Distribution Theory and Simulations for Tests of Outliers in Regression

Anthony C. ATKINSON and Marco RIANI

This article provides distributional results for testing multiple outliers in regression. Because direct simulation of each combination of number of observations and number of parameters is too time consuming, three straightforward methods using truncated simple samples are described for approximating the pointwise distribution of the test statistic. Scaling factors are found to adjust for the number of parameters. The same simulations also provide a powerful method of calibrating pointwise inferences for simultaneous tests for an unknown number of outliers. Analysis of data on fidelity cards reveals an unexpected group of outliers.

Key Words: Bonferroni inequality; Fast methods; Forward search; Efficient simulation; Pointwise bounds; Robust methods; Simultaneous inference.

1. INTRODUCTION

This article provides distributional results for testing multiple outliers in regression. The likelihood ratio test is equivalent to testing a deletion residual and so is the test used, for example, by Hadi and Simonoff (1993). We use distributional arguments for trimmed samples, together with simple and quick simulations, to provide very good approximations to the distribution of the test when a known number of outliers is present. We then make efficient use of our simulations to derive the simultaneous properties of the series of tests that occurs in the practical case when the number of outliers is not known.

There is, of course, a vast literature on the detection of multiple outliers in regression. See, for example, Beckman and Cook (1983) or Barnett and Lewis (1994). A serious problem is that of "masking": if there are several outliers, least squares estimation of the parameters of the model may lead to small residuals for the outlying observations. Single-deletion methods (e.g., Cook and Weisberg 1982; Atkinson 1985) may fail and the outliers will go undetected. Hawkins (1983) argued for exclusion of all possibly outlying observations,

Anthony C. Atkinson is Emeritus Professor, The London School of Economics, London WC2A 2AE, UK (E-mail: *a.c.atkinson@lse.ac.uk*). Marco Riani is Professor, Dipartimento di Economia, Università di Parma, Italy (E-mail: *mriani@unipr.it*).

Journal of Computational and Graphical Statistics, Volume 15, Number 2, Pages 460–476
DOI: 10.1198/106186006X113593

which are then tested sequentially for reinclusion.

The drawback to this procedure is that it is unclear how many observations should be deleted and, because of masking, which ones, before reinclusion and testing begin. A detailed discussion of the problem and many references were given by Billor, Hadi, and Velleman (2000). The tables of Wisnowski, Montgomery, and Simpson (2001) contain the results of an extensive simulation comparison of many outlier detection rules, in terms of detection and false alarm rates. The purpose of this article is to present the null distribution of the test statistic, together with simple approximations which are presented in cogent plots.

In order to find the null distribution we use the forward search as described by Atkinson and Riani (2000). It is an objective procedure of the type suggested by Hawkins (1983). It starts from a small, robustly chosen, subset of the data and fits subsets of increasing size. Each newly introduced observation can be tested for outlyingness before it is included in the fitted subset. By repeated simulations we can find the distribution of the test statistic when m of the n observations are fitted and the remaining $n - m$ are tested, provided m is known. Reuse of the simulations provides cheap calculation of significance levels for simultaneous inference when m is unknown.

The article is organized as follows. Section 2 briefly reviews the forward search and robust estimation; the implications for data analysis are stressed in Section 2.5. The forward search depends on estimators from trimmed samples. We therefore write the outlier test explicitly in terms of such samples and show how simulations using samples from trimmed distributions with no regression structure can be used to provide good approximations to the pointwise distribution of the statistic when m is known. Section 4 introduces a correction for the effect of regression. We then show that the Bonferroni bound is irrelevant except for finding the null distribution when testing for a single outlier. Section 6 uses an efficient simulation method to find the simultaneous content of our pointwise bounds: the pointwise significance may be as much as 10 times too large. Section 7 presents the analysis of a regression example containing more than 500 units. We conclude with comments on further work. Mathematical details are relegated to the Appendix. The emphasis throughout is methodological: we find easily calculated approximations to the distribution of the test statistic for both pointwise and simultaneous inference. Similar methods could be used to investigate the properties of the non-null distribution or the distribution of robust estimators where observations are either downweighted or trimmed.

2. THEORY

2.1 LEAST SQUARES

We start with standard results from 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 iid $N(0, \sigma^2)$.

With $\hat{\beta}$ the least squares estimator of β the vector of least squares residuals is

$$e = y - \hat{y} = y - X\hat{\beta} = (I - H)y, \tag{2.1}$$

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 mean square estimator of σ^2 can be written

$$s^2 = e^T e/(n-p) = \sum_{i=1}^{n} e_i^2/(n-p).$$

The likelihood ratio test for the agreement of observation y_i with the remaining $n-1$ observations under the normal theory assumptions is the deletion residual

$$r_i^* = \frac{y_i - x_i^T \hat{\beta}}{\sqrt{s_{(i)}^2 (1-h_i)}} = \frac{e_i}{\sqrt{s_{(i)}^2 (1-h_i)}}, \tag{2.2}$$

where $s_{(i)}^2$ is the estimate of σ^2 with observation i deleted. When the observation y_i comes from the same population as the other observations, r_i^* has a t distribution on $n-p-1$ degrees of freedom. See, for example, Atkinson and Riani (2000, p. 24). However, if y_i is chosen to be tested because it has the most extreme value of r_i^*, the observed value needs to be compared with the extreme order statistics from the t distribution. Grubbs (1950) gave the theory for a simple sample from a normal population. There is, in addition, an effect due to the slight under-estimation of σ^2 from deletion of the largest value of $|r_i^*|$ (see Figure 4, p. 469).

2.2 THE FORWARD SEARCH

Let $\mathcal{M}$ be the set of all subsets of size m of the n observations. The forward search fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. We discuss the starting point of the search in Section 2.3.

Let $S_*^{(m)} \in \mathcal{M}$ be the optimum subset of size m. Least squares applied to this subset 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 can from (2.1) be written as

$$e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*). \tag{2.3}$$

The search moves forward with the subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m^*)$. When $m < n$ the estimates of the parameters are based on only those observations giving the central m residuals; $\hat{\beta}(m^*)$ and $s^2(m^*)$ are calculated from truncated samples.

