Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Chapter 1

Introduction

Linear multiple regression has since long been a major data analytic tool in many scientific fields (for instance, behavioral sciences, social sciences, biological sciences, medical sciences, psychometrics, and econometrics) for predicting a response variable from a linear combination of predictor variables. The linear regression model arises from assuming a linear relation between a response variable and a set of predictor variables. In a graphical representation, the plot of the response variable against the linear combination of the predictors is assumed to show a linear trend. Often, however, the response-predictors relation is not linear, for example, the plot of the relation between the number of car accidents a driver has caused and the age of the driver shows a u-shaped trend: younger and older drivers cause more car accidents than middle-aged drivers. When a linear regression model is applied to such data, the conclusion would wrongly be that there is no relation between the response and the predictor, because linear regression is restricted to reveal only relations showing a linear trend. Because of this restriction, regression methods that can deal with nonlinear relations have become more and more popular in the last few decades. Estimation of regression models for nonlinear relations is (sometimes much) more computationally complex and intensive than for the linear regression model. So, the rising popularity of modern regression techniques is also due to the fast increase in availability of efficient computational equipment.

1.1. Regression methods for nonlinearly related data

Three major approaches to regression incorporating nonlinear relations can be distinguished: *Nonlinear Regression*, *Generalized Linear Models*, and *Regression with Transformation*. In Nonlinear Regression, nonlinear relations are described with a nonlinear model, i.e., a model that is not a linear function of the parameters (the derivatives of the model with respect to the model parameters depends on one or more parameters). The other two approaches use linear models and describe nonlinear relations by transformations that aim at linearizing the relation between the response and the predictors. Another distinction between Nonlinear Regression and the other two approaches concerns the measurement level of the variables and the distribution of the error. Nonlinear regression assumes continuous variables and normally distributed error, while in the other approaches the assumptions are weaker.

Nonlinear Regression (Draper and Smith 1966; Seber and Wild 1989) models the response as a nonlinear parametric function of the predictors. Nonlinear Regression models are truly nonlinear because no linearization of the relation between the response and the predictors is involved. In using Nonlinear Regression, a model has to be choosen, based on theoretical considerations, that has to be expressed as a mathematical function. Nonlinear Regression is particulary suited for data describing a state or process, as frequently encountered in physics, chemistry or biological science. The parameters of a nonlinear model can usually directly be interpreted in terms of the process described by the model. For example, the exponential growth model has one parameter that is a rate constant. Nonlinear regression models are also used when the parameters do not have an interpretable meaning, but just serve to define a flexible function (in neural networks, for example).

Generalized Linear Models (GLM, McCullagh and Nelder 1989; Nelder and Wedderburn 1972) model the response as a nonlinear function of a linear combination of the predictors. Thus, these models are linear in the parameters, and the nonlinear relation between the response and the predictors is linearized through a nonlinear function, that is, the relation between the response and the nonlinearly transformed linear combination of the predictors is linear. The nonlinear function, that links the response to the predictors, is called the link function, and is one extension of the linear model. The second extension is that the response variable need not be continuous and normally distributed, but can have a distribution function belonging to an exponential family other than the normal. For example, linear multiple regression can be regarded as a generalized linear model with a continuous and normally dis-

tributed response and the identity function as the link. Another example is logistic regression, which is a generalized linear model for a binary response with the logit function as the link. Also, feed foreward Neural Networks (Ripley 1996) and Projection Pursuit Regression (Friedman and Stuetzle 1981), which is equivalent to a particular form of feed foreward Neural Networks, are closely related to GLM.

