Tutz, Scholz:

Ordinal regression modelling between proportional odds and non-proportional odds

Sonderforschungsbereich 386, Paper 304 (2003)

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

Ordinal regression modelling between proportional odds and non-proportional odds

Gerhard Tutz, Torsten Scholz
University of Munich, Institute of Statistics
Akademiestrasse 1, 80799 München
tutz@stat.uni-muenchen.de scholz@stat.uni-muenchen.de

Abstract

The proportional odds model has become the most widely used model in ordinal regression. Despite favourable properties in applications it is often an inappropriate simplification yielding bad data fit. The more flexible non-proportional odds model or partial proportional odds model have the disadvantage that common estimation procedures as Fisher scoring often fail to converge. Then neither estimates nor test statistics for the validity of partial proportional odds models are available. In the present paper estimates are proposed which are based on penalization of parameters across response categories. For appropriate smoothing penalized estimates exist almost always and are used to derive test statistics for the assumption of partial proportional odds. In addition, models are considered where the variation of parameters across response categories is constrained. Instead of using prespecified scalars (Peterson & Harrell, 1990) penalized estimates are used in the identification of these constrained models. The methods are illustrated by various applications. The application to the retinopathy status in chronic diabetes shows how the proposed test statistics may be used in the diagnosis of partial proportional odds models in order to prevent artefacts.

KEYWORDS: Partial proportional odds model, ordinal regression, penalized estimation, penalized test statistics, constrained models.

1 Introduction

Since McCullagh's (1980) paper on regression models for ordinal data cumulative type models have become a standard tool in ordinal regression. When the outcomes of Y are ordinal with assigned values $1, \ldots, k$ the general cumulative logit model has the form

$$\log\big(P(Y \leq r \,|\, \boldsymbol{x})/P(Y > r \,|\, \boldsymbol{x})\big) = \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}_r, \tag{1}$$

$r = 1, \ldots, q = k-1$, where $\boldsymbol{\gamma}_r' = (\gamma_{1r}, \ldots, \gamma_{pr})$ is a category-specific parameter which depends on the response category. The simpler version is the proportional odds model

$$\log\big(P(Y \leq r \,|\, \boldsymbol{x})/P(Y > r \,|\, \boldsymbol{x})\big) = \gamma_{0r} + \boldsymbol{x}'\boldsymbol{\gamma}, \tag{2}$$

$r = 1, \ldots, q$, which assumes that the parameter $\boldsymbol{\gamma}' = (\gamma_1, \ldots, \gamma_p)$ does not depend on the category; $\boldsymbol{\gamma}$ is a so-called global parameter vector. The simpler version has several advantages. The observable response Y may be seen as a coarser version of an underlying unobservable continuous variable $\tilde{Y} = -\boldsymbol{x}'\boldsymbol{\gamma} + \varepsilon$ with $\gamma_{01} < \ldots < \gamma_{0q}$ denoting the category boundaries on the latent continuum that define the levels of Y by $Y = r$ if $\gamma_{0,r-1} < \tilde{Y} \leq \gamma_{0r}$. Basically one has an unidimensional regression model (for the latent variable) with straightforward interpretation of parameters. In addition, model (2) has the property of stochastic ordering (McCullagh, 1980), here in the form of proportional odds, meaning that for two populations characterized by covariates $\boldsymbol{x}_1, \boldsymbol{x}_2$ one obtains that the proportion of cumulative odds does not depend on the response category, i.e.

$$\log\left(\frac{P(Y \leq r \,|\, \boldsymbol{x}_1)/P(Y > r \,|\, \boldsymbol{x}_1)}{P(Y \leq r \,|\, \boldsymbol{x}_2)/P(Y > r \,|\, \boldsymbol{x}_2)}\right) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)'\boldsymbol{\gamma}.$$

As compared to the general model the proportional odds model makes economic use of parameters which is always a recommendable modelling strategy. The drawback is that often the simple model does not fit the data well. Moreover, assuming falsely that the odds are proportional can lead to invalid results. Therefore one has to investigate if the assumption of proportional odds holds for all or part of the covariates and to provide estimates for the general model. Candidates for tests are likelihood ratio tests, score tests and the Wald test (see Peterson & Harrell, 1990). Although the likelihood ratio test has desireable statistical properties, maximization of likelihood functions has to be performed twice. More severe is the problem of numerical instability. While the proportional odds model (2) only postulates $\gamma_{01} < \ldots < \gamma_{0q}$ the restriction that probabilities are from the interval $[0, 1]$ yields for the general model the restriction $\gamma_{01} + \boldsymbol{x}'\boldsymbol{\gamma}_1 < \ldots < \gamma_{0q} + \boldsymbol{x}'\boldsymbol{\gamma}_q$ which has to hold for all possible $\boldsymbol{x}$-values. That means that in applications Fisher scoring often fails to converge. The alternative is to fit $k-1$ separate binary models with $y = 1$ if $Y \leq r$ and $y = 0$ if $Y > r$. Although asymptotic covariances may be derived (Brant, 1990) the separate maximum likelihood estimates cannot be used to find estimates for the closed general model (1). Moreover, they do not correspond to the overall maximum likelihood estimates for the general model. Because of these problems with the likelihood ratio test, most often the score test has been used (e.g. in the SAS implementation). But what if the score test rejects the proportional odds model? With the existence of maximum likelihood estimates for the non-proportional model being questionable the further analysis of the data is questionable.

An example where several of these problems occur concerns the effect of diabetes and smoking on the retinopathy status (Bender & Grouven, 1998). With the retinopathy status given in three categories and three further covariates, one of them with a quadratic effect (for details see Section 3), the score test for the assumption of proportional odds is 16.693 which, when compared to a χ^2–distribution with 5 degrees of freedom (*df*), shows that the proportional odds model is inadequate. Likelihood ratio and Wald test cannot be applied since estimates for the non-proportional odds model do not exist. In cases like that, if category-specific parameters are not available, practitioners often interpret the global effects, even if the proportional odds model does not hold. In the case of the retinopathy data the investigation of the smoking effect yields the values 1.685 (score test), 1.715 (Wald test) and 1.706 (likelihood ratio test). Comparison to the χ^2–distribution with 1 *df* shows that smoking is not significant, but based on a model which is not appropriate. As will be shown in Section 3 the non-significance of smoking is an artefact which is due to the fitting of the wrong model (see also Bender & Grouven, 1998, who investigate dichotomized responses).

Even with the score test there are difficulties when trying to find a simpler model. The score test looks like a good device for testing the adequacy of the simpler proportional odds model since it is only necessary to fit the simpler model for which estimates for most data sets exist. But this holds only for the fit of the proportional odds model itself, not for the fit of the partial proportional odds model where only part of the variables have global weights.

To be more specific let $P = \{1, \ldots, p\}$ denote the index set for all of the p variables and $S \subset P$ denote a subset. The most general model considered is the non-proportional odds model (NPOM). Let the partial proportional odds model PPOM(S) be defined as the model where the variables x_j, $j \in S$, are global variables, i.e. the hypothesis

$$H_S: \ \gamma_{j1} = \ldots = \gamma_{jq}, \ j \in S,$$

is assumed to hold for all $j \in S$. Obviously the proportional odds model is equivalent to PPOM(P) and the non-proportional odds model is equivalent to PPOM($\emptyset$). The models PPOM(S) are only partially ordered with NPOM being the most general and PPOM(P) being the most restrictive model. In order to find a simple structure one has to test not only H_P but the hypotheses H_S, $S \subset P$. But when fitting model PPOM(S) one has to fit category-specific parameters γ_{jr}, $j \notin S$, which means that only in the trivial case $S = P$, one can avoid the fitting of category-specific parameters.

In the present paper the numerical difficulties are circumvented by using penalized likelihood methods. However, in contrast to common penalization approaches where penalization is used to obtain smooth effects of covariates (Eilers & Marx, 1996, Ruppert & Carroll, 1999, Ruppert, 2000) penalization in the following refers to the variation of effects across response categories. It may be considered as a form of 'vertical' smoothing (across the response) whereas smoothing across the values of the predictor may be seen as a form of 'horizontal' smoothing (across the values of explanatory variables). As a by-product alternative models with a simple structure between the proportional and the non-proportional odds model may be obtained.

