RR 92-10

M-ESTIMATORS IN MULTIPLE REGRESSION WITH OPTIMAL SCALING

Peter Verboon

Department of Data Theory

University of Leiden

This research was supported by a PSYCHON grant (560-267-029) of the Netherlands organization for scientific research (NWO).

M-ESTIMATORS IN MULTIPLE REGRESSION WITH OPTIMAL SCALING

Abstract

In the linear regression problem M-estimators are frequently applied as an alternative to least squares estimators to obtain robustness when there are outliers in the dependent variable. When we allow for nonlinear transformations of the variables (Young, 1981), the effects of outliers can still be large. In this paper it is shown that M-estimators can also be used for nonlinear multiple regression. Two types of M-estimators are used: the Huber and biweight estimator.

Permutation tests are applied to obtain significance levels for the multiple correlation coefficients. Furthermore, an exploratory procedure is proposed, which is useful in finding out which tuning constant is optimal for the given problem. In a Monte Carlo study the stability of the different solutions and the influence of the outliers is examined by using the jack-knife. It was found that the biweight function was the most robust and eliminated the influence of the outliers.

Key words: multiple regression, nonlinear, optimal scaling, M-estimators, robustness, outliers.

1. Introduction

After the problem of estimating a location parameter in an univariate distribution, the regression problem is, without doubt, the most thoroughly studied topic in the context robustness and resistance. Why is the classical regression problem, dealing

with the prediction of a dependent or criterion variable by one or more independent or predictor variables, so popular with people, who are concerned with robustness? The main reason is, of course, because the technique of regression analysis is among the most important ones in statistics and data analysis. Furthermore, the unwanted effects of using a non-robust criterion, such as least squares, to fit the regression model on contaminated data can quite easily be visualized and understood. Moreover, robust techniques clearly outperform the classical least squares approach in these situations, which is not necessarily true for other data analyses techniques.

Most research on robust regression has been concerned with the linear regression case. In this paper robustness aspects of the *nonlinear* multiple regression problem are studied. By nonlinear we mean that variables are used which do not necessarily have numerical measurement levels, but can have ordinal or nominal levels, and that these variables may be nonlinearly transformed. Therefore, in addition to the regression weights, we will also estimate the nonlinear transformation functions for the variables. Throughout this paper the variables are also assumed to be discrete with a limited number of categories, which implies that finding transformation functions boils down to replacing the original category values by new values, called *quantifications*. This process is called optimal scaling (Young, 1981; Gifi, 1990). The computed quantifications are optimal in the sense that they optimize the multiple correlation coefficient. Optimal scaling is applied to both dependent and independent variables.

In the present paper we will examine how outliers affect the least squares results in nonlinear multiple regression. It will be shown that M-estimators can also be used to fit the nonlinear regression model and in a simulation study the performance of the M-estimators will be compared with least squares for two statistics: the weighted multiple correlation coefficient and a robust measure of association based on sums and differences of rescaled variables. It is illustrated how permutation tests can be used to obtain significance levels for the statistics. Furthermore, some attention is paid to optimally choosing the so-called tuning constants of the robust functions. Finally, the stability of the solutions is compared and influential points are inspected.

2. Robustness in Linear Regression

Consider the classical linear regression model:

$$\mathbf{y} = \mathbf{Zb} + \mathbf{r}, \tag{1}$$

where $\mathbf{y}$ (order n) is the dependent or criterion variable, $\mathbf{Z}$ a matrix with p independent or predictor variables and $\mathbf{b}$ contains the p regression parameters. The error or residuals are represented by the vector $\mathbf{r}$. The classical least squares approach minimizes the sum of squared residuals $\mathbf{r}$.

With the development of robust regression procedures, three main classes of problems can be distinguished. The first and oldest problem is concerned with extreme observations in the directions of the criterion variable, which are called outliers. Since the residuals, which are the deviations from the fitted regression model, are always measured in the direction of the criterion variable, we may also say that this class of problems deals with the effects of large residuals. It is well known that large residuals have a great impact upon the ordinary least squares criterion, for which reason alternative loss function have been developed. The most important ones of these are the so-called M-estimators, which minimize the following function

$$\min \sum_{i=1}^{n} \rho(r_i), \tag{2}$$

where r_i are the residuals and $\rho(.)$ a function corresponding with some M-estimator, such as the Huber or the biweight. An overview of M-estimators can be found in Goodall (1983). These estimators are based on loss functions that have a higher tolerance towards large residuals, which implies that outliers in the criterion variable have less influence upon the solution, in other words they bound the influence of residuals. The name M-estimator for these functions stems from the fact that they are *Maximum Likelihood* estimators. The least squares function is also a M-estimator.

In addition to the first class, the second class of problems deals with what is sometimes called the influence of position. Here the situation is considered in which there are extreme values in the space of the predictor variables. The classical way to determine the influence of a point in the predictor space is to examine the diagonal

elements of the *hat* matrix (see Cook & Weisberg, 1982) or to look at the Mahalanobis distances in the predictor space or robust versions of these distances. Points with large Mahalanobis distances are called *leverage* points: that is, they lever the regression plane. For this reason, the effect of leverage points can be very dramatic, for instance, see Hawkins *et al.* (1984). Leverage points may be viewed as either good or bad, depending on whether they fit the structure of the majority of the points or not. Since M-estimators are not capable of bounding the influence of leverage points, new methods have been developed, such as the Mallows and Schweppe type estimators, which are called *generalized* M-estimators. The key idea in these proposals is to attribute weights to the observations, where the weights are defined by some decreasing function of the robust Mahalanobis distances between the observations in the predictor space and the centre of this space. The Mallows type estimator down weights leverage points regardless of the magnitude of the corresponding residual, whereas the Schweppe type estimator only down weights leverage points if the corresponding residual is large. The Mallows proposal has the following form

$$\min \sum_{i=1}^{n} \rho(r_i)\omega(\mathbf{z}_i), \tag{3}$$

where $\mathbf{z}_i$ is the ith row of $\mathbf{Z}$, $\omega(.)$ is some decreasing function of the distances in the predictor space and $\rho(.)$ as before, a function corresponding with some M-estimator. The Schweppe proposal is

$$\min \sum_{i=1}^{n} \rho\left(\frac{r_i}{\omega(\mathbf{z}_i)}\right) \omega(\mathbf{z}_i). \tag{4}$$

Since the Schweppe estimator takes into account the good leverage points, it leads to better efficiency compared to the Mallows estimator.

