ADAC (2007) 1:123–141
DOI 10.1007/s11634-007-0007-y

REGULAR ARTICLE

Fast calibrations of the forward search for testing multiple outliers in regression

Marco Riani · Anthony C. Atkinson

Published online: 22 June 2007

Abstract The paper considers the problem of testing for multiple outliers in a regression model and provides fast approximations to the null distribution of the minimum deletion residual used as a test statistic. Since direct simulation of each combination of number of observations and number of parameters is too time consuming, methods using simple normal samples are described for approximating the pointwise distribution of the test statistic. One approximation is based on adjustments to the results of simple simulations. The other uses properties of order statistics from folded t distributions to move outside the significance levels available by simulation. Analyses of data with beta errors and of transformed data on survival times demonstrate the usefulness in graphical methods of the inclusion of our bounds.

Keywords Beta-distributed errors · Deletion residual · Multiple outliers · Order statistics · Robust methods · Truncated t distribution

1 Introduction

To test for outliers requires division of the data into two groups. The parameters are estimated from m observations, believed to come from the uncontaminated distribution model. The estimates are then used in constructing the test for the remaining $n - m$ observations. Here we apply this idea to the likelihood ratio test for outliers

M. Riani (✉)
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: mriani@unipr.it

A. C. Atkinson
The London School of Economics, London WC2A 2AE, UK
e-mail: a.c.atkinson@lse.ac.uk

in regression and propose fast simulations that give very good approximations to the null distribution of the test statistic.

The literature on the detection of multiple outliers in regression is large (see, for example, Beckman and Cook 1983 or Barnett and Lewis 1994). A recent review and comparison of methods is Wisnowski et al. (2001). The central problem is that of "masking": if there are several outliers, least squares estimation of the parameters of the model from all n observations will not lead to identification of the m uncontaminated observations. Single deletion methods (for example, Cook and Weisberg 1982; Atkinson 1985) will also fail. Hawkins (1983) argues for exclusion of all possibly outlying observations, which are then tested sequentially for reinclusion.

The forward search as described in Atkinson and Riani (2000) is an objective procedure of the type suggested by Hawkins. It starts from a small, robustly chosen, subset of the data that is clear of outliers and fits subsets of increasing size. Each observation is tested for outlyingness before it is included in the fitted subset. The likelihood ratio test for agreement of the new observation with those already in the subset reduces to the well-known deletion residual.

The focus of this paper is on the null distribution of this series of test statistics. The simplest way of generating distributional envelopes is the parametric bootstrap that uses repeated simulations of the forward search to find the distribution of the test statistic for each m for a given n and p, the number of regressors. However, there are only 10,080 minutes in a week; unless the time for simulation of an individual forward search is small, the simulation of the 10,000 searches needed to find the tail probabilities for values of the test statistic will take an excessive time. Atkinson and Riani (2006) suggest several faster approximate methods. One is the simulation of n observations without regression, the envelopes from which have to be corrected. They roughly adjust the resulting upper percentage points of the null distribution of the outlier test for p, using a correction that does not depend on n. One contribution of this paper is to derive fast, accurate corrections for both the upper and lower percentage points that depend on both p and n. We thus require only one set of fast simulations for any n, regardless of the dimension p of the fitted linear model.

A second contribution of this paper is to use results about order statistics to provide a second very fast and potentially highly accurate set of bounds. Atkinson and Riani (2006) suggest a method based on the asymptotic normality of the ordered residuals. The resulting approximation is however only useful when m is not far from n. Here we use the distribution of order statistics from a folded t distribution. Except for very precise work, there is an appreciable range of values of p and n over which these envelopes can be used without any further correction for all m of interest. The great advantage of this method is that it can provide extreme significance levels beyond those available by simulation.

The paper is organised as follows. In Sect. 2 we briefly review the forward search and the outlier test. Section 3 presents the analysis of a set of data accompanied by envelopes from 10,000 simulated forward searches. The general shape of the envelopes for the null distribution of the statistic is indicated in Sect. 4 with the details of our fast simulation approximation in Sect. 5. The importance of the lower bounds for the test statistic is illustrated in Sect. 6 by the analysis of observations with beta distributed errors. The envelopes based on order statistics are derived in Sect. 7 and used in Sect. 8

in a brief analysis of data on survival times. This analysis underlines the importance of being able to calculate extreme percentage points of the distribution. Throughout the paper the focus is on the distribution of the outlier test, with the forward search used as a framework for simulations and inference. The emphasis is on the construction of methods useful in the analysis of data.

2 Theory

2.1 Least squares

In the regression model $y = X\beta + \epsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, where $H = X(X^T X)^{-1} X^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^T e/(n - p) = \sum_{i=1}^n e_i^2/(n - p)$.

2.2 The forward search

The forward search fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. Let $S_*^{(m)}$ be the subset of size m found by the forward search, for which the matrix of regressors is $X_*(m)$. Least squares on this subset of observations yields parameter estimates $\hat{\beta}_*(m)$ and $s_*^2(m)$, the mean square estimate of σ^2 on $m - p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S_*^{(m)}$. The n resulting least squares residuals are

$$e_{i*}(m) = y_i - x_i^T \hat{\beta}_*(m). \tag{1}$$

The search moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m + 1$ smallest absolute values of $e_{i*}(m)$. The estimates of the parameters are based only on those observations giving the central m residuals.

To start we take $m_0 = p$ and search over subsets of p observations to find the subset, out of 3,000, that yields the least median of squares (LMS) estimate of β (Rousseeuw 1984). Although this initial estimator is not $\sqrt{n}$-consistent (Hawkins and Olive 2002) our results show that the initial estimator is not important, provided masking is broken. Identical inferences are obtained using the least trimmed squares estimator except sometimes when m is small and $n/p < 5$. Random starting subsets also yield indistinguishable results over the last one third of the search. Examples for multivariate data are in Atkinson et al. (2006). The forward search, adding, and sometimes deleting, observations provides a bridge between the initial estimate and $\sqrt{n}$-consistent parameter estimates for the uncontaminated observations as the sample size goes to infinity, in a similar way to the estimators discussed in Maronna and Yohai (2002). We are, however, concerned here with finite sample results and simulation based inference.

