Prediction accuracy and stability of regression with optimal scaling transformations

Kooij, A.J. van der

Chapter 2

Local Minima in Regression with Optimal Scaling Transformations

CATREG is a program for categorical multiple regression, applying optimal scaling methodology to quantify categorical variables, including the response variable, simultaneously optimizing the multiple regression coefficient. The scaling levels that can be applied are nominal, nonmonotonic spline, ordinal, monotonic spline or numerical. When ordinal or monotonic spline scaling levels are applied, local minima can occur. With ordinal or monotonic spline scaling levels, the transformations are required to be monotonically increasing, but this can also be achieved by reflecting a monotonic decreasing transformation. A monotonic transformation is obtained by restricting a nonmonotonic transformation, but the direction of the monotonic restriction (increasing or decreasing) is undefined, and it will be shown that this is the cause of local minima. Several strategies to obtain the global minimum for the ordinal scaling level will be presented. Also, results of a simulation study to asses the performance of these strategies are given. The simulation study is also used to identify data conditions under which local minima are more likely to occur and are more likely to be severe. It was found that local minima more often occur with low to moderately low R^2 values, with higher number of categories and with higher multicollinearity.

Reprinted from Computational Statistics and Data Analysis, 50, Van der Kooij, A.J., Meulman, J.J., & Heiser, W.J., Local Minima in Categorical Multiple Regression, 446–462, Copyright (2006), with permission from Elsevier.

2.1. Introduction

With numerical data multiple regression is the most often used method to predict a dependent or response variable from a set of predictor variables. CATREG is a nonparametric method to perform multiple regression when data are categorical or a mix of numerical and categorical variables. CATREG allows for nonlinear transformations of the variables, including the response variable. CATREG can also be used with numerical data to explore the existence of nonlinear relationships. The program is available in SPSS (SPSS Inc. 1998; Van der Kooij and Meulman 1999).

Transformation of variables has become an important tool in data analysis over the last decades. Various models and programs have been developed, for example the Box-Cox model (Box and Cox 1964), using parametric families of transformations. More general methods include monotone transformations (Kruskal 1965), MORALS (Young et al. 1976) (implemented in TRANSREG (SAS/STAT 1989)), spline transformations (Winsberg and Ramsay 1980; Ramsay 1988), Projection Pursuit Regression (Friedman and Stuetzle 1981), ACE (Breiman and Friedman 1985), and GAM (Hastie and Tibshirani 1990) for fitting generalized additive models using smoothers; see also Gaudart, Giusiano, and Huiart (2004) for a comparison of Neural Networks and linear regression.

CATREG applies the optimal scaling methodology as developed in the Gifi system (Gifi 1990) to quantify categorical variables, simultaneously optimizing the multiple regression coefficient. In the quantification process, information in the observed variable is retained in the quantified variable. The kind of information that is retained, and thereby the form of the transformation, depends upon the scaling level. The numerical scaling level results in a linear transformation, that is, the data are treated as numerical, and are only transformed into standardized variables. The non-numerical scaling levels allow for nonlinear transformations: the nominal and nonmonotonic spline scaling levels allow for nonmonotonic transformations; the ordinal and monotonic spline scaling levels allow for monotonic transformations. The scaling level can be chosen for each variable separately.

When all or some of the scaling levels are of the ordinal or monotonic spline type, local, nonglobal minima (suboptimal solutions) can occur. Ordinal or monotonic spline quantifications are obtained by applying a restriction to unrestricted (nominal, nonmonotonic) quantifications. In this chapter, we argue that the occurrence of local minima in CATREG is due to the fact that the direction of the restriction is undefined. CATREG is closely related to MORALS (Young et al. 1976), TRANSREG (SAS/STAT 1989) and ACE

(Breiman and Friedman 1985). With both these methods we have observed local minima as well. MORALS, TRANSREG, and ACE also use monotonic restriction of nonmonotonic transformations to obtain monotonic transformations, so it is very likely that the cause of local minima in MORALS and ACE is the same as in CATREG.

In this chapter we present a strategy to obtain the global minimum, using systematic multiple starts. That is, each start uses the same initial values for the quantifications, but with different signs. Applying a negative sign results in a monotonic increasing transformation that is equal to a reflected decreasing transformation. Using all possible sign patterns will yield the global minimum. However, the number of possible sign patterns is a power of two, where the power is the number of predictor variables with ordinal or monotonic spline scaling level. So, for each additional predictor the number of systematic multiple starts is doubled. To reduce the number of starts, we have adapted the strategy to use not all but only some sign patterns. To determine which sign patterns to use, a loss of variance criterion is used. We also developed a second strategy to find the optimal sign pattern by using a hierarchical method. A third strategy is a combination of these two strategies. The performance of the reduced strategies is assessed by a simulation study. In addition, the simulation study is used to asses the possible effect of data conditions on the incidence and severeness of local minima. For this purpose, we simulated data with different sizes of R^2, different amounts of multicollinearity, and different number of categories.

2.2. Model

Multiple regression is a linear technique to study the relationship between a response variable and a set of predictor variables. Categorical multiple regression is a nonlinear technique, where the nonlinearity is in the transformation of the variables. CATREG requires the data to be categorical, thus consisting of integer values. For non-integer data CATREG offers various discretization options for recoding (binning) or linearly transforming the data into integer values.