The focus here is on cumulative type models. Alternative models for ordinal regression and comparisons between different types of models have been considered by Cox (1988), Brant (1990), Tutz (1991), Armstrong & Sloan (1989), Greenland (1994), Albert & Chib (2001). Overviews are found in Barnhart & Sampson (1994) and from a more general view in Agresti (1984, 1999).

2 Penalized estimates for non-proportional models

Let more generally the cumulative type model for observations $(Y_i, \boldsymbol{x}_i)$ be given by

$$P(Y_i \leq r \mid \boldsymbol{x}_i) = F(\eta_{ir}),$$

where F is a strictly monotone distribution function. The general model has predictor

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i' \boldsymbol{\gamma}_r, \tag{3}$$

$r = 1, \ldots, q = k - 1$, whereas the model with homogeneous effects (the proportional odds model if F is the logistic distribution) has predictor

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i'\boldsymbol{\gamma}.$$

A common framework for these models is the multivariate generalized model which has the form

$$\boldsymbol{\pi}_i = h(Z_i\boldsymbol{\beta}) \quad \text{or} \quad g(\boldsymbol{\pi}_i) = Z_i\boldsymbol{\beta},$$

where $\boldsymbol{\pi}_i' = (\pi_{i1}, \ldots, \pi_{iq})$ with components $\pi_{ir} = P(Y_i = r \,|\, \boldsymbol{x}_i)$ and g is the link function, whereas $h = g^{-1}$ is the inverse link or response function. These functions are determined by F. For the model with homogenous effects the design matrix Z_i and the parameter vector are given by

$$Z_i = [I_{q\times q}, 1_{q\times 1} \otimes \boldsymbol{x}_i'], \quad \boldsymbol{\beta}' = (\boldsymbol{\gamma}_0', \boldsymbol{\gamma}'),$$

with $I_{q\times q}$ denoting the unit matrix, $1_{q\times 1}' = (1, \ldots, 1)$ and $\boldsymbol{\gamma}_0' = (\gamma_{01}, \ldots, \gamma_{0q})$. For the general model (3) one uses instead

$$Z_i = [I_{q\times q}, I_{q\times q} \otimes \boldsymbol{x}_i'], \quad \boldsymbol{\beta}' = (\boldsymbol{\gamma}_0', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_q'),$$

with $\boldsymbol{\gamma}_r' = (\gamma_{1r}, \ldots, \gamma_{pr})$, $r = 1, \ldots, q$. For further details see Fahrmeir & Tutz (2001).

2.1 Penalized likelihood

In the following fitting procedures are considered for the general partial proportional-odds model PPOM(S) which has global effects for x_j, $j \in S$, and category-specific effects for x_j, $j \in \bar{S} = P \backslash S$. In obvious notation the predictor is given by

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_{i,S}'\boldsymbol{\gamma}_S + \boldsymbol{x}_{i,\bar{S}}'\boldsymbol{\gamma}_{\bar{S},r},$$

where in $\boldsymbol{x}_{i,S}$ the variables x_{ij}, $j \in S$, and in $\boldsymbol{x}_{i,\bar{S}}$ the variables x_{ij}, $j \notin S$, are collected. With $\boldsymbol{\beta}$ corresponding to the parameters $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_S, \{\boldsymbol{\gamma}_{\bar{S},r}\}$, $r = 1, \ldots, q$, the proposed penalized log-likelihood has the form

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \cdot \tau(\boldsymbol{\beta}), \tag{4}$$

where

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{r=1}^{k} y_{ir} \log(\pi_{ir}),$$

is the usual log-likelihood. The penalty term is given by

$$\tau(\boldsymbol{\beta}) = \sum_{j \in \{0\} \cup \bar{S}} \lambda_j \sum_{r=2}^{q} (\Delta \gamma_{jr})^2,$$

where Δ is the difference operator operating on adjacent response categories, i.e. $\Delta\gamma_{jr} = \gamma_{jr} - \gamma_{j,r-1}$. Thus τ penalizes the differences of the parameters across response categories. The amount of smoothing is determined by the smoothing parameters λ_j, $j \in \{0\} \cup \bar{S}$. If $\lambda_j = 0$, $j \in \{0\} \cup \bar{S}$, maximization of the penalized likelihood yields the unconstrained maximum likelihood by use of the iterative Newton procedure with Fisher scoring. For small values of the smoothing parameters and many category-specific variables the iterative procedure often fails to converge since the restrictions imposed by ordering are not fullfilled. In the following the non-existence of estimates means that the Newton procedure with Fisher scoring fails. For the special case $\lambda_j \to \infty$, $j \in \bar{S}$, the model with homogeneous effects is fitted. Maximization of (4) yields the estimation equation $s_p(\boldsymbol{\beta}) = 0$ where $s_p(\boldsymbol{\beta}) = \partial l_p(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the penalized score function which has the form

$$s_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\pi}_i) - P\boldsymbol{\beta},$$

with $D_i = \partial h(\boldsymbol{\eta}_i)/\partial\boldsymbol{\eta}$, $\boldsymbol{\eta}_i' = (\eta_{i1}, \ldots, \eta_{iq})$ and $\boldsymbol{y}_i' = (y_{i1}, \ldots, y_{iq})$ where $y_{ir} = 1$ if $Y_i = r$, $y_{ir} = 0$ otherwise. Z_i is the design matrix corresponding to model PPOM(S) and the matrix P represents the penalization and the smoothing parameters (see appendix). The corresponding fitting procedure is iterative Newton with modified Fisher scoring, i.e. the penalized Fisher matrix is used as weight. Starting with $\hat{\boldsymbol{\beta}}^{(k)}$ the procedure obtains an update $\hat{\boldsymbol{\beta}}^{(k+1)}$ by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + F_p\left(\hat{\boldsymbol{\beta}}^{(k)}\right)^{-1} s_p\left(\hat{\boldsymbol{\beta}}^{(k)}\right), \tag{5}$$

where $F_p(\boldsymbol{\beta}) = F(\boldsymbol{\beta}) + P$, $F(\boldsymbol{\beta}) = \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1} D_i' Z_i$. If the amount of smoothing is increased the estimates are closer to the usual maximum likelihood estimates of the proportional odds model. Since these exist under much weaker conditions, in most cases limiting smoothing parameters λ_{l0} exist such that (5) converges for $\lambda_j > \lambda_{j0}$.

2.2 Existence of estimates: a small simulation study

In a small simulation study the potential of smoothed estimates is evaluated. It is assumed that a non-proportional odds model holds with predictors

$$\begin{aligned}
\eta_{i1} &= -0.8 + 0.7 \cdot x_i \\
\eta_{i2} &= -0.4 + 0.4 \cdot x_i \\
\eta_{i3} &= 0.2 + 0.1 \cdot x_i
\end{aligned}$$

where x_i, $i = 1, \dots, n$, are drawn from an uniform distribution on $[-1, 1]$.

For sample size $n = 300$ Fisher scoring without penalization failed in 17 from 100 simulations when the non-proportional odds model (NPOM) was assumed. For these 17 simulations the improvement is investigated which results from fitting the model with penalization as compared to the fit of the proportional odds model (POM) for which estimates exist.

The loss functions which are considered are mean squared error loss

$$\text{MSEL} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} (\pi_{ir} - \hat{\pi}_{ir})^2,$$

mean relative squared error loss

$$\text{MRSEL} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} \frac{(\pi_{ir} - \hat{\pi}_{ir})^2}{\pi_{ir}},$$

and mean entropy or Kullback-Leibler loss

$$\mathrm{MEL} = \frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{k}\pi_{ir}\log\left(\frac{\pi_{ir}}{\hat{\pi}_{ir}}\right).$$

Figure 1 shows the resulting losses for the 17 simulations when both models, POM and NPOM, are fitted. For the fitting of the NPOM the smoothing parameter is chosen to be the minimal value where the model may be estimated by penalized Fisher scoring. Figure 1 shows that the fitting of the NPOM based on penalization yields distinctly smaller losses in all cases with the exception of two simulations where Kullback-Leibler loss is not improved.

Table 1 summarizes these results by showing the mean values of the loss functions. The mean losses for the POM are about twice as much as for the penalized NPOM. Averaging the ratios of the losses yields a factor of at least 0.53 by which the loss can be reduced when fitting the penalized NPOM instead of the POM.

It should be noted that improvement in comparison to the proportional odds model is of interest only if the non-proportional odds model cannot be fitted. Therefore the consideration is restricted to these cases.