2.3 Testing for outliers

To test for outliers the deletion residual is calculated for the $n - m$ observations not in $S_*^{(m)}$. These residuals are

$$r_{i*}(m) = \frac{y_i - x_i^T \hat{\beta}_*(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}} = \frac{e_{i*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}}, \tag{2}$$

where $h_{i*}(m) = x_i^T \{X_*(m)^T X_*(m)\}^{-1} x_i$; the leverage of each observation depends on $S_*^{(m)}$. Let the observation nearest to those constituting $S_*^{(m)}$ be $i_{\min}$ where

$$i_{\min} = \arg \min_{i \notin S_*^{(m)}} |r_{i*}(m)|,$$

denotes the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$. To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{i_{\min}*}(m) = \frac{e_{i_{\min}*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i_{\min}*}(m)\}}}, \tag{3}$$

as a test statistic. If the absolute value of (3) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S_*^{(m)}$. Our paper is mainly devoted to the determination of the percentage points of the null distribution of (3).

3 An example of bootstrap envelopes: Hawkins's Data

The fundamental method of finding the distribution of the outlier test statistic (3) during the forward search, without approximations, is the parametric bootstrap involving the simulation of 10,000, or some such large number, of forward searches. In general the bootstrap simulations will depend on the estimates of the parameter of the model fitted to the data. But, for normal theory regression data the distribution of the deletion residuals in (2) is independent of the values of the variance σ^2 and the parameters β used in the simulation. We therefore take them respectively as one and zero.

Accordingly, in analysing a specific set of normal theory regression data, we keep the structure of the X matrix, in order to give residuals with the same hat matrix as the data. If, as in most of this paper, we are concerned with inference for an unspecified model of known dimension, the X matrix is generated, once for the set of simulations, with a column of ones and, if $p > 1$, the remaining columns sampled from standard normal random variables. The responses are also sampled from a standard normal distribution, but once for each search. The empirical percentage points of the simulated distribution are found for each m and displayed as a function of m, when they form an envelope with the required pointwise size.

The set of simulated data analysed in this section were intended by Hawkins to be misleading for standard regression methods. An analysis is given by

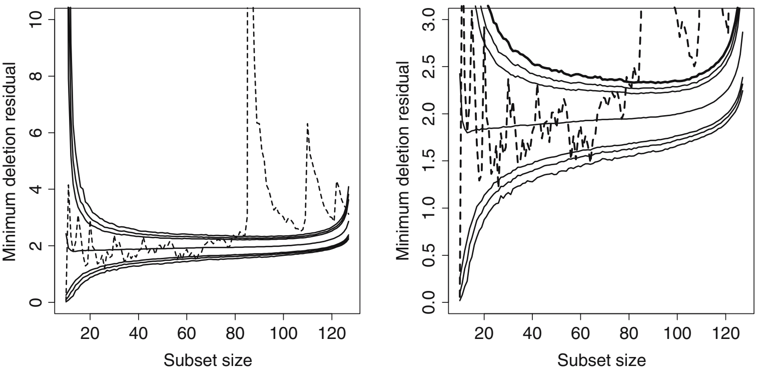

Fig. 1 Hawkins's Data: forward plot of minimum deletion residuals. *Left-hand panel*, three clear peaks and masking at the end of the search; *right-hand panel*, zoom showing large values just before $m = 86$. Bootstrap envelopes

(Atkinson and Riani, 2000, Sect. 3.1). There are 128 observations and nine explanatory variables. An advantage of this simulated data set as a first example is that it is free of the complications of real data that we confront in Sect. 8.

The left-hand panel of Fig. 1 shows the forward plot of the minimum deletion residual among observations not in the subset, that is the outlier test statistic (3), together with the bootstrap percentage points of the distribution based on 10,000 simulations. The right-hand panel shows a detail of this plot. The envelopes plotted with continuous lines in both panels are the 1, 2.5, 5, 50, 95, 97.5 and 99% points of the empirical distribution of the outlier test statistic (3) during forward searches simulated without outliers. The envelopes are of a kind we shall see in all simulations. Initially they are very broad, corresponding to distributions with high trimming and few degrees of freedom for the estimation of error. In the central part of the search the band is virtually horizontal and gradually narrows. Towards the end of the search there is rapid increase as we test the few largest residuals. The left-hand-panel of the figure shows the four groups of observations in the data, with the first sharp peak at $m = 86$, revealing observation 87 as a very remote outlier. As further observations enter, the deletion residual decreases as the estimates of β and σ^2 change, accommodating the cluster of outliers. The plot repeats this pattern twice more, although less strongly, showing the four groups of observations revealed by Atkinson and Riani (2000).

It is not clear from the zoom of the search in the right-hand panel of Fig. 2 whether the last few observations to enter before $m = 86$ may not also be outliers as they lie outside the simulation envelope. However, the simulation envelopes increase at the end of the search, increasingly so from $m = 120$. We accordingly repeat the simulation with $n = 86$. Figure 2 shows the resulting envelopes with the forward plot of the deletion residuals for the first 86 observations to enter. Now the increasing last part of the curve is seen to lie within the simulation envelope. There are no outliers at the end of the search. There is however a slightly low patch around $m = 60$. This serves as a reminder that we are finding pointwise bounds. The results of Atkinson and Riani (2006) indicate that the probability of one observation lying outside the 1% bound during the last half of such a search is almost 20%.

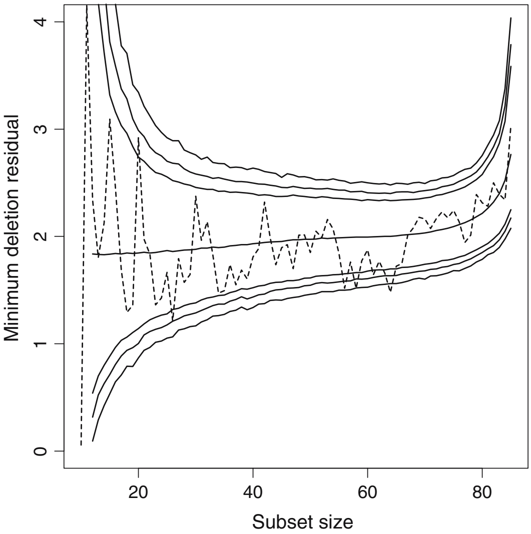

Fig. 2 Hawkins's Data: forward plot of minimum deletion residuals, 86 observations. The last observations now lie within the bootstrap envelope

4 The structure of bootstrap upper bounds

We now consider a fast alternative to the bootstrap envelopes. Atkinson and Riani (2006) suggest deriving envelopes from simulations of ordered observations, that is the case $p = 1$. These envelopes then have to be adjusted for the number of parameters p. Although we simulate for the desired n, the adjustment will depend on the values of both n and p. For expository purposes we first consider only the upper bounds.

