we extend the GEE1 method with odds ratio parametrization (Lipsitz et al.,1991) to ordinal responses, using global cross–ratios (Dale, 1986) as measures for association.

Consider two ordinal responses Y_{ij}, Y_{ik} in the same cluster. For each pair of categories l and m of Y_{ij} and Y_{ik}, the global cross–ratio (GCR), given the covariate $\mathbf{X}_i$, is defined as

$$\psi_{jk}^{(i)}(l,m) = \frac{\mathrm{pr}(Y_{ij} \leq l, Y_{ik} \leq m|\mathbf{X}_i)\mathrm{pr}(Y_{ij} > l, Y_{ik} > m|\mathbf{X}_i)}{\mathrm{pr}(Y_{ij} > l, Y_{ik} \leq m|\mathbf{X}_i)\mathrm{pr}(Y_{ij} \leq l, Y_{ik} > m|\mathbf{X}_i)}, \quad l,m = 1,\ldots,q\,. \tag{7}$$

The bivariate cumulative probability function $F_{jk}^{(i)}(l,m) = \mathrm{pr}(Y_{ij} \leq l, Y_{ik} \leq m|\mathbf{X}_i)$ of Y_{ij}, Y_{ik} given $\mathbf{X}_i$, can be expressed in terms of $\psi_{jk}^{(i)}(l,m)$ and marginal cumulative probabilities $F_j^{(i)}(l) = \mathrm{pr}(Y_{ij} \leq l|\mathbf{X}_i)$, $F_k^{(i)}(m) = \mathrm{pr}(Y_{ik} \leq m|\mathbf{X}_i)$:

$$F_{jk}^{(i)}(l,m) = \begin{cases} F_j^{(i)}(l)F_k^{(i)}(m) & \text{if} \quad \psi_{jk}^{(i)}(l,m) = 1 \\ \frac{\kappa - \sqrt{\kappa^2 + 4\psi_{jk}^{(i)}(l,m)(1-\psi_{jk}^{(i)}(l,m))F_j^{(i)}(l)F_k^{(i)}(m)}}{2(\psi_{jk}^{(i)}(l,m)-1)} & \text{if} \quad \psi_{jk}^{(i)}(l,m) \neq 1 \end{cases} \tag{8}$$

where $\kappa = 1 + (F_j^{(i)}(l) + F_k^{(i)}(m))(\psi_{jk}^{(i)}(l,m) - 1)$. Second order moments $E(y_{jl}^{(i)} y_{km}^{(i)})$ can be expressed in terms of $F_j^{(i)}(l), F_k^{(i)}(m), \psi_{jk}^{(i)}(l,m)$ through the relation

$$E(y_{jl}^{(i)} y_{km}^{(i)}) = \begin{cases} F_{jk}^{(i)}(l,m) & l = m = 1 \\ F_{jk}^{(i)}(l,m) - F_{jk}^{(i)}(l,m-1) & l = 1, m > 1 \\ F_{jk}^{(i)}(l,m) - F_{jk}^{(i)}(l-1,m) & l > 1, m = 1 \\ F_{jk}^{(i)}(l,m) - F_{jk}^{(i)}(l,m-1) - F_{jk}^{(i)}(l-1,m) & \\ \quad + F_{jk}^{(i)}(l-1,m-1) & l > 1, m > 1\,. \end{cases} \tag{9}$$

Cross–ratios $\psi_{jk}^{(i)}(l,m)$ are modelled loglinearly, possibly including covariates $z_{jk}^{(i)} = z(\mathbf{x}_{ij}, \mathbf{x}_{ik})$:

$$\log \psi_{jk}^{(i)}(l,m) = \alpha_{lm} + z_{jk}^{(i)'} \gamma_{jk}\,. \tag{10}$$

