Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Chapter 3

Prediction Accuracy of Regression with Optimal Scaling Transformations: The .632 Bootstrap with CATREG

Methods for nonlinear multiple regression are studied in terms of prediction accuracy. The focus is on CATREG, a method that is based on optimal scaling. Although CATREG can deal with numeric variables, it has its roots in the area of categorical variables. CATREG maximizes the multiple correlation over optimal quantifications, and this measure is inversely related to the apparent prediction error. The latter is contrasted to expected prediction error. In this chapter it is shown that CATREG compares favorably to other methods well-known in statistics, but can do even better when the number of observations in numeric variables is reduced to a much smaller of categories.

3.1. Introduction

When numeric data are available, linear multiple regression is the most often used method to predict a response or output variable from a set of predictor

This chapter has been submitted for publication as Van der Kooij, A.J. & Meulman, J.J. (2006). Prediction Accuracy of Regression with Optimal Scaling Transformations: The .632 Bootstrap with CATREG.

or input variables. Over the years, a number of nonlinear generalizations of multiple regression have appeared in the psychometric as well as the more mainstream statistical literature. These nonlinear generalizations had two purposes: to accomodate categorical variables in multivariate analysis, and to linearize nonlinear relationships in numeric data. The psychometric contribution to the area has been innovative, which has been acknowledged in the statistical literature as well (for example, see Buja (1990)). In more recent years, new nonlinear regression methods have been developed actively in computer science, in the area of data mining. These techniques are usually subsumed under the name machine learning. A shift in emphasis from computer science to a more statistical approach, can be deduced from the use of the term statistical learning (Hastie et al. 2001). A major difference between the psychometric literature and the statistical learning literature is the study of the prediction accuracy of a nonlinear regression technique, focusing on either the apparent or the expected error rate. In the psychometric literature, the emphasis has been on the prediction of the observed response variable, minimizing the apparent error rate to obtain robust, stable estimates of the regression coefficients. In statistical learning, one is mainly interested in predicting future outcome variables. In the latter, the observed values on the predictor and outcome variables are only used to obtain estimates for the parameters, to be applied to new observations to predict future, unknown, outcomes. Here the emphasis is on the expected prediction error rate.

In this chapter, we will focus on a particular nonlinear method to perform multiple regression when the data consist of ordered or unordered categorical variables. The method, called CATREG, uses the optimal scaling methodology as developed in the Gifi system (Gifi 1990) to quantify categorical variables according to a particular scaling level, thus "transforming" categorical variables into numeric variables. However, optimal scaling can be applied to numeric data as well, and thus the optimal scaling methodology is also very suitable when nonlinear relationships exist between numeric predictor and response variables. In the optimal scaling process the multiple regression coefficient is maximized by optimally quantifying categorical variables, under the restriction that specific information in the observed variables has to be preserved in the quantified variables. The kind of information that should be retained, and thereby the form of the transformation, determines the optimal scaling level that is chosen for each variable. The term "scaling level" is different from "measurement level", which describes the scale properties of the observed variable. Scaling level is used to refer to the level at which a particular variable is analyzed, and describes the assumed relation (transformation) between the values of the observed variable and the quantified variable. Scal-

ing levels can be ordered in a hierarchy, going from the most restrictive level (numeric, linear) at the top, via somewhat less restrictive levels (monotonic splines, ordinal step functions), to the least restrictive levels (nonmonotonic splines, nominal step functions). While the nominal and ordinal scaling levels produce stepwise transformations, the spline options produce smooth piecewise polynomial transformations. The results with the nominal and ordinal scaling levels will be strictly invariant under one-to-one nominal respectively ordinal transformations of the observed variables; for spline transformations, this is generally not the case.

As mentioned above, a variety of nonlinear prediction models and programs that include transformations of the variables in the optimization process, have been developed over the last decades, starting with the Box-Tidwell (Box and Tidwell 1962) and the Box-Cox models (Box and Cox 1964), using parametric families of transformations. The optimal scaling methodology originated in psychometrics with Kruskal's nonmetric version of multidimensional scaling (Kruskal 1964a,b), next applied to analysis of variance Kruskal (1965), which approach was adopted by ADDALS (De Leeuw et al. 1976) and MORALS (Young et al. 1976). The collective work by the Leiden group at the department of Data Theory resulted in Gifi (1990). Winsberg and Ramsay (1980) introduced monotonic splines in multiple regression (for a review, see Ramsay (1988)). Optimal scaling appeared in the mainstream statistical literature with ACE (Alternating Conditional Expectations) in Breiman and Friedman (1985) and Buja (1990). CATREG also has a relationship with projection pursuit regression Friedman and Stuetzle (1981), generalized additive models (GAM), extensively described in Hastie and Tibshirani (1990), and the methods described in Buja et al. (1989) and Hastie, Tibshirani, and Buja (1994).

This chapter will concentrate on the CATREG approach to nonlinear multiple regression, because of its unique emphasis on categorical variables; the model and algorithm will be described in Section 3.2. The major part of the chapter focuses on the optimality of the quantifications that are obtained from the observed data to predict future responses, i.e., the expected prediction error, studied by the .632 bootstrap procedure Efron (1983), described in Section 3.3. The prediction accuracy for CATREG will be compared with the results for various other methods for nonlinear multipe regression in Section 3.4. Prediction accuracy will also used to compare different scaling levels and number of categories (Section 3.5). Section 3.6, finally, reports the effect of sample size on the prediction accuracy.

3.2. CATREG: Regression with optimal scaling transformations

3.2.1 CATREG model

The CATREG model fits the classical linear regression model with nonlinear transformations of the variables, written as

$$\varphi_r(\mathbf{y}) = \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j) + \mathbf{e}, \tag{3.1}$$

by minimizing the least squares loss function

$$L(\varphi_r; \varphi_1, \ldots, \varphi_J; \beta_1, \ldots, \beta_J) = N^{-1} \|\varphi_r(\mathbf{y}) - \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j)\|^2, \tag{3.2}$$

with N the number of observations, J the number of predictor variables, $\{\beta_j\}$, $j = 1, \ldots, J$, the regression coefficients, $\varphi_r(\mathbf{y})$ the transformation for the response variable $\mathbf{y}$, $\varphi_j(\mathbf{x}_j)$ the transformations for predictor variables $\{\mathbf{x}_j\}$, $j = 1, \ldots, J$, and $\mathbf{e}$ the error vector, and where $\|\cdot\|^2$ denotes the squared Euclidean norm. Loss function (4.13) is minimized over $\{\beta_j\}$, $\{\varphi j(\mathbf{x}_j)\}$, and $\varphi_r(\mathbf{y})$ to maximize the least squares fit between $\varphi_r(\mathbf{y})$ and the linear combination $\sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j)$. Because the transformed variables $\varphi_r(\mathbf{y})$ and $\{\varphi_j(\mathbf{x}_j)\}$ are centered and normalized to have sum of squares equal to N, loss function (4.13) maximizes the (squared) multiple correlation. (An alternative formulation of multiple regression with optimal scaling was given in Van der Kooij and Meulman (1997).)

