Volume 23 (2), pp. 123–136
http://www.orssa.org.za

ORiON
ISSN 0529-191-X

Variable selection in multiple linear regression: The influence of individual cases

SJ Steel* DW Uys†

Received: 31 May 2006; Revised: 23 April 2007; Accepted: 25 April 2007

Abstract

The influence of individual cases in a data set is studied when variable selection is applied in multiple linear regression. Two different influence measures, based on the C_p criterion and Akaike's information criterion, are introduced. The relative change in the selection criterion when an individual case is omitted is proposed as the selection influence of the specific omitted case. Four standard examples from the literature are considered and the selection influence of the cases is calculated. It is argued that the selection procedure may be improved by taking the selection influence of individual data cases into account.

Key words: Akaike's information criterion, influential data cases, Mallows' C_p criterion, multiple linear regression, variable selection.

1 Introduction

The literature on statistical variable selection, frequently one of the first steps in a multiple linear regression analysis, is considerable. Recent references are Burnham & Anderson (2002) and Murtaugh (1998). The purpose of regression variable selection is to reduce the predictors to some "optimal" subset of the available regressors. This may be important because (i) a smaller set of predictors may provide more accurate predictions of future cases, or, (ii) it may be important to identify those predictor variables significantly influencing the response (for example, in a clinical trial). There are many techniques which are routinely used for variable selection in multiple linear regression: stepwise routines such as forward selection (Miller, 2002), Bayesian techniques (an extensive list of references is given in Burnham & Anderson, 1998, p.127), cross-validation selection techniques (Liu *et al.*, 1999), or an all possible subsets approach based, for example, on Mallows' C_p criterion (Mallows, 1973, 1995), or Breiman's little bootstrap (Breiman, 1992; Venter &

*Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa.

†Corresponding author: Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, South Africa, email: dwu@sun.ac.za.

Snyman, 1997). In this paper attention is restricted to selection using an all possible subsets approach based on either the C_p criterion or Akaike's information criterion (Akaike, 1973).

The literature on measures of the influence of individual cases in a data set on a multiple linear regression fit is equally impressive. Some contributions include Cook (1977, 1986), and Belsley *et al.* (1980). These and other contributions on influence measures assume that the predictor variables are identified beforehand. If an initial selection step takes place, the influence measures are therefore conditional, *i.e.* the specific predictors in the model are given.

Only a few papers dealing with the influence of individual data cases in regression explicitly take an initial variable selection step into account. In this context, Léger & Altman (1993) distinguish between conditional and unconditional selection versions of Cook's distance. To explain the difference between the two versions, let V be the set of indices corresponding to the predictor variables selected from the full data set, and let $\widehat{y}\,(V)$ be the prediction vector based on the selected variables and calculated from the full data set. Also, let $\widehat{y}_{(-i)}\,(V)$ be the prediction vector based on the variables corresponding to V, but calculated from the full data set without case i. Note that $\widehat{y}_{(-i)}\,(V)$ contains a prediction for case i, although this case is not used in calculating $\widehat{y}_{(-i)}\,(V)$. The conditional Cook's distance for the i-th case is

$$\left\|\widehat{y}(V) - \widehat{y}_{(-i)}\,(V)\right\|^2,$$

appropriately scaled. Here, $\|\bullet\|$ denotes the Euclidean norm. To obtain the unconditional Cook's distance, Léger & Altman (1993) argue that it is necessary to repeat the variable selection using the data without case i. This selection yields a subset $V_{(-i)}$ of indices, with $V_{(-i)}$ possibly different from V. The unconditional distance is

$$\left\|\widehat{y}(V) - \widehat{y}_{(-i)}\left(V_{(-i)}\right)\right\|^2,$$

appropriately scaled. Léger & Altman (1993) discuss the differences between these two selection versions of Cook's distance and argue that the unconditional version is preferable, since it explicitly takes the selection effect into account.

In this paper we introduce two new measures of the selection influence of an individual data case in multiple linear regression analysis. Our point of view is that such a measure should provide an indication of whether the fit of the selected model improves or deteriorates owing to the presence of the case. This is the main contribution of the paper. Section 2 is devoted to a brief exposition of Mallows' C_p statistic and Akaike's information criterion (AIC). The new measures of selection influence, based on C_p and AIC respectively, are introduced in Section 3. It is indicated how variable selection based on C_p or on AIC may be modified, and hopefully improved, by making use of the influence measures. Four practical examples are discussed in Section 4. We close with some recommendations in Section 5.