In the *Regression with Transformations* approach the predictors and/or the response variable themselves are nonlinearly transformed and no distributional assumptions are made. So, here the relation between the response and the predictors is linearized through separate nonlinear transformations of the variables, allowing for flexible modeling of nonlinar relations. The transformation function can be any parametric or non-parametric function. The best known transformation model using parametric transformations is the Box-Cox model (Box and Cox 1964). Examples of methods using non-parametric transformations are MONANOVA (Kruskal 1965), ADDALS (De Leeuw, Young, and Takane 1976), MORALS (Young, De Leeuw, and Takane 1976) (implemented in TRANSREG (SAS/STAT 1989)), and Monotone transformations to additivity using splines (Winsberg and Ramsay 1980; Ramsay 1988). Hastie and Tibshirani (1990) have developed *Generalized Additive Models* (GAM), generalizing additive models (Stone 1985) in the same way as GLM generelizes linear models. In an additive model the response is modeled as a linear combination of separately transformed predictors using smoothers. Thus, GAM generalizes the additive model by incorporation of a link function, linking the response to the sum of the smoothly transformed predictors, and by allowing the response to have any distribution in the exponential family. A different generalization of the additive model is the inclusion of a smooth transformation of the response (ACE, Breiman and Friedman (1985)). In the Statistical Learning literature Support Vector Machines (SVMs, Vapnik (1996) and Hastie, Tibshirani, and Friedman (2001)) are developed for predictor transformations. A SVM extends the set of predictor variables with an extremely large number of transformations of the predictors and then "picks out", using a very clever "trick", the transformations most suited for the prediction.

1.2. The CATREG method

The method that is the subject of study in this monograph, CATREG, takes the Regression with Transformation approach, applying the optimal scaling methodology as developed in the Gifi system (Gifi 1990) to transform both the response and the predictor variables. The CATREG model is a special case of the CANALS model (Van der Burg and De Leeuw 1983) for non-

linear canonical correlation analysis. CATREG was developed as a method for linear regression for categorical variables. Categorical variables can be unordered (nominal, for example religion or marital status) or ordered (ordinal, for example preferences, judgements, or Likert scales). The categories of nominal variables have labels and ordinal categories have rank numbers or ordered labels, such as low, medium, high, or never, sometimes, often, always, neither of which can be regarded as numeric values. In the following both labels and rank numbers will be referred to as category values. Optimal scaling is a method to find optimal numeric values to replace category values, thus transforming categorical data to numeric data. In the optimal scaling terminology this transformation process is called "quantification" (the process is called "quantifying qualitative data" in Young 1981). The transformations (quantifications) of categorical variables are estimated simultaneously with the estimation of the regression coefficients, using an alternating least squares procedure that maximizes the squared multiple regression coefficient, R^2, for linear regression on the transformed variables. As a result of this optimization criterion, the optimal scaling transformations linearize the relation between the response and the predictors (as will be illustrated in the next section). Thus, the CATREG method results in transformed categorical variables that have values with numeric properties and that are optimal for describing the relation between the response and the predictors. Quantification of categorical variables usually results in nonlinear transformations, that can be nonmonotonic or, by applying restrictions, monotonic or linear. Such restrictions are specified by choosing an optimal scaling level (described in the next section).

In the optimal scaling methodology, numeric variables are treated as categorical variables, with the number of categories equal to the number of distinct values of the variable (thus, values of numeric variables will also be referred to as category values). Choosing the numeric scaling level for a numeric variable results a linear transformation to standard scores. By including linear transformations, CATREG can also be applied to data containing numeric variables. A numeric variable can also be nonlinearly transformed, in which case the relative spacing of the category values will not be respected. Thus, optimal scaling is applicable to both categorical variables (for quantification) and to numeric variables (for nonlinear transformation). Optimal scaling can be applied more generally with analysis techniques other than regression, for example with principal components analysis and canonical correlation analysis; for an extensive overview see Gifi (1990).

