Gieger:

Marginal Regression Models with Varying Coefficients for Correlated Ordinal Data

Sonderforschungsbereich 386, Paper 177 (1999)

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

Projektpartner

Marginal Regression Models with Varying Coefficients for Correlated Ordinal Data

Christian Gieger[1]

Institut für Statistik, Ludwig-Maximilians-Universität München

Ludwigstr. 33, 80539 München, Germany

SUMMARY

This paper discusses marginal regression models for repeated or clustered ordinal measurements in which the coefficients of explanatory variables are allowed to vary as smooth functions of other covariates. We model the marginal response probabilities and the marginal pairwise association structure by two semiparametric regressions. To estimate the fixed parameters and varying coefficients in both models we derive an algorithm that is based on penalized generalized estimating equations. This allows to estimate the marginal model without specifying the entire distribution of the correlated categorical response variables. Our implementation of the estimation algorithm uses an orthonormal cubic spline basis that separates the estimated varying coefficients into a linear part and a smooth curvature part. By avoiding an additional backfitting step in the optimization procedure we are able to compute a robust approximation for the covariance matrix of the final estimate. We illustrate our method by an application to longitudinal data from a forest damage survey. We show how to model the dependence of damage state of beeches on non-linear trend functions and time-varying effects of age.

[1]email: gieger@stat.uni-muenchen.de

Keywords: Ordinal response, semiparametric marginal regression models, varying coefficients models, penalized generalized estimating equations, smoothing splines, forest damage data.

1 Introduction

Recently, marginal regression models for correlated ordinal outcomes have been proposed by several authors, e.g. Heagerty and Zeger (1996), Molenberghs and Lesaffre (1994, 1999), and Fahrmeir and Pritscher (1996). These models are multivariate extensions of univariate models for ordinal outcomes, e.g. cumulative logistic regression models (McCullagh, 1980). Fahrmeir and Pritscher (1996) developed multicategorical generalized estimating equations (GEE1) for ordinal responses by extending a generalized estimating equations approach for binary responses. Their GEE1 approach is defined by two distinct parametric regressions for marginal means as well as for marginal odds ratios that are measures of the pairwise association structure. A cumulative logistic model is applied to estimate the mean structure, whereas the logarithms of global odds ratios are used to construct a link function for the marginal pairwise association structure. For binary data their approach reduces to a GEE1 with a marginal logistic parameterization of the mean structure and a marginal odds ratio parameterization of the association structure (e.g. Lipsitz, Laird and Harrington, 1991).

In this paper, we extend the parametric marginal model of Fahrmeir and Pritscher (1996) to a semiparametric marginal model with varying coefficients. This means we allow that the effects of some or even all covariates in both regressions vary smoothly as functions of other covariates. For instance in the forest damage study considered in this paper, we allow

that the time-trend and the effect of age of the trees on the damage state depends on the the observed year. What we get, is a special kind of multiplicative interaction between the covariates age and calendar time. For the marginal pairwise association of two responses, we assume a model that specifies the marginal odds ratios as smooth functions of the time-lag. Models of this type for cross-sectional data have been first considered by Hastie and Tibshirani (1993). This approach yields a very flexible modeling framework with semiparametric predictors for both, the mean structure and the association structure. For joint estimation of fixed effects and varying coefficients in both models, we derive penalized generalized estimating equations (PGEE1). A PGEE1 approach is a good choice if the focus of attention is directed to a correct specification of the marginal mean model. This means for categorical outcomes that we are interested in the influence of covariates on the marginal probabilities of the response categories. For instance in our application, we model the marginal probability of trees being in a particular damage class. As in GEE1 for the estimation of fixed effects, PGEE1 approaches require only the parametrization of first and second order moments. For this reason they allow to estimate the effects even when large clusters or many repeated measurements are observed. By way of contrast this is computationally not possible for marginal categorical models based on full likelihoods. These models with parametric (see Heumann, 1996, 1997) or semiparametric predictors (see Gieger, 1998) are only suitable for a moderate number of correlated responses.

There are several related models with nonparametric components proposed for correlated categorical outcomes. Wild and Yee (1996) presented an additive extension of a generalized estimating equation approach for correlated binary data. To describe the mean and association structure they used an additive logit model and an additive model for the log odds ratios. Their approach for binary outcomes is a special case of our multivariate

semiparametric model. Berhane and Tibshirani (1998) presented a more general additive model in the context of exponential family models. They discussed how generalizations of quasi-likelihood methods can be used to estimate additive models for correlated responses. Semiparametric modeling of predictors in estimating equations based on local regression techniques has recently been considered by Carroll, Ruppert and Welsh (1998) and more specifically for longitudinal data with ordinal responses by Kauermann (1999). Fahrmeir, Gieger and Heumann (1999) discussed semiparametric marginal modeling of dependent ordinal responses by penalty approaches in the context of a clinical study. They showed how the model which we present here in detail can be adapted to situations with an isotonic response pattern. Finally, Heagerty and Zeger (1998) proposed a nonparametric model for the association if scientific interest is focused on the dependence structure. Apart from the fact that we consider a very general semiparametric model with varying coefficients, our model differs from the others in the implementation of the method: The estimation algorithm is based on an orthonormal cubic spline basis that separates the estimated varying coefficients into a linear part and a smooth spline part. This spline basis was first proposed by Demmler and Reinsch (1975). By avoiding a backfitting step in the optimization procedure we are immediately able to compute a robust approximation for the covariance matrix of the final estimate.

In section 2 of this paper we describe the semiparametric model and the PGEE1 approach for the marginal mean structure. The model for the association structure is considered in section 3. Section 4 describes a (quasi-)Fisher-Scoring algorithm which allows to estimate both regression models simultaneously. In section 5 the semiparametric marginal modeling approach is illustrated by an analysis of data from a forest damage study. Finally, we give a conclusion in section 6. In the following we discuss the idea of

semiparametric marginal modeling in the context of a longitudinal study but extensions to more general clustered settings are obvious (see Gieger, 1998).

2 Marginal regression models for the mean

2.1 Model specification for the mean

We suppose that a study has been conducted with N subjects. For each subject i we observe T responses Y_{it} with $q+1$ ordered categories together with covariate information $x_{it} = (x_{it1}, \ldots, x_{itp})'$. Thus the data are given by (Y_{it}, x_{it}), $i = 1, \ldots, N$, $t = 1, \ldots, T$. Without loss of generality we denote the ordered categories of Y_{it} by $1, \ldots, q+1$. In addition we assume that the covariates are either non-stochastic or stochastic but external, i.e. their values are not influenced by outcomes of the response variables.

As in our application, the influence of covariates x_{it} on the marginal probabilities of the response categories $\pi_{itr} = \mathrm{pr}\,(Y_{it} = r)$ with $r = 1, \ldots, q$ is often of prime interest, whereas the association between the responses can be regarded as a nuisance parameter. Consequently, we are mainly interested in a correct specification of the model for the marginal mean structure of Y_{it}. As link function for the mean structure, we use a cumulative logistic link

$$\mathrm{logit}\,(\mathrm{pr}(Y_{it} \leq r)) = \log\left(\frac{\mathrm{pr}(Y_{it} \leq r)}{1 - \mathrm{pr}(Y_{it} \leq r)}\right) = \eta_{itr}, \qquad r = 1, \ldots, q\,, \tag{1}$$

where η_{itr} is the predictor of category r. Instead of a logistic link, one can also use other link functions known from the analysis of ordinal data with generalized linear models (see Fahrmeir and Tutz, 1997, and Molenberghs and Lesaffre, 1999 for more detailed

discussions). Inversion of the link function (1) yields the response function

$$\pi_{it1} = \frac{\exp(\eta_{it1})}{1+\exp(\eta_{it1})} \quad \text{and} \quad \pi_{itr} = \frac{\exp(\eta_{itr})}{1+\exp(\eta_{itr})} - \frac{\exp(\eta_{itr-1})}{1+\exp(\eta_{itr-1})}, \quad r = 2, \ldots, q, \qquad (2)$$

which allows to compute the marginal probabilities $\pi_{itr} = \text{pr}(Y_{it} = r)$ in terms of the predictor η_{itr}.