In Figure 1 a schematic overview is given on how the different functions affect the observations. The figure illustrates that M-estimators bound the influence of the residuals, which is indicated by the dashed lines; the Mallows estimator bounds, in addition, the influence in the predictor space, independent of the fact how well the extreme points fit the regression line. The curved line in the Schweppe type estimator illustrates the fact that only the influence of badly fitting leverage points is bounded.

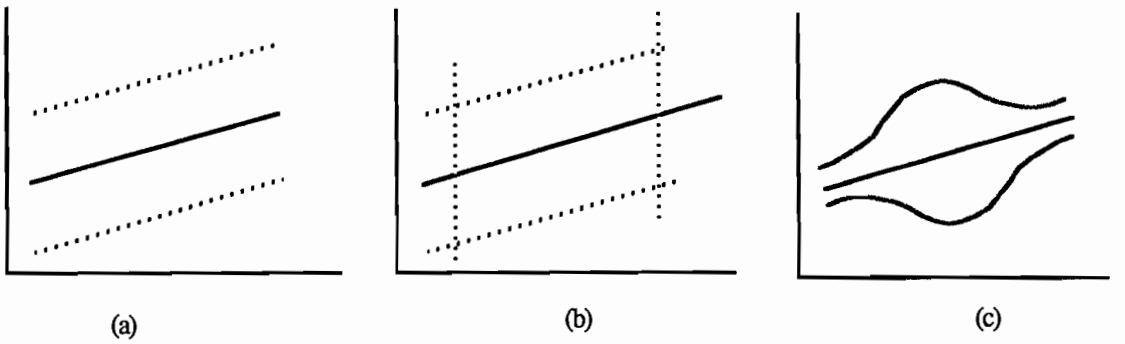

Figure 1. *Schematic overview of M-estimators (a), Mallows-type (b) and Schweppe-type (c) in simple regression.*

An important drawback of all functions, discussed so far, is that their *breakdown* point depends on the number of predictor variables. The breakdown point of an estimator is defined as the smallest percentage of contamination in the data that may cause the estimator to take on values, which are arbitrarily far away from the "true" estimator. With true estimator we mean the value that is based on the majority of the 'good points'. It is very unfortunate that the breakdown point of M-estimators increases with increasing dimensionality of the predictor space, since outliers are more likely to occur in an high dimensional space than in a low dimensional one.

To deal with problems of high percentage of contamination, the so-called high-breakdown regression estimates have been proposed. As one of the most important high-breakdown methods we mention the least median of squares (Rousseeuw, 1984; Rousseeuw & Leroy, 1987), which is defined by

$$\min \text{med}_i \, (r_i)^2. \tag{5}$$

Instead of replacing the square sign, as with the M-estimators, the least median of squares replaces the summation operation by the taking the median. The breakdown point of this estimator is 50%, meaning that up to half of the points in the data may be outliers without affecting the results. The method is computationally very intensive, which can be seen as a serious drawback. Recently, some improved algorithms are proposed (Rousseeuw, 1992), which should make the method computationally more efficient. Another drawback, mentioned by Hettmansperger and Sheather (1992), is

that under certain conditions the method can be very sensitive to small changes in the data.

In this section only the most important functions are mentioned, just to glance over the broad field of robust regression. For more discussion on this subject, we refer to Hampel *et al.* (1986), Huber (1981), Li (1985) and the literature already mentioned.

3. Nonlinear Regression

To generalize the linear regression model with nonlinear extensions, we consider residuals, which are defined as

$$r_i = f(y_i) - \sum_{j=1}^{p} g_j(z_{ij}), \tag{6}$$

where f(.) and $g_j(.)$ are transformation functions, which satisfy the measurement levels of the variables and are chosen to maximize the multiple correlation coefficient. To simplify notation, we define $q_i = f(y_i)$ and $x_{ij} = g_j(z_{ij})$, which yields a vector $\mathbf{q}$, representing the transformed criterion variable and a matrix $\mathbf{X}$ (nxp) with transformed predictor variables. In the least squares framework, minimizing these residuals, has been described in Young *et al.* (1976). For each variable an unrestricted update is computed which is then transformed such that it satisfies the measurement level of this variable. The restricted updates are alternatingly computed until the solution converges. In order to avoid degenerations (for instance, all values equal to zero, which would give a perfect fit), the transformed criterion variable is centered and normalized as $\mathbf{q}'\mathbf{q} = n$. During the algorithm, there is no need to estimate slope parameters, since these slopes are incorporated in the lengths of the predictor variables. After convergence, one could also normalize the predictor variables to n, and use the reciprocal of the normalization factor as slope estimates, given in $\mathbf{b}$.

It is of some interest to study whether robustness is also important for the nonlinear regression problem. The categorical character of the data assumes that there are no extreme values possible in the original codings of both criterion and predictor variables. In the first place, the range of the categories is known and bounded, so that it is easily checked whether an observations does not fall into one of the categories.

Secondly, even if there were an extreme observation, which is not in the range of the categories - for instance, typing errors, which have not been checked out - the use of nominal or ordinal measurement levels might reduce the damage.

However, by using the classical set of admissible transformations functions, it is not guaranteed that leverage points and extremes in the criterion variable will vanish. In fact, there is some potential danger, that leverage points are induced by the transformations. Thus, the original coding may be quantified with an extreme value. It is well known that optimal scaling may easily yield extreme quantifications, especially when a few categories have a low frequency of occurrence. In this situation extreme values in the criterion variable, and what is even more concerning, extreme values in the predictor space, leverage points, are likely to occur. In the present study we try to circumvent this problem by taking care that all categories are well filled. In the practice of data analysis, however, badly filled categories frequently occur and may cause real problems. Problems of this kind are usually solved by manipulating the data, such as merging several categories to obtain a smaller number of well filled categories or, even more drastic, to eliminate complete points (objects) which cause the problems.

To overcome these problems without manipulating the data, one could apply Schweppe and Mallows estimators for the nonlinear case. Another approach is to restrict the set of admissible transformations for the predictor variables. In this situation one could restrict the values of $\mathbf{X}$ to be in a specific range, that is, the transformations are bounded. If M-estimators are applied, bounding the transformations of the criterion variable should not be necessary, since extreme values in the criterion variable would lead to large residuals, which can be dealt with by M-estimators.

Although we assume that there are no extreme values in the variables, there is still the possibility that there are points that do not fit the overall structure in the data, which will yield large residuals, assuming the majority of the data is well fitted. A solution based on the least squares criterion may also be influenced by this type of outliers. Like in the linear regression case, we could replace the least squares criterion by one of the M-estimators. This would guard against outliers in the residuals.

