Robust Diagnostic Data Analysis: Transformations in Regression

Marco RIANI

Dipartimento di Economia
Sezione di Statistica
Università di Parma
Italy
(*mriani@unipr.it*)

Anthony C. ATKINSON

Department of Statistics
London School of Economics
London WC2A 2AE
United Kingdom
(*a.c.atkinson@lse.ac.uk*)

We introduce a very general "forward" method of data analysis that starts from a small, robustly chosen subset of the data and shows the effect of adding observations by a forward search. Powerful diagnostic procedures result: The observations are ordered by their agreement with the proposed transformation, masking is overcome, and the inferential effect of each observation is clear. We apply the resulting method to the transformation of both univariate and multivariate data. Other applications of the forward search are mentioned.

KEY WORDS: Box–Cox transformation; Deletion diagnostic; Forward search; Least median of squares; Masking; Multivariate normality; Score statistic; Transformation to normality; Very robust methods.

We introduce a powerful new robust diagnostic method for the analysis of data. In contrast to a "backward" method, which fits a model to all data and then deletes suspicious observations, we use a "forward" method, which starts from a small, robustly chosen subset of the data and monitors the effect of adding observations to the subset until finally all the data are fitted. We thereby avoid the masking effect of multiple outliers that can defeat backward methods.

The method is clearly very general. In Section 8, we mention applications to linear and nonlinear regression, generalized linear models, and multivariate analysis. Here we focus on transformation of the response in regression.

Our forward search provides an ordering of the data from that most in agreement with the proposed transformation to that furthest from it. The ordering is then used to understand the contribution of each observation to the estimated transformation. Because observations that appear as outlying in untransformed data may not be outlying once the data have been transformed, and vice versa, we employ the forward search on data subject to various transformations, as well as on untransformed data.

We emphasize exploratory methods so that cogent graphical summaries of our analyses are at a premium. In this way our approach derives from that of Tukey (1977). Like Cook and Weisberg (1994), however, many of the quantities we plot come from standard statistical analyses. For example, a typical output from our analysis is a plot of the score statistic for transformation as we increase the number of observations used in fitting the model. The plot, which we call a "fan" plot, enables us to see the influence of individual observations, not just outliers, on the evidence for a transformation. We can then link the effect of each observation back to plots of the data, information that often enriches our interpretation of scatterplots of the original or transformed data. Even if no outliers are present, our procedure provides an elegant graphical summary of the evidence for a transformation.

We first consider univariate transformations and then multivariate ones. Section 1 briefly derives the score statistic for transformations. The forward search is defined in Section 2. We first discuss the inferential justification for the procedure and then consider in detail three aspects of the search—the starting point, the choice of the observations to be added, and the monitoring of important quantities during the search. The first example in Section 3 illustrates the null behavior of our procedure for the transformation of the response and shows how multiple masked outliers can indicate an incorrect transformation. The outliers are not revealed by single-deletion methods but are revealed by the forward search, which leads to the correct transformation. Sections 4 and 5 parallel Sections 1 and 2, describing the forward search for multivariate data using Mahalanobis distances and the transformation of such data. We then give an example of the transformation of multivariate data and one of regression with a multivariate response. Comment and further discussion are in Section 8.

1. TRANSFORMATIONS OF THE UNIVARIATE RESPONSE IN REGRESSION

For the linear regression model $E(y) = x^T\beta$, with $xp \times 1$, let n be the number of observations. Box and Cox (1964) analyzed the normalized power transformation

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \dot{y}^{\lambda-1}, & \lambda \neq 0 \\ \dot{y} \log y, & \lambda = 0, \end{cases} \quad (1)$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\sum \log y_i/n)$. If the observations are normally dis-

tributed with $R(\lambda)$ the residual sum of squares of the $z(\lambda)$, the profile log-likelihood of the observations, maximized over β and σ^2, is

$$L_{\max}(\lambda) = -(n/2)\log\{R(\lambda)/(n-p)\} \quad (2)$$

so that $\hat{\lambda}$ minimizes $R(\lambda)$. For inference about the transformation parameter λ, Box and Cox suggested the likelihood ratio test statistic

$$\begin{aligned} T_{\mathrm{LR}} &= 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_o)\} \\ &= n[\log\{R(\lambda_0)/R(\hat{\lambda})\}]. \end{aligned} \quad (3)$$

A disadvantage of this likelihood ratio test is that a numerical maximization is required to find the value of $\hat{\lambda}$. For regression models, a computationally simpler alternative test is the approximate score statistic derived by Taylor-series expansion of (1) as

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)w(\lambda_0), \quad (4)$$

where

$$w(\lambda_0) = \left.\frac{\partial z(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0}.$$

The combination of (4) and the regression model $y = x^T\beta + \varepsilon$ yields the model

$$\begin{aligned} z(\lambda_0) &= x^T\beta - (\lambda - \lambda_0)w(\lambda_0) + \varepsilon \\ &= x^T\beta + \gamma w(\lambda_0) + \varepsilon. \end{aligned} \quad (5)$$

Because (5) is again a regression model with an extra variable $w(\lambda_0)$ derived from the transformation, the new variable is called the constructed variable for the transformation. The approximate score statistic for testing the transformation, $T_p(\lambda_0)$, is the t statistic for regression on $w(\lambda_0)$ in (5). This can be calculated either directly from the regression in (5) or from the formulas of Atkinson (1985, chap. 6) in which multiple regression on x is adjusted for the inclusion of the constructed variable.

Similar ideas can be used for transformation of the explanatory variables (Box and Tidwell 1962). Constructed variables for the joint transformation of the response and of explanatory variables, together with score tests, were given by Atkinson (1985, sec. 8.4). Whatever the formal statistical outcome of such procedures, the interpretability of the results is also of importance in finding a sensible transformation.

2. THE FORWARD SEARCH

2.1 General Principles

Like most methods for outlier detection, our method divides the data into two parts, a larger "clean" part used for parameter estimation and the outliers. The simplest division is into one potential outlier and the rest of the data, leading to single-deletion diagnostics. Standard books on regression diagnostics, such as those of Cook and Weisberg (1982), Atkinson (1985), and Chatterjee and Hadi (1988), include formulas for multiple-deletion diagnostics, in which a small number, perhaps two or three, of potential outliers are considered at once. But there is a combinatorial explosion of the number of cases that have to be considered by such backward working. Here we employ very robust methods. In the resampling algorithms for least median of squares (LMS) regression (Rousseeuw 1984) and minimum volume ellipsoid estimation for multivariate data (Rousseeuw and van Zomeren 1990), the model is fitted to p observations, when the remaining $n-p$ observations can be tested to see if any outliers are present. The resulting parameter estimates are very robust but are defined by the algorithm that produces them, a crucial distinction with standard statistical methods such as least squares. For example, LMS estimates could be found by searching over larger subsamples, perhaps with $m = p+1$ or $p+2$. The disadvantage is that the probability of subsamples with outliers increases. Woodruff and Rocke (1994) showed, however, that such estimators, while remaining very robust, have lower variance than those based on smaller subsets. They are therefore more reliable when used in outlier-detection procedures.

In the forward search, such larger subsamples of outlier-free observations are found by starting from small subsets and moving to larger subsets containing observations that have small residuals and so are unlikely to be outliers. The method was introduced by Hadi (1992) for the detection of outliers from a fit using approximately half the observations. Different versions of the method were described by Hadi and Simonoff (1993, 1994) and Atkinson (1994). This literature emphasizes the use of the forward search to find a single set of parameter estimates and outliers, determined by the point at which the algorithm stops, which may be either deterministic or data dependent. The emphasis in our article is very different: At each stage of the forward search we use information such as parameter estimates and residual plots to guide us to a suitable model.

