Computational Statistics and Data Analysis 54 (2010) 3300–3312

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust model selection with flexible trimming

Marco Riani [a], Anthony C. Atkinson [b,*]

[a] Università di Parma, Via Kennedy 6, I-43100, Parma, Italy

[b] Department of Statistics, London School of Economics, London WC2A 2AE, UK

ARTICLE INFO

Article history:
Received 31 January 2009
Received in revised form 18 February 2010
Accepted 9 March 2010
Available online 20 March 2010

Keywords:
Candlestick plot
C_p
$C_p(m)$
Distributional robustness
F distribution
Forward search
M-estimation

ABSTRACT

The forward search provides data-driven flexible trimming of a C_p statistic for the choice of regression models that reveals the effect of outliers on model selection. An informed robust model choice follows. Even in small samples, the statistic has a null distribution indistinguishable from an F distribution. Limits on acceptable values of the C_p statistic follow. Two examples of widely differing size are discussed. A powerful graphical tool is the generalized candlestick plot, which summarizes the information on all forward searches and on the choice of models. A comparison is made with the use of M-estimation in robust model choice.

1. Introduction

Mallows' C_p is widely used for the selection of a model from among many non-nested regression models. However, the statistic is a function of two residual sums of squares; it is an aggregate statistic, a function of all the observations. Thus C_p suffers from the well-known lack of robustness of least squares and provides no evidence of whether or how individual observations or unidentified structure are affecting the choice of model. In our paper we use the robustness of the data-driven flexible trimming provided by the forward search to choose regression models in the presence of outliers. Our tools are new distributional results on C_p in the forward search, and a powerful new version of the C_p plot, which we call a generalized candlestick plot. Software for our analysis is referenced in Section 6.

We present aggregate C_p and its scaled and shifted F distribution in Section 2. In Section 3.1 we recall the forward search, which achieves robustness by fitting models to carefully chosen subsets of observations of increasing size, generically m. This flexible trimming provides the forward search statistic $C_p(m)$ which reveals how the choice of a model depends upon the observations included in the subset. A second robust method due to Ronchetti and Staudte (1994), using M estimates, is described in Section 3.2. Theoretical and numerical results on the distribution of $C_p(m)$ are in Section 4. Surprisingly, the ordering of the observations and the truncation of their distribution has no observable impact on the null distribution of $C_p(m)$, which is effectively scaled and shifted F.

Our major example is the analysis of 80 observations on ozone concentration in Section 5. The robust and non-robust aggregate analyses are in Section 5.1. In Section 5.2 we provide forward plots of $C_p(m)$ for one value of p. These plots can be hard to interpret. In order to summarize this information and to allow comparisons over a range of values of p we introduce, in Section 5.3, a powerful new graphic which we call a "generalized candlestick" plot. In Section 5.4 we use this plot to

* Corresponding author.
E-mail addresses: mriani@unipr.it (M. Riani), a.c.atkinson@lse.ac.uk (A.C. Atkinson).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.03.007

identify the inferential effect of the observations entering at the end of the searches for the various models. There seem to be two observations that have a large effect on the value of $C_p(m)$ and so on the choice of model. These observations are confirmed as outliers in Section 5.5. The analysis concludes in Section 5.6 with comparisons of the best models when the outliers are excluded.

A much larger example, with 1000 observations, is mentioned in Section 6, where we discuss methods for problems with many explanatory variables. The paper concludes in Section 7 with an indication of extensions, in particular to the analysis of time series data. We stress throughout, that the purpose is not to choose a single model, but to obtain understanding of the interaction between model choice and outliers, so that we are guided to the selection of a suitable model.

2. Aggregate C_p

There are n univariate observations y. For the linear multiple regression model $y = X\beta + \epsilon$, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. The residual sum of squares from fitting this model to the data is $R_p(n)$.

In the selection of regression variables, σ^2 is estimated from a large regression model with an $n \times p^+$ matrix X^+, $p^+ > p$, of which X is submatrix. The unbiased estimator of σ^2 comes from regression on all p^+ columns of X^+ and can be written

$$s^2 = R_{p^+}(n)/(n - p^+). \tag{1}$$

Then

$$C_p = R_p(n)/s^2 - n + 2p = (n - p^+)R_p(n)/R_{p^+}(n) - n + 2p. \tag{2}$$

One derivation of C_p (Mallows, 1973) is that it provides an estimate of the mean squared error of prediction at the n observational points from the model with p linear parameters, provided the full model with p^+ parameters and the reduced model with p parameters yield unbiased estimates of σ^2. Then $\mathrm{E}\{R_p(n)\} = (n-p)\sigma^2$, $\mathrm{E}(s^2) = \sigma^2$ and $\mathrm{E}(C_p)$ is approximately p.

Models with small values of C_p are preferred. Statements are often made that those models with values of C_p near p are acceptable. However, it can be helpful also to consider the distribution of the statistic.

This distribution is given, for example, by Mallows (1973) and by Gilmour (1996). From (2) we require the distribution of the ratio of two nested residual sums of squares. It is straightforward to show that the required distribution is

$$C_p \sim (p^+ - p)F + 2p - p^+, \quad \text{where } F \sim F_{p^+-p,\, n-p^+}. \tag{3}$$

In short, if

$$F^* \sim F_{\nu_1,\nu_2}, \quad \mathrm{E}(F^*) = \nu_2/(\nu_2 - 2). \tag{4}$$

Then, from (3),

$$\mathrm{E}(C_p) = p + 2\frac{p^+ - p}{n - p^+ - 2}. \tag{5}$$

As $n \to \infty$, $\mathrm{E}(C_p) \to p$. Gilmour comments that when $n - p^+$ is small, $\mathrm{E}(C_p)$ can be appreciably greater than p. Hurvich and Tsai (1989) use Taylor series expansions to find a correction for this bias for very small n.

3. Robust approaches

3.1. The forward search and forward C_p

Individual outliers and influential observations in regression models can be detected by the single deletion methods described in the books of Cook and Weisberg (1982) and of Atkinson (1985). However, these procedures may fail to reveal multiple outliers, due to masking in which the outliers so affect the parameter estimates as to seem part of the main body of the data. Multiple outliers can be detected by use of the Forward Search. This method moves from fitting small, robustly chosen, subsets of the data to fitting all n observations in such a way that unsuspected structure is revealed and outliers, if any, enter the subset to be fitted towards the end of the search.

The forward search for a single regression model fits subsets of observations of size m to the data, with $m_0 \le m \le n$. 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 derived from the residual sum of squares $R_p(m)$. Residuals $e_i(m)$ can be calculated for all observations including those not in $S^{(m)}$. The search moves forward with the augmented subset $S^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$, $i = 1, \ldots, n$.

