Open access • Journal Article • DOI:10.1111/J.1751-5823.2009.00088.X

Penalized regression with ordinal predictors — Source link

Jan Gertheiss, Gerhard Tutz

Institutions: Ludwig Maximilian University of Munich

Published on: 01 Dec 2009 - International Statistical Review (Blackwell Publishing Ltd)

Topics: Ordinal regression, Ordinal data, Regression analysis, Generalized linear model and Linear regression

Jan Gertheiss & Gerhard Tutz

Penalized Regression with Ordinal Predictors

Technical Report Number 015, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Penalized Regression with Ordinal Predictors

Jan Gertheiss & Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
{jan.gertheiss,tutz}@stat.uni-muenchen.de

January 9, 2008

Abstract

Ordered categorial predictors are a common case in regression modeling. In contrast to the case of ordinal response variables, ordinal predictors have been largely neglected in the literature. In this article penalized regression techniques are proposed. Based on dummy coding two types of penalization are explicitly developed; the first imposes a difference penalty, the second is a ridge type refitting procedure. A Bayesian motivation as well as alternative ways of derivation are provided. Simulation studies and real world data serve for illustration and to compare the approach to methods often seen in practice, namely linear regression on the group labels and pure dummy coding. The proposed regression techniques turn out to be highly competitive. On the basis of GLMs the concept is generalized to the case of non-normal outcomes by performing penalized likelihood estimation.

Keywords: Bayesian Methodology, Classical Linear Model, Dummy Coding, Generalized Linear Models, Generalized Ridge Regression, Ordinal Predictors, Penalized Likelihood Estimation

1 Introduction

Categorial variables that have more than two categories are often measured on ordinal scale level, so that the events described by the category numbers or class labels $1, \dots, K$ can be considered as ordered but not as equally-spaced. Following Anderson (1984) one may distinguish between two major types of ordinal categorial variables, *"grouped continuous variables"* and *"assessed ordered categorial variables"*. The first type is a mere categorized version of an underlying continuous variable, which in principle may be observed itself, e.g. if age or income are only given in categories. A variable of the second type arises when an assessor processes an indeterminate amount of information before providing his judgement on the grade of the ordered categorial variable, cf. Anderson (1984). In both cases, however, it should be kept in mind that only the ordering is meaningful.

The case of ordinal response variables has been well investigated. Starting with McCullagh's (1980) seminal paper various modeling approaches have been suggested, see for example Armstrong and Sloan (1989), Peterson and Harrell (1990), Cox (1995) for frequentistic approaches, or Albert and Chib (2001) for a Bayesian modeling approach. A more recent overview on ordered categorical response models has been given by Liu and Agresti (2005). Less work has been done concerning ordinal predictors, although ordinal predictors are often found in regression modeling. In social sciences where attitudes are measured in categories as well as in biostatistics, for example in dose-responses analyses, independent variables with discrete ordered categories are quite common. Especially for the latter case Walter et al. (1987) developed a coding scheme for ordinal predictors. For the K (ordered) levels of the independent variable $K-1$ dummy variables which describe the *"between-strata differences"* are defined. As Walter et al. point out in the case where all dummies are used it is always possible to *"convert"* from one coding scheme to another, for example to the well known dummy coding with reference category. So both coding schemes share the feature that they do not explicitly use the predictor's ordinal structure in the estimation procedure. Of course the Walter et al. scheme may offer better parameter interpretation, if e.g. *"the objective is to identify contrasts in the dependent (...) variable between successive levels of the independent variable"*. Nevertheless by using the ordinal scale level only for coefficient interpretation the method still faces the problem of overfitting and non-existence of estimates, in particular if the predictor has many categories and all dummy variables are taken into account.

In order to avoid the problems linked to dummy coding many researches prefer treating ordinal variables as metric ones. Applying methods for continuous variables to ordinal ones is particularly seen in social sciences, cf. Winship and Mare (1984). Consequently, the discussion if methods for interval-level variables in general can be used for ordinal variables as well has a long tradition. For example Labowitz (1970) supports doing so, whereas Mayer (1970, 1971) disagrees.

The problem with using continuous regressor methods is that scores have to be assigned to the categories of the predictor. If the categories represent subjective judgements like 'strong agreement', 'slight agreement', ... 'strong disagreement'' the assigned scores are typically artificial. Interpretation depends on the assigned scores which are to a certain extent arbitrary and different sets of scores usually yield different effect strengths. One may advocate the use of scores if the ordinal scale of the factor is due to an underlying continuous variable. When the intervals on which the categories are built are known, one might build mid-point scores. But even that approach has its difficulties when the upper bound is not known. If for example income is given in intervals it is hard to know what values hide in the highest (unlimited) interval. Then mid-point scales are a mere guess. In addition, the score of the highest category is at the boundary of the predictor space and therefore tends to be highly influential.

In the present paper we suggest a simple procedure to incorporate the ordinal scale level and obtain stable estimates without using assigned scores. It is proposed to use penalized estimates where the penalty explicitly uses the ordering of categories. The procedure can be described as a penalized regression technique, but we give a Bayesian motivation as well. Similar procedures have presented before. For the case of ordered predictors (as for example in signal regression) Land and Friedman (1997) introduced a lasso type penalty on differences of adjacent regression coefficients. The penalty typically yields a piecewise constant coefficient curve. When this so-called *"variable fusion"* is applied to dummy coded ordinal predictors the result is variable fusion, i.e. grouping of some classes. The use of the fused lasso (Tibshirani et al., 2005) would have a similar effect. Our goal is different, we want to conserve the given class structure. The objective of the paper is to demonstrate the usefulness of simple (quadratic) penalization techniques for ordered categorial covariates. We will start with the classical regression problem with a metric normally distributed response (Section 2) and consider the extension to generalized linear models in Section 6.

2 Penalized Regression for Ordinal Predictors

2.1 Coefficient Smoothing

Let the one-dimensional predictor x be ordinal with ordered categories $1, \ldots, K$. For the relationship between x and a normal response y we assume the classical linear model

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K + \epsilon, \tag{1}$$

with $\epsilon \sim N(0, \sigma^2)$ and $x_1, \ldots, x_K$ denoting dummy variables for the categories of x. The (0/1) dummy variables are given by

$$x_k = \begin{cases} 1 & x = k \\ 0 & \text{otherwise} \end{cases} .$$

For means of identifiability, parameters have to be constrained, for example by specifying a reference category. Let $k = 1$ be chosen as the reference category, so that $\beta_1 = 0$. For simplicity in we preliminary assume $\alpha = 0$, i.e. the mean in the reference category is assumed to be zero. But an intercept or an α that specifies the effect of some other (metric) covariates z, in terms of $\alpha = \alpha(z) = z^T\alpha$, can be easily incorporated into the proposed concept.

Rather than estimating the parameters by simple maximum likelihood methods we propose to penalize differences between coefficients of adjacent categories in the estimation procedure. The rationale behind is as follows: the response y is assumed to change slowly between two adjacent categories of the independent variable. In other words, we try to avoid high jumps and prefer a smoother coefficient vector. To be more concise, let the linear regression model be given in matrix notation by

$$y = X\beta + \epsilon, \tag{2}$$

where X denotes the $N \times (K-1)$ design matrix with full rank $K-1$, $y^T = (y_1, \ldots, y_N)$ is the response vector and $\epsilon^T = (\epsilon_1, \ldots, \epsilon_N)$ is the noise vector with independent normally distributed components $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, N$. The penalized log-likelihood that is proposed is given by

$$l_p(\beta) = -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{\psi}{2}J(\beta), \tag{3}$$

with the penalty term given by $J(\beta) = \sum_{j=2}^{K}(\beta_j - \beta_{j-1})^2$. In matrix notation it has the form

$$J(\beta) = \beta^T U^T U \beta = \beta^T \Omega \beta,$$

with

$$U = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \tag{4}$$

and $\Omega = U^T U$. Maximization of (3) yields the generalized ridge estimator

$$\hat{\beta}^* = (X^T X + \lambda\Omega)^{-1} X^T y, \tag{5}$$

with penalty matrix Ω and tuning parameter $\lambda = \psi\sigma^2$.