2.3 ROBUST ESTIMATION AND THE START OF THE SEARCH

The search starts from a subset of p observations $S_*^{(p)}$ that is chosen to provide a very robust estimator of the regression parameters. We use least median of squares (LMS,

Rousseeuw 1984; Rousseeuw and Leroy 1987, sec. 4.4) that is defined as minimizing the hth ordered squared residual $e^2_{[h]}(\beta)$ with respect to β. Rousseeuw and Hubert (1997) showed that taking h as the integer part of $(n+p+1)/2$ is optimum. Algorithms for LMS find an approximation to the estimator. They compare regression surfaces through all (or randomly chosen) subsets of p observations and calculate the hth ordered residual $e^2_{[h]}(\beta)$ with respect to all n observations. The subset $S^{(p)}_*$ is the set of p observations, out of the subsets searched, that yields the minimum value of $e^2_{[h]}(\beta)$. In this article we search 1,000 subsets. Provided any masking of outliers is broken, the search is insensitive to the starting method. For instance, provided $n >> p$, which is the case in our simulations, we obtain indistinguishable results if LMS is replaced by least trimmed squares (Rousseeuw 1984). What is important for our present purpose is that the search uses parameter estimates based on a central part of the sample.

2.4 TESTING FOR OUTLIERS

To test for outliers the deletion residual is calculated for the $n-m$ observations not in $S^{(m)}_*$. Analogously to (2.2) we obtain

$$r^*_i(m^*) = \frac{y_i - x_i^T\hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1+h_i(m^*)\}}} = \frac{e_i(m^*)}{\sqrt{s^2(m^*)\{1+h_i(m^*)\}}}. \tag{2.4}$$

The notation $h_i(m^*)$ serves as a reminder that the leverage of each observation depends on $S^{(m)}_*$. Let the observation nearest to those constituting $S^{(m)}_*$ be $i_{\min}$ where

$$i_{\min} = \arg\min |r^*_i(m^*)| \quad \text{for} \quad i \notin S^{(m)}_*,$$

the observation with the minimum absolute deletion residual among those not in $S^{(m)}_*$. If observation $i_{\min}$ is an outlier, so will be all other observations not in $S^{(m)}_*$.

To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r^*_{i_{\min}}(m^*) = \frac{e_{i_{\min}}(m^*)}{\sqrt{s^2(m^*)\{1+h_{i_{\min}}(m^*)\}}}. \tag{2.5}$$

The distribution of this statistic is the subject of this article.

2.5 IMPLICATIONS

The purpose of the robust starting point in Section 2.3 is to provide a "clean" subset of the data, free of outliers. The procedure for moving forward in Section 2.2 ensures that we always take the observations with the m smallest residuals, so that the set remains outlier free for as long as possible.

At each step of the search we test whether observation $i_{\min}$ is an outlier relative to the m observations $\in S^{(m)}_*$. Because we have constructed $S^{(m)}_*$ to be outlier free we are testing

whether one observation is an outlier relative to a set of m nonoutlying observations. The presence of the leverage measure $h_{i_{\min}}(m^*)$ in (2.5) ensures that the test only depends on whether the outlier has a shifted mean relative to the observations in $S_*^{(m)}$. The position of the remaining $n - m - 1$ potential outliers relative to the rest of the data affects the joint properties of the residuals for these observations. But, by moving forward until we encounter the first outlier we are breaking this complicated structure by extracting only the least significant outlier. The unraveling effect of the forward search on a complicated outlier structure is shown, for example, by Atkinson and Riani (2000, fig. 3.6).

In most applications there will only be a few outliers and interest will be in the end of the search. However, the data may contain unsuspected subsets. For example, Neter, Kutner, Nachtsheim, and Wasserman (1996, p. 334, p. 347) introduced two sets of 54 observations on the logged survival time of patients undergoing liver surgery. Interest in a forward search to check that the same model is suitable for the two groups of patients focuses on the center of the search through the 108 combined observations. Again, in most examples, it will be the upper simulation envelope that is important. However, an incorrect error distribution may make the lower envelope important; a regression model with U-shaped beta errors is an example.

3. SIMULATING THE DISTRIBUTION

The null distribution of (2.5) can be found by simulating numerous forward searches. However, if a single search, with the starting procedure described in Section 2.3, takes one minute, 10,000 searches take almost exactly one week. Our three alternatives rely on simple, and much faster, ways of simulating variables with approximately the same distribution as the $e_i(m^*)$. When $m = n$ these residuals are distributed $N\{0, (1-h_i)\sigma^2\}$. But with $m < n$ the estimates of the parameters come from truncated samples. We use methods based on truncated simple samples, when all $h_i(m^*) = 1/m$, prove results about the equivalence of the distribution to the required one and introduce a small sample correction factor that allows for the effect of p. We thus require only one set of simulations for any n, regardless of the dimension of the fitted linear model.

3.1 THE EMPIRICAL DISTRIBUTION

Figure 1 shows the distribution of the outlier test, the absolute minimum deletion residual $|r^*_{i_{\min}}(m^*)|$, from 10,000 simulations of a regression model when $n = 100$ and $p = 3$. The values of the explanatory variables were independent standard normal random variables as were the errors ϵ. Because least squares residuals do not depend on the value of the parameters β, we took these as zero. The curves in the figure are the 1, 2.5, 5, 50, 95, 97.5, and 99% points of the empirical distribution. This shape is characteristic of all those we shall see: initially the bands are wide where the variance is estimated on a few degrees of freedom. The central part of the simulation is very stable, trending up only slightly. At the end of the search the observations with large residuals enter the search and both lower and