To start we take $m_0 = p + 1$ and so search over subsets of $p + 1$ observations to find the subset, out of 3000, that yields the least median of squares (LMS) estimate of β (Rousseeuw, 1984).

The C_p criterion (2) for all observations is a function of the residual sums of squares $R_p(n)$ and $R_{p^+}(n)$. For a subset of m observations we can define the forward value of C_p as

$$C_p(m) = (m - p^+)R_p(m)/R_{p^+}(m) - m + 2p. \tag{6}$$

For each m we calculate $C_p(m)$ for all models of interest. In our extended example the models of interest are all those including an intercept term.

However, some care is needed in interpreting this definition. For each of the models with p parameters, the search may be different, so that the subset $S(m)$ will depend on which model is being fitted. This same subset is used to calculate $R_{p^+}(m)$, so that the estimate s^2 in (1) will also depend on the particular model being evaluated as well as on m. Since the searches are for different models, outliers will not necessarily enter in the same order for all models.

3.2. M Estimation

Ronchetti and Staudte (1994) use M-estimation to replace least squares in both the submodel of interest and the full model. They argue that the use of robust estimates requires not only the replacement of the residual sums of squares in (2) by their robust versions, but also adjustment to the coefficient $-(n - 2p)$.

An M estimator $\tilde{\beta}$ is the solution of the equation

$$\sum_{i=1}^{n} \eta(x_i, y_i - x_i^T\beta)x_i = 0, \tag{7}$$

for some function $\eta(x, e)$. The weight attached to each observation is $w(x, e) = \eta(x, e)/e$, with outlying observations appreciably downweighted. The parameter estimate $\tilde{\beta}$ yields estimated weights $\hat{w}$ and the analogue of the residual sum of squares $R_p(n)$ in (2) is

$$W_p = \sum_{i=1}^{n} \hat{w}_i^2 r_i^2 = \sum_{i=1}^{n} \hat{w}_i^2 (y_i - \hat{y}_i)^2. \tag{8}$$

The robust version of C_p, by analogy with (2) is written

$$RC_p = W_p/\hat{\sigma}^2 - (U_p - V_p) = (n - p^+)W_p/W_{p^+} - (U_p - V_p). \tag{9}$$

In (9) U_p and V_p are constants that depend on the particular M estimator that is used.

Although the development followed by Ronchetti and Staudte (1994) is more general, in our comparisons in Section 5 we use the Huber function in which

$$\eta(x, \epsilon) = \psi_c(\epsilon) = \max\{-c, \min(\epsilon, c)\}. \tag{10}$$

Then $V_p \simeq p$. However the expression for U_p is more complicated, depending on the expectation of products of ψ and its first derivative. When all weights are identically one, RC_p reduces to Mallows' C_p. Computational reasons for our choice of the Huber function are mentioned in Section 6.

4. The distribution of C_p in the forward search

The distributional results of Section 2 apply when C_p is calculated from the full sample. But, in the forward search with $m < n$ we order the observations during the search and take the central m residuals $e_i(m)$ to calculate the sums of squares $R_{p^+}(m)$ and $R_p(m)$. These sums of squares are accordingly based on truncated samples and will have smaller expectations than those based on a full sample of m observations. Specifically $\mathrm{E}\{s^2(m)\} < \sigma^2$. However, we now show that the full sample distribution holds to a good approximation with n replaced by m. That is (3) becomes

$$C_p(m) \sim (p^+ - p)F + 2p - p^+, \quad \text{where } F \sim F_{p^+-p,\, m-p^+}, \tag{11}$$

so that the only change is in the degrees of freedom of the F distribution. We explore this indication and provide results supporting (11) as the distribution of $C_p(m)$.

The argument is in two parts. First we ignore the departure from normality caused by truncation and consider the effect of the ordering of the observations on the estimation of σ^2. Then we combine simulations with results on the robustness of F tests to departures from normality to show how well the approximation (11) holds.

In the absence of outliers the estimates of the parameters β remain sensibly constant during the search whereas the value of s^2 increases as the search progresses and more remote observations enter the subset $S(m)$. The result is a forward plot of t statistics that decrease precipitously from large values at the beginning of the search. Figure 3.12 of Atkinson and Riani (2000) shows an example.

To obtain a statistic with a distribution closer to the standard t distribution, Atkinson and Riani (2002) introduced a forward added variable t test that is a special case of $C_p(m)$. In the general case of $C_p(m)$ let

$$X^+ = \left(X \vdots X^-\right). \tag{12}$$

In the t test X^- is one of the columns of X^+. The forward search is performed using the $p^+ - 1$ variables in X. Calculation of the t test for the vector X^- can use the added variable for X^- (Atkinson, 1985, Section 5.2). Because of the orthogonality of the various residuals involved in the calculation, the resulting t statistic has a null t distribution if the truncation of the distribution of the residuals is ignored.

To extend this argument to the distribution of $C_p(m)$ we make the usual analysis of variance breakdown of sums of squares in (6) and write

$$R_p(m) = R_{p^+|p}(m) + R_{p^+}(m), \tag{13}$$

so that $R_{p^+|p}(m)$ is the reduction in the residual sum of squares in going from the model with p parameters to that with p^+. Then the stochastic part of the distribution of $C_p(m)$ in (6) can be written

$$R_p(m)/R_{p^+}(m) = R_{p^+|p}(m)/R_{p^+}(m) + 1, \tag{14}$$

proportional to the F test for the addition of the extra $p^+ - p$ variables not in the reduced model. This is the multivariate generalization of the added variable t test of Atkinson and Riani (2002) and the orthogonality of the distribution to the search can again be shown by the orthogonality of the residuals of the added variables. The difference is that now the adjustments using the variables are not restricted to the addition of single terms to the model.

In the calculation of added variable t tests a single search is used for each model and the added variable statistic constructed. Likewise, in the calculation of $C_p(m)$, there is a single search for each model which is used for the calculation of the two sums of squares $R_{p^+}(m)$ and $R_p(m)$. As a consequence, the distributional results for $C_p(m)$ are a direct multivariate generalization of those for the t test.

We now need to consider the effect of the truncation of the distribution of the errors on the sums of squares $R_{p^+}(m)$ and $R_p(m)$. For central sets of m residuals we will have

$$\mathrm{E}\{s^2(m)\} = \kappa(m)\sigma^2. \tag{15}$$

The factor $\kappa(m)$ depends on the particular central m observations chosen by the forward search. Riani et al. (2009) use the results of Tallis (1963) on truncated distributions to calculate related factors for elliptically truncated multivariate distributions. However the value of the factor is unimportant in the present context, as $\kappa(m)$ occurs in both the numerator and the denominator of (14) and so cancels.