We use least squares to estimate the parameters from the selected subset of m observations, with m running from p to n. From these parameter estimates, we calculate a set of n residuals e_m. There will then be $n-m$ observations not used in fitting that may contain outliers. We do not seek to identify these outliers by a formal test but instead plot the residuals and the score statistic for transformations $T_p(\lambda)$, monitoring changes associated with the introduction of a particular observation into the subset m used for fitting. There are three particular consequences of this procedure:

1. *Stability.* In the absence of outliers and systematic departures from the model, the parameter estimates for each m are unbiased estimators of the same quantity. So both parameter estimates and residuals should remain approximately constant during the forward search.

2. *Ordering.* If there are k outliers and we start from a clean subset, the forward procedure will include these outliers toward the end of the search, usually in the last k steps. Until then, residual plots and parameter estimates will remain approximately constant. Figure 4, Section 3, is such a plot of residuals.

3. *Transformations.* Outliers in one transformed scale may not be outliers in another scale. If the data are analyzed using the wrong transformation, the k outliers may enter the search well before the end.

The forward-search algorithm has three steps—the choice of an initial subset, the forward-search process, and the monitoring of statistics during the search. In the following subsections we consider these three aspects separately.

2.2 Step 1: Choice of the Initial Subset

If the model contains p parameters, our forward-search algorithm starts with the selection of a subset of p units. Observations in this subset are intended to be outlier free, but it is enough that they contain no masked outliers. Let $Z = (X, y)$, where X is the $n \times p$ matrix of explanatory variables so that Z is $n \times (p+1)$. If n is moderate and $p \ll n$, the choice of initial subset can be performed by exhaustive enumeration of all $\binom{n}{p}$ distinct p-tuples $S^{(p)}_{i_1,\ldots,i_p} \equiv \{z_{i_1}, \ldots, z_{i_p}\}$, where $z^T_{i_1}$ is the i_1th row of Z for $1 \le i_1, \ldots, i_p \le n$ and $i_j \ne i_{j'}$. Specifically, let $\iota^T = [i_1, \ldots, i_p]$ and let $e_{i,S^{(p)}_\iota}$ be the least squares residual for unit i given observations in $S^{(p)}_\iota$. We take as our initial subset the p-tuple $S^{(p)}_*$, which satisfies

$$\tilde{e}^2_{[\text{med}],S^{(p)}_*} = \min_\iota [\tilde{e}^2_{[\text{med}],S^{(p)}_\iota}], \tag{6}$$

where $\tilde{e}^2_{[l],S^{(p)}_\iota}$ is the lth ordered squared residual among $\tilde{e}^2_{i,S^{(p)}_\iota}, i = 1, \ldots, n$,

$$\text{med} = [(n+p+1)/2], \tag{7}$$

and $[(n+p+1)/2]$ denotes the integer value of $(n+p+1)/2$. If $\binom{n}{p}$ is too large, we use instead some large number of samples—for example, 1,000. Criterion (6) provides an LMS method for regression models with independent errors (Rousseeuw 1984; Hawkins 1993). The breakdown point of this estimator is asymptotically 50%.

2.3 Step 2: Adding Observations During the Forward Search

Given a subset of dimension $m \ge p$, say $S^{(m)}_*$, the forward search moves to dimension $m+1$ by selecting the $m+1$ units with the smallest squared least squares residuals, the units being chosen by ordering all squared residuals $\tilde{e}^2_{i,S^{(m)}_*}, i = 1, \ldots, n$.

The forward-search estimator $\hat{\beta}_{\text{FS}}$ is the collection of least squares estimators in each step of the forward search; that is,

$$\hat{\beta}_{\text{FS}} = (\hat{\beta}^*_p, \ldots, \hat{\beta}^*_n), \tag{8}$$

with $\hat{\beta}^*_m$ the least squares estimator from subset $S^{(m)}_*$.

In most moves from m to $m+1$, just one new unit joins the subset. Two or more units may join $S^{(m)}_*$ as one or more leave, however, which sometimes happens at the beginning of a search, particularly with multivariate data, if the subset contains some unmasked outliers. Later in the search it occurs only when the search includes one unit that belongs to a cluster of outliers. At the next step, the remaining outliers in the cluster seem less outlying, so several may be included at once. Of course, several other units then have to leave the subset. Step 2 of the forward search is repeated until all units are included in the subset.

In this approach, we use a highly robust method and at the same time least squares (i.e., fully efficient) estimators. A consequence of the zero breakdown point of least squares estimators is that the introduction of atypical influential observations is signaled by sharp changes in the curves that monitor parameter estimates or other statistics at every step. We can thus analyze the inferential effect of the atypical units on aspects of the fitted model.

Initial Subset. The method is not sensitive to the method used to select an initial subset. For example, the LMS criterion (6) can be replaced by the least trimmed squares criterion in which the sum of the squares of the smallest half of the residuals, as defined by med in (7), is minimized. The search is often able to recover from a start that is not very robust due to the inclusion of unmasked outliers. A regression example was given by Atkinson and Mulira (1993). In Section 4, we use bivariate boxplots to provide an initial subset for multivariate data from which the grossest outliers are removed. The first few steps of the search are very active because potentially outlying observations are identified and removed. But the final, informative, third of the forward search is insensitive to the precise selection of the initial subset.

Residuals and the Search. Forward searches allowing for the variances of the residuals were employed by Hadi and Simonoff (1993) and by Atkinson (1994), who used studentized residuals, whereas we use raw residuals. Our comparisons show that, although the choice of residual has a slight effect on the forward search, particularly at the beginning, the search using raw residuals is more stable in that usually only one observation is added at a time rather than several being interchanged. For monitoring the effect of individual observations on statistics and parameter estimates, it is helpful to connect particular effects with particular observations. Both methods respond to a cluster of outliers with multiple exchanges.

2.4 Step 3: Monitoring the Search

The estimate of σ^2 does not remain constant during the forward search as observations that have small residuals are sequentially selected. Thus, even in the absence of outliers, the residual mean squared estimate $s^2_{S^{(m)}_*} < s^2_{S^{(n)}_*} = s^2$ for $m < n$. One useful plot monitors all residuals at each step of the forward search. Large values of the residuals among cases not in the subset indicate the presence of outliers. Because of the strong dependence of $s^2_{S^{(m)}_*}$ on m, we standardize all residuals by the final root mean squared estimate s.

3. EXAMPLES OF TRANSFORMATION OF THE RESPONSE

We show results from a single forward search from a carefully selected starting point. The alternative of using several searches from random starting points seems to yield similar results but is more cumbersome. More important is whether the selection of the subset and the forward search

are carried out on untransformed data or on each individual value of λ. Individual searches give appreciably clearer plots and are preferable.

We also have found that test statistics are more informative than parameter estimates: If the likelihood is flat, the estimates can vary widely without conveying any useful information about the transformation.

3.1 Poison Data

We begin with the poison data from Box and Cox (1964). The observations are time to death of animals in a 3×4 factorial experiment with four observations at each factor combination. Box and Cox suggested the reciprocal transformation ($\lambda = -1$) so that death rate, rather than survival time, has a simple structure.

