J. Japan Statist. Soc.
Vol. 38 No. 1 2008 3–14

A ROBUST AND DIAGNOSTIC INFORMATION CRITERION FOR SELECTING REGRESSION MODELS

Anthony C. Atkinson* and Marco Riani**

We combine the selection of a statistical model with the robust parameter estimation and diagnostic properties of the Forward Search. As a result we obtain procedures that select the best model in the presence of outliers. We derive distributional properties of our method and illustrate it on data on ozone concentration. The effect of outliers on the choice of a model is revealed. Although our example is for regression, the connection with AIC is stressed.

Key words and phrases: AIC, C_p, Forward Search, outliers.

1. Introduction

Professor Akaike's 1974 paper on model selection (Akaike (1974)) is one of the most highly cited papers in statistics and control engineering. The basic idea, certainly familiar to virtually all readers of our paper, is that in choosing between non-nested statistical models there needs to be a trade-off between the improved fit of a larger model and the increased number of parameters. Akaike's elegant solution penalizes the maximized log-likelihood by twice the number of parameters in the model. However, the loglikelihood is an aggregate statistic, a function of all the observations. AIC provides no evidence of whether or how individual observations or unidentified structure are affecting the model choice. It is the purpose of the present paper to use the forward search, a graphics rich robust procedure, to reveal the effect of such sources on model choice.

We focus on regression models, our paper combining regression diagnostics with model selection. In the next section we introduce AIC and relate it to C_p. The use of whole-sample C_p for choice of a regression model is exemplified in Subsection 2.2 for 80 observations on ozone concentration. Our diagnostic analysis begins in Subsection 3.1 where we introduce the forward search. In Subsection 3.2 we derive versions of AIC and C_p that are calculated from subsets of the data and illustrate their properties. Section 4 considers the distribution of C_p in the forward search while Section 5 applies our procedure to the ozone data. We find that the choice of model is highly sensitive to the presence of two outliers. Some discussion of the literature is in Section 6.

Accepted August 4, 2007.

*Department of Statistics, London School of Economics, London WC2A 2AE, U.K. Email: a.c.atkinson@lse.ac.uk

**Dipartimento di Economia, Università di Parma, Italy. Email: mriani@unipr.it

2. Aggregate statistics

2.1. AIC and C_p

The loglikelihood of n observations y, a function of the $p \times 1$ vector of parameters β is $L(\beta; y)$. If $\hat{\beta}$ is the maximum likelihood estimator of β, AIC is defined as

$$AIC = -2L(\hat{\beta}; y) + 2p. \tag{2.1}$$

That model is selected for which AIC is a minimum, a choice that is unaffected by any constants depending on y and n.

We are interested in the special case of the AIC for the linear multiple regression model $y = X\beta + \epsilon$, in which X is an $n \times p$ full-rank matrix of known constants, with i-th 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 $R_p(n)$ and, for known σ^2,

$$AIC_\sigma = n \log(2\pi) + n \log \sigma^2 + R_p(n)/\sigma^2 + 2p. \tag{2.2}$$

If, as is usually the case, σ^2 is not known, the maximum likelihood estimator is

$$\hat{\sigma}^2 = R_p(n)/n. \tag{2.3}$$

With this internal estimate of σ^2 the criterion (2.2) becomes

$$AIC_I = n \log(2\pi) + n \log\{R_p(n)/n\} + n + 2p, \tag{2.4}$$

a form often used in the selection of non-nested time series models with normally distributed errors. Some references are in §9 of Tong (2001). Venables and Ripley (1997, p. 221) use a Taylor series expansion to show the close relationship between (2.2) and (2.4). In these derivations we have, unusually, included all constants as we shall be interested in comparing model selection criteria across different subsets of $m < n$ observations.