The CATREG model is the classical linear regression model, applied to transformed variables:

$$\varphi_r(\mathbf{y}) = \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j) + \mathbf{e}, \tag{2.1}$$

with loss function:

$$L(\varphi_r; \varphi_1, \ldots, \varphi_j; \beta_1, \ldots, \beta_j) = \|\varphi_r(\mathbf{y}) - \sum_{j=1}^{J} \beta_j \varphi_j(\mathbf{x}_j)\|^2, \tag{2.2}$$

with J the number of predictor variables, $\mathbf{y}$ the observed or discretized response variable, $\mathbf{x}_j$ the observed or discretized predictor variables, β_j the regression coefficients, φ_r the transformation for the response variable, φ_j the transformations for predictor variables, and $\mathbf{e}$ the error vector. All transformed variables are centered and normalized to have sum of squares equal to N, and $\| \cdot \|^2$ denotes the squared Euclidean norm. (An alternative formulation of multiple regression with optimal scaling was given in Van der Kooij and Meulman (1997).)

The form of the transformations depends upon the optimal scaling level, which can be chosen for each variable separately, and is independent of the measurement level. The scaling level defines which part of the information that is in an observed or discretized variable (depending upon the measurement level), is retained in the transformed variable. With the numerical scaling level, the category values of a variable are treated as quantitative. Then all information is retained and the only transformation applied is standardization, thus resulting in a linear transformation. So, when for all variables numerical scaling level is applied, the result of CATREG is equal to the result of linear multiple regression with standardized variables. With the nonnumerical scaling levels, the category values are treated as qualitative, and are transformed into quantitative values. In this case, some part of the information in the observed or discretized variable is dropped, allowing for nonlinear transformations. With the ordinal or monotonic spline scaling level, the interval information is dropped, and only grouping and ordering information has to be retained, allowing for a monotonic transformation. With the nominal and nonmonotonic spline scaling level only grouping information has to be retained, resulting in nonmonotonic transformations. By applying nonlinear scaling levels, nonlinear relationships between the response variable and the predictor variables are linearized, hence the term linear regression model is still applicable.

2.3. Algorithm

In CATREG the observed or discretized variables are 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 $\mathbf{v}_m$:

$$\varphi_r(\mathbf{y}) = \mathbf{G}_r\mathbf{v}_r, \text{ and } \varphi_j(\mathbf{x}_j) = \mathbf{G}_j\mathbf{v}_j, \tag{2.3}$$

where $\mathbf{v}_r$ is the vector of category quantifications for the response variable, and $\mathbf{v}_j$ the vector of category quantifications for a predictor variable. So, the CATREG model with the transformed variables written in terms of indicator matrices and category quantifications is:

$$\mathbf{G}_r\mathbf{v}_r = \sum_{j=1}^{J} \beta_j \mathbf{G}_j\mathbf{v}_j + \mathbf{e}, \tag{2.4}$$

with the associated least squares loss function:

$$L(\mathbf{v}_r; \mathbf{v}_1, \ldots, \mathbf{v}_j; \beta_1, \ldots, \beta_j) = \|\mathbf{G}_r\mathbf{v}_r - \sum_{j=1}^{J} \beta_j \mathbf{G}_j\mathbf{v}_j\|^2. \tag{2.5}$$

The loss function (2.5) is minimized by an alternating least squares algorithm, alternating between the quantification of the response variable on the one hand, and the quantification of the predictor variables and estimation of the regression coefficients on the other hand. First, the quantifications and the regression coefficients have to be initialized. CATREG has two ways of initialization: random and numerical. Random initialization uses standardized random values 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.

In the first step of the algorithm, the quantifications of the predictor variables and the regresssion coefficients are held fixed. With numerical scaling level the quantifications $\mathbf{v}_r$ of the response variable are the category values of the centered and normalized observed or discretized variable. With a non-numerical scaling level the quantifications are updated in the following way:

$$\tilde{\mathbf{v}}_r = \mathbf{D}_r^{-1}\mathbf{G}_r' \sum_{j=1}^{J} \beta_j \mathbf{G}_j\mathbf{v}_j, \tag{2.6}$$

where $\mathbf{D}_r = \mathbf{G}_r'\mathbf{G}_r$. The quantifications $\tilde{\mathbf{v}}_r$ are the unstandardized quantifications for the nominal scaling level. For ordinal, and nonmonotonic or monotonic spline scaling level, a restriction is applied to $\tilde{\mathbf{v}}_r$, 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 ordinal and spline scaling levels. Then $\mathbf{v}_r^*$ is standardized:

$$\mathbf{v}_r^+ = N^{1/2}\mathbf{v}_r^*(\mathbf{v}_r^{*\prime}\mathbf{D}_r\mathbf{v}_r^*)^{-1/2}. \tag{2.7}$$

In the second step of the algoritm, the quantification of the response variable is held fixed, and the quantifications $\mathbf{v}_j$ of predictor variables with non-numerical scaling level, and the regression coefficients are updated for one variable at a time, this is sometimes called backfitting, and was applied, a.o., in Kruskal (1965), De Leeuw et al. (1976), Friedman and Stuetzle (1981), Hastie and Tibshirani (1990), and Gifi (1990)). The approach works at follows. First the N-vector of predicted values is computed:

$$\mathbf{z} = \sum_{j=1}^{J} \beta_j \mathbf{G}_j \mathbf{v}_j. \tag{2.8}$$

For updating the quantifications of variable j, the contribution of variable j to the prediction (the weighted linear combination of transformed predictors) is subtracted from $\mathbf{z}$:

$$\mathbf{z}_j = \mathbf{z} - \beta_j \mathbf{G}_j \mathbf{v}_j. \tag{2.9}$$

The unrestricted quantifications are updated as:

$$\tilde{\mathbf{v}}_j = \text{sign}(\beta_j)\mathbf{D}_j^{-1}\mathbf{G}_j'(\mathbf{G}_r\mathbf{v}_r^+ - \mathbf{z}_j). \tag{2.10}$$