The bootstrap bounds in Fig. 1 for n = 128 and those in Fig. 2 for n = 86 are similar in shape. In general the upper bootstrap bounds have a similar shape as n and p vary for each value of α in the upper tail. Figure 3 shows this dependence for the empirical 99% point for $n = 200$ as p ranges from 3 to 13. As p increases the curves move upwards. The shapes are similar, with the largest difference for small m. As $m \to n$ the curves move closer together, dependence on p virtually disappearing.

Since we are looking for a common correction for all n we also require that the curves have a similar structure as n varies. Figure 4 again shows the 99% bootstrap bounds, but now for $p = 3$ as n varies from 50 to 300, plotted against 100 m/n, that is the percentage of observations in the subset. The shapes are indeed similar, so that it should be possible to find a correction for a wide range of n.

5 Simulation envelopes from ordered observations

5.1 Background

In the forward search the n observations are ordered for each value of m. In the absence of outliers this ordering does not change much during the search. To approximate the distribution of the statistics we simulate some large number K of sets of n observations

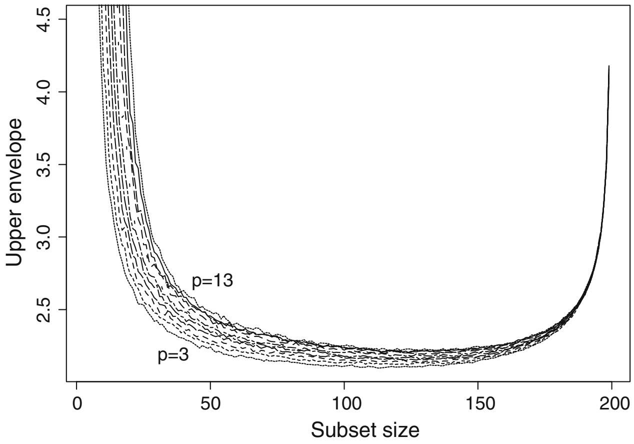

Fig. 3 Analysis of dependence of bootstrap upper envelopes on p: 99% envelope when $n = 200$ for p from 3 to 13

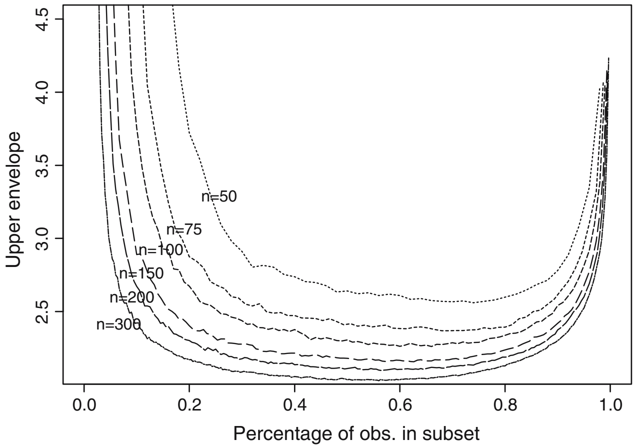

Fig. 4 Analysis of dependence on n: 99% bootstrap envelopes for n from 50 to 300 and $p = 3$ plotted against percentage of observations in the subset

Z from the normal distribution and correct for the mean to give raw residuals $e^{(z)}(k)$ ($k = 1, \ldots, K$). We order the absolute values of each set of n residuals $e^{(z)}(k)$ just once.

For each value of m we use the m smallest values of each set $|e^{(z)}(k)|$ to calculate the outlier test statistic (3) with $i_{\text{min}} = i_{[m+1]}$, the index of the $(m + 1)$st ordered value of $|e^{(z)}(k)|$. This procedure, which avoids a forward search for each simulation, is repeated K times to give the empirical distribution of the outlier test statistic. With good corrections for regression we require only one set of such simulations for any n, regardless of the dimension p of the fitted linear model.

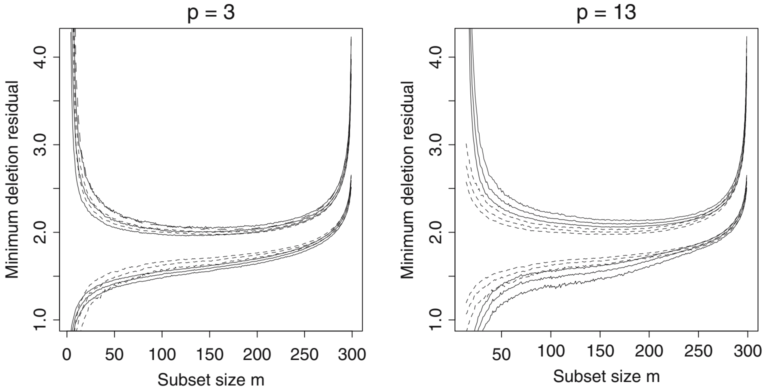

Fig. 5 Forward plots of 1, 2.5, 5, 95, 97.5 and 99% simulation envelopes for the outlier test when $n = 300$. *Continuous lines*, bootstrap; *dotted lines*, ordered observations. The agreement of the upper envelopes when $p = 3$ is striking

5.2 Adjustment for regression

Figure 5 shows forward plots of the percentage points of the distribution of the test when $n = 300$. The continuous lines are from 10,000 bootstrap simulations of a forward search, with X a matrix of random normal regressors, whereas the dotted lines uses the faster approximation of ordered observations. In the left-hand panel, $p = 3$, and there is excellent agreement between the empirical method and the approximation for the upper percentage points over the whole range of m. The agreement for the lower percentage points, and for both upper and lower points when $p = 13$ is not satisfactory.