1.2.1 Optimal scaling levels

In the quantification process certain properties of the data are preserved in the transformations. The properties that are chosen to be preserved are specified by choosing an optimal scaling level for the variables. It is important to realize that the optimal scaling level is the level on which a variable is analyzed, which does not need to be the same as the measurement level of the variable. The properties of data that are distinguished in the CATREG approach are grouping, ordering and equal relative spacing. Depending on the measurement level (nominal, ordinal, or interval), variables have one, two, or all of these properies. Variables with nominal measurement level only have the grouping property, that is, the category values only serve to code the observations into classes. Ordinal variables have the properties of grouping and ordering. Interval (numeric) variables have all three properties. If the researcher wants to preserve all of the (possibly assumed) properties of the measured variables in the quantified variables, the scaling level should be chosen in accordance with the measurement level of the variable. With nominal scaling level, only the grouping property is preserved, ordinal scaling level preserves grouping and ordering, and the numeric scaling level preserves grouping, ordering, and equal relative spacing. Note that choosing the numeric scaling level for a categorically measured variable implies that in the analysis category values are treated as numeric values (and when all variables are treated numerically, CATREG is equivalent to standard linear regression). The shape of the curve when plotting the quantified values against the category values is related to the scaling level: with the nominal scaling level the transformation curve can go up an down since the ordering of the quantified values need not be the same as the ordering of the original category values. For the ordinal scaling level, the ordering of the quantified values and of the original category values is the same, resulting in a monotone transformation curve. The numeric scaling level results in a straight line, because the intervals between quantifications for consecutive categories are proportional to the intervals between the category values.

The scaling level, and thus the shape of the transformation curve, is also related to the number of degrees of freedom (DF) of the transformation, and thereby to the fit (R^2) of the model. Transformations with more freedom result in less smooth transformations and better fit, while more restrictive transformations are smoother but result in less fit. So, there is a trade-off between preserving the properties of the data and preserving the relational information in the data: restricting the transformations, preserving more properties of the data, goes at the cost of fit and loosing relational information. The trans-

formation with the maximum freedom is the one resulting from the nominal scaling level, where the number of DF is equal to the number of categories minus one. The ordinal scaling level requires an order restriction on the category quantifications, resulting in the number of DF equal to the number of categories with different quantified values minus 1. Numeric scaling imposes an interval restriction in addition to the order restriction and has 1 DF.

Nominal and ordinal scaling level result in transformations that are step functions, which are suited for variables with a rather small number of categories. For variables with a large number of categories, spline functions are more appropiate, among these we distinguish nonmonotonic splines for unordered transformations and monotonic splines for ordered transformations. Spline functions are piecewise polynomial functions, which are more restrictive than step functions, resulting in smoother transformation curves, but in a lower fit. To obtain a spline transformation, the range of a variable is divided into a number of intervals, equal to a specified number of knots minus 1. Knots are the points at the interval boundaries. Then polynomial functions of a specified degree are fitted in each interval and joined at the knots. The smoothness and the number of DF of a spline transformation curve depends on the number of knots and the degree of the polynomial functions.

In terms of restrictiveness, and thus smoothness of the transformation curve and fit, nonmonotonic spline transformation is in between a nominal and a linear transformation: with the number of interior knots equal to the number of categories minus two and a first degree polynomial, the spline transformation is the same as the nominal transformation; with the number of interior knots equal to zero and a first degree polynomial, the spline transformation is the same as the linear transformation. In the same way, a monotonic spline transformation is in between an ordinal and a linear transformation. This is illustrated in Figure 1.1, which displays transformation plots for prediction of clarity of diamonds from their carat. The data consist of a sample of 239 diamond stones (from tradeshop.com; diamonds with an IGI certificate) with seven clarity grades (Internally Flawless to Slight Inclusions invisible to the eye) and carat*100 ranging from 16 to 432 (binned into the customary size ranges). Clarity has been analyzed ordinally, and a variety of scaling levels has been applied to carat. With the nominal scaling level for carat, the quantification curve is rather jagged, resembling an inverse u-shape. So, the trend for this sample is that in the lower and the highest size ranges, diamonds have lower clarity grades, while the highest clarity is found for diamonds in the 96–99 carat category. Applying a nonmonotonic spline transformation (2nd degree, with 6 internal knots), the irregularities are smoothed out somewhat, and even more for 1 internal knot. Because