There remains the effect of the shape of the truncated distribution. In the literature on the properties of tests in least squares regression when the errors are non-normal, Box and Watson (1962) show that the sensitivity of F tests like (14) to non-normality of the errors depends not only on the distribution of y (which they take to be symmetrical, so that kurtosis is the dominant property) but also on the distribution of X^+. The results rely on a central-limit theorem for the sufficient statistics entering the calculation of $R_{p^+}(m)$ and $R_p(m)$; the more skewed the distribution of X^+ values, the larger does the sample have to be for the F distribution to be an excellent approximation to that of the calculated statistic.

For standard experimental designs, such as factorials, they conclude that "an unimportant degree of sensitivity is to be expected" to departures from normality, and, that when X^+ is drawn from the normal distribution, the sensitivity is further reduced. The numerical results of Prentice (1974) confirm these results. His examples are for appreciably smaller sample sizes than we monitor at the beginning of the search and for much more extreme conditions in X^+ than we expect to encounter in regression analysis.

To confirm these results on the lack of effect of symmetrical truncation on the distribution of $C_p(m)$ we ran an exhaustive simulation study. In the results reported here we calculated C_p for linear regression against quadratic regression on a single variable x. There were ten observations in each set of data, so that the F distribution in (11) was on 1 and 7 degrees of freedom. We generated standard normal samples of ten observations and, separately, samples of ten from the truncated normal distribution

$$Y = \Phi^{-1}(U_{\mathrm{trunc}}) \quad \text{with } U_{\mathrm{trunc}} \sim U(0.4, 0.6).$$

The truncation corresponds to $m/n = 0.2$, implying that we are interested in properties over the last 80% of the search. The values of x were the same for all samples.

Two features are of interest. One is whether the values of $C_p(m)$ for the normally distributed errors have a noticeably different distribution from those from the truncated normal. The second is whether the distribution in (11) holds for one or both sets of errors so that

$$C_p(10) - 1 \sim F_{1,7}. \tag{16}$$

There were two further sets of sample sizes in our simulation. We repeated the calculation of the value of $C_p(10)$ a thousand times and then calculated the empirical 1%, 10%, 25%, 50%, 75%, 90% and 99% of this distribution of the statistic. In order to obtain an idea of the reliability of these estimates of quantiles we repeated this calculation ten thousand times.

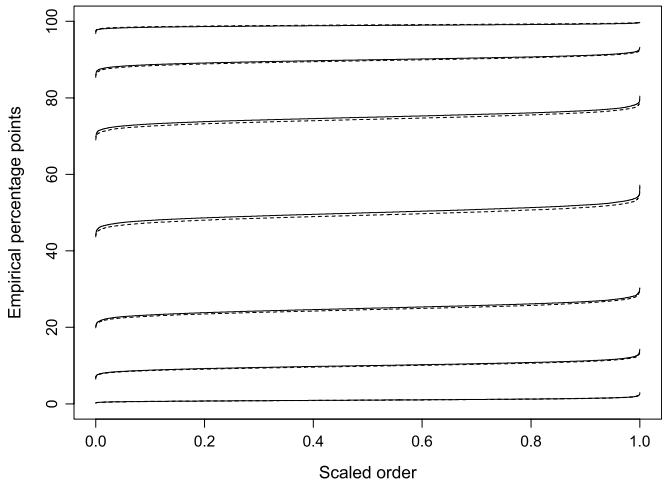

Fig. 1. Small sample distribution of forward $C_p(10)$ from 10,000 simulations. Empirical distribution of 1%, 10%, 25%, 50%, 75%, 90% and 99% points compared with the theoretical distribution (16). Continuous line, normal observations, dotted line, truncated observations. Agreement with theory is excellent.

The results of this appreciable sized simulation study are summarized in Fig. 1. We present the empirical distribution functions of the seven percentage points of $C_p(10)$ converted using (16) into percentage points of the theoretical distribution of $C_p(10)$. This apparently uninteresting plot is in fact full of valuable information about the distribution of our statistic.

The results for the normal distribution are shown as continuous curves and those for the 80% truncated distribution as dashed lines. There is remarkably little difference between the two distributions of the percentage points. This is especially so in the tails of the distribution, which is the region of inferential importance. The second point is that the medians of the empirical distributions are close to the nominal values; not only does truncation have no effect, but the small sample effects mentioned by Hurvich and Tsai (1989) and Gilmour (1996) do not affect even this severe a truncation. In selecting models we will usually be interested in the last part of the search, where truncation is much less. Further simulations, with asymmetrical truncations and various sample sizes, confirmed these findings.

5. The ozone data

5.1. Background and aggregate model selection

As a comparatively small example with sufficiently many potential explanatory variables to be interesting, we look at part of the data on ozone concentration used by Breiman and Friedman (1985) when introducing the *ACE* algorithm. These are a series of 330 daily measurements, from the beginning of the year, of ozone concentration and eight meteorological variables in California.

Atkinson and Riani (2000, Section 3.4) analyse the first 80 observations, so starting in January. They find that the data should be transformed by taking logs and that a linear time trend should be considered as one of the explanatory variables. Together with the constant term, we therefore have $p^+ = 10$.

Fig. 2 is a plot of aggregate C_p for these 80 observations from the ozone data in which the smaller values of C_p for subset models are plotted against p, together with the 1% and 99% points of the scaled and shifted F distribution (3). The plot has a typical shape. Initially, for small p, all models have values of C_p lying well outside the envelope, and so these small models are not satisfactory. The best relatively small models are for $p = 6$ and 7. All models we will discuss include a constant and the time trend, so we will often not explicitly mention them. The model with smallest C_p for $p = 6$ also includes variables 2, 5, 6 and 8. This is the model selected by Atkinson and Riani (2000, p. 70). In the second-best model for $p = 6$, variable 4 replaces variable 6, giving the model including variables 2, 4, 5 and 8. The best model for $p = 7$ includes both these variables. Good models for larger values of p add further variables to the model for $p = 7$, giving rise to larger values of C_p. These results are summarized in Table 1.