There are several optimization strategies for M-estimators, one of which is iteratively reweighted least squares (IRLS) (Holland & Welsch, 1977; Verboon, 1990). Using IRLS, actually gives us the so-called W-estimators, which are equal to M-estimators when the solution is unique. In case of multiple solutions, which may occur with redescending M-estimators, such as the biweight, different optimalization strategies may yield different solutions and the IRLS solution is just one of them. Theoretically, the extension with nonlinear transformations, will increase the possibility of multiple solutions.

In the present study, the Huber and biweight loss functions will be used for the nonlinear regression problem. These functions will be minimized by applying an IRLS algorithm. For some fixed set of weights, gathered in a diagonal matrix **V**, we minimize

$$(\mathbf{q}, \mathbf{X}) = (\mathbf{q} - \mathbf{X})'\mathbf{V}(\mathbf{q} - \mathbf{X}) \tag{7}$$

with **q** and **X** restricted by their corresponding measurement levels and under the normalization constraints $\mathbf{1}'\mathbf{q} = 0$ and $\mathbf{q}'\mathbf{q} = n$. The weighted least squares problem, given in (7) is solved analogously to the unweighted least squares solution, since the diagonal weights matrix **V** causes no additional problems. Only the constraints on **q** cause an additional problem, because these are defined in the metric **I**, while the minimization problem is defined in the metric **V**. A procedure to solve this problem is proposed in Heiser (1987). After (7) is minimized, the residuals are computed according to (6) and with these residuals, new weights are obtained (cf. Verboon, 1990). With these weights (7) is minimized again and these steps are repeated until the solution convergences. After convergence, small weights are assigned to residuals corresponding to outliers, while the good fitting points have weights equal or nearly equal to one. It follows that these weights can be used as diagnostics for badly fitting points.

4. An illustrative example

For a generated data set we compared the two robust M-estimators, Huber (HUB) and biweight (BIW), with ordinary least squares (LSQ) for the regression problem

with optimal scaling. To reduce the probability to converge to local minima for BIW, the optimal values for $\mathbf{X}$ from the HUB solution were taken as starting values for BIW. In the comparison between the three criteria the interest is in the influence of the outliers upon two statistics. The first statistic is the squared weighted multiple correlation defined as

$$R^2 = 1 - \frac{\mathbf{r}'\mathbf{V}\mathbf{r}}{\mathbf{q}'\mathbf{V}\mathbf{q}}, \tag{8}$$

where $\mathbf{r}$ contains the residuals and the diagonal matrix $\mathbf{V}$ contains the weights as computed by (6) or (7). If the technique has succeeded in down weighting possible outliers, then R^2 will obviously be less affected by these outliers than the unweighted multiple correlation. The second statistic, denoted as S, is a robust measure of association, introduced by Gnanadesikan and Kettenring (1972). To compute S we start standardizing q and $\hat{\mathbf{q}}$ (the prediction of $\mathbf{q}$) in a robust way. Next, we compute the sum and difference vector of these two standardized vectors. With the robust spread measures of these sums (σ_s) and differences (σ_d) S is computed as

$$S = \frac{\sigma_s - \sigma_d}{\sigma_s + \sigma_d}. \tag{9}$$

The median absolute deviation (MAD) was used as the robust spread measure. In Devlin *et al.* (1975) the association measure S was studied among other measures of association and appeared to have good robustness properties.

Data

The data used for this illustration consisted of one criterion and three predictor variables for forty objects. The three predictors and an error variable consisted of randomly generated values, taken from a uniform distribution. The criterion variable was constructed with these random variables by the equality: $\mathbf{y} = 4\mathbf{x}_1 + 2\mathbf{x}_2 + \mathbf{x}_3 + \mathbf{e}$. Both the criterion and the predictor variables were divided into five categories with equal frequencies. For all variables, we assumed ordinal measurement restrictions. Thus, the optimal scaling assigns new values (quantifications) to the original coding, in such a way that the quantifications are monotonic with the original coding.

Analyzing these data with LSQ yielded R^2 = .917, S = .737, and a vector of regression parameters **b** = {.78, .35, .30}. We took these results as target values, against which all other results were compared.

Next, four objects (2,4,18,23) were randomly selected and turned into outliers by replacing their scores on the criterion variable by extreme ones (one or five), in such a way that the correlation decreased. In this way, we have constructed a data set with 10% outliers and about 10% random error (error vector plus rounding error).

Results

Applying LSQ to this contaminated data set, yielded fits of R^2 = .390 and S = .636, which are much smaller than the target values. The residuals from the LSQ analysis are given in Figure 2. The outliers appear to have somewhat larger residuals, but the picture is not very clear, because the distribution of the residuals is not very extreme, that is, the outliers do not really stick out. There are even a few good points (objects 16 and 37) with residuals as large as the outliers. Furthermore, it is clear that the outliers cause a dramatic drop in the weighted multiple correlation coefficient and a small decrease in the robust measure of association.

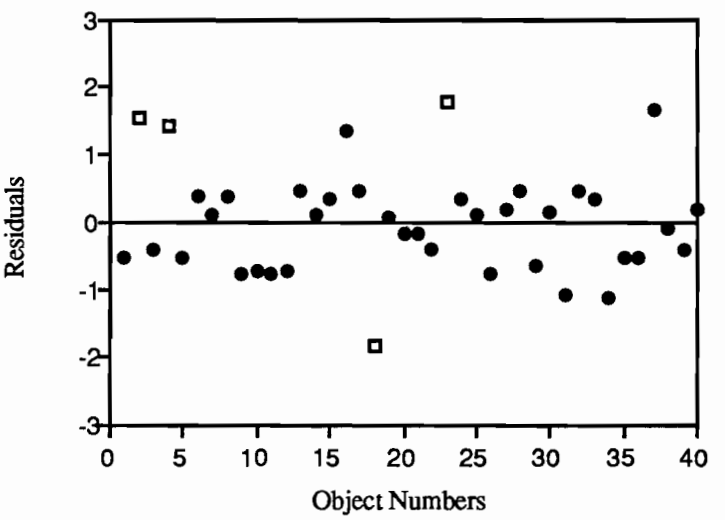

Figure 2. *Least squares residuals per object. Outliers are indicated with squares.*

The next step is to analyze these data using the Huber and biweight loss functions. For the Huber function we set the tuning constant equal to $2/3\sigma$, where σ is a spread measure defined as

$$\sigma = \text{MED}\,(|\mathbf{r}|) + 4\ \text{MAD}\,(|\mathbf{r}|). \tag{10}$$

The vector **r** contains residuals from the least squares analysis and the MAD is the median absolute deviation. With the Huber function the weighted multiple correlation improves to $R^2 = .546$, which is somewhere between the values from the contaminated and the uncontaminated solutions. The robust measure S = .752, which is even somewhat higher than the target value. The residuals from this analysis are given in Figure 3. There is a clear difference between this plot and the one for LSQ. The residuals for the outliers in HUB are larger than for LSQ, while the other objects have smaller residuals. The residuals of object 16 and 37 are now clearly distinguished from the outliers.