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

$$s^2 = R_K(n)/(n - K). \tag{2.5}$$

With this estimate the criterion (2.2) is

$$\begin{aligned} AIC \; = \; & n \log(2\pi) + n \log\{R_K(n)/(n-K)\} \\ & + (n-K)R_p(n)/R_K(n) + 2p. \end{aligned} \tag{2.6}$$

In the standard application of model selection procedures both n and s^2 are fixed, the variable factors being the value of p and the regressors that are being considered. Then choice of the model minimizing (2.6) is identical to the choice of model minimizing

$$C_p = R_p(n)/s^2 - n + 2p = (n-K)R_p(n)/R_K(n) - n + 2p. \tag{2.7}$$

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 parameters provided the full model with K parameters yields an unbiased estimate 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. In Section 4 we consider the distribution of values of C_p and try to make this statement more precise. We stress that, in our opinion, the mechanical use of C_p is to be avoided. Any model selected by use of C_p should be subject to customary statistical checks, such as tests of the significance of the included terms.

2.2. The ozone data

As an example with sufficiently many potential explanatory variables to be interesting, we look at the data on ozone concentration used by Breiman and Friedman (1985) when introducing the *ACE* algorithm. These are a series of 300 daily measurements, from the beginning of the year, of ozone concentration and eight meteorological variables in California. Atkinson and Riani (2000, §3.4) analyse the first 80 observations. They find that the data should be transformed by taking logs and that a time trend should be considered as one of the explanatory variables. Together with the constant term, we therefore have $K = 10$.

Figure 1 is a C_p plot for the ozone data in which the smaller values of C_p for subset models are plotted against p. It shows a typical shape. Initially, for small p, all models have values of C_p much greater than p, and so these small models

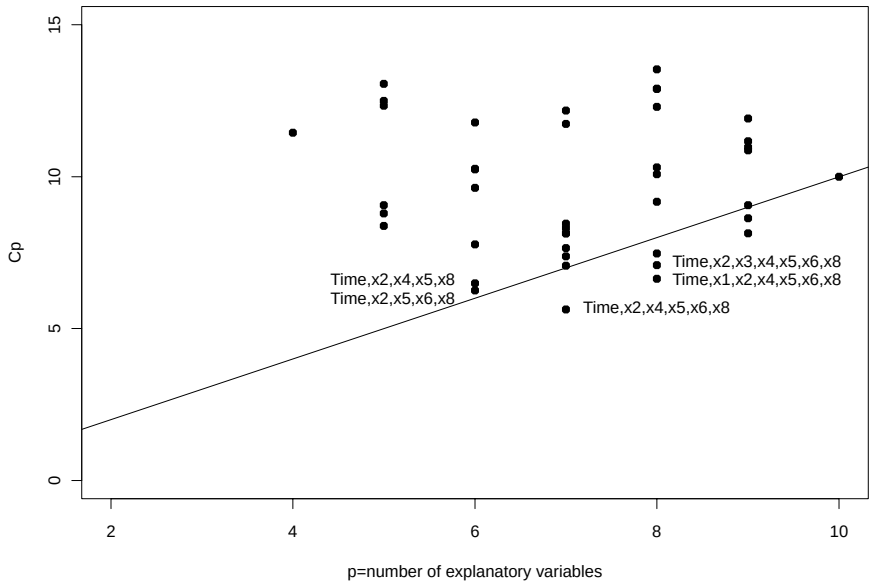

Figure 1. C_p plot for the ozone data. The combination of the two best models for $p = 6$ yields the best model for $p = 7$.

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. 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.

The model with minimum C_p in Fig. 1 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 determine how the choice of model is influenced by outliers or other unsuspected structure.

3. A robust approach

3.1. The forward search

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. Atkinson and Riani (2000) give several examples of the failure of deletion diagnostics and introduce instead 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.

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

$$e_{i*}(m) = y_i - x_i^T \hat{\beta}_*(m). \tag{3.1}$$

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

To start we take $m_0 = p + 1$ and so search over subsets of $p + 1$ observations to find the subset, out of 3,000, that yields the least median of squares (LMS) estimate of β (Rousseeuw (1984)).