Since dummy coding is used, the vector $X^T y$ just contains the class-wise sums of the response values, i.e. $X^T y = (n_2\bar{y}_2, \ldots, n_K\bar{y}_K)^T$, with $\bar{y}_j$ denoting the (observed) mean of y in class j and n_j the number of observations in class j. Consequently every coefficient $\hat{\beta}_j$ is a *shrunken weighted average* of $\bar{y}_2, \ldots, \bar{y}_K$. For illustration let us consider a simple example with $K = 4$ and a balanced design, i.e. $n_j = n \; \forall j$. Then one has

$$(X^T X + \lambda\Omega) = \begin{pmatrix} n + 2\lambda & -\lambda & 0 \\ -\lambda & n + 2\lambda & -\lambda \\ 0 & -\lambda & n + \lambda \end{pmatrix},$$

with inverse

$$(X^TX+\lambda\Omega)^{-1} = c^{-1}\begin{pmatrix} (n+2\lambda)(n+\lambda)-\lambda^2 & \lambda(n+\lambda) & \lambda^2 \\ \lambda(n+\lambda) & (n+2\lambda)(n+\lambda) & \lambda(n+2\lambda) \\ \lambda^2 & \lambda(n+2\lambda) & (n+2\lambda)^2-\lambda^2 \end{pmatrix}$$

and constant $c = n^3 + 5n^2\lambda^2 + 6n\lambda^2 + \lambda^3$. Shrinkage means that every row-sum of $(X^TX + \lambda\Omega)^{-1}$ is less than c/n for all $\lambda > 0$. In the given example the explicit forms of the parameter estimates $\hat{\beta}_j$, $j = 2, 3, 4$, are derived as

$$\begin{aligned} c\hat{\beta}_2 &= n\lambda^2(\bar{y}_2 + \bar{y}_3 + \bar{y}_4) + n^2(n+3\lambda)\bar{y}_2 + n^2\lambda(\bar{y}_3 + 0), \\ c\hat{\beta}_3 &= n\lambda^2(\bar{y}_2 + 2\bar{y}_3 + 2\bar{y}_4) + n^2(n+3\lambda)\bar{y}_3 + n^2\lambda(\bar{y}_2 + \bar{y}_4), \\ c\hat{\beta}_4 &= n\lambda^2(\bar{y}_2 + 2\bar{y}_3 + 3\bar{y}_4) + n^2(n+3\lambda)\bar{y}_4 + n^2\lambda(\bar{y}_3 + \bar{y}_4). \end{aligned}$$

The first term is a (weighted) average of the observed class-wise means with lower weight for already passed classes, the second term is the observed mean in the corresponding class and the last term is the mean of the neighboring classes' means. Since the mean in the reference category is assumed to be zero, this value is inserted in the first line. A fifth class does not exist, so $\bar{y}_5$ is replaced by $\bar{y}_4$. But since explicit forms as shown above become unmanageable when K is increased, matrix notation is preferred in the following.

2.2 Bayesian Motivation

To derive a prior distribution for the vector $\beta = (\beta_2, \ldots, \beta_K)^T$ we assume that the coefficients $\beta_1, \ldots, \beta_K$ are generated by a short and very simple random walk with properties as follows:

- The differences $\delta_k = \beta_{k+1} - \beta_k$ are stationary and normal: $\delta_k \sim N(0, \tau^2)$ for all $k \in \mathbb{N}$ and $k \leq K - 1$.
- The differences $\beta_{k_2} - \beta_{k_1}, \ldots, \beta_{k_n} - \beta_{k_{n-1}}$ are independent for all $1 \leq k_1 < k_2 < \ldots < k_n \leq K$, $n \geq 3$, $k_r \in \mathbb{N}$.
- $\beta_1 = 0$.

The parameter vector $\beta = (\beta_2, \ldots, \beta_K)^T$ is multivariate normally distributed, i.e. $\beta \sim N(0, \tau^2\Gamma)$ with

$$\Gamma = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2 & 3 & \cdots & K-1 \end{pmatrix}. \qquad (6)$$

Let us now consider the classical linear normal model (2) from a Bayesian perspective. Assuming that $N(\nu, \tau^2\Gamma)$ is the prior $\pi(\beta)$ one obtains the posterior density

$$\pi(\beta|y) = c(y)f(y|\beta)\pi(\beta) = \tilde{c}(y)h(\beta|y),$$

with $c(y)$ and $\tilde{c}(y)$ denoting normalizing constants and the (not normalized) posterior density given by

$$h(\beta|y) = \exp\left(-\frac{1}{2}(\sigma^{-2}(y - X\beta)^T(y - X\beta) + \tau^{-2}(\beta - \nu)^T\Gamma^{-1}(\beta - \nu))\right).$$

The pure Bayes point estimate $\hat{\beta}_B$ of β is given by the posterior mode, i.e. $\hat{\beta}_B = \text{argmax}_\beta\{\pi(\beta|y)\} = \text{argmax}_\beta\{h(\beta|y)\}$, which can be found by minimizing the function

$$g(\beta) = (y - X\beta)^T(y - X\beta) + \frac{\sigma^2}{\tau^2}(\beta - \nu)^T\Gamma^{-1}(\beta - \nu). \tag{7}$$

Simple derivation yields

$$\hat{\beta}_B = \left(X^TX + \frac{\sigma^2}{\tau^2}\Gamma^{-1}\right)^{-1}\left(X^Ty + \frac{\sigma^2}{\tau^2}\Gamma^{-1}\nu\right). \tag{8}$$

If $\nu = 0$ the Bayes estimate equals the generalized ridge estimator $\hat{\beta} = (X^TX + \lambda\Lambda)^{-1}X^Ty$ with penalty matrix $\Lambda = \Gamma^{-1}$ and smoothing parameter $\lambda = \sigma^2/\tau^2$. Alternatively $\Lambda = \tau^{-2}\Gamma^{-1}$ and $\lambda = \sigma^2$ may be set. Therefore every generalized ridge regression with regular penalty matrix Λ can be interpreted as Bayesian approach with normal sample and prior distribution and (up to a constant) prior covariance matrix Λ^{-1}. In the special case of $N(0, \tau^2)$ iid coefficients the ordinary ridge estimator is obtained with λ equal to the ratio σ^2/τ^2 of sample and prior variance, see e.g. Hastie et al. (2001). As equation (7) shows, in general $\hat{\beta}_B$ can be seen as penalized least squares estimation with the penalty given by the Mahalanobis distance to the prior mean ν.

It is easily derived that Bayes estimators are strongly linked to coefficient smoothing as considered in the previous subsection. The inverse of Γ from (6) is

$$\Gamma^{-1} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}. \tag{9}$$

Simple matrix multiplication shows $\Gamma^{-1} = U^TU$, with U from (4). When the prior mean is set to zero, the Bayes estimate is equivalent to the generalized ridge estimate $\hat{\beta}^* = (X^TX + \lambda\Omega)^{-1}X^Ty$, with $\Omega = \Gamma^{-1} = U^TU$.

2.3 Ridge Reroughing

Rather than focusing on the penalty matrix Ω (or Γ^{-1}) we consider the general Bayes estimate (8) again and alternatively concentrate on the prior mean ν. By focusing on ν we derive an alternative estimate that is linked to scoring approaches. Especially when the ordinal predictor has many categories, it is often seen that analysts prefer treating the categorial variable as a metric one and perform simple linear regression, e.g. on the class labels. Strictly speaking this procedure is not correct, since it ignores the lower scale level of an ordinal variable, but it can be seen as a first step - when the resulting estimate is seen as a kind of prior mean ν. Therefor the class labels are tentatively treated as scores, i.e. realizations of an interval scaled predictor. With slope $\hat{\theta}$ from the corresponding linear model we can set

$$\hat{\nu} = (1, 2, \ldots, K-1)^T \hat{\theta} = R\hat{\theta}.$$

For estimating β one may use the general Bayes estimate (8). For simplicity we assume independence concerning the different β_j and set $\Omega = \Gamma^{-1} = I$, with I denoting the identity matrix, and $\lambda = \sigma^2/\tau^2$. With G denoting the design matrix for estimating θ and replacing ν by $\hat{\nu}$ one obtains the estimate

$$\hat{\beta}^{**} = (X^T X + \lambda I)^{-1}(X^T + \lambda R(G^T G)^{-1} G^T)y. \tag{10}$$

To derive a prior for ν the Bayes estimate $\hat{\beta}^{**}$ uses the simplest scoring scheme imaginable, namely the class labels. If $\alpha = 0$ is assumed, G is just a vector of length N with $G_i = k - 1$, if the ith observation is from class k.

It should be noted that the estimate $\hat{\beta}^{**}$ can also be derived without any knowledge of the Bayesian approach. Suppose we have obtained $\hat{\nu}$ via linear modeling with the class labels representing the independent variable, i.e. $\hat{\nu} = R\hat{\theta}$. When the response is predicted only using the rigorous linear model the observed errors are $y - X\hat{\nu}$. Now we can try to improve our model by fitting these residuals. Of course this approach would fail if we still assumed the same linear model as before. So one gives up the severe linear restrictions, e.g. by using dummy coding. Since overfitting should be avoided, ridge regression may be chosen. The resulting new coefficient vector is

$$\hat{\gamma} = (X^T X + \lambda I)^{-1} X^T (y - X\hat{\nu}).$$

Since we have just fitted residuals we can create an "updated" coefficient vector for the original model by adding $\hat{\nu}$ and $\hat{\gamma}$ obtaining

$$\begin{aligned}
\hat{\nu} + \hat{\gamma} &= \hat{\nu} + (X^T X + \lambda I)^{-1} X^T (y - X\hat{\nu}) \\
&= \hat{\nu} + (X^T X + \lambda I)^{-1} X^T y - (X^T X + \lambda I)^{-1} X^T X \hat{\nu} \\
&- (X^T X + \lambda I)^{-1} \lambda I \hat{\nu} + (X^T X + \lambda I)^{-1} \lambda I \hat{\nu} \\
&= (X^T X + \lambda I)^{-1} (X^T y + \lambda I \hat{\nu}) \\
&= \hat{\beta}^{**}.
\end{aligned}$$

Thus the Bayes estimate $\hat{\beta}^{**}$ is equivalent to a two-step estimate that uses specific assumptions. Fitting of residuals has already been proposed by Tukey (1977) under the name "reroughing", or "twicing" as a special case of reroughing. Therefore we we refer to $\hat{\beta}^{**}$ as "ridge reroughing". Today Tukey's reroughing, resp. twicing is often seen as a predecessor of boosting approaches, see for example Schapire (1990), Freund (1995), Freund and Schapire (1996), or Bühlmann and Yu (2003).