Fig. 3 gives a similar plot for the robust $RC_p(n)$ statistic of Ronchetti and Staudte (1994), but without envelopes, since we do not know the distribution of this statistic. The RC_p values were calculated with the S-Plus routine `RCp` using Huber's ψ function (10) with c equal to 1.345. A small point is that the figure from library `robust` has been redrawn to include dots for the value of robust RC_p for each model, in addition to the somewhat extensive labels. We were surprised at the improvement in interpretability afforded by this small change. In Fig. 3 the minimum RC_p is for a model with $p = 8$ and variables 1, 3, 4, 5, 6 and 8, whereas the best model for $p = 7$ is for variables 1, 4, 6, 7 and 8, with 1, 4, 5, 6 and 8 giving a slightly larger value of RC_p. These models are again listed in Table 1. In this case the best model for $p = 8$ does not include the best model for $p = 7$, although the second-best model does.

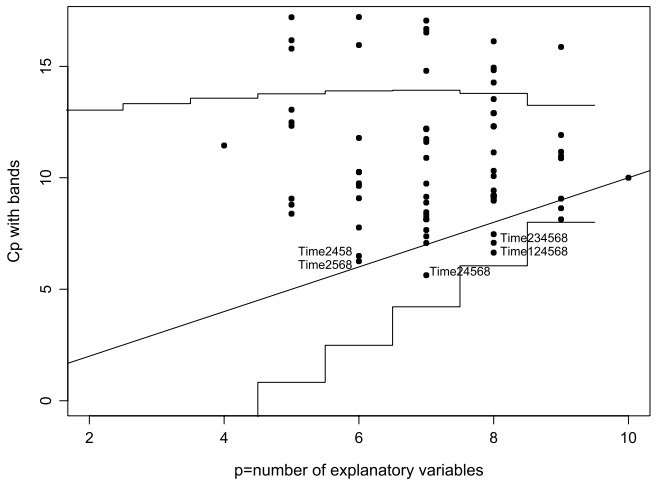

Fig. 2. Ozone data: C_p plot. The combination of the two best models for $p = 6$ yields the best model for $p = 7$. Bands are the 2.5% and 97.5% points of the scaled and shifted F distribution (3).

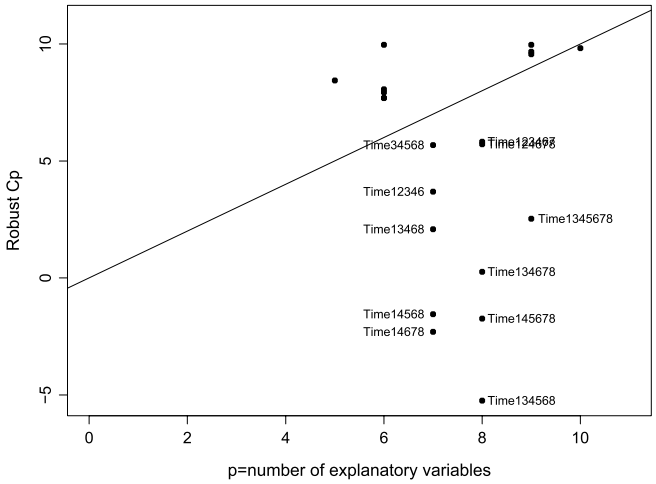

Fig. 3. Ozone data: plot of robust RC_p of Ronchetti and Staudte. Note the very sharp minimum compared with that in the non-robust C_p plot of Fig. 2.

Table 1
Ozone data: some models, including an intercept and time, selected by robust and non-robust aggregate C_p.

p	Variables	
Non-robust		
7	2, 4, 5, 6, 8	
6	2, 5, 6, 8	2, 4, 5, 8
Robust		
8	1, 3, 4, 5, 6, 8	1, 4, 5, 6, 7, 8
7	1, 4, 6, 7, 8	1, 4, 5, 6, 8

There are three general comments about Fig. 3. One is that the robust method has selected models with more variables than the non-robust method. The second is that RC_p has a much sharper increase in values away from the minimum; the model appears to be selected with less ambiguity. The third is that the small values of the robust statistic are lower than those in Fig. 2; the minimum, for $p = 8$, is -5.24. As (3) shows, the lower limit of non-robust C_p is $2p - p^+ = 6$ when $p = 8$. The occurrence of this negative value rules out the scaled and shifted F distribution (3) for the robust C_p statistic.

In contrast, the model with minimum C_p in Fig. 2 is for $p = 7$ and includes the constant, the trend and variables 2, 4, 5, 6 and 8. However, this model may be too large, since the t values for x_4 and x_6 are respectively -1.64 and 1.71. Our purpose is to use the forward search to determine how the choice of model is influenced by outliers or other unsuspected structure.

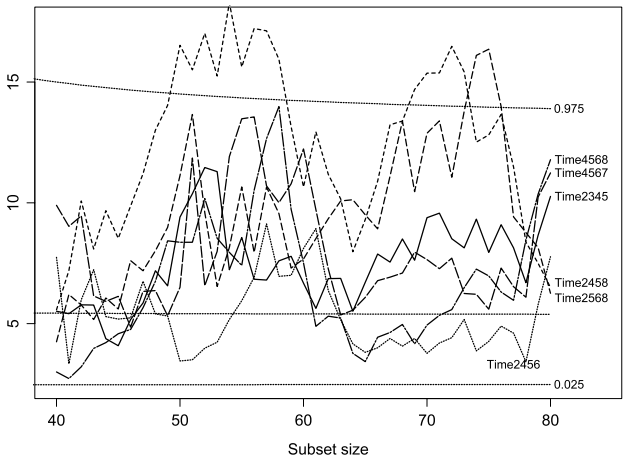

Fig. 4. Forward plots of $C_p(m)$ for the ozone data when $p = 6$ for m from 40 to 80. There is a dramatic change in the ordering of the models in the last two steps of the search. Bands are the 2.5% and 97.5% points of the scaled and shifted F distribution (11).

5.2. Forward C_p plots

When we apply the forward search to model selection we obtain a forward plot of $C_p(m)$ for each model. Thus the points in Figs. 2 and 3 are replaced by the curves of forward plots for all values of m that are of interest. The resulting quantity of graphical output can be overwhelming. We first illustrate the structure of forward plots of $C_p(m)$ and then, in Section 5.3, introduce a new plot that cogently summarizes this information.