2 Variable selection using C_p or *AIC*

Let $y_1, \ldots y_n$ be observations of the response in a multiple linear regression, with corresponding observations x_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, p$ of p explanatory variables. We make the customary assumption that

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{1}$$

for $i = 1, \ldots, n$. In (1), $\varepsilon_1, \ldots, \varepsilon_n$ are independent normal $\left(0; \sigma^2\right)$ random variables, and $\beta_0, \ldots, \beta_n$ and σ^2 are unknown parameters. We assume that $\{x_{i1}, \ldots, x_{ip}\}$ may contain redundant variables, so that it is advisable to perform variable selection as a first step in the analysis. Let RSS be the residual sum of squares from the least squares fit of (1). Then $\hat{\sigma}^2 = RSS/(n - p - 1)$ is the commonly used unbiased estimator of σ^2. Consider a subset V of $\{1, \ldots, p\}$, and let $RSS(V)$ be the residual sum of squares from the least squares fit using only the regressors corresponding to the indices in V, together with an intercept. The C_p statistic for the corresponding model is

$$C_p(V) = \frac{RSS(V)}{\hat{\sigma}^2} + 2(v + 1) - n, \tag{2}$$

where v is the number of indices in V. Variable selection based on (2) entails calculating $C_p(V)$ for each subset of $\{1, \ldots, p\}$, and selecting the variables corresponding to $\hat{V}$, the subset minimizing (2). This approach is based on the fact that for a given V, $\hat{\sigma}^2 C_p(V)$ is an estimate of the expected squared error if a (multiple) linear regression function based on the variables corresponding to V is used to predict Y^*, a new (future) observation of the response random vector, Y. Therefore, choosing $\hat{V}$ to minimize (2) is equivalent to selecting the variables which minimize the estimated expected prediction error.

The *AIC* is based on the maximized log-likelihood function of the model under consideration. Given independent normal errors, and ignoring constant terms, the maximized log-likelihood for the model corresponding to a subset V is given by $-n \log\left[RSS(V)/n\right]/2$. This is a non-decreasing function of the number of selected regressors. Akaike (1973) therefore included a penalty term, *viz.* $v + 2$, which equals the number of parameters which have to be estimated. Multiplying the resulting expression by -2 yields

$$\widetilde{AIC}(V) = n \log\left(\frac{RSS(V)}{n}\right) + 2(v + 2). \tag{3}$$

For an up-to-date discussion and further motivation of (3), see Burnham and Anderson (2002). It is known that $AIC(V)$ does not perform well when the number of parameters to be estimated is large compared to the sample size (typically cases where $40\,(v + 2) > n$). In such cases a modified version of (3) should be used, *viz.*

$$AIC(V) = n \log\left(\frac{RSS(V)}{n}\right) + \frac{2n\,(v + 2)}{n - v - 3}. \tag{4}$$

Variable selection based on (3) or (4) entails calculating the criterion for each subset V of $\{1, \ldots, p\}$, and selecting the variables corresponding to the minimizing subset. This is equivalent to selecting the variables which maximize a penalized version of the maximum log-likelihood.

3 New measures of selection influence

It is standard statistical practice in multiple linear regression to study the influence of a single data case in an analysis as follows: analyse the complete data set and calculate a (summary) measure, say M; repeat the analysis after omitting the case under consideration and calculate $M_{(-i)}$; quantify the influence of case i in terms of a function, $f(\bullet)$, of M and $M_{(-i)}$. An example is provided by Cook's statistic, where M is the vector of predictions of the response, and $f(x) = \|x\|^2 / RSS$.

The measures of selection influence which we propose are also based on a leave-one-out strategy. The following questions need to be resolved: (i) What is meant by an analysis of the complete or reduced data set? (ii) What measure M should be used? (iii) How should we define $f(\bullet)$? Regarding the first question, in a variable selection context an analysis of a data set entails applying a given variable selection technique, and fitting the model corresponding to $\hat{V}$ to the data. Consequently, if we wish to study the influence of a single data case in such an analysis, it is necessary to apply the selection technique under consideration to the full data set and again to the reduced data set. This is in line with the unconditional approach recommended by Léger and Altman (1993). Turning to the second question, different choices of M can be made, depending on the aspect of the fitted model which is of interest. In variable selection the number of selected variables and the lack of fit of the corresponding model are typically of interest. These quantities are combined in selection criteria such as C_p and the AIC. It therefore seems reasonable to take M equal to the criterion employed in the selection method. This implies that $f(\bullet)$ has to be based on the difference in the value of the selection criterion before and after omitting case i. This difference, $M - M_{(-i)}$, may then be divided by M in order to calculate the relative change in the selection criterion. The proposed selection influence measure for the i-th case is therefore given by

$$f\left(M, M_{(-i)}\right) = \frac{M - M_{(-i)}}{M}, \tag{5}$$

where M denotes the selection criterion under consideration. Note that (5) may be calculated for all selection criteria where the particular criterion is a combination of some sort of goodness-of-fit measure and a penalty function (such a penalty function usually includes the number of predictors of the particular selected model as one of its components, see Kundu and Murali (1996)).