The off–diagonal elements of $\mathbf{V}_i$ in the GEE (3) for β are determined by $\mathrm{cov}(y_{jl}^{(i)}, y_{km}^{(i)}) = E(y_{jl}^{(i)} y_{km}^{(i)}) - \pi_{jl}^{(i)} \pi_{km}^{(i)}$. To estimate association parameters $\alpha = \{\alpha_{lm}, \gamma_{jk} \; : \; l, m = 1, \ldots, q \, , j, k = 1, \ldots, n\}$ jointly with β, the GEE (3) is augmented by a second GEE for α,

$$\sum_{i=1}^{I} \mathbf{C}_i' \mathbf{U}_i^{-1} (\mathbf{w}_i - \nu_i) = 0 \, . \tag{11}$$

In (11) $\mathbf{w}_i = (\ldots, w_{jl,km}^{(i)}, \ldots)$ contains the products $w_{jl,km}^{(i)} = (y_{jl}^{(i)} - \pi_{jl}^{(i)})(y_{km}^{(i)} - \pi_{km}^{(i)})$, $l, m = 1, \ldots, q \, , j < k = 1, \ldots, n$ and $\nu_i = \nu_i(\alpha, \beta) = (\ldots, \nu_{jl,km}^{(i)}, \ldots)$ is the vector of expectations

$$\nu_{jl,km}^{(i)} = E(w_{jl,km}^{(i)}) = E(y_{jl}^{(i)} y_{km}^{(i)}) - \pi_{jl}^{(i)} \pi_{km}^{(i)} \tag{12}$$

for observations $\{y_{jl}^{(i)}, y_{km}^{(i)}\}$ in cluster i. The matrix $\mathbf{C}_i$ is the Jacobian $\partial \nu_i / \partial \alpha$, obtained from inserting (9),(10) in (12) and differentiating with respect to α. The matrix $\mathbf{U}_i$ is a further working covariance matrix, now for the 'observations' $\mathbf{w}_i$. For the GEE1 approach the following two simple diagonal specifications are useful:
As in the binary case (Prentice, 1988) the simplest choice is the identity matrix specification

$$\mathbf{U}_i = \mathbf{I}.$$

Another choice is

$$\mathbf{U}_i = \mathrm{diag}(\mathrm{var}(w_{jl,km}^{(i)})_{l,m=1,\ldots,q \, , j<k=1,\ldots,n}) \, , \tag{13}$$

where

$$\mathrm{var}(w_{jl,km}^{(i)}) = \pi_{jl}^{(i)}(1 - \pi_{jl}^{(i)}) \pi_{km}^{(i)}(1 - \pi_{km}^{(i)}) - (\nu_{jl,km}^{(i)})^2 + \nu_{jl,km}^{(i)}(1 - 2\pi_{jl}^{(i)})(1 - 2\pi_{km}^{(i)}) \, .$$

The parameters α and β are computed by a (quasi-) Fisher scoring algorithm, switching between the iterations

$$\begin{aligned} \hat{\beta}_{GCR}^{m+1} &= \hat{\beta}_{GCR}^{m} + \left(\sum_{i=1}^{I} \mathbf{X}_i \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{X}_i' \right)^{-1} \left(\sum_{i=1}^{I} \mathbf{X}_i \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \pi_i) \right) \\ \hat{\alpha}^{m+1} &= \hat{\alpha}^{m} + \left(\sum_{i=1}^{I} \mathbf{C}_i' \mathbf{U}_i^{-1} \mathbf{C}_i \right)^{-1} \left(\sum_{i=1}^{I} \mathbf{C}_i' \mathbf{U}_i^{-1} (\mathbf{w}_i - \nu_i) \right) \, . \end{aligned}$$

until convergence. Under regularity assumptions $\hat{\beta}_{GCR}$ is consistent and asymptotically normal as in (5) if marginal probabilities are correctly specified, i.e. indeed $E(\mathbf{y}_i) = \pi_i$ with π_i as in (1) or (2). As can be shown by asymptotic theory for misspecified models (see e.g. Fahrmeir, 1990), it does not matter whether the model for $\mathbf{w}_i$ is correct, i.e. whether $E(\mathbf{w}_i) = \nu_i$ holds or not. This is the main advantage of the GEE1 method if one is only interested in correct inference on covariate effects for marginal probabilities.