The form of the transformations $\varphi(\mathbf{y})$ and $\varphi_j(\mathbf{x}_j)$ depends upon the optimal scaling level that is chosen for each variable, and the scaling level defines which part of the information that is present in the observed predictor and/or response variables should be retained in the transformed variable. With the numerical scaling level, a variable is treated as quantitative, and a linear transformation is applied, so that all information is preserved. With the nonnumerical scaling levels, the variables are treated as qualitative, and the optimal scaling process turns them into quantitative variables, retaining as much information in the observed variables as is determined by the scaling level. The ordinal scaling level and monotonic splines retain only the grouping and ordering information, and result in a monotonic transformation, in the form of a step function or a spline function, respectively. The nominal scaling level and nonmonotonic splines retain only the grouping information (objects with equal values in the observed data will obtain equal optimal scale

values), resulting in a nonmonotonic transformation, again either in the form of a step function or a nonmonotonic spline function. Nonnumerical scaling levels allow for nonlinear transformations, and possible nonlinear relationships between the response variable and the predictor variables will be linearized.

3.2.2 CATREG algorithm

In the CATREG approach to nonlinear multiple regression, the data are assumed to be categorical, thus consisting of discrete, integer values. Having discrete data, a variable can be coded into an $N \times C_m$ indicator matrix $\mathbf{G}_m$, where N is the number of observations and C_m denotes the number of categories of variable m, $m = 1, \ldots, M$, where M is the total number of variables, thus, $M = J + 1$. An entry $g_{ic(m)}$ of $\mathbf{G}_m$, where $c = 1, \ldots, C_m$, is 1 if observation i is in category c of variable m, and zero otherwise. Then the transformed variables can be written as the product of the indicator matrix $\mathbf{G}_m$ and a C_m-vector of category quantifications $\mathbf{v}_m$

$$\varphi_r(\mathbf{y}) = \mathbf{G}_r\mathbf{v}_r, \text{ and } \varphi_j(\mathbf{x}_j) = \mathbf{G}_j\mathbf{v}_j, j = 1, \ldots, J. \tag{3.3}$$

So, the CATREG model with the transformed variables written in terms of indicator matrices and category quantifications is

$$\mathbf{G}_r\mathbf{v}_r = \sum_{j=1}^{J} \beta_j \mathbf{G}_j\mathbf{v}_j + \mathbf{e}, \tag{3.4}$$

with the associated least squares loss function

$$L(\mathbf{v}_r; \mathbf{v}_1, \ldots, \mathbf{v}_J; \beta_1, \ldots, \beta_J) = N^{-1}\|\mathbf{G}_r\mathbf{v}_r - \sum_{j=1}^{J} \beta_j \mathbf{G}_j\mathbf{v}_j\|^2. \tag{3.5}$$

Note that in the CATREG approach a continuous variable with N different values is regarded as a categorical variable with N different categories. In that case, the indicator matrix will be a permutation of the identity matrix. So, basically, a category in CATREG is defined as a distinct value of a variable.

The loss function (3.5) is minimized by an alternating least squares algorithm, alternating between estimation of the quantification of the response variable on the one hand, and estimation of the quantifications and regression coefficients of the predictor variables on the other hand. First, the quantifications and the regression coefficients have to be initialized. In the CATREG

algorithm, two ways of initialization have been implemented: random and numerical. Random initialization uses standardized random values from a univariate distribution for the initial quantifications, and the initial regression coefficients are the zero order correlations of the randomly quantified response variable with the randomly quantified predictor variables. The initial values with numerical initialization are obtained from an analysis with numerical scaling level for all variables (thus, from a linear multiple regression analysis).

The CATREG algorithm consists of two steps. In the first step, the quantification of the response variable is estimated, keeping the quantifications of the predictor variables and the regression coefficients fixed,

$$\tilde{\mathbf{v}}_r = \mathbf{D}_r^{-1}\mathbf{G}_r' \sum_{j=1}^{J} \beta_j \mathbf{G}_j \mathbf{v}_j, \tag{3.6}$$

where $\mathbf{D}_r = \mathbf{G}_r'\mathbf{G}_r$, a diagonal matrix with the marginal frequencies of the categories of the response variable. The category quantifications $\tilde{\mathbf{v}}_r$ are the unstandardized quantifications for the nominal scaling level. For nonnominal scaling levels, $\tilde{\mathbf{v}}_r$ is restricted according to the scaling level, yielding $\mathbf{v}_r^*$. Thus, $\mathbf{v}_r^* = \tilde{\mathbf{v}}_r$ for the nominal scaling level, and $\mathbf{v}_r^* = \text{restricted}(\tilde{\mathbf{v}}_r)$ for the nonnominal scaling levels. Then $\mathbf{v}_r^*$ is standardized, resulting in the updated category quantifications:

$$\mathbf{v}_r^+ = N^{1/2}\mathbf{v}_r^*(\mathbf{v}_r^{*\prime}\mathbf{D}_r\mathbf{v}_r^*)^{-1/2}. \tag{3.7}$$

In the second step of the algorithm, the quantification of the response variable is held fixed, and the quantifications of the predictor variables and the regression coefficients are estimated for one variable at a time. This is sometimes called backfitting (Friedman and Stuetzle 1981; Buja et al. 1989), and was applied, a.o., in Kruskal (1965), De Leeuw et al. (1976), Gifi (1990), Breiman and Friedman (1985), Buja et al. (1989), and Hastie and Tibshirani (1990). Some recent developments in the backfitting methodology are described in Härdle and Hall (1993), Opsomer and Ruppert (1997), Mammen, Linton, and Nielsen (1999), Hastie and Tibshirani (2000), and Nielsen and Sperlich (2005). The approach works at follows. First the N-vector of predicted values is computed as

$$\mathbf{z} = \sum_{j=1}^{J} \beta_j \mathbf{G}_j \mathbf{v}_j. \tag{3.8}$$

For updating the quantification of variable j, the contribution of variable j to the prediction (3.8) is subtracted from $\mathbf{z}$,

$$\mathbf{z}_j = \mathbf{z} - \beta_j \mathbf{G}_j \mathbf{v}_j, \tag{3.9}$$

and the category quantifications are computed as

$$\tilde{\mathbf{v}}_j = \text{sign}(\beta_j)\mathbf{D}_j^{-1}\mathbf{G}_j'(\mathbf{G}_r\mathbf{v}_r^+ - \mathbf{z}_j), \tag{3.10}$$

yielding the unrestricted quantifications for the nominal scaling level. So, the transformation for a predictor variable j is fitted to the partial residual vector $(\mathbf{G}_r\mathbf{v}_r^+ - \mathbf{z}_j)$, that is, the category quantifications for a predictor variable are updated from the quantified response variable corrected for the contribution of the other predictor variables. Because of this, the predictor variables are transformed towards independence (an illustration of this property is given in Section 3.6). Next, for variables with nonnominal scaling level, $\tilde{\mathbf{v}}_j$ is restricted according to the scaling level, and normalized as in (3.7), yielding the updated category quantifications $\mathbf{v}_j^+$. Next the regression coefficient β_j is updated:

$$\beta_j^+ \ = N^{-1}\tilde{\mathbf{v}}_j'\mathbf{D}_j \ \mathbf{v}_j^+. \tag{3.11}$$

Then, the updated contribution of variable j to the prediction is added to $\mathbf{z}_j$:

$$\mathbf{z} = \mathbf{z}_j + \beta_j^+\mathbf{G}_j\mathbf{v}_j^+, \tag{3.12}$$

and the algorithm continues with updating the category quantifications and regression coefficient for the next predictor variable, until all predictor variables are updated. Finally, the loss is computed as $N^{-1}\|\mathbf{G}_r\mathbf{v}_r^+ - \mathbf{z}\|^2$. The two steps of the algorithm are repeated until a user-specified convergence criterion is met.