The proposed influence measure for the i-th case when the C_p criterion is used becomes

$$f(C_p\left(V\right), C_p\left(V_{(-i)}\right)) = \frac{C_p\left(V\right) - C_p\left(V_{(-i)}\right)}{C_p\left(V\right)}, \tag{6}$$

where $C_p\left(V_{(-i)}\right)$ is calculated as in (2), but with the i-th case omitted. Note that in calculating $C_p\left(V_{(-i)}\right)$ the estimator for the error variance is obtained from the full data set. The use of this error variance estimator is supported by considerations given by Léger and Altman (1993) for using $\hat{\sigma}^2$ in the denominator of the unconditional Cook's distance. The proposed influence measure based on the C_p criterion in (6) is large if the relative difference between $C_p\left(V\right)$ and $C_p\left(V_{(-i)}\right)$ is large. If this is true for an omitted data

case i, the particular case is considered possibly selection influential. Note that negative values of $C_p(V)$ may occur. These negative values may cause misrepresentation of the relative difference between $C_p(V)$ and $C_p\left(V_{(-i)}\right)$, *i.e.* the relative differences for certain data cases may now be incorrectly larger than others if, for example, $C_p(V) - C_p\left(V_{(-i)}\right)$ in the numerator, and $C_p(V)$ in the denominator of (6) are negative. We overcome this difficulty by omitting the subtraction of n and $(n-1)$ in the calculation of $C_p(V)$ and $C_p\left(V_{(-i)}\right)$ respectively. Thus, for instances where V and $V_{(-i)}$ are similar, or where V and $V_{(-i)}$ have the same number of indices, the sign of the numerator in (6) depends on the sizes of $RSS(V)$ and $RSS(V_{(-i)})$ respectively. Using Akaike's information criterion (AIC) for variable selection, the corresponding influence measure for the i-th case in (5) is

$$f\left(AIC(V), AIC\left(V_{(-i)}\right)\right) = \frac{AIC(V) - AIC\left(V_{(-i)}\right)}{AIC(V)}. \tag{7}$$

The value of $AIC\left(V_{(-i)}\right)$ in (7) is obtained by using either (3) or (4), but with the i-th case omitted.

4 Illustrative examples

The proposed influence measures in (6) and (7) are applied to four example data sets in this section.

4.1 The fuel data

Consider the fuel data given by Weisberg (1985, pp. 35, 36 and 126). There are 50 cases (one case for each of the 50 states in the USA). The response is the 1972 fuel consumption in gallons per person. The four predictor variables are:

x_1 : amount of tax on a gallon of fuel in cents;
x_2 : percentage of the population with a driver's license;
x_3 : average income in thousands of dollars; and
x_4 : total length of roads in thousands of miles.

Applying C_p selection to the full data set resulted in variables x_2 and x_3 being selected. The proposed influence measure in (6) is calculated for each data case i. The resulting values, together with the selected variables, Cook's unconditional selection distances, and estimated expected squared prediction errors are shown in Table 1 for some of the omitted data cases. The estimated average prediction errors for the reduced data sets were obtained by randomly selecting 39 of the 49 cases as a training data set and the other 10 cases as a test data set. Applying the C_p criterion to the training data set, the selected model was used to calculate an average squared prediction error for the 10 cases of the test data set. Random selection of a training data set and calculation of the average squared prediction error for the test data set, based on the selected model from the training data set, was repeated 20 000 times. Record was kept of the 20 000 average prediction errors. Their average was calculated in order to obtain an estimate of the expected squared prediction error. For the full data set (using 40 observations in the training data sets and 10 observations in the test data sets) this estimate equals 9 740.