To complete the mean model, we have to specify the functional form of the predictor η_{itr}. In this paper, we choose a semiparametric model

$$\eta_{itr} = u'_{itr}\alpha + w'_{itr}f(v_{it}), \qquad r = 1, \ldots, q, \qquad (3)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots)'$ is a vector of fixed effects, $f(v_{it}) = (f_1(v_{it1}), f_2(v_{it2}), \ldots)'$ is a vector of unknown smooth functions, and u_{itr}, w_{itr} are category-specific design vectors constructed from basic covariates in x_{it}. In (3) we have supplemented a parametric predictor $u'_{itr}\alpha$ by a nonparametric predictor $w'_{itr}f(v_{it})$, which allows that the effects of the design variables w_{itr} vary in dependence of continuous covariates $v_{it} = (v_{it1}, v_{it2}, \ldots)'$. This additional set v_{it} of covariates is formally a subset of x_{it}. If we delete the nonparametric part, then (3) reduces to the predictor of a parametric marginal model (e.g. Fahrmeir and Pritscher, 1996). For $v_{it1} = v_{it2} = \ldots = t$ with time points $t = 1, \ldots, T$ we can regard the model as a dynamic marginal regression model with parameters changing with time. If all design variables in w_{itr} are constants, e.g. $w_{itr} = (1, 1, \ldots)'$, then we get a (marginal) additive model in the terminology of Hastie and Tibshirani (1990) or Berhane and Tibshirani (1998). If w_{itr} is a vector of 0/1–variables generated by binary coding of a factor variable, then there is an important way of thinking about the model term $w'_{itr}f(v_{it})$: A separate curve corresponds to each of the levels (with exception of the reference level) of the factor variable. In fact, it is often useful to group a continuous covariate into a number of intervals to explore interactions with other continuous covariates in this special

way. In the case of repeated measurements we can specify nonparametric influences of covariates which vary over time by this binary coding procedure. In the same way we get separate predictor components for each category of the response by introducing binary indicators depending on the actual category. For instance this is the way how we can specify time-dependent threshold functions, say $\theta_1(t), \ldots, \theta_q(t)$, for the transition from one category to the next higher category of the cumulative model.

2.2 Design of the mean model

A complete description of the structural component requires the construction of a design matrix. We indicate in this section how smoothness assumptions and roughness penalties for the unknown functions in the semiparametric predictor (3) lead to a finite dimensional design.

To simplify the description, we first consider the problem of estimating a logistic model for a single binary response. We assume a logistic mean structure of the form

$$\text{logit}(\text{pr}(Y_i = 1)) = \log\left(\frac{\text{pr}(Y_{it} = 1)}{1 - \text{pr}(Y_{it} = 1)}\right) = \alpha + w_i f(v_i), \qquad i = 1, \ldots, N, \qquad (4)$$

where α is a fixed intercept parameter, w_i is a design variable, and f is an unknown function of a covariate v_i. A popular strategy to estimate the unspecified function f nonparametrically is to assume that the function belongs to a class of smooth functions. In this paper we suppose that the function f has continuous first and second derivatives f', f'', and f'' being quadratically integrable. Then an appropriate estimating function can be supplemented by an additive roughness penalty which measures the curvature and penalizes a too rough behavior of f. A natural roughness penalty for our specific space

of smooth functions is the integrated squared second derivative

$$J_\lambda(f) = \frac{\lambda}{2} \int_{-\infty}^{\infty} \{f''(v)\}^2 \, dv. \tag{5}$$

This penalty assesses the total curvature in f, or alternatively, the degree to which f departs from a straight line. The smoothing parameter λ controls the influence of the penalty term on the estimating criterion. A large value of λ gives large weight to the penalty term, therefore enforcing a smooth function with small variance but possibly high bias. For a small λ the penalty term has less influence and we allow more faith with the data, which is measured by the estimating function. The result is a rough estimate with possibly high variance but reduced bias. There are several methods to determine the smoothing parameter λ automatically, e.g. crossvalidation (see Gieger,1998). But these methods typically work only for models with a moderate number of smoothing parameters. For this reason, we use a more explorative method: In our application we choose a reasonable value for each smoothing parameter after trying out several alternative values.

It is well known (e.g. Eubank, 1988) that the solution of a penalized estimating criterion with roughness penalty (5) is a natural cubic spline with knots at each distinct point of v_i, $i = 1, \ldots, N$. For the moment we can assume that each observation v_i in the sample is unique. Then the space of natural cubic splines is finite dimensional with dimension N equal to the number of observations. Consequently, we can write the function f as linear combination $f(v) = \phi(v)'\gamma$ of spline basis functions $\phi(v) = (\phi_1(v), \ldots, \phi_N(v))'$. Now the estimation problem is finite dimensional and we have to estimate the basis coefficients $\gamma = (\gamma_1, \ldots, \gamma_N)'$ instead of the function f. In this paper, we choose a cubic spline basis

of orthonormal functions $\phi_k(\cdot)$, $k = 1, \ldots, N$. They are defined by

$$\sum_{i=1}^{N} \phi_k(v_i)\phi_m(v_i) = I(k = m), \tag{6}$$

$$\int_{-\infty}^{\infty} \phi_k''(v)\phi_m''(v)dv = \rho_k I(k = m), \tag{7}$$

with $k, m = 1, \ldots, N$. This basis was first considered by Demmler and Reinsch (1975). It has the property that it decomposes the spline space into a space of linear functions in v and a curvature part formed by centered cubic splines. To illustrate this behaviour we assume that the basis functions are ordered by their roughness $\rho_k = \int_{-\infty}^{\infty} \{\phi_k''(v)\}^2 dv$, so that we get $0 = \rho_1 = \rho_2 < \rho_3 < \cdots < \rho_N$. Figure 1 (a) shows for $N = 100$ and $v_i = i/100$ the functions $\phi_1(\cdot)$, $\phi_2(\cdot)$, $\phi_3(\cdot)$, $\phi_5(\cdot)$, $\phi_{10}(\cdot)$ and $\phi_{50}(\cdot)$. The first and second basis element (solid lines) define the linear part of the spline. They are not penalized by the roughness penalty (5). The remaining functions are natural cubic splines, which oscillate around the abscissa with increasing frequency. Demmler and Reinsch (1975) indeed showed that for $k \geq 3$ the number of sign changes in the k-th function $\phi_k(\cdot)$ is exactly $k - 1$. This behaviour of the Demmler-Reinsch basis results in a clear interpretation of the coefficients $\gamma = (\gamma_1, \ldots, \gamma_N)'$.

– Figure 1 about here –

For the computation of the spline we have to give a guidance how to determine the basis functions $\phi_k(\cdot)$ and the roughness measures ρ_k. A well known finite-dimensional representation of the penalty (5) in terms of function evaluations $f = (f(v_1), \ldots, f(v_N))'$ is given by $J_\lambda(f) = \lambda f' K f$ where the $N \times N$ penalty matrix K is defined by second differences. The computation of K has been described for example by Green and Silverman (1994). The required basis functions $\phi_k(\cdot)$ are the interpolated eigenvectors of an eigendecomposition of K and the roughness penalties ρ_k are the corresponding eigenvalues.