2.4 Selection of Smoothing Parameter λ

One way to chose an appropriate penalty parameter λ is to employ a corrected version of the Akaike information criterion (AIC) as proposed by Hurvich et al. (1998). The corrected AIC is given by

$$\mathrm{AIC}_c = \log(\hat{\sigma}^2) + \frac{1+\mathrm{tr}(H)/N}{1-(\mathrm{tr}(H)+2)/N} = \log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(H)+1)}{N-\mathrm{tr}(H)-2}, \qquad (11)$$

with H denoting the hat matrix which maps the response vector y into the space of fitted values, i.e. $\hat{y} = X\hat{\beta} = Hy$. The discrepancy between data y and fit $\hat{y}$ is measured by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = y^T(I-H)^T(I-H)y.$$

The trace of H can be interpreted as the effective number of parameters used in the smoothing fit, cf. Hurvich et al. (1998) or Hastie et al. (2001). From (5) and (10) we obtain the hat matrix corresponding to $\hat{\beta}^*$ and $\hat{\beta}^{**}$ respectively. For the coefficient smoothing approach one has

$$H^* = X(X^TX + \lambda\Omega)^{-1}X^T,$$

with $\Omega = U^TU$, and for ridge reroughing

$$H^{**} = X(X^TX + \lambda I)^{-1}(X^T + \lambda R(G^TG)^{-1}G^T)$$

is obtained. For the latter hat matrix an alternative form is given by

$$H^{**} = H_1 + H_2(I - H_1),$$

with

$$H_1 = G(G^TG)^{-1}G^T,\ H_2 = X(X^TX + \lambda I)^{-1}X^T.$$

This results from the procedure's interpretation as "fitting of the residuals". Smoothing parameters may be obtained by minimizing the AIC_c on a grid of possible λ-values.

It should be noted that from a Bayesian perspective the estimate $\hat{\lambda}$ is an estimate of the ratio σ^2/τ^2. Therefore when $\hat{\lambda}$ is plugged in, the coefficient vectors $\hat{\beta}^*$ and $\hat{\beta}^{**}$ become empirical Bayes estimators.

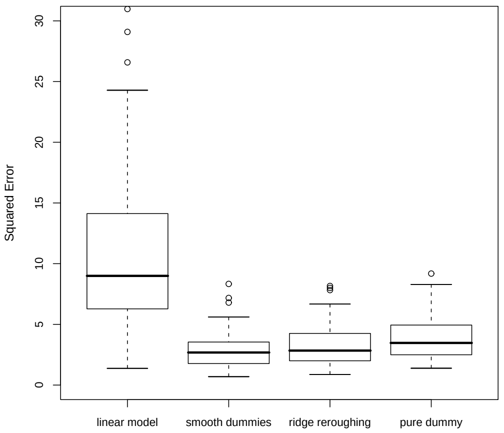

Figure 1: Squared Error for 100 simulations with $\sigma^2 = 4$.

3 Simulation Studies

3.1 Imitating the Bayesian Perspective

For our first simulation scenario we assume one ordinal independent variable with K categories and a balanced design with $N = 10K$ observations, so that in each category one has 10 observations. Let the coefficient vector β be created by a random walk as described in Section 2.2; more precisely we set

$$\beta_1 = 0;\ \beta_j = \beta_{j-1} + b_j,\ b_j \sim N(0,1)\ \text{(iid)},\ j = 2,\ldots,K.$$

Note that in this setting the coefficient vector changes from one simulation to another, whereas the design matrix remains fixed. For the design vectors x_i we use dummy coding and create the corresponding response $y_i = x_i^T\beta + \epsilon_i,\ \epsilon_i \sim N(0,\sigma^2),\ i = 1,\ldots,N$. The penalty parameters are determined by minimizing the corrected AIC. Figure 1 shows the results in terms of the squared error

$$\text{SE} = \sum_{j=2}^{K}(\hat{\beta}_j - \beta_j)^2 \tag{12}$$

for $K = 11$, $\sigma^2 = 4$ and 100 simulation runs; in case of the linear model 2 outliers are not shown. The distinct winner are the smooth dummy coefficients $\hat{\beta}^*$, followed by ridge reroughing. The first finding is not surprising, since the Bayes estimator $\hat{\beta}^*$ is the theoretically best estimate in the present situation. However, λ has to be chosen, which seems to be done quite well by minimizing the corrected AIC. The good results for ridge reroughing are somewhat unexpected. Although

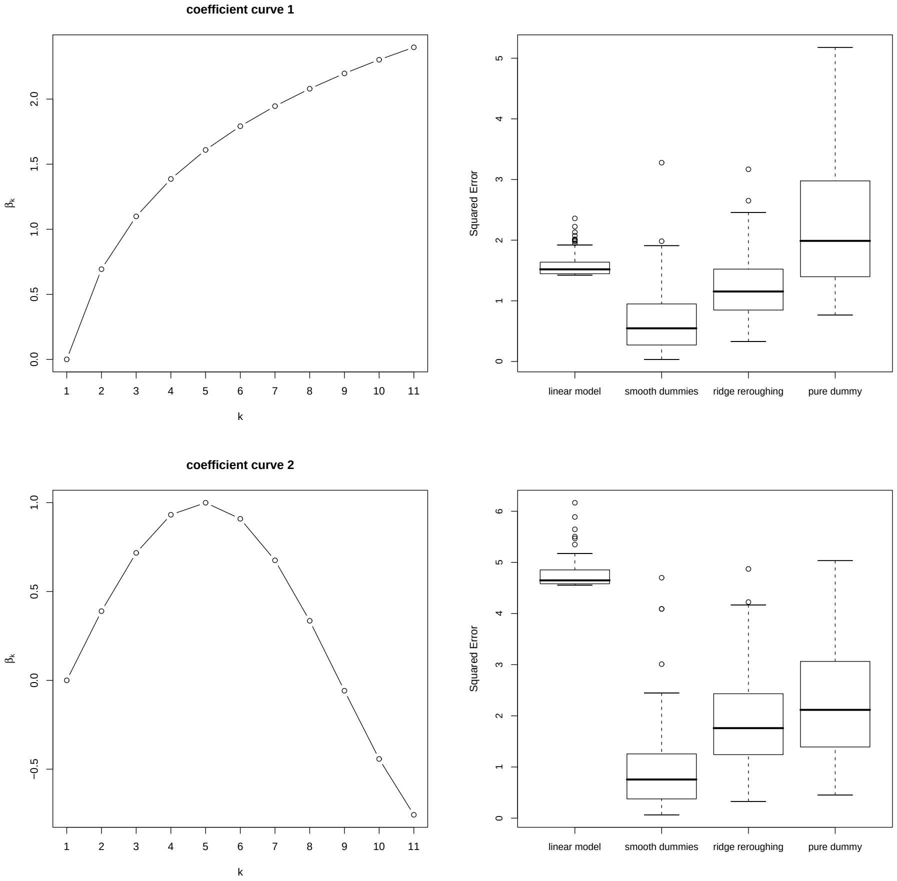

Figure 2: True coefficient vectors (left) and Squared Error for the considered methods after 100 simulation runs with $\sigma^2 = 2$.

the performance of the linear model is very bad, penalizing the distance to the corresponding coefficients apparently improves the quality of dummy coding.

3.2 Fixed Coefficient Vectors

In the following we return to the frequentistic point of view and fix the true coefficient vector β as shown in Figure 2 (left), but randomly generate a new design matrix with $N = 110$ observations in every simulation run, that means for every observation the class label is chosen at random. But data generation is not completely at random, since we only use data sets with at least one observation in each class. So the expected number of observations is 10 for each class. Figure 2 (right) shows the Squared Error for the methods on 100 simulation runs with $\sigma^2 = 2$. We chose a slightly curved coefficient vector (top) and one that is obviously nonlinear (bottom). As before the smooth dummy coefficients perform best; ridge reroughing is worse but still distinctly better than pure dummy coding. It is seen that even when the curve is approximately linear, the performance of the linear model is rather bad. Only simple dummy coding performs worse on average. Although it represents the true model, due to the high variability the number of free parameters to be estimated is too large for the estimates to be competitive.

