J. Japan Statist. Soc.
Vol. 32 No. 1 2002 43–55

ASSESSMENT OF INFLUENCE OF INDIVIDUAL OBSERVATIONS ON PREDICTION MEAN SQUARE ERRORS IN VARIABLE SELECTION PROBLEMS

Hidekazu Takeuchi*

A new influence measure is proposed to assess the influence of individual observations on prediction mean square errors (PMSE) in variable selection problems. It is based on the estimated PMSE which consists of Cook's distance and Mallows' C_P statistic. Another interpretation of Cook's distance is also given through the expression of the new influence measure. Illustrative examples show the effectiveness of the new influence measure.

Key words and phrases: Cook's distance, influence measure, influential observation, Mallows' C_P statistic, regression diagnostics, sensitivity analysis.

1. Introduction

Here we consider the detection of influential observations in variable selection problems in linear regression. Many studies have been published on the detection of influential observations. Cook and Weisberg (1982) and Chatterjee and Hadi (1988), for example, propose some influence measures for each observation in case of fixed variable subsets. A representative influence measure is Cook's distance proposed by Cook (1977).

Some papers deal with the detection of influential observations when the variable subsets are not fixed. Weisberg (1981) derives an influence measure based on C_P statistic suggested by Mallows (1973). The influence measure allocates C_P value to individual observations and consists of residual and leverage parts, as is usual in standard regression diagnostics. Léger and Altman (1993) propose an influence measure, which is Cook's distance computed from the difference between predicted values of the response variable, based on selected variable subsets with all observations and without one observation. They give a sensitivity analysis combined with the variable selection problem, where they take up Mallows' C_P statistic and step-wise regression procedures such as forward selection and backward elimination. Gupta and Huang (1996) introduce an influence measure as an alternative to Cook's distance to detect influential observations. They derive a measure of goodness of fit for the fitted models to select important variable subsets.

In this paper, we extend the above influence measures to assess the influence of an observation on prediction mean square errors for the selected variable subset. In usual regression diagnostics with Cook's distance, we detect influential

Received June 10, 2000. Revised December 4, 2001. Accepted December 25, 2001.

*Department of Contemporary Law, Tokyo Keizai University, 1-7, Minamicho, Kokubunji-shi, Tokyo 185-8502, Japan.

observations based on the difference between predicted values of the response variable with and without one observation as in e.g. Léger and Altman (1993). From the different viewpoint with the Léger and Altman's measure we derive a new influence measure. Based on the estimated prediction mean square errors we assess the influence of individual observations. We derive the new influence measure along the same line as in the derivation of Mallows' C_P statistic in the variable selection problem. So through the new influence measure based on the C_P type criterion we propose an influence detection procedure to assess the influence of individual observations on the estimated prediction mean square errors with relation to the variable selection problem. Surprisingly the new influence measure consists of Cook's distance and Mallows' C_P statistic. Based on the expression of the new influence measure, we also give another interpretation of Cook's distance, which may be of independent interest.

In Section 2 we give notation and definition. In Section 3 we propose the new influence measure on the estimated prediction mean square errors. We also propose an assessment procedure using the new influence measure. In Section 4 we give illustrative examples. Finally in Section 5 we give some comments on the new influence measure.

2. Notation and definition

We consider a standard linear regression model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ vector of a response variable, $\boldsymbol{X}$ is a full-rank $n \times q$ known matrix of predictor variables, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors with mean vector, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, and variance-covariance matrix, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. Here σ^2 is an unknown variance and $\boldsymbol{I}$ is a unit matrix. The ordinary least squares estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and an unbiased estimator of the variance σ^2 is given by $\hat{\sigma}^2 = \boldsymbol{e}'\boldsymbol{e}/(n-q)$, where $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ is the vector of residuals. In this definition $\boldsymbol{H}$ is the hat matrix defined as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, of which the i-th diagonal element is denoted by h_{ii}. Following the usual notation in the case deletion diagnostic procedure let a subscript (i) denote the omission of the i-th observation. For example we have the estimator of the variance with the i-th observation deleted as

$$\hat{\sigma}^2_{(i)} = \frac{n-q-t_i^2}{n-q-1}\hat{\sigma}^2,$$

where $t_i = e_i/(\hat{\sigma}\sqrt{1-h_{ii}})$ with the i-th element of $\boldsymbol{e}$, e_i.