Our analysis is based on five values of λ—$-1, -.5, 0, .5$, and 1. In all examples these values are sufficient to indicate a satisfactory transformation. The data are transformed and a starting point is found by LMS for each of five forward searches, which then proceed independently for each λ using the transformed data. In this example we found the five initial subsets by searching 500 subsets for each λ, although this detail does not affect our general results. Table 1 gives the last six observations to enter in each search, together with the ordering of the observations; observation 20 is the largest.

For $\lambda = .5$ and 1, the largest observations are the last to enter the subset used for fitting because they agree least with the model, whereas, for $\lambda = -1$, all the large observations enter earlier in the search than $m = 43$. If a correct transformation has been found, small and large observations should both enter the search throughout, including at the end, as they do here for $\lambda = -.5$.

Figure 1 is the fan plot of the approximate score statistic $T_p(\lambda)$ for each search as the subset size m increases. The central horizontal bands on the figure are at ± 2.58, containing 99% of a standard normal distribution. For data without outliers, the curves for the different values of λ fan out as they do here: If outliers are present, as they are in Figure 3, Section 3.2, the curves may cross several times. But the final order always has $\lambda = -1$ at the top and $\lambda = 1$ at the bottom. Initially, in Figure 1, for small subset sizes there is no evidence against any transformation. During the whole forward search there is never any evidence against either $\lambda = -1$ or $\lambda = -.5$ (for all the data $\hat{\lambda} = -.75$). The log transformation is also acceptable until the last four observations are included by the forward search. As the table shows, these include some of the largest observations in order. The plot shows how evidence against the log transformation depends critically on this last 8% of the data. Evidence that $\lambda \neq 1$ is spread throughout the data: Less than half of the observations are sufficient to indicate the need for some transformation. There are no jumps in this curve, just an increase in evidence against $\lambda = 1$ as each observation is introduced into the subset. The relative smoothness of the curves reflects the lack of outliers and exceptionally influential cases.

Table 1. Poison Data: Last Six Observations to Enter the Five Separate Searches and Numbers of Six Largest Observations

	λ					
m	-1	$-.5$	0	.5	1	Largest observations
43	27	44	14	43	28	13
44	28	37	28	28	43	15
45	37	28	37	14	17	17
46	44	8	17	17	14	42
47	11	20	20	42	42	14
48	8	42	42	20	20	20

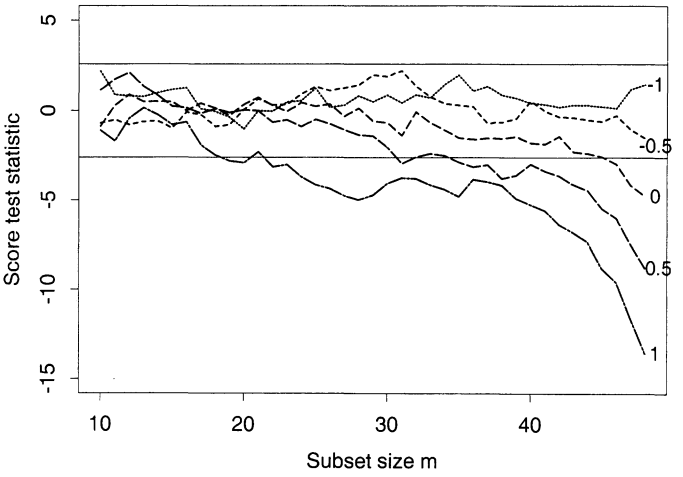

Figure 1. Poison Data: Fan Plot of Score Statistic for Power Transformation $T_p(\lambda_o)$ as the Subset Size m Increases. The parameter values are $\lambda = 1$, — · — · —; $\lambda = .5$, — — —; $\lambda = 0$, — — —; $\lambda = -.5$, - - - - -; $\lambda = -1$, · · · · · · · · · · . Both $\lambda = -.5$ and $\lambda = -1$ are acceptable.

3.2 Modified Poison Data: An Example of Masking

We now modify the poison data to create four masked outliers that are not revealed by single-deletion diagnostics and that indicate an incorrect transformation. The outliers and their influential effect are revealed by our forward analysis.

Table 2 shows how the four masked outliers were created by making four small observations smaller. These modifications should have little effect when the data are analyzed on the original scale but are very evident when $\lambda = -1$ and so influence the transformation away from -1 toward 1.

The maximum likelihood estimate $\hat{\lambda}$ is .274 for an additive model without interactions. To explore the flatness of the likelihood surface and to indicate a confidence region for λ, five values of the approximate score test for the transformation are

λ	-1	$-.5$	0	.5	1
$T_p(\lambda)$	22.08	10.01	2.87	-2.29	-8.41.

All transformations are rejected at the 5% level, although neither the log nor the square-root transformation are

Table 2. Modified Poison Data: The Four Modified Observations

Observation	Original	Modified
6	.29	.14
9	.22	.08
10	.21	.07
11	.18	.06

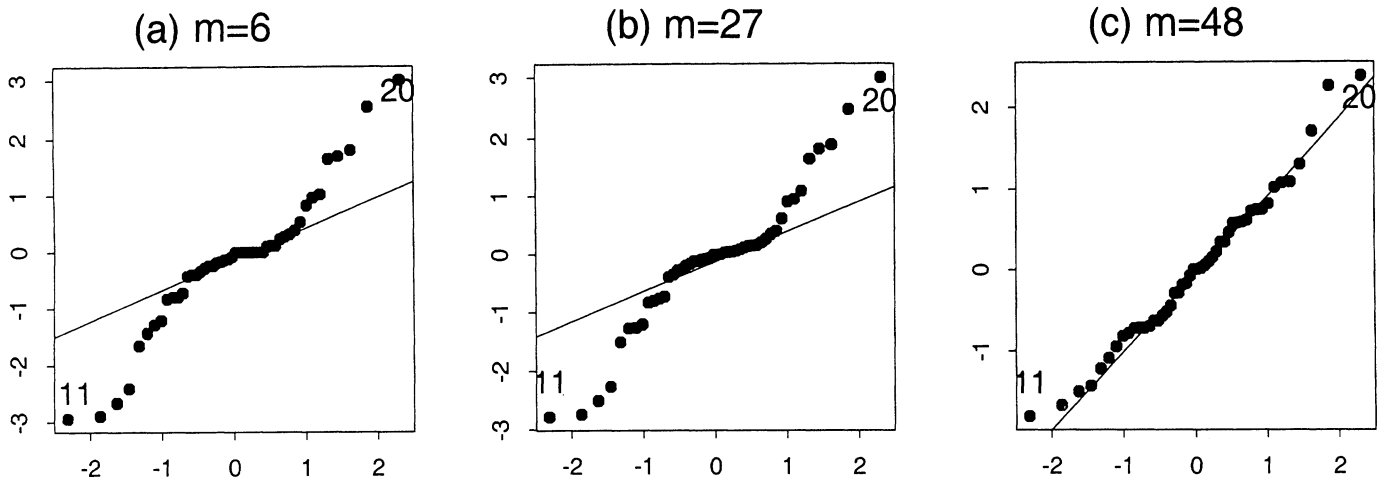

Figure 2. Modified Poison Data: QQ Plots of Scaled Residuals at Three Stages of the Forward Search with λ = 1/3: (a) m = 6, (b) m = 27 – "Half" the Data Used in Fitting, (c) m = 48 – Least Squares Residuals From Fitting All the Data.

strongly rejected. A value between them yields $T_p(1/3) = -.59$, making $1/3$ acceptable. The largest effect of an observation is from the deletion of observation 20; whether or not it is included, the data indicate the third root transformation, an unusual transformation, except for volumes. When $\lambda = .5$, deletion of observations 20 or 42 makes the square-root transformation acceptable, taking the analysis even further from the value appropriate to the majority of the data.