Case omitted	Variables selected	Influence measure	Unconditional Cook's distance	Average prediction error
1	2,3	0.0016	0.0017	9 909
12	2,3	0.0004	0.0007	10 061
13	2,3	0.0006	0.0004	10 100
19	2,3	0.0557	0.1986	9 513
28	2,3	0.0009	0.0005	9 972
32	2,3	0.0005	0.0008	9 924
33	2,3	0.0224	0.0486	9 811
39	2,3	0.0192	0.0402	9 756
40	2,3,4	0.2038	0.8756	8 013
45	2,3	0.0652	0.1627	9 329
49	2,3,4	0.1024	1.0664	8 688
50	1,2,3	0.2658	2.4029	6 057

Table 1: *Fuel data: Variable selection with Mallows' C_p criterion.*

It is clear from Table 1 that the proposed influence measure and Cook's unconditional distance are relatively large when case 40, 49 or 50 is omitted. Both reach a maximum when data case 50 is omitted. We observe a sharp reduction in the estimated average prediction error (from 9 740 to 6 053) for the reduced data set where case 50 is omitted. The estimated average prediction errors for the reduced data sets without case 40 or 49 are also considerably smaller than that of the full data set, but not as low as that of the reduced data set with case 50 omitted. It would therefore seem to be advisable to omit data case 50 before performing variable selection and model fitting on the fuel data. Note that variables x_1, x_2 and x_3 are selected if case 50 is omitted.

There is a strong correspondence between the values of the proposed influence measure and the estimated average prediction errors. This is reflected in the correlation coefficient of -0.9716 between these two sets of numbers. The correlation between Cook's unconditional selection distance and the estimated average prediction error is -0.9657. Finally, the correlation between the proposed influence measure and Cook's unconditional distance is 0.9276, confirming a strong positive relationship between these two measures for this example.

Very similar results are obtained in Table 2 when the influence measure based on AIC in (7) is applied to the fuel data. We report on the same reduced data sets as in Table 1. Variable selection on the full data set by means of AIC also resulted in variables x_2 and x_3 being selected. With case 50 omitted, variables x_1, x_2 and x_3 are selected. Clearly, case 50 is once again selection influential. For all the other reduced sets (also those not shown) AIC selection results in the same variables selected as in the full data set. The proposed influence measure also identifies case 50 as the most influential. Cook's unconditional distance (also suggesting case 50 as most influential) remains unchanged from Table 1, except for the reduced data sets where cases 40 and 49 are respectively omitted. The estimated average prediction error for the full data set equals 9 613. The following correlations are also reported: between the proposed influence measure and the estimated average prediction error: -0.9809; between Cook's unconditional distance and the estimated average prediction error: -0.9199; between the proposed influence measure and Cook's unconditional distance: 0.8634.

Case omitted	Variables selected	Influence measure	Unconditional Cook's distance	Average prediction error
1	2,3	0.0176	0.0017	9 791
12	2,3	0.0174	0.0007	9 917
13	2,3	0.0174	0.0004	9 974
19	2,3	0.0244	0.1986	9 380
28	2,3	0.0175	0.0005	9 830
32	2,3	0.0174	0.0008	9 785
33	2,3	0.0201	0.0486	9 687
39	2,3	0.0198	0.0402	9 624
40	2,3	0.0471	0.3401	7 937
45	2,3	0.0257	0.1627	9 245
49	2,3	0.0309	0.3634	8 630
50	1,2,3	0.0575	2.4029	6 027

Table 2: *Fuel data: Variable selection with AIC.*

The proposed influence measures in (6) and (7) show that the maximum relative difference between M and $M_{(-i)}$ is obtained if case 50 is omitted. Cook's unconditional selection distance confirms the implication that case 50 is selection influential. The low values obtained for the estimated average prediction errors when case 50 is omitted supply strong evidence that omitting this case before performing variable selection improves the predictive power of the resulting model.

The following question arises: How should one judge the significance of a value of the proposed influence measure? In other words, should one recommend to a practitioner analyzing this data set to omit case 50 before performing variable selection? We attempt to answer this question by comparing the influence measures of data case 50 (0.2658 in (6) and 0.0575 in (7)) with the largest influence measures obtained in 10 000 residual bootstrap samples (Efron and Tibshirani, 1993). The bootstrap samples are obtained in the following way: Determine the vector of residuals of the form $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, once a linear regression model has been fitted to the complete data set. Random selection of 50 of these residuals with replacement yields a bootstrap vector of residuals, denoted by $\boldsymbol{r_b}$. If we calculate $\boldsymbol{r_b} + \hat{\boldsymbol{Y}}$ a new bootstrap response vector, denoted by $\boldsymbol{Y_b}$, is obtained. This newly formed bootstrap vector and the original set of unchanged predictor variable values constitute the bootstrap sample. Consider now 10 000 of these bootstrap samples, each of size 50 for the fuel data. The proposed influence measures in (6) and (7) are calculated for every single omitted data case in each of the 10 000 bootstrap samples. By keeping record of the largest values of (6) and (7) in every bootstrap sample, we are provided with two sets of values with which we may compare the largest influence measures (*i.e.* if data case 50 is omitted) of the fuel data set. The distributions of these 10 000 largest bootstrap influence measures, based on (6) and on (7) respectively, are shown in Figure 1 and 2.