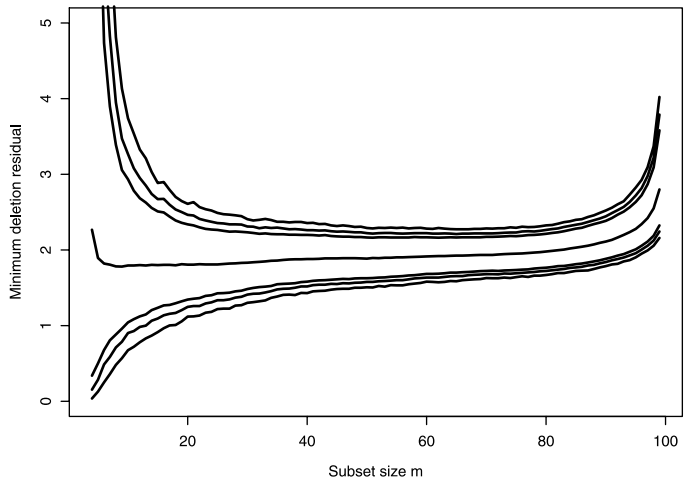

Figure 1. Empirical percentage points of the test for outliers for different levels of trimming: 10,000 forward searches, n = 100, p = 3.

upper bands increase rapidly. As we shall see the values for $m = n-1$ can be approximated by the Bonferroni inequality.

3.2 METHOD 1: TRUNCATED SAMPLES

The statistic is a function of the m residuals $e_i(m^*) \in S_*^{(m)}$ and of $e_{i_{\min}}(m^*)$. In the absence of outliers, these will be the observations with the $m+1$ smallest values of $|e_i(m^*)|$ (see Appendix A.1). As the $e_i(m^*)$ are residuals, their distribution does not depend on the parameters β of the linear model. Since all observations in the simple sample have the same leverage, all residuals have the same variance, which is to be estimated. To find the required distribution we therefore simulate from a truncated normal distribution and repeatedly calculate the value of the outlier test for such samples. The steps are:

Step 1. Obtain a random sample of $m+1$ observations U_i from the uniform distribution on $[0.5-(m+1)/2n, 0.5+(m+1)/2n]$.

Step 2. Use the inversion method to obtain a sample of $m+1$ from the truncated normal distribution:

$$z_i = \Phi^{-1}(U_i),$$

where Φ is the standard normal cdf

Step 3. Find the most outlying observation:

$$z_{i_{\min}} = \max |z_i| \quad i = 1, \ldots, m+1.$$

Then $S_z^{(m)} = \{z_i | i \neq i_{\min} = 1, \ldots, m+1\}$.

Step 4. Estimate the parameters. Let $\bar{z}(m)$ be the mean of the m observations in $S_z^{(m)}$ and $s_z^2(m)$ be the mean square estimate of the variance.

Step 5. Calculate the simulated value of the outlier test in (2.5):

$$r^z_{i_{\min}}(m) = \frac{z_{i_{\min}} - \bar{z}(m)}{s_z(m)\sqrt{\{(m+1)/m\}}} = \frac{e^z_{i_{\min}}(m)}{s_z(m)\sqrt{\{(m+1)/m\}}}, \tag{3.1}$$

where $e^z_i(m) = z_i - \bar{z}(m)$.

The simulation of the truncated normal distribution using the inversion method in Steps 1 and 2 is straightforward in S-Plus or R. Since in (3.1) we have estimated the sample mean, rather than a regression model, $h_{i_{\min}}(m) = 1/m$.

In (3.1) $r^z_{i_{\min}}$ is based on a random sample, rather than the regression residuals in (2.4). We have the following:

Theorem 1. *The distributions of the outlier tests $r^z_{i_{\min}}(m)$ (3.1) and $r^*_{i_{\min}}$ (2.5) tend at the same rate to the same limit as $n \to \infty$ with m/n fixed.*

The Theorem is proved in Appendix A.1.

3.3 Method 2: Ordered Observations

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. As a second method of approximating the distribution of the statistics, we simulate sets of n observations from the normal distribution, correct for the mean to give residuals $e^z(n)$ and order the absolute values of the residuals just once, before calculations begin. For each value of m we use the m smallest values of the $|e^z(n)|$ to estimate the parameters; in the outlier test (2.5) $i_{\min} = i_{[m+1]}$. The procedure, which avoids a forward search for each simulation, is repeated many times to give the empirical distribution of the outlier test.

3.4 Method 3: Order Statistics

Since (3.1) is a function of order statistics, we can use expected values of normal order statistics to approximate the confidence intervals found by simulation.

Theorem 2. *Approximate $100(1-\alpha)\%$ confidence intervals for the outlier test from (3.1) are given by*

$$\frac{\zeta_{m+1,n} \pm \sigma_\xi \Phi^{-1}(\alpha/2)}{\sigma_T \sqrt{\{(m+1)/m\}}}. \tag{3.2}$$

where

$$\zeta_{m+1,n} = \Phi^{-1}\left\{\frac{8n+8m+7}{4(4n+1)}\right\},$$

$$\sigma^2_\xi = \frac{(8m+5)(8n-8m-3)}{4n(4n+1)^2\phi^2\left\{\Phi^{-1}\left(\frac{8m+5}{2(4n+1)}\right)\right\}},$$