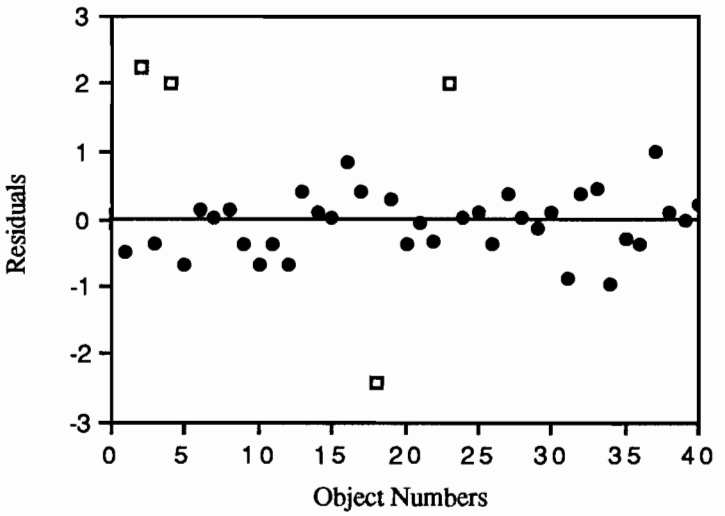

Figure 3. *Residuals per object from analysis with Huber function. Outliers are indicated with squares.*

For the biweight function the tuning constant was set to 2σ. The weighted multiple correlation is now very close to the target value: $R^2 = .857$, while S = .718. The plot of residuals (Figure 4) very clearly identifies the outliers, while the residuals for all

other objects are close to zero. In fact, the residuals of object 16 and 37 are now completely within the range of all other objects.

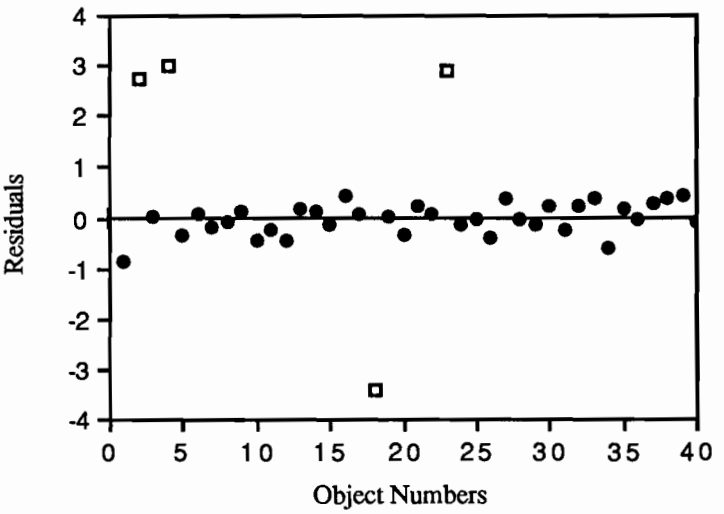

***Figure 4**. Residuals per object from analysis with biweight function. Outliers are indicated with squares.*

In Table 1 the results of all analyses are summarized.

Table 1

Summary of results nonlinear regression analyses.

Criterion	R^2	S	b_1	b_2	b_3	weights for outliers			
LSQ	.390 (.12)	.636 (.00)	.64	.10	.12				
HUB	.546 (.04)	.752 (.00)	.70	.21	.08	.49	.55	.46	.55
BIW	.857 (.11)	.718 (.00)	.81	.36	.29	.11	.03	.00	.05
target	.917 (.00)	.737 (.00)	.78	.35	.30				

The analyses using the robust functions yield superior results compared to the least squares analysis. The R^2 and S are closer to the target values and are significant at $\alpha < .05$, except for R^2 computed with BIW for which α is about .11. The

significance levels are obtained by permutation tests, which will be discussed in the next section. For the biweight the influence of the outliers is almost completely erased from the analysis, because all weights have become close to zero for the outliers. Weights assigned to good objects are all larger than .95. The improvements for the Huber function are more gradual, which can be expected because the function has been defined less radically than the biweight. The weights corresponding to the outliers are about .5, which means that the outliers contribute much less to the solution than with least squares, but that they still have some influence.

To study the effect of the outliers on the optimal scaling the category quantifications (transformations) obtained in the analysis without outliers are plotted against the transformations obtained in the three analyses of the contaminated data. In Figure 5 these plots are shown for the criterion and the three predictor variables. Outliers have no influence on the transformations, if the values are positioned approximately on the imaginary line y=x. In that case the obtained transformations are similar to the target transformations.

For the criterion variable we see that the BIW transformations are on the line y=x, while LSQ and HUB are slightly deviant from this line and thus from the target values. The same can be said about the first and most important predictor variable. In both pictures it is seen that HUB is doing a little better than LSQ. For the second and third predictor variable, the category quantifications for BIW are still close to the target quantifications, but the category quantifications for HUB and especially for LSQ are completely different from the target values.

We may conclude that the influence of the outliers is quite large on the quantifications obtained with the least squares function. Their influence is almost completely erased for the biweight, but the outliers still have some influence on the quantifications obtained with the Huber function.