Also for a variable subset, P, with p predictor variables, $\boldsymbol{X}_P$, we get some statistics as follows: We have $\hat{\sigma}^2_P = \boldsymbol{e}'_P\boldsymbol{e}_P/(n-p) = RSS_P/(n-p)$, where $\boldsymbol{e}_P = \boldsymbol{y} - \boldsymbol{X}_P\hat{\boldsymbol{\beta}}_P = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ with $\hat{\boldsymbol{\beta}}_P = (\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{X}'_P\boldsymbol{y}$ and $\boldsymbol{P} = \boldsymbol{X}_P(\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{X}'_P$. We denote $\hat{\boldsymbol{\beta}}_P$ with the i-th observation deleted as

$$\hat{\boldsymbol{\beta}}_{P(i)} = (\boldsymbol{X}'_{P(i)}\boldsymbol{X}_{P(i)})^{-1}\boldsymbol{X}'_{P(i)}\boldsymbol{y}_{(i)} = \hat{\boldsymbol{\beta}}_P - \frac{e_{Pi}}{1-p_{ii}}(\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{x}'_{Pi},$$

where $\boldsymbol{x}_{Pi}$ is the i-th row vector of $\boldsymbol{X}_P$, e_{Pi} is the i-th element of $\boldsymbol{e}_P$ and p_{ii} is the i-th diagonal element of $\boldsymbol{P}$.

A representative influence measure is Cook's distance proposed by Cook (1977). For a variable subset P, Cook's distance is defined by

$$\begin{aligned} CD_{Pi} &\equiv \frac{(\hat{\boldsymbol{\beta}}_P - \hat{\boldsymbol{\beta}}_{P(i)})'\boldsymbol{X}'_P\boldsymbol{X}_P(\hat{\boldsymbol{\beta}}_P - \hat{\boldsymbol{\beta}}_{P(i)})}{p\hat{\sigma}_P^2} \\ &= \frac{(\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_{P(i)})'(\hat{\boldsymbol{y}}_P - \hat{\boldsymbol{y}}_{P(i)})}{p\hat{\sigma}_P^2} \\ &= \frac{t_{Pi}^2}{p} \cdot \frac{p_{ii}}{1-p_{ii}}, \end{aligned} \tag{2.1}$$

where $\hat{\boldsymbol{y}}_P = \boldsymbol{X}_P\hat{\boldsymbol{\beta}}_P$, $\hat{\boldsymbol{y}}_{P(i)} = \boldsymbol{X}_P\hat{\boldsymbol{\beta}}_{P(i)}$ and $t_{Pi} = e_{Pi}/(\hat{\sigma}_P\sqrt{1-p_{ii}})$. Cook's distance has some interpretations as follows: From the first expression of (2.1) Cook's distance, CD_{Pi} is based on the weighted distance between $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_{P(i)}$. From the second expression of (2.1), CD_{Pi} is interpreted as an influence measure based on the difference between the predicted values $\hat{\boldsymbol{y}}_P$ and $\hat{\boldsymbol{y}}_{P(i)}$. Note that Léger and Altman (1993) utilize the second expression of (2.1) to introduce Cook's distance as a diagnostic measure conditionally on the selected model. In the unconditional case they propose another Cook's distance measuring the influence of individual observations based on the difference between the predicted values of the response variable with and without the i-th observation as

$$D_i^u \equiv \frac{(\hat{\boldsymbol{y}}^S - \hat{\boldsymbol{y}}^{S(i)})'(\hat{\boldsymbol{y}}^S - \hat{\boldsymbol{y}}^{S(i)})}{p\hat{\sigma}^2}, \tag{2.2}$$

where $\hat{\boldsymbol{y}}^S$ is the predicted vector based on the p variables selected with all observations and $\hat{\boldsymbol{y}}^{S(i)}$ is the predicted vector based on the variable subset selected with the i-th observation deleted. The third expression of (2.1) shows that CD_{Pi} enables us to detect influential observations through outlier and leverage effects.