3.2. Forward AIC and forward C_p

The information criteria (2.6) and (2.7) for all observations are functions of the residual sums of squares $S_p(n)$ and $S_K(n)$. For a subset of m observations we can then define the forward values of these criteria as, for example,

$$C_p(m) = (m-K)R_p(m)/R_K(m) - m + 2p. \tag{3.2}$$

For each m we calculate $C_p(m)$ for all models of interest. 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 $R_K(m)$, so that the estimate s^2 in (2.5) may also depend on the particular model being evaluated as well as on m.

To show the similarities and differences between $C_p(m)$ and the analogous forward version AIC(m) of AIC, we give forward plots of these two quantities for the wool data of Box and Cox (1964). After log transformation of the response, the data are adequately described by a first-order model in the three factors x_1, x_2 and x_3. To these we add a fourth variable which is white noise and so has no effect on the model.

The left-hand panel of Fig. 2 shows the forward plot of AIC(m) for the four four-parameter models, that is three variables and a constant. The values for the model with x_1, x_2 and x_3 are always the smallest and increase smoothly with m. This is clearly the preferred model. The right-hand panel of the figure repeats the forward plot now for the values of $C_p(m)$. The ordering of the models and the detailed behaviour of the two sets of curves are, of course, the same, since

$$AIC(m) - C_p(m) = m\log(2\pi) + m\log\{R_K(m)/(m-K)\} + m. \tag{3.3}$$

However, the values of $C_p(m)$ are the more easily interpreted since, as with C_p, the expected value of the statistic for a satisfactory model is around p, a result we demonstrate in Section 4. This is indeed the virtually constant value for the

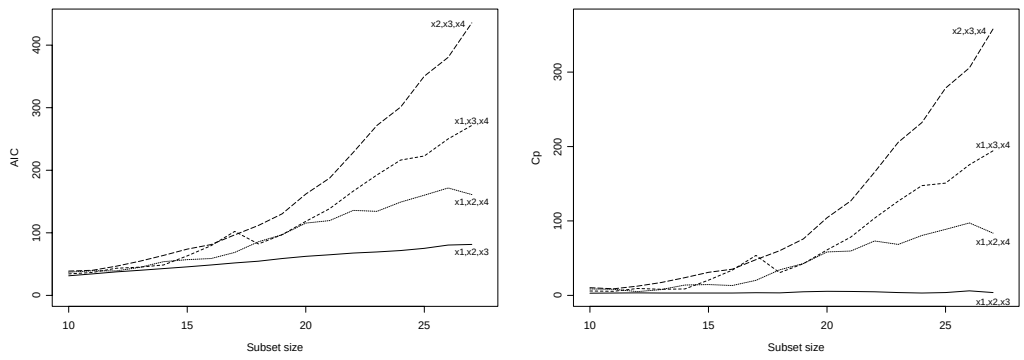

Figure 2. Wool data: three explanatory variables plus 1 noise variable. Forward plots of AIC(m) and $C_p(m)$ for $p = 4$.

model with x_1, x_2 and x_3. There are no jumps in the curve and so no indication of outliers that affect the choice of this best model. As we shall see, the curves for values of $C_p(m)$ for the ozone data in Fig. 4 are appreciably less smooth.

Although the interpretation of the right-hand panel of Fig. 2 seems straightforward, in more complicated examples with many variables, it is helpful to provide information about the distribution of the statistic. We investigate the distribution of $C_p(m)$ in the next section. The results are much simpler than those for AIC(m) so, for the rest of the paper, we focus our attention on the properties and use of $C_p(m)$.

4. The distribution of C_p in the forward search

The distribution of C_p is given, for example, by Mallows (1973) and by Gilmour (1996). From (2.7) we require the distribution of the ratio of two nested residual sums of squares. It is straightforward to show that the required distribution is

$$C_p \sim (K-p)F + 2p - K, \qquad \text{where} \quad F \sim F_{K-p,n-K}. \tag{4.1}$$

Gilmour comments that when $n-K$ is small, E(C_p) can be appreciably greater than p. In our example, with $n = 80$, this is not the case. In interpreting Gilmour's results, note that his k is the number of regressors, not the number of parameters, in the full model, so that our $K = k+1$.