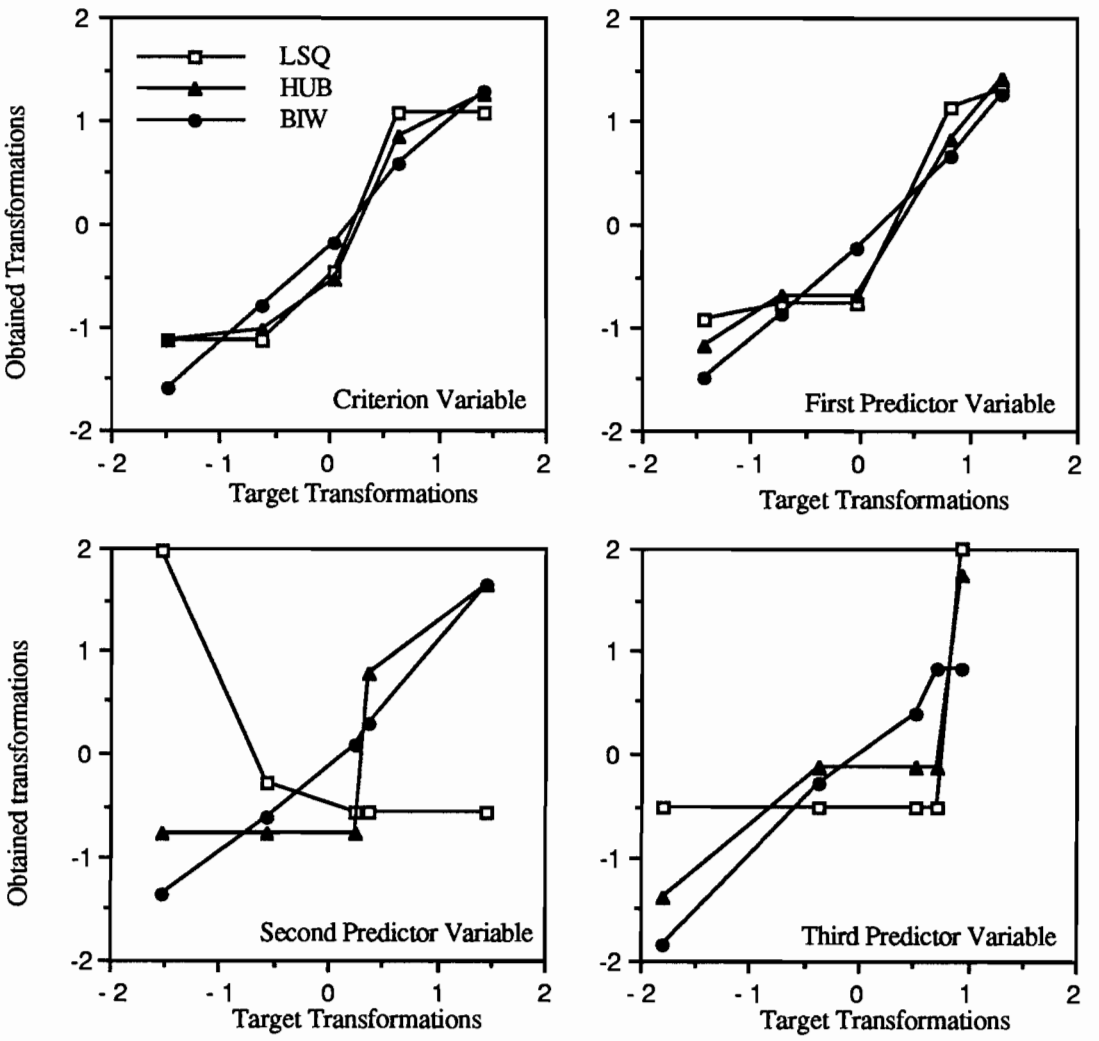

Figure 5. *Category quantifications of the four variables. Results from analysis without outliers (target) versus results from analyses with outliers for LSQ, HUB and BIW.*

Randomization

Since optimal scaling will always yield larger multiple correlations than analyses without optimal scaling, even when there is no structure at all, it is of some interest to obtain significance levels for the obtained estimates of the multiple correlations. To this end, we applied randomization tests (Edgington, 1987) for all criteria, using random data permutations (the number of permutations was set to 200). In each permutation a complete analysis was performed, meaning that if the correlations obtained after permutation are larger than zero, then this effect is completely due to random error and optimal scaling. The significance levels derived from these randomization tests are reported Table 1. The average values, obtained in the permutation tests, together with the standard deviations are presented in Table 2.

Table 2

Average values of R^2 and S with standard deviations obtained from 200 permutations.

Criterion	R^2	S
LSQ	.275 (.093)	.103 (.021)
HUB	.316 (.117)	.102 (.143)
BIW	.450 (.242)	.097 (.153)

Note that the biweight already yields rather high values of R^2, even with no structure at all in the data. The spread of the biweight results is also quite large. The S statistic seems to be somewhat more reliable, because its value is low with random data and is about the same for the three functions.

5. The choice of the tuning constant

In the results we have presented so far, differences have been studied between the three loss functions (criteria). But, in fact, the M-estimators represent a whole family of criteria, each member corresponding to one particular choice of the tuning constant. Of course, for a large number of tuning constants (for instance, the very large ones) the results will be exactly the same, but in some critical region in the set of admissible choices, we would like to choose the optimal tuning constant. A tuning constant is considered optimal if an analysis with this constant clearly distinguishes the outliers from the good points. In other words, small weights (ideally zero) should be assigned to outliers and large weights (ideally one) to good points. When the tuning constant is chosen too small, then good points will also be down weighted, on the other hand, when the tuning constant is chosen too large, then the solution largely resembles least squares and the outliers will be assigned weights that are not small enough to bound their influence. In the previous analyses we have chosen the tuning constant as a function of the spread of the least squares residuals. In an exploratory context this way of choosing the tuning constant seems to be the most natural approach (cf. Devlin *et al.*, 1981; Denby & Mallows, 1977).

In the linear regression case Denby and Mallows (1977) proposed to use diagnostic plots for studying the effect of the tuning constant. They plotted the residuals for each point and the regression coefficients as a function of the tuning constant. Such plots could give an indication about the optimal choice of the tuning constant. However, their study was restricted to the HUB criterion. An efficient algorithm is proposed by them, which quickly computes residuals belonging to a new tuning constant based on the results from another tuning constant, without redoing the complete analysis. Since we are dealing with optimal scaling, we have to repeat the analysis each time, for which reason we have selected only a limited number of tuning constants.

In the first place our interest is in the distribution of the weights in order to find an optimal tuning constant. The weights are a function of the residuals, which makes our approach related to the one in Denby and Mallows (1977). Additional permutations were performed for BIW(σ) and BIW(3σ), and the number of weights smaller than a

particular value were recorded. These results, given as cumulative percentages, can be presented as in Figure 6.