The vertical line drawn in the class interval (0.26;0.29] on the histogram in Figure 1 shows the position of 0.2658 in the distribution. The proportion of bootstrap influence measures that are smaller than 0.2658, equals 0.8824. This implies that the value 0.2658 lies close to the 90th percentile of the bootstrap distribution. Similarly, the largest influence measure (0.0575) lies above the 83rd percentile in Figure 2. Providing this information to any practitioner who analyses the fuel data, will surely be helpful in the decision as to whether

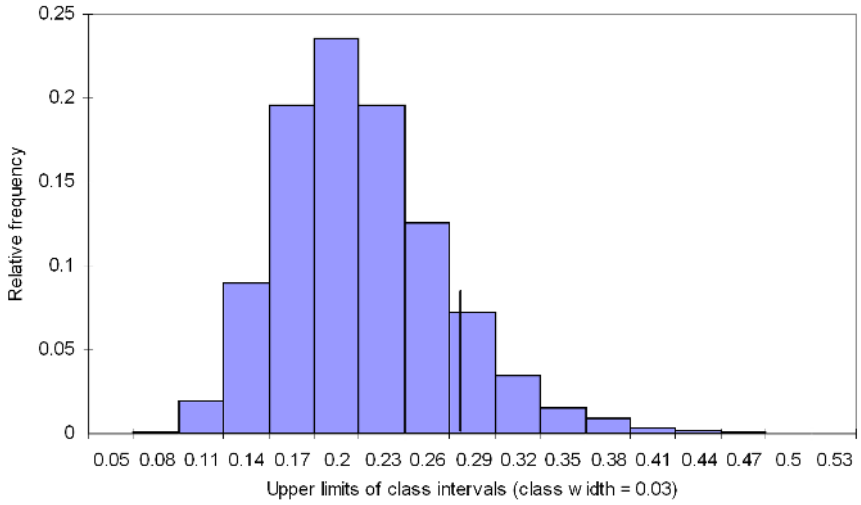

Figure 1: *Histogram of largest C_p influence measures in 10 000 bootstrap samples.*

case 50 should be omitted before subsequent analysis is performed.

It is important to bear in mind that the proposed influence measures only identify individual selection influential data cases. If it is, for example, decided to reject case 50 from the fuel data set, the influence measures should be recalculated on the $n-1$ remaining observations to identify other possibly selection influential data cases.

4.2 The evaporation data

We also apply the proposed influence measures to the evaporation data given by Freund (1979). Ten independent predictor variables were measured on 46 consecutive days, in accordance with the amount of evaporation from the soil, which represents the response.

The ten predictor variables are:

x_1 : maximum daily soil temperature;
x_2 : minimum daily soil temperature;
x_3 : integrated area under the soil temperature curve;
x_4 : maximum daily air temperature;
x_5 : minimum daily air temperature;
x_6 : integrated area under the daily temperature curve;
x_7 : maximum daily relative humidity;
x_8 : minimum daily relative humidity;
x_9 : integrated area under the daily humidity curve; and
x_{10} : total wind measured in miles per day.

Variable selection with C_p on the full data set yields variables x_1, x_3, x_6, x_8 and x_9 as the selected variables. Table 3 shows the following for some of the reduced data sets: the selected variables; values of the proposed influence measure in (6); Cook's unconditional distances; and the estimated expected squared prediction errors. Random selection of 37

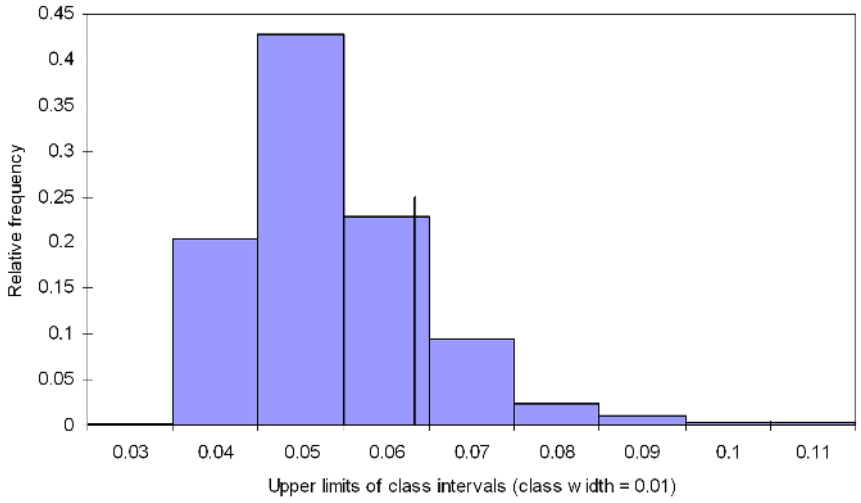

Figure 2: *Histogram of largest AIC influence measures obtained in 10 000 bootstrap samples.*

cases from the full data set in order to constitute a training data set, and using the model selected from this set to determine the average squared prediction error for the remaining 9 cases, resulted in an estimated average prediction error of 71, in 20 000 repetitions. The estimated values for the reduced data sets are based on 20 000 repetitions, each time using 36 cases in the training data set and 9 cases in the test data set.