4 Application to forest damage study of Flossenbürg

To analyze the data of the survey described in Section 2, we assumed a cumulative model (1), with $q = 2$ relevant categories for the marginal probabilities of light or distinct damage **D**. In addition to the thresholds θ_1, θ_2, the model includes five main effects (canopy density **C**, mixture of stand **M**, utilization method **U**, site **S** and altitude **A**) given in effect–coding, and some interaction terms like $\mathbf{C} * \mathbf{A}$ for interactions between **C** and **A**. For each covariate, the last category is taken as the reference category. In the working correlation matrix (6) for exchangeable association, **Q** is a 2×2 matrix with four unknown elements $q_{11}, q_{12}, q_{21}, q_{22}$. Correspondingly, cross–ratios are parametrized by $\log \psi_{jk}^{(i)}(l,m) = \alpha_{lm}\,, l, m = 1, 2$, i.e. $\psi_{jk}^{(i)}(l,m) = \psi(l,m)$ is constant for each pair of observations within each cluster. No covariate effects γ_{jk} are included, since covariates give no information on association in our application. Note however that working covariances $\mathrm{cov}(y_{jl}^{(i)}, y_{km}^{(i)}) = E(y_{jl}^{(i)} y_{km}^{(i)}) - \pi_{jl}^{(i)} \pi_{km}^{(i)}$ and working correlations depend on cluster i through $\pi_{jl}^{(i)} \pi_{km}^{(i)}$. Table 1 gives parameter estimates and standard errors for the independence model, the exchangeable correlation model and the global cross–ratio model. Additionally, naive standard errors, corresponding to unmodified maximum likelihood estimation for the independence model are given in the second column. Estimates for other interaction effects have been omitted since they turned out to be nonsignificant.

4.1 Comparison of results

Since IEE and ML equations, based on independent observations, are identical, point estimates $\hat{\beta}_{IEE}$ and ML estimates are identical (column 1).

Comparing, however, naive standard errors obtained from unmodified ML estimation to robust standard errors, we see that naive standard errors are distinctly smaller. Therefore naive ML estimation will lead to overinterpretation of results. In particular interactions effects, mixture of stands $*$ altitude and site $*$ altitude, would be falsely considered as significant. Therefore they are omitted in Table 1.

Parameter estimates obtained from the GEE model with exchangeable correlations (6) are smaller in absolute value. Standard errors are smaller than for the IEE model, but still mostly larger than naive standard errors, leading to similar conclusions for significance or nonsignificance of effects. In contrast to the IEE model however, we encountered problems of convergence in the estimation procedure due to large estimates of $\mathbf{Q}$ in (6) during iterations. These convergence problems can be overcome by resetting $\mathbf{Q} := \mathbf{Q}/c$ for some chosen constant $c > 1$, e.g. $c = 3$ in our application. Note that this does not affect consistency and asymptotic normality of $\hat{\beta}_{GEE}$, but efficiency .

Such convergence problems can be largely avoided by global cross–ratio modelling, at least with the choice $\mathbf{U}_i = \mathbf{I}$ for the second GEE as used here. We had similar experience with other data sets (Spatz, 1995). Parameter estimates are often quite near to those for the exchangeable correlation model, while standard errors are higher, but still smaller than for the IEE model due to improved efficiency.

In this application, all three models lead to very similar conclusions, see next section. Both the IEE and the exchangeable correlation model are simple to implement, results from the global cross–ratio model are more reliable due to higher efficiency and lack of convergence problems.