These results apply to C_p which is calculated from the full sample. However, in the forward search with $m < n$ we take the central m residuals (3.1) to calculate the sums of squares $R_K(m)$ and $R_p(m)$. These sums of squares are accordingly based on truncated samples and will have smaller expectations than those based

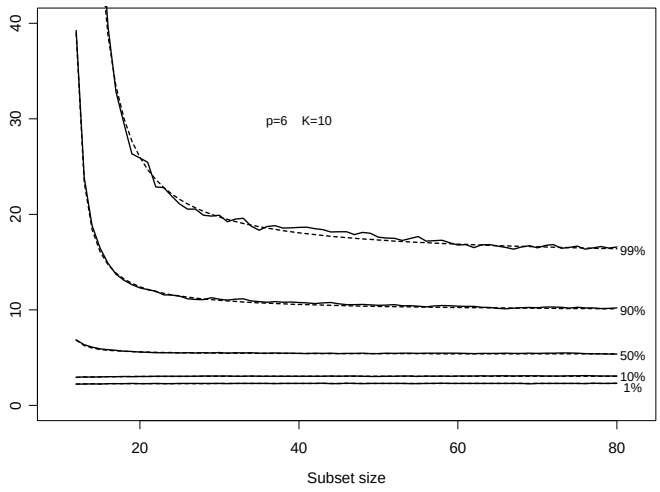

Figure 3. Comparison between empirical (solid line) and theoretical (dotted line) envelopes for $C_p(m)$ based on the F distribution (4.1) when $n = 80$, $p = 6$ and $K = 10$: 1%, 10%, 50%, 90% and 99% quantiles.

on a full sample of m observations. Specifically $\mathrm{E}\{s^2(m)\} < \sigma^2$. We conducted a small simulation study to check the effect of this truncation on the distribution of $C_p(m)$.

Figure 3 shows a forward plot of the empirical distribution from 10,000 simulations of 80 observations with $p = 6$ and $K = 10$. We give the empirical 1%, 10%, 50%, 90% and 99% points as n varies from 12 to 80, together with those calculated from the full sample distribution of C_p defined in (4.1). Amazingly, the distribution of $C_p(m)$ during the search is indistinguishable from that of the full sample statistic for sample size m. Accordingly, we can use (4.1) directly to provide envelopes for our forward plots.

5. Forward model selection for the ozone data

5.1. Forward C_p plots

We examine model selection by a forward plot for each plausible value of p. From Fig. 1 it seems that $p = 6$ is a good choice, that is a constant, the trend and four explanatory variables. We also check other values of p.

Figure 4 shows the forward plots of $C_p(m)$ from $m = 59$ for p from 4 to 7, including only those models that have small values of $C_p(m)$ in this region of the search. These plots confirm our earlier choice of $p = 6$. However, a feature for all values of p is that many of the curves increase in the last two steps. The