For the restriction according to the ordinal scaling level, weighted monotonic regression (Kruskal 1965; Barlow et al. 1972) of the nominal quantifications on the observed values is applied. To satisfy the restrictions involved in splines, weighted regression of nominal quantifications on an I-spline basis (Ramsay 1988) is applied, with additional nonnegativity restrictions on the spline coefficients when the transformation should be monotonic.

When all or some of the scaling levels are ordinal or involve monotonic splines, local minima may occur. The severeness of this local minimum problem depends on a number of factors (Van der Kooij, Meulman, and Heiser 2006). In general, obtaining a local minimum is not very likely when the fit of the model is reasonable, when the number of categories is not very high, and when there is not much multicollinearity. In Van der Kooij et al. (2006) it was also shown that the global minimum can be obtained by a strategy that involves multiple systematic starts.

3.3. Estimation of expected prediction error

To determine how well a fitted regression model can be generalized to future observations, the expected prediction error needs to be estimated. The apparent prediction error usually does not provide a good estimate for the expected prediction error. For linear regression, the apparent prediction error is the average loss for the observed data,

$$\overline{\text{err}} = N^{-1}\|\mathbf{y} - \sum_{j=1}^{J} \beta_j \ \mathbf{x}_j\|^2, \tag{3.13}$$

that is minimized over the regression weights $\{\beta_j\}$. Using the apparent error as an estimate of the expected prediction error is usually too optimistic because then the expected prediction error is estimated from the same data that were used for fitting the model. To obtain a better estimate, resampling methods such as cross validation or a bootstrap procedure could be used. With cross validation, a part of the data (the training set) is used for fitting the model (giving a set of regression coefficients $\{\beta_j\}$) and next the fitted model is applied to the other part of the data (the test set) to estimate the expected prediction error. With the bootstrap, N observations are drawn randomly from the data with replacement, repeating this process a number of times to obtain B bootstrap samples. Each bootstrap sample serves as a training set to fit a model to, and this fitted model is then applied to the observed data (in this case the test set). The prediction error resulting from applying the fitted model to the test set is averaged over the number of bootstrap samples, resulting in the simple bootstrap estimate of expected prediction error. A more refined approach is to estimate the optimism in the apparent prediction error and then add the optimism to the apparent error. The optimism is estimated as the difference between the simple bootstrap prediction error and the prediction error resulting from applying the fitted model to the bootstrap data itself. According to Efron and Tibshirani (1993), simulation experiments show that cross validation is roughly unbiased but can show large variability, whereas the simple and the refined bootstrap estimates have lower variability but can be severely biased downward, because the training set and the test set have observations in common. However, a modified bootstrap estimation method, called the .632 bootstrap (Efron 1983), corrects for the downward bias and has been shown to perform better than cross validation and the simple and refined bootstrap estimation methods.

3.3.1 The .632 bootstrap with linear regression

The .632 bootstrap (Efron 1983) provides an improved estimate of expected prediction error by improving the estimation of the optimism. Applying the fitted model from each bootstrap sample to the test set (observed data) and computing the prediction error is not done for all observations in the test set (as in the simple bootstrap estimate), but only for those observations that were not in the bootstrap sample. So, for a particular bootstrap sample, the observations that were not drawn for that sample serve as the test set, hence the test set is different for each bootstrap sample. This procedure yields an estimate of expected prediction error that is called the "leave-one-out" bootstrap in Efron and Tibshirani (1997): each observation being predicted using regression coefficients from a particular bootstrap sample, was not in (was "left out" from) that bootstrap sample. The improved estimate of the optimism is a fraction of the difference between the apparent prediction error and the leave-one-out bootstrap estimate of prediction error. The idea underlying the .632 bootstrap estimate is that it is harder to predict the response for an observation using regression coefficients from a model that was fitted to data that do not contain that particular observation (as is also the case in cross-validation). The factor .632 arises because it is approximately the probability for an observation to appear in a bootstrap sample of size N.

So, the .632 bootstrap estimate of expected prediction error is

$$\widehat{\mathrm{Err}}^{(.632)} = \overline{\mathrm{err}} + \widehat{\mathrm{OP}}, \tag{3.14}$$

where the optimism is estimated as

$$\widehat{\mathrm{OP}} = .632(\overline{\mathrm{Err}}^{(1)} - \overline{\mathrm{err}}), \tag{3.15}$$

and $\overline{\mathrm{Err}}^{(1)}$, the leave-one-out bootstrap estimate of prediction error is

$$\overline{\mathrm{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}(y_i - \sum_{j=1}^{J}\beta_j^b x_{ij})^2, \tag{3.16}$$

with C^{-i} being the set of indices of the bootstrap samples b that do not contain observation i, and $|C^{-i}|$ is the number of such samples. (Slightly different definitions are given in Hastie and Tibshirani (1990, p. 298): $\overline{\mathrm{Err}}^{(1)} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|C^{-b}|}\sum_{i\in C^{-b}}(y_i - \sum_{j=1}^{J}\beta_j^b x_{ij})^2$, with C^{-b} being the set of indices of the observations not in bootstrap sample b; and in Efron (1983): $\overline{\mathrm{Err}}^{(1)} = \frac{1}{\sum_{i=1}^{N}|C^{-i}|}\sum_{i=1}^{N}\sum_{b\in C^{-i}}(y_i - \sum_{j=1}^{J}\beta_j^b x_{ij})^2)$. In Efron and Tibshirani (1997) it

is noted that this last definition agrees with the definition of Efron and Tibshirani (1993) given in (3.16) as $B \rightarrow \infty$ and that they produced nearly the same results in simulations.)

3.3.2 The .632 bootstrap with CATREG

With CATREG estimates of both the regression weights $\{\beta_j\}$ *and* of the category quantifications $\mathbf{v}_r$ and $\{\mathbf{v}_j\}$ are obtained. Thus, applying the model fitted to a bootstrap sample to the observations in the test set involves applying both regression weights and substituting categories values with the applicable category quantifications. The CATREG apparent prediction error is written as

$$\overline{\text{err}} = N^{-1} \| \mathbf{G}_r \mathbf{v}_r - \sum_{j=1}^{J} \beta_j \mathbf{G}_j \mathbf{v}_j \|^2. \tag{3.17}$$

In computing the CATREG leave-one-out bootstrap estimate of prediction error, a complication arises for variables with nominal or ordinal scaling level when a category does not occur in a bootstrap sample: then an estimate of the quantification for that category is not obtained. When numerical or spline transformations have been used, this is not a problem, because the quantifications for the non-occurring category can easily be obtained from the transformation function by *interpolation* (which for splines is determined by the spline coefficients). For variables with nominal or ordinal scaling level, however, this is not possible, since their transformation functions are not parametric. In that case, estimating the quantification of a non-occurring category might introduce substantive error, especially since categories that do not appear in a particular bootstrap sample are likely to be categories with a low marginal frequency in the observed data. So, if an observation i has a category on a variable with nominal or ordinal scaling level that does not occur in a particular bootstrap sample, that bootstrap sample is excluded from the estimation of the expected prediction error for observation i. There also is a complication for variables with spline transformations because the latter sometimes require *extrapolation* to estimate the quantification of a non-occurring category. Since spline transformations using extrapolation are likely to be unstable, bootstrap samples that require extrapolation for an observation i are also not included in the estimation of the expected prediction error for observation i. The index numbers of the remaining bootstrap samples are collected in the index set C^{-i}. Taking all this into account, the CATREG leave-one-out bootstrap estimate of prediction error is written as

$$\overline{\text{Err}}^{(1)} = \frac{1}{N^{(1)}} \sum_{i=1}^{N} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} (\mathbf{G}_r \mathbf{v}_r^b - \sum_{j=1}^{J} \beta_j^b \mathbf{G}_j \mathbf{v}_j^b)^2 \text{ for} |C^{-i}| \neq 0, \quad (3.18)$$

where C^{-i} is the set of indices of the bootstrap samples b that
(a) do not contain observation i,
(b) do contain the categories that apply to observation i for variables with nominal or ordinal transformations,
(c) do not require extrapolation for observation i for variables with spline transformations.
$N^{(1)}$ is the number of observations for which $|C^{-i}|$ is not zero. $|C^{-i}|$ may become zero if, for example, observation i has one of the extreme categories on a variable with a spline transformation, and this category has a frequency of one. Then each bootstrap sample that does not contain this observation, also does not contain the extreme category; thus, for observation i all bootstrap samples are excluded.