To get a finite-dimensional version of the model we insert the evaluated basis functions $\phi_{ik} = \phi_k(v_i)$ into the mean model (4) and the penalty (5). This yields the model

$$\text{logit}(\text{pr}(Y_i = 1)) = z_i'\beta \tag{8}$$

with the design vector $z_i = (1, w_i\phi_{i1}, \ldots, w_i\phi_{iN})'$ and a high–dimensional parameter vector $\beta = (\alpha, \gamma_1, \ldots, \gamma_N)'$. Now the penalty term has the finite-dimensional form

$$J_\lambda(f) = \frac{\lambda}{2} \sum_{k=1}^{N} \rho_k \gamma_k^2 \,. \tag{9}$$

Figure 1 (b) shows the positive values of the roughness measures ρ_k on a logarithmic scale. Together with (9) we see that coefficients γ_k corresponding to basis functions with a high frequency are penalized strongly by the estimating criterion. We refer to Hastie (1996) for a more detailed discussion of this shrinkage behavior of a spline smoother.

Now we can return to the more general problem of finding the finite dimensional design of the semiparametric predictor (3). First we represent each function in (3) as a linear combination $f_j(v_j) = \phi_j(v_j)'\gamma_j,\ j = 1, \ldots, p$ of cubic spline basis functions $\phi_j(v_j) = (\phi_{j1}(v_j), \ldots, \phi_{jN_j}(v_j))'$ and corresponding coefficients $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jN_j})'$. In general the dimensions N_j, $j = 1, \ldots, p$, of the spline spaces are different depending on the numbers of distinct values of the covariates in the sample. In practical use the occurrence of repeated observations often results in a considerable reduction of the dimension of the estimation problem.

After defining the design vectors $z_{itrj} = (w_{itrj}\phi_{itj1}, \ldots, w_{itrj}\phi_{itjN_j})'$ in terms of evaluated basis functions $\phi_{itjk} = \phi_{jk}(v_{itj})$, we get the finite-dimensional representation

$$\eta_{itr} = u_{itr}'\alpha + z_{itr1}'\gamma_1 + \ldots + z_{itrp}'\gamma_p, \qquad r = 1, \ldots, q \tag{10}$$

of the semiparametric predictor (3). Gathering all vectors together yields the new design vector $z_{itr} = (u_{itr}', z_{itr1}', \ldots, z_{itrp}')'$ and the parameter vector $\beta = (\alpha', \gamma_1', \ldots, \gamma_p')'$. Now we

can write the complete predictor $\eta_i = (\eta_{i1}', \ldots, \eta_{it}', \ldots, \eta_{iT}')'$ with $\eta_{it} = (\eta_{it1}, \ldots, \eta_{itq})'$ of the multivariate response $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ as

$$\eta_i = Z_i \beta \tag{11}$$

with the design matrix $Z_i = (Z_{i1}', \ldots, Z_{it}', \ldots, Z_{iT}')'$ defined by $Z_{it} = (z_{it1}, \ldots, z_{itq})'$. What we get is again a parametric predictor η_i with a high-dimensional parameter vector β.

Often there are situations where we have to modify the construction of the design matrix Z_i slightly. To illustrate the problem we consider again a simple logit model

$$\text{logit}\left(\text{pr}(Y_i = 1)\right) = \alpha + \gamma_1(v_{i1}) + \gamma_2(v_{i2}), \tag{12}$$

with a fix intercept parameter α and two additive functions $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$. The construction of the design for this model as described yields a design matrix in which the constant basis functions of $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are redundant. To get an unique solution we have to reduce the dimensions of the bases, i.e. we have to delete both constant basis functions. Here it pays off that the linear parts of the splines occur in the Demmler-Reinsch bases as elements. In general we have to reduce the sets of basis functions which define the linear parts of the splines until there are no more redundant elements. More formally, Hastie and Tibshirani (1993) and Wahba (1990) have shown that an unique solution exists if the corresponding embedded parametric model, where we have restricted each function $\gamma_j(\cdot)$ to be linear, has an unique solution.

Finally, we consider the penalty term

$$J_{\lambda_1,\ldots,\lambda_p}(f_1, \ldots, f_p) = \sum_{j=1}^{p} \frac{\lambda_j}{2} \int_{-\infty}^{\infty} \left\{ f_j''(v_j) \right\}^2 dv_j \tag{13}$$

corresponding to the semiparametric predictor (3). After inserting the finite-dimensional

representations $\sum_{k=1}^{N_j} \rho_{jk}\gamma_{jk}^2$, $j = 1, \ldots, p$, into (13) we get the penalty in matrix notation

$$J_{\lambda_1,\ldots,\lambda_p}(\beta) = \frac{1}{2}\beta' \Lambda \mathrm{P} \beta \,, \tag{14}$$

where the diagonal matrix Λ contains the smoothing parameters. Furthermore, the diagonal penalty matrix P has zeros on the diagonal if the corresponding parameters are not penalized, i.e. for fixed effects and linear basis functions. Otherwise the diagonal elements of P are equal to the roughness measures $\rho_{jk} = \int_{-\infty}^{\infty} \left\{ \phi_{jk}''(v_j) \right\}^2 dv_j$.

2.3 Penalized generalized estimating equations for the mean

To estimate the unknown parameters and functions of the marginal mean model we propose using penalized generalized estimating equations (PGEE1). In the following, y_{it} is the vector of indicator variables $y_{itr} = I(Y_{it} = r)$ for the observed categories $r = 1, \ldots, q$ of subject i at time t and π_{it} is the corresponding vector of probabilities $\pi_{itr} = \mathrm{pr}(y_{itr} = 1)$ derived from model (3). Then the PGEE1 for the marginal mean model is

$$u(\beta) = \sum_{i=1}^{N} Z_i' D_i V_i^{-1} (y_i - \pi_i) - \Lambda \mathrm{P} \beta = 0. \tag{15}$$

with $y_i = (y_{i1}', \ldots, y_{iT}')'$ and $\pi_i = (\pi_{i1}', \ldots, \pi_{iT}')'$. The first term in (15) has the common form of GEE's, where D_i is a blockdiagonal matrix with the first derivatives $\partial \pi_{it} / \partial \eta_{it}$, $t = 1, \ldots, T$, on the diagonal and V_i is a working–covariance matrix. The second term is the first derivative of the quadratic penalty term $\frac{1}{2}\beta' \Lambda \mathrm{P} \beta$ in (14).

In a PGEE1 it is not necessary for V_i to be equal to the true covariance matrix, say Σ_i, of the multivariate response y_i. If $V_i = \Sigma_i$ holds equation (15) is equal to a penalized estimating equation for the marginal mean derived from a likelihood-based model (see Gieger, 1998). In general however $V_i \neq \Sigma_i$ resulting in some loss of efficiency depending on the degree of misspecification. In the binary case a GEE1 or PGEE1 analysis is often

based on the working assumptions of independence. Letting $V_{it} = \text{diag}(\pi_{it}) - \pi_{it}\pi_{it}'$ denote the covariance matrix of the response y_{it}, we obtain by setting $V_i = \text{diag}(V_{i1}, \ldots, V_{iT})$ the simplest model for the covariance structure which assumes independence between the observations of subject i. We do not use this assumption in our multicategorical setting because we have found that this model yields a too crude approximation. In contrast to the binary case where an independence assumption usually has only minor effects on the point estimates, this is often not true for more than two categories (see also Fahrmeir, Gieger and Heumann, 1999). Instead of using the independence model, we supplement the marginal mean model by a second model for the pairwise association structure.