Now we concisely introduce Mallows' C_P statistic given by

$$C_P \equiv \frac{RSS_P}{\hat{\sigma}^2} + 2p - n, \tag{2.3}$$

since we employ the similar derivation to have a new influence measure. Mallows (1973) derives (2.3) as an estimator of

$$\Gamma_P = \frac{\sum_j^n E(\hat{y}_{Pj} - \theta_j)^2}{\sigma^2} = \frac{SSB_P + \sum_j^n V(\hat{y}_{Pj})}{\sigma^2}, \tag{2.4}$$

where $\hat{y}_{Pj}$ is the j-th element of $\hat{\boldsymbol{y}}_P$, θ_j is the j-th element of $\boldsymbol{\theta} = E(\boldsymbol{y})$ and $SSB_P = (\boldsymbol{\eta} - \boldsymbol{\theta})'(\boldsymbol{\eta} - \boldsymbol{\theta}) = \boldsymbol{\theta}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\theta}$ is the sum of squared biases with $\boldsymbol{\eta} = E(\hat{\boldsymbol{y}}_P) = \boldsymbol{P\theta}$. (2.4) is based on the total sum of mean squares of the prediction errors, $\sum_j^n E(\hat{y}_{Pj} - \theta_j)^2$, and hereafter we denote it as PMSE (prediction mean square errors). Since $E(RSS_P) = (n-p)\sigma^2 + SSB_P$ we replace SSB_P by $RSS_P - (n-p)\sigma^2$. Furthermore, from $V(\hat{y}_{Pj}) = p_{jj}\sigma^2$, if we replace σ^2 by $\hat{\sigma}^2$ in (2.4), then we obtain (2.3) as an estimator of (2.4).

3. A new influence measure

Léger and Altman (1993) propose an influence detection procedure using Cook's distance based on the difference between the predicted values based on selected models with all observations and without one observation out of them. Although they emphasize the use of the squared distance based on the above predicted values, we take an influence detection procedure based on the estimated PMSE as in Section 3.1. Since our influence measure enables us to assess the influence of individual observations on the estimated PMSE, it will give a different diagnosis than the one of Léger and Altman (1993). We will give some comments on the difference between their and our influence assessment procedures through illustrative examples given in Section 4. Furthermore, we will see that the proposed influence measure surprisingly consists of Cook's distance and the usual Mallows' C_P statistic.

3.1. Main results

In the similar way to the derivation of the C_P statistic given by (2.3), we derive a new influence measure as follows: For the case when a variable subset P is selected, if the i-th observation is deleted, we define

$$\text{(3.1)} \qquad \Gamma_{P(i)} \equiv \frac{\sum_j^n E(\hat{y}_{Pj(i)} - \theta_j)^2}{\sigma^2} = \frac{SSB_{P(i)} + \sum_j^n V(\hat{y}_{Pj(i)})}{\sigma^2},$$

where $\hat{y}_{Pj(i)} = \boldsymbol{x}_{Pj}\hat{\boldsymbol{\beta}}_{P(i)}$. (3.1) corresponds to (2.4). Note that the numerator of (3.1) is PMSE when the i-th observation is deleted and that it is decomposed as

$$\text{(3.2)} \qquad \sum_j^n E(\hat{y}_{Pj(i)} - \theta_j)^2 = \sum_{j\neq i}^n E(\hat{y}_{Pj(i)} - \theta_j)^2 + E(\hat{y}_{Pi(i)} - \theta_i)^2,$$

where $\hat{y}_{Pi(i)} = \boldsymbol{x}_{Pi}\hat{\boldsymbol{\beta}}_{P(i)}$. The first term of (3.2) is PMSE other than the i-th observation and the second term is the one for the i-th observation when the i-th observation is deleted from the data set to estimate $\boldsymbol{\beta}_P$.