4 Some Bias-Variance Calculations

4.1 Biased Estimation by Coefficient Smoothing

In this section we focus on the frequentistic approach and (theoretically) examine the covariance matrix $V(\hat{\beta}^*)$ and the expectation $E(\hat{\beta}^*)$, resp. the bias of the proposed estimator $\hat{\beta}^*$. From definition (5) for smoothed dummy coefficients it follows directly

$$\begin{aligned} E(\hat{\beta}^*) &= (X^TX+\lambda\Omega)^{-1}X^TX\beta = \beta - \lambda(X^TX+\lambda\Omega)^{-1}\Omega\beta, && (13)\\ V(\hat{\beta}^*) &= \sigma^2(X^TX+\lambda\Omega)^{-1}X^TX(X^TX+\lambda\Omega)^{-1}. && (14)\end{aligned}$$

So smoothed dummy coefficients are biased with

$$Bias(\hat{\beta}^*) = \lambda(X^TX+\lambda\Omega)^{-1}\Omega\beta. \qquad (15)$$

As for the original, or *standard* ridge estimator from Hoerl and Kennard (1970) the bias depends on the design, the true β-vector and the amount of shrinkage, resp. smoothing. The covariance matrix additionally depends on the variance σ^2 of course, but not on the true coefficients vector. The expected squared distance $E((\hat{\beta}^*-\beta)^T(\hat{\beta}^*-\beta))$ from the estimated to the true coefficient vector is the trace of the MSE-matrix

$$\begin{aligned} M(\hat{\beta}^*) &= V(\hat{\beta}^*) + Bias(\hat{\beta}^*)Bias(\hat{\beta}^*)^T\\ &= (X^TX+\lambda\Omega)^{-1}(\sigma^2X^TX+\lambda^2\Omega\beta\beta^T\Omega)(X^TX+\lambda\Omega)^{-1}. && (16)\end{aligned}$$

This trace is sometimes also called (scalar) MSE. It can be computed by

$$\text{MSE}(\hat{\beta}^*) = \text{tr}(V(\hat{\beta}^*)) + Bias(\hat{\beta}^*)^TBias(\hat{\beta}^*).$$

Balanced Designs

For illustration we explicitly treat the case of a balanced design with n observations in each of $K = 11$ classes, but with restriction $\alpha = 0$. Since now $X^TX = nI$, we have

$$\text{MSE}(\hat{\beta}^*) = (\sigma^2/n)\text{tr}((I+(\lambda/n)\Omega)^{-2}) + (\lambda/n)^2\beta^T\Omega(I+(\lambda/n)\Omega)^{-2}\Omega\beta.$$

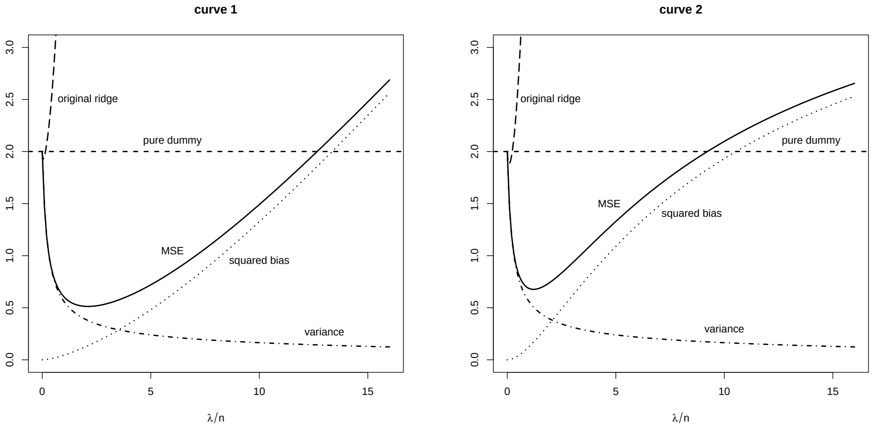

Figure 3: Scalar MSE, variance and squared bias of smoothed dummy coefficients as a function of λ/n for balanced designs with $\sigma^2/n = 0.2$ and true β-vectors from Figure 2 (the left panel corresponds to the first curve, the right panel to the second curve); additionally the MSEs of standard ridge and a pure dummy model are shown.

For σ^2/n we choose 0.2. This value is equal to the ratio of variance and mean class size in the second simulation setting in Section 3. The true coefficient vectors considered there (see Figure 2) are used here as well. Figure 3 shows the resulting MSE, the squared bias $Bias(\hat{\beta}^*)^T Bias(\hat{\beta}^*)$ and $\mathrm{tr}(V(\hat{\beta}^*))$ (denoted as variance) as a function of λ/n. For comparison we also marked the MSE of the original ridge estimator and the MSE of the unbiased pure dummy model. It is seen again that the latter can be dramatically improved by the biased estimator $\hat{\beta}^*$. In contrast standard ridge regression is not very helpful, since it was developed for non-orthogonal problems, i.e. for regression problems where the columns of the design matrix are far away from being orthogonal. In the present case of dummy coded categorial predictors, however, these columns are perfectly orthogonal.

4.2 Correction of the Standard Ridge Bias

From (10) it is seen that also the ridge reroughing estimator $\hat{\beta}^{**} = Z_\lambda y$ is a linear estimator, with $Z_\lambda = (X^TX + \lambda I)^{-1}(X^T + \lambda R(G^TG)^{-1}G^T)$. So we have

$$\begin{aligned} E(\hat{\beta}^{**}) &= Z_\lambda X \beta \\ &= \beta - \lambda A_\lambda \beta + \lambda A_\lambda R(G^TG)^{-1}G^T X\beta, \quad (17) \\ V(\hat{\beta}^{**}) &= \sigma^2 Z_\lambda Z_\lambda^T = \sigma^2 A_\lambda Q_\lambda A_\lambda^T, \quad (18) \end{aligned}$$

with $A_\lambda = (X^TX+\lambda I)^{-1}$ and $Q_\lambda = (X^T+\lambda R(G^TG)^{-1}G^T)(X^T+\lambda R(G^TG)^{-1}G^T)^T$. It is seen that the bias $-\lambda A_\lambda \beta$ of the standard ridge estimator is additively corrected by the term $\lambda A_\lambda R(G^TG)^{-1}G^TX\beta$.

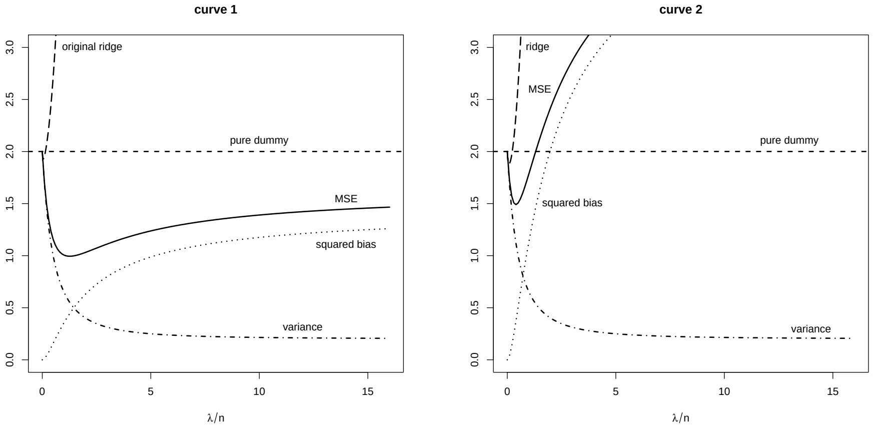

Figure 4: Scalar MSE, variance and squared bias of ridge reroughing as a function of λ/n for balanced designs with $\sigma^2/n = 0.2$ and true β-vectors from Figure 2 (the left panel corresponds to the first curve, the right panel to the second curve); additionally the MSEs of standard ridge and a pure dummy model are shown.

Balanced Designs

In case of a balanced design with n observations in each of K classes, and restriction $\alpha = 0$, one has $X^T G = nR$, $G^T G = n\kappa$, with $\kappa = \sum_{k=1}^{K-1} k^2$, and consequently

$$\begin{aligned} Q_\lambda &= (X^T + (\lambda/n)\kappa^{-1}RG^T)(X^T + (\lambda/n)\kappa^{-1}RG^T)^T \\ &= nI + n(\lambda/n)^2\kappa^{-1}RR^T + 2n(\lambda/n)\kappa^{-1}RR^T \\ &= n(I + (\lambda/n)\kappa^{-1}(\lambda/n + 2)RR^T). \end{aligned}$$

With $A_\lambda = n^{-1}(I + (\lambda/n)I)^{-1}$:

$$Bias(\hat{\beta}^{**}) = ((\lambda/n)^{-1} + 1)^{-1}(\kappa^{-1}RR^T - I)\beta$$

So the (scalar) MSE($\hat{\beta}^{**}$) = tr($V(\hat{\beta}^{**})$)+$Bias(\hat{\beta}^{**})^T Bias(\hat{\beta}^{**})$ again only depends on the true β and σ^2/n, and can be plotted as a function of λ/n. This is done in Figure 4 (with the same settings as in the previous subsection). As we see, ridge reroughing improves the pure dummy model, but not to the same extent as the smoothed coefficients. Especially if the true β is only slightly curved, ridge reroughing works quite well. The limit for $\lambda \to \infty$ is MSE($\hat{\nu}$), the mean squared error of the linear model. For a balanced design one has MSE($\hat{\nu}$) = $(\sigma^2/n)\kappa^{-1}\mathrm{tr}(RR^T) + \|(\kappa^{-1}RR^T - I)\beta\|_2^2$. With the slightly curved β, and the chosen σ^2/n, MSE($\hat{\nu}$) is only 1.622. So in this special case ridge reroughing can be expected to outperform the pure dummy model for all $\lambda > 0$.