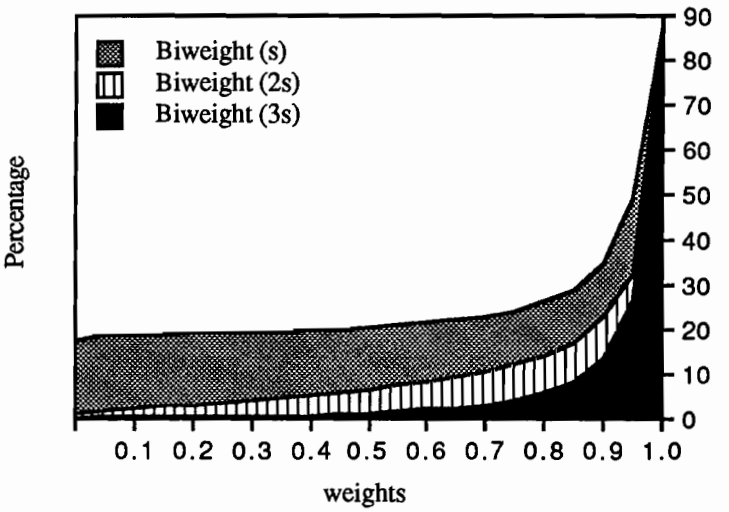

Figure 6. *Cumulative percentage weights for different tuning constants of BIW.*

The figure shows the distribution of the weights in unstructured data. This plot is useful for comparing the obtained distribution of the weights with the distribution under the no-structure condition. For instance, in the present analysis with BIW(2σ), we obtained four zero weights, while all others weights were larger than .95. Comparing this distribution with the one given in Figure 6 indicates that the computed distribution of weights is significantly different from the no-structure condition. The probability of getting a weight equal to zero is about 3%, which means that it is reasonable to assume that points, to which zero weights are assigned, are outliers. The figure also tells us that we should be cautious in interpreting small weights, when a small tuning constant has been used. When there is no structure in the data, about 30% of the objects will obtain weights smaller than .80 and 20% even smaller than .10 (for TC=σ and with these data). Based on Figure 6, a tuning constant of 2σ seems to be a good choice for the biweight, even if there is very little structure in the data.

In Figure 7 the percentage cumulative weights are shown for three tuning constants of the Huber function.

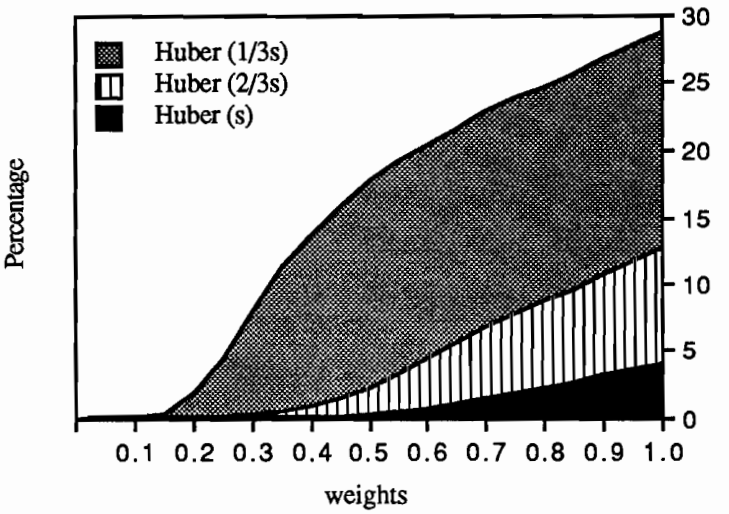

Figure 7. *Cumulative percentage weights for different tuning constants of HUB.*

In HUB(2/3σ) the obtained weights for the four outliers were all about .50. Figure 7 shows that the chance of obtaining weights smaller than .45 in the no-structure condition for this tuning constant is less than 3%. Based on this plot, for HUB(2/3σ) all weights corresponding to the outliers appear to be significant at 5% (see Table 1). Both figures bear the message that in data with little structure, many small weights will occur and that therefore the solution may become very unstable, when the tuning constant is relatively small. This should warn us against choosing the tuning constant too small. Of course, for large tuning constants, the robustness properties of the Huber and biweight will vanish.

In this example we knew the structure of the data and therefore we could easily decide which tuning constant satisfied the optimality criterion. With empirical data, however, we don't know what the true model is and therefore, it is less easy to choose an optimal tuning constant. A relatively small tuning constant will filter the effect of points with large residuals and increase the fit of the other points. However, the total weighted sum of squares will be smaller, which means that a smaller part of the original data is accounted for. In comparing two different tuning constants, both the multiple correlation coefficient, and the sum of squares should be considered. The smaller tuning constant will increase the multiple correlation, but usually decrease the

weighted sums of squares. To estimate the optimal tuning constant (TC), the following measure of difference D, is defined:

$$D = \frac{R^2(a) - R^2(b)}{SSQ(b) - SSQ(a)}, \qquad (SSQ(a) \neq SSQ(b)) \qquad (11)$$

where $R^2(a)$ is the multiple correlation and SSQ(a) the weighted sums of squares, both computed for TC = *a*. If D is small, then the gain in explained variance is small, compared to the loss in weighted sums of squares, when the tuning constant is changed from *b* to *a*. On the other hand, a large D value indicates that the gain in explained variance is large, at the cost of a relatively small amount of weighted sums of squares. In Figure 8, the D measure is plotted for the biweight function, applied to the same data set.

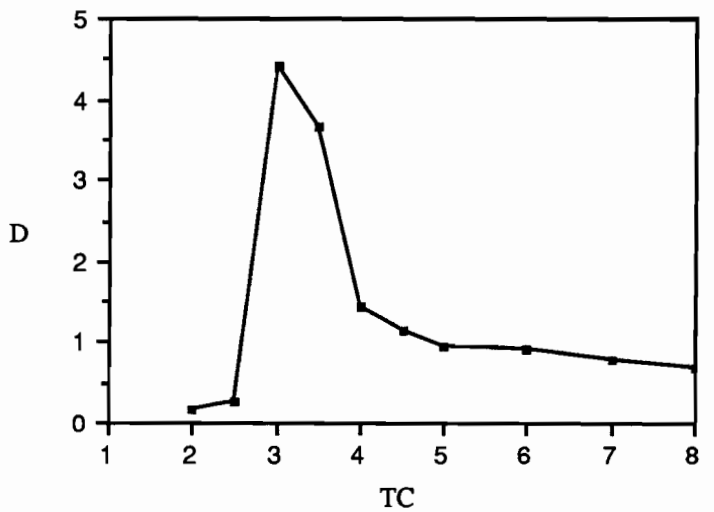

***Figure 8.** D measure for different tuning constants in BIW.*

The figure should be read from right to left. For instance, when we change the TC from 5 to 4.5, D becomes 1.15, which is relatively small. Changing the TC from 4 to 3.5, causes a large increase in D (3.7), which means that that taking TC=3.5, instead of TC=4, highly improves the solution. The figure shows that a tuning constant of 3 or 3.5 will be the best choice for analyzing this data, which corresponds to the value that we used, since for these data $\sigma = 1.66$. For tuning constants smaller than 2, we found degenerate solutions.

For any practical use we recommend to choose the tuning constant as we did in this example and to check whether a better choice is possible by applying the procedure given above. This way of obtaining an optimal value for the tuning constant resembles some of the adaptive procedures discussed in Hogg (1974).