In this chapter we used the definition of $\overline{\text{Err}}^{(1)}$ of Efron and Tibshirani (1993) given in (3.16), computed the Monte Carlo standard error (MCSE) as $[\frac{1}{N^2} \sum_i (\overline{\text{Err}}_i^{(1)} - \overline{\text{Err}}^{(1)})^2]^{1/2}$, and used 200 bootstrap samples. But, for the comparison of the CATREG results to the results of the methods reported in Hastie and Tibshirani (1990, p. 299), we used the definition of $\overline{\text{Err}}^{(1)}$ of Hastie and Tibshirani (1990, p. 298), given in Section 3.3.1, and hence computed MCSE as $[\frac{1}{B^2} \sum_b (\overline{\text{Err}}_b^{(1)} - \overline{\text{Err}}^{(1)})^2]^{1/2}$, and used 50 bootstrap samples (with the other definitions of $\overline{\text{Err}}^{(1)}$ and/or a different number of bootstrap samples, the results given in Hastie and Tibshirani (1990, p. 299) could not be reproduced).

3.4. Performance of CATREG and six other nonlinear regression methods: Prediction accuracy in the analysis of the Ozone data

The ozone data have been frequently analyzed in literature to illustrate nonlinear transformations (a.o. Breiman and Friedman (1985), Hastie and Tibshirani (1990), and Lin and Zhang (2003)). The data set is available by ftp from the Department of Statistics, University of California at Berkeley (ftp.stat.berkeley.edu/pub/users/breiman). The data consist of 330 daily measurements of atmospheric ozone concentration in the Los Angeles basin in 1976 and eight daily meteorological measurements. The response variable

is the log of the daily maximum of the hourly-average ozone concentrations (linearly transformed to integer values when using CATREG). The predictors are the eight meteorological measurements and day of the year (Table 3.1).

3.4.1 Comparison between CATREG and Box-Tidwell Method, Full Additive Model, Stepwise Full Additive Model, TURBO, and BRUTO

In Hastie and Tibshirani (1990, sec. 10.3) the ozone data are used for a comparative assessment of five prediction methods in terms of their prediction accuracy. The five methods that were compared are:
(i) Linear regression after transformations using the Box and Tidwell method as applied by Hawkins (1989). Power transformations were used for all predictor variables except dpg (quadratic transformation) and doy (sine and cosine bases at three frequencies);
(ii) Full additive model using smoothing splines with four degrees of freedom.
(iii) A backward stepwise strategy applied to a full additive model. The backward stepwise strategy selects the number of degrees of freedom for each variable from the regimen of $df = 0, 1, 4$ (with $df = 0$ the predictor is excluded; with $df = 1$ the predictor is included linearly; with $df = 4$ the predictor is transformed using smoothing splines with four df).
(iv) Additive regression splines (TURBO (Friedman and Silverman 1989); this is an adaptive knot selection (number and location) strategy, using piecewise-linear basis functions for locating the knots. Once the knot positions are determined, the piecewise-linear functions are converted into piecewise-cubic functions).
(v) Automatic backfitting using smoothing splines (BRUTO Hastie (1989) and Hastie and Tibshirani (1990); this method combines backfitting and adaptive smoothing).

With CATREG we performed 50 bootstraps, using the numerical scaling level for the response variable and nonmonotonic splines (based on second degree polynomials, with two interior knots) for the predictor variables. So, like Hastie and Tibshirani (1990, sec. 10.3), we have four degrees of freedom for each predictor. In the Hastie and Tibshirani (1990, sec. 10.3) study, the estimated prediction error for the raw data is reported. Using the CATREG algorithm, standardized quantifications and regression coefficients (β's) are obtained. Thus, the apparent error and the estimated prediction error obtained with CATREG are for standardized transformed data. Because the response variable has been treated as a numeric variable (standardized), the CATREG standardized error can be converted to the error for the raw data

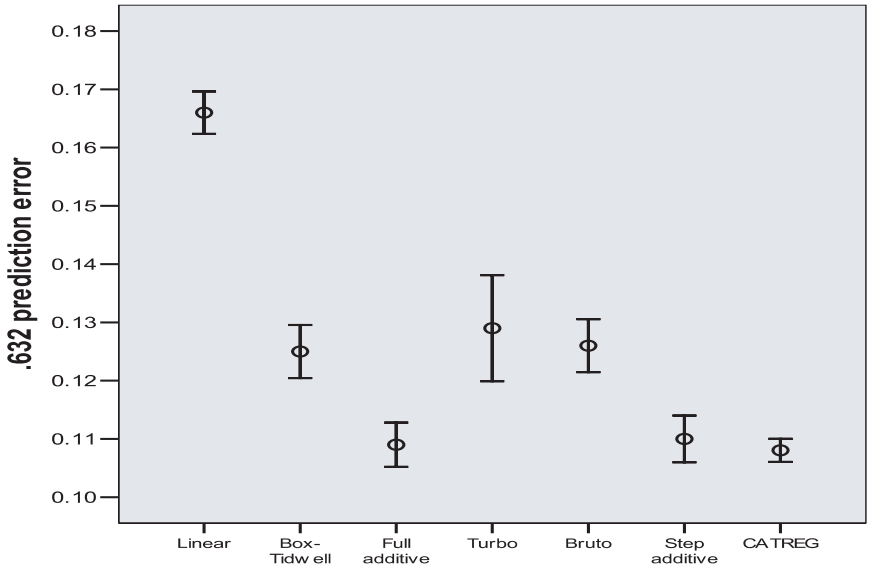

Figure 3.1. .632 *Bootstrap estimates of prediction error. The bars around the estimates represent the estimated Monte Carlo standard error.*

by multiplying the CATREG standardized error with the variance of the raw response variable: the apparent error $\overline{\text{err}}$ (3.17) is multiplied with the variance of $\mathbf{y}$, and in the leave-one-out bootstrap estimate of prediction error (3.18) the squared error term is multiplied with the variance of $\mathbf{y}^b$.

The .632 bootstrap prediction error for linear regression, for the five nonlinear methods reported in Hastie and Tibshirani (1990, p. 299), and for CATREG are displayed in Figure 3.1. The CATREG .632 bootstrap estimate of prediction error is .108, which is about the same as the estimate resulting from the full additive model (.109), but the MCSE with CATREG (.002) is lower than with the full additive model (.008).