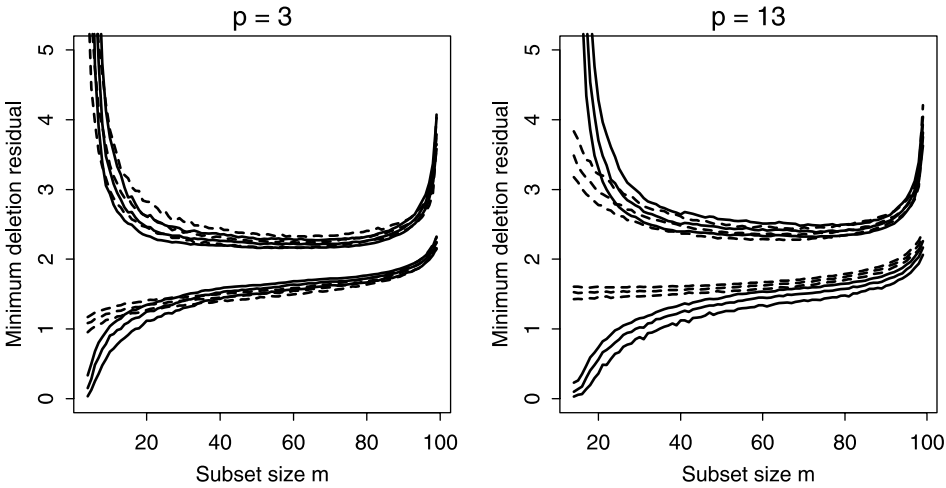

Figure 2. Percentage points of the test for outliers for different levels of trimming: 10,000 forward searches, n = 100. Dotted lines, Method 1, adjusted for regression. Continuous lines, empirical curves as in Figure 1. The adjustment works well for the upper percentage points.

and

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

The proof is in Appendix A.2.

4. ADJUSTMENT FOR REGRESSION

In the three methods of Section 3 we estimate the sample mean, rather than a regression model so $h_{i_{\min}}(m) = 1/m$. Simulations show that the shapes of the sets of curves are similar to those in Figure 1, but that the variance is too small when we are analyzing regression data, for which the average value of the h_i is p/m. Multiplication of (3.1) by $\sqrt{(1+p/m)}$ results in a slight overcorrection. Empirically we find good agreement for the upper percentage points of the distribution by using the adjusted statistic

$$r_{i_{\min}}^{\text{adj}}(m) = \sqrt{\frac{m+\theta p}{m}}\,\frac{z_{i_{\min}} - \bar{z}(m)}{s_z(m)},$$

with $\theta = 0.7$. The lower percentage points are not improved by this correction. However, our interest is in detecting outliers, which give large values of the statistic. As m increases, the effect of the correction decreases.

Figure 2 shows forward plots of the percentage points of the distribution of the test when $n = 100$. The continuous lines are from 10,000 empirical simulations of a forward search, whereas the dotted lines uses the faster approximation of Method 1. In the left-hand panel, $p = 3$, and there is good agreement between the empirical method and the approximation, except for the lower percentage points when m is small. In the right-hand panel, for $p = 13$,

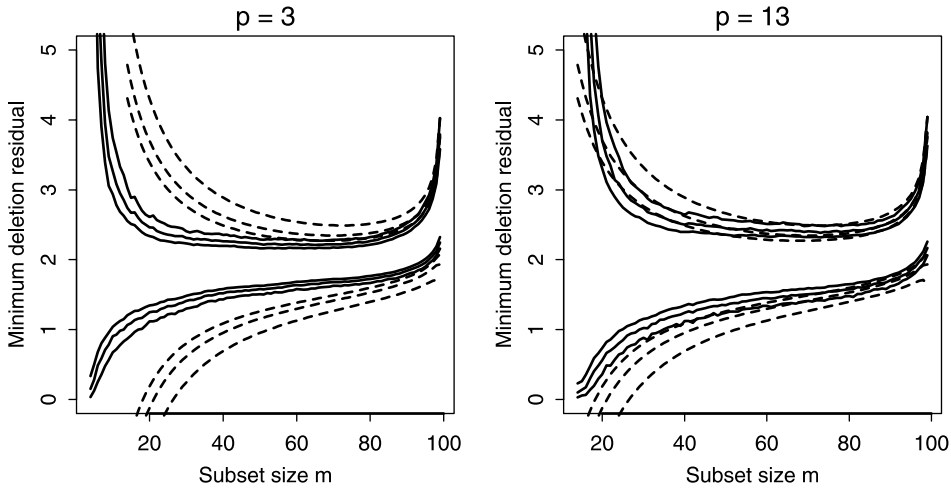

Figure 3. Percentage points of the test for outliers for different levels of trimming: 10,000 forward searches, n = 100. Dotted lines, Method 3, adjusted for regression. Continuous lines, empirical curves as in Figure 2. The adjustment works well for the upper percentage points when p = 13.

the approximation is only satisfactory for the upper percentage points, improving steadily with m.

A similar plot in Figure 3, using Method 3, the expected values of order statistics, shows that this method also works well for the upper percentage points when $p = 13$. The plot for Method 2 is virtually indistinguishable from that for Method 1 and so is not given here.

5. THE IRRELEVANCE OF BONFERRONI BOUNDS

The statistic (2.5) tests the $m+1$st largest residual for outlyingness. Hadi and Simonoff (1994) used a Bonferroni bound to allow for the ordering of the residuals during their forward search and compare (2.5) with the percentage points of $t_{m-p}\{\alpha/2(m+1)\}$ using $s^2(m^*)$ to estimate the error variance.

The left-hand panel of Figure 4 shows the resulting 1, 2.5, and 5% bounds when $p = 3$ and $n = 100$, which are unrelated to the true distribution, apart from the last step of the search; due to the low correlation of the residuals the bound is almost exact when $m = n-1$, except for the slight effect of underestimation of σ^2 mentioned in Section 2.1. Earlier in the search the bounds are far too large, because $s^2(m^*)$ is treated as an estimate from a full sample, rather than from the truncated sample that arises from the ordering of the residuals. The surprising flatness of the bounds is caused by compensating changes in the percentage point $\alpha/2(m+1)$ and the degrees of freedom $m-p$.