For variables with non-numerical scaling level $\tilde{\mathbf{v}}_j$ is restricted according to the scaling level, and normalized as in (2.7), yielding $\mathbf{v}_j^+$. For variables with numerical scaling level $\mathbf{v}_j^+$ contains the category values of the centered and standardized observed or discretized data. Next the regression coefficient β_j is updated:

$$\beta_j^+ = N^{-1}\tilde{\mathbf{v}}_j'\mathbf{D}_j\mathbf{v}_j^+. \tag{2.11}$$

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

$$\mathbf{z} = \mathbf{z}_j + \beta_j^+ \mathbf{G}_j \mathbf{v}_j^+, \tag{2.12}$$

and the algorithm continues with updating the quantification for the next predictor variable, until all predictor variables are updated. Then the loss is computed as $\|\mathbf{G}_r\mathbf{v}_r^+ - \mathbf{z}\|^2$. These two steps are repeated until a user-specified convergence criterion is met.

For the ordinal scaling level, weighted monotonic regression (Kruskal 1965; Barlow et al. 1972) of the nominal quantifications on the observed or discretized variable is used. For the restriction according to the spline scaling levels we use weighted regression of nominal quantifications on an I-spline basis (Ramsay 1988), with additional nonnegativity restrictions for the monotonic spline scaling level. At this point, a complication may occur. A monotonically increasing restriction can sometimes result in a transformed variable with constant values. For example, when the values of $\tilde{\mathbf{v}}$ are monotonically decreasing, except for the first or the last value, the restricted quantifications are the mean of $\tilde{\mathbf{v}}$ for all categories. In this case, transformation to a constant can be avoided by allowing for a monotonically *decreasing* function, and reflecting the result to make the quantifications increasing. In CATREG we accomplish this by applying monotonic (spline)regression to the reflected unrestricted quantifications (see also section 2.4.1).

2.4. Local minima

When using multiple random starts (that is, a number of different starts, each with random initial values), ordinal or monotonic spline scaling levels for the predictor variables, and numerical scaling level for the response variable, we may obtain different values of R^2. Also, the use of the same random start while changing the order in which the predictor variables enter the backfitting algorithm, can result in a different R^2 value. The maximum number of different R^2 values obtained with multiple random starts is 2^q, with q the number of predictor variables with ordinal or monotonic spline scaling level. All the 2^q solutions have different sign patterns for the quantifications and associated β's: there is one solution with all β's positive, q solutions with one negative β, $\binom{q}{r}$ solutions with r negative β's, and one solution with all β's negative. The values of the initial quantifications are not important: using different initial values but with the same sign pattern for the regression coefficients results in the same R^2.

2.4.1 A probable cause of local minima

Our conjecture is that local minima with ordinal or monotonic spline scaling level occur because the direction of the monotonic restrictions is undefined. Ordinal or monotonic spline quantifications are required to be monotonically increasing, which is accomplished in CATREG by applying a monotonically increasing restriction to nominal quantifications. But this can also be accomplished by applying a monotonically decreasing restriction to nominal quantifications and reflecting the result. Applying a monotonically decreasing restriction and reflecting the result is equivalent to applying a monotonically increasing restriction to reflected nominal quantifications. The order of nominal quantifications is meaningful, but they need not to be ordered according to the data, so reflection is allowed. For a predictor variable, reflection of the nominal quantifications, without applying a restriction, results in an updated β that only differs in sign from the updated β using unreflected quantifications. For the response variable the only difference is a change of sign for all β's. But ordinal or monotonic spline restriction on the reflected nominal quantification results in different restricted quantifications, and thus β's that differ both in sign and value.

Table 2.1 and Figure 2.1 give an example of ordinal restriction for a predictor variable on nominal quantifications and on reflected nominal quantifications. The first column displays the categories of a predictor variable, with all categories having equal frequency. The nominal quantifications are in the second column. The quantification for category 2 is lower than the one for category 1, resulting in the restricted quantification for both categories equal to their mean (note that here and in the following we have to use weighted means when the frequencies are not equal). This mean is 0.40, which is higher than the nominal quantification for category 3, and this results in a restricted quantification for categories 1, 2 and 3 equal to their mean, and this mean (-0.15) is lower than the nominal quantification for category 4; hence this quantification remains unchanged. The fourth column displays the reflected nominal quantifications. Here the nominal quantifications for categories 1, 2, and 3 are correct, that is they increase; but for category 4 the quantification decreases. Therefore the quantifications of category 3 and 4 are averaged. So, in the example, ordinal restriction of the nominal quantifications results in

Table 2.1. Ordinal restrictions on unreflected and reflected nominal quantifications.