3 Marginal regression models for the association

3.1 Model specification

Common measures for the association between two ordinal responses Y_{is} and Y_{it} are global odds ratios (Dale, 1986, Fahrmeir and Pritscher, 1996). For each pair of categories $l, r = 1, \ldots, q$ they are given by

$$\psi_{i,st,lr} = \frac{\text{pr}(Y_{is} \leq l, Y_{it} \leq r)\,\text{pr}(Y_{is} > l, Y_{it} > r)}{\text{pr}(Y_{is} > l, Y_{it} \leq r)\,\text{pr}(Y_{is} \leq l, Y_{it} > r)}. \tag{16}$$

This means that the $(q+1) \times (q+1)$ contingency table of probabilities $\pi_{i,st,lr} = \text{pr}(Y_{is} = l, Y_{it} = r)$ is collapsed at each cutpoint (l, r) into a 2×2 contingency table. Then common odds ratios are computed out of these coarser tables. Together with marginal cumulative probabilities $\xi_{isl} = \text{pr}(Y_{is} \leq l)$ and $\xi_{itr} = \text{pr}(Y_{it} \leq r)$, global odds ratios form an unique reparametrization of the bivariate distribution of Y_{is} and Y_{it}. By solving (16) we can express the bivariate cumulative probability function $\xi_{i,st,lr} = \text{pr}(Y_{is} \leq l, Y_{it} \leq r)$ in terms

of corresponding global odds–ratios $\psi_{i,st,lr}$ and the marginal cumulative probabilities ξ_{isl} and ξ_{itr}. This yields

$$\xi_{i,st,lr} = \begin{cases} \xi_{isl}\xi_{itr} & , \text{if} \quad \psi_{i,st,lr} = 1, \\ \dfrac{\kappa - \sqrt{\kappa^2 + 4\psi_{i,st,lr}(1-\psi_{i,st,lr})\xi_{isl}\xi_{itr}}}{2(\psi_{i,st,lr}-1)} & , \text{if} \quad \psi_{i,st,lr} \neq 1, \end{cases} \tag{17}$$

where $\kappa = 1 + (\xi_{isl} + \xi_{itr})(\psi_{i,st,lr} - 1)$. With this formula we can calculate bivariate probabilities $\pi_{i,st,lr} = E(y_{isl}y_{itr}) = \text{pr}(Y_{is} = l, Y_{it} = r)$ in terms of univariate marginal cumulative probabilities ξ_{isl}, ξ_{itr} and global odds ratios $\psi_{i,st,lr}$ through the relation

$$\pi_{i,st,lr} = \begin{cases} \xi_{i,st,lr} & , l = r = 1 \\ \xi_{i,st,lr} - \xi_{i,st,lr-1} & , l = 1, r > 1 \\ \xi_{i,st,lr} - \xi_{i,st,l-1r} & , l > 1, r = 1 \\ \xi_{i,st,lr} - \xi_{i,st,lr-1} - \xi_{i,st,l-1r} + \xi_{i,st,l-1r-1} & , l > 1, r > 1. \end{cases} \tag{18}$$

This means with (17) and (18) it is possible to compute the off–diagonal elements of the model covariance matrix V_i with $\text{cov}(y_{isl}, y_{itr}) = \pi_{i,st,lr} - \pi_{isl}\pi_{itr}$.

Now we can complete the parametrization of the marginal model by using the logarithms of the global odds ratios as link function for the marginal association structure of Y_{is} and Y_{it}, $s < t = 1, \ldots, T$. We get

$$\log(\psi_{i,st,lr}) = \tilde{\eta}_{i,st,lr}\,, \qquad l, r = 1, \ldots, q, \tag{19}$$

where $\tilde{\eta}_{i,st,lr}$ is the predictor of the association model. Additionally, with (17) and (18) we have explicit formulas for the response function.

Again we assume a semiparametric model

$$\tilde{\eta}_{i,st,lr} = \tilde{u}'_{i,st,lr}\tilde{\alpha} + \tilde{w}'_{i,st,lr}\tilde{f}(\tilde{v}_{i,st})\,, \tag{20}$$

where $\tilde{u}_{i,st,lr}$, $\tilde{w}_{i,st,lr}$ are design vectors of the association structure, $\tilde{\alpha} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots)'$ is a vector of fixed association parameters, and $\tilde{f}(\tilde{v}_{i,st}) = (\tilde{f}_1(\tilde{v}_{i,st,1}), \tilde{f}_2(\tilde{v}_{i,st,2}), \ldots)'$ is a vector

of unknown smooth functions in terms of covariates $\tilde{v}_{i,st} = (\tilde{v}_{i,st,1}, \tilde{v}_{i,st,2}, \ldots)'$, which in general are constructed from basic covariates x_{is} and x_{it}. By using spline representations of the unspecified functions in (20) we can again construct a finite dimensional predictor

$$\tilde{\eta}_i = \tilde{Z}_i' \delta, \tag{21}$$

with an appropriate design matrix $\tilde{Z}_i$ and a parameter vector δ of the association structure.

3.2 Penalized generalized estimating equations for the association

To estimate the model, we augment the PGEE1 for the mean (15) by a second PGEE1 for the association structure. We get

$$u(\delta) = \sum_{i=1}^{N} \tilde{Z}_i' C_i U_i^{-1} (\tilde{y}_i - \nu_i) - \Delta\Omega\,\delta = 0, \tag{22}$$

where $\tilde{y}_i = (\ldots, \tilde{y}_{i,st,lr}, \ldots)'$ is the vector of centered products $\tilde{y}_{i,st,lr} = (y_{isl} - \pi_{isl})(y_{itr} - \pi_{itr})$, $l, r = 1, \ldots, q$, $s < t = 1, \ldots, T$ and $\nu_i = (\ldots, \nu_{i,st,lr}, \ldots)'$ is the vector of covariances $\nu_{i,st,lr} = \pi_{i,st,lr} - \pi_{isl}\pi_{itr}$. The matrix C_i is the first derivative of ν_i with respect to $\tilde{\eta}_i$ and U_i is a further working covariance matrix. As in the binary case (see Prentice, 1988) a simple but useful working assumption is $U_i = \text{diag}(\text{var}(\tilde{y}_i))$ with diagonal elements

$$\text{var}(\tilde{y}_{i,st,lr}) = \pi_{isl}(1 - \pi_{isl})\pi_{itr}(1 - \pi_{itr}) - \nu_{i,st,lr}^2 + \nu_{i,st,lr}(1 - 2\pi_{isl})(1 - 2\pi_{itr}).$$

The penalty term of (22) consists of a second diagonal matrix Δ of smoothing parameters and a diagonal matrix Ω of penalty terms corresponding to the association model.

4 Estimation of the marginal model

To estimate the marginal model we can adapt the Fisher-Scoring procedure for the estimation of a parametric model (see Fahrmeir and Pritscher, 1996) to the nonparametric case. The (quasi)–Fisher–Scoring step for the PGEE1 approach is

$$\left(\sum_{i=1}^{N} Z_i' D_i^{(k)} \left(V_i^{(k)}\right)^{-1} D_i^{(k)'} Z_i + \Lambda \mathrm{P}\right) \left(\beta^{(k+1)} - \beta^{(k)}\right) = u\left(\beta^{(k)}\right) \tag{23}$$

$$\left(\sum_{i=1}^{N} \tilde{Z}_i' C_i^{(k)} \left(U_i^{(k)}\right)^{-1} C_i^{(k)'} \tilde{Z}_i + \Delta \Omega\right) \left(\delta^{(k+1)} - \delta^{(k)}\right) = u\left(\delta^{(k)}\right) . \tag{24}$$

We get the PGEE1 estimates $(\hat{\beta}, \hat{\delta})$ by switching between the equations (23) and (24) until convergence. Note here that the solution of the PGEE1 for β depends on the association parameters δ only through the working covariance V_i. Therefore like in parametric GEE1 approaches the estimator for the marginal mean model should be robust against a moderate misspecification of the association structure.