6. A Monte Carlo Study

The stability of the solution and the influence of individual objects upon the multiple correlation can be studied by performing a jack-knife (Miller, 1974). The interest is in the stability of the M-estimators (including least squares) in the nonlinear regression context, and whether outliers are less influential for the M-estimators than for least squares with respect to the two statistics R^2 and S. Thus far, we have only used one single data set to demonstrate several aspects of M-estimators in nonlinear regression. To study the generality of the results a Monte Carlo experiment was executed with varying levels of random error (ε) and a varying number of predictor variables (p). There were two error levels: 10% and 25%, and two levels of p: p=1 and p=3. To decrease possible effects due to a specific data set the jack-knife will be applied in an experiment with replications.

Procedure

Four large data pools were generated; two of them consisted of scores on four variables and the other two of scores on two variables, with each variable containing seven categories. One of the variables was taken as the criterion variable and the remaining one or three as the predictors. The pools were generated such that they contained 10% and 25% error, respectively for each level of p. In each condition sets were randomly drawn from the structured data pool, with a random number of objects (n_t) in each draw under the restriction $\{25 \leq n_t \leq 50\}$. This yields in each condition sets $\mathbf{Z}_t$ (t=1,...,T) of order n_t x (p+1), were T is the number of replications, chosen to be 5. Next, each set was contaminated with three outliers, which yielded contamination percentages between 6% and 12%. The outliers were constructed by replacing the scores of three randomly selected rows with new scores which consisted of series of

ones and sevens, chosen in such a way that they did not fit the linear model of the majority of the objects. In analyzing all data sets ordinal measurement restrictions for the variables were used.

Results

Before and after the outliers were added to the data, each $\mathbf{Z}_t$ was analysed with the three loss functions. The weighted multiple correlations computed in these analyses were averaged over the five replications and are given in Table 3.

Table 3

R^2 averaged over 5 replications.

		p=1		*p*=3	
	outliers	ε = .10	ε = .25	ε = .10	ε = .25
LSQ	no	.964	.748	.961	.846
	yes	.611	.417	.534	.487
HUB	no	.991	.880	.961	.848
	yes	.957	.768	.675	.636
BIW	no	.999	.964	.963	.856
	yes	.999	.958	.921	.797

With three predictors and no outliers HUB and BIW give about the same results as LSQ, but with one predictor BIW and HUB yield somewhat higher values than LSQ. When outliers are added the BIW results are only slightly affected, while LSQ is clearly the most affected. Obviously, more random error yields smaller correlations for all functions. The number of predictors does not seem to have a consistent effect upon R^2. In general, more predictors yield somewhat smaller R^2 values.

In Table 4 results are given for the association measure S.

Table 4

S averaged over 5 replications

		p=1		p=3	
	outliers	ε = .10	ε = .25	ε = .10	ε = .25
LSQ	no	.669	.519	.712	.663
	yes	.000	.067	.301	.275
HUB	no	.605	.473	.712	.659
	yes	.000	.103	.411	.276
BIW	no	.497	.454	.713	.665
	yes	.134	.345	.548	.623

Without outliers and with three predictors S is about the same for the three functions. With one predictor BIW has the smallest S and LSQ the highest. The outliers affect S for all functions, that is, S decreases when outliers are added. For LSQ this effect is very large, while for BIW the effect is smallest. The use of one predictor decreases S compared to three predictors. It is peculiar that with one predictor S is smaller with less random error. In fact S becomes even zero for HUB and LSQ, while the weighted multiple correlation was still rather high under this condition.

Table 5

Average values R^2 with average standard deviations obtained in the Jackknife

	p=1		p=3	
	ε = .10	ε = .25	ε = .10	ε = .25
LSQ	.617 (.034)	.417 (.033)	.544 (.045)	.484 (.054)
HUB	.964 (.014)	.726 (.084)	.678 (.060)	.625 (.078)
BIW	.993 (.036)	.938 (.060)	.903 (.081)	.860 (.082)

The jack-knife results (Table 5) show that under all conditions the weighted multiple correlation, averaged over the *n* jack-knifes and over the five replications, is close to the values from Table 3. The standard deviations were computed as the spread of the R^2 in the jack-knife analyses, which were then averaged over the five replications. LSQ has the smallest standard deviations and BIW the largest. More error increases the standard deviations and one predictor yields generally more stable results than three predictors.

The values of the statistic S (Table 6) are much smaller than R^2. For S the results are consistent with Table 4. Again BIW appears to be the least stable criterion. More error and one predictor usually leads to less stable results for S.

Table 6

Average values S with average standard deviations obtained in the Jackknife

	p=1		*p*=3	
	ε = .10	ε = .25	ε = .10	ε = .25
LSQ	.000 (.000)	.103 (.095)	.287 (.116)	.250 (.121)
HUB	.010 (.034)	.104 (.051)	.332 (.148)	.269 (.118)
BIW	.361 (.084)	.344 (.138)	.593 (.227)	.428 (.172)

In Table 7 the average weights are given. For the outliers the average of all weights was computed, which was also done for all other objects. For both functions the outliers are clearly distinguished from the other objects, because the weights are much smaller. In general, the weights are somewhat smaller when there is much error, except for the weights assigned to the outliers when there is one predictor.

Table 7

Loss weights averaged over outliers and other objects, respectively

		p=1		p=3	
		ε = .10	ε = .25	ε = .10	ε = .25
HUB	others	.946	.919	.947	.948
	outliers	.075	.304	.567	.497
BIW	others	.944	.898	.936	.924
	outliers	.009	.044	.182	.114

Table 8

Influence on R^2 averaged over outliers and other objects, respectively

		p=1		p=3	
		ε = .10	ε = .25	ε = .10	ε = .25
LSQ	others	.012	.019	.026	.029
	outliers	.100	.085	.050	.064
HUB	others	.009	.069	.036	.043
	outliers	.020	.087	.090	.116
BIW	others	.005	.047	.056	.065
	outliers	.069	.060	.041	.043

To study the influence of individual objects on the two statistics an influence measure IFL was computed. For R^2 it is defined as

$$\mathrm{IFL}(i,t) = R^2_{(i)t} - \frac{\sum_{i=1}^{n_t} R^2_{(i)t}}{n_t} . \qquad (12)$$

where $R^2_{(i)t}$ is the squared weighted multiple correlation computed in the tth replication with the ith object omitted. The absolute values of IFL(i,t) are averaged over i and t. This was done separately for the three outliers and for the (n_t - 3) other objects. Results given in Table 8 show that for LSQ and HUB the influence of the outliers is larger than of the other objects. For BIW the outliers are more influential than the others for p=1 and less influential than the others for p=3.