2.3 Test Statistics

The usual likelihood ratio test is restricted to cases where estimates exist. Thus the test for the hypothesis $H_S:\ \gamma_{j1} = \ldots = \gamma_{jq},\ j \in S$, which investigates if the variables $x_j,\ j \in S$, may be considered as global variables, often fails.

An alternative which is based on penalized estimates uses the maximal penalized log-likelihood for the NPOM $l_p\big(\hat{\boldsymbol{\gamma}}_0, \{\hat{\boldsymbol{\gamma}}_{P,r}\}\big)$, $r = 1, \ldots, q$ and the maximal penalized log-likelihood $l_p\big(\tilde{\boldsymbol{\gamma}}_0, \tilde{\boldsymbol{\gamma}}_S, \{\tilde{\boldsymbol{\gamma}}_{\bar{S},r}\}\big)$, $r = 1, \ldots, q$, for the PPOM(S). One considers

$$LR_p = -2\big\{l_p\big(\tilde{\boldsymbol{\gamma}}_0, \tilde{\boldsymbol{\gamma}}_S, \{\tilde{\boldsymbol{\gamma}}_{\bar{S},r}\}\big) - l_p\big(\hat{\boldsymbol{\gamma}}_0, \{\hat{\boldsymbol{\gamma}}_{P,r}\}\big)\big\}.$$

Let $\hat{\boldsymbol{\beta}}$ denote the estimates of the more general model, which in this case is the NPOM, and $\tilde{\boldsymbol{\beta}}$ denote the estimates of the reduced model. With the corresponding fits $\widehat{\boldsymbol{\pi}}_i' = (\hat{\pi}_{i1}, \dots, \hat{\pi}_{ik})'$ and $\widetilde{\boldsymbol{\pi}}_i' = (\tilde{\pi}_{i1}, \dots, \tilde{\pi}_{ik})'$ one obtains

$$LR_p = 2 \sum_{i=1}^{n} \sum_{r=1}^{k} y_{ir} \log \left(\frac{\hat{\pi}_{ir}}{\tilde{\pi}_{ir}} \right) + \tau(\tilde{\boldsymbol{\beta}}) - \tau(\hat{\boldsymbol{\beta}}).$$

When using LR_p different smoothing may be used for the models which are fitted. A closer look at the penalty suggests that the same smoothing should be used for both models. With $\lambda_{j,S}$ ($\lambda_{j,P}$) denoting the smoothing parameters for the sub model (general model) one has

$$\begin{aligned} \tau(\tilde{\boldsymbol{\beta}}) - \tau(\hat{\boldsymbol{\beta}}) &= \sum_{j \in \{0\} \cup \bar{S}} \lambda_{j,S} \sum_{r=2}^{q} (\Delta \tilde{\gamma}_{jr})^2 - \sum_{j=0}^{p} \lambda_{j,P} \sum_{r=2}^{q} (\Delta \hat{\gamma}_{jr})^2 \\ &= \sum_{j \in \{0\} \cup \bar{S}} \sum_{r=2}^{q} \left(\lambda_{j,S} (\Delta \tilde{\gamma}_{jr})^2 - \lambda_{j,P} (\Delta \hat{\gamma}_{jr})^2 \right) - \sum_{j \in S} \lambda_{j,P} \sum_{r=2}^{q} (\Delta \hat{\gamma}_{jr})^2. \end{aligned}$$

If $\lambda_{j,S} = \lambda_{j,P}$ is chosen for $j \in \{0\} \cup \bar{S}$, then the first term is very small since $\tilde{\gamma}_{jr} \approx \hat{\gamma}_{jr}$ for $r = 1, \dots, q$. Thus the essential term comes from the smoothing of parameters which are connected to the variables x_j, $j \in S$, for which it is investigated if the effects are category-specific. If estimates are not penalized, i.e. $\lambda_{0,S} = \lambda_{1,S} = \dots = \lambda_{|\bar{S}|,S} = \lambda_{0,P} = \lambda_{1,P} = \dots = \lambda_{p,P} = 0$, one obtains $\tau(\tilde{\boldsymbol{\beta}}) = \tau(\hat{\boldsymbol{\beta}}) = 0$ and LR_p has the usual asymptotic χ^2-distribution. The same asymptotic behaviour is obtained if the smoothing parameters converge to zero at an appropriate rate. However, since in the finite sample case asymptotic quantiles may not be trustworthy, Monte Carlo simulations are used for the penalized likelihood ratio statistic LR_p. The significance of the observed value l of LR_p is thereby obtained by simulating N samples $(\widetilde{\boldsymbol{Y}}_1, \dots, \widetilde{\boldsymbol{Y}}_n)$ of independent multinomially distributed random variables from a model with response probability matrix $[\widetilde{\boldsymbol{\pi}}_1 \cdots \widetilde{\boldsymbol{\pi}}_n]$. That proportion of the N corresponding values of LR_p which exceed l will be taken as the p-value (Firth, Glosup & Hinkley, 1991).

An alternative is the Wald test. Let C denote the matrix $C\boldsymbol{\beta} = \mathbf{0}$ which corresponds to the linear hypothesis $\gamma_{j1} = \ldots = \gamma_{jq}$, $j \in S$. With $\hat{\boldsymbol{\beta}}$ denoting the penalized estimate for the NPOM one considers the Wald type statistic

$$W_p = \hat{\boldsymbol{\beta}}' C' \left(C \,\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\, C' \right)^{-1} C \hat{\boldsymbol{\beta}},$$

where $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = F_p(\hat{\boldsymbol{\beta}})^{-1} F(\hat{\boldsymbol{\beta}}) F_p(\hat{\boldsymbol{\beta}})^{-1}$ is the sandwich matrix given in the appendix.

In contrast to the Wald test the score type test uses the estimate $\tilde{\boldsymbol{\beta}}$ for the reduced model by computing

$$s_p = s(\tilde{\boldsymbol{\beta}})' F(\tilde{\boldsymbol{\beta}})^{-1} s(\tilde{\boldsymbol{\beta}}).$$

For penalized estimates s_p may be motivated from

$$s_p = \left(s_p(\tilde{\boldsymbol{\beta}}) - E\big(s_p(\tilde{\boldsymbol{\beta}})\big) \right)' F(\tilde{\boldsymbol{\beta}})^{-1} \left(s_p(\tilde{\boldsymbol{\beta}}) - E\big(s_p(\tilde{\boldsymbol{\beta}})\big) \right),$$

where $s_p(\tilde{\boldsymbol{\beta}})$ denotes the penalized score function evaluated at $\tilde{\boldsymbol{\beta}}$ (see appendix). Simple derivation shows the equivalence of these two forms.

Two tests are of specific importance. The first is the *global test* where the NPOM is compared to the proportional odds model PPOM(P). For the global test and smoothing parameters set to zero one obtains the asymptotic χ^2–distribution with $p(q-1)$ df. With smoothing parameters unequal to zero only smoothing parameters for the general model NPOM have to be chosen. The second type of test considers the parameters for one fixed variable j by investigating the hypothesis $H_{\{j\}}: \ \gamma_{j1} = \ldots = \gamma_{jq}$.

The family of models PPOM(S), $S \subset P$, is a non-hierarchical family. Although PPOM(S_1) $\subset$ PPOM(S_2) holds if $S_2 \subset S_1$ the models are only partially ordered. Thus when looking for an adequate model one has to consider all the hypotheses H_S within a multiple test problem. Instead of a stepwise procedure we propose to use a simple strategy by testing all the hypotheses $H_{\{j\}}$, $j \in P$, with the full model as the background model.

Thus the fit of model PPOM($\{j\}$) is compared to NPOM = PPOM($\emptyset$) by penalized likelihood ratio statistics.

2.4 Comparison of power functions

Judging the quality of the introduced test statistics can effectively be done by means of the corresponding power functions. In a simulation study the following predictor specification was assumed to hold

$$\begin{aligned}
\eta_{i1} &= -0.8 + (0.4 + \Delta) \cdot x_i \\
\eta_{i2} &= -0.4 + \ 0.4 \cdot x_i \\
\eta_{i3} &= \ \ 0.2 + (0.4 - \Delta) \cdot x_i
\end{aligned}$$

where x_i, $i = 1, \ldots, n$, are drawn from an uniform distribution on $(-1, 1)$. Depending on the specified value of $\Delta \geq 0$ the true model is a POM ($\Delta = 0$) or a NPOM ($\Delta > 0$). The magnitude of Δ determines the “distance” between the POM and the NPOM in the way that an increasing Δ causes an increasing non-proportionality of the odds.