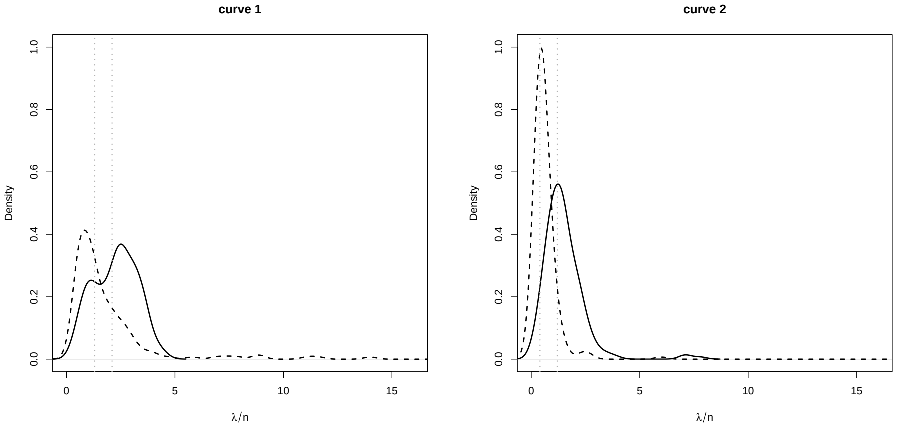

Figure 5: Kernel density estimates for estimated λ-values for smooth dummy coefficients (solid line) and ridge reroughing (dashed line), dotted grey lines mark the respective theoretically optimal λ from Figure 3 and 4; curve 1 corresponds to the first coefficient curve in Figure 2, curve 2 to the second one.

For very large values of λ/n the performance of penalized estimates is (often) worse than for simple dummy coding (see Figure 3 and 4). So it should be investigated what λ-values are actually chosen in applications. When the theoretic results from this and the previous subsection are compared to the simulations with fixed coefficient vector and approximatively balanced design in Section 3, the good performance of the AIC based tuning parameter determination procedure is confirmed. The averaged squared errors are quite close to the corresponding optimum MSEs in Figure 3 and 4. Only the narrow minimum in the right panel of Figure 4 is a little bit harder to detect when a rough grid of λ-values is used. For a specific investigation of the AIC based procedure we run a simulation that is very similar to the second one in Section 3. We explicitly have a balanced design with $n = 10$ observations in each class now and a rather fine grid is used for λ candidates. All other specifications are kept unchanged. The choice $\sigma^2 = 2$ for example means that the ratio $\sigma^2/n = 0.2$ takes the same value as in the theoretic illustration above (Figure 3 and 4). The simulation is run 200 times for each of the two coefficient curves from Figure 2. The estimated λ-values are summarized in Figure 5. It is seen from the kernel density estimate, that selected λ-values tend to be close to the optimum marked by the dotted lines and are far away from regions where bias becomes a problem (see Figure 3 and 4). The only exception is reroughing when used for curve 1. In this case some large λ-values occur (not shown). But as it is seen from the left panel of Figure 4 in this special case even with $\lambda \to \infty$ ridge reroughing is still distinctly superior to pure dummy coding. So in this special situation even estimates of λ which are much higher than the optimum are not really a problem.

5 Applications to Real World Data

In general a constant should be included when real world data is investigated. This can be done very easy by centering the data or by expanding the design matrix by a (first) column consisting of ones and the penalty matrix by a (first) column and (first) row consisting of zeros. Thus the constant is not penalized. If one wants to penalize the constant, the penalty matrix has to be modified accordingly. But we prefer not to penalize the constant. In Bayesian words, for the constant we employ an improper constant prior.

5.1 The Relationship between Age and Income

It is often fonud that there is a dependence between age and income, often assumed in terms of 'the older you are the more you earn', and modeled by a simple linear model. But before doing so you have to answer two questions. First, is the relationship monotone at all? And if it is, secondly, is it really linear?

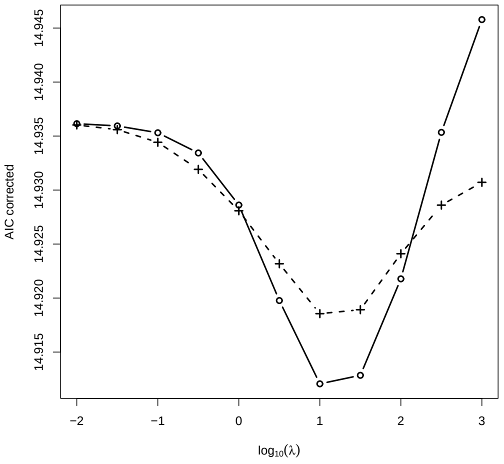

Figure 6: Curves of the corrected AIC for smooth dummy coefficients (○) and ridge reroughing (+); age/income data.

If age is only available in the form of age groups, the independent variable 'age' becomes ordinal. To answer the two questions mentioned above it may help to treat categorized variables as they are - categorial. The data set investigated here consists of $n = 190$ female scientists who are between 20 and 60 years old and living in Germany. The grouping of age a is given by: (1) $20 < a \leq 25$, (2) $25 < a \leq 30$, (3) $30 < a \leq 35$, (4) $35 < a \leq 40$, (5) $40 < a \leq 45$, (6) $45 < a \leq 50$, (7) $50 < a \leq 55$, (8) $55 < a \leq 60$. The data is taken from the Socio-Economic Panel Study (SOEP), a representative longitudinal study of private households in

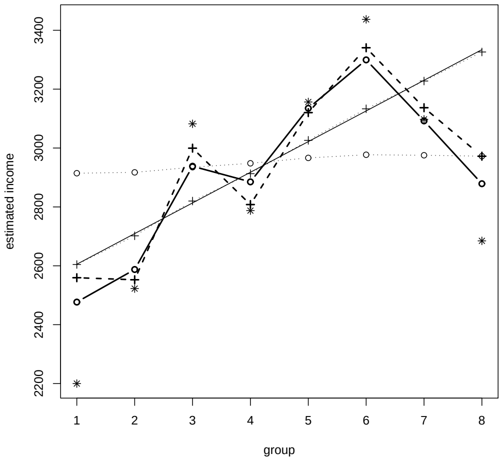

Figure 7: Estimated mean income for all age groups employing a linear model with group labels as predictors (solid line), penalized regression yielding smooth dummy coefficients (◦ and solid lines), ridge reroughing (+ and dashed lines) and a pure dummy model (∗); dotted lines with ◦ or + mark smooth, resp. ridge reroughing coefficients when $\lambda = 10^3$ is chosen.

Germany. In Figure 6 curves of the corrected AIC are shown for a wide logarithmic grid of λ-values. For both smooth dummy coefficients and ridge reroughing the criterion is minimized by $\lambda = 10$. In Figure 7 the corresponding estimated mean income for all age groups is given. For comparison we also give estimates for a linear model on the group labels, a pure dummy model, as well as estimates when the extreme value $\lambda = 10^3$ is set and differences are penalized, respectively ridge reroughing is applied. If differences between coefficients of adjacent groups are penalized, the estimates are shrunken away from the stars and towards a constant. With increasing penalty parameter λ estimates of adjacent groups are more and more alike, as seen from the dotted ◦ curve. Ridge reroughing instead shrinks estimates towards the estimates obtained by simple linear regression on the group labels (see the + curves). Since the age groups' midpoints are equally spaced, treating the group labels as independent variable is equivalent to a regression on the one-dimensional predictor 'age' composed of the midpoints of corresponding age groups. Procedures like that are often seen when the ordinal predictor may be characterized as grouped continuous variable.

Comparisons between methods with respect to prediction accuracy

When we look at the learning data only, of course pure dummy coding generally shows better fit than linear regression on the group labels, since the latter method

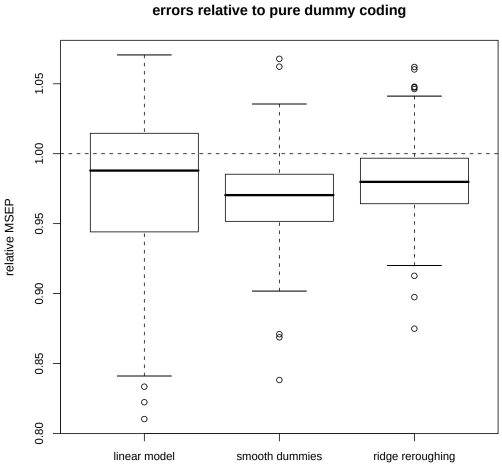

Figure 8: MSEP for the age/income data after 200 random splits into training ($m = 90$) and test ($n = 100$) data for the linear model, smoothed dummy coefficients and ridge reroughing; errors are relative to pure dummy coding.

just means imposing some severe linear restrictions on the dummies' coefficients. With an adequately chosen penalty parameter - concerning prediction accuracy - the proposed penalized regression approach should be somewhere between the other two methods.