Category (Equal freq.)	Nominal quantification	Ordinal Restriction	Reflected nominal quantification	Ordinal Restriction
1	1.05	-0.15	-1.05	-1.05
2	-0.25	-0.15	0.25	0.25
3	-1.25	-0.15	1.25	0.40
4	0.45	0.45	-0.45	0.40

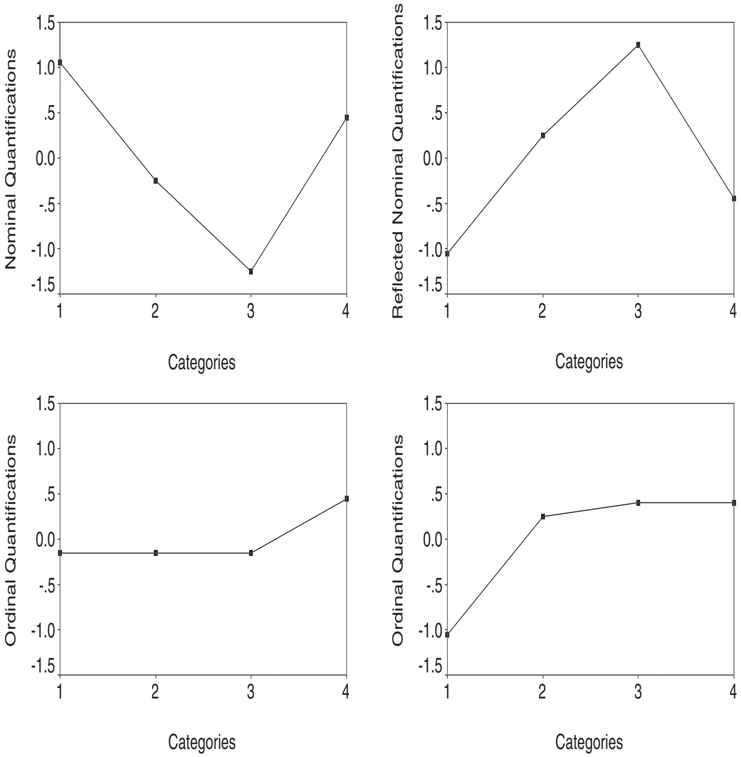

Figure 2.1. Transformation plots. Top left: nominal transformation. Top right: reflected nominal transformation. Bottom left: ordinal transformation, (restricted nominal transformation). Bottom right: ordinal transformation (increasing monotonic restriction on reflected nominal transformation, or reflected decreasing monotonic restriction on unreflected nominal transformation).

equal quantifications for categories 1, 2, and 3, and a positive β, while ordinal restriction of the reflected nominal quantifications results in equal quantifications for categories 3 and 4, and a negative β. These two sets of quantifications have different variance, which eventually results in different local minima.

2.4.2 Strategies to obtain the global minimum

Since there are 2^q possible different solutions, with 2^q sign patterns for the regression coefficients, we can obtain the global minimum using 2^q starts, all with the same initial values, but with different sign patterns. The signs are applied to the unrestricted quantifications (replacing $\text{sign}(\beta_j)$ in (2.10)) and remain fixed during iteration. As explained above, applying a sign to the unrestricted quantifications, results in a β with the same sign. This strategy of multiple systematic starts works fine if the number of ordinal or monotonic spline predictors is relatively small. Since the number of required starts is a power of two, this number doubles with each additional ordinal or monotonic spline predictor, and therefore complete enumeration is often infeasible. So we searched for a more direct method to find the optimal sign pattern, and developed several strategies for this purpose. Although none is completely successful, their performance is much better than with the numerical or random start. That is, using a sign pattern determined in some reasonable way results in less local minima than using the numerical or random start.

Iterative strategies

Iterative strategies attempt to find the optimal signs during iteration. Then for each variable with the scaling level 'ordinal'or 'monotonic spline', in each iteration the sign is selected. With these strategies we used numerical initialization. Strategy 1 is to select the sign according to the loss due to restriction for the unreflected and the reflected nominal quantifications (thus, the loss for the monotonic increasing and the monotonic decreasing transformation; this strategy is also applied in ACE (although this is not reported in the Breiman and Friedman, 1985 paper, this approach is implemented in Friedman's ACE function in S-PLUS)). The loss due to restriction is $\|\mathbf{G}\tilde{\mathbf{v}} - \mathbf{G}\mathbf{v}^*\|^2$. Strategy 2 is to select the sign for a variable according to the R^2 resulting from the unreflected and reflected updates for the variable to be transformed, with temporarily updates for the other variables. Thus, this strategy takes into account the effect of the sign for the variable to be updated on the other variables. In the simulation study, the iterative strategies reduce the number of local minima with about 50% (strategy 1), and 75% (strategy 2). Strategy 1 is not always successful in finding the global mininum, because it can happen that by applying the sign that results in a higher loss (due to the monotonic restriction), this higher loss is 'later'compensated by a lower loss due to monotonic restrictions for other variables. Strategy 2 was developed to take this into account. The reason strategy 2 is not always successful in finding the global mininum either, is that it can happen that in the earlier iterations applying particular signs results in the highest R^2, while these signs eventually result in a local minimum. So, then the global minimum can only be found with suboptimal updates in the earlier iterations.

Non-iterative strategies

We have also tried a number of non-iterative strategies. These strategies fix the signs that are selected in the first iteration using one of four criteria. With these strategies we also used numerical initialization. The criteria we used to select the signs were as follows. The sign for a variable is set to negative if: the loss due to restriction on $-\mathbf{v}$ is equal to or less than the loss due to restriction on $\mathbf{v}$ (strategy 3); the number of different category quantifications is half or less than half of the number of categories (strategy 4); Spearman's rank-order correlation (that is, the correlation between the ranked values of $\tilde{\mathbf{v}}$ and the ranked observed or discretized data) is zero or negative (strategy 5); the weighted sum of differences between consecutive values in $\tilde{\mathbf{v}}$ is zero or negative (strategy 6). It turned out that none of these strategies worked by themselves (they gave more local minima than with the numerical start, except for strategy 3), but as it turned out, they can be used to reduce the number of multiple systematic starts.