The power functions of the test statistics are obtained by estimating the probability to reject the assumption of proportional odds for increasing Δ. Starting with $\Delta = 0$ we drew 160 samples of $n = 250$ independent repetitions of a multinomially distributed random variable according to the given predictor specification. For each sample the test statistics LR_p, W_p and s_p have been calculated. Their significance was determined by Monte Carlo simulation drawing $N = 200$ data sets respectively. Based on the obtained 160 p-values the rejection probability was estimated by relative frequencies. The procedure was repeated for stepwise increased Δ with a maximum value of $\Delta = 0.4$ to guarantee that the ordering restriction $\eta_{i1} < \eta_{i2} < \eta_{i3}$ is fulfilled for all observations x_i.

It should be noted that s_p coincides with the usual score statistic s when testing the POM against the NPOM. The obtained results can therefore be used to compare the score statistic as the common test tool with our proposed alternatives. Figure 2 shows the estimated power functions of LR_p and s_p. The vertical lines indicate the proportion of failures when fitting the NPOM with and without vertical penalization. Obviously the chosen smoothing parameter $\lambda_{1,P} = \exp(4)$ significantly reduces the proportion of samples where Fisher scoring failed.

In Figure 3 the estimated power functions of LR_p and s_p are superimposed. The graphs of the power functions are almost identical. A significant difference is hardly to verify so that LR_p and s_p are of comparable quality. Further simulations showed that this statement can be generalized for any amount of vertical penalization.

3 Application to retinopathy

In a 6-year follow up study on diabetes and retinopathy status reported by Bender & Grouven (1998) the interesting question is how the retinopathy status is associated with risk factors. The considered risk factor is smoking ($SM = 1$: smoker, $SM = 0$: non-smoker) adjusted for the known risk factors diabetes duration (DD) measured in years, glycosylated hemoglobin (GH) which is measured in percent and diastolic blood pressure (BP) measured in mmHg. The response variable retinopathy status has three categories (1: no retinopathy, 2: nonproliferative retinopathy, 3: advanced retinopathy or blind). Bender & Grouven pointed out that an appropriate model should contain the linear (DD) and the quadratic effect (DDQ) of diabetes duration. Therefore first we consider the non-proportional odds model

$$\eta_{ir} = \gamma_{0r} + SM_i \cdot \gamma_{SM,r} + DD_i \cdot \gamma_{DD,r} + DDQ_i \cdot \gamma_{DDQ,r} + GH_i \cdot \gamma_{GH,r} + BP_i \cdot \gamma_{BP,r}\,,$$

for patients $i = 1, \ldots, n = 613$ and categories $r = 1, 2$. Without penalization Fisher scoring fails when trying to fit this model. Even after rescaling of the variables such that observations lie in the intervall $[0, 1]$ the algorithm does not converge. Fisher scoring always breaks off with an error after a few iterations, caused by an ill-conditioned Fisher matrix. However, convergence of the algorithm is obtained by imposing a small penalty $\lambda_{DD,P}^{\min} = \lambda_{DDQ,P}^{\min} = 1$ on the parameters $\gamma_{DD,r}$ and $\gamma_{DDQ,r}$, $r \in \{1, 2\}$, respectively. Then no further penalization is needed for the remaining parameters.

Submodels of the general model are investigated by testing PPOM($\{j\}$) against the NPOM for all variables $j \in P = \{SM, \{DD, DDQ\}, GH, BP\}$. Thereby the linear and the quadratic effect of diabetes duration are considered simultaneously. Again Fisher scoring fails for these models, except for the PPOM with DD and DDQ having global effects. Imposing the penalty $\lambda_{DD,S} = \lambda_{DDQ,S} = 1$ ensures the existence of estimates even for the critical settings. The test statistics LR_p, W_p and s_p are calculated and the significance of their observed values is determined by Monte Carlo simulation.

Figure 4 and Figure 5 give significance traces where the smoothing parameters are plotted against p-values. Significance traces as proposed by Azzalini & Bowman (1993) (see also Bowman & Azzalini (1997)) are a helpful tool to investigate effects across a wide range of smoothing parameters. We take $\lambda_{j,S} = \lambda_{j,P} = \lambda_{j,P}^{\min}$ for $j \in \bar{S}$, whereas the smoothing parameter $\lambda_{j,P}$ for variable j, $j \in S$, varies in the range $[\lambda_{j,P}^{\min}, \exp(14)]$, with $\lambda_{j,P}^{\min} = 0$ except for diabetes duration where $\lambda_{j,P}^{\min} = 1$.

From Figures 4 and 5 it is seen that the p-values remain quite stable across the range of the smoothing parameter. With p-values close to 0.5 the plots distinctly indicate proportional odds for glycosylated hemoglobin and diastolic blood pressure. With p-values close to 0.05 for smoking and below 0.02 for diabetes duration (linear and quadratic effect) there is some evidence that these effects are category-specific.

In addition the lower panels in Figures 4 and 5 show the corresponding deviations of the estimates $\hat{\gamma}_{j,1}$, $\hat{\gamma}_{j,2}$, $j \in S$, from the assumption of global effects. Projection of the vector $(\hat{\gamma}_{j,1}, \hat{\gamma}_{j,2})'$ to the hyperplane defined by $\gamma_{j,1} = \gamma_{j,2}$ yields

$$||\Delta||_2 \;=\; \sqrt{0.5} \cdot |\hat{\gamma}_{j,1} - \hat{\gamma}_{j,2}|$$

as a measure of discrepancy.

These additional plots are helpful to investigate if the considered range of smoothing parameters covers the range of interest. It is seen that for smoothing parameters $\log(\lambda) > 6$ global effects are fitted. Thus the interesting area where the models with category-specific and global effects are distinguished is the range where $\log(\lambda) \leq 6$. But although there is some variation in this range the questions if effects are substantial or not are consistent across varying smoothing parameters.

Based on the significance traces the model with predictor

$$\eta_{ir} = \gamma_{0r} + GH_i \cdot \gamma_{GH} + BP_i \cdot \gamma_{BP} + SM_i \cdot \gamma_{SM,r} + DD_i \cdot \gamma_{DD,r} + DDQ_i \cdot \gamma_{DDQ,r}\,,$$

with global effects for glycosylated hemoglobin and diastolic blood pressure is considered as sufficiently complex.

Table 2 gives the estimates based on the smoothing parameters $\lambda_{j,P}^{\min}$ for smoking and diabetes duration. Tests for the significance of effects of variables may be performed by using appropriate versions of the penalized test statistics in Section 2.3. The only modification is that the submodel is now defined by the omission of effects of variables. Whether these effects are global or category-specific depends on the fitted model. If $\lambda_{j,P}^{\min} = 0$ or is close to it, the simple χ^2-approximation may be used when performing the tests. Table 2 shows the results based on Monte Carlo simulation. $N = 1000$ data sets were thereby simulated to obtain the significance probabilities. For all tests smoking status shows a significant negative effect for the first category, indicating that smokers bear a higher risk to develop at least nonproliferative retinopathy. A significant effect of smoking on the development of advanced retinopathy can not be derived. For all other variables the penalized test statistics indicate high significance of their (global or category-specific) effects.

It should be remarked that a simple analysis with the proportional odds model shows no significant effect of smoking. Thus the straightforward analysis based on an ill-fitting model yields very misleading results. This effect has already been pointed out by Bender & Grouven (1998) who investigated dichotomized response variables and noted that the proportional odds model is not appropriate.

4 Constrained models

The identification of variables with global effects is gainful since these effects are easier to interpret. Thus one wants to find the most simple model, with as many as possible global effects. In the literature alternative ways to find a simply structured model have been proposed. Peterson & Harrell (1990) consider constrained partial proportional odds models which have predictor

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}_i'\boldsymbol{\gamma} + \sum_{j \in \bar{S}} x_{ij}\tilde{\gamma}_j\delta_{jr},$$

where δ_{jr}, $j \in \bar{S}$, $r = 1, \ldots, q$, are prespecified scalars. That means the variables from $\bar{S}$ have category-specific parameters $\gamma_j + \tilde{\gamma}_j\delta_{jr}$ with unknown parameters γ_j, $\tilde{\gamma}_j$. For example the choice $\delta_{j1} = 1$, $\delta_{j2} = 2, \ldots$ means that the effect increases in a known form across response categories. Thus simply structured models are obtained, however at the cost of assuming rather arbitrary constants.