Case omitted	Variables selected	Influence measure	Unconditional Cook's distance	Average prediction error
1	1,3,6,8,9	0.002	0.0048	77
4	1,3,6,8,9	0	0	74
8	1,3,6,9,10	0.1039	0.7118	62
21	1,3,6,8,9	0.0423	0.0389	71
22	1,3,6,9,10	0.0359	0.6148	73
26	1,3,5,7,8,9	0.021	0.5225	72
31	1,3,4,8,9	0.034	1.3462	69
32	1,3,6,9,10	0.0324	0.8613	73
33	1,3,6,9,10	0.1612	0.8349	60
40	6,9,10	0.0512	1.9332	71
41	1,3,6,9	0.2054	0.7768	52
46	1,3,6,9,10	0.0102	0.547	73

Table 3: *Evaporation data: Variable selection with Mallows' C_p criterion.*

The largest influence measures are obtained when case 41 is omitted, followed by case 33. The reduced data sets when cases 41 and 33 are respectively omitted also give the smallest values for the estimated average prediction error. The strong correspondence between the proposed influence measure and the estimated average prediction error for all 46 reduced data sets is reflected in a correlation coefficient of -0.9712. Note that Cook's unconditional distance is a maximum if a different case (*i.e.* case 40) is omitted. A correlation coefficient of only -0.4532 is obtained when the relationship between Cook's unconditional selection distance and the estimated average prediction error is considered. Finally, as expected,

the weak relationship between the proposed influence measure and Cook's unconditional distance is reflected in a correlation coefficient of 0.5022 for this example.

Applying AIC to the full data set results in variables x_1, x_3, x_6 and x_9 being selected, with the estimated average prediction error for the full data equal to 72. The corresponding AIC results, for the same reduced data sets as in Table 3, are shown in Table 4. The proposed influence measure in (7) also identifies cases 41 and 33 as most influential. The correlation between the influence measure in (7) and the estimated average prediction error (-0.9747) is higher than the correlation between Cook's unconditional distance and the estimated average prediction error (-0.334). The correlation between the proposed influence measure and Cook's unconditional distance equals 0.3664.

Case omitted	Variables selected	Influence measure	Unconditional Cook's distance	Average prediction error
1	1,3,6,9	0.0151	0.0076	77
4	1,3,6,9	0.0146	0.0011	75
8	1,3,6,9,10	0.0502	0.6989	63
21	1,3,6,9	0.028	0.0474	71
22	6,9,10	0.0272	2.2452	73
26	1,3,7,8,9	0.0193	2.0713	74
31	1,3,4,8,9	0.0249	2.4136	70
32	1,3,6,9,10	0.0243	0.9236	74
33	1,3,6,9,10	0.0731	0.935	61
40	6,9,10	0.0345	2.3096	70
41	1,3,6,9	0.0904	0.471	51
46	1,3,6,9	0.017	0.0079	74

Table 4: *Evaporation data: Variable selection with AIC.*

Omission of case 41 yields the largest values of measures (6) and (7) for the evaporation data. The low values obtained for the estimated average prediction error when case 41 is omitted provide strong evidence that omitting this case before doing variable selection improves the predictive power of the resulting model. Even stronger evidence is acquired when this value of the influence measure, if data case 41 is omitted, is compared with the largest influence measures obtained from 10 000 residual bootstrap samples. The histograms in Figure 3 and 4 show the distribution of these values for selection with C_p and AIC respectively. The vertical lines show the values of the influence measure if case 41 is omitted. This value lies above the 97th percentile in Figure 3 and above the 86th percentile in Figure 4.