Reduced multiple systematic starts strategy using non-iterative strategies to select signs patterns

This approach uses non-iterative strategies to reduce the number of multiple systematic starts by selecting sign patterns. We select sign patterns by allowing a negative sign only for those variables that have a negative sign in the pattern found with one of the non-iterative strategies (strategies 3 through 6). Or, stated differently, sign patterns that contain a negative sign for variables with a positive sign in the pattern found with one of the non-iterative strategies, are exluded. In this way the number of multiple systematic starts is reduced (if p is smaller than q) to 2^p, with p the number of negative signs obtained with a non-iterative strategy. This strategy works reasonably well, but using other methods to select sign patterns worked better.

Reduced multiple systematic starts strategies using other methods to select sign patterns

An alternative approach to select sign patterns to reduce the number of multiple systematic starts, is a hierarchical one (strategy 7): first the solution with no negative signs and the q solutions with one negative sign are computed. If R^2 with the zero negative signs pattern is the highest, this sign pattern is selected, and the procedure stops. On the other hand, if the highest R^2 is obtained with one of the one-negative-sign patterns, we continue by computing the $q - 1$ solutions for this one-negative-sign pattern with one more negative sign added. If none of the two-negative-sign patterns results in a higher R^2 than the one for the one-negative-sign pattern, the one-negative-sign pattern is selected, and the procedure stops. Otherwise we continue by adding one more negative sign to the best two-negative-sign pattern, etc. With this method the maximum number of solutions to be computed is reduced from 2^q to $1+\sum_{i=1}^{q} i$. Another, more flexible, method is the use of a percentage of loss criterion to select sign patterns (strategy 8). With this strategy only variables with a loss of variance due to monotonic increasing restriction higher than a certain percentage, are allowed to have a negative sign. The loss of variance is determined in the first iteration after a nominal initial solution, that is, a solution with nominal scaling level for all variables, since the nominal scaling level is the least restrictive. Thus, sign patterns in which variables that do not meet the percentage of loss criterion have a negative sign, are excluded. By varying the percentage value, we can manipulate the number of solutions to be computed, and thereby the chance of obtaining a local minimum. With a percentage of zero, all sign patterns are selected, so the global minimum will certainly be obtained. With a higher percentage value, less sign patterns are selected, thus the chance of obtaining a local minimum is increased. Combining strategies 7 and 8 results in strategy 9: in the hierarchical strategy only variables selected with the percentage of loss strategy are allowed to have a negative sign.

2.5. Simulation study

We have performed a simulation study to test our strategies to reduce the number of starts, and to investigate if and how data conditions effect the incidence and the severeness of local minima.

2.5.1 Design

The data conditions we used as design factors (see Table 2.2) are four ranges of R^2 (low, moderately low, moderately high, and high), five numbers of categories (3, 7, 15, 25, and in the range 40–50), and five levels of multicollinearity (correlation between two predictor variables of 0.0, 0.25, 0.50, 0.75, and 0.95, other predictor intercorrelations close to zero). Summarizing, we have 100 conditions; since we generated 50 data sets for each condition, we have 5000 samples in total. The number of objects was fixed at 200, and the number of predictor variables was fixed at five. The predictor variables were drawn from a normal distribution with specified population correlation (see the five levels of multicollinearity).

Table 2.2. Data conditions.

R^2		No. Categories predictors	Multicollinearity
low:	0.00 – 0.30	3	0.00
mlow:	0.35 – 0.55	7	0.25
mhigh:	0.60 – 0.75	15	0.50
high:	0.80 – 1.00	25	0.75
		40 – 50	0.95

Table 2.3. Difference R^2 resulting from numerical start and optimal R^2. The frequencies give the number of local minima obtained in 5000 samples.

	Frequency						
Difference with	R^2						Cumulative
optimal R^2	low	mlow	mhigh	high	Total	%	%
.000 – .005	180	123	24	20	347	38.0	38.0
.005 – .015	195	88	22	4	309	33.9	71.9
.015 – .025	76	32	15	0	112	13.5	85.4
.025 – .035	36	20	2	0	58	6.4	91.8
.035 – .045	23	8	2	0	32	3.6	95.4
.045 – .055	16	8	1	0	25	2.7	98.1
.055 – .065	2	4	0	0	6	0.7	98.8
.065 – .075	1	1	0	0	2	0.2	99.0
.075 – .085	2	0	0	0	2	0.2	99.2
.085 – .095	1	0	0	0	1	0.1	99.3
.095 – .105	2	0	0	0	2	0.2	99.6
.108	0	1	0	0	1	0.1	99.7
.115 – .125	2	0	0	0	2	0.2	99.9
.142	1	0	0	0	1	0.1	100.0
Total	537	285	66	24	912	100.0	

The response variable was computed as a weighted linear combination of the five predictor variables and an error vector. Random sign patterns were applied to these weights by drawing a random integer from the interval [1,32] ($32 = 2^5$ is the total number of sign patterns with five predictor variables). The values of the weights were chosen such that R^2 was in the desired range. Then the variables were discretized by grouping them into 3, 7, 15, or 25 categories, such that the resulting variable is approximately normal. For the range 40–50 categories condition, and for the response variable, we applied a linear transformation resulting in the required number of categories. We used numerical scaling level for the response variable and ordinal scaling level for the predictor variables.