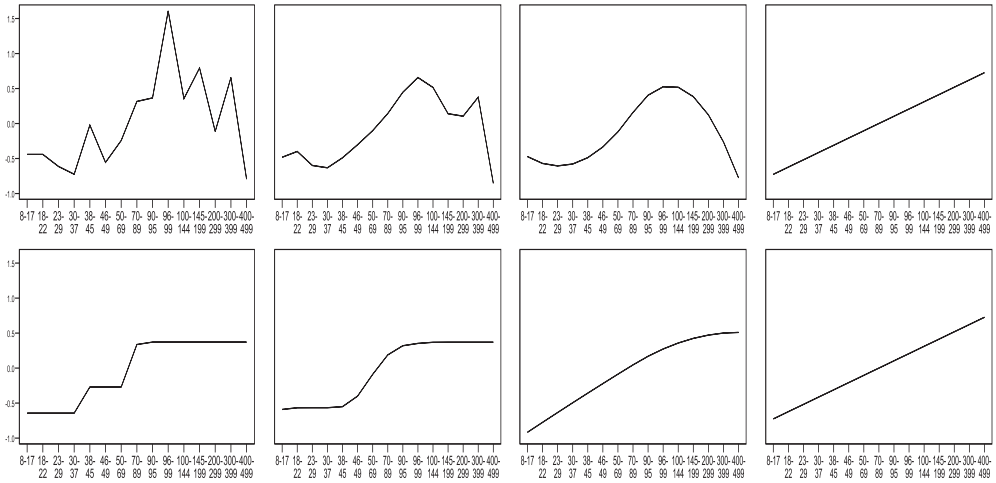

Figure 1.1. Transformations of the carat of diamonds for predicting the clarity. The y-axis represents the quantified values multiplied with the regression coefficient. Top row from left to right: nominal ($R^2 = 0.226$), nonmonotonic spline of degree 2 with 6 ($R^2 = 0.161$) and 1 ($R^2 = 0.157$) internal knot, numeric ($R^2 = 0.093$). Bottom row: ordinal ($R^2 = 0.153$), monotonic spline of degree 2 with 6 ($R^2 = 0.131$) and 1 ($R^2 = 0.112$) internal knot, numeric.

ordinal transformations are obtained by averaging of nominal quantifications that are in the wrong order (see also next section), applying the ordinal scaling level results in a transformation that restricts all quantified values but one to three plateau's. When applying a monotonic transformation (2nd degree, with 6 knots), the middle plateau disappears and with 1 internal knot the transformation is almost linear. (Although the the nominal and ordinal transformations are step functions, in the plots the points are connected with a straight line to facilitate graphical inspection of the transformation trends.)

In Figure 1.2 the linearization of the regression is illustrated. On the left, clarity is plotted against the original values of the carat variable, the middle plot displays the nominal transformation of carat, and on the right clarity is plotted against transformed carat. In the plots on the left and the right, the linear regression line is included, as well as the line connecting the means of the values of clarity for each category of carat; this line indicates the trend in the scatter of data points, which is clearly nonlinear with the untransformed carat variable. When carat is transformed, however, this line is linear, coinciding with the regression line. Thus, the nonlinear relation between the variables, displayed in the plot on the left, is replaced by a linear relation, displayed in

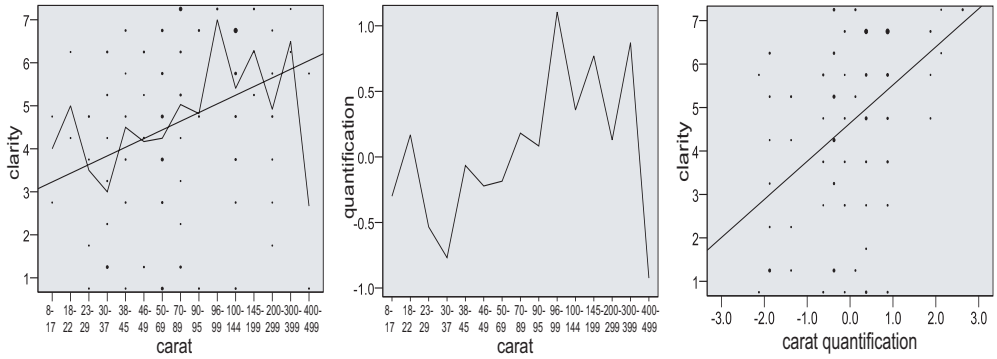

Figure 1.2. Left: Clarity against original values of carat ($R^2 = 0.077$). Middle: Carat transformation. Right: Clarity against transformed carat ($R^2 = 0.169$). (The size of the point markers indicate density: the biggest marker represents about 20 points, decreasing with about 5 points.) .

the plot on the right. This is achieved by a nonlinear transformation of the carat variable, which follows the trend in the plot on the left.