Within the framework of penalized estimates one may consider simple models with the "scores" δ_{jr} chosen by the data. Let us for simplicity consider the NPOM with

$$\eta_{ir} = \gamma_{0r} + \sum_{j=1}^{p} x_{ij}\gamma_{jr}.$$

Instead of assuming $\gamma_{jr} = \gamma_j + \tilde{\gamma}_j\delta_{jr}$ with known δ_{jr} the reduction of parameters is obtained by fitting models with different degrees of differences. The underlying concept is that strong penalization yields polynomial fits with the fit depending on the order of the penalty. If one uses the penalty of order d

$$\tau(\boldsymbol{\beta}) = \sum_{j=0}^{p} \lambda_j \sum_{r=d+1}^{q} (\Delta^d \gamma_{jr})^2, \quad 1 \leq d \leq q-1\,,$$

where $\Delta^d\gamma_{jr} = \Delta(\Delta^{d-1}\gamma_{jr})$ and $\Delta^1 = \Delta$, for $\lambda_j \to \infty$ the fitted parameters $\hat{\gamma}_{jr}$ follow a polynomial of degree $d-1$. For example if $d = 2$ the fitted parameters are on a straight line, i.e. $\hat{\gamma}_{jr} = \alpha_{j0} + \alpha_j r$. Thus the effective number of parameters is reduced to two: intercept α_{j0} and slope α_j. The predictor for variable x_j becomes

$$x_j\hat{\gamma}_{jr} = x_j\alpha_{j0} + x_j(\alpha_j r),$$

with the "scores" $\delta_{jr} = r$. For differences of higher order the scores become more complex but still involve only few parameters. If the order of difference d yields a large amount of smoothing for variable x_j then γ_{jr} may be described by a polynomial of degree $d-1$.

4.1 Application to Severity of Nausea Data

A simple example which has also been considered by Peterson & Harrell (1990) is based on a data set given by Farewell (1982). Farewell (1982) investigates the severity of nausea for patients receiving chemotherapy with and without cisplatin on an ordinal scale with six categories (see Table 3).

Farewell (1982) rejects the assumption of proportional odds for the data set. Peterson & Harrell (1990) fit a non-proportional odds model and conclude from the fitting that the predictor

$$\eta_{ir} = \gamma_{0r} + x_i\gamma + x_i\tilde{\gamma}\delta_r$$

with the scores $\delta_1 = \ldots = \delta_4 = 0$ and $\delta_5 = 1$ is appropriate. Then the hypothesis $H_0 : \delta_1 = \ldots = \delta_4 = 0,\ \delta_5 = 1$ is tested and the restricted model is fitted. Although the testing of a hypothesis that has been generated from the data seems not too informative, the method yields an appropriate simplification.

The alternative approach which is proposed here is to let the parameters δ_r unspecified after the assumption of proportional odds is rejected. We fit the model with penalties of varying degree. Figure 6 shows the AIC criterion for degrees $d = 1, 2, 3$. It is seen that for increasing amount of smoothing penalties of degree 1 and 2 show distinct minima whereas for $d = 3$ the AIC criterion decreases with increasing amount of smoothing. The extreme amount of smoothing $\lambda = \exp(8)$ is equivalent to restricting the parameters to follow a polynomial of degree $d-1$. Thus for $\lambda = \exp(8)$ and $d = 1$ one fits the proportional odds model, for $d = 2$ the fit is as measured by the AIC criterion even worse, but for $d = 3$ the AIC criterion has strongly improved. Therefore the model with a polynomial of second degree for the parameters yields a satisfying fit. The corresponding log-likelihood is -371.54 which is slightly larger than the log-likelihood for the model with $\delta_1 = \ldots = \delta_4 = 0$ and $\delta_5 = 1$ which as reported by Peterson & Harrell (1990) is -372.19.

4.2 Application to injuries of the knee

In a clinical study focussing on the healing of sports related injuries of the knee $n = 123$ patients have been treated. By random design one of two therapies were chosen. In the treatment group an anti-inflammatory spray was used while in the placebo group a spray without active ingredients was used. After ten days of treatment with the spray, the mobility of the knee was investigated in a standardized experiment during which the knee was actively moved by the patient. The pain Y occuring during the movement was assessed on a five point scale ranging from 1 for no pain to 5 giving severe pain. In addition to treatment (*Treat*) the covariate age (*Age*) is given. From previous analysis it is known that *Age* should be included as a quadratic function. Thus linear and quadratic effects of age are considered.

Investigation of the data shows that *Treat* has category-specific effects whereas *Age* and Age^2 can be considered as global variables. Figure 7 shows the AIC criterion for the model where the effects of treatment are penalized with a penalty of degree $d \in \{1, 2, 3\}$. The value of AIC for all of the models is distinctly smaller than AIC for the proportional odds model. But with increasing amount of smoothing for $d = 1$ and $d = 2$ AIC increases whereas it obtains its smallest value for $d = 3$. Thus the model with linear predictor

$$\eta_{ir} = \gamma_{0r} + Age_i \cdot \gamma_A + Age_i^2 \cdot \gamma_{A^2} + Treat_i \cdot \gamma_{T,r}$$

with $\gamma_{T,r}$ being determined by a polynomial of degree 2 yields the best fit. The estimated parameters are given in Table 4.

5 Concluding remarks

This paper presents tools for the modelling of partial and non-proportional odds models. Test statistics are used to evaluate if all or part of the covariates have global effects. If some of the variables are category-specific still estimates are available based on the concept of penalization across response categories. This may be seen as a way of finding estimates but also as a form of nonparametric modelling by assuming smoothness of parameters across response categories.

Simple parametric models are obtained by fitting models with penalties of differing degrees together with strong smoothing. Considering the case of maximal smoothing is equivalent to the fitting of parametric models where the effects are determined by a polynomial degree. The approach avoids problems which arise when one looks for simple models by using assigned scores. Usually the assigned scores are rather arbitrary. If simple scores are found from the data, testing faces the problem of using the same data twice, once for finding the hypotheses, then for investigating them. Alternatively, we propose first to test which effects are category-specific, but then to find a simple parametric model by use of the AIC criterion.

Appendix

We consider the PPOM(S), $S \subset P$ with predictor

$$\eta_{ir} = \gamma_{0r} + \boldsymbol{x}'_{i,S}\boldsymbol{\gamma}_S + \boldsymbol{x}'_{i,\bar{S}}\boldsymbol{\gamma}_{\bar{S},r}.$$

It is helpful to restructure the design matrix such that the global variables come first. One uses

$$Z_i = \begin{pmatrix} 1 & & 0 & \boldsymbol{x}'_{i,S} & \boldsymbol{x}'_{i,\bar{S}} & & \\ & \ddots & & \vdots & & \ddots & \\ 0 & & 1 & \boldsymbol{x}'_{i,S} & & & \boldsymbol{x}'_{i,\bar{S}} \end{pmatrix}$$

and $\boldsymbol{\beta}' = (\boldsymbol{\gamma}'_0, \boldsymbol{\gamma}'_S, \boldsymbol{\gamma}'_{\bar{S},1}, \ldots, \boldsymbol{\gamma}'_{\bar{S},q})$ to obtain $\eta_{ir} = Z_i\boldsymbol{\beta}$.

The penalty of first order has the form

$$\tau(\boldsymbol{\beta}) = \sum_{j \in \{0\} \cup \bar{S}} \lambda_j \sum_{r=2}^{q} (\Delta\gamma_{jr})^2.$$

This may be simplified by using the difference matrix

$$\tilde{D}_q = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & & \ddots & & \\ & & & & -1 & 1 \end{pmatrix},$$

which is a $(q-1) \times q$ matrix. With $D = \tilde{D}_q \otimes \mathrm{Diag}\big(\lambda_{1,S}^{1/2}, \ldots, \lambda_{|\bar{S}|,S}^{1/2}\big)$ one obtains for the penalty of order $d = 1$ in closed matrix form

$$\tau(\boldsymbol{\beta}) = \boldsymbol{\beta}' P \boldsymbol{\beta},$$

with $P = \mathrm{Diag}\big(\lambda_{0,S}\tilde{D}'_q\tilde{D}_q, 0_{|S|\times|S|}, D'D\big)$, where $0_{|S|\times|S|}$ is a matrix with zeros having dimension $|S| \times |S|$. For penalties of general order d, $1 \le d \le q-1$, the matrix $\tilde{D}_q$ has to be replaced by the product $\tilde{D}_{q-d+1} \cdot \ldots \cdot \tilde{D}_{q-1} \cdot \tilde{D}_q$.