2.5.2 Results

Using all possible systematic starts the global minimum is always found, requiring $2^5 = 32$ starts for each sample. This gives the optimal R^2 for each sample. With a numerical start (that is, initial values and signs obtained from a solution with numerical scaling level for all variables) 912 analyses (18.2%) resulted in a local minimum. The difference between the optimal R^2 and the R^2 resulting from the numerical start indicates the severeness of the local minima. A local minimum is less severe if the difference with the optimal R^2 is small. In Table 2.3, the frequencies for the differences are displayed for each R^2 condition. Fortunately, we see frequencies go down with increasing severeness. Also, the more severe local minima are mainly in the low and moderately low R^2 conditions.

The results for the strategies 1 through 6 and the reduced multiple starts strategy using strategies 3 through 6 are in Table 2.4. The number of local minima is reduced, but not as much as one would wish. Strategies 7, 8, and 9 give better results, that are given in Tables 2.5, 2.6, and 2.7 respectively.

With the hierarchical strategy (strategy 7), 39 analyses resulted in a local minimum (see Table 2.5), requiring 13.6 starts on average. The reduction of the incidence and severeness of local minima is satisfactory, and the number of starts is reduced to about half. The same is true for the results of the strategy that uses the percentage of loss criterion (strategy 8, see Table 2.6). These results illustrate that the average number of starts is decreased with increasing percentage values, along with an increasing number of local minima. Although the total number of local minima is higher with higher percentage values, the increase is mainly in the local minima that are not severe. With a percentage value of zero, all starts are used, thus finding the optimal R^2 is guaranteed.

Table 2.4. Results of the strategies 1 through 6: number of local minima obtained in 5000 samples. (With the numerical start, 912 local minima were obtained).

Strategy	# local minima	Mean # starts	Std. dev. # starts
Iterative strategies			
Strategy 1	441	N.A.	
Strategy 2	229	N.A.	
Non-iterative strategies			
Strategy 3	555	N.A.	
Strategy 4	2093	N.A.	
Strategy 5	1548	N.A.	
Strategy 6	1354	N.A.	
Reduced multiple systematic starts using non-iterative strategies			
Using strategy 3	508	7.7	6.4
Using strategy 4	87	16.8	12.1
Using strategy 5	557	7.3	6.1
Using strategy 6	505	7.7	6.4

Table 2.5. Results of the hierarchical strategy (strategy 7).

Difference with optimal R^2	Frequency
.000 – .005	23
.005 – .015	14
.015 – .025	2
Total	39
Mean no. starts	13.6
Std. dev. no. starts	2.2
(max. no. starts is 16)	

Table 2.6. Results of the percentage of loss strategy (strategy 8).

	Frequency					
Difference with	loss due to restriction					
optimal R^2	$\geq$ 60%	$\geq$ 50%	$\geq$ 40%	$\geq$ 30%	$\geq$ 10%	$\geq$ 0%
.000 – .005	28	9	7	2	2	0
.005 – .015	19	9	5	3	0	0
.015 – .025	3	2	1	0	0	0
.025 – .035	4	2	1	1	0	0
Total	54	22	14	6	2	0
Mean no. starts	13.0	14.2	15.3	16.5	20.0	32.0
Std. dev. no. starts	10.2	10.8	11.2	11.6	11.9	0.0
(max. no. starts 32)						

Table 2.7. Results of the hierarchical strategy combined with the percentage of loss strategy (strategy 9).

	Frequency					
Difference with	loss due to restriction					
optimal R^2	$\geq$ 60%	$\geq$ 50%	$\geq$ 40%	$\geq$ 30%	$\geq$ 10%	$\geq$ 0%
.000 – .005	48	31	29	25	25	23
.005 – .015	31	21	17	16	14	14
.015 – .025	5	4	3	2	2	2
.025 – .035	4	2	1	1	0	0
Total	88	58	50	44	41	39
Mean no. starts	8.2	8.5	8.9	9.2	10.3	13.6
Std. dev. no. starts	4.2	4.3	4.4	4.4	4.5	2.2
(max. no. starts 16)						

The results of the strategy that combines the hierarchical strategy with the percentage of loss strategy (strategy 9) are in Table 2.7. With this strategy a percentage value of zero does not guarantee finding the optimal R^2, because here a percentage value of zero gives the results of the hierarchical strategy (strategy 7). So, the minimum total number of local minima with the combined strategy is equal to the total number of local minima obtained with the hierarchical strategy. The number of local minima with the combined strategy is approximately equal to the sum of the number of local minima with the hierarchical strategy and the percentage of loss strategy. For example, in Table 2.7, 25 in the column $\geq 30\%$, is the sum of 23 (Table 2.5) and 2 (Table 2.6), and 16 in the same column is the sum of 14 (Table 2.5) and 3 (Table 2.6) minus 1. So, with the same percentage value, the combined strategy results in more local minima than the percentage of loss strategy, although the increase is mainly in the not severe local minima. But with the combined strategy the average number of starts is reduced more than with the percentage of loss strategy. For example, with the percentage of loss strategy using a percentage value of 60, the total number of local minima is 54, requiring 13 starts on average. With the combined strategy, the same number of local minima is obtained using a percentage value between 50 and 40, but requiring between 8.5 and 8.9 starts on average, which is substantially less.