Wisnowski et al. (2001, p. 360) reported that the procedure of Hadi and Simonoff (1993) has a low detection rate for moderate and small outliers and an abnormally low false alarm rate. The shape of the bounds in the left-hand panel of the figure shows that this is

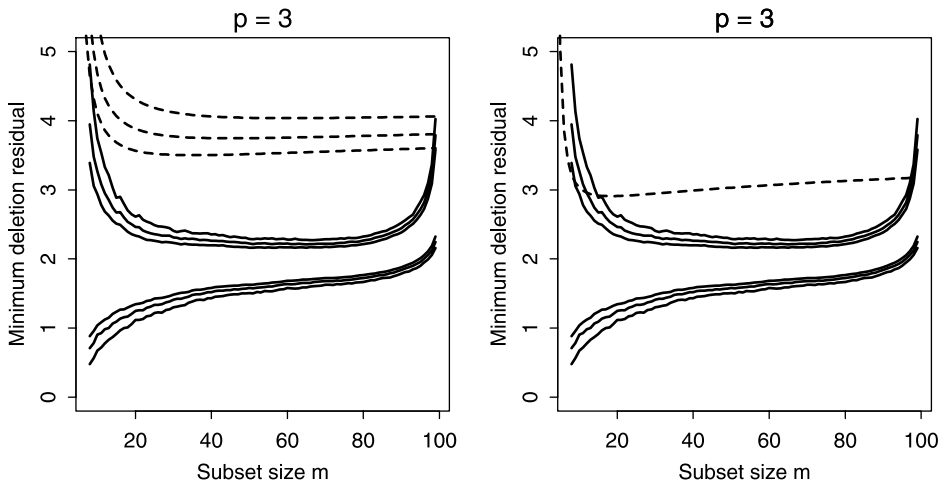

Figure 4. Percentage points of the test for outliers for different levels of trimming when $p = 3$. Dotted lines, Bonferroni bounds; continuous lines, empirical curves as in Figure 1. Left-hand panel: 1, 2.5, and 5% bounds; right-hand panel: 20% bound recommended by Wisnowski et al. (2001).

caused by rejection bands that are far too wide. Wisnowski et al. (2001) suggested using a value of 20% for α. The right-hand panel of Figure 4 shows the resulting bounds. These are still too wide in the centre of the search, although less so than before, but are now too narrow at the end. As would be expected from this plot, the power of the procedure increases, but so does the false alarm rate, as the bound is too tight when there are few, or no, outliers. The conclusion is that Bonferroni bounds do not provide a useful guide for the values of the statistic for testing a general number of outliers.

6. SIMULTANEOUS INFERENCE

Our methods provide approximations to the point-wise confidence intervals found in Section 3.1 from 10,000 simulations of the forward search. We now use and extend an argument by Buja and Rolke (2003) to make simultaneous inference about the significance of the values of the test statistic observed over the search, or part of it. We find there is a surprisingly high probability that, for example, any simulated curve is significant at the 5% point at least once.

We perform N forward searches on simulated data. The ith simulation yields a curve of simulated values of the statistic (2.5). The pointwise bounds in, for example, Figure 1 (p. 465) are found by ranking these values for each m. The ranked results of the ith simulation are then a curve of values $R_{im}, m \in [p, n-1], i = 1, \ldots, N$. For each i we find the maximum of R_{im} for $m \in \mathcal{F}$, where $\mathcal{F}$ can be only part of the forward search. Let this value for the ith search be $R_i^{\mathcal{F}}$. We thus obtain the empirical distribution of the $R_i^{\mathcal{F}}$, that is of the maximum nominal significance level of $R_i^{\mathcal{F}}$ for each search. The nominal percentage significance level of these values is, to a good approximation, $100\{(R_i^{\mathcal{F}} - 3/8)/(N + 1/2)\}$.

Table 1. Simultaneous Inference: Nominal and Actual Significance Levels for the Last Half of the Search When $n = 100$

	Rank of statistic: j =				*Consecutive: k =*	
Nominal α	*1*	*2*	*3*	*6*	*2*	*3*
0.01	0.196	0.112	0.066	0.019	0.091	0.046
0.05	0.552	0.410	0.320	0.164	0.352	0.232

The empirical percentage level depends on the number of searches N_i which have values of $R_l^{\mathcal{F}} \geq R_i^{\mathcal{F}}, l = 1, \ldots, i-1, i+1, \ldots, N$. The empirical percentage level is therefore $100\{(N_i - 3/8)/(N + 1/2)\}$. These sets of percentage points give Figure 5 and Table 1 when $\mathcal{F}$ is the last half of the search.

These results for 10,000 simulations when $n = 100$ and $p = 3$ show the extremely large effect of simultaneous inference. The topmost curve in the left-hand panel of Figure 5 shows the true significance level rapidly increasing with the nominal level. The right-hand panel plots the same curve for the nominal values of up to 10% that are of interest in statistical inference. The entries in Table 1 show that actual levels are at least ten times the nominal levels.

In addition to the largest value of R_{im} for each i we can also find the jth largest value. In Figure 5 we plot values up to $j = 6$. Increasing j decreases the true probability level, but even the 1% nominal level for the sixth largest value has a true probability that is almost twice the nominal value. The discrepancy is greater for the 5% point.

In analyzing forward plots we are often interested in the interpretation of broad peaks. The method extends straightforwardly to successive k-tuples. Let $m_k = (m, \ldots, m+k-1)$.

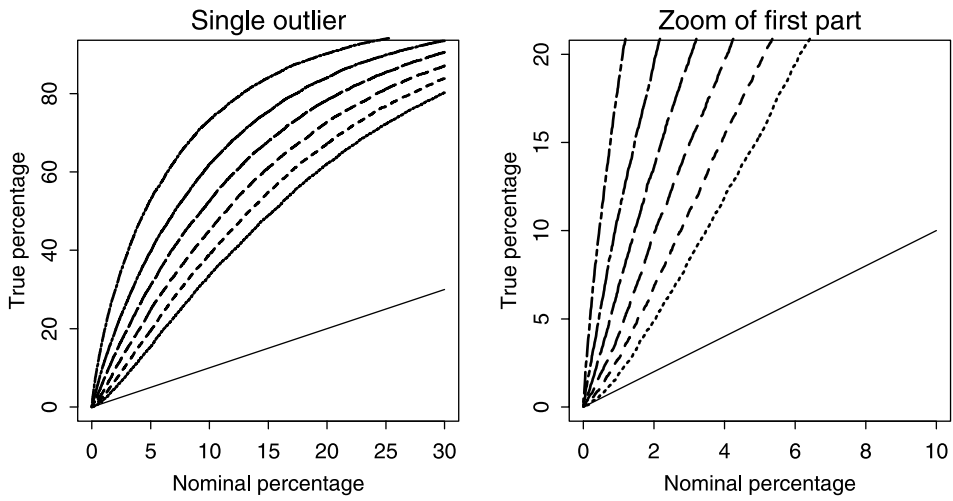

Figure 5. Nominal and simultaneous significance levels of the outlier test for a single maximum, n= 100, p = 3, $\mathcal{F}$ is the last half of the search. The curves, from the top downwards, are for the jth largest value, j = 1,. . .,6. The nominal and true values would be the same on the straight line. Right-hand panel: zoom for values useful in inference.