	Working association model						
	Independence			Exchangeable			
				Correlation		Gl. cross–ratio	
Covariates	Estim.	S.E. (naive)	S.E. (robust)	Estim.	S.E. (robust)	Estim.	S.E. (robust)
Threshold parameters							
	-0.635	0.067	0.117	-0.566	0.078	-0.576	0.097
	2.765	0.082	0.134	2.047	0.081	2.513	0.109
Canopy density							
very low	-0.685	0.157	0.313	-0.648	0.199	-0.644	0.252
low	0.182	0.082	0.137	0.159	0.097	0.160	0.117
medium	0.373	0.070	0.122	0.353	0.082	0.357	0.101
high	0.129	0.104	0.166	0.136	0.106	0.128	0.136
Mixture of stand							
coniferous	0.104	0.029	0.037	0.094	0.029	0.101	0.035
mixed	-0.104	0.029	0.037	-0.094	0.029	-0.101	0.035
Utilization Method							
2. commercial thinning	-0.544	0.078	0.134	-0.438	0.085	-0.514	0.109
1. commercial thinning	0.262	0.078	0.124	0.181	0.083	0.211	0.104
precommercial thinning	0.281	0.106	0.216	0.256	0.140	0.303	0.175
Site							
bedrock	-0.082	0.049	0.067	-0.047	0.052	-0.061	0.062
granite weathering	-0.134	0.060	0.080	-0.114	0.057	-0.122	0.072
gneiss weathering	-0.021	0.068	0.084	-0.029	0.064	-0.042	0.079
soil with water surplus	0.238	0.059	0.073	0.191	0.058	0.224	0.068
Altitude							
500-600m	0.342	0.086	0.106	0.296	0.081	0.318	0.100
601-650m	-0.106	0.059	0.078	-0.078	0.058	-0.093	0.071
651-700m	0.013	0.054	0.072	-0.015	0.056	-0.001	0.066
701-750m	-0.034	0.068	0.088	0.000	0.068	-0.009	0.080
750-900m	-0.215	0.079	0.107	-0.204	0.079	-0.215	0.097
Utilization Method $*$ Canopy density							
2.c.th./very low	0.600	0.170	0.328	0.599	0.208	0.556	0.265
1.c.th./very low	0.465	0.198	0.330	0.330	0.218	0.360	0.272
p.c.th./very low	-1.066	0.283	0.615	-0.930	0.389	-0.916	0.492
2.c.th./low	0.044	0.099	0.158	0.001	0.108	0.025	0.133
1.c.th./low	0.019	0.097	0.146	-0.007	0.105	0.011	0.126
p.c.th./low	-0.064	0.148	0.260	0.006	0.185	-0.036	0.221
2.c.th./medium	-0.305	0.091	0.147	-0.254	0.096	-0.257	0.121
1.c.th./medium	-0.091	0.087	0.136	-0.031	0.095	-0.048	0.116
p.c.th./medium	0.396	0.117	0.226	0.286	0.152	0.305	0.186
2.c.th./high	-0.339	0.164	0.255	-0.346	0.158	-0.324	0.206
1.c.th./high	-0.394	0.135	0.198	-0.291	0.130	-0.323	0.165
p.c.th./high	0.733	0.142	0.251	0.637	0.165	0.647	0.206

Table 1: Results

4.2 Interpretation of effects

Due to effect coding, parameter estimates for the categories of each covariate sum up to zero. High (positive) values indicate a positive influence on minor damage, while low (negative) values show positive influence on damage.

Mixture of stand

In the particular survey area, the probability for low damage is significantly higher for coniferous stands compared to mixed stands. For example, the estimated effect 0.101 in the cross–ratio model leads to an odds ratio increase of $1.1 = \exp(1.01)$ for low damage in coniferous stands.

Site

As to be expected, soil with water surplus is significantly beneficial for low damage of spruce. There is no significant difference between the influence of the remaining categories of site.

Altitude

Altitudes below 600 m are most favourable, on the other hand probability of high damage increases significantly above 750 m. Other altitudes have no significant influence.

Utilization method

The main effect of utilization shows that stands with second commercial thinning corresponding to higher age, have a clearly higher portion of damaged spruce. Although the difference between first and precommercial thinning is not significant, there is evidence of increasing damage with an increase of age.

Canopy density

Stands with very low density have distinctly increased probability for high damage, while medium canopy density is clearly beneficial.