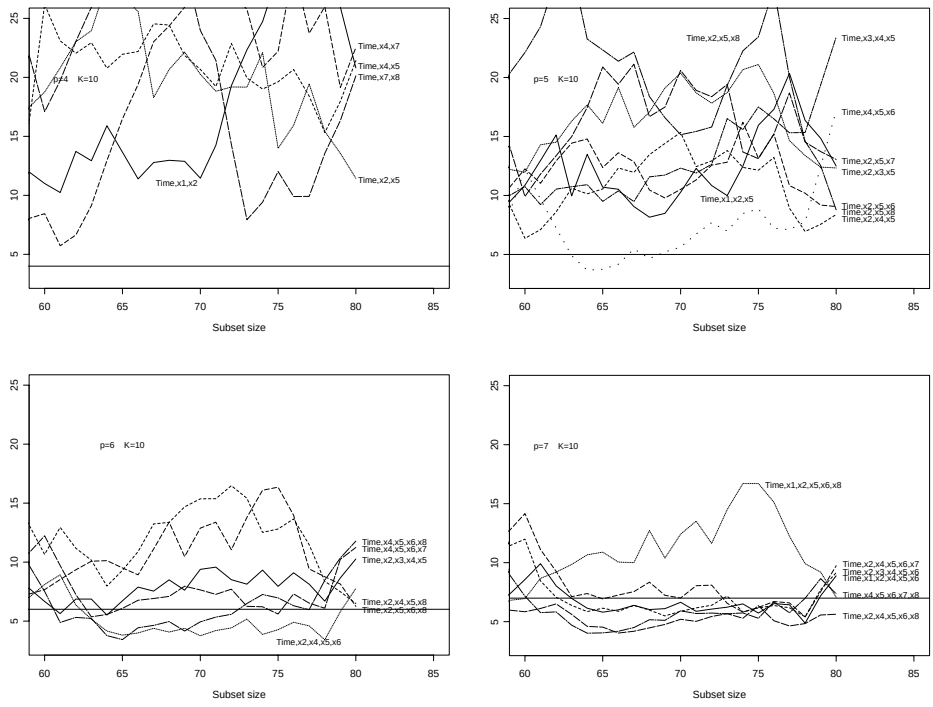

Figure 4. Ozone data: forward plots of $C_p(m)$ when $p = 4$, 5, 6 and 7. The last two observations to enter the subset have a clear effect on model choice.

plot for $p = 6$ shows that, when $m = 78$, minimising the value of C_p leads to the choice of model with terms in x_2, x_4, x_5 and x_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.

5.2. Outlier detection

The last two observations to enter $S_*(m)$ are 56 and 65; these also seem to be outlying in the plot of residuals against trend in Figure 3.36 of Atkinson and Riani (2000). To detect outliers we calculate the deletion residual for the $n - m$ observations not in $S_*^{(m)}$. These residuals are

$$r_{i*}(m) = \frac{y_i - x_i^T \hat{\beta}_*(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}} = \frac{e_{i*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}}, \tag{5.1}$$

where $h_{i*}(m) = x_i^T \{X_*(m)^T X_*(m)\}^{-1} x_i$; the leverage of each observation depends on $S_*^{(m)}$. Let $i_{\min}$ denote the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$, that is

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

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

$$r_{i_{\min}*}(m) = \frac{e_{i_{\min}*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i_{\min}*}(m)\}}}, \tag{5.2}$$

as a test statistic. If the absolute value of (5.2) is too large, the 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 distribution of the test statistic (5.2). We use simulation to find envelopes for the small value of n for the ozone data.

The left-hand panel of Fig. 5 shows a forward plot of the minimum deletion residual for all 80 observations when the model contains variables 2, 4, 5 and 6, together with 1%, 50% and 99% simulation 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 right-hand panel of Fig. 5. There is no evidence of any further outlying observations.

We now return to model selection. Figure 6 gives the last part of the forward plot of $C_p(m)$ for $n = 78$ when $p = 6$, together with 2.5%, 50% and 97.5% quantiles calculated from (4.1). We give the curves only for those models that are one of the three best at some point for the last ten values of m. The model with variables 2, 4, 5 and 6 is clearly the best; unlike any other model its value of $C_p(m)$ lies in the lower half of the distribution for $m > 63$. There are many alternative six-parameter models with values of $C_p(78)$ lying below the 97.5%

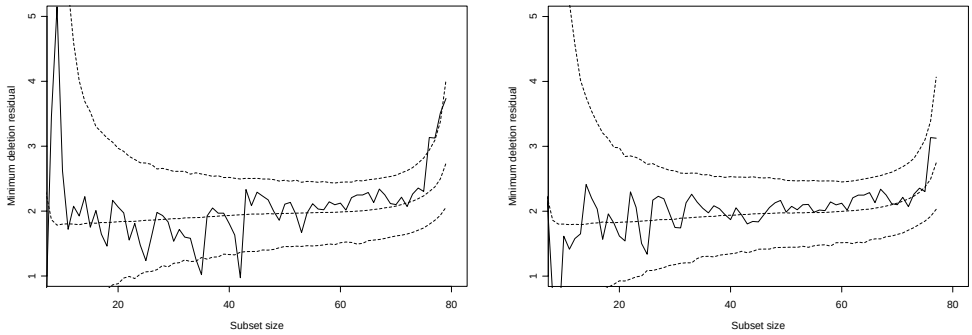

Figure 5. Ozone data: monitoring the minimum deletion residual (5.2). Left-hand panel, $n = 80$, right-hand panel, $n = 78$. There are two outlying observations.