4.3 Two further examples

In this section we present two further examples illustrating application of the proposed criteria. For the sake of brevity the results are merely described in the text, without a full summary in tables. Consider the Scottish hill racing data which have been examined by several authors such as Atkinson (1986) and Hoeting *et al.* (1996). The data investigate the relationship between the record-winning times for 35 hill races in Scotland with the following predictors:

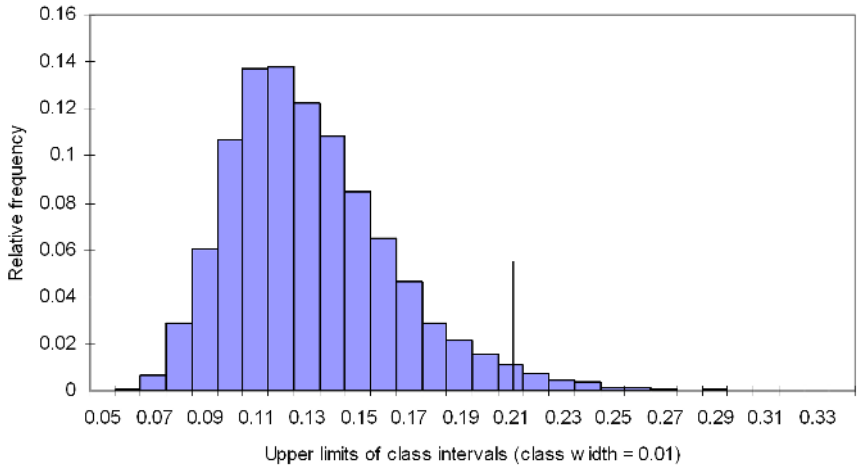

Figure 3: *Histogram of largest C_p influence measures in 10 000 bootstrap samples.*

x_1 : distance covered in miles; and
x_2 : elevation climbed during the race.

According to Atkinson (1986) the data contain a known error: observation 18 should be 18 minutes rather than 78 minutes. Here we specifically use the data set containing the error in order to evaluate the performance of the proposed influence measures in (6) and (7). Throughout, the results obtained when either C_p or AIC are applied to the data are very similar. Both variables are selected if the two selection techniques are applied to the full data set. The one-at-a-time omission of case 7 or 11 leads to variable x_1 being selected. For all the other reduced data sets the same variables as in the full data set are selected.

The largest influence measures in (6) and (7) are obtained when data case 18 is omitted, followed by cases 11 and 7. The influence measures for these cases (especially case 18) are significantly larger than when other cases are omitted. Cook's unconditional distance reaches a maximum when case 7 is omitted, with also relatively large values when cases 11 and 18 are omitted. There is a sharp reduction in the estimated average prediction error when case 18 is omitted. The estimated average prediction error for the reduced data sets without cases 11 and 7 are also significantly smaller than that of the full data set. For the full data set 28 observations are used in the training data sets and 7 observations in the test data sets.

Finally, consider the stack loss data which also have been examined by several authors (Brownlee, 1965; Atkinson, 1985; Hoeting *et al.*, 1996). The data investigate the relationship between the percentage of unconverted ammonia that escapes from a plant in 21 days, and the following three explanatory variables:

x_1 : air flow that measures the rate of operation of the plant;
x_2 : inlet temperature of cooling water circulating through coils in the tower; and
x_3 : value proportional to the concentration of acid in the tower.

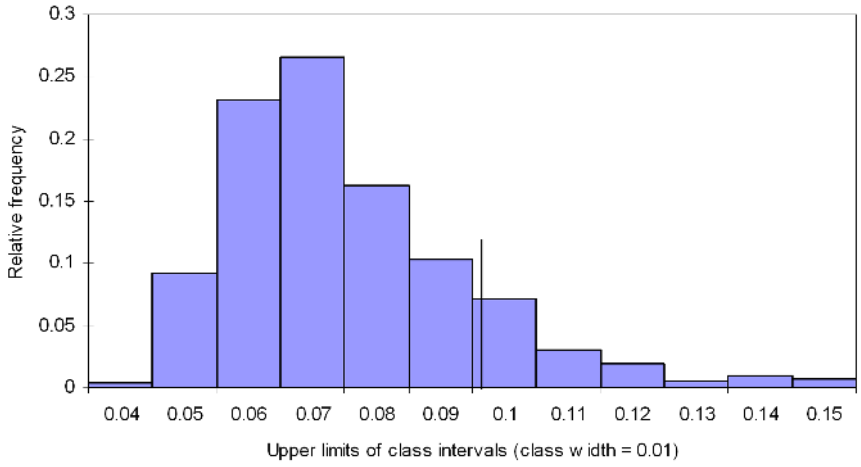

Figure 4: *Histogram of largest AIC influence measures obtained in 10 000 bootstrap samples.*

Again, as with the hill racing data, results obtained when either C_p or AIC are applied, coinside throughout. Variable selection with both techniques on the full and all reduced data sets consistently result in variables x_1 and x_2 as being selected. The influence measures show a relative maximum when case 21 is omitted, followed by cases 4, 3 and 1. The corresponding unconditional Cook's distances of these cases are also relatively larger than those of the other cases. In the same way the estimated average prediction errors if these cases are omitted are significantly smaller than the estimated average prediction error calculated on the full data set. For the full data set 16 observations are used in the training data sets and 5 observations in the test data sets.