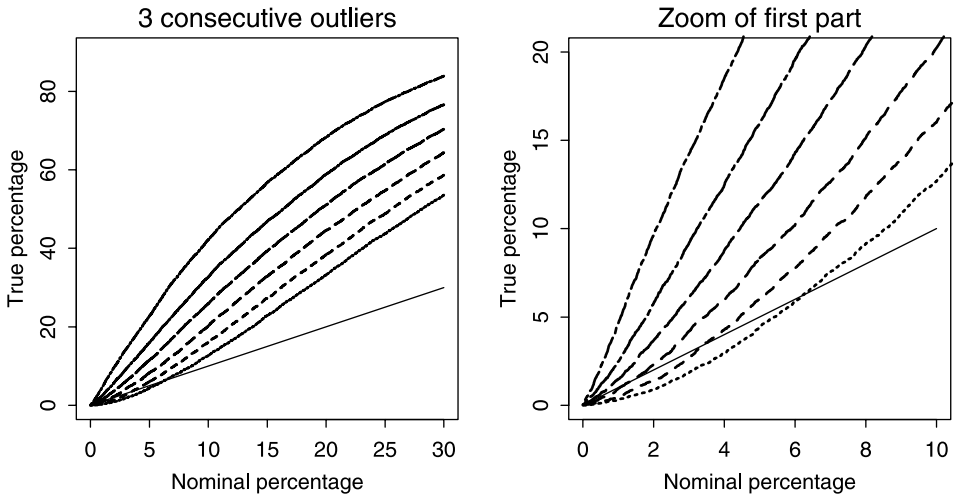

Figure 6. Nominal and simultaneous significance levels of the outlier test for peaks of three observations, $n = 100$, $p = 3$, $\mathcal{F}$ is the last half of the search. The curves, from the top downwards, are for the jth largest value, $j = 1, \ldots, 6$. The nominal and true values would be the same on the straight line. Right-hand panel: zoom for values useful in inference.

Then for each $m_k \in \mathcal{F}$ we find $R_{ik}^{\mathcal{F}}$, the minimum value of R_{im}. These k-tuples are overlapping. We then proceed as before, but finding the empirical distribution of the jth largest value of $R_{ik}^{\mathcal{F}}$ rather than of $R_i^{\mathcal{F}}$. We base $R_{ik}^{\mathcal{F}}$ on the minimum value of R_{im} over each m_k because we are interested in the least significant member of the k-tuple.

The results for $k = 3$ are in Figure 6 and Table 1. Comparison of the left-hand panels of the two plots shows that the true significance for triples increases appreciably more slowly with the nominal level than it does for single values, although the numbers are still large. The right-hand panel indicates that the nominal 1% level for the fourth largest triple is close to the true level: however, at the 5% level, the sixth largest triple is closest to correct.

7. LOYALTY CARDS

As an example to show the use of our simulation envelopes we take 509 observations on the behaviour of customers with loyalty cards from a supermarket chain in Northern Italy. The data are themselves a random sample from a larger database. The sample of 509 observations is available at *www.riani.it/trimmed*. The response is the amount, in euros, spent at the shop over six months and the explanatory variables are: x_1, the number of visits to the supermarket in the six month period; x_2, the age of the customer and, x_3, the number of members of the customer's family.

The data need transformation to achieve constant variance. We use the Box-Cox power transformation. If we confine our attention to values of the transformation parameter which are ratios of small integers (Box and Cox 1964), the indication is that transformation to the

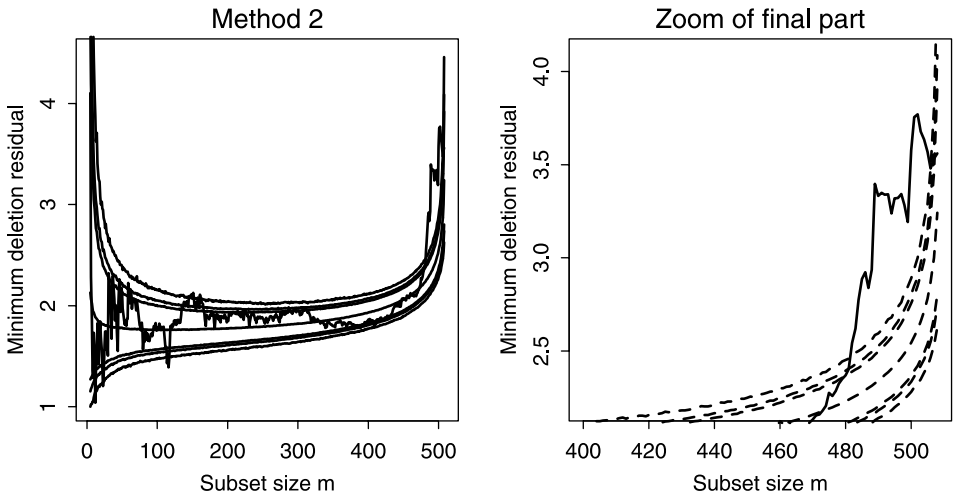

Figure 7. Loyalty cards, transformed data. The test for outliers with envelopes constructed with Method 2. The zoom in the right-hand panel shows the effect of the cluster of outliers which are masked when $m = n$.

power 1/3 is needed; we work with this transformation for the rest of our analysis. However, the forward plot of the asymptotically standard normal test for this transformation shows that this value would be rejected for m around $n - 30$. For a description of the plot (called a "fan" plot) see Atkinson and Riani (2000, sec. 4.3), with distributional results in Atkinson and Riani (2002).