The corresponding penalized score function has the form

$$s_p(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) - P\boldsymbol{\beta} \;=\; s(\boldsymbol{\beta}) - P\boldsymbol{\beta},$$

the penalized Fisher matrix is given by

$$F_p(\boldsymbol{\beta}) \;=\; E\left(-\frac{\partial l_p}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) \;=\; \sum_{i=1}^{n} Z_i' D_i \Sigma_i^{-1} D_i' Z_i + P \;=\; F(\boldsymbol{\beta}) + P.$$

The covariance matrix may be approximated by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) \;=\; F_p(\hat{\boldsymbol{\beta}})^{-1} F(\hat{\boldsymbol{\beta}}) F_p(\hat{\boldsymbol{\beta}})^{-1}.$$

The latter expression may be derived by simple Taylor expansion. Corresponding approximations have been used by Eilers & Marx (1996) for univariate generalized linear models and by Tutz (2003).

Fisher scoring has the form

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \Big(Z'W\big(\hat{\boldsymbol{\beta}}^{(k)}\big)Z + P\Big)^{-1} Z'W\big(\hat{\boldsymbol{\beta}}^{(k)}\big)\tilde{\boldsymbol{\eta}}\big(\hat{\boldsymbol{\beta}}^{(k)}\big)$$

where $Z' = (Z_1', \ldots, Z_n')$, $W(\hat{\boldsymbol{\beta}}) = \text{Diag}\Big(\big(\partial h(Z_i\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\eta}'\big)\Sigma_i(\hat{\boldsymbol{\beta}})^{-1}\big(\partial h(Z_i\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\eta}\big)\Big)$ and $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = Z\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n}\big(\partial h(Z_i\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\eta}'\big)\big(\boldsymbol{y}_i - h(Z_i\hat{\boldsymbol{\beta}})\big)$. At convergence one obtains

$$\hat{\boldsymbol{\beta}} = \big(Z'W(\hat{\boldsymbol{\beta}})Z + P\big)^{-1} Z'W(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$$

which yields the hat matrix

$$H = Z\big(Z'W(\hat{\boldsymbol{\beta}})Z + P\big)^{-1} Z'W(\hat{\boldsymbol{\beta}}) = ZF_p(\hat{\boldsymbol{\beta}})^{-1} Z'W(\hat{\boldsymbol{\beta}})$$

and

$$AIC = -2\left(l(\hat{\boldsymbol{\beta}}) - \text{trace}(H)\right).$$

Acknowledgement:

Support from Deutsche Forschungsgemeinschaft (SFB 386) is gratefully acknowledged. We thank R. Bender who let us use the retinopathy data.

References

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. (1999). Modelling Ordered Categorical Data: Recent Advances and Future Challenges. *Statistics in Medicine* **18**, 2191–2207.

Albert, J. and Chib, S. (2001). Sequential Ordinal Modeling with Applications to Survival Data. *Biometrics* **57**, 829–836.

Armstrong, B. and Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* **129**, 191–204.

Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society* **B 55**, 549–557.

Barnhart, H. and Sampson, A. (1994). Overview of multinomial models for ordinal data. *Communications in Statistics* **23**, 3395–3416.

Bender, R. and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology* **51**, 809–816.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Statistical Science Series: Clarendon Press.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **46**, 1171–1178.

Cox, C. (1988). Multinomial regression models based on continuation ratios. *Statistics in Medicine* **7**, 435–442.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science* **11**, 89–121.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed.). New York: Springer.

Farewell, V. T. (1982). A note on regression analysis of ordinal data with variability of classification. *Biometrika* **69**, 533–538.

Firth, D., Glosup, J., and Hinkley, D. V. (1991). Model checking with non-parametric curves. *Biometrika* **78**, 245–252.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine* **13**, 1665–1677.

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

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics* **39**, 205–217.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Cornell University. Preprint.

Ruppert, D. and Carroll, R. J. (1999). Spartially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–223.

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

Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics* (forthcoming).

	MSEL	MRSEL	MEL
POM	0.0093	0.0542	0.0251
NPOM	0.0043	0.0208	0.0135
NPOM/POM	0.4336	0.3801	0.5277

TABLE 1: *Mean loss functions and proportions averaged over all simulations where Fisher scoring without penalization failed.*

	estimate	standard error	p-values LR_p	W_p	s_p
γ_{01}	6.031	0.523			
γ_{02}	7.642	0.587			
γ_{GH}	-2.932	0.627	0.000	0.000	0.000
γ_{BP}	-3.685	0.605	0.000	0.000	0.000
$\gamma_{SM,1}$	-0.409	0.207	0.012	0.012	0.032
$\gamma_{SM,2}$	0.059	0.244	0.127	0.114	0.286
$\gamma_{DD,1}$	-11.263	1.656	0.000	0.000	0.000
$\gamma_{DD,2}$	-11.880	1.701			
$\gamma_{DDQ,1}$	8.265	1.837	0.000	0.000	0.000
$\gamma_{DDQ,2}$	7.319	1.822			

TABLE 2: *Estimated effects and standard errors for the PPOM($\{GH, BP\}$). On the right p-values of the penalized test statistics are given.*

				Severity of nausea			*Total*
	None	*Mild*		*Moderate*		*Severe*	
	1	*2*	*3*	*4*	*5*	*6*	
No cisplatin	43	39	13	22	15	29	161
Cisplatin	7	7	3	12	15	14	58

Table 3: *Data on severity of nausea for patients receiving chemotherapy.*

	no restriction ($\lambda = 0$)		smoothed ($\log(\lambda) = 8$) with $d = 3$	
γ_{01}	2.131	(2.063)	2.127	(2.063)
γ_{02}	2.699	(2.068)	2.703	(2.068)
γ_{03}	3.685	(2.080)	3.668	(2.078)
γ_{04}	5.956	(2.141)	5.985	(2.138)
γ_A	-0.256	(0.138)	-0.256	(0.138)
γ_{A^2}	0.004	(0.002)	0.004	(0.002)
$\gamma_{T,1}$	0.123	(0.419)	0.129	(0.418)
$\gamma_{T,2}$	1.412	(0.392)	1.390	(0.377)
$\gamma_{T,3}$	1.541	(0.466)	1.580	(0.432)
$\gamma_{T,4}$	0.773	(0.887)	0.698	(0.804)