*Utilization method * Canopy density*

Since there are significant interactions between categories of utilization method and canopy density, main and interaction effects should be interpreted together by adding them up. For example, stands with precommercial thinning with high or medium canopy density have a clearly lower probability of damage than indicated by the main effect alone. On the other side, the rather rare combination of precommercial thinning and very low canopy density is very unfavourable. In contrast, stands of higher age (first and second commercial thinning) are less affected by very low canopy density. Going through the other interactions, one may summarize as follows: Lower canopy density becomes more favourable with increasing age.

5 Conclusions

Marginal models for ordinal responses provide a useful tool for regression analyses in surveys where dependence among trees in clusters or plots has to be expected. If dependence is not of interest in itself, but is only regarded as a nuisance, the IEE method and the cross–ratio GEE1 method developed in this paper are particularly recommended, the former because of simplicity and the latter because of increased efficiency and reliability. The approach can also be applied to longitudinal data, and extensions to other settings, e.g., mixed discrete–continuous responses, should be of interest for future research.

References

Dale, J.R. (1986) Global cross–ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909-917.

Diggle, P.J., Liang,K.-L. and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* London, Chapman and Hall.

Fahrmeir, L. (1990) Maximum likelihood estimation in misspecified generalized linear models. *Statistics* **21**, 487-502.

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

Kublin, E. (1987) Statistische Auswertungsmodelle für Waldschadensinventuren – Methodische Überlegungen. *Forstwissenschaftliches Centralblatt* **106**, 57-68.

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

Liang, K.-Y., Zeger, S. and Qaqish, B. (1992) Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society* **B 54**, 3-40.

Lipsitz, S., Laird, N., and Harrington, D. (1991) Generalized estimation equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

Miller, M., Davies, S., and Landis, J., (1993) The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033-1044.

Molenberghs, G. and Lesaffre, E. (1994) Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* **89**, 633-644.

Mössmer, R. et al. (1991) Waldschäden im Luftbild: CIR–Luftbildauswertung 1991: Forstamt Flossenbürg. *Bayrische Forstliche Versuchs- und Forschungsanstalt*, Report, München.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1084.

Pritscher, E. (1992) Marginale Regressionsmodelle für multivariate kategoriale Variablen. Diplomarbeit, Institut für Statistik, Universität München.

Pritscher, E. , Bäumler, A. and Fahrmeir, L. (1994) Marginale Regressionsmodelle für ordinale Waldschadensdaten mit räumlicher Korrelation. *Forstwissenschaftliches Centralblatt* **113**, 367-378.

Quednau, H.D. (1989) Statistische Analyse von Waldschadensdaten aus Luftbildern mit Berücksichtigung von Nachbarschaftseffekten. *Forstwissenschaftliches Centralblatt* **108**, 96-102.

Spatz, R. (1995) Marginale Modellierung und Analyse kategorialer Längsschnittdaten. Diplomarbeit, Institut für Statistik, Universität München.

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

author's footnote

Ludwig Fahrmeir is Professor of Statistics at the Institute of Statistics, University of Munich, and chairman of the Sonderforschungsbereich 'Statistical analysis of discrete structures with applications in biometrics and econometrics', sponsored by the German Science Foundation DFG. He is also involved in joint projects and statistical consulting on medical, epidemiological and environmental studies as well as in forestry and social science applications.

Seminar für Statistik, Universität München, Ludwigsstr.33/II, D-80539 München, Germany, email: fahrmeir@stat.uni-muenchen.de

Lisa Pritscher is research assistant at the University of Munich with support of the German Science Foundation DFG. While preparing her master thesis she co-operated with the Bavarian State Institute of Forestry. She also was occupied with epidemiological projects.

Seminar für Statistik, Universität München, Ludwigsstr.33/III, D-80539 München, Germany, email: lisa@stat.uni-muenchen.de

Acknowledgement: We are grateful to A. Bäumler, providing the data, and R. Spatz, for computational assistance.