Atkinson and Riani (2008) in their analysis of the ozone data give separate plots of forward C_p for values of p from 4 to 7. In Fig. 4 we show only $C_6(m)$ from $m = 40$, including only those models that have small values of $C_6(m)$ at some point in this region of the search. Many of the curves increase in the last two steps, suggesting the presence of two influential observations. The plot shows that, when $m = 78$, minimizing the value of $C_p(78)$ leads to the choice of model with terms 2, 4, 5 and 6, although this is only the third best model of this size when $m = n$. This plot clearly and elegantly shows how the choice of model is being influenced by the last two observations to enter the forward search. It also shows that models such as 2, 4, 5 and 8 and 2, 5, 6 and 8 behave very differently from 2, 4, 5 and 6. Although these two models have small values of $C_6(80)$ that are indeed less than the values at $m = 78$, the values of $C_6(m)$ go above the 97.5% bound of the scaled and shifted F distribution in the region of m shown in the plot. The unstable behaviour with m of some of these curves is indicative of masking, that is of the hiding of the inferential behaviour of one outlier by the presence of another. Since different models may lead to different searches through the data, the value of m at which the outliers, if any, enter the subset will depend on the model. The plots of $C_p(m)$ in Fig. 4 do indeed show jumps at a variety of values of m. However, several models show a sharp increase in value in the last two steps of the search.

5.3. The generalized candlestick plot

We now generalize C_p plots such as Fig. 2 to contain a summary, for each model, of the information in the trajectory of the forward plots of $C_p(m)$ in Fig. 4. Our starting point is the "candlestick" plot used to summarize such quantities as the high, low and closing values of stocks. Google provides many references. However, we need a generalization of this plot. Since we expect any outliers to enter the search towards the end, the last few values of $C_p(m)$ are of particular interest, as is the comparison of these values with earlier average behaviour.

Fig. 5 shows our generalized candlestick plot for 71 models in the range $p = 3$ to 8. All these models have trajectories that, over the last 20% of the search (that is in the last 16 steps) had one of the five smallest values of $C_p(m)$ for their particular value of p. The vertical lines in the plot summarize the values of $C_p(m)$ for each model in the central part of the search, that is for $m = n - 16$ through $n - 4$. Individual symbols are reserved for the last four values. The definition of the candlesticks is:

Lowest Value; minimum in the central part of the search;
Highest Value; maximum in the central part of the search;
Central Box; mean and median of the values in the central part of the search; filled if mean $<$ median;
Stars; the values in steps $n - 3$, $n - 2$ and $n - 1$ if these lie outside the box;
Unfilled Circle; the final value.

Thus each point in the standard non-robust C_p plot such as Fig. 2 is replaced by a single vertical line and up to four extra symbols.

The general shape of the plot in Fig. 5 is similar to that of the non-robust C_p plot in Fig. 2. For small values of p all models have large values of $C_p(m)$ over the last 16 values of m. Conversely, for large p there are many models with small values

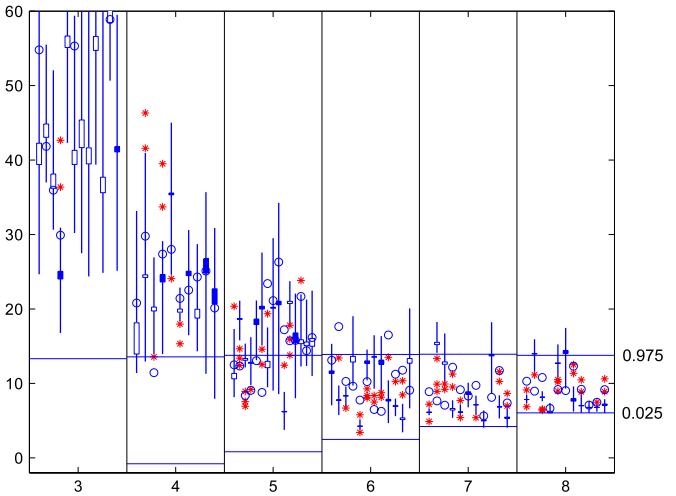

Fig. 5. Generalized candlestick plot for $C_p(m)$, $m = 64 - 80$, for the 71 best models in this range of m; 2.5% and 97.5% points of the distribution of C_p from (3).

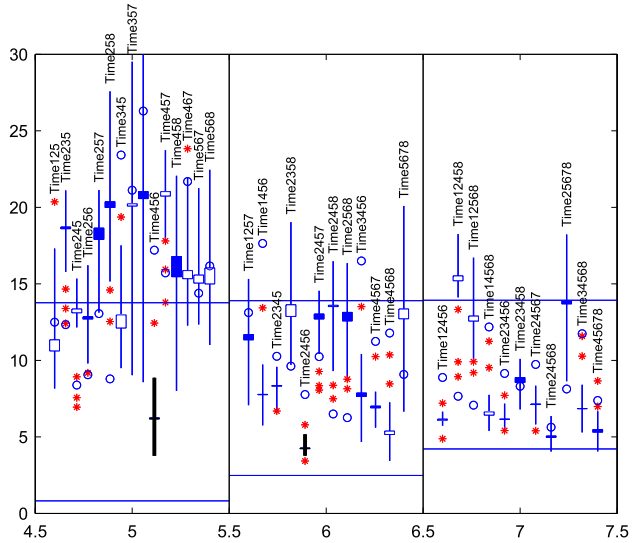

Fig. 6. Ozone data: generalized candlestick plot for $C_p(m)$, $m = 64 - 80$, for the 38 best models in this range of m with $p = 5$, 6 and 7; 2.5% and 97.5% points of the distribution of C_p (detail of Fig. 5). Note in particular the highlighted models 4, 5 and 6 and also 2, 4, 5 and 6 (all including time), plotted with a thicker vertical line.

of $C_p(m)$ over most of the range. What we have not seen before is the decrease in variability in the values of $C_p(m)$ as p increases. When p is too small, the values of $C_p(m)$ respond with extreme sensitivity to the addition of extra observations.

The indication of Fig. 5 is that values of p of 5, 6 or 7 should yield a satisfactory model. Accordingly, we reproduce in Fig. 6 the generalized candlestick plot of Fig. 5 for these three values, adding labels to identify the models.

Two models are highlighted in this figure. The first, for $p = 5$ has, in addition to time and an intercept, the variables 4, 5 and 6. This has small and stable values of $C_p(m)$, except for the last two steps of the search. The other, for $p = 6$ adds variable 2. This model, with stable values of $C_p(m)$ over the range studied was not included in Table 1 because, as the unfilled circles in the plot show, it had only the third smallest value of C_6 at the end of the search. We now consider the effect of individual observations at the end of the search on the values of $C_p(m)$.

5.4. Individual observations and $C_p(m)$

The ordering of observations by the forward search enables us to pinpoint the influential effect of individual observations. Table 2 lists the last observation to enter the subset for the 71 models plotted in Fig. 5. For the range of plausible values of p, observation 65 is the last to enter in all but one search and may well be an outlier. We can however obtain more incisive information from the generalized candlestick plot.