But just looking at the learning data when comparing different methods means ignoring the problem of overfitting. So we create a training data set consisting of $m = 90$ randomly chosen observations, the remaining $n = 100$ samples serve as test set. To make sure that the pure dummy model can be fitted we restrict the analysis to training samples which contain observations from every group. The training data is used for simple linear regression with the group labels (wrongly) treated as a metric predictor, the proposed penalized regression techniques (including tuning parameter estimation by minimizing the corrected AIC) and a linear model based on pure dummy coding of the categorial predictor, i.e. ignoring its ordinal structure. Now each of these models can be used to predict the response in the test set. By comparing these predictions $\hat{y}_i$ to the true values y_i, $i = 1, \ldots, n$, one gets an idea of the considered method's true prediction accuracy. The results can be summarized in terms of the *Mean Squared Error of Prediction*

$$\mathrm{MSEP} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The procedure is repeated 200 times. Since pure dummy coding gives unbiased estimates, we take this method as reference and examine relative errors. Figure 8

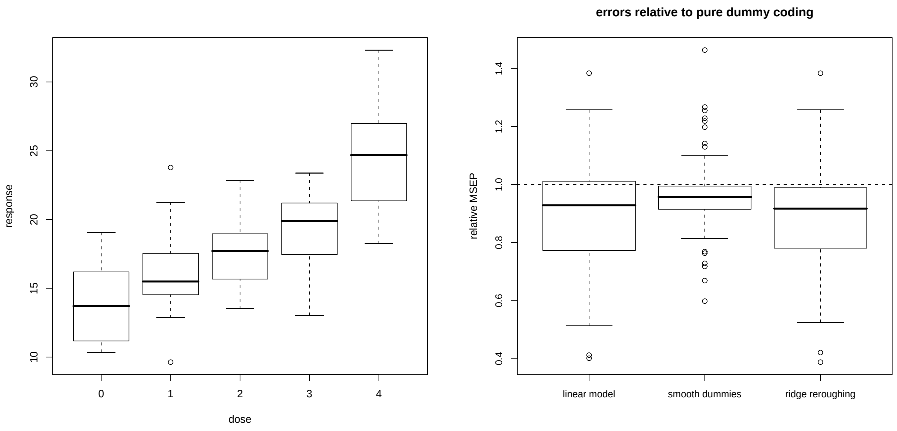

Figure 9: Summary of angina data (left) and MSEP (relative to pure dummy coding) after 200 random splits into training ($m = 20$) and test ($n = 30$) data (right).

shows a graphical summary of the observed MSEP-values (relative to pure dummy coding) for the linear model, smoothed dummy coefficients and ridge reroughing, accumulated over all random splits. It is seen that smoothed dummy coefficients and ridge reroughing yield lower MSEP-values than pure dummy coding in more that 75% of all cases. T-tests would be highly significant with p-values less than 2.2×10^{-16}.

5.2 Dose Response Analysis

The second example considered here is a dose response study of an angina drug. The data is taken from Westfall et al. (1999), respectively the R packages `multcomp` or `mratios`, see R Development Core Team (2007) for further information. The independent variable is treatment, ordinally scaled with levels 0 to 4. The response is metric: change from pretreatment as measured in minutes of pain-free walking. The left panel of Figure 9 gives a graphical summary of the data at hand. Except the last group the relationship seems to be almost linear. So the linear model can be expected to perform best. Indeed, after splitting the data into training ($m = 20$) and test set ($n = 30$), computing the MSEP and repeating this procedure as described before, the linear model can be called a winner - but together with ridge reroughing. This is shown in the right panel of Figure 9. As before we consider MSEP-values relative to those obtained with pure dummy coding. Apparently the linear model and ridge reroughing mostly outperformed simple dummy coding. Obviously penalizing the distance to the linear model worked quite well. Moreover our selection procedure to find the right penalty parameter seems to be reliable. Finally the performance of a dummy model is even improved by smoothing coefficients via a difference penalty. Of course, great differences between methods cannot be seen but small differences do exist.

6 Handling Non-Normal Responses

6.1 Estimation by Penalized Likelihood

In many applications the response y is not normally distributed, e.g. if y is dichotomous. Le Cessie and van Houwelingen (1992) considered the special case of ridge estimators in logistic regression. Such a logit model can be embedded in the context of generalized linear models (McCullagh and Nelder, 1989). Here the fundamental assumptions are as follows. Given the predictors x_j, the response y belongs to a simple exponential family. The mean μ of this distribution is linked to the linear predictor $\eta = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$ by $\mu = h(\eta)$, respectively $\eta = g(\mu)$, where h is a known one-to-one, sufficiently smooth response function, and g is the link function, i.e. the inverse of h.

The concept of generalized linear models (GLMs) serves for a generalization of the proposed penalized regression approach. The restriction $\beta_1 = 0$ still holds and for simplicity we assume at first $\alpha = 0$; in the logit model, for example, this means $P(y = 1) = 0.5$ for the reference category. But a not penalized constant can be included in analogy to the previous section, i.e. just the upper left element of the penalty matrix has to be set to zero.

The prior for $\beta = (\beta_2, \ldots, \beta_K)^T$ is the (multivariate) normal distribution with mean ν and variance/covariance matrix $\tau^2\Omega^{-1}$. As before we have to maximize the posterior density

$$\pi(\beta|y) = c(y) f(y|\beta)\pi(\beta),$$

or alternatively

$$\begin{aligned} \log(\pi(\beta|y)) &= \log(c(y)) + \log(f(y|\beta)) + \log(\pi(\beta)) \\ &= \tilde{c}(y) + l(y;\beta) - \frac{1}{2}\tau^{-2}(\beta-\nu)^T\Omega(\beta-\nu), \end{aligned}$$

with $l(y;\beta) = \log(f(y|\beta))$ denoting the log-likelihood. That means, with $\lambda = \tau^{-2}$, we have to maximize the penalized likelihood

$$l_p(\beta) = l(y;\beta) - \frac{\lambda}{2}(\beta-\nu)^T\Omega(\beta-\nu).$$

Derivatives yield

$$\frac{\partial l_p(\beta)}{\partial\beta} = s(\beta) - (\lambda\Omega\beta - \lambda\Omega\nu) = s(\beta) - \lambda\Omega\beta + \lambda\Omega\nu,$$

with $s(\beta) = \partial l(y;\beta)/\partial\beta$ denoting the score function. Now we use Fisher-Scoring, i.e. the scoring step from the current estimate $\hat{\beta}^{(k)}$ to $\hat{\beta}^{(k+1)}$, $k = 0, 1, 2, \ldots$, is given by

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (F(\hat{\beta}^{(k)}) + \lambda\Omega)^{-1}(s(\hat{\beta}^{(k)}) - \lambda\Omega\beta + \lambda\Omega\nu),$$

with $F(\beta) = E(-\partial s(\beta)/\partial\beta)$ denoting the expected Fisher information matrix. Score function and Fisher matrix are explicitly given (for example in Fahrmeir and Tutz, 2001):

$$s(\beta) = X^T D(\beta)\Sigma^{-1}(\beta)[y - \mu(\beta)], \quad F(\beta) = X^T W(\beta)X,$$

with $y = (y_1, \dots, y_N)^T$, $\mu(\beta) = (\mu_1(\beta), \dots, \mu_N(\beta))^T$, $\Sigma(\beta) = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$, $D(\beta) = \text{diag}(D_1(\beta), \dots, D_N(\beta))$, $W(\beta) = \text{diag}(w_1(\beta), \dots, w_N(\beta))$, with $D_i(\beta) = \partial h(x_i^T\beta)/\partial\eta$ and $w_i(\beta) = D_i^2(\beta)\sigma_i^{-2}(\beta)$; $\sigma_i^{-2}(\beta)$ denotes the (fitted) variance of observation i.

Choice of Penalty Parameter λ

In the case of non-normal outcomes a corrected version of the AIC is not available. Hence we employ the traditional AIC given by

$$\text{AIC} = D + 2 \cdot \text{tr}(H), \tag{19}$$

where D is the deviance of model $\hat{\mu} = h(\hat{\eta})$. Another possibility would be using a cross-validation criterion as done e.g. by Le Cessie and van Houwelingen (1992). The deviance is defined by (see e.g. Fahrmeir and Tutz, 2001)

$$D = -2\phi \sum_{i=1}^{N} (l_i(\hat{\mu}_i) - l_i(y_i)),$$

with $l(y_i)$ denoting the individual log-likelihood where μ_i is replaced by y_i (the maximum likelihood achievable). Moreover, one has to use the generalized hat matrix. In case of smoothed dummy coefficients at convergence the estimate has the form

$$\hat{\beta} = (X^T W(\hat{\beta})X + \lambda\Omega)^{-1} X^T W(\hat{\beta})\tilde{y}(\hat{\beta}),$$

with "working observations"

$$\tilde{y}(\beta) = X\beta + D^{-1}(\beta)(y - \mu(\beta)).$$

The estimate $\hat{\beta}$ is a weighted generalized Ridge estimator of the linear problem

$$\tilde{y}(\hat{\beta}) = X\beta + \epsilon,$$

The hat matrix corresponding to this model has the form

$$H = X(F(\hat{\beta}) + \lambda\Omega)^{-1} X^T W(\hat{\beta}).$$

If ridge reroughing is performed we can only give an approximative version of the generalized hat matrix. As in the normal response case we have the prior mean $\hat{\nu} = R\hat{\theta}$, with $\hat{\theta}$ estimated by Fisher scoring when the class label is taken as metric predictor; the corresponding design matrix is denoted by G. Now at convergence one has

$$\begin{array}{rcl}\hat{\beta} & = & (X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})\tilde{y}(\hat{\beta}) + (X^TW(\hat{\beta})X + \lambda I)^{-1}\lambda I\hat{\nu} \\ & = & (X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})\tilde{y}(\hat{\beta}) + (X^TW(\hat{\beta})X + \lambda I)^{-1}\lambda I\hat{\nu} \\ & + & (X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})X\hat{\nu} - (X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})X\hat{\nu} \\ & = & (X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})(\tilde{y}(\hat{\beta}) - X\hat{\nu}) + \hat{\nu},\end{array}$$

with the already introduced "working observations" $\tilde{y}(\beta) = X\beta + D^{-1}(\beta)(y - \mu(\beta))$. The estimated linear predictor is

$$X\hat{\beta} = X(X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})(\tilde{y}(\hat{\beta}) - X\hat{\nu}) + X\hat{\nu}.$$