By defining working-observations $y_i^* = Z_i\beta + (D_i^{-1})'(y_i - \pi_i)$ and $\tilde{y}_i^* = \tilde{Z}_i\alpha + (C_i^{-1})'(\tilde{y}_i - \nu_i)$ we can rewrite (23) and (24) as iteratively reweighted penalized least-squares estimators

$$\left(\sum_{i=1}^{N} Z_i' D_i^{(k)} \left(V_i^{(k)}\right)^{-1} D_i^{(k)'} Z_i + \Lambda \mathrm{P}\right) \beta^{(k+1)} = \sum_{i=1}^{N} Z_i' D_i^{(k)} \left(V_i^{(k)}\right)^{-1} D_i^{(k)'} \tilde{y}_i^{(k)} \tag{25}$$

$$\left(\sum_{i=1}^{N} \tilde{Z}_i' C_i^{(k)} \left(U_i^{(k)}\right)^{-1} C_i^{(k)'} \tilde{Z}_i + \Delta \Omega\right) \alpha^{(k+1)} = \sum_{i=1}^{N} \tilde{Z}_i' C_i^{(k)} \left(U_i^{(k)}\right)^{-1} C_i^{(k)'} \tilde{y}_i^{*(k)} . \tag{26}$$

By switching between (25) and (26) we get an estimation procedure which is simple to implement and numerically stable. The algorithm for the computation of the estimates $(\hat{\beta}, \hat{\delta})$ can be summarized as follows:

1. Compute the basis functions of the natural cubic splines and construct the design matrices Z_i and $\tilde{Z}_i$. Use the eigenvalues to form the penalty matrices P and Ω.

2. Obtain initial values $(\beta^{(0)}, \delta^{(0)})$. One can use $\beta^{(0)}$ resulting from a regression assuming independence and $\delta^{(0)} = 0$.

3. Use (2), (17), and (18) to obtain the marginal probabilities of first and second order. These probabilities can be used to get current estimates of $V_i^{(k)}$.

4. Get an update $\beta^{(k+1)}$ of $\beta^{(k)}$ by taking a (quasi–)Fisher–Scoring step (23) or alternatively by solving the iterative reweighted penalized least–squares criterion (25).

5. Repeat step 3 to obtain updated estimates for the marginal probabilities. These probabilities can be used to get current estimates of $U_i^{(k)}$ and $\nu_i^{(k)}$

6. Get an update $\delta^{(k+1)}$ of $\delta^{(k)}$ by taking a (quasi–)Fisher–Scoring step (24) or alternatively by solving the iterative reweighted penalized least–squares criterion (26).

7. Iterate steps 3 to 6, until a convergence criterion is fulfilled.

A rigorous asymptotic theory for models using roughness penalty approaches is still not available. To get a robust approximation for the covariance matrix of the final PGEE1 estimate $\hat{\beta}$ we use a nonparametric version of the well–known sandwich matrix $V_{rob}(\hat{\beta}) = H^{-1}V^*H^{-1}$ with $H = \sum_{i=1}^N Z_i D_i' V_i^{-1} D_i Z_i' + \Lambda \mathrm{P}$ and with an empirical covariance estimator $V^* = \sum_{i=1}^N Z_i D_i' V_i^{-1} (y_i - \hat{\pi}_i) (y_i - \hat{\pi}_i)' V_i^{-1} D_i Z_i'$. The empirical covariance V^* is a correction term in the case of a misspecification of the association structure (see Royall, 1986). As McCullagh in a discussion of an article by Fitzmaurice, Laird and Rotnitzky (1993), we take the view that in the case of a clear misspecification of the association structure which is indicated by a distinct correction of the naive covariance approximation $V(\hat{\beta}) = H^{-1}$, we should look for a better model for the association. This means that we compare the model–based or naive standard error based on $V(\hat{\beta})$ and the robust standard error based on $V_{rob}(\hat{\beta})$, and interpret the difference as degree of misspecification.

5 Application to a forest damage study

Since 1983 a yearly visual forest damage inventory is carried out in a forest district in the northern part of Bavaria. There are 80 observation points with occurrence of beeches spread over the whole area. In this damage study we analyze the influence of covariates, e.g., age of the trees, pH value of the soil, and canopy density at the stand, on the defoliation of beeches at the stand. A detailed survey and data description can be found in Göttlein and Pruscha (1996).

We use the degree of defoliation as an indicator for damage state of the trees. Due to the survey design, responses must be assumed to be serially correlated. The ordinal response variable, Y_t, "damage state" at time t is measured in 3 categories: none ($Y_t = 1$), light ($Y_t = 2$), and distinct/strong ($Y_t = 3$ = reference) defoliation. Figure 2 shows the relative frequencies of the damage categories in the sample for the years 1983 to 1994.

– Figure 2 about here –

Due to the ordinal scale of the response, we can use a cumulative logistic model to relate the marginal probabilities of "damage state" to the following covariates:

A Age of the trees at the beginning of the study with categories: below 50 years (=1), between 50 and 120 years (=2), and above 120 years (=reference).

PH PH value of the soil in 0-2 cm depth. The measures range from a minimum of 3.3 to a maximum of 6.1.

CD Canopy density at the stand with categories: low (=1), medium (=2), and high (=reference).

The covariates pH value and canopy density vary for each stand over time, while the variable age is constant over time by construction. In particular, we assume for the marginal cumulative probabilities of no damage ($r = 1$) and none or light damage ($r = 2$) the following model

$$\text{logit}(\text{pr}(Y_t \leq r)) = \theta_r(t) + f_3(t)A^{(1)} + f_4(t)A^{(2)} + f_5(PH_t) + \alpha_6 CD_t^{(1)} + \alpha_7 CD_t^{(2)},$$

where $A^{(1)}, A^{(2)}, CD_t^{(1)}, CD_t^{(2)}$ are dummy variables for the categorical covariates A and CD. To capture the time trend in the data we allow the threshold functions and the effects of the time constant variable age to vary smoothly with time t. Due to a lack of information about the form of the influence, it is reasonable to model the effect of pH value nonparametrically by an unspecified smooth function. The effects of canopy density are assumed to be fixed like in an ordinary parametric model.

– Figure 3 about here –

Figures 3 shows the estimated threshold functions $\hat{\theta}_1(t)$ (left plot) and $\hat{\theta}_2(t)$ (right plot). Both curves decrease up to the year 1988 with a more pronounced decrease of the first threshold $\hat{\theta}_1(t)$. This indicates a shift to higher probabilities for the categories light and distinct/strong damage up to this year. After an improvement, i.e. a shift to the none damage category, up to 1992 there is another increase in damage up to 1994. This result is true for beeches above 120 years, i.e. for the reference category of age. For the other two categories of age we have in addition to consider the effects of the corresponding dummy variables $A^{(1)}$ and $A^{(2)}$. Both effects are positive over the 12 years (Figure 4, left plot). This indicates a positive influence on minor damage, i.e. younger beeches are less damaged. The positive effect of the category with below 50 years old trees (upper curve) is greater and the increase of the effect after 1988 corrects the change to the worse after

1992 indicated by the threshold functions. These interpretations are further illustrated by Figure 5, where the estimated distributions stratified by age are plotted against time. While the state of the younger tree population is very well recovered after a period with light damage, the state of older trees stays on a bad level.