5 Recommendations

Quantifying the influence of an individual data case in a selection context is important. The proposed influence measures may easily be calculated for any multiple regression sample that has to be analyzed. Once the influence measures have been obtained, a decision has to be taken as to whether to omit the data case with the largest influence measure, before selection is repeated on the reduced data set. The magnitude by which the estimated average prediction error decreases, if calculated for the complete and reduced data set, provides us with a good indication of whether the data case should be omitted. The illustrated bootstrap approach may also be utilized to judge the significance of the largest proposed influence measure. We strongly recommend that both these aspects (*i.e.*, the estimated average prediction error and the bootstrap distribution) should be taken into consideration before the data case with the largest influence measure is merely excluded from the regression sample.

The proposed influence measure is based on the C_p and AIC selection criteria. It may easily be extended to other selection criteria where the particular criterion is a combination of some sort of goodness-of-fit measure and a penalty function.

Since omission of individual cases does not address the problems of masking and swamping, two or more cases may be omitted at a time. This approach, however, causes difficulties with respect to computing time.

Acknowledgements

The authors would like to thank the anonymous referees whose valuable comments led to an improved version of the paper.

References

[1] Akaike H, 1973, *Information theory and an extension of the maximum likelihood principle*, The 2^{nd} International Symposium on Information Theory, Akademia Kiadó, Budapest, pp. 267–281.

[2] Atkinson AC, 1985*, Plots, transformations, and regression*, Clarendon Press, Oxford.

[3] Atkinson AC, 1986, Comments on *"Influential observations, high leverage points, and outliers in linear regression*, Statistical Science, **1**, pp. 397–402.

[4] Belsley DA, Kuh E & Welsch RE, 1980, *Regression diagnostics*, Wiley, New York (NY).

[5] Breiman L, 1992, *The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error*, Journal of the American Statistical Association, **87**, pp. 738–754.

[6] Brownlee KA, 1965, *Statistical theory and methodology in science and engineering*, 2^{nd} Edition, Wiley, New York (NY).

[7] Burnham KP & Anderson DR, 2002, *Model selection and multi-model inference*, Springer, New York (NY).

[8] Cook RD, 1977, *Detection of influential observations in linear regression*, Technometrics, **19**, pp. 15–18.

[9] Cook RD, 1986, *Assessment of local influence*, Journal of the Royal Statistical Society, Series B, **48**, pp. 133–169.

[10] Choongrak K & Hwang S, 2000, *Influential subsets on the variable selection*, Communications in Statistics: Theory and Methods, **29**, pp. 335–347.

[11] Efron B & Tibshirani RJ, 1993, *An introduction to the bootstrap*, Chapman and Hall, New York (NY).

[12] Freund RJ, 1979, *Multicollinearity etc.: Some "new" examples*, Proceedings of the Statistical Computing Section, American Statistical Association, pp. 111–112.

[13] Hoeting J, Raftery AE & Madigan D, 1996, *A method for simultaneous variable selection and outlier identification in linear regression*, Computational Statistics and Data Analysis, **22**, pp. 251–270.

[14] Kundu D & Murali G, 1996, *Model selection in linear regression*, Computational Statistics and Data Analysis, **22**, pp. 461–469.

[15] Léger C & altman N, 1993, *Assessing influence in variable selection problems*, Journal of the American Statistical Association, **88**, pp. 547–556.

[16] Liu H, Weiss RE, Jennrich RI & Wegner NS, 1999, *PRESS model selection in repeated measures data*, Computational Statistics and Data Analysis, **30**, pp. 169–184.

[17] Mallows CL, 1973, *Some comments on C_p*, Technometrics, **15**, pp. 661–675.

[18] Mallows CL, 1995, *More comments on C_p*, Technometrics, **37**, pp. 362–372.

[19] Miller AJ, 2002, *Subset selection in regression*, 2^{nd} Edition, Chapman and Hall, London.

[20] Murtaugh PA, 1998, *Methods of variable selection in regression modeling*, Communications in Statistics: Simulation and Computation, **27**, pp. 711–734.

[21] Venter JH & Snyman JLJ, 1997, *Linear model selection based on risk estimation*, Annals of the Institute of Statistical Mathematics, **49**, pp. 321–340.

[22] Weisberg S, 2005, *Applied linear regression*, 3^{rd} Edition, Wiley, New York (NY).