Table 4: *Fitted parameters for knee data (standard deviations in brackets).*

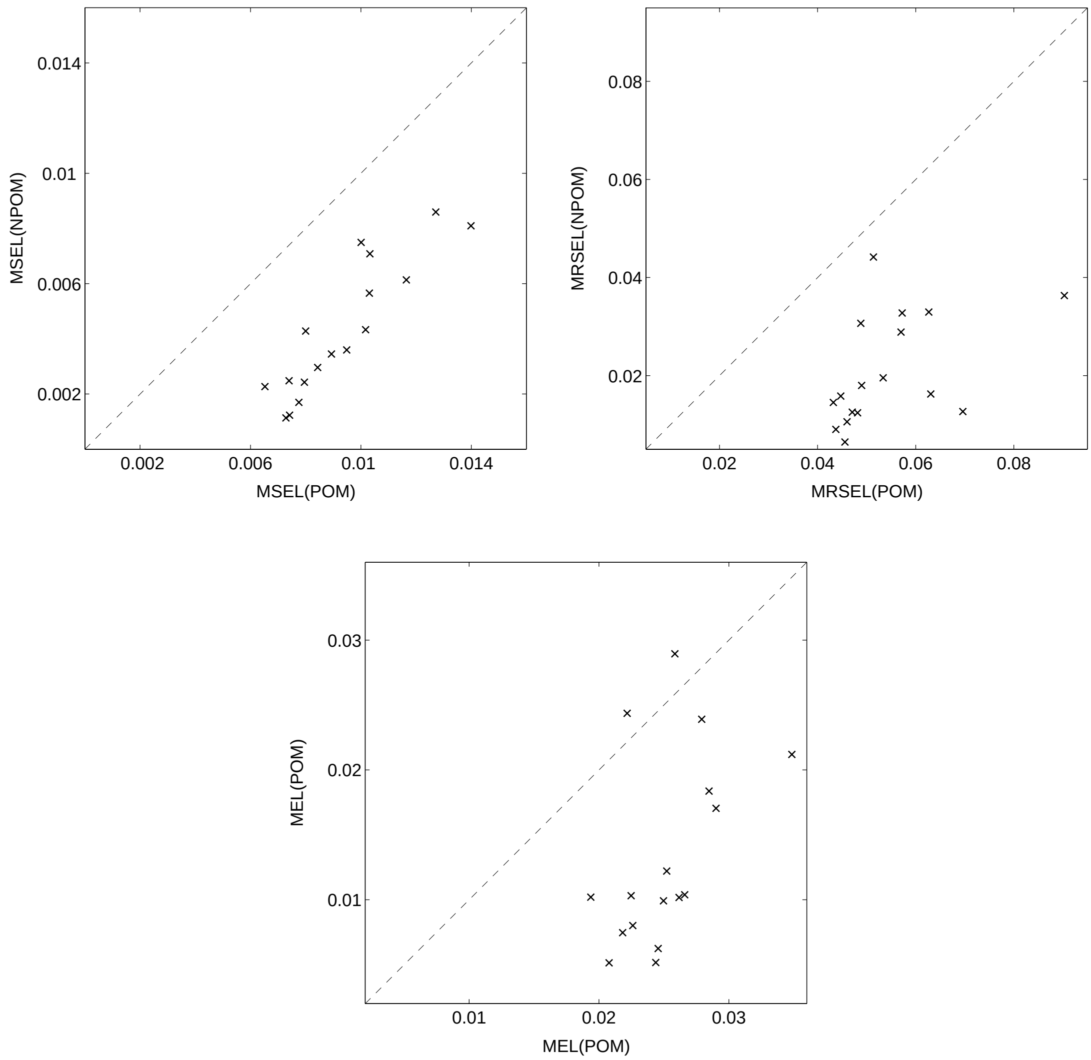

FIGURE 1: *Loss functions of the POM plotted against the corresponding loss function of the NPOM for the simulations where Fisher scoring without penalization failed. Dashed lines have intercept zero and slope one.*

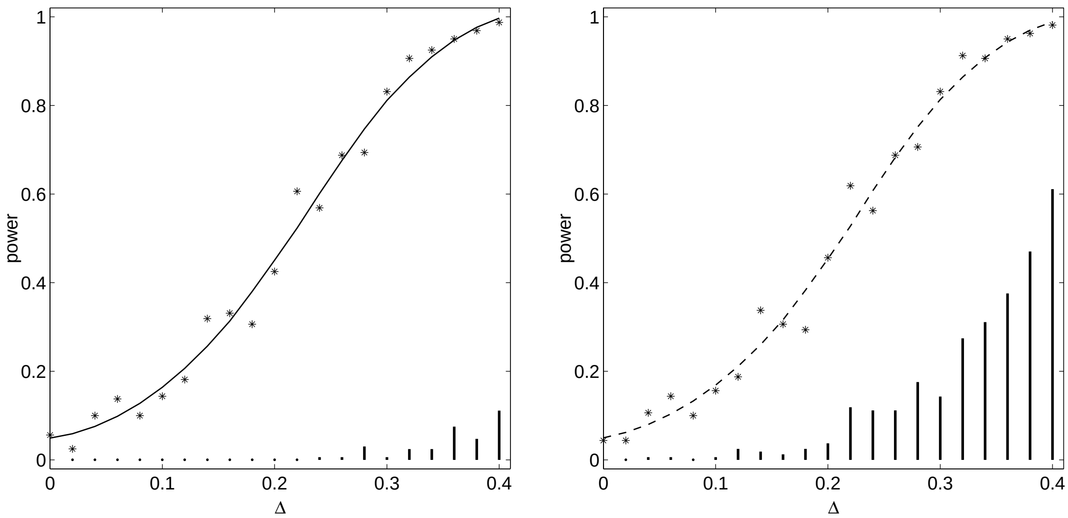

FIGURE 2: *Estimated power functions of LR_p (left) and s_p (right). Lines are local quadratic fits of the estimated points. Vertical lines indicate proportion of failures (left: penalization with $\lambda_P = \exp(4)$, right: no penalization).*

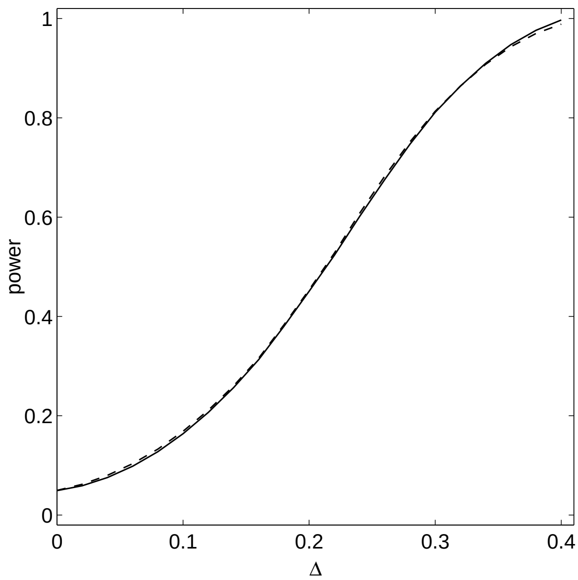

FIGURE 3: *Superimposed power functions of LR_p (solid line) and s_p (dashed line).*

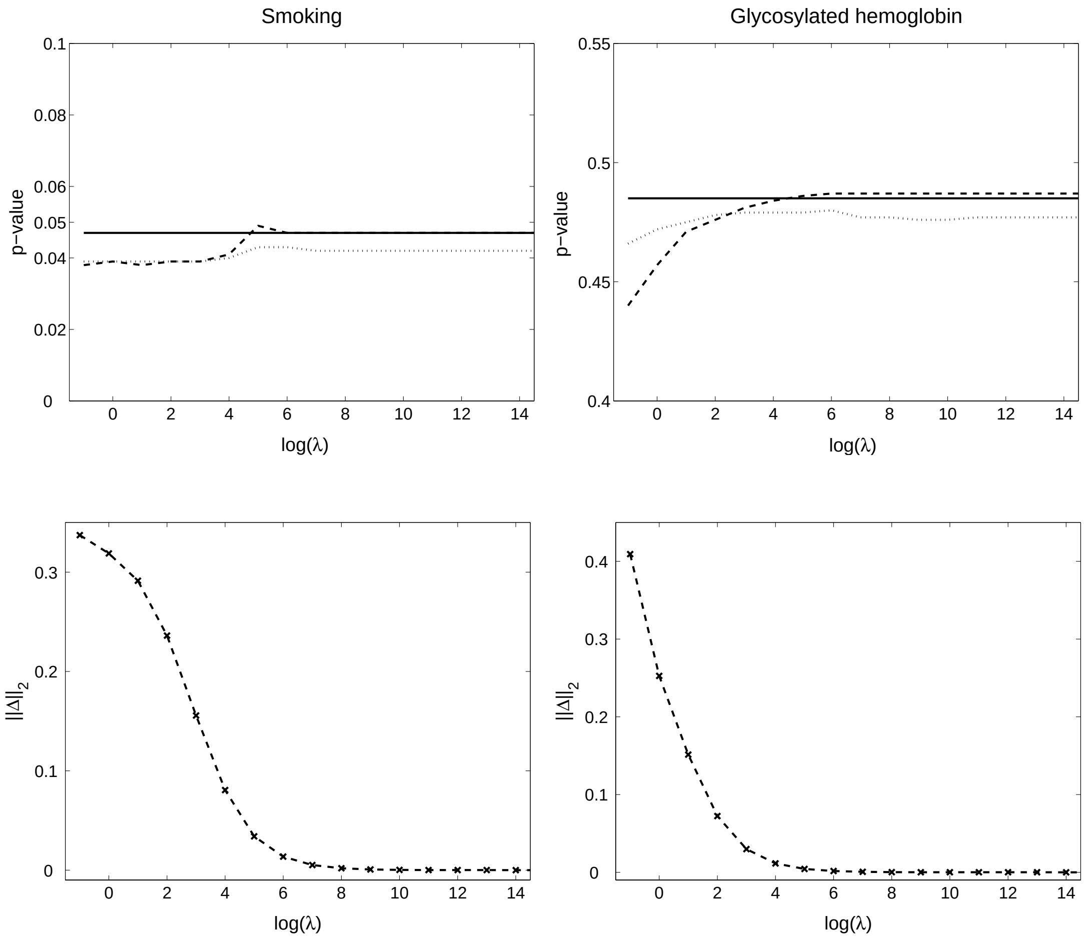

FIGURE 4: *Upper panel shows estimated p-values for testing PPOM($\{SM\}$) (left) and PPOM($\{GH\}$) (right) against the NPOM for the test statistics s_p (solid line), W_p (dashed line) and LR_p (dotted line) and different amounts of penalization. Lower panel shows deviations of penalized category-specific effects of smoking (left) and glycosylated hemoglobin (right) from the corresponding assumption of global effects.*