– Figure 4 about here –

The estimated function for the influence of pH value is more or less linear over the range of observed pH values (Figure 4, right plot). Stands with low pH values have a negative influence on damage state compared to stands with less acid soils, i.e. low pH values aggravate the condition of the trees. But due to the flat course of the estimated curve, there is some doubt that pH–value has an influence at all. Finally, we get the following parameter estimates for the effects of canopy density together with model based and robust standard errors:

Covariate	Estimate	SE (model)	SE (robust)	p–value (model)	p–value (robust)
$CD_t^{(1)}$	-1.2822	0.3587	0.3104	0.0003	0.0000
$CD_t^{(2)}$	-0.5318	0.2481	0.2196	0.0320	0.0153

Both estimates are negative. This means that stands with low ($CD_t^{(1)}$) or medium ($CD_t^{(2)}$) canopy density have an increased probability for high damage compared to stands with a high canopy density. The reason could be that lower canopy densities result in rougher

conditions for the tree population connected with stronger physiological, aerodynamic and physical stress.

– Figure 5 about here –

With $T = 12$ measures per stand we have 66 time pairs at which we measure the pairwise association by global odds ratios. A preliminary descriptive analysis with empirically estimated global odds ratios indicates different values of the odds ratios for each cutpoint (l, r) and a decline in association with the time distance between the visits to the stand. Thus the association structure is parameterized by a logarithmic global odds ratio model of the form

$$\log(\psi_{i,st}^{(lr)}) = \tilde{f}_{lr}(|t - s|) \qquad l, r = 1, 2,$$

i.e. we do not force the dependence on the time lag $|t - s|$ into a specific parametric form. The estimates for the association functions (Figure 6) are quite similar in their global shape but the levels are different. We also recognize a distinct temporal structure. There is a decrease in the association between two responses as the time distance increases. Remarkable is the outlier for combination $(l = 2, r = 1)$ which is caused by a very irregular filled table. This value can not be covered by the association model but in total the serial structure seems to be appropriate modeled. The model based and robust standard errors for the curves and fixed parameters of the marginal mean model are close together. If we interpret the difference between both estimates as indicator for the degree of misspecification of the association structure then we have found an appropriate model for the association.

– Figure 6 about here –

This application shows that compared to purely parametric modeling, a semiparametric approach allows a refined and more flexible specification of the mean structure and a gain in efficiency due to an improved working association. We have found non-linear trend functions which would have been difficult to recognize and to model with parametric approaches. Time-dependent modeling of the age effects showed us a distinct different behaviour over calendar time of the age categories. Concering the influence of PH value our model allowed us to let the data decide about the appropriate form.

6 Conclusions

Inclusion of nonparametric predictors is an important aspect for adequate modelling in marginal regression. We discussed it for a PGEE1 approach but extensions to other settings are conceptually immediate. In particular, PGEE2 and full likelihood approaches (see Gieger, 1998) are interesting topics. But an important limitation is that these methods even in the parametric case only work for a moderate number of correlated observations. These means in practice that time-series like in our application with 12 observations per unit are computationally not manageable.

Acknowledgments

This work has been done as part of the author's PhD thesis at the University of Munich. During this time the author was supported by the Deutsche Forschungsgemeinschaft, SFB 386: "Statistical Analysis of Discrete Structures; Modelling and Application in Biometrics and Econometrics". I thank Axel Göttlein for making the forest damage data available, and Ludwig Fahrmeir, Christian Heumann, Geert Molenberghs, and Patrick Heagerty for helpful discussions.

References

BERHANE, K, TIBSHIRANI, R.J. (1998). Generalized Additive Models for Longitudinal Data. *The Canadian Journal of Statistics*, 26, 517–535.

CARROLL, R.J., RUPPERT, D., WELSH, A.H. (1998). Local Estimating Equations. *Journal of the American Statistical Association*, 93, 214–227.

DALE, J.R. (1986). Global Cross–Ratio Models for Bivariate, Discrete, Ordered Responses. *Biometrics*, 42, 909–917.

DEMMLER, A., REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik* 24, 375–382.

EUBANK, R.L. (1988). *Smoothing Splines and Nonparametric Regression.* Dekker, New York.

FAHRMEIR, L., PRITSCHER, L. (1996). Regression Analysis of Forest Damage by Marginal Models for Correlated Ordinal Responses. *Journal of Environmental and Ecological Statistics*, 3, 257–268.

FAHRMEIR, L., GIEGER, C., HEUMANN, C. (1999). An Application of Isotonic Longitudinal Marginal Regression to Monitoring the Healing Process. *Biometrics*, 55, 951–956.

FAHRMEIR, L., TUTZ, G. (1994, 1997 corrected third printing ed.). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer-Verlag.

FITZMAURICE, G.M., LAIRD, N.M., ROTNITZKY, A.G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science*, 8, 284–309.

GIEGER, C. (1997). Non- and Semiparametric Marginal Regression Models for Ordinal Response. *Discussion paper 71*, SFB 386, Ludwig–Maximilians Universität, München, URL: ftp://ftp.stat.uni-muenchen.de/pub/sfb386/paper71.ps.Z.

GIEGER, C. (1998). Marginale Regressionsmodelle mit variierenden Koeffizienten. Dissertation, Institut für Statistik, Universität München.

GÖTTLEIN, A., PRUSCHA, H. (1996). Der Einfluß von Bestandskenngrößen, Topographie, Standort und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch. *Forstwirtschaftliches Centralblatt*, 114, 146–162.

GREEN, P.J., SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

HASTIE, T. (1996). Pseudo Splines. *Journal of the Royal Statistical Society*, B 58, 379–396.

HASTIE, T., TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HASTIE, T., TIBSHIRANI, R. (1993). Varying–coefficient Models. *Journal of the Royal Statistical Society*, B 55, 757–796.

HEAGERTY, P., ZEGER, S. (1996). Marginal Regression Models for Clustered Ordinal Measurements. *Journal of the American Statistical Association*, 91, 1024–1036.

HEAGERTY, P., ZEGER, S. (1998). Lorelogram: A Regression Approach to Exploring Dependence in Longitudinal Categorical Responses. *Journal of the American Statistical Association*, 93, 150–162.

HEUMANN, C. (1996). Marginal Regression Modeling of Correlated Multicategorical Response: A Likelihood Approach. *Discussion paper 19*, SFB 386, Ludwig–Maximilians Universität, München,

URL: ftp://ftp.stat.uni-muenchen.de/pub/sfb386/paper19.ps.Z.

HEUMANN, C. (1997). Likelihoodbasierte marginale Regressionsmodelle für korrelierte kategoriale Daten. Dissertation, Institut für Statistik, Universität München.

KAUERMANN G. (1999). Modeling Longitudinal Data with Ordinal Response by Varying Coefficients. Forthcoming in *Biometrics*.

LIPSITZ, S., LAIRD, N., HARRINGTON, D. (1991). Generalized Estimation Equations for Correlated Binary Data: Using The Odds Ratio as a measure of Association. *Biometrika*, 78, 153–160.

MC CULLAGH, P. (1980). Regression Model for Ordinal Data (with discussion). *Journal of the Royal Statistical Society*, B 42, 109–127.

MOLENBERGHS, G., LESAFFRE, E. (1994). Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *Journal of the American Statistical Association*, 89, 633–644.