Figure 3.2 displays the CATREG nonmonotonic spline transformations for the predictor variables. CATREG usually plots category quantifications (on the vertical axis) versus categories (on the horizontal axis), but here the quantifications were multiplied with the standardized regression coefficients, so that the vertical range gives an indication of the importance of a predictor, and the direction of the transformation reflects the direction of the relation between the predictor and the response. Comparing the CATREG transformation plots to the transformation plots for the five methods applied in Hastie and Tibshirani (1990, sec. 10.3), we notice that also with respect to

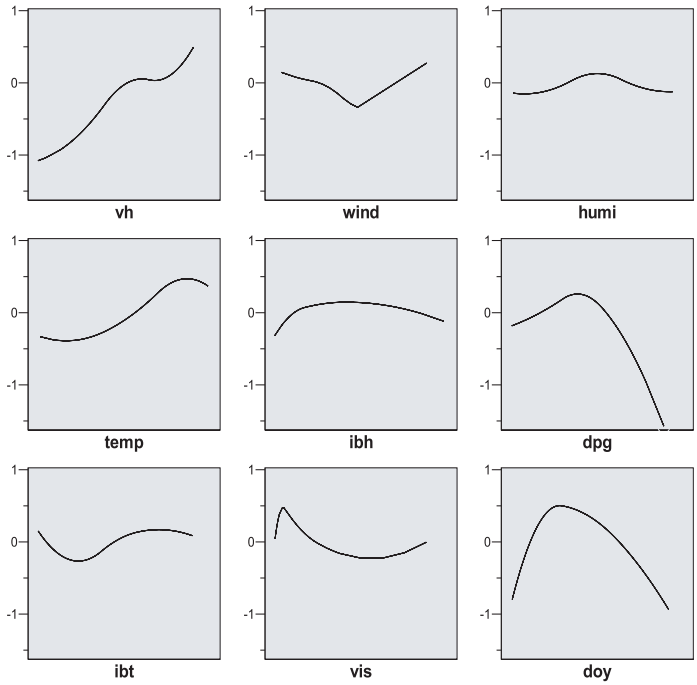

Figure 3.2. CATREG nonmonotonic spline transformations for the ozone data. The y-axis represent the quantifications multiplied with the regression coefficient for the predictor.

the form of the transformations, CATREG is very close to the full additive model (op.cit. p. 297).

3.4.2 Comparison between CATREG and ACE

As mentioned before, CATREG is very closely related to ACE (Breiman and Friedman 1985). The main difference is in the method to obtain smooth transformations. For smooth nonlinear transformations (nonmonotonic and monotonic) ACE uses the supersmoother (a variable span local linear smoother, (Friedman 1984)), while CATREG uses spline transformations. Like CATREG, ACE also provides numerical and nominal transformations; unlike CATREG, ACE does not provide ordinal transformations (step functions) and provides transformations for circular (periodic) variables. We used the ACE function of S-plus (MathSoft, Inc. 1999), which gives the transformed variables as output, but unfortunately not the transformation parameters, so we were not able to apply interpolation if this was necessary in computing $\overline{\text{Err}}^{(1)}$. Therefore, we

Table 3.1. Original and recoded number of categories predictor variables ozone data (ibh was recoded by rounding the original values divided by 100, dpg by rounding the original values divided by 5, ibt by rounding the original values divided by 10, and doy was recoded to months).

	No. of categories	
Predictor	Original	Recoded
500 millibar pressure height (vh)	53	53
Wind speed (wind)	12	12
Humidity (humi)	65	65
Temperature (temp)	63	63
Inversion base height (ibh)	196	42
Pressure gradient (dpg)	128	33
Inversion base temperature (ibt)	193	37
Visibility (vis)	24	24
Day of the year (doy)	330	12

computed $\overline{\text{Err}}^{(1)}$ for ACE as in (3.18) but without β_j^b, because in the ACE program β_j^b is absorbed in $\varphi_j^b(x_{ij})$ (thus, the bootstrap model is applied to the test set by substituting the categories in the test set with the applicable transformed variable values, and a bootstrap sample is excluded in predicting observation i in the test set if observation i has a category that did not appear in the bootstrap sample). With the original ozone data, $|C^{-i}| = 0$ is obtained for all i, because for all observations inter- and/or extrapolation was needed for at least one predictor. This is due to the fact that some of the variables have a lot of unique or very low frequency categories. For such categories the probability to be left out from a bootstrap sample is very high. For this reason some of the predictor variables were recoded so that the resulting variables had a smaller number of categories. The number of categories of the original data and of the recoded data are given in Table 3.1.

The ACE transformations are displayed in Figure 3.3. The transformations from ACE are somewhat less smooth than the nonmonotonic spline transformations resulting from CATREG (Figure 3.2; the nonmonotonic spline transformations for the recoded data are not displayed because they are almost the same as the transformations for the original data in Figure 3.2). Comparing the predictive accuracy (Table 3.2), we see that with ACE the apparent prediction error is somewhat lower than with CATREG and the expected prediction error is somewhat higher. On the other hand, the supersmoother transformations are much smoother than the CATREG nominal transforma-

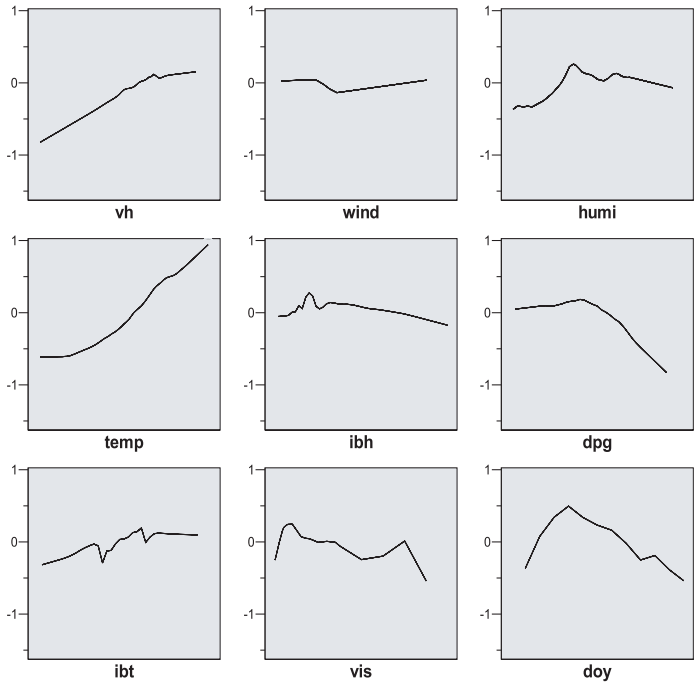

Figure 3.3. ACE supersmoother transformations for the recoded ozone data.

tions (Figure 3.4) and they have considerably less estimated prediction error. So, we conclude that the supersmoother gives transformations somewhere in between nominal and nonmonotonic spline transformations both in terms of smoothness and expected error rate.

Table 3.2. Prediction error for the recoded ozone data ($N = 330$, 200 bootstraps).