1.2.2 Estimation of transformations

In the CATREG method, the regression model and the quantifications are estimated simultaneously in an iterative process using *alternating least squares*. The algorithm alternates between estimation of the transformation of the response variable and estimation of the transformations and regression weights of the predictor variables. The transformation of the response in an iteration is estimated from the linear combination of the transformed predictors from the previous iteration. Backfitting (Friedman and Stuetzle 1981; Buja, Hastie, and Tibshirani 1989) is used for estimation of the transformations of the predictor variables. In the backfitting procedure, the transformations are estimated from the partial residual, that is, the error when the response is predicted from all predictors, except the predictor for which the transformation is being estimated. For the predictor first estimated in an iteration, the partial residual is computed using the transformations for the other predictors from the previous iteration. Then, in computing the partial residual for the second predictor, the updated transformation for the first predictor is used. For the partial residual for the third predictor, the updates for the first and second predictors are used, etc.

Nominal quantifications are the starting point (and the end point if the scaling level is nominal) in estimating restricted quantifications. The nomi-

nal quantification for a category is the mean of the predicted values for the category when the response is estimated and the mean of the partial residual values for the category when a predictor is estimated. If the scaling level is not nominal, these quantifications are restricted according to the scaling level. The restriction is imposed by applying weighted (weighting with category frequencies) regression of the nominal quantifications; on the category values for ordinal and numeric scaling level, and on an I-spline basis (Ramsay 1988) for the spline transformations, with nonnegativity restrictions for the monotonic splines. For the ordinal scaling level, weighted monotonic regression (Kruskal 1965; Barlow, Bartholomew, Brenner, and Brunk 1972) is used, which boils down to weighted averaging of nominal quantifications for categories that are in the wrong order. With the numeric scaling level, the category values are converted to standard scores, which is equivalent to weighted linear regression of the nominal quantifications on the category values. Finally, the quantified variable is normalized, and for a predictor variable the regression coefficient is estimated (which, after the algorithm has converged, is equal to the normalization factor). In the CATREG method a monotonic transformation is always increasing with the category values. If the scaling level for a predictor is ordinal or monotonic spline, and the relation with the response (after removing the influence of the other predictors) is monotonically decreasing, this is expressed in a negative regression coefficient.

To describe the quantification process graphically, we use the same diamonds data set as before, predicting the clarity now from two variables: the carat (with monotonic spline transformation of 2nd degree with 6 internal knots) and the cut (with nominal transformation). The top row of Figure 1.3 shows the predicted values plot (left) for the dependent variable and the partial residuals plot for the predictors. The bottom row shows the transformations. The partial residuals are $\mathbf{e}^r_{\text{carat}} = \phi^r(\mathbf{y}) - \beta^{r-1}_{\text{cut}}\phi^{r-1}_{\text{cut}}(\mathbf{x})$ and $\mathbf{e}^r_{\text{cut}} = \phi^r(\mathbf{y}) - \beta^r_{\text{carat}}\phi^r_{\text{carat}}(\mathbf{x})$, where the superscript r denotes the last iteration number.

In the plots at the top of Figure 1.3, the means of the values for each category are connected by a line. These means are the unstandardized nominal quantifications. So, the transformation curve for cut is equal to the curve in the partial residuals plot (the transformation plots display the quantifications multiplied with the regression coefficient, which equals the standardization factor, so the transformed values here are the unnormalized quantifications). The transformation for carat is a somewhat smoothed version of the curve in the partial residuals plot. The transformation for clarity is obtained by first restricting the means for categories VS1 and VS2 in the predicted values plot to their weighted mean, because the quantifications for these categories are in