Table 2
Ozone data: frequency of the last observation to enter the forward search for the 71 models of Fig. 5 as a function of p.

p	Observation number					
	13	31	53	56	65	80
3	1	1	3	2	3	2
4	1	1	0	3	5	0
5	0	0	0	1	14	0
6	0	0	0	0	12	0
7	0	0	0	0	11	0
8	0	0	0	0	11	0

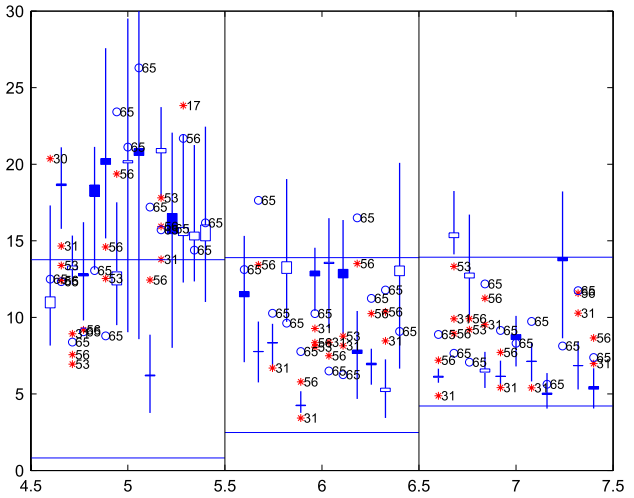

Fig. 7. Ozone data: generalized candlestick plot of $C_p(m)$ for the 38 models in Fig. 6. The observations entering in the final part of the individual searches are marked. All models include a term in time.

Fig. 7 is the generalized candlestick plot for the 38 models of Fig. 6 with, where appropriate, the last four values of $C_p(m)$ labelled with the observation entering the subset. observation 65 is last to enter in all but one search and this observation occurs with the circles indicating the final value of $C_p(m)$. For the model 4, 5 and 6 highlighted in Fig. 6 we can see that deletion of observation 65 causes the value of C_p to fall below the 97.5% envelope; further deletion of observation 56 causes another large decrease. The second model highlighted in Fig. 6 is 2, 4, 5 and 6. Although deletion of observations 65 and 56 does cause a decrease in the values of $C_p(m)$, this effect is less than that for the smaller model. The values for the two models with the smallest values of C_p at the end of the search both increase slightly when observation 65 is deleted.

It is clear from this plot that deletion of observations 65 and 56 has a significant effect on model selection. We now consider whether these observations can be treated as outlying and so removed from the analysis of the data.

5.5. Outlier detection

To detect outliers we scale the least squares residuals $e_i(m)$ of Section 3.1 to give the deletion residuals $r_i(m)$ for the $n-m$ observations not in $S^{(m)}$, the scaling depending on the leverage of the observation and on the estimate of σ^2. Let $i_{\min}$ denote 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 $r_{i_{\min}}(m)$ as a test statistic. If the absolute value is too large, observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^{(m)}$. Riani and Atkinson (2007) give further details and discuss the calculation of approximations to the pointwise distribution of the test statistic, which we use.

The upper panel of Fig. 8 shows a forward plot of the minimum deletion residual $r_i(m)$ for all 80 observations when the model contains variables 4, 5 and 6, together with 1%, 50% and 99% envelopes. The last two observations are clearly revealed as outlying. If they are removed and the envelopes recalculated for $n = 78$ we obtain the plot in the lower panel of Fig. 8. There is no evidence of further outlying observations. We accordingly continue our analysis on the set of 78 observations.

Table 3 lists the t statistics for the terms in the two highlighted models in Fig. 6 when $n = 78$. When the model includes 2, 4, 5 and 6 neither x_2 nor x_6 are quite significant at the 1% level. However, when x_2 is removed from the model, all remaining terms have absolute t values > 3. The standard backward analysis of these 78 data points based on t values would lead to the model 2, 4 and 5 for $p = 5$. Fig. 6 shows that further omission of two more observations leads to an appreciable value of $C_p(m)$ for such a model.

Atkinson and Riani (2007) use forward plots of the added variable t statistics of Atkinson and Riani (2002) to build a model for the ozone data. Their Fig. 5 shows the evolution of the t statistics during the forward search for the six parameter

Table 3
Ozone data: t statistics of the two best models after removing outliers ($n = 78$).

Term	Model	
	Time 4 5 6	Time 2 4 5 6
Intercept	−7.81	−5.74
Time	9.15	8.99
x_2	–	−2.57
x_4	−3.98	−3.01
x_5	8.91	6.80
x_6	3.03	2.39

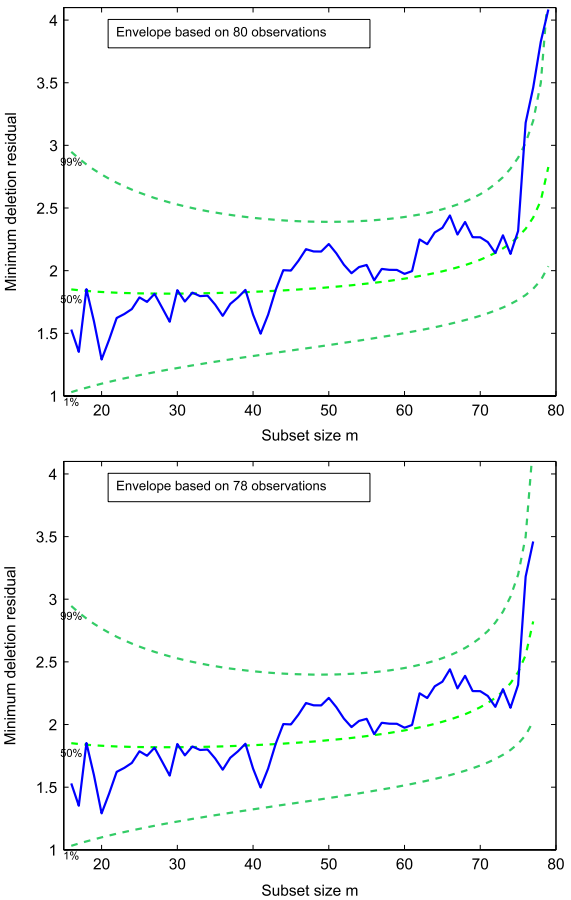

Fig. 8. Ozone data, outlier detection, model with variables time and 4, 5 and 6. Upper panel, forward plot of minimum deletion residual $r_i(m)$, with 1%, 50% and 99% envelopes for $n = 80$. Lower panel, forward plot after deletion of observations 56 and 65; envelopes for $n = 78$.

model. Around $m = 75$ both x_6 and x_4 have t values close to those in Table 3. However, the term in x_2 has a decreasing t value as observations are deleted. Their plot does serve to complement the information obtained from the generalized candlestick plot. However, unless the explanatory variables are nearly orthogonal, one plot with $p - 1$ curves will be needed for each model and a summary, such as that of Fig. 6, becomes essential for building models with several explanatory variables. Of course, if the explanatory variables are nearly orthogonal, perhaps from a designed experiment, it will not be necessary to use formal model selection techniques such as C_p.