Method	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
ACE supersmoother	284	.084	.129	53.7	.013
CATREG Nonmonotonic spline	325	.090	.114	27.2	.009
CATREG Nominal	284	.004	.358	8842.5	.033

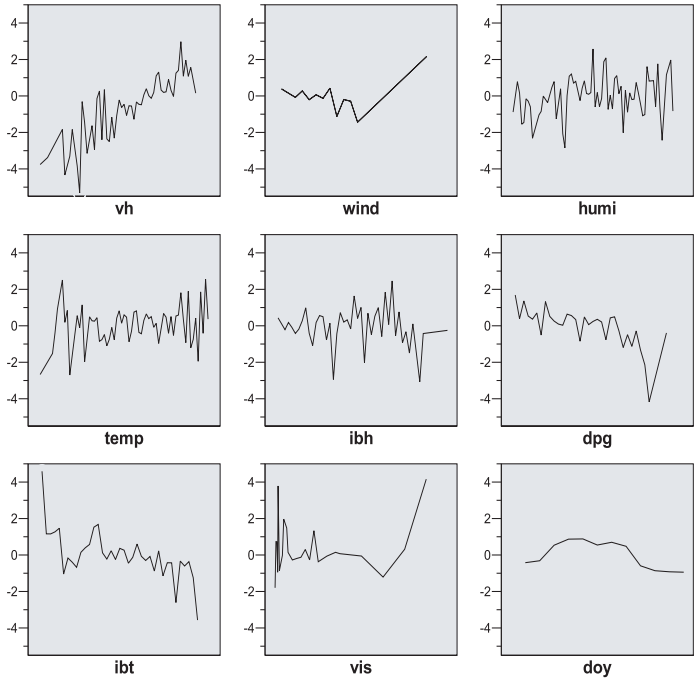

Figure 3.4. CATREG nominal transformations for the recoded ozone data. The y-axis represent the quantifications multiplied with the regression coefficient for the predictor.

3.5. Prediction accuracy for different scaling levels in CATREG

As explained in Section 3.2.1, the scaling levels differ with respect to the degree of restrictiveness in the quantification process, which is related to the number of degrees of freedom of the transformations. More restricted transformations result in higher apparent error rate by definition, but they can give better results in terms of expected error rate. To compare the prediction accuracy of the scaling levels, the recoded ozone data as described in Section 3.4.2 were used. Although the variable "Day of the year" shows a transformation that clearly reflects a seasonal trend and should therefore in practice not be restricted to a monotonic transformation, a monotonic transformation was included here for reason of comparison. In Figures 3.4, 3.5, and 3.6, nominal, ordinal, and monotonic spline transformations are displayed. (Again, the nonmonotonic spline transformations for the recoded data are not displayed

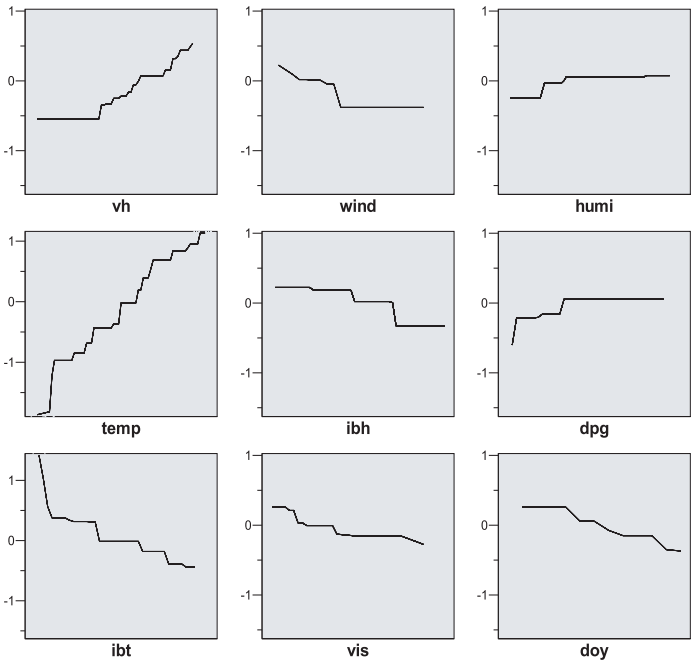

Figure 3.5. CATREG ordinal transformations for the recoded ozone data. The y-axis represent the quantifications multiplied with the regression coefficient for the predictor.

Table 3.3. CATREG prediction error for the recoded ozone data (splines, based on second degree polynomials, with two interior knots, $N = 330$, 200 bootstraps, numeric initialization, no multiple systematic starts).

Scaling level	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
Nominal	284	.004	.358	8842.5	.033
Nonmonotonic spline	325	.090	.114	27.2	.009
Ordinal	284	.106	.162	53.1	.017
Monotonic spline	325	.137	.155	13.2	.013

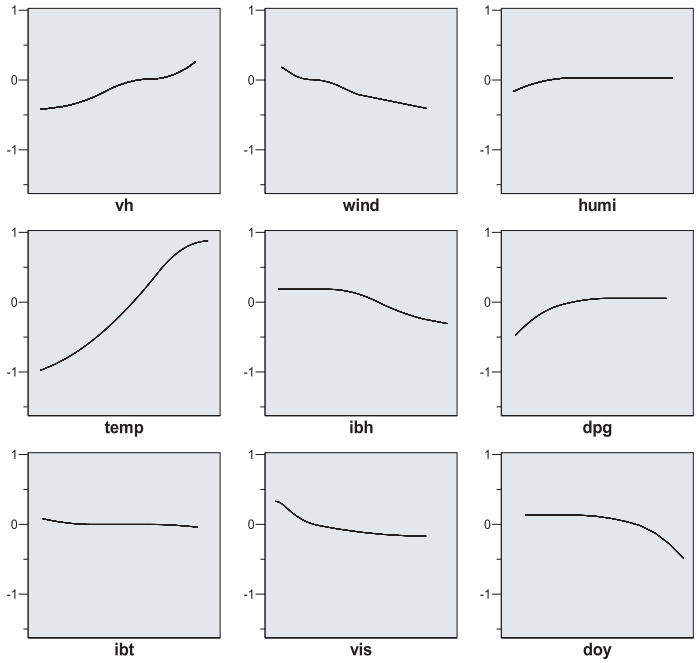

Figure 3.6. CATREG monotonic spline transformations for the recoded ozone data. The y-axis represent the quantifications multiplied with the regression coefficient for the predictor.

because they are very close to the transformations for the original data displayed in Figure 3.2.)

In Table 3.3, the prediction accuracy results for the different scaling levels are given, again obtained from 200 bootstrap samples, using numerical initialization for all scaling levels (thus, without applying multiple systematic starts for the ordinal and monotonic spline transformations; the results when applying multiple systematic starts will be discussed below).

The nominal scaling resulted in the lowest apparent error, but in the highest prediction error. This is as expected, because the nominal scaling level is the least restrictive level and there are many categories; the number of degrees of freedom for a variable with the nominal scaling level is the number of categories minus one. So, with nominal scaling level, the freedom in quantifying the categories resulted in an almost perfect fit for the training set, but this is obtained at the cost of much less accuracy for future predictions; this situation clearly indicates overfitting.

Comparing the spline transformations to the nominal and ordinal step functions, we see that for the spline transformations the apparent prediction error is higher, but the expected error rate is lower. The same pattern is observed when the more restrictive ordinal transformations are compared to the nominal transformations. However, comparing the monotonic spline transformations to the nonmonotonic spline transformations, we see that both the apparent error and the expected error rate are higher for the more restrictive monotonic spline transformations.

Summarizing, the apparent error rate for the ordinal and spline transformations is higher than for the nominal transformations, but their expected error rate is considerably lower. Also, the apparent error rate for the monotonic spline transformation is higher than for the ordinal transformation, but the expected error rate is lower. This phenomenon is well-known in the machine learning literature, where suboptimal results for the observed data are accepted to obtain superior results for predicting future observations.

A more restrictive scaling level does not necessarily result in more prediction accuracy, however. When there are important nonmonotonic relationships between the predictors and the response, monotonic transformation is too restrictive, as the ozone data illustrate: the less restrictive nonmonotonic spline transformations perform better than the monotonic spline transformations. However, when there are only slight nonmonotonicities in the data, monotonic transformation might result in lower expected prediction error than nonmonotonic transformation. Similarly, in the case of only slight nonmonotonicities nominal transformation may perform better than ordinal transformation when the number of categories is low.