The working observations $\tilde{y}(\hat{\beta})$ and $\tilde{y}(\hat{\nu}) = \tilde{y}(\hat{\theta})$ are just first-order Taylor approximations of $g(y)$, i.e.

$$g(y) \approx g(\mu(\beta)) + \frac{\partial g(\mu(\beta))}{\partial \mu}(y - \mu(\beta)) = X\beta + D^{-1}(\beta)(y - \mu(\beta)) = \tilde{y}(\hat{\beta})$$

and

$$g(y) \approx g(\mu(\nu)) + \frac{\partial g(\mu(\nu))}{\partial \mu}(y - \mu(\nu)) = X\nu + D^{-1}(\nu)(y - \mu(\nu)) = \tilde{y}(\hat{\nu}).$$

Since $X\hat{\nu} = G\hat{\theta}$, we have

$$X\hat{\beta} \approx H_2(I - H_1)\tilde{y}(\hat{\theta}) + H_1\tilde{y}(\hat{\theta}),$$

with $H_1 = G(G^TW(\hat{\theta})G)^{-1}G^TW(\hat{\theta})$ and $H_2 = X(X^TW(\hat{\beta})X + \lambda I)^{-1}X^TW(\hat{\beta})$. Consequently the approximate generalized hat matrix is defined in analogy to the normal response case by

$$H = H_1 + H_2(I - H_1).$$

6.2 Simulations

In analogy to the normal response case we compare the proposed penalized regression approaches to a (generalized) linear model that takes the group labels as (metric) independent variable, and to a GLM based on pure dummy coding. Probably the most famous GLM is the logit model, i.e. we assume

$$P(y = 1) = \frac{\exp(x^T\beta)}{1 + \exp(x^T\beta)}.$$

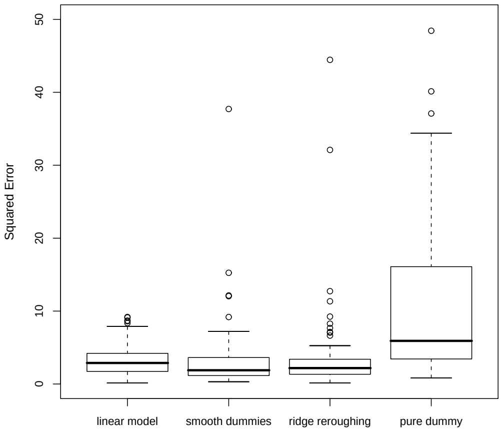

Figure 10: Squared Error for the considered methods over 100 simulation runs.

As in the simulation study above the first scenario is as follows: The design is balanced, whereas the coefficient vector β is generated by a random walk with $N(0, 0.5^2)$ distributed steps (see also Section 3). With the true probabilities $\pi_i = \exp(x_i^T\beta)/(1 + \exp(x_i^T\beta))$ the dichotomous response y is generated by the corresponding binomial distribution, i.e. $y_i \sim B(1, \pi_i)$. Sometimes complete data separation may happen. In these cases the pure dummy model and in a very unlikely case even the linear model cannot be estimated. Here we set $\hat{\beta} = 0$, $\hat{\nu} = 0$ respectively. Figure 10 shows the squared error SE (as defined in (12)) for the considered methods over 100 simulation runs; in case of pure dummy coding 4 outlier are not shown. Pure dummy coding is clearly outperformed by the other three methods. Penalized regression for smoother coefficients as well as ridge reroughing are better than the linear regression on the group labels.

As a second scenario we fix the true β as shown in Figure 2 (top) but shrink by factor 0.5 and randomly generate the design matrix with $N = 330$ observations. The 0/1-coded response is generated as before. In case of complete data separation we proceed as described above. Figure 11 (left) shows the results in terms of the Squared Error. As in the normal response case we finally assume an obviously nonlinear coefficient vector (see Figure 2, bottom). The results are visualized in Figure 11 (right). Smoothed dummy coefficients distinctly outperform the other methods in both situations. Also ridge reroughing is clearly better than dummy coding. Not surprisingly the linear model performs very bad in case of a highly curved coefficient vector.

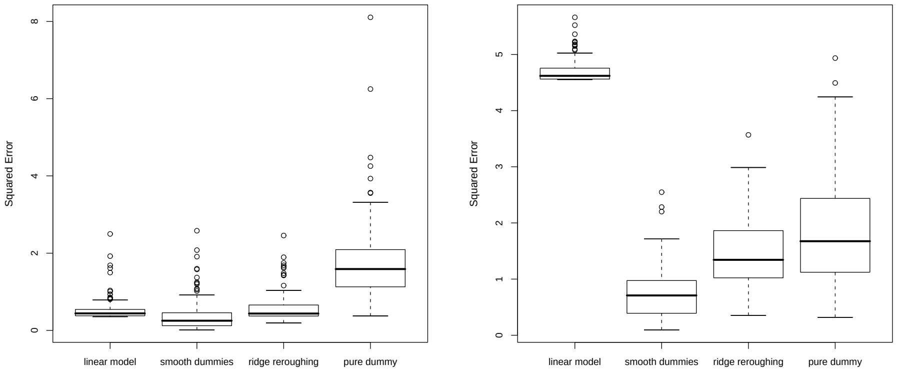

Figure 11: Squared Error for the considered methods over 100 simulation runs; left: true coefficient vector as shown in Figure 2 (top) but shrunken by 0.5, right: true coefficient vector as shown in Figure 2 (bottom).

6.3 Application to Real World Data

The data investigated here is a subsample from a study about coffee drinkers. The (dichotomous) response is coffee brand, which is only separated into cheap coffee from a German discounter and real branded products. The (ordinal) predictors are monthly income (in a categorized version), social class and age group. A more precise description can be found in the table below.

variable	group	description
age group	1	0 to 24 years
	2	25 to 39 years
	3	40 to 49 years
	4	50 to 59 years
	5	60 years or older
social class	1	lower class
	2	lower middle class
	3	medium middle class
	4	upper middle class
	5	upper class
monthly income	1	0 to 749 Euro
	2	750 to 1249 Euro
	3	1250 to 1749 Euro
	4	1750 Euro or more

Figure 12 shows the marginal distributions of the independent variables in the data set. The light-colored parts correspond to consumers of the considered

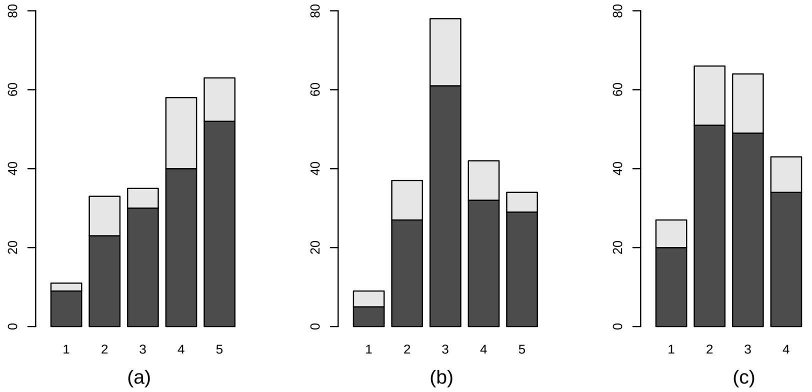

Figure 12: Marginal distributions of the independent variables: (a) age group, (b) social class and (c) monthly income in the data set; the light-colored parts correspond to drinkers of cheap coffee.

cheap coffee brand. A kind of structure can be seen for example with respect to the social class. As expected, consumers from the upper class rather buy brand products - compared with middle and lower class. Further data analysis however is done employing the proposed penalized regression approaches.

So far we assumed a single independent variable, but models with several predictors are an obvious extension. Only the penalty matrix has to be modified. For smoothed coefficients now a block diagonal structure is given, because differences between coefficients belonging to different predictors should not be penalized. If ridge reroughing is performed, the penalty matrix stays the same as before, but the prior mean results from a GLM (a logit model in the case investigated here) with more than one independent variable. Since all categorial predictors are measured on the same scale (because of dummy coding), a single penalty parameter λ may be sufficient for a initial modeling approach.