5.6. Robust and forward model building

To conclude our analysis of the ozone data we compare the two models of Table 3 with the models indicated by the robust procedure of Ronchetti and Staudte (1994) given in Table 1, for which $p = 7$ and 8. The generalized candlestick plot for these higher values of p is given in Fig. 9, with three models highlighted. The fourth model, 1, 4, 6, 7 and 8, does not qualify for inclusion.

It is not clear from this plot why these models have been selected. Although the three shown have small and stable values of $C_p(m)$ over the central part of the search, so do several other models with these values of p. In addition,

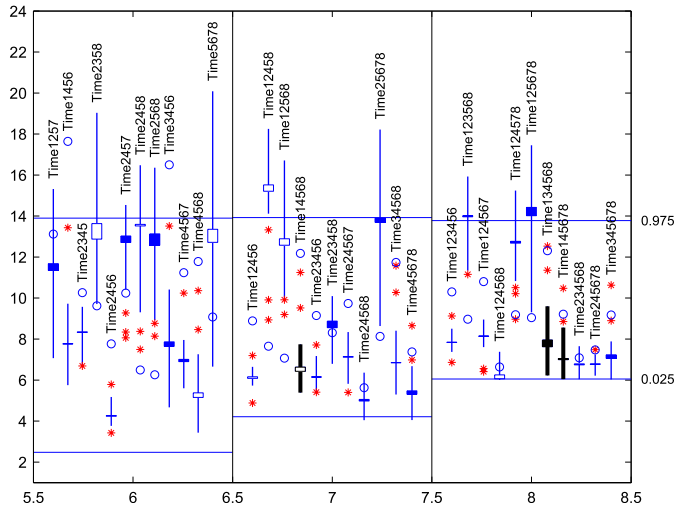

Fig. 9. Ozone data: generalized candlestick plot for $C_p(m)$, $m = 64 - 80$, for the best models with $p = 6, 7$ and 8. The three plausible models from Table 1 found by the robust procedure of Ronchetti and Staudte are highlighted.

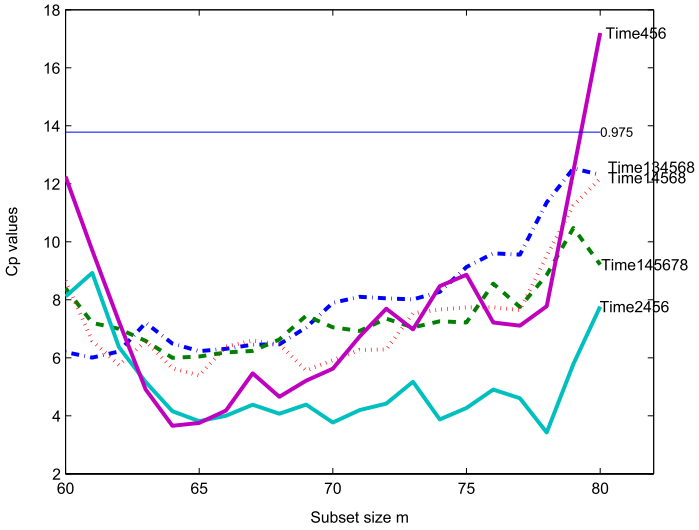

Fig. 10. Ozone data: forward plots of $C_p(m)$ for selected models from Tables 1 and 3.

we have found smaller models with small values of C_p which have stable behaviour once the two outliers have been deleted.

We expand the information on these five models from the generalized candlestick plot by showing, in Fig. 10, the forward plots of the values of $C_p(m)$ on which the candlesticks were based. The effect of deletion of the two outliers on the values for the model 4, 5 and 6 is evident; once observations 56 and 65 are deleted this model has a small value of $C_p(m)$. For most of the values of m covered by the plot, the values for the model 2, 4, 5 and 6 are the smallest, although showing, to a lesser extent, the effect of deletion. The three models indicated by use of the robust C_p have similar, more stable and rather larger values of $C_p(m)$. These values are less affected by the outliers than are those of the two smaller models. Fig. 5 indicates that this form of robustness may be a consequence of fitting a model with many terms.

6. Larger data sets

For a larger and more complex example we analysed data on factors influencing the use of credit and other cards. There were 1000 observations on the customers of a bank, a single response and nine explanatory variables. The results were similar to those obtained here. The larger number of observations did not cause any problems with computational time.

Problems with size will however occur as the number of explanatory variables p^+ increases. One solution is to group the variables, preferably into groups reflecting common factors, and to build the model including and excluding groups of variables. Those groups that contain no variables of interest (that is, those for which $C_p(m)$ in the central part of the search

is always above the 99% threshold) can be dropped and the grouping repeated with smaller groups until a few individual factors emerge as candidates for inclusion in a model.

The Matlab software used in this paper is part of the FSDA (Forward Search Data Analysis) toolbox which can be downloaded, free of charge, from the webpage www.riani.it in the section "Matlab code". Full documentation is included. The importance of software is highlighted by our choice of robust method in Section 5.6 which was constrained by the available software.

7. Discussion

An appreciable number of robust selection criteria for the comparison of sets of models have already been presented in the statistical literature. We have indeed already compared our results with some of those of Ronchetti and Staudte (1994). Ronchetti (1997) and Ronchetti et al. (1997) augment this work with robust versions of AIC, as well as C_p and cross-validation. Sommer and Huggins (1996) proposed model selection based on Wald tests. Qian and Künsch (1998) introduced a new criterion for robust model selection in linear models based on generalized Huberization and on the theory of stochastic complexity. Agostinelli (2002) proposed a robust modification of AIC and C_p using a weighted likelihood methodology. Morgenthaler et al. (2003) constructed a selection technique for simultaneous identification of both the correct model structure and of unusual observations. Müller and Welsh (2005) introduced a selection criterion that, for each model, combines a measure of goodness of fit, a penalty term for the number of parameters and the expected prediction error, conditional on the observed sample. Maronna et al. (2006) proposed a robust Final Prediction Error criterion. Salibian-Barrera and Van Aelst (2008) based robust model selection criteria on *MM*-estimators for the prediction error of each model by using the fast robust bootstrap. Lutz et al. (2008) suggested five robustifications of L_2 Boosting for linear regression and considered various robustness properties. McCann and Welsch (2007) and Khan et al. (2007a) robustified the stepwise algorithm for Least Angle Regression, while Khan et al. (2007b) investigated robust and computationally efficient versions of classical forward selection and stepwise methods for model selection.