Applying multiple systematic starts to obtain the global optimal solution with ordinal and monotonic spline transformations (see Table 3.4) resulted in both lower apparent error, and lower expected prediction error for the monotonic spline transformations, but in higher expected error rate for the ordinal step function. This result suggests that when ordinal transformations are appropriate (for example, when variables have only a limited number of ordered categories), using multiple systematic starts is needed only if the primary objective of the analysis is to explain the data at hand. On the other hand, we would not need to perform multiple systematic starts for ordinal transformations if the aim of the analysis is to predict future responses. However, the differences are rather small in this example because the response variable can be rather well predicted (with the linear model for the unrecoded data R^2 is .717), and as was shown in Van der Kooij et al. (2006), local minima become less frequent and less severe with increasing R^2.

Table 3.4. CATREG prediction error for the recoded ozone data (splines based on second degree polynomials with two interior knots, $N = 330$, 200 bootstraps, multiple systematic starts).

Scaling level	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
Ordinal	284	.096	.171	78.7	.016
Monotonic spline	325	.128	.148	15.9	.012

Table 3.5. CATREG prediction error for the recoded ozone data (splines based on second degree polynomials with two interior knots, $N = 330$, 200 bootstraps, random initialization, no multiple systematic starts).

Scaling level	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
Nominal	284	.004	.460	11126.8	.040
Nonmonotonic spline	325	.090	.114	27.2	.009
Ordinal	284	.190	.205	8.0	.018
Monotonic spline	325	.148	.159	7.9	.013

When CATREG is used to maximize R^2 for the data at hand, it is usually recommended to use random initialization for nominal and nonmonotonic spline transformations, and numerical initialization for ordinal and monotonic spline transformations. The comparison of the results obtained with numerical initialization to those obtained with random initialization (see Table 3.5), shows that there is hardly a difference in the apparent error rate for the nominal scaling level, but that the expected prediction error is considerably higher when random initialization was used. For the nonmonotonic spline transformations, there is no difference, neither in the apparent prediction error nor in the expected prediction error. For the ordinal and monotonic spline transformations, numerical initialization gives the best results, both for apparent and for expected error rates.

The high expected prediction error compared to the apparent error for the nominal scaling level is likely due to the relatively high number of categories (thus, a high number of degrees of freedom) for most of the predictors. To confirm this, the predictor variables were recoded to decrease the number of categories in three steps and the effect on the error rates for the nominal scaling level was examined. The results are given in Table 3.6. For the 25-categories-analysis, only the predictor variables with more than 25 categories, were recoded, for the 12-categories-analysis only the predictor variables with

Table 3.6. CATREG prediction error for the ozone data with nominal scaling level (predictors recoded to 25, 12, and 7 categories, $N = 330$, 200 bootstraps).

	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
numerical initialization					
25 categories	320	.050	.285	465.6	.021
12 categories	328	.083	.161	95.3	.013
7 categories	330	.102	.145	42.1	.012
random initialization					
25 categories	320	.050	.286	467.6	.022
12 categories	328	.083	.162	96.4	.013
7 categories	330	.102	.145	42.1	.012

more than 12 categories, and finally, for the 7-categories-analysis all predictors were recoded. As could be expected, reducing the number of categories increases the apparent prediction error. However, the expected prediction error and the Monte Carlo standard error were substantially reduced going from 25 via 12 to 7 categories. The fact that the original continuous response variable is best predicted from categorical predictor variables with a (severely) reduced number of categories is encouraging for the data analytical approach taken in CATREG. With a limited number of categories, the difference in expected prediction error between the results obtained with numerical initialization and random initialization almost disappeared.

3.6. The effect of the number of observations on the expected prediction error

The high expected prediction error for the nominal scaling level relative to the apparent error might well be due to the rather small number of observations in the ozone data compared to the number of categories. To investigate the effect of the number of observations on the expected prediction error, data were used that come from 9409 questionnaires containing 502 questions filled out by shopping mall customers in the San Francisco Bay area (Impact Resources, Inc., Columbus, OH, 1987). An extract from this survey consisting of 14 demographic attributes was used, listed in Table 3.7; the extract of the data can be obtained from http://www-stat.stanford.edu/~tibs/ElemStatLearn/. To avoid the introduction of additional error by imputing missing data, only the 6876 observations without missing values were used. The response variable is the annual household income.

Table 3.7. Number of categories demographic data and standardized regression coefficients when using nominal scaling level for all predictors.

	No. of categories	β
Annual household income (response)	9	
Sex	2	-.063
Marital status	5	-.176
Occupation	9	-.244
Dual incomes	3	.011
Householder status	3	-.127
Type of home	5	-.080
Ethnic classification	8	.030
Language in home	3	-.038
Age	7	.278
Education	6	.119
Years lived in Bay Area	5	.029
Number of persons in household	9	.038
Number of persons <18 in household	9	-.027

With ordinal transformation of the response and nominal transformations of the predictors, an R^2 of .520 is obtained. The transformations are displayed in Figure 3.7. The variables "Age", "Education", "Years lived in Bay Area", "Number of persons in household", and "Number of persons <18 in household" are of ordinal measurement level; analyzing these variables at the ordinal scaling level results in only a slightly lower R^2 of .518. This small decrease in R^2 is explained by the fact that the ordinal variables "Age" and "Education" are important predictors and their transformations using the nominal scaling level are only slightly nonmonotonic ("Age") or monotonic ("Education"). So, using the ordinal scaling level for these predictors does not require much restriction, and consequently R^2 does not decrease much. The transformations of the variables "Number of persons in household", and "Number of persons <18 in household" require more restriction when applying the ordinal scaling level, but these predictors are not very important, thus restricting their transformations to be monotonic will not affect R^2 much either. Although the variables "Number of persons in household" and "number of persons <18 in household" are not very important, their transformations nicely illustrate the backfitting property of transformation towards independence mentioned in Section 3.2.2: The observed variables have a high correlation (.71), while the correlation of the transformed variables is almost zero (.01).