Since the maximum number of starts is only 32 in this study, the percentage of loss strategy would be the best choice here. But with more variables with ordinal or monotonic spline scaling level, the number of starts rapidly increases for the percentage of loss strategy. To give an impression: when q (the number of variables with ordinal or monotonic spline scaling level) is 15, and all variables have seven categories (in CATREG CPU time depends upon the number of categories, not the number of objects), the number of all starts is 32768, which requires 4 minutes on a 2.2 Ghz computer. When q is 20, the number of all starts is 1048576, requiring 4.5 hours, and when q is 21, the number of all starts is 2097152, requiring 11.5 hours.

When the number of predictor variables is large, the combined strategy might be preferable, since the maximum number of starts is less, and this number does not grow as fast as the percentage of loss strategy.

Effect of data conditions on incidence of local minima

To investigate the effect of the data conditions on the incidence of local minima, we counted the number of local minima and performed analysis of variance on the counts, treating the data conditions as random factors. Table 2.8 displays the results. The main effects, the two-way interaction effect between

Table 2.8. Incidence of local minima. Analysis of variance, $N = 5000$.

Source	F	Sig.
R^2	21.769	.000
No. Categories	5.414	.011
Multicollinearity	9.120	.008
R^2 × No. Categories	7.742	.000
R^2 × Multicollinearity	1.421	.190
No. Categories × Multicollinearity	0.785	.694
R^2 × No. Categories × Multicollinearity	2.365	.000

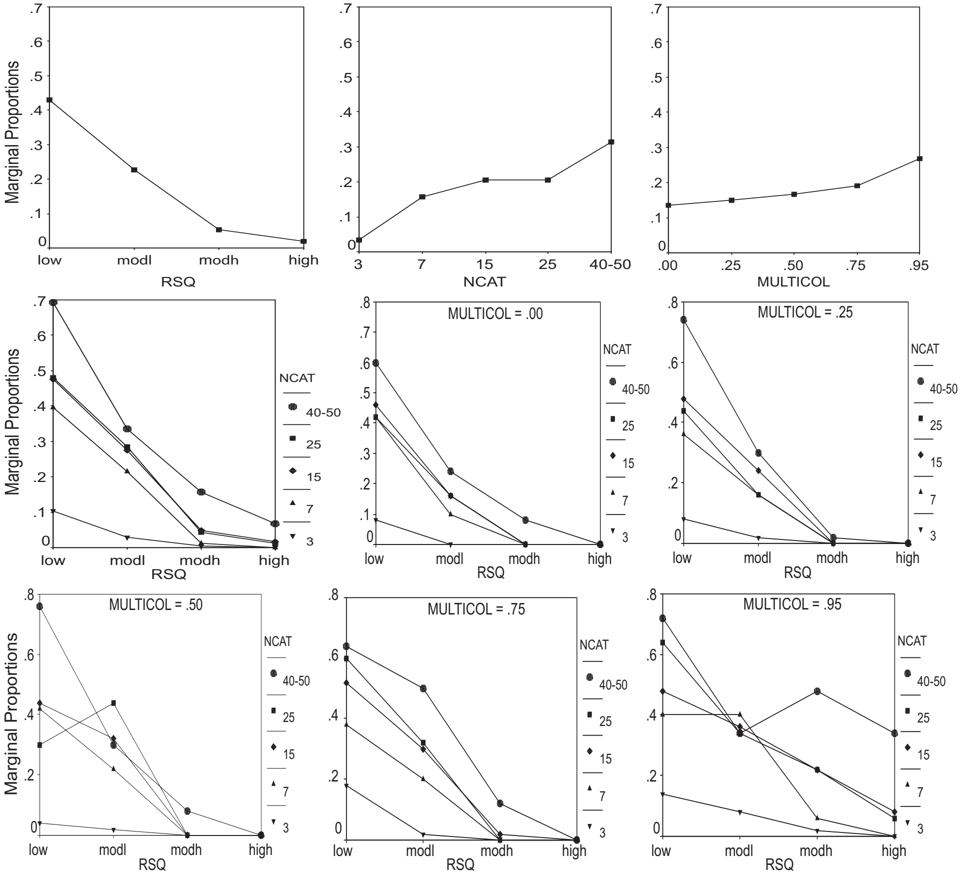

Figure 2.2. Incidence of local minima. Main effects for R^2, number of categories, and multicollinearity (row 1), interaction between R^2 and number of categories (row 2, column 1), and three-way interaction (row 2, columns 2 and 3, and row 3).

R^2 and the number of categories, and the three-way interaction effect are significant. In Figure 2.2 the main effects and the interaction between R^2 and the number of categories have been displayed graphically. We see that local minima occur more often when R^2 is low or moderately low, more often with higher number of categories, and also with higher levels of multicollinearity. In the two-way interaction plot we notice that with three categories, local minima are rare and that they almost only occur when R^2 is low or moderately low. When R^2 is moderately high or high, local minima seldomly occur with three and seven categories. When R^2 is high, local minima only seem to occur when the number of categories is in the range 40–50.

Effect of data conditions on severeness of local minima

To investigate the effect of the data conditions on the severeness of local minima, analysis of variance was performed on the 912 local minima that resulted from the numerical start, with the output variable the difference between the optimal R^2 and the suboptimal R^2 resulting from the numerical start. In this analysis the design is unbalanced and there are empty cells, which has been accounted for in the type of sum of squares used in the analysis. The results are in Table 2.9.