Figure 2 exhibits three quantile-quantile (QQ) plots of residuals for $\lambda = 1/3$. Figure 2(a) shows scaled residuals from an LMS fit to an elemental set, found by searching over 10,000 randomly selected subsets of size $m = 6$, the number of parameters in the linear model. The plot shows the typically long-tailed distribution from a very robust fit, as does Figure 2(b), where $m = 27$; both plots might suggest either many outliers or none. The least squares fit to all the data, Figure 2(c), however indicates no particular outliers. Thus we would conclude from this analysis that the $1/3$ transformation is reasonable on statistical grounds, although lacking a physical interpretation, and we have no indication of the effect of the four outliers. The example shows that, if data are analyzed on the wrong transformation scale, even the application of very robust methods such as LMS fails to highlight outliers and influential observations.

In contrast, Figure 3 is the fan plot of the score statistics for transformation. Instead of a series of curves that either remain horizontal or steadily diverge, as in Figure 1, some curves are within the bounds for most subsets and then increase rapidly at the end; others go outside the 1% boundary, only to return at the end. Both forms of behavior are associated with influential outliers.

For $\lambda = -1$, addition of the last four observations (the four outliers) causes a rapid increase in the value of the score statistic from 1.16 to 22.1, providing strong evidence against $\lambda = -1$. The behavior of the curve for $\lambda = -.5$ is similar but much less extreme: The four outliers are again included at the end. The curve for $\lambda = 0$ first goes below the boundary but then rises above the upper threshold when the four contaminated observations are included, again in the last four steps of the forward search. The statistic for $\lambda = .5$ lies on or below the boundary when $22 \leq m \leq 37$; the final value of $T_p(.5)$ is -2.29. In this scale, the four contaminated observations are not extreme: Three enter at $m = 38, 39$, and 40 and cause the appreciable upward jump in the statistic.

To confirm that $\lambda = -1$, we look at the plot of standardized residuals during the forward search for this value of λ. Figure 4 shows the four outliers, observations 6, 9, 10, and 11, that enter in the last four steps of the forward search. Until this point, the pattern of residuals remains remarkably constant, as Section 2.1 indicates. The pattern changes appreciably only in the last four or five steps, when the outliers and observation 8 are introduced.

The results of the forward search in Figures 3 and 4 clearly show the masked outliers and their effects, which were not revealed by single-case deletion methods nor by the residual plots for $\lambda = 1/3$ of Figure 2. We believe this comparison exhibits the power of our method. For the rest

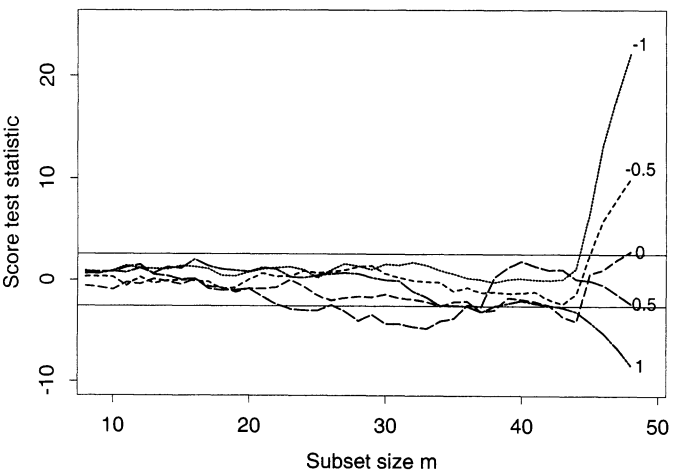

Figure 3. Modified Poison Data; Fan Plot of Score Statistic for Power Transformation $T_p(\lambda_o)$ as the Subset Size m Increases; Individual Searches for Each λ. The parameter values are λ = 1, — - — - —; λ = .5, — — —; λ = 0, — — — —; λ = −.5, - - - - -; λ = −1, ... The upward jumps result from the introduction of the masked outliers.

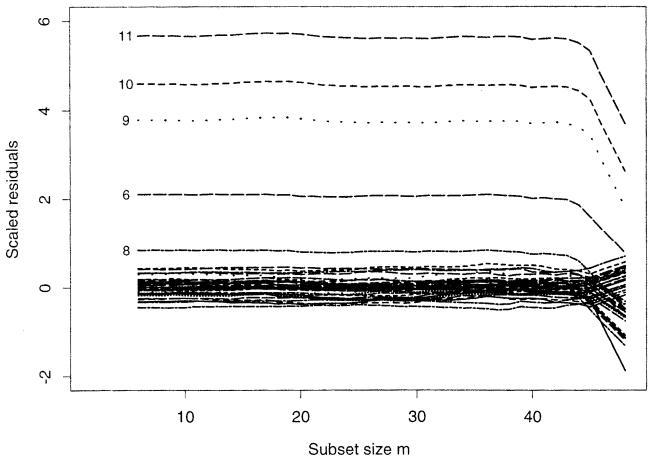

Figure 4. Modified Poison Data: Scaled Residuals During the Forward Search With $\lambda = -1$. The outliers are clearly visible and have constant residuals for most of the search. The changes in residuals in the last steps of the search are caused by the introduction of the outliers.

of the article, we consider examples with original, unmodified data so that the outliers are not known.

4. MULTIVARIATE OUTLIERS

The extension of the forward search of Section 2 to multivariate data replaces squared residuals with Mahalanobis distances based on residuals from the regression model, perhaps after transformation of the responses. For the complete data, let y_i^T be the ith of n observations on a v-variate response with normally distributed errors and let $\hat{\mu}_{i,S_*^{(n)}} = \hat{\mu}_i$ be the estimated response using all the observations. If there is no regression, $\hat{\mu}_{i,S_*^{(n)}} = \bar{y}_{S_*^{(n)}} = \bar{y}$, the vector of mean responses. The squared Mahalanobis distance for observation i is

$$d_i^2 = (y_i - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (y_i - \hat{\mu}_i) = e_i^T \hat{\Sigma}^{-1} e_i, \tag{9}$$

where $\hat{\Sigma} = \hat{\Sigma}_{S_*^{(n)}}$ is the sample covariance matrix, with

$$\begin{aligned} \hat{\Sigma}_{jk} &= \sum_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik})/(n-p) \\ &= \sum_i e_{ij} e_{ik}/(n-p), \end{aligned} \tag{10}$$

and p is the dimension of the vector of regression parameters. Asymptotically the d_i^2 follow a chi-squared distribution on v df. If $\hat{\mu}$ and $\hat{\Sigma}$ were not estimates but were known population parameters, outlying values of y_i would yield large values of the squared distance d_i^2. The effect of such values on the estimation of $\hat{\mu}$ and $\hat{\Sigma}$, however, leads to the rapid breakdown of the Mahalanobis distance for the detection of outliers, particularly if several outliers are present.