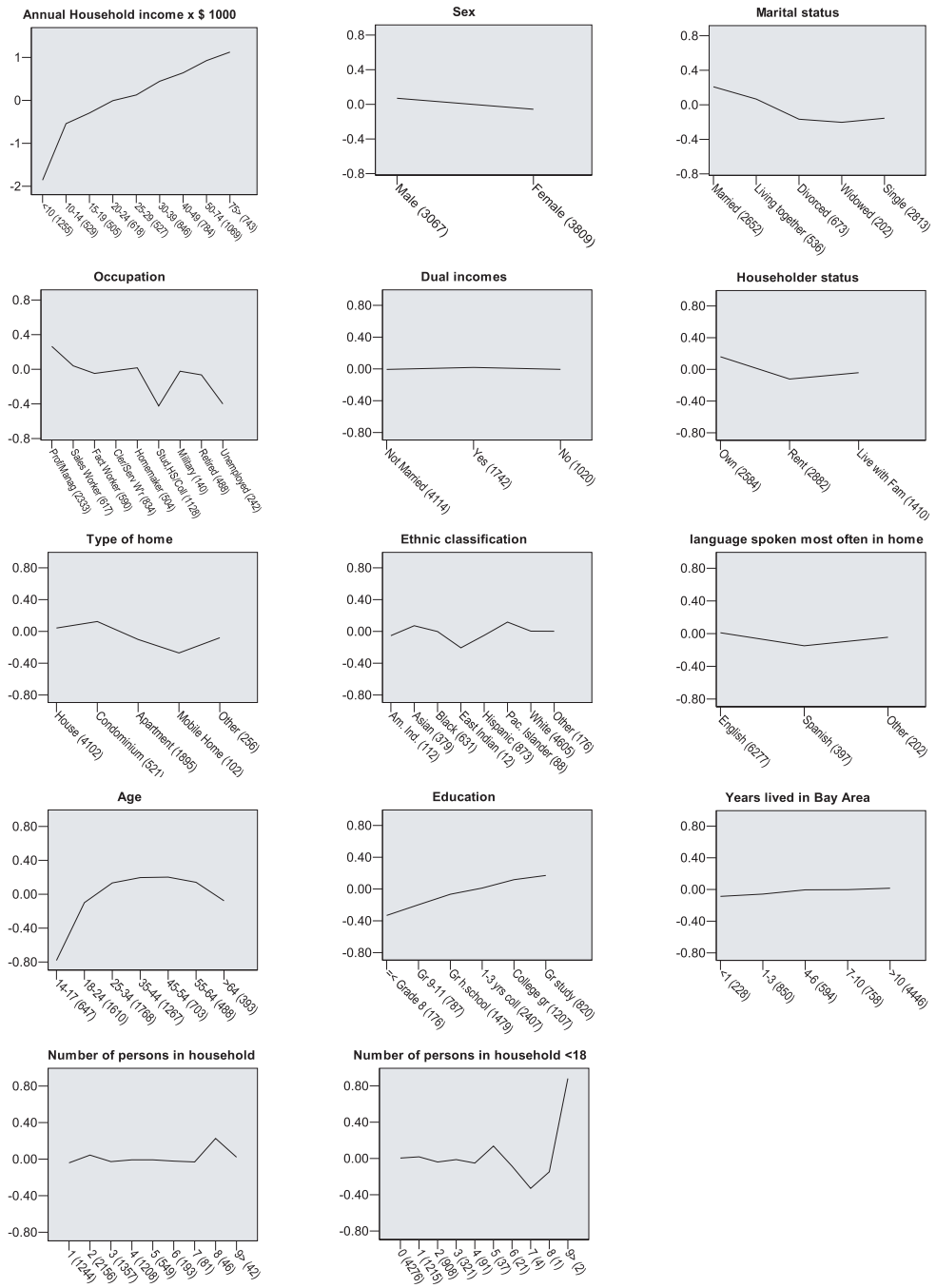

Figure 3.7. CATREG transformations for the demographic data (response ordinal, predictors nominal). The y-axis represent the quantifications multiplied with the regression coefficient for the predictor. The figures in brackets following the category labels are the category frequencies.

Table 3.8. CATREG prediction error for the demographic data (splines based on second degree polynomials with one interior knot, numerical initialization, 200 bootstraps). The reported results are for standardized variables.

Scaling level	$N^{(1)}$	$\overline{\text{err}}$	$\widehat{\text{Err}}^{(.632)}$	Increase	MCSE
N=6876					
Nominal	6875	.480	.522	8.8	.011
Nonmonotonic spline	6876	.492	.520	5.8	.011
N=3300					
Nominal	3298	.475	.535	12.7	.016
Nonmonotonic spline	3299	.490	.528	7.8	.015
N=1000					
Nominal	1000	.473	.602	27.2	.034
Nonmonotonic spline	1000	.494	.578	17.1	.032
N=330					
Nominal	326	.435	.740	70.0	.066
Nonmonotonic spline	328	.476	.631	32.6	.056

Three random subsamples of specified size from the data were drawn to obtain data sets with reduced number of observations, with sample sizes of 3300, 1000, and 330. The results are given in Table 3.8. As expected, the apparent error does not change much with different number of observations, but the expected prediction error is affected by the number of observations, for both the nominal scaling level and the more restrictive nonmonotonic splines. There is not much increase in the expected prediction error going from $N = 6876$ to $N = 3300$, but reducing N to 1000 the expected prediction error considerably increases, and even more so when N is reduced to 330. Also, with a higher number of observations (6876 and 3300), there is not much difference in the expected prediction error for the nominal and nonmonotonic spline transformations, while with a smaller number of observations (1000 and 330) the difference is considerable.

3.7. Conclusions

In this chapter the CATREG approach to nonlinear multiple regression was described, which involves regression with optimal scaling according to a variety of scaling levels. The procedure fits nominal, nonmonotonic spline, ordinal, and monotonic spline transformations of the predictors and/or the response

that maximize the multiple correlation coefficient and minimize the apparent error rate. The CATREG procedure was studied in terms of its prediction accuracy by estimating the expected error rate, using the .632 bootstrap proposed by Efron (1983), applying the definition of the leave-one-out bootstrap estimate of prediction error as defined in Efron and Tibshirani (1993).

First, the prediction accuracy of CATREG was compared with the results of five other methods for nonlinear prediction methods as reported in Hastie and Tibshirani (1990, sec. 10.3). For the particular data set that was used (the ozone data), the results show that in terms of expected error rate, CATREG with nonmonotonic spline transformations as (in fact, a little better than) the full additive model with smoothing spline transformations, and the latter was the best performing method in the comparison by Hastie and Tibshirani. In addition, the Monte Carlo standard error for CATREG was much smaller. Next, CATREG was compared to the very closely related ACE procedure by Breiman and Friedman (1985). It turned out that in terms of prediction accuracy, the results of ACE with nonmonotonic transformations using a supersmoother were somewhat less than the results obtained by CATREG with nonmonotonic spline transformations, but much better than CATREG with (non-smooth) nominal transformations (which are step functions). In general, nominal transformations may result in a close to perfect apparent prediction error rate, but in a far from perfect expected error rate when the number of observations is small and the number of categories is large (so that the marginal frequencies are small). By reducing the original number of categories, or binning continuous predictors in a small number of categories, much better prediction accuracy may be obtained, as was shown in the analysis of the ozone data. Both the prediction accuracy increased and the Monte Carlo standard error decreased when the original response variable measuring ozone concentrations was predicted from categorical data with a limited number of categories derived from the original meteorological measurements with number of categories ranging from 12 to 65. This is a very positive result for the categorical approach to multiple regression.

Because CATREG also includes options for ordinal and monotonic spline transformations, their prediction accuracy was studied as well. CATREG solutions that turned out to be local minima (which may occur with ordinal and monotonic spline transformations (Van der Kooij et al. 2006)), thus with a higher apparent error (by definition), appeared to have lower expected error rates compared to the global minimum solutions. These results put the local minimum problem in a new perspective. The ordinal and monotonic spline transformations were also compared to the results for nominal and nonmonotonic spline transformations. In general, we conjecture that more

restricted models lead to higher apparent error rates (again, by definition), but to lower expected error rates. There is a major exception to this latter rule, however, and this is when relationships between predictors and the response variable are severely nonlinear and nonmonotonic. In the latter case, using ordinal and/or monotonic spline transformations will certainly reduce the prediction accuracy.

3.8. Computational note

For completeness, it should be remarked that the indicator matrices introduced in the algorithm section (3.2.2), are only used in the equations, and of course not in the actual implementation of the algorithm. The CATREG algorithm has been implemented in a user-friendly computer program in the CATEGORIES package in SPSS (Meulman et al. 1999). For non-integer data, the CATREG program (SPSS 8.0 onwards (SPSS Inc. 1998; Van der Kooij and Meulman 2004) offers an option for linearly transforming the data into integer values, and various discretization options for recoding continuous data (binning).