We now see whether all observations agree with this model. Figure 7 is the forward plot of the test for outliers. Because of the large number of observations we need to use one of our approximate methods. Here we show results for Method 2 in which simulated samples are ordered only once. Use of Method 1 gives plots that seem to us indistinguishable from these. It is clear that, at the end of the search, there are several outliers, even when allowing for the results of Section 6 on simultaneous inference for k-tuples. The outliers consist of the last observations to enter the search which lie above the 99% simulation envelope. What is interesting in the right-hand panel is that the final value of the statistic when $m = n - 1$ lies inside the envelope, so that the outliers are masked. The left-hand panel shows that for the second half of the search, apart from the end, the outlier statistics wanders around in a way that the results of Section 6 indicate are not significant.

The observations we have found are outlying in an interesting way. The left-hand panel of Figure 8 is a scatterplot of y against x_1. There is both some evidence of a relationship between the variables and evidence of heteroscedasticity. The right-hand panel of the figure is the scatterplot of the transformed response $y^{1/3}$ against x_1 (frequency) with the last 26 units to enter the forward search highlighted. We have identified a subset of individuals, most of whom are behaving in a strikingly different way from the majority of the population. The existence of such a group, who is spending less than would be expected, will be important in any further modeling of the data.

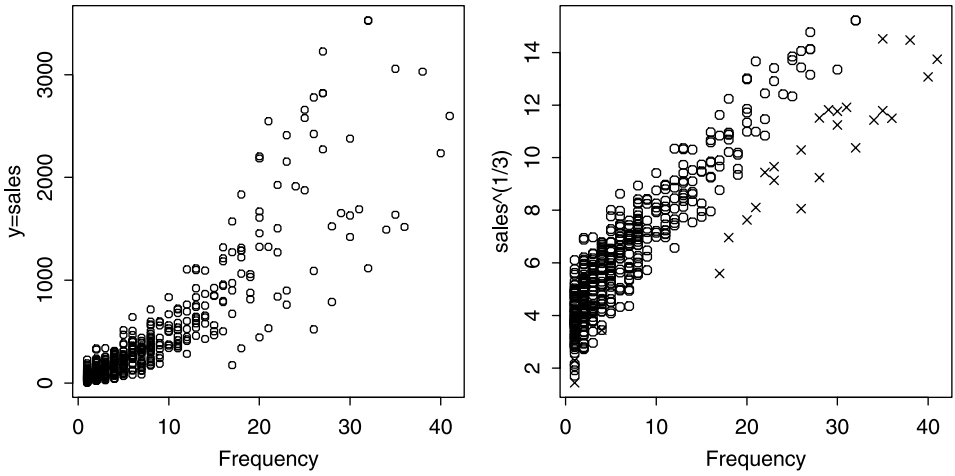

Figure 8. Loyalty cards: response against x_1 (frequency). Left-hand panel, original data; right-hand panel, transformed response, clearly showing the outliers. The last 26 units to enter the subset are plotted with crosses.

8. DISCUSSION

We have used three simple simulation methods to find the pointwise distribution of the outlier test in regression and then reused our simulations to investigate the simultaneous properties of these intervals. Our method could be used to confirm the number of outliers by, in the example of Section 7, simulating envelopes for the remaining $509 - 26 = 483$ observations and seeing whether the increase in the observed values visible in the right-hand panel of Figure 7 around $m = 480$ lies within the sharply increasing envelope at the end of this search.

In general, the forward search has provided an insightful framework for understanding the series of tests that occur when the number of outliers is unknown. We could extend our simulations in many ways, for example to the null distribution of the test defined by the boundaries of the modified Bonferroni rule in Figure 4 (p. 469). The simulations can also explore the distribution of many other quantities such as robust estimators that are derived from truncated samples. We also see important applications in the development and study of rules for the automatic clustering and classification of multivariate data. Finally, we note that detection of multiple outliers in regression is as complicated as, and has much in common with, the use of control charts for the detection of outliers from a mean over time; a variety of rules are required depending on whether one or a few large outliers are to be detected as opposed to a smaller change that affects several observations.

APPENDIX: PROOFS OF THEOREMS

A.1 THEOREM 1, METHOD 1: TRUNCATED SAMPLES

Progression of the Search. In normal progression $S_*^{(m+1)}$ is formed by augmenting

$S_*^{(m)}$ with observation $i_{\min}$ and the residuals come from the central part of a truncated distribution. However if, for example, the first of a group of outliers is included in $S_*^{(m)}$, some other members of the outlying group $\notin S_*^{(m)}$ may have small residuals $e_i(m^*)$. Due to this masking several of these observations may be included in $S_*^{(m+1)}$, while other observations are excluded. Such interchanges only occur in the presence of outliers and so can be ignored when we are simulating to find the null distribution of r_i^*. It is therefore appropriate, as we have done, to sample from the center of a truncated distribution. A thorough discussion of interchanges in the forward search was given by Atkinson, Riani, and Cerioli (2004, pp. 68–69).

Distribution of Residuals. When $m = n$, the residuals $e_i(n)$ in (2.3) are distributed $N\{0, (1 - h_i)\sigma^2\}$. For $m < n$ the least squares estimates $\hat{\beta}(m^*)$ are linear combinations of order statistics of normal random variables and are unbiased estimators of β, so that $E\{e_i(m^*)\} = 0$. We can therefore approximate the distribution of the $e_i(m^*)$ by sampling from a distribution with zero mean. However, the distribution of the $\hat{\beta}(m^*)$ is not exactly normal, so, although the y_i are normal, the residuals $e_i(m^*)$ will only be asymptotically normal. For results on the distribution of linear combinations of order statistics see, for example, Csörgö (1983) or Bening (2000).