The close agreement between the upper bootstrap bands for $p = 3$ and the approximation provides the basis for our adjustment of the upper bounds for regression. For any n approximate upper bounds are found from ordered samples, which only requires adjustment for the mean and ordering the absolute residuals. These are then corrected for p.

5.3 A simple correction

The curves in Fig. 3 are not only similar in shape but seem to be virtually in a common ratio as p increases. Accordingly, for each m for a particular n we can calculate the ratio of the bound for $p+1$ to that for p, for p going from 3 to 12. The resulting plots, for six values of n from 50 to 300 and $\alpha = 0.99$ with $p = 3$ are plotted in Fig. 6. There is some randomness in the curves since, even with 10,000 simulations, the extreme percentage points are only approximately determined and the curves are formed from ratios of such values. However, the structure of this plot is clear. The curves coincide around $m = 70$ and thereafter decline, approximately linearly, to 1 at $m = n - 1$.

Our simple correction comes from this graph. The multiplicative factor is found by linear interpolation from a value of 1.008 when $m = 70$ to 1 when $m = n - 1$.

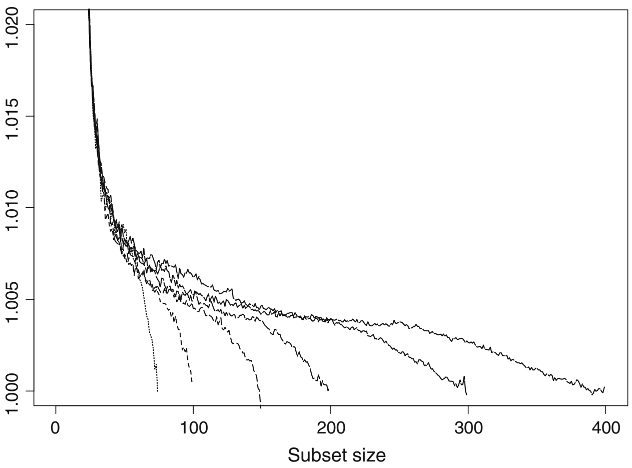

Fig. 6 95% bootstrap bounds. Average increase in going from p to $p+1$ for various sample sizes n, labeled by the subset size m at the end of each search

The steps are:

Step 1 Simulate 10,000 (or some other large number) of approximations to the forward search using ordered samples, that is $p = 1$. Find the empirical 95, 97.5 and 99% points of the distribution for each m. Let such a point be $\zeta(m)$.

Step 2 Let $k = m/(n-1)$ and $k_0 = 70/(n-1)$. Then the correction factor is

$$\gamma(k) = 1 + 0.008\frac{(1-k)}{(1-k_0)}. \tag{4}$$

Step 3 The corrected point on the envelope is therefore

$$\zeta^c(m) = \zeta(m) \times \gamma(k)^{(p-3)}. \tag{5}$$

Remark This adjustment is intended for the upper tail of the distribution ($\alpha > 0.5$) and applies when $n > 70$.

Figure 7 shows how well this simple correction works for $n = 258$ and $p = 8$. The approximate and exact curves are virtually indistinguishable from $m = 150$. The adjustment is best for the 99% bound and works well down to $m = 70$. In practice, the differences between the curves will not affect inferences drawn from the data.

5.4 A possible more sophisticated correction

Although Fig. 7 shows how well our simple correction behaves, a slight systematic error is apparent in the figure; for small values of k the curves are slightly too low, that is too small a factor has been used whereas, for large k, that is as m approaches n, the correction is too large, since the value of $\gamma(k)$ is slightly too large. This behaviour is

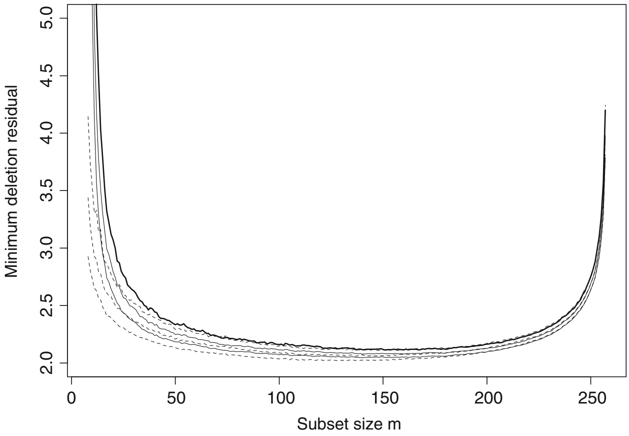

Fig. 7 Corrected upper bounds from ordered samples (*dotted lines*) for $n = 258$ and $p = 8$. The agreement with the bootstrap envelopes (*continuous lines*) becomes excellent as m increases

explained by our correction, which is a linear approximation, the same for all α and p, to the slightly curved plots in Fig. 6. We now indicate how a more sophisticated correction might be obtained. As an example, we assume interest is in inference over the last half of the search.

Figure 8 shows fitted quadratic curves of $\gamma(k)$ for four values of n with $\alpha = 0.95$. These smooth curves explain all the structure in Fig. 6, as do similar curves for other values of α. Using orthogonal polynomials for fitting these quadratics, we obtain parameter estimates that vary in a smooth way with α and n.

Parameterization of this variation would then give a response surface from which $\gamma(k)$ could be found for the desired n and α. However, such a procedure depends on the region of the search for which improved envelopes are required. For example, for the last 20% of the search, a linear approximation suffices. We do not here follow this route further, since the results in Fig. 7 show the adequacy of our simple linear approximation for practical purposes. Instead we consider approximations to the lower tail of the distribution.

5.5 The structure of the lower bounds

It is clear from Fig. 5 that the lower bounds from bootstrap simulations are somewhat different from those provided by the ordering of simulated observations, especially for large p. However, the curves from bootstrap simulations with $p = 1$, which are not shown in the figure, do have a shape similar to the bootstrap curves for general p. We accordingly base our procedure for general n on scaling up bootstrap simulations with $p = 1$. Our approximate lower intervals therefore require more computation than those for the upper bound in which the forward search is avoided. We proceed as before to find an approximation based on averaging changes in the curves. This time we have a rule which varies with α, since the curves are not equispaced, and which also depends, in form, on the values of m and n.