To study the influence upon S an influence measure was computed analogously to (12). Table 9 shows that for LSQ the outliers are less influential with respect to S than the other objects, which may be expected with a robust measure. For BIW and HUB the outliers have more influence upon S than the other objects, which is explained by the fact that S is computed with the unweighted variables. When the weights are small for the outliers, their quantifications may take on any value as long as the measurement restrictions are satisfied. For LSQ with one predictor and ε=.10 the influence is zero, which could also be predicted from Table 6. Under this condition S is always zero, while R^2 is rather large. This discrepancy is caused by peculiar transformations.

Table 9

Influence on S averaged over outliers and other objects, respectively

		p=1		p=3	
		ε = .10	ε = .25	ε = .10	ε = .25
LSQ	others	.000	.068	.091	.079
	outliers	.000	.048	.092	.070
HUB	others	.005	.024	.098	.084
	outliers	.055	.065	.131	.157
BIW	others	.008	.108	.155	.139
	outliers	.096	.097	.216	.155

7. Conclusions and Discussion

In the paper it is shown that like in ordinary linear regression, M-estimators are also useful in nonlinear multiple regression. They bound the influence of large residuals and increase the explained proportion of variance compared to least squares. In the present study, tuning constants of 2σ for the biweight and $(2/3)\sigma$ for the Huber function seem to yield fairly good results, where σ was chosen as a robust spread measure of the least squares residuals. The outliers were detected and in case of the biweight their influence was almost completely erased. Furthermore, permutation tests showed that the obtained weighted multiple correlations and the robust association measures were highly significant.

A diagnostic plot was proposed for finding an optimal tuning constant. The proportion of explained weighted variance was compared with the weighted sum of squares. The idea behind this plot is that the weighted sums of squares should not become too small. In the present example this approach clearly indicated the optimal tuning constant for BIW. Denby and Mallows (1977) have proposed a somewhat different method for studying the effect of the tuning constant. They use diagnostic plots of residuals and regression coefficients as a function of the tuning constant. Such plots may reveal at which value of the tuning constant the pattern of coefficients or residuals changes.

In the Monte Carlo study it was found that random error affected all three loss functions equally strong. With one predictor there were sometimes large differences between S and R^2 for both least squares and Huber. This is due to very unsmooth quantifications for both predictor and criterion variable in which the seven categories were transformed to only two or three distinct values. Such peculiar transformations may lead to very different values for S and R^2.

It seems that the least squares results are the most stable and the biweight results the least stable, but this conclusion needs some nuance. The large stability of least squares is mainly due to the large influence of the three outliers. When omitting an outlier in the jack-knife, the other two outliers still dominate the solution. Thus, least squares is stable at the wrong solution.

For the majority of the data sets the biweight was highly stable at the good solution, with approximately zero weights for the outliers. However, for a few data sets the variance of the jack-knife results was very large. Repeating these analyses with different starting values showed that these large variances could at least partly be explained by the presence of local minima. This instability occurred more often when the data sets were relatively small. If we do not use the average of the standard deviations over the five replications, but take the median instead, then the biweight appears to be equally stable as least squares.

For all practical use we should be aware of local minima in applying the biweight, especially with small data sets and little structure. Using different starting values may occasionally improve the solution. Further research may be useful to see whether more refined optimization methods have something to offer here.

The association measure S is robust against extreme values in one of the variables. However, in this study we considered outliers which did not fit the model of the majority of the points; these outliers had no a-priori extreme values. The measure S appeared to be sensitive to this type of outliers. Since S is computed with the unweighted variables, it is rather peculiar that S has been less affected by the outliers under the biweight than under least squares. This may be explained by the fact that the computed quantifications under the biweight yield relatively high correlations. After transformation the outliers will be more extreme for the biweight than for least squares, but these extremes are ignored by assigning small weights to them. The same extreme values will also be ignored by S, which therefore yields higher values than for least squares.

In this study outliers in ordinal data were defined as points with scores in the extreme categories, which did not fit the linear model of the majority of the points. It was provided that the extreme categories were sufficiently well filled, so that the quantification of such a category could not become unduly large. The optimal quantifications computed for least squares changed under influence of the outliers, but this change did not make the influence upon the multiple correlation disappear. It follows that optimal scaling is not able to bound the influence of this type of outliers.

With the biweight the quantifications were approximately as in the uncontaminated situation, because the outliers were down weighted.

8. References

Cook, R. D. & Weisberg, S. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

Denby, L. & Mallows, C.L. (1977). Two diagnostic displays for robust regression analysis. *Technometrics, 19*, 1-13.

Devlin, S. J., Gnanadesikan, R. & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika, 62, 3*, 531-545.

Devlin, S. J., Gnanadesikan, R. & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76,* 354-362.

Edgington, E.S. (1987). *Randomization tests. 2nd Ed.* New York: Dekker.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

Gnanadesikan, R. & Kettenring, J.R. (1972). Robust estimates, residuals and outlier detection in discriminant analysis. *Biometrics, 28*, 81-124.

Goodall, C. (1983). M-estimators of location: an outline of the theory. In: D.C. Hoaglin, F. Mosteller & J.W. Tukey (eds.), *Understanding robust and exploratory data analysis.* New York: Wiley.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics: the approach based on influence functions.* New York: Wiley.

Hawkins, D. M., Bradu, D. & Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics, 26*, 3, p. 197-208.

Heiser, W.J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5*, 337-356.

Hettmansperger T.P. & Sheather S.J. (1992). A cautionary note on the method of least median of squares. *The American Statistician, 46*, 79-83.

Hogg, R.T. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69,* 909-927.

Holland, P. W. & Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Comm. in Statistics, A6,* 813-827.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Li, G. (1985). Robust Regression. In D.C. Hoaglin, F. Mosteller & J.W. Tukey (Eds), *Exploring data tables, trends and shapes,* pp. 281-341. New York: Wiley.

Miller, R.G. (1974). The jack-knife - a review. *Biometrika, 61,* 1-15.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79,* 871-880.

Rousseeuw, P.J. (1992). Computational aspects of certain robust estimators. *Proceedings workshop "Data Analysis and Robustness",* Ascona, Switzerland.

Rousseeuw, PJ. & Leroy, A.M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Verboon, P. (1990). *Majorization with iteratively reweighted least squares: a general approach to optimize a class of resistant loss functions.* Research Report 90-07. Leiden: Department of Data Theory.

Young, F.W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46,* 347-388.

Young, F.W., De Leeuw, J. & Takane, Y. (1976). Regression with qualitative and quantitative variables: an alternative least squares method with optimal scaling features. *Psychometrika, 41,* 505-528.