Now suppose that a subset $S^{(m)}_{i_1,\ldots,i_m}$ of m observations is used to estimate the regression and covariances. Let the estimates be $\hat{\mu}_{i,S_*^{(m)}}$ and $\hat{\Sigma}_{S_*^{(m)}}$, yielding the set of squared Mahalanobis distances

$$d^2_{i,S_*^{(m)}} = (y_i - \hat{\mu}_{i,S_*^{(m)}})^T \hat{\Sigma}^{-1}_{S_*^{(m)}} (y_i - \hat{\mu}_{i,S_*^{(m)}}). \tag{11}$$

Our forward search uses (11) in place of the least squares residuals of Section 2. For multivariate data, we find a larger initial subset than that for regression by again transforming the data, followed by robust analysis of the matrix of bivariate scatterplots, using the procedure of Zani, Riani, and Corbellini (1998). The initial subset consists of those observations that are not outlying on any scatterplot, found as the intersection of all points lying within a robust contour containing a specified proportion of the data. The size of the subset can be adjusted by changing the level of the contour. Examples of the boxplots were given by Atkinson and Riani (1997). The initial steps of the resulting forward search sometimes involve the removal and introduction of several observations as multivariate outliers are identified that are not outlying in the marginal two-dimensional plots. We again perform the forward search once for each vector λ.

5. MULTIVARIATE TRANSFORMATIONS TO NORMALITY

The parametric family of power transformations was extended to multivariate data by Andrews, Gnanadesikan, and Warner (1971) and by Gnanadesikan (1977). Given the difficulty of visualizing multivariate data, diagnostic methods for multivariate transformations are even more important than those for univariate data; yet few have been developed. Velilla (1993) compared marginal and joint transformations and gave further references to related work. Velilla (1995) developed deletion diagnostics and robust estimates of the transformation parameter, exemplified on simulated data without regression structure. As in the univariate transformation of Section 1, we treat the general case in which the means of the observations may have a regression structure.

Let y_{ij} be the ith observation on response j. In the extension of the Box and Cox (1964) family to multivariate responses, the normalized transformation of y_{ij} is

$$\begin{aligned} z_{ij}(\lambda_j) &= (y_{ij}^{\lambda_j} - 1)/\lambda_j \dot{y}_j^{\lambda_j - 1}, \quad (\lambda \neq 0) \\ &= \dot{y}_j \log y_{ij}, \qquad\qquad (\lambda = 0), \end{aligned} \tag{12}$$

where $\dot{y}_j$ is the geometric mean of the jth response. If the transformed observations are normally distributed with mean μ_i for the ith observation and covariance matrix Σ, twice the profile log-likelihood of the observations is given by

$$\begin{aligned} 2L_{\max}(\lambda) &= \text{const} - n \log |\hat{\Sigma}(\lambda)| \\ &\quad - \sum_{i=1}^{n} \{z_i(\lambda) - \hat{\mu}_i(\lambda)\}^T \hat{\Sigma}^{-1}(\lambda) \{z_i(\lambda) - \hat{\mu}_i(\lambda)\} \\ &= \text{const} - n \log |\hat{\Sigma}(\lambda)| \\ &\quad - \sum_{i=1}^{n} e_i(\lambda)^T \hat{\Sigma}^{-1}(\lambda) e_i(\lambda). \end{aligned} \tag{13}$$

In (13), $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ are derived from least squares estimates for fixed λ and $e_i(\lambda)$ is the $v \times 1$ vector of residuals.

The calculation of $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ is simplified when, as in the examples in this article, the matrix of explanatory variables X is the same for all responses. As a result, the

least squares estimates are found by independent regressions for each response, yielding the $p \times v$ matrix of parameter estimates $\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z(\lambda)$. Then, in the usual way,

$$(n-p)\hat{\Sigma}(\lambda) = \sum_{i=1}^{n} e_i(\lambda) e_i(\lambda)^T = \{z(\lambda) - X\hat{\beta}(\lambda)\}^T \{z(\lambda) - X\hat{\beta}(\lambda)\}. \quad (14)$$

When these estimates are substituted in (13), the profile log-likelihood reduces to

$$2L_{\max}(\lambda) = \text{const}' - n \log |\hat{\Sigma}(\lambda)|. \quad (15)$$

To test the hypothesis $\lambda = \lambda_0$, the statistic

$$T_{\text{LR}} = n \log\{|\hat{\Sigma}(\lambda_0)|/|\hat{\Sigma}(\hat{\lambda})|\} \quad (16)$$

is compared with the χ^2 distribution on v df, the generalization of the univariate statistic (3). In (16), $\hat{\lambda}$ is the vector of v parameter estimates maximizing (13), which is found by numerical search. As in Section 2, constructed variables can be used to define diagnostic tests that avoid numerical maximization of the likelihood.

For multivariate transformation, the regression model for the jth response when the constructed variable is included is

$$z_j(\lambda_0) = X\beta_j(\lambda_0) + w_j(\lambda_0)\gamma_j. \quad (17)$$

Testing that λ_0 is the correct transformation of the response is equivalent to testing that the γ_j in (17) are 0. Due to the presence of the constructed variables, the explanatory variables are no longer the same for all responses and the simplification of the regression used in the calculations for (13) no longer holds: The covariance Σ between the v responses has to be used in estimation, and ordinary least squares is replaced by generalized least squares. In the particular form (17), the parameters for each response are different and the estimates are related only through covariances of the y_j. This special structure, that of seemingly unrelated regression (Zellner 1962), was used by Atkinson (1995) to obtain deletion diagnostics for multivariate transformations. We do not explore these methods here because some of the simplicity of deletion diagnostics is lost when straightforward regression can no longer be used. Instead we use the likelihood ratio statistic (16), noting that analogues of the statistic $T_p(\lambda)$ could also be combined with the forward search to provide information about the effect of outliers on multivariate transformations.

6. MULTIVARIATE DATA

Soil Data

Mulira (1992) gave 57 readings on five properties of soil samples [listed by Atkinson and Riani (1997)]—two measurements of pH, which are highly correlated, and three measures of available phosphorus, potassium, and magnesium, whose marginal distributions show appreciable skewness.

The maximum likelihood estimates of the five transformation parameters are $\hat{\lambda}_1 = -.24, \hat{\lambda}_2 = -.03, \hat{\lambda}_3 = .01, \hat{\lambda}_4 = -.86$, and $\hat{\lambda}_5 = -.25$. The null hypothesis of no transformation yields a value of 140.1 for the likelihood ratio statistic (16), so the data should be transformed. For the hypothesis of the log transformation for all five variables, the corresponding value is 12.1, compared with 11.07 for the 95% point of χ^2 on 5 df. If λ_4 is taken as -1, with all other $\lambda_j = 0$, the statistic has the value 2.1. Because the values of pH are already the logarithms of hydrogen ion concentration, it seems surprising that they should be logged again. The same transformation of pH is indicated in the analysis of data on the acidity of lakes by Richardson and Green (1997), following earlier lognormal analyses. The test for the hypothesis $\lambda = (1, 1, 0, -1, 0)$, however, is 3.9. To confirm that this conclusion is not dependent on outliers, we use a forward search on untransformed data, monitoring likelihood ratio statistics (16) for three values of the λ vector. Figure 5 shows that the statistic for testing no transformation of y_1 and y_2 and the log transformation of the other three variables—that is, $\lambda = (1, 1, 0, 0, 0)$—is above the 95% point of the χ^2_5 distribution from $m = 37$ and around the 99% point of the distribution (15.09) from $m = 39$. The statistics for $\lambda = (1, 1, 0, -1, 0)$ and $(0, 0, 0, -1, 0)$, parameter vectors that are based on the reciprocal transformation for y_4 but differ in whether or not y_1 and y_2 are transformed, are also shown in the figure. Overall the figure shows that the reciprocal transformation for y_4 deserves further investigation but that transformation of the pH variables is not important.