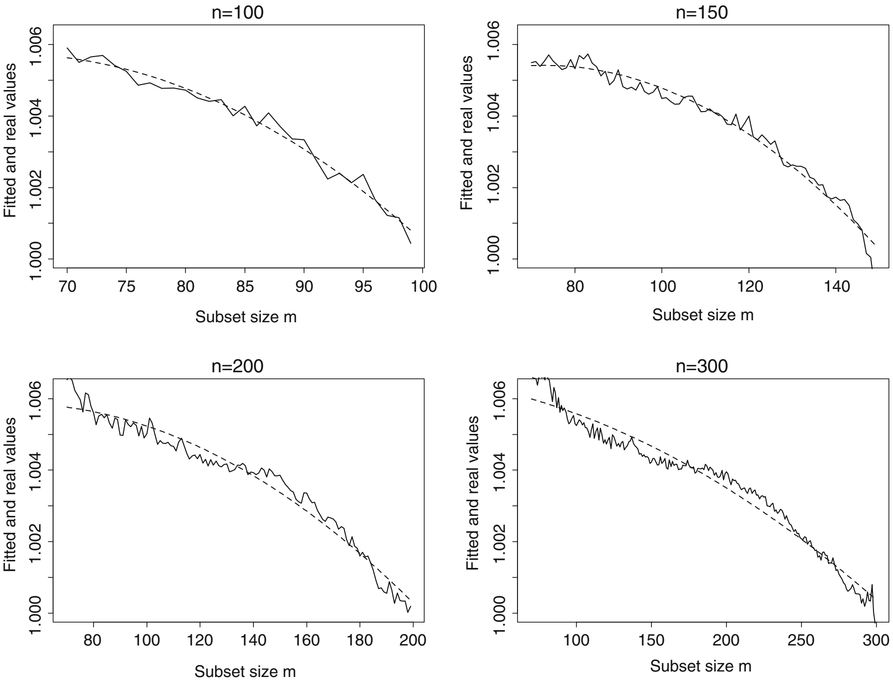

Fig. 8 Approximations, using a quadratic curve, to the ratios of empirical 95% curves in Fig. 6 for four values of *n*

Figure 9 shows the average decrease in the lower bounds of the bootstrap envelopes in going from p to $p+1$ for p from 1 to 13 when $\alpha = 0.01$. It is the analogue of Fig. 6, which gave the increase for the upper bounds. A new feature of these curves is that, for large m and n, the correction factor tends to one, that is there is no correction. In order to accommodate this "bent stick" relationship we propose the following steps:

Step 1 Simulate 10,000 bootstrap approximations to the forward search with $p = 1$. Find the empirical 1, 2.5 and 5% points of the distribution for each m. Let such a point be $\zeta(m)$.

Step 2 Let $m_{\max}$ be the value of subset size above which no correction is to be made. Then $k = m/m_{\max}$ and $k_0 = 70/m_{\max}$. The correction factor is

$$\gamma(k) = \begin{cases} 1 - \tau(\alpha)\frac{(1-k)}{(1-k_0)} & m \le m_{\max} \\ 1 & \text{otherwise.} \end{cases} \quad (6)$$

The values of $\tau(\alpha)$ are

α	$\tau(\alpha)$
0.05	0.997
0.025	0.992
0.01	0.983

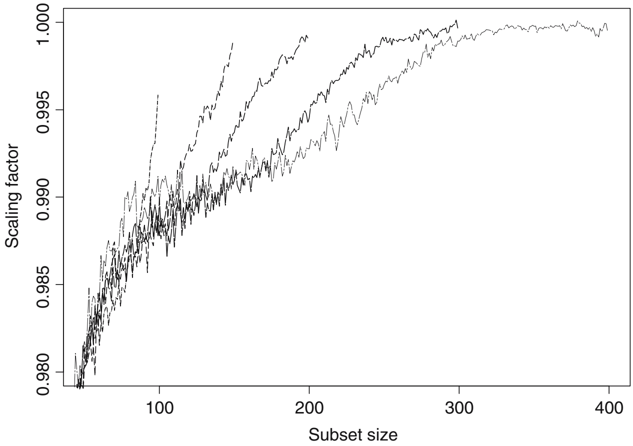

Fig. 9 1% bounds. Average decrease in bootstrap envelopes in going from p to $p+1$ for various sample sizes

and $m_{\max}$ is given by

n	$m_{\max}$
< 200	n
$200 \leq n \leq 600$	$[n\{1-(n-200)/1000\}]$
> 600	360

where $[\cdot]$ denotes integer part.

Step 3 The corrected point on the envelope is therefore

$$\zeta^c(m) = \zeta(m) \times \gamma(k)^{(p-1)}. \tag{7}$$

Remark Note that in (7) the power in the exponent is $p-1$, whereas in (5) it is $p-3$.

Figure 10 shows how this simple correction works for $n = 258$ and $p = 8$, the values used for the upper tail in Fig. 7. The correction is excellent for values of m above 100. We have also tested the approximation on many other values of m and n, with similarly satisfactory results. For smaller n the correction will work for smaller values of m. For really small n bootstrap simulations can be used directly.

6 Lower tails and distributional form

In general interest is in the upper tail of the distribution, where we expect outliers to be evidenced and where the correction is speedier since we simulate and order random samples. However, systematic departures from the model can lead to low values of the statistic. In this case we can use the method of Sect. 5.5, scaling the bootstrap envelopes from $p = 1$. As an example we show in Fig. 11 a forward search from data in which the errors were generated by a beta distribution with parameters 1/2, 1/2. The figure includes the signs of the outlier tests.