Theorem 3. *The variance of $e_i(m^*)$ depends on the leverage $h_i(m^*)$ as does the correlation of $e_i(m^*)$ and $e_j(m^*)$ on the off-diagonal element $h_{ij}(m^*)$ of the hat matrix $H(m^*)$. Now $trH(m^*) = p$, so that $\bar{h}(m^*)$, the average value of $h_i(m^*)$, is p/m. Let $\psi = m/n$, which is taken to stay constant with increasing n. Then $\bar{h}(m^*) = p/(n\psi)$, which decreases with increasing n. The off-diagonal elements $h_{ij}(m^*)$ likewise decrease with n.*

For the simulated sample $\{z_i\}$, $h_i(m) = h_{ij}(m) = 1/m = 1/(n\psi)$, so that the mean and variance of the $e_i^z(m)$ converge at the same rate as the $e_i(m^*)$ to the mean and variance of the $e_i(m^*)$. Both sets of variables are also asymptotically normal as the effect of estimation of the parameters decreases with n.

We are not directly concerned with the distribution of these residuals themselves, but with convergence of the distribution of the test statistic $r^z_{i_{\min}}$ to that of $r_{i_{\min}}$. Since the two statistics are the same function of the two sets of residuals, they will converge at the same rate to the same limit. Our theorem is thus proved. Our simulated results show that this convergence happens for relatively small sample sizes.

A.2 THEOREM 2, METHOD 3: ORDER STATISTICS

We use normal order statistics to approximate the individual terms of (3.1).

In the numerator of (3.1) $z_{i_{\min}}$ is the $m + 1$st largest order statistic of absolute values from a sample of size n from a standard normal distribution. Its expectation is (Cox and Hinkley 1974, p. 470) approximately

$$\zeta_{m+1,n} = \Phi^{-1}\{0.5 + 0.5(m + 1 - 3/8)/(n + 1/4)\}.$$

If we let $p_\xi = (m + 1 - 3/8)/(n + 1/4)$ and $\xi_{m+1,n} = \Phi^{-1}(p_\xi)$, the variance of z_{imin}

(Stuart and Ord 1987, p. 331) is

$$\sigma_\xi^2 = p_\xi(1-p_\xi)/\{n\phi^2(\xi_{m+1,n})\}.$$

Because $E(\bar{z}) = 0$, the asymptotic $100(1-\alpha)\%$ normal theory confidence limits for the numerator of (3.1) are

$$\zeta_{m+1,n} \pm \sigma_\xi \Phi^{-1}(\alpha/2). \tag{A.1}$$

We also require $E\{s_z^2(m)\}$, the estimated variance of the truncated sample containing the central m/n portion of the full distribution. Let

$$a^- = 0.5 - m/(2n) \quad \text{and} \quad a^+ = 0.5 + m/(2n),$$

so that

$$y^- = \Phi^{-1}(a^-) \quad \text{and} \quad y^+ = \Phi^{-1}(a^+).$$

Because of symmetry $y^- = -y^+$. The truncated normal distribution then has density

$$\phi_T(y) = \phi(y)/(a^+ - a^-) = n\phi(y)/m, \qquad -y^+ \le y \le y^+.$$

The mean is zero and the variance

$$\sigma_T^2 = \left\{1 - \frac{y^+\phi(y^+) - y^-\phi(y^-)}{\Phi(y^+) - \Phi(y^-)}\right\} = \{1 - 2ny^+\phi(y^+)/m\}. \tag{A.2}$$

Asymptotically the numerator and denominator of (3.1) will be independent. If we take the expectation of the ratio to be the ratio of the expectations, substitution of (A.1) and (A.2) in (3.1) yields our required result (3.2).

[Received September 2004. Revised July 2005.]

REFERENCES

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

Atkinson, A. C., and Riani, M. (2000), *Robust Diagnostic Regression Analysis*, New York: Springer–Verlag.

——— (2002), "Tests in the Fan Plot for Robust, Diagnostic Transformations in Regression," *Chemometrics and Intelligent Laboratory Systems*, 60, 87–100.

Atkinson, A. C., Riani, M., and Cerioli, A. (2004), *Exploring Multivariate Data with the Forward Search*, New York: Springer–Verlag.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data* (3rd ed.), New York: Wiley.

Beckman, R. J., and Cook, R. D. (1983), "Outlier.....s" (with discussion), *Technometrics*, 25, 119–163.

Bening, V. E. (2000), *Asymptotic Theory of Testing Statistical Hypotheses: Efficient Statistics, Optimality, Power Loss, and Deficiency*, Dordrecht: VSP.

Billor, N., Hadi, A. S., and Velleman, P. J. (2000), "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators," *Computational Statistics and Data Analysis*, 34, 279–298.

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

Buja, A., and Rolke, W. (2003), "Calibration for Simultaneity: (Re)sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data," Technical report, The Wharton School, University of Pennsylvania.

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

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Csörgö, M. (1983), *Quantile Processes with Statistical Applications*, Philadelphia: SIAM.

Grubbs, F. E. (1950), "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, 21, 27–57.

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. (1983), Discussion of "Outlier.....s" by Beckman and Cook, *Technometrics*, 25, 155–156.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models* (4th ed.), New York: McGraw-Hill.

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

Rousseeuw, P. J., and Hubert, M. (1997), "Recent Developments in PROGRESS," in L_1-*Statistical Procedures and Related Topics*, Volume 31 of *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, ed. Y. Dodge, Hayward, CA: IMS, pp. 201–214.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Stuart, A., and Ord, K. J. (1987), *Kendall's Advanced Theory of Statistics* (Vol. 1, 5th ed.), London: Griffin.

Wisnowski, J. W., Montgomery, D. C., and Simpson, J. R. (2001), "A Comparative Analysis of Multiple Outlier Detection Procedures in the Linear Regression Model," *Computational Statistics and Data Analysis*, 36, 351–382.