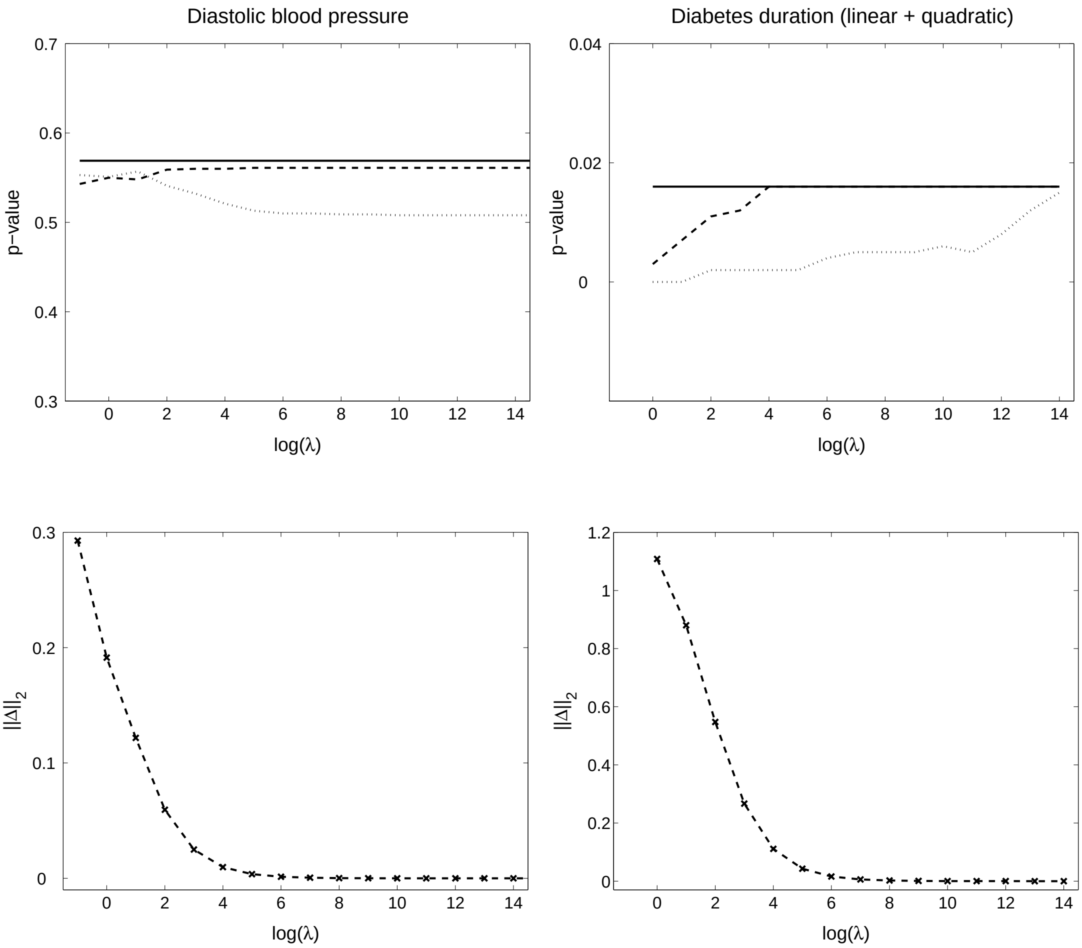

FIGURE 5: *Upper panel shows estimated p-values for testing PPOM($\{BP\}$) (left) and PPOM($\{DD, DDQ\}$) (right) against the NPOM for the test statistics s_p (solid line), W_p (dashed line) and LR_p (dotted line) and different amounts of penalization. Lower panel shows deviations of penalized category-specific effects of diastolic blood pressure (left) and diabetes duration (linear + quadratic, right) from the corresponding assumption of global effects.*

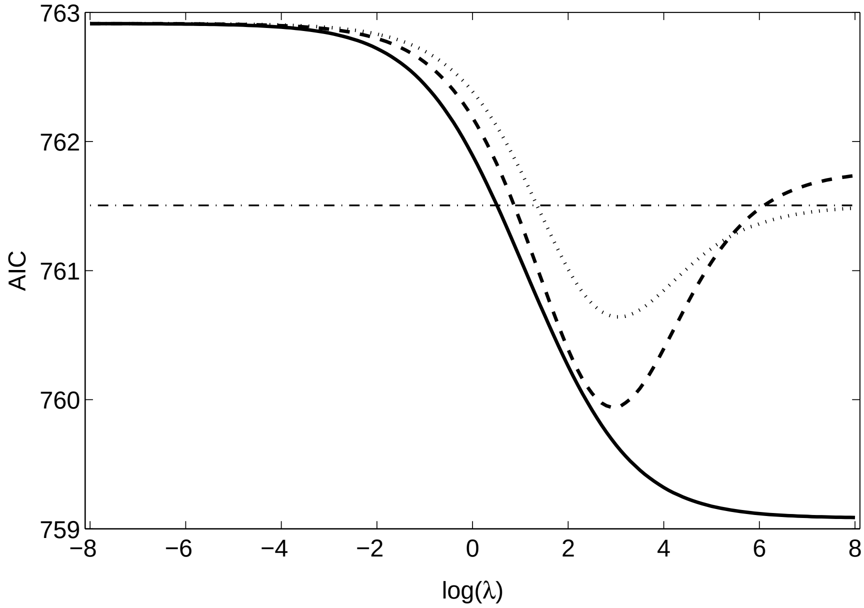

FIGURE 6: *AIC of the NPOM for varying smoothing parameter and different degrees of vertical penalization (1: dotted, 2: dashed, 3: solid). Dash-Dot line indicates the AIC of the corresponding POM.*

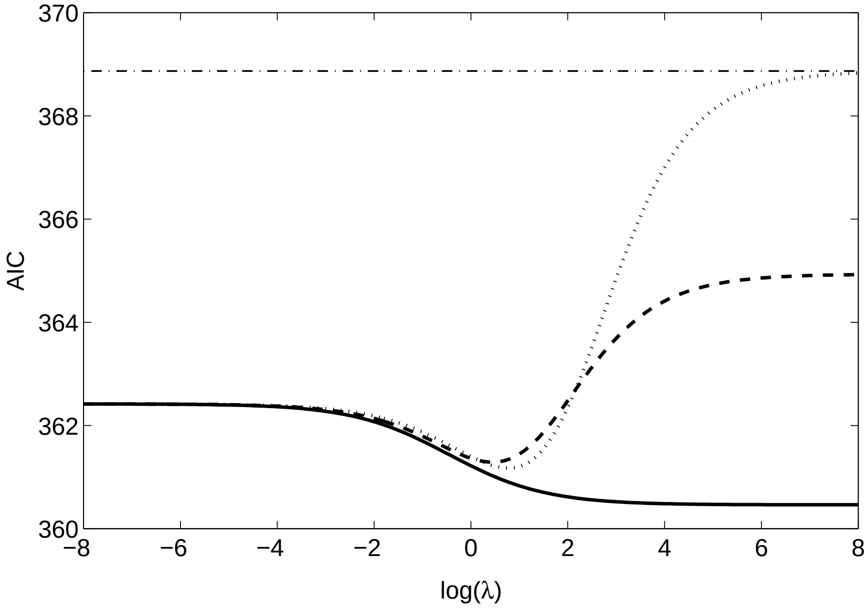

FIGURE 7: *AIC of the PPOM($\{Age, Age^2\}$) for varying smoothing parameter and different degrees of vertical penalization (1: dotted, 2: dashed, 3: solid). Dash-Dot line indicates the AIC of the corresponding POM.*