The observations show a pattern that is initially gently rising, before it later declines. In the latter part of the search nearly all of the values are negative and many are too

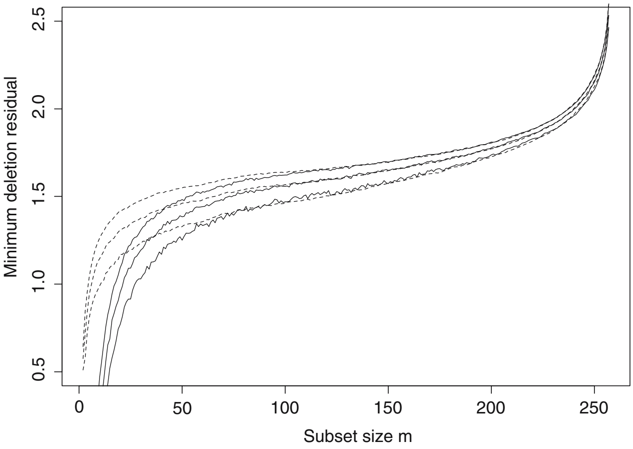

Fig. 10 Lower bootstrap bounds for $p = 1$ and $n = 258$ corrected for $p = 8$. The correction works well above $m = 100$. Continuous lines, bootstrap bounds with $p = 8$

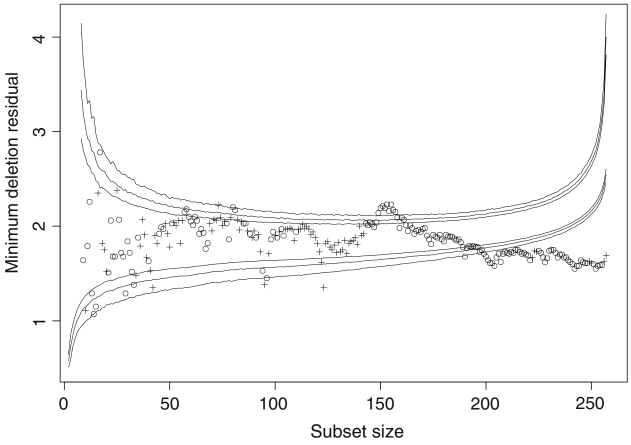

Fig. 11 Envelopes for normal distribution for $n = 258$ and $p = 8$ as in Figs. 7 and 10, with y generated from a beta distribution: + positive residuals, ○ negative residuals

small. An explanation is that the distribution of errors is U-shaped. Initially the search fits observations in the centre and then those from the upper half of the distribution. Observations in the other half of the distribution are remote and so give rise to the large negative values around $m = 150$. As further observations from the lower half of the distribution continue to be fitted, these observations become less remote and, from m around 180, the residuals are both negative and too small for a normal sample because the distribution does not have any tail: many observations are concentrated just above minus one.

The structure evident in Fig. 11 is highlighted by the envelopes. In their absence this rather flat plot might be thought not to show anything. Certainly a plot of the statistics is essential to any elucidation of the structure of the data.

7 Envelopes from order statistics

Since the test statistic (3) is the $(m+1)$st ordered value of the deletion residuals, we can use distributional results for order statistics as an alternative method for obtaining approximate envelopes for our plots. Atkinson and Riani (2006) suggest a method based on normal order statistics which provided poor approximations. Here we explore envelopes using the t-distribution of our statistic that have greatly improved properties. Since these envelopes do not require simulation in their calculation, we can use them for much more extreme points of the distribution than would be possible for bootstrap intervals without massive simulations.

Let $Y_{(k)}$ be the kth order statistic from a sample of size n from a distribution with c.d.f. $G(y)$. Then the c.d.f of $Y_{(k)}$ is given exactly by

$$P\{Y_{(k)} \leq y\} = \sum_{j=k}^{n} \binom{n}{j} \{G(y)\}^j \{1 - G(y)\}^{n-j} \tag{8}$$

(see, for example, Lehmann 1991, p. 353). In our case we are considering the absolute values of the deletion residuals. If the c.d.f of the t distribution on ν degrees of freedom is written as $T_\nu(y)$, the absolute value has the c.d.f.

$$G(y) = 2T_\nu(y) - 1, \quad 0 \leq y < \infty. \tag{9}$$

To find percentage points of $Y_{(k)}$ we numerically solve (8) for the value of y giving the required percentage point. For a subset of size m we put $k = m + 1$. With $G(y)$ given by (9) and $\nu = m - p$ we obtain $V_{m,\alpha}$ as the $100\alpha\%$ point of the distribution of the absolute value of a t random variable, that is of the t distribution folded at the origin.

If we had an unbiased estimator of σ^2 the envelopes would be given by $V_{m,\alpha}$ for $m = m_0, \ldots, n-1$. However, the estimator $s_*^2(m)$ is based on the central m observations from a normal sample (strictly the m observations with smallest squared residuals based on the parameter estimates from $S_*^{(m-1)}$, which may not always be quite the same thing.) The variance of the truncated normal distribution containing the central m/n portion of the full distribution is

$$\sigma_T^2(m) = 1 - \frac{2n}{m} \Phi^{-1}\left(\frac{n+m}{2n}\right) \phi\left\{\Phi^{-1}\left(\frac{n+m}{2n}\right)\right\},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the standard normal density and c.d.f. (see, for example, Johnson et al. 1994, pp. 156–162). Since the outlier tests we are monitoring are divided by an estimate of σ^2 that is too small, we need to scale up the values of the order statistics to obtain the envelopes