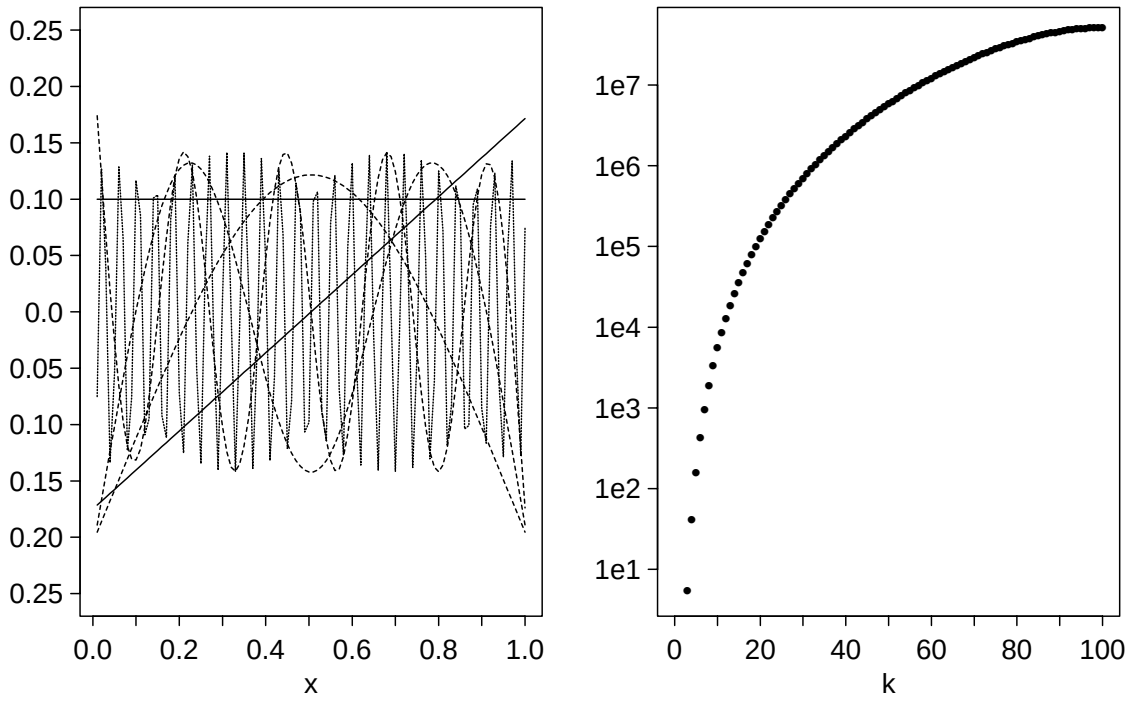
MOLENBERGHS, G., LESAFFRE, E. (1999). Marginal Modeling of Multivariate Categorical Data. *Statistics in Medicine*, 18, 2237–2255.

PRENTICE, R.L. (1988). Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics*, 44, 1033–84.

ROYALL, R.M. (1986). Model–Robust Confidence Intervals Using Maximum Likelihood Estimators. *International Statistical Review*, 54, 221–226.

WAHBA, G. (1990). *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.

WILD, C.J., YEE, T.W. (1996). Additive Extensions to Generalized Estimating Equation Methods. *Journal of the Royal Statistical Society*, B 58, 711–725.



(a) (b)

Figure 1: Demmler-Reinsch basis functions ϕ_1, ϕ_2, ϕ_3, ϕ_5, ϕ_{10}, and ϕ_{50} for $x_i = 1/100$, $i = 1, \ldots, 100$ (a) and positive roughness measure ρ_k of ϕ_k $k = 3, \ldots, 100$ (b).

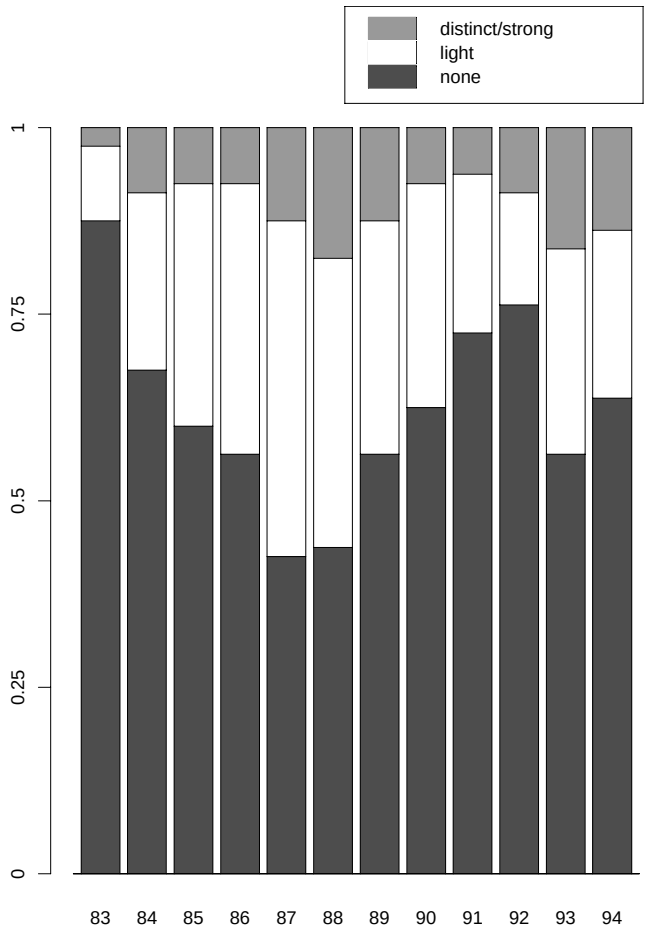


Figure 2: Damage class distribution by time.

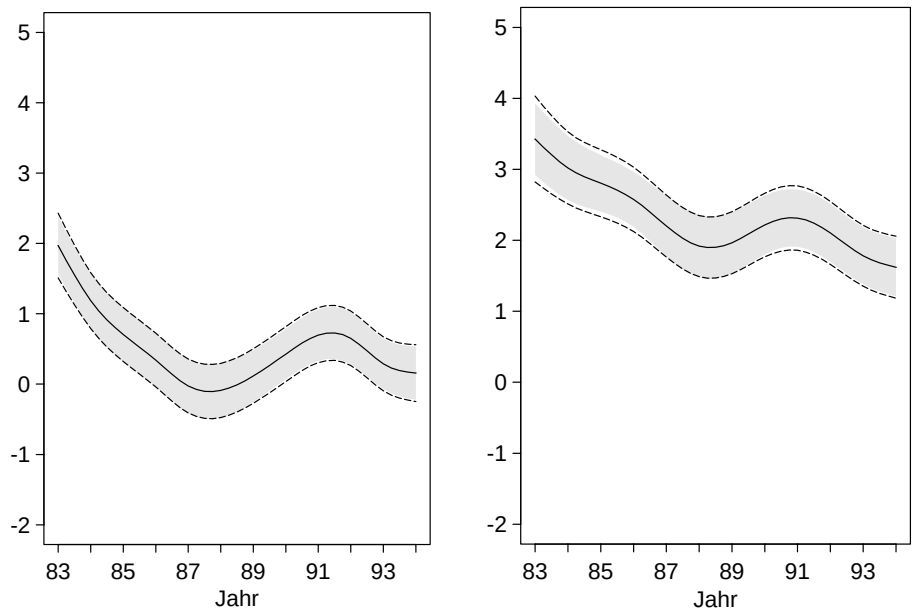


Figure 3: Estimated thresholds $\hat{\theta}_1(t)$ (left plot) and $\hat{\theta}_2(t)$ (right plot) with pointwise standard error bands (model based - dashed line, robust - boundary of shaded region).