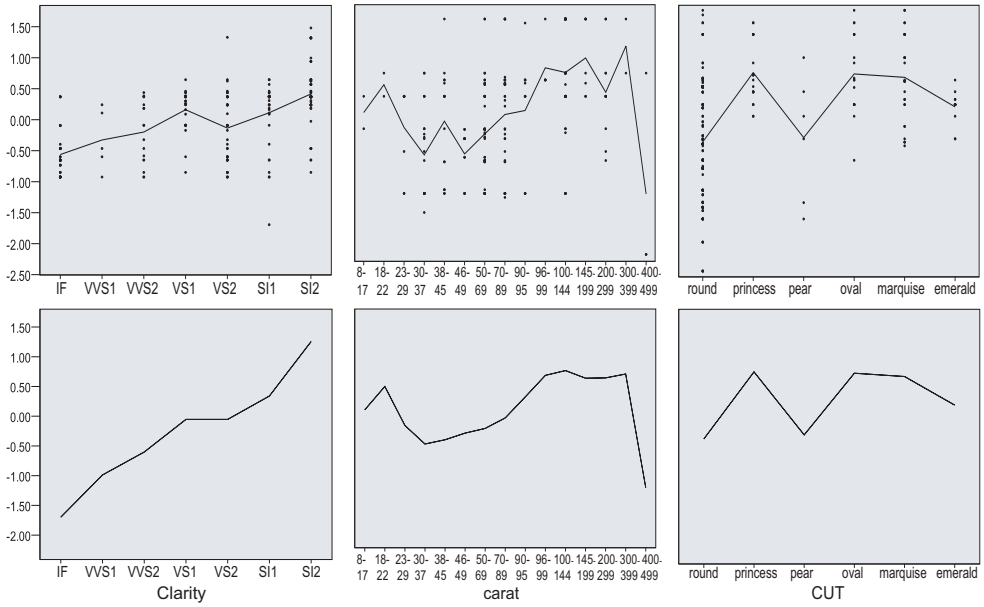

Figure 1.3. Prediction of clarity from carat and cut ($R^2 = 0.331$, tolerance = 0.902). Top row: Partial residuals. Bottom row: Transformations (including regression coefficient for predictors).

the wrong order, and then standardizing the result.

1.2.3 Local minima

All three approaches to regression for nonlinearly related data mentioned above use iterative algorithms for estimation. Iterative estimation methods require starting values, and depending on these starting values the methods may not arrive at the global optimal solution, but at a local optimal solution, that is, a solution for which the error function is at a local minimum. With the CATREG method, local minima can only occur with ordinal and monotonic spline transformations. This results from the fact that a monotonic transformation can be either increasing or decreasing. For a variable with nominal or nonmonotonic spline transformation, the transformation can be reflected without consequence for the contribution to the prediction, because reflecting the quantifications will only cause a sign change of the regression coefficient. On the other hand, reflecting nominal quantifications during the iteration process for ordinal and monotonic spline variables has consequences, because a monotonic restriction on the reflected nominal quantifications results in a

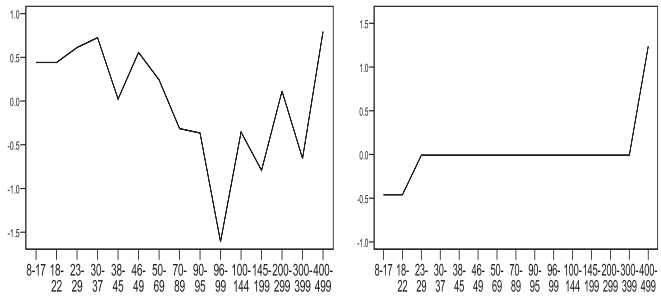

Figure 1.4. Transformations of the carat variable for predicting the clarity of diamonds. The y-axis represents the quantified values multiplied with the regression coefficient. Left: Reflected nominal ($R^2 = 0.226$). Right: Ordinal, estimated from reflected unrestricted quantifications ($R^2 = 0.024$).

different transformation. For example, if for carat with ordinal scaling level the unrestricted nominal quantifications would have been reflected during iteration, the transformation would be as shown in Figure 1.4, resulting in a local minimum (the ordinal transformation in Figure 1.1 resulted in the global minimum).