$$V_{m,\alpha}^* = V_{m,\alpha}/\sigma_T(m).$$

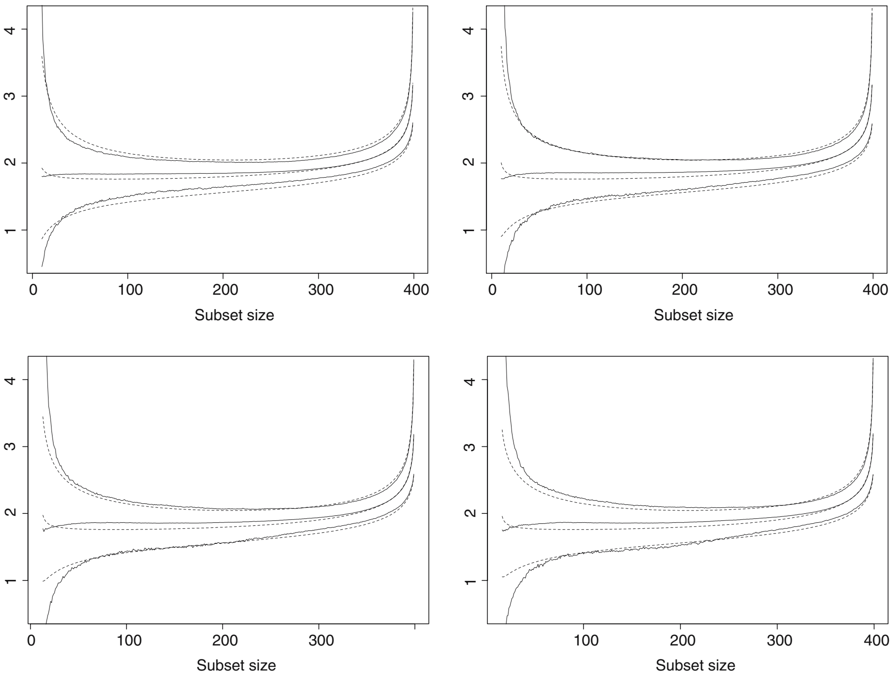

Fig. 12 Order statistic envelopes for $n = 400$ and four values of p: 4, 7 (*top*), 10 and 13 (*bottom*). Envelopes shown are 1, 50 and 99%. *Continuous lines*, bootstrap envelopes

The four panels of Fig. 12 show the order-statistic envelope together with the bootstrap envelopes for $n = 400$ and four values of p: 4, 7, 10 and 13. The approximation is very serviceable over virtually the whole of the range of m. The agreement for the frequently used values of p, such as 7 and 10 is slightly better than that for $p = 4$, where the envelopes are a little too broad, and for $p = 13$, where they are a little narrow.

8 Surgical unit data

In this section we continue to explore the properties of the order statistic envelopes $V^*_{m,\alpha}$ and give a practical example of their use with probability levels that could not feasibly be obtained by simulation.

Neter et al. (1996, p. 334) introduce data on the logged survival time of 54 patients undergoing liver surgery, together with four potential explanatory variables. On p. 437 another 54 observations are introduced to check the model fitted to the first 54. Their comparison suggests there is no systematic difference between the two sets.

Figure 13 is a forward plot of the test for outliers for all 108 observations, together with bootstrap envelopes and those from the method of ordered observations, Sects. 5.3 and 5.5. The two forms of envelopes agree closely, especially above $m = 60$. Different symbols are used to plot the two sets of observations and the figure seems to show

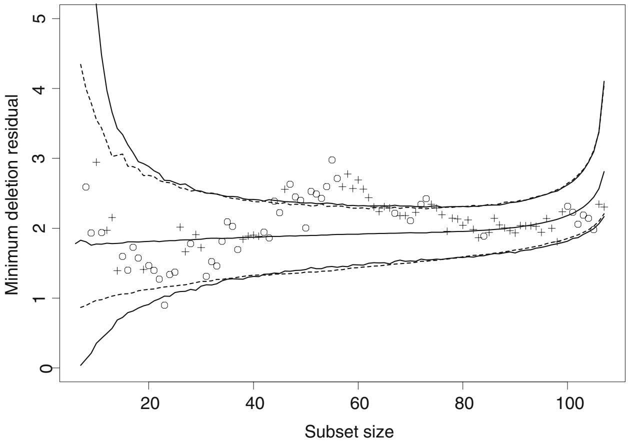

Fig. 13 Surgical Unit Data: forward plot of minimum deletion residuals with bootstrap envelopes (*continuous line*) and envelopes from the method of ordered observations, Sects. 5.3 and 5.5. The two sets of data are represented by different symbols

evidence of two groups. The extreme value of the statistic at the centre of the plot is caused by the entry of observations from one group, represented by circles, followed by the entry of several observations from the other group, represented by crosses. If the two groups were indistinguishable, entry would be random for observations from the groups.

Figure 14 repeats Fig. 13 but with bootstrap and order-statistic based envelopes, Sect. 7. One feature of the plots is that the approximate envelopes, while not quite as good as those from adjusting ordered observations, provide a more than adequate guide to inference. In addition, we have included the 99.99% order-statistic envelope to indicate the strong evidence that this peak is not a random fluctuation. Use of the method of Sect. 5.3 to generate such an extreme envelope would require an impractically vast simulation to estimate this extreme quantile. At the end of the search the statistic is nudging the lower envelope. The general shape recalls that of Fig. 11.

The maximum value of the statistic in Fig. 13 is at $m = 55$. As Fig. 13 shows, the units that enter after $m = 57$ nearly all come from the group of confirmatory observations, an indication that this second group of observations differs systematically from the original 54 units. The similarity with Fig. 11 arises because the errors seem to come from a mixture of two normal distributions differing in location. Accordingly we introduce a dummy variable for the two sets and repeat the analysis. This variable is highly significant, with a t value of -7.83 at the end of the search. However, the resulting forward plot, Fig. 15, still has a highly significant peak in the centre, even if the significance is now closer to 0.01%. The curve remains at a high level in the second half of the search compared to Fig. 13. Some remaining structure is indicated that is more than a simple shift in mean.

As a last example of the use of envelopes for the outlier test, we consider the two groups separately. The forward plots of the test for outliers are in Fig. 16. The plot for the second group of observations, the right-hand panel, indicates that that group is homogeneous. However, the left-hand panel strongly indicates that the first

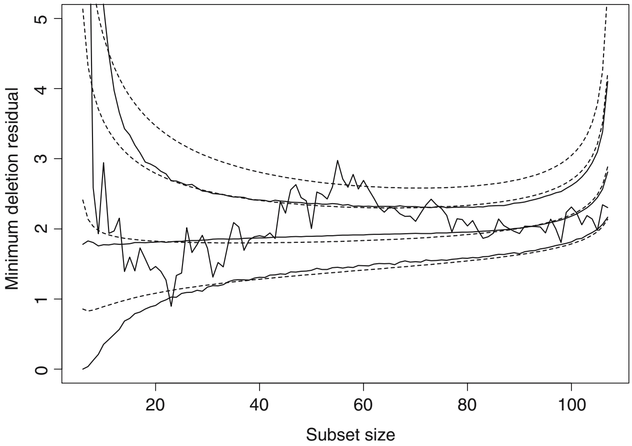

Fig. 14 Surgical Unit Data: forward plot of minimum deletion residuals with bootstrap envelopes (*continuous line*) and 1, 50, 99 and 99.99% envelopes using the method of order statistics, Sect. 7. The maximum of the statistic in the centre of the search is significant well beyond 0.01%, according to the approximation. The difference in the two groups of observations is highly significant