The preliminary analysis also indicates the reciprocal transformation for just one of y_3, y_4, and y_5. Because they are all concentrations of chemical elements, we might expect them all to have the same transformation. Five parameters each with five values of λ would require $5^5 = 3,125$ forward searches, so we consider only plausible combinations of the λ from the vector $\lambda = (1, 1, 0, -1, 0)$ and thus replace the 5^5 factorial search by a one-variable-at-a-time search with five factors (the λ's) each at five levels. We test the transformation by use of the likelihood ratio for each

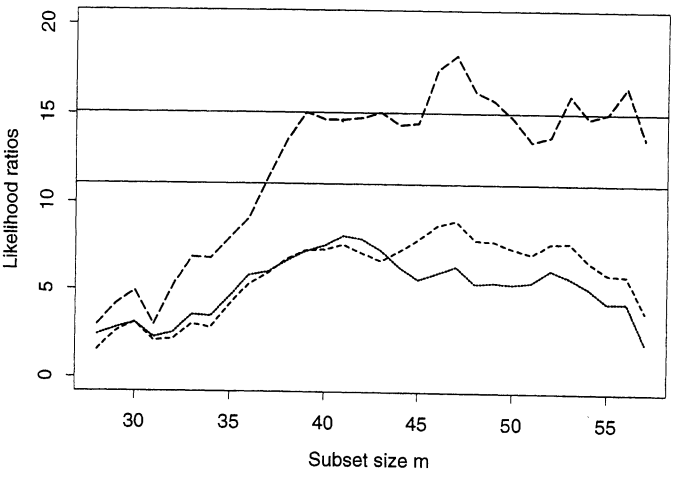

Figure 5. Soil Data: Likelihood Ratio Tests for Three Transformations: λ = (1, 1, 0, 0, 0), — — — —; λ = (1, 1, 0, −1, 0), - - - - - -; λ = (0, 0, 0, −1, 0), ·····. The horizontal lines are the 95% and the 99% points of the χ^2_5 distribution.

value of λ_k, which is on 1 df. Calculation of the likelihood ratio statistic for each value of λ_k in the forward search requires a numerical maximization. The fan plots of the 25 forward searches are given in Figure 6, in which the five panels for the five variables plot the signed square root of the likelihood ratio statistic for the usual five values of λ to indicate whether lower or higher values of λ are preferred.

Figure 6 shows that the value of 1 is acceptable for λ_1 and λ_2 and, as the smoothness of the curves indicates, does not depend on any particular observations. Only the log transformation is possible for variable 3; this forward search ends with the addition of either observation 24 or 20, which, respectively, cause rejection of $\lambda = -.5$ and $\lambda = .5$, the values on either side of 0. For variable 5, either $-.5$ or 0 are possible and, for variable 4, either -1 or $-.5$. These plots show that we cannot find a common transformation for y_3, y_4, and y_5 for m greater than 53. The four observations to be deleted to achieve this are 19, 20, 24, and 55, the last four to be added, in various orders, in all searches leading to acceptable transformations. The source of these observations should thus be checked for anomalies and transcription errors. Whether or not they are deleted, $\lambda = 1$ is acceptable for y_1 and y_2.

The forward searches through the data identified a set of four influential observations. For comparison, the repeated application of backward deletion diagnostics presented only one influential observation. This example illustrates how our method provides a coherent approach to the identification of multiple outliers and their influence on transformations.

7. MULTIVARIATE MULTIPLE REGRESSION DATA

Dyestuffs Data

The data, arising in a study of dyestuffs manufacture, are taken from Box and Draper (1987, pp. 114–115). There are 64 observations at the points of a 2^6 factorial and three responses—strength, hue, and brightness. Box and Draper found that only three of the six variables have a significant effect on the three responses. Their plots of residuals (p. 123) arguably indicate that y_2 should be transformed, but no transformation of either y_1 or y_3 was suggested.

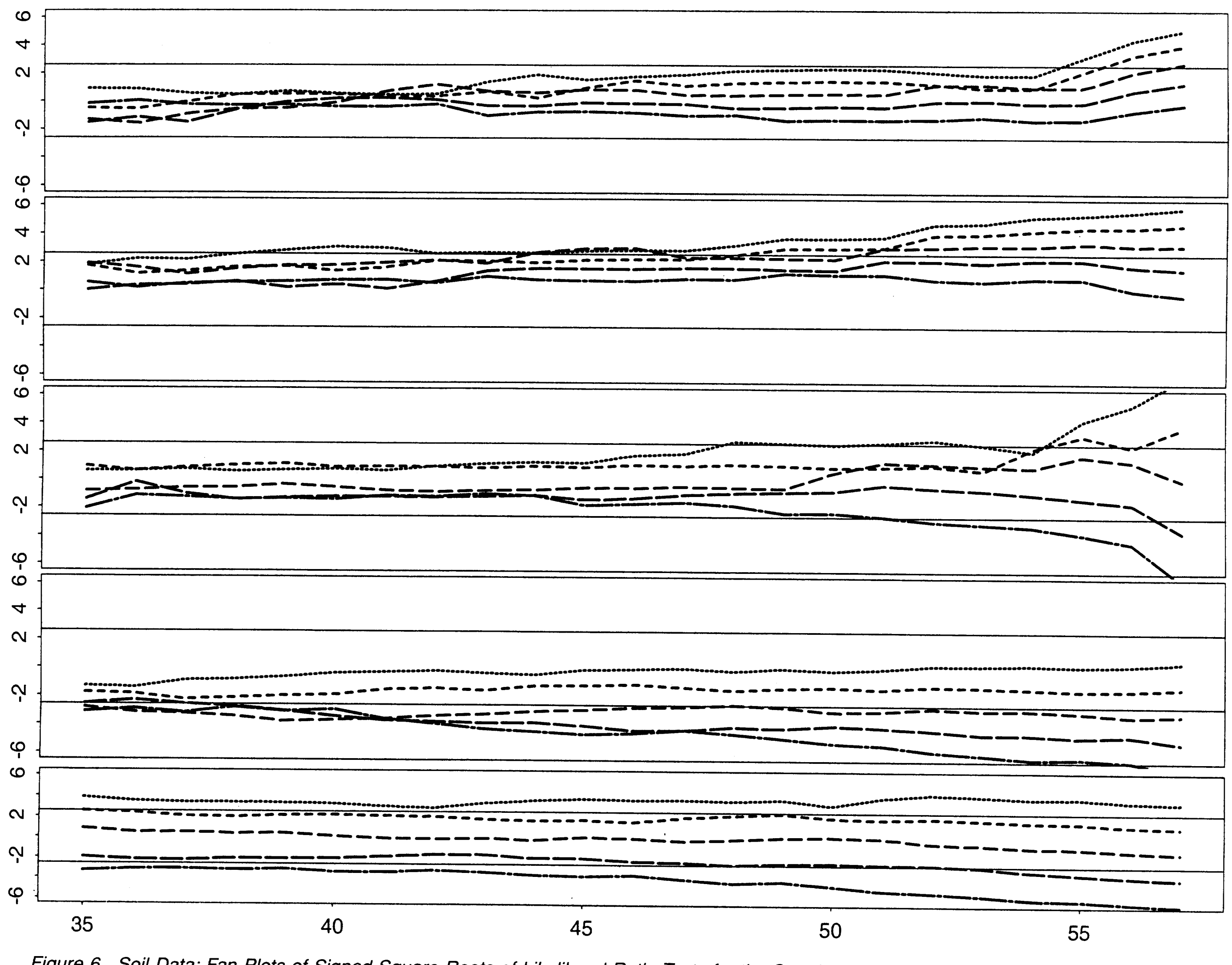

Figure 6. Soil Data: Fan Plots of Signed Square Roots of Likelihood Ratio Tests for the Standard Five Values for Each Component of λ When the Other Four Values Are the Relevant Members of (1, 1, 0, −1, 0). The top panel is for λ_1. In each panel, the parameter values are λ = 1, — - — - —; λ = .5; — — —; λ = 0, — — — —; λ = −.5, - - - - - -; λ = −1, · · · · · · · · ·. In all, 25 searches were required.