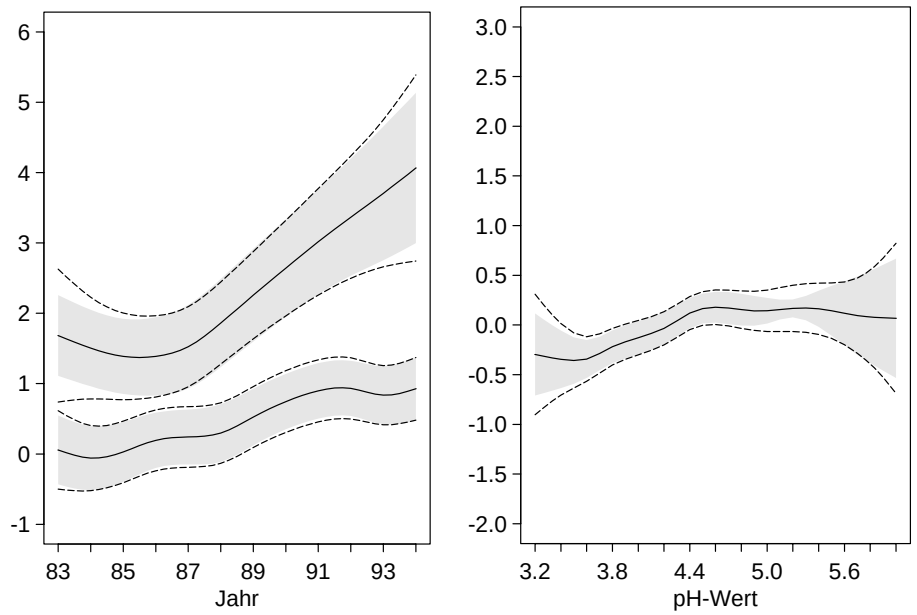


Figure 4: Estimated effects of age (left) with $A^{(1)}$ (upper curve), $A^{(2)}$ (lower curve), and pH value (right) with pointwise standard error bands (model based - dashed line, robust - boundary of shaded region).

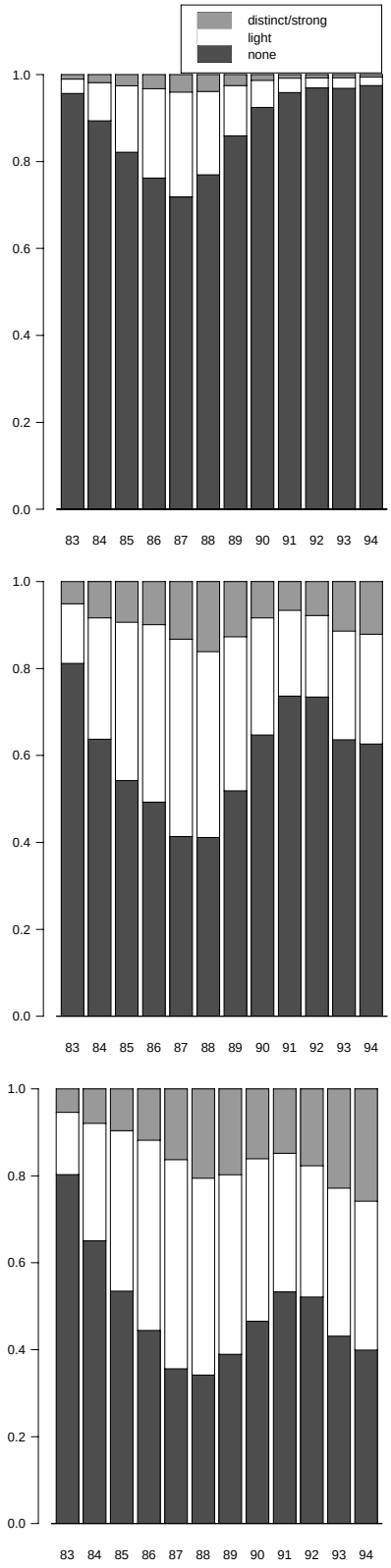


Figure 5: Estimated probabilities $\text{pr}(Y_t = 1), \text{pr}(Y_t = 2)$, and $\text{pr}(Y_t = 3)$ for age. From top to bottom: up to 50 years, between 50 and 120 years, above 120 years.

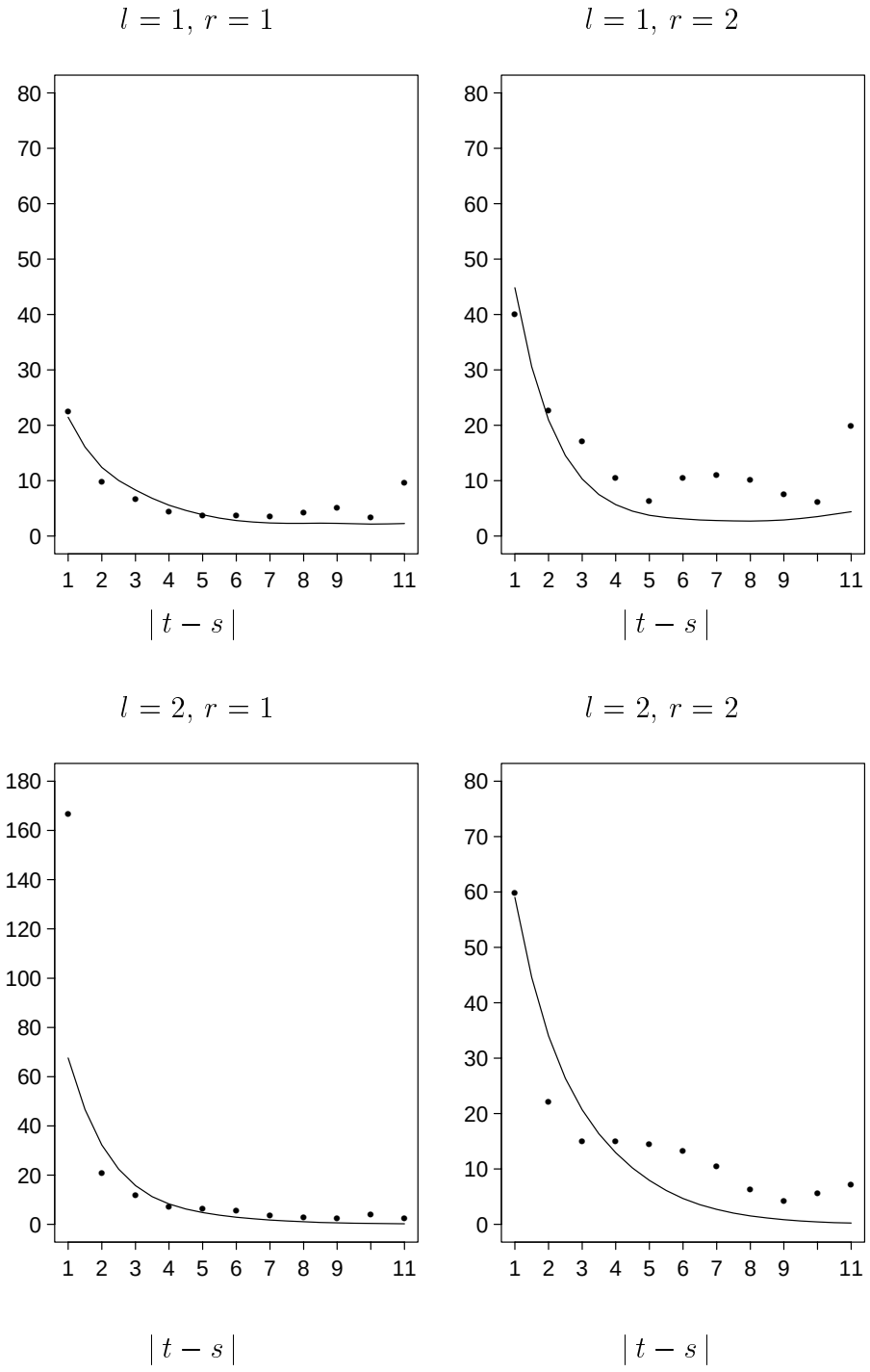


Figure 6: Estimated global odds ratios (lines) and empirically observed global odds ratios (points). Note that there is a different scale for combination $l = 2, r = 1$.