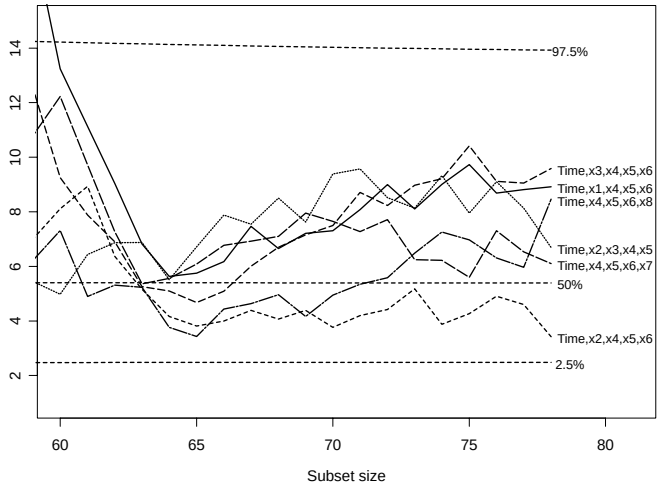

Figure 6. Ozone data without outliers: forward plots of $C_p(m)$ when $p = 6$, together with 2.5%, 50% and 97.5% quantiles from (4.1). The model including variables 2, 4, 5 and 6 is preferred.

quantile. Plots for five such are shown in Fig. 6. All however fall in the upper half of the distribution.

It is also interesting to consider the effect of deleting observations 56 and 65 on the properties of the final model. Table 1 lists the t-statistics for the six terms in the model and their significance both for all observations and for the 78 observations after deletion of the two outliers. When $n = 80$ neither x_4 nor x_6 are significant when they are both in the model. But deletion of the outliers causes the variables to be jointly significant, one at 2% and the other well past the 1% level.

We have based our argument on the plot for $p = 6$. In Fig. 7 we reproduce the C_p plot of Fig. 1 for all values of p after the two outliers have been removed.

Table 1. Ozone data: effect of deletion of outliers on significance of terms in model with variables 2, 4, 5 and 6.

	All 80 observations		$n = 78$	
Term	t	p-value	t	p-value
Constant	−4.83	0.000	−5.74	0.000
Time	7.16	0.000	8.99	0.000
x_2	−3.34	0.001	−2.57	0.012
x_4	−1.79	0.077	−3.01	0.004
x_5	5.75	0.000	6.80	0.000
x_6	1.60	0.114	2.39	0.019
R^2	0.67		0.74	