1.2.4 Prediction accuracy, stability, and regularization

The focus of the CATREG method is on minimizing the apparent prediction error, that is, the prediction error for the data set at hand. To asses the prediction accuracy for predicting future observations, the expected prediction error should be studied. For this purpose, the ideal case is to have a training data set to estimate a model and a test data set to estimate the expected prediction error, applying the model found with the training set to the test set. Often, a test data set is not available and resampling methods like cross validation or the bootstrap are used. A particular resampling method, called the .632 bootstrap (Efron 1983), could be regarded as a combination of bootstrapping and cross validation, and has been shown to give better results than cross validation and other bootstrap methods. Using the .632 bootstrap with CATREG requires some adjustment, because both model parameters (regression coefficients) and transformations (quantifications) are involved.

Often, ordinary least squares (OLS) regression does not perform well with respect to prediction accuracy, due to (highly) instable estimates of the regression coefficients, sometimes referred to as *bouncing beta's*. Instability arises when predictors are (highly) correlated, or when we have a large number of predictors relative to the number of observations. Regularization methods

started with Ridge regression, (Hoerl and Kennard 1970a,b), reducing the variability in the β estimates by adding a penalty parameter to the regression model, which results in shrunken regression coefficients. Ridge regression gradually shrinks the coefficients towards zero, but in this process they never become exactly zero and all predictors are kept in the model, which does not enhance interpretability. The Lasso (Least Absolute Shrinking and Selection Operator; Tibshirani 1996) was developed to address both prediction accuracy and model complexity. The Lasso uses an L_1 penalty, in stead of the L_2 penalty used in Ridge regression, resulting also in gradual shrinkage of the coefficients towards zero, but following different paths. The main difference between the shrinkage paths of Ridge regression and the Lasso is that on the Lasso paths, regression coefficients are shrunken to exactly zero, and some earlier than others. Thus, the Lasso combines shrinkage with subset selection. With the Pathseeker algorithm of Friedman and Popescu (2004), the paths are not found by applying a fixed penalty function, but are found directly, providing a flexible regularization technique. The combination of shrinkage and subset selection is also a feature of the recently developed Elastic Net (Zou and Hastie 2005). In contrast to the Lasso, the Elastic net can select more predictors than there are observations, and results in shrinking groups of predictors. Ridge regression, the Lasso, and the Elastic Net were developed for regularization of linear regression, but can easily be incorporated into the CATREG method. Recently, two methods were introduced for regularization of regression with nonmonotonic transformations: the Group-Lasso (Yuan and Lin 2006) and Blockwise Sparse Regression (Kim, Kim, and Kim 2006). These methods use groups or blocks of dummies to represent variables and turn out to be equivalent to CATREG with nonmonotonic transformations.

1.3. Outline

Chapters 2 to 5 of this monograph are based on either published or submitted papers. This causes some inevitable overlap, because in all these chapters the basic elements of the CATREG method are described. Chapter 2 focuses on the problem of local minima, presenting several strategies to obtain the global minimum for the ordinal scaling level and the results of a simulation study to asses the performance of these strategies. The simulation study was also used to identify data conditions under which local minima are more likely to occur and are more likely to be severe. The topic of chapter 3 is the assessment of the prediction accuracy of CATREG, using the .632 bootstrap. The differences in prediction accuracy for the optimal scaling levels are studied, as well as the effect of the number of observations and the number of cate-

gories. The prediction accuracy of CATREG is compared with the prediction accuracy of a variety of other Regression with Transformation methods. In Chapter 4, the implementation of Ridge regression, the Lasso and the Elastic Net in CATREG is developed and illustrated. Also, the equivalence of the Group Lasso and Blockwise Sparse Regression with CATREG using non-monotonic transformations is established. Chapter 5 presents an application of the CATREG method incorporating the Lasso and the .632 bootstrap, for data from the psychotherapeutic field. Finally, chapter 6 concludes this monograph with a summary and a discussion.

The CATREG method is available in the Categories package of SPSS (Meulman, Heiser, and SPSS Inc. 1999, 2004). Implementation and documentation of the CATREG progam was done for a major part by the author of this monograph. The appendices contain a detailed description of the CATREG algorithm, as well as the CATREG chapters from the SPSS Categories manual.