We begin with univariate analyses of the responses to indicate the need for a transformation and then use the extra power of multivariate transformations to confirm our results. Throughout, we use the three-variable model (x_1, x_4, and x_6) used by Box and Draper, the evidence for which is not affected by the transformations considered. The three resulting fan plots of the score test are gathered together in Figure 7. For the first response y_1 (strength) there is a jump in four out of five score statistics at the end of the search due to the inclusion of observation 1, the smallest observation. The effect is largest on the reciprocal transformation and negligible on the acceptable λ values of 1 and .5. The conclusion is that y_1 does not require transformation. For y_2, hue, the structure of the plot is similar, except that the square-root transformation is indicated. The large increases in the statistics for $\lambda = 0, -.5$, and -1 at the end of the search are caused by inclusion of the two smallest observations. Only $\lambda = .5$ is acceptable throughout the search. The plots for brightness, y_3, are devoid of sudden jumps, all observations indicating no need for transformation.

To confirm the findings from univariate analyses that the vector transformation parameter λ is (1, .5, 1), a forward search is based on the Mahalanobis distances calculated from the residuals from the regressions. Figure 8 shows plots of likelihood ratio tests from searches on untransformed data. The upper curve is the likelihood ratio for testing $\lambda = (1, 1, 1)$ against an unrestricted alternative. The plot shows the 95% and 99% points of this χ^2_3 distribution: The hypothesis of no transformation is clearly rejected. Because the search is on untransformed data, the initial part of the search includes observations that support the null hypothesis. The lower curve is again for testing the hypothesis of no transformation, but with the alternative $\lambda = (1, \lambda_2, 1)$ so that only transformation of y_2 is considered. The two tests are virtually indistinguishable, showing that all the evidence for transformation is in y_2 and does not depend on one or a few observations.

Finally Figure 9 plots the signed square root of the likelihood ratio test, on 1 df, for testing $\lambda = (1, .5, 1)$. No other transformation is indicated because the statistic is less than .7 in absolute value for $m \geq 22$. The only feature of interest is the downward jump at the beginning of the search followed by the upward jump at $m = 21$, caused by the exclusion of observation 10 from the initial subset and its

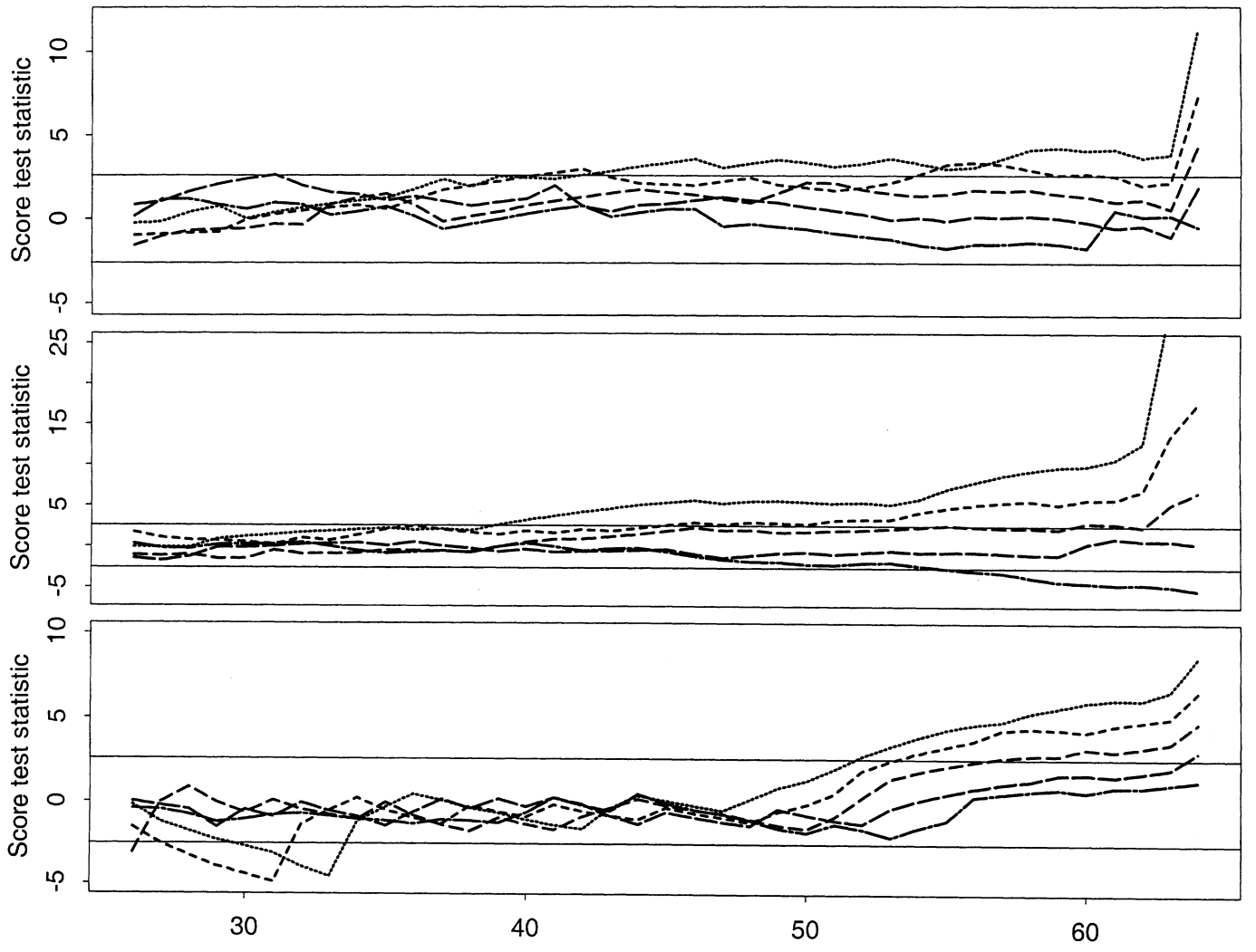

Figure 7. Dyestuffs Data: Fan Plots of Score Statistics $T_P(\lambda_o)$ for Marginal Power Transformation of Each Response: Top Panel y_1, Bottom Panel y_3: Individual Searches for Each λ. The parameter values are λ = 1, — - — - —; λ = .5, — — —; λ = 0, — — — —; λ = −.5, - - - - -; λ = −1, · · · · · · · · · · . Only y_2 needs transforming.