The main effects, the two-way interaction effects between R^2 and the number of categories, and between the number of categories and the multicollinearity, are significant. In Figure 2.3 the main effects and these two-way interactions are displayed. We see that local minima are more severe with lower R^2, with higher number of categories, and with higher multicollinearity. In the plot of the interaction between R^2 and the number of categories, we see that with 40–50 and 25 categories the severeness of local minima is about equal in the low and moderately low R^2 conditions, and we see a drop in severeness going from the moderately low to the moderately high R^2 condition that is greater than that for the other number of categories. We saw earlier that in the high R^2 condition, local minima almost only occur with 40–50 categories (Figure 2.2); here we see that these local minima are not very severe. The plot of the interaction between number-of-categories and multicollinearity shows that in the highest multicollinearity condition there is a high increase in severeness going from the 25 to the 40–50 categories condition. In the .75 and .25 multicollinearity conditions, there is a high increase in severeness going from the 15 to the 25 categories condition. In contrast, in the .50 and .00 multicollinearity conditions, the increase in severeness levels off when going from the 15 to the 25 categories condition.

Table 2.9. Severeness of local minima. Analysis of variance, $N = 912$.

Source	F	Sig.
R^2	13.889	.009
No. Categories	5.016	.010
Multicollinearity	5.944	.003
R^2 × No. Categories	2.626	.025
R^2 × Multicollinearity	1.369	.286
No. Categories × Multicollinearity	6.872	.000
R^2 × No. Categories × Multicollinearity	0.354	.991

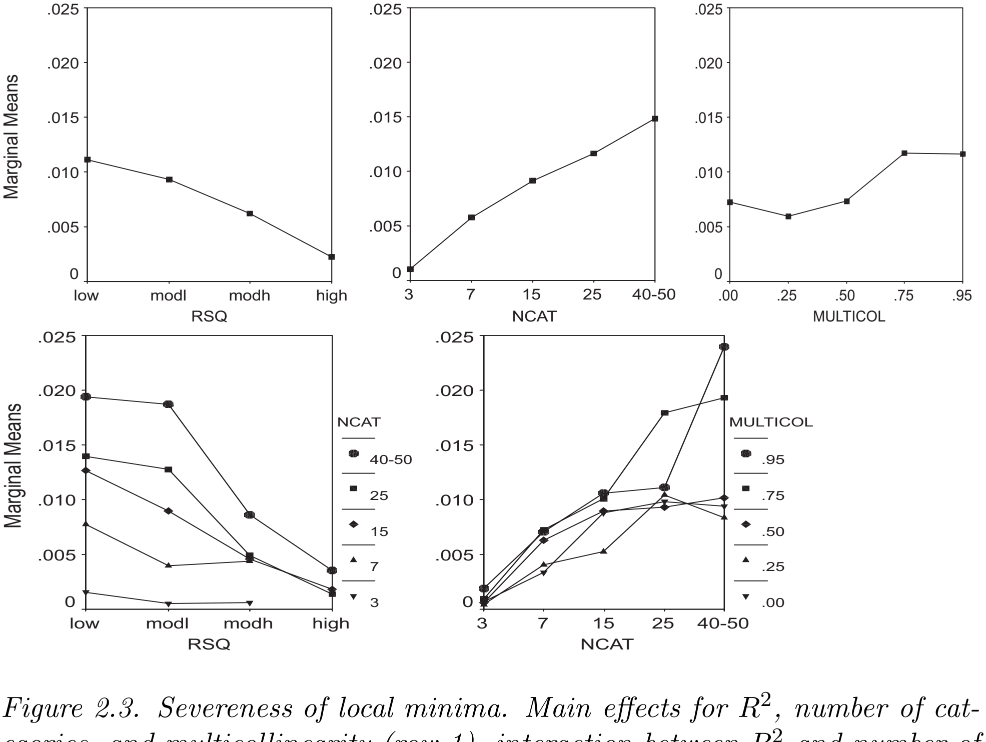

Figure 2.3. Severeness of local minima. Main effects for R^2, number of categories, and multicollinearity (row 1), interaction between R^2 and number of categories (row 2, column 1), and interaction between number of categories and multicollinearity (row 2, column 2).

2.6. Conclusion

In this chapter we detailed the CATREG algorithm for multiple regression with optimal scaling. Monotonic (spline) transformations may lead to suboptimal solutions. We have done a simulation study to investigate the effect of particular data conditions on the incidence and severeness of these local minima. In this study we found that local minima more often occur with low to moderately low R^2 values, with higher number of categories and with higher multicollinearity. We identified the cause of the occurrence of local minima to be the undefined direction of the monotonic restriction, which is equivalent to the sign of nominal quantifications being undefined, and we developed a strategy using multiple systematic starts to obtain the global minimum. Since this strategy requires a number of starts that is a power of two, and since complete enumeration is not feasible with a large number of predictors, we also developed a) several strategies that attempt to find the optimal signs in a more direct way, and b) three strategies that attempt to find the optimal signs by using a reduced number of multiple systematic starts. The simulation study was also used to asses the performance of these strategies. The reduced multiple systematic starts strategies proved to work well. Both the hierarchical strategy and the percentage of loss strategy considerably diminish both the incidence and the severeness of local minima. The strategy that combines these two strategies has both the advantage of the flexibility of the percentage of loss strategy, and the advantage of a smaller maximum number of starts, while at the same time this number does not grow very fast.