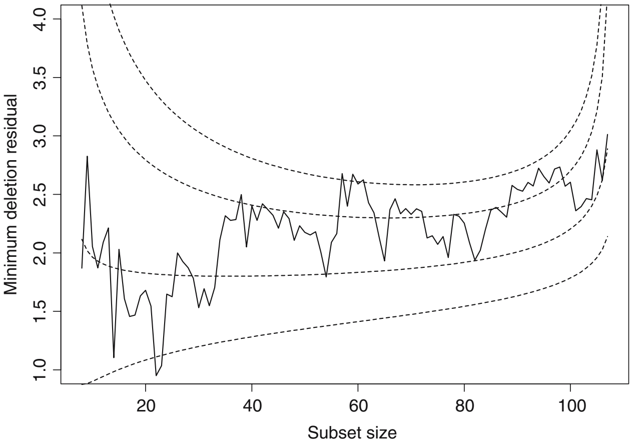

Fig. 15 Surgical Unit Data: forward plot of minimum deletion residuals including a dummy variable for the two groups with order-statistic envelopes. The maximum of the statistic in the centre of the search is now significant just beyond 0.01%, according to the approximation

group contains at least one identifiable subgroup that needs to be disentangled before further analysis is undertaken. Because there are only 54 observations in each group we give the bootstrap envelopes found by direct simulation using the X matrix of the observations including all variables.

9 Discussion

Several authors have used versions of the forward search for detecting outliers in regression with the minimum deletion residual (2). See Billor et al. (2000) for references

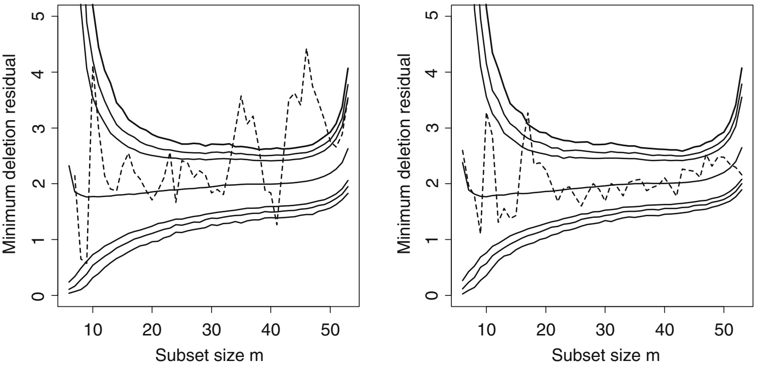

Fig. 16 Surgical Unit Data: forward plot of minimum deletion residuals for the first and second groups of 54 observations. There is strong evidence that the first group is not homogeneous

and a full discussion. However, these authors do not consider forward plots of the statistic such as those in Atkinson and Riani (2000). Our contribution here is to provide easily calculated bounds for the statistic during the forward search.

All three of our examples illustrate the importance of the bounds in inference about outliers. Figure 1 shows that, at the end of the search for Hawkins data, the value of the statistic lies within the simulation bounds. There is strong masking; considering the deletion of single observations would fail to reveal any of the outlier structure. Similar comments can be made about the surgical unit data in Fig. 13. It is during the earlier part of the search that the existence of more than one group becomes apparent. The importance of the bounds for the beta data in Fig. 11 is rather different. In the absence of bounds, the evidence of lack of fit would probably be overlooked, since this evidence rests on values of the statistic that are systematically too small for the larger residuals expected at the end of the search under the null normal distribution.

The bounds found here have approximately the correct pointwise content for known m. Atkinson and Riani (2006) describe a method that makes efficient use of the individual simulations to derive the simultaneous properties of the series of tests that occurs in the practical data-analytical case when the number of outliers is not known. This method requires the full set of simulations; the information cannot be retrieved from the approximations which are the subject of this paper.

Acknowledgments We are very grateful to the referees, editors and, especially to Professor Hans-Hermann Bock, for their very thorough reading of our paper and constructive comments which have led to an appreciable improvement in the presentation of our work.

References

Atkinson AC (1985) Plots, transformations, and regression. Oxford University Press, Oxford

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M (2006) Distribution theory and simulations for tests of outliers in regression. J Comp Graphical Statist 15:460–476

Atkinson AC, Riani M, Cerioli A (2006) Random start forward searches with envelopes for detecting clusters in multivariate data. In: Zani S, Cerioli A, Riani M, Vichi M (eds) Data analysis, classification and the forward search. Springer, Berlin, pp 163–171

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, New York

Beckman RJ, Cook RD (1983) Outlier..........s (with discussion). Technometrics 25:119–163

Billor N, Hadi AS, Velleman PJ (2000) BACON: blocked adaptive computationally efficient outlier nominators. Comp Statist Data Anal 34:279–298

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman and Hall, London

Hawkins DM (1983) Discussion of the paper by Beckman and Cook. Technometrics 25:155–156

Hawkins DM, Olive DJ (2002) Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). J Am Statist Assoc 97:136–159

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions - 1, 2nd edn. Wiley, New York

Lehmann E (1991) Point estimation, 2nd edn. Wiley, New York

Maronna RA, Yohai VJ (2002) Discussion of Hawkins and Olive (2002). J Am Statist Assoc 97:154–155

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models, 4th edn. McGraw-Hill, New York

Rousseeuw PJ (1984) Least median of squares regression. J Am Statist Assoc 79:871–880

Wisnowski JW, Montgomery DC, Simpson JR (2001) A comparative analysis of multiple outlier detection procedures in the linear regression model. Comp Statist Data Anal 36:351–382