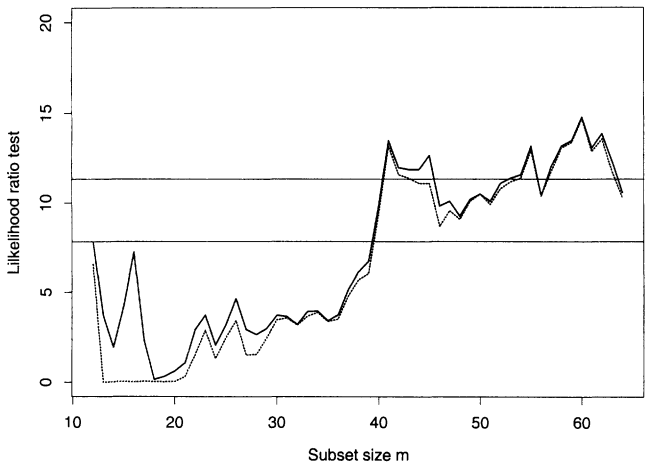

Figure 8. Dyestuffs Data: Likelihood Ratio Tests for the Null Hypothesis λ = (1, 1, 1) Against the Unrestricted Alternative (λ_1, λ_2, λ_3), ·····, and the Restricted Alternative λ = (1, λ_2, 1), - - - - - -. The horizontal lines are the 95% and the 99% points of the χ^2_3 distribution. All evidence for a transformation is provided by y_2.

later inclusion. This observation is not outlying in y_1 and y_3 but has the smallest value of y_2 in the sample. It therefore provides appreciable information about the transformation at the beginning of the search when m is small but has negligible influence when other observations, which also support this transformation, are included. Although this perturbation is not statistically significant, it is gratifying to note that it can be traced to a physical source. There is no evidence of the influence of any other observations, other searches giving similar results to those of Figure 9 for the last one-third of the search.

8. DISCUSSION AND OTHER PROBLEMS

In this article, the ordering from the forward search has been used to provide information about transformations and for the detection of outliers and influential observations in regression, providing simple graphical representations of residuals from the forward search and of the evolution of t statistics and likelihood ratio tests. Extensions of the method to nonlinear least squares and generalized linear models were described by Atkinson and Riani (2000).

The forward search has also been applied to data analyzed by standard multivariate methods. For discriminant analysis of multivariate normal data, Atkinson and Riani (1997) used the forward search to show the evolution of classification probabilities and to identify outliers. Another application in multivariate analysis is to the detection of multiple outliers in spatial prediction models, for which a forward-search algorithm was given by Cerioli and Riani (1999).

By the standards of data mining, most of these examples would be considered "tiny." There are two ways in which our procedure could be modified for larger datasets: (1) The selection of an initial subset becomes slower and less reliable as n increases. The results of Woodruff and Rocke (1994) show that there are advantages both in time and in the robustness of the estimators in dividing large datasets into several smaller groups and finding an initial robust estimate from each. (2) There could be larger steps in the forward search: Instead of incrementing our searches by one observation at a time, moving from m to $m+1$, the algorithm is basically unchanged if the move is to a subset of size $m+s$. The value of s could be chosen adaptively, to be relatively large at the beginning of the search but smaller at the end when changes in the statistics are more likely. Furthermore, if records are kept of the progress of the search, it can be restarted from a selected point and run with a different value of s. An important aim in any such extensions would be to produce plots that are as simple to understand as those given here.

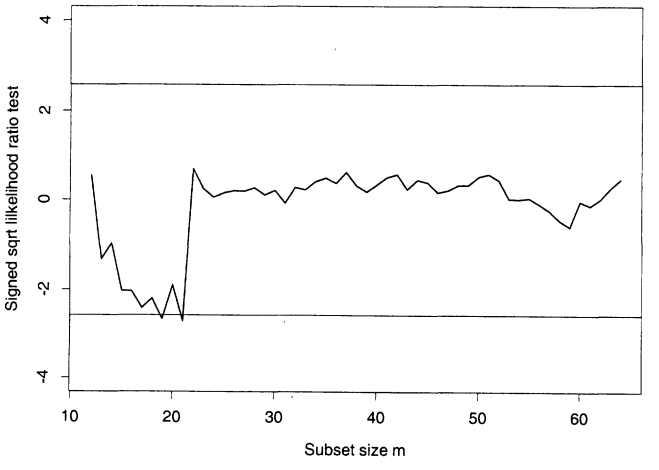

Figure 9. Dyestuffs Data: Signed Square Root of Likelihood Ratio Test for λ_2 = .5 With Other Components Untransformed: An Uneventful Confirmatory Plot, Showing That the Correct Transformation Has Been Found and That There Are no Influential Outliers.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the Staff Research Fund of the London School of Economics.

[Received November 1997. Revised March 2000.]

REFERENCES

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971), "Transformations of Multivariate Data," *Biometrics*, 27, 825–840.

Atkinson, A. C. (1985), *Plots, Transformations, and Regression*, Oxford, U.K.: Oxford University Press.

——— (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, 1329–1339.

——— (1995), "Multivariate Transformations, Regression Diagnostics and Seemingly Unrelated Regression," in *MODA 4—Advances in Model-Oriented Data Analysis*, eds. C. P. Kitsos and W. G. Müller, Heidelberg: Physica-Verlag, pp. 181–192.

Atkinson, A. C., and Mulira, H.-M. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers," *Statistics and Computing*, 3, 27–35.

Atkinson, A. C., and Riani, M. (1997), "Bivariate Boxplots, Multiple Outliers, Multivariate Transformations and Discriminant Analysis: The 1997 Hunter Lecture," *Environmetrics*, 8, 583–602.

——— (2000), *Robust Diagnostic Regression Analysis*, New York: Springer-Verlag.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211–246.

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley.

Box, G. E. P., and Tidwell, P. W. (1962), "Transformations of the Inde-

pendent Variables," *Technometrics*, 4, 531–550.

Cerioli, A., and Riani, M. (1999), "The Ordering of Spatial Data and the Detection of Multiple Outliers," *Journal of Computational and Graphical Statistics*, 8, 239–258.

Chatterjee, S., and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, New York: Wiley.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

——— (1994), *An Introduction to Regression Graphics*, New York: Wiley.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: Wiley.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 761–771.

Hadi, A. S., and Simonoff, J. S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, 1264–1272.

——— (1994), "Improving the Estimation and Outlier Identification Properties of the Least Median of Squares and Minimum Volume Ellipsoid Estimators," *Parisankhyan Sammikkha*, 1, 61–70.

Hawkins, D. M. (1993), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association*, 88, 580–589.

Mulira, H.-M. (1992), *Computational Methods for Transformations to Multivariate Normality*, unpublished Ph.D. thesis, London School of Economics, Dept. of Statistical and Mathematical Sciences.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 731–792.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Velilla, S. (1993), "A Note on the Multivariate Box–Cox Transformation to Normality," *Statistics and Probability Letters*, 17, 259–263.

——— (1995), "Diagnostics and Robust Estimation in Multivariate Data Transformations," *Journal of the American Statistical Association*, 90, 945–951.

Woodruff, D., and Rocke, D. M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association*, 89, 888–896.

Zani, S., Riani, M., and Corbellini, A. (1998), "Robust Bivariate Boxplots and Multiple Outlier Detection," *Computational Statistics and Data Analysis*, 28, 257–270.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368.