It is our opinion that none of the above-listed robust methods, at least as presented, enables us to appreciate the inferential effect that each unit, outlier or not, leverage or not, exerts on the model selection procedure. In this paper we have shown that our forward analysis using the values of $C_p(m)$, and their summary in the generalized candlestick plot, provide a powerful method of ascertaining the effect of individual observations and outliers on the choice of a model using C_p. We are thus led to a robust and informed choice of model. The emphasis in the robust methods tends, on the contrary, to be on the properties of the estimates obtained. We suspect that much of the information we obtain on the effects of individual observations can be recovered by careful analysis of the weights that come from the various robust procedures. We look forward to seeing such analyses.

Forward methods, similar to those of our paper, can be applied to model choice with AIC (Akaike, 1974), of which C_p is a special case for regression. A general disadvantage of AIC is that the distribution of the statistic is only known asymptotically. As we have seen, knowledge of the distribution of $C_p(m)$ is helpful in calibrating our plots.

A major area of employment of AIC is in the selection of time series models (Tong, 2001, Section 9). For autoregressive models an alterative is the straightforward use of C_p, which is however not possible for moving average models. Riani and Atkinson (submitted for publication) accordingly develop a C_p statistic with known distributional properties for the selection of time series ARMA models with regressors. The extension of the diagnostic and robust methods of the current paper to this time series C_p statistic is straightforward, but does pose appreciable numerical problems as, during the search, the Kalman filter has to be run many times with a large number of missing observations.

The starting point for our paper was the desire to go beyond methods of model choice based on aggregate statistics. We note that Bayesian methods, such as Green (1995), which work on the likelihood and a prior distribution, also use aggregate statistics. The problem of finding a forward procedure when each sample size has to be evaluated using MCMC methodology seems to us to be extremely challenging. Our approach can be easily extended to more general models such as those considered by Cantoni and Ronchetti (2001).

Finally we stress that, in our opinion, the mechanical use of C_p, or of its forward version, is to be avoided. Any model selected by use of these methods should be subject to customary statistical checks, not only tests of outliers and of the significance of the terms in the model, but to a full diagnostic forward analysis (Atkinson and Riani, 2000).

References

Agostinelli, C., 2002. Robust model selection in regression via weighted likelihood methodology. Statistics and Probability Letters 56, 289–300.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.
Atkinson, A.C., 1985. Plots, Transformations, and Regression. Oxford University Press, Oxford.
Atkinson, A.C., Riani, M., 2000. Robust Diagnostic Regression Analysis. Springer-Verlag, New York.
Atkinson, A.C., Riani, M., 2002. Forward search added variable t tests and the effect of masked outliers on model selection. Biometrika 89, 939–946.
Atkinson, A.C., Riani, M., 2007. Building regression models with the forward search. Journal of Computing and Information Technology-CIT 15, 287–294. doi:10.2489/cit.1001135.
Atkinson, A.C., Riani, M., 2008. A robust and diagnostic information criterion for selecting regression models. Journal of the Japanese Statistical Society 38, 3–14.
Box, G.E.P., Watson, G.S., 1962. Robustness to non-normality of regression tests. Biometrika 49, 93–106.
Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and transformation (with discussion). Journal of the American Statistical Association 80, 580–619.

Cantoni, E., Ronchetti, E., 2001. Robust inference for generalized linear models. Journal of the American Statistical Association 96, 1022–1030.
Cook, S., Weisberg,, 1982. Residuals and Influence in Regression. Chapman and Hall, London.
Gilmour, S.G., 1996. The interpretation of Mallows's C_p-statistic. The Statistician 45, 49–56.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.
Khan, J., Van Aelst, S., Zamar, R., 2007a. Building a robust linear model with forward selection and stepwise procedures. Computational Statistics and Data Analysis 52, 239–248.
Khan, J., Van Aelst, S., Zamar, R., 2007b. Robust linear model selection based on least angle regression. Journal of the American Statistical Association 102, 1289–1299.
Lutz, R., Kalisch, M., Bühlmann, P., 2008. Robustified L2 boosting. Computational Statistics and Data Analysis 52, 3331–3341.
Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.
Maronna, R.A., Martin, D.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, New York.
McCann, L., Welsch, R.E., 2007. Robust variable selection using least angle regression and elemental set sampling. Computational Statistics and Data Analysis 52, 249–257.
Morgenthaler, S., Welsch, R., Zenide, A., 2003. Algorithms for robust model selection in linear regression. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser-Verlag, Basel, pp. 101–109.
Müller, S., Welsh, A.H., 2005. Outlier robust model selection in linear regression. Journal of the American Statistical Association 100, 1297–1310.
Prentice, R.L., 1974. Degrees-of-freedom modifications for F tests based on nonnormal errors. Biometrika 61, 559–563.
Qian, G., Künsch, H.R., 1998. On model selection via stochastic complexity in robust linear regression. Journal of Statistical Planning and Inference 75, 91–116.
Riani, M., Atkinson, A.C., 2007. Fast calibrations of the forward search for testing multiple outliers in regression. Advances in Data Analysis and Classification 1, 123–141. doi:10.1007/s11634-007-0007-y.
Riani, M., Atkinson, A.C., 2010. The selection of time series models perhaps with regressors (submitted for publication).
Riani, M., Atkinson, A.C., Cerioli, A., 2009. Finding an unknown number of multivariate outliers. Journal of the Royal Statistical Society, Series B 71, 447–466.
Ronchetti, E., 1997. Robustness aspects of model choice. Statistica Sinica 7, 327–338.
Ronchetti, E., Field, C., Blanchard, W., 1997. Robust linear model selection by cross-validation. Journal of the American Statistical Association 92, 1017–1023.
Ronchetti, E., Staudte, R.G., 1994. A robust version of Mallows's C_p. Journal of the American Statistical Association 89, 550–559.
Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.
Salibian-Barrera, M., Van Aelst, S., 2008. Robust model selection using fast and robust bootstrap. Computational Statistics and Data Analysis 52, 5121–5135.
Sommer, S., Huggins, R.M., 1996. Variable selection using the Wald test and robust C_p. Journal of the Royal Statistical Society Series C (Applied Statistics) 45, 15–29.
Tallis, G.M., 1963. Elliptical and radial truncation in normal samples. Annals of Mathematical Statistics 34, 940–944.
Tong, H., 2001. A personal journey through time series in Biometrika. Biometrika 88, 195–218.