Table 1 shows the estimated coefficients of corresponding dummy variables, when logit modeling with group labels as predictors, the two proposed penalized regression approaches, as well as logit modeling based on pure dummy coding is performed. With $\lambda = 10$ the methods' characteristics can be nicely illustrated. It is seen that coefficients from ridge reroughing (column 3) are shrunken towards the coefficients in the first column. The latter have a strict linear structure, i.e. coefficients of dummy variables belonging to the same predictor specify a linear function. This results from our choice to use the class labels (as independent variables) to build this first reference model. Midpoints cannot be taken to create a pseudo interval scaled predictor. Either there are no midpoints, because the variable (social class in that case) is not a categorized version of a metric variable, or some classes do not have sharp limits, e.g. age group number 5.

		linear model	smooth dummies	ridge RR	pure dummy
intercept		−0.36	−0.81	−0.35	−0.38
age group	2	−0.10	0.06	0.03	0.79
	3	−0.20	−0.09	−0.40	−0.29
	4	−0.30	0.05	−0.03	0.84
	5	−0.40	−0.18	−0.50	−0.04
social class	2	−0.28	−0.14	−0.31	−0.92
	3	−0.56	−0.31	−0.66	−1.39
	4	−0.84	−0.39	−0.75	−1.28
	5	−1.12	−0.56	−1.17	−1.96
income	2	0.02	−0.05	−0.07	−0.13
	3	0.03	−0.02	0.08	0.29
	4	0.05	−0.04	0.03	0.17

Table 1: Coefficients of corresponding dummy variables, estimated by the use of a (generalized) linear model, i.e. logit model, with group labels as predictors, penalized regression ($\lambda = 10$*) yielding smooth dummy coefficients, ridge reroughing (*$\lambda = 10$*) and a logit model based on pure dummy coding.*

Coefficient smoothing is done by penalizing differences between coefficients of adjacent groups. That is why in column 2 coefficients between the horizontal lines are quite similar - especially compared to the pure dummy model in the last column.

To investigate the methods' performance in terms of prediction accuracy, as before, the data is randomly split into training ($m = 100$) and test ($n = 100$) data. The training data is for penalty parameter determination and model fitting, the test set for evaluation only. As a measure of prediction accuracy we take the Sum of Squared Deviance Residuals (SSDR) on the test set. In the special case of a logit model we have (with convention $0 \cdot \log(0) = 0$)

$$\begin{aligned} \text{SSDR} &= \sum_{i=1}^{n} \left(y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{\pi}_i}\right)\right) \\ &= \sum_{i:y_i=1} \log\left(\frac{1}{\hat{\pi}_i}\right) + \sum_{i:y_i=0} \log\left(\frac{1}{1 - \hat{\pi}_i}\right). \end{aligned}$$

Le Cessie and van Houwelingen (1992) call the summands, i.e. the squared deviance residuals, "*minus log-likelihood errors*".

Figure 13 (left) summarizes the results in terms of SSDR after 200 random splits. Quite often the pure dummy model could not be fitted due to complete data separation. (In the reported study this happened exactly in 64 of 200 realizations.) Therefore the boxplot for the pure dummy model is only based on

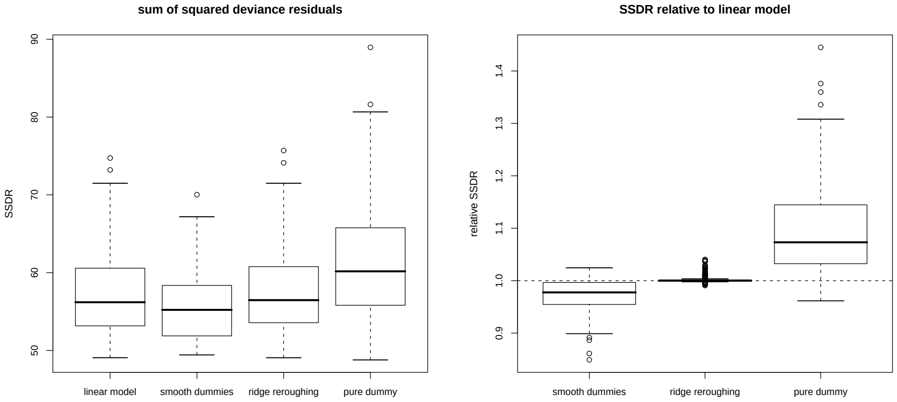

Figure 13: Performance (in terms of SSDR) of a (generalized) linear regression on the group labels, penalized regression for smoother dummy coefficients, ridge reroughing and a pure dummy model (for the latter only the 68% successful estimates have been used); left: observed values for all considered methods, right: SSDR-values relative to linear model.

the cases when the corresponding maximum likelihood estimates did exist. The results for ridge reroughing and the linear model are almost equal, but smoothing dummy coefficients tends to yield smaller values of SSDR. The pure dummy model instead performs clearly worst. So SSDR-values relative to pure dummy coding would not provide more insight, with disregarding 32% of the results. Hence the linear model is rather taken as reference (see Figure 13, right). The equal performance of ridge reroughing and the linear model as well as the superior performance of smooth dummies is confirmed.

7 Summary and Discussion

We propose a penalized regression technique to handle ordinal predictors. We started from a classical linear model for normal outcomes and dummy coded one-dimensional predictors. A Bayesian motivation was given but it was also illustrated how the estimation procedure can be derived without any use of Bayesian methodology. Two major types of penalized regression were developed. The first means penalizing the differences between coefficients of adjacent groups, the second can be described as a kind of refitting or reroughing (Tukey, 1977) procedure. Since hat matrices are given, both approaches can be seen as linear smoothers of a normal response variable. Hence the penalty parameter could be determined via a corrected version of the AIC as proposed by Hurvich et al. (1998). In a second step the approach was generalized to non-normal outcomes by employing the concept of generalized linear models (McCullagh and Nelder, 1989) and penalized likelihood estimation.

Our approach was compared to 'standard procedures', namely linear regression on the group labels and pure dummy coding. In both simulation studies and real world data evaluation the proposed regression approaches turned out to be competitive with respect to prediction accuracy. Except for the angina data coefficient smoothing performed best in all settings. Ridge reroughing was mostly worse than the penalized differences approach but always better than (or at least as good as) the considered reference methods. So - compared to the latter - performing ridge reroughing would have never been a mistake. An explanation could be as follows: Due to its construction ridge reroughing is a (data driven) tradeoff between the two generally seen extremes, a rigorous linear model that (wrongly) assumes interval scaled data and the flexible dummy coding that faces the problem of overfitting and ignores the labels' ordering. By an automatic penalty determination procedure, to a certain extend, the linear model is corrected away from linearity, but dependent on the data at hand. Nevertheless in the vast majority of applications the model may be further improved by coefficient smoothing.

Nonparametric regression on the group labels should have similar results as the proposed technique, particularly in the simulation settings with fixed coefficient curves. The procedure proposed in this article, however, can been interpreted as a nonparametric method. Nonparametric regression is usually done via basis expansion of e.g. regression splines. But choosing the adequate number and placing of basis functions, resp. knots is a complex task. A common procedure is to use a relative large number of equally spaced knots and a penalty on the basis coefficients, see for example Eilers and Marx (1996). Since an ordinal categorial predictor can only take some discrete values, the (estimated) regression function only needs to be defined on these values and dummy coding can be seen as a special, but somewhat natural and most flexible basis expansion of the underlying regression function.

References

Albert, J. H. and S. Chib (2001). Sequential ordinal modelling with applications to survival data. *Biometrics 57*, 829–836.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society B 46*, 1–30.

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

Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association 98*, 324–339.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine 14*, 1191–1203.

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

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.).

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation 121*, 256–285.

Freund, Y. and R. E. Schapire (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, 148–156.

Hastie, T., R. Tibshirani, and J. H. Friedman (2001). The elements of statistical learning. *Springer-Verlag, New York, USA.*

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hurvich, C. M., J. S. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B 60*, 271–293.

Labowitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review 35*, 515–524.

Land, S. R. and J. H. Friedman (1997). Variable fusion: A new adaptive signal regression method. Technical report 656, Department of Statistics, Carnegie Mellon University Pittsburg.

Le Cessie, S. and J. C. van Houwelingen (1992). Ridge estimators in logistic regression. *Applied Statistics 41*, 191–201.

Liu, Q. and A. Agresti (2005). The analysis of ordinal categorical data: An overview and a survey of recent developments. *Test 14*, 1–73.

Mayer, L. S. (1970). Comment on "the assignment of numbers to rank order categories". *American Sociological Review 35*, 916–917.

Mayer, L. S. (1971). A note on treating ordinal data as interval data. *American Sociological Review 36*, 519–520.

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

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall.

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

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*, 197–227.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness vie the fused lasso. *Journal of the Royal Statistical Society B 67*, 91–108.

Tukey, J. (1977). *Explanatory Data Analysis*. Addison Wesley.

Walter, S. D., A. R. Feinstein, and C. K. Wells (1987). Coding odinal independent variables in multiple regression analysis. *American Journal of Epidemiology 125*, 319–323.

Westfall, P. H., R. D. Tobias, D. Rom, R. D. Wolfinger, and Y. Hochberg (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute Inc.

Winship, C. and R. D. Mare (1984). Regression models with ordinal variables. *American Sociological Review 49*, 512–525.