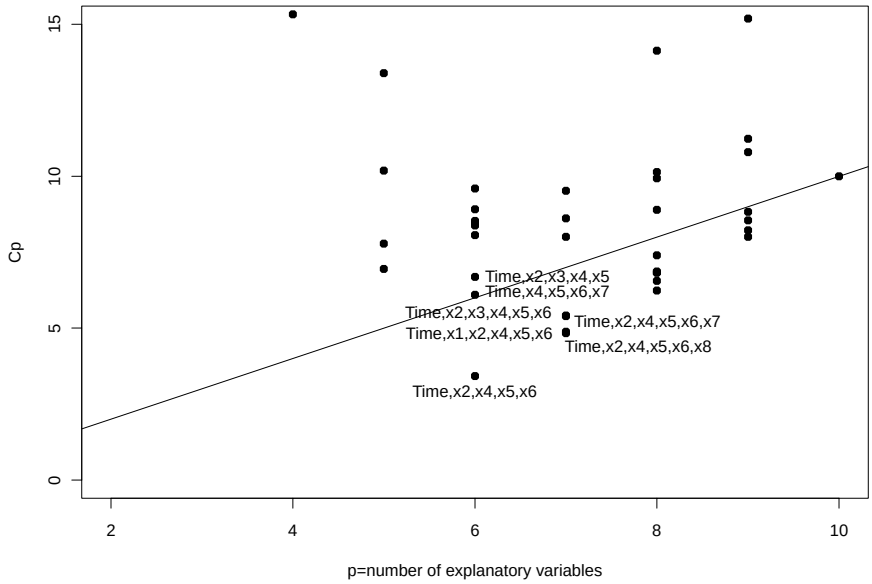

Figure 7. C_p plot for the ozone data after deletion of the two outliers. One model with $p = 6$ is now clearly best. In comparison, the best model in Fig. 1, which had $p = 7$, was less sharply revealed.

The comparison is instructive. Now the model with variables 2, 4, 5 and 6 has an appreciably smaller value of C_p than the next best six-parameter model. In addition, this value is less than that for the best seven-parameter model. By detection and deletion of the outliers we have not only changed the selected model but have sharpened the choice of the best model.

The distributional results in Fig. 7 indicate some other potential models. Whether we need to be concerned to have more than one model depends on the purpose of model fitting. If the model is to be used to predict over the region over which the data have been collected and the system is unlikely to change, so that the correlations between the explanatory variables remain sensibly constant,

then any of these models will give almost equally good predictions. If however the relationships between the variables may change, or predictions are needed in new regions where data are sparse or non-existent, then the outcomes of all satisfactory models, as selected here by $C_p(m)$, must be taken into account. The possible effects of climate change on ozone concentration in the Californian desert indicate that the consequences of several well-fitting models should be explored.

6. Literature

There is such a vast literature on AIC and C_p that choice of a few references is invidious. However, Shibata (1976) shows that use of minimum AIC leads, on average, to overestimation of model order. Atkinson (1980) explores the use of other values of the penalty multiplier than the 2 of (2.1). Hurvich and Tsai (1989) suggest a bias correction that improves the probability of selection of the model of correct order. A Bayesian model choice criterion, known as BIC, with a penalty function that increases as $\log n$, was introduced by Schwarz (1978). Kuha (2004) commends the use of both AIC and BIC in the same model selection problem and gives a list of recent references. Whether AIC, BIC or some other penalty multiplier is used, the results of Subsection 3.2 show calculation of the forward version of the criterion to be straightforward.

There is appreciably less work on the effect of individual observations on model selection. Kitagawa (1979), for unstructured samples, uses AIC to make the choice between models with outliers and those without. Weisberg (1981) breaks down the value of C_p to give the contribution of each observation and Ronchetti and Staudte (1994) use robust parameter estimates to yield an adjusted form of C_p that, in the analysis of the ozone data leads to selection of an eight-parameter model. Unlike our method that uses a series of parameter estimates to evaluate the effect of individual observations on model choice, these procedures are all based on a single estimate, robust or otherwise.

It is a great privilege to salute Professor Hirotugu Akaike on the occasion of his 80th birthday and to celebrate both the man and his scientific achievements.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model, *Biometrika*, **67**, 413–418.

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

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

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.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and transformation (with discussion), *Journal of the American Statistical Association*, **80**, 580–619.

Cook, R. D. and Weisberg, S. (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.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

Kitagawa, G. (1979). On the use of AIC for the detection of outliers, *Technometrics*, **21**, 193–199.

Kuha, J. (2004). AIC and BIC. Comparisons of assumptions and performance, *Sociological Methods and Research*, **33**, 188–229.

Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **15**, 661–675.

Riani, M. and 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.

Ronchetti, E. and 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.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117–126.

Tong, H. (2001). A personal journey through time series in Biometrika, *Biometrika*, **88**, 195–218.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus* (*2nd Edition*), Springer-Verlag, New York.

Weisberg, S. (1981). A statistic for allocating C_p to individual cases, *Technometrics*, **23**, 27–31.