We note that we include PMSE of the i-th observation in (3.2) in the similar way to the derivation of some influence measures such as Cook's distance given by (2.1) or (2.2) proposed by Léger and Altman (1993). To make the fair comparison of the influence of individual observations, we do think it necessary for us to have the second term in (3.2).

Now we propose a new influence measure as an estimator of (3.1). We derive the estimator to separate the expression into some parts with relation to influence detection measures in the same way as in Takeuchi (1994). As is shown in Appendix, we can show that an estimator of (3.1) is given by

$$\text{(3.3)} \qquad C_{Pi}^T = C_P + p\frac{\hat{\sigma}_P^2}{\hat{\sigma}^2}CD_{Pi},$$

where C_P is the same Mallows' C_P statistic for the selected variable subset P as (2.3) and CD_{Pi} is the same Cook's distance as (2.1). Surprisingly the measure consists of Cook's distance and Mallows' C_P statistic in itself. Note that C_P is independent of the deletion of the observation and that for the common variable subset P, C_{Pi}^{T} is not less than C_P because $CD_{Pi} \geq 0$.

In a variable selection procedure, we must identify an optimal variable subset P to detect the influence of individual observations. Then we specify a variable selection criterion to select the optimal variable subset. Léger and Altman (1993) select the optimal variable subset with each observation deleted by using a specified variable selection criterion in the unconditional Cook's distance. Note that in the conditional Cook's distance the variable subset for each observation is common since the optimal one is selected in advance by using a specified variable selection criterion. Therefore, for each observation we select an optimal variable subset P in all variable subsets by using a specified variable selection criterion in the variable selection procedure. Our variable selection procedure basically uses the same C_P criterion in the unconditional Cook's distance.

It is quite interesting that we can give another interpretation of Cook's distance given by (2.1) through the expression of (3.3). Suppose that the variable subset P in (3.3) is fixed. Then the first term, C_P statistic, is a constant on the right-hand side of (3.3). The coefficient of the second term, $p\hat{\sigma}_P^2/\hat{\sigma}^2$, is also a positive constant. We can consider that Cook's distance measures the influence of the i-th observation on the estimated PMSE. It will mean that PMSE in (3.1) with the i-th observation deleted is larger as CD_{Pi} is larger.

In addition we may say that by employing (3.3) we assess the influence of individual observations on the squared distance from the predicted value based on the selected variable subset P to the true model since (3.3) is an estimator of (3.1). Usual regression diagnostics with Cook's distance may detect influential observations on the difference between the predicted values of the response variable with and without the i-th observation as in Léger and Altman (1993). Various influence measures based on the predicted values will be constructed in this manner. Thus our new influence measure will give a different diagnosis than theirs.

3.2. Assessment of influence

There are some approaches to assessing the influence of individual observations in variable selection problems. Léger and Altman's (1993) influence assessment procedure, which is one of representative approaches, consists of two stages. In the first stage, using a specified variable selection criterion, they select an optimal variable subset with each observation deleted. In the second stage, they calculate Cook's distance given by (2.1) or (2.2) for each observation to assess the influence of individual observations.

We employ C_{Pi}^{T} given by (3.3) to assess the influence of individual observations on the estimated PMSE. Our influence assessment procedure consists of the following three steps.

Step 1 We select an optimal variable subset P with the i-th observation deleted by using a specified variable selection criterion for all observations.

Step 2 For the selected variable subset P we have C^T_{Pi} by calculating C_P, $\hat{\sigma}^2_P$ and CD_{Pi}.

Step 3 We may judge the i-th observation to be influential if the value of C^T_{Pi} can be regarded as large.

Note that if for all observations a common variable subset happens to be selected, then we actually take the usual diagnostic procedure based on Cook's distance only.

In Step 1 we will use the usual Mallows' C_P criterion to select an optimal variable subset for each observation. In Step 2 we calculate all fundamental statistics with relation to the new influence measure C^T_{Pi} given by (3.3). In Step 3 we assess the influence of individual observations for the selected variable subset through the new influence measure. Therefore Steps 1 and 3 correspond to the variable selection and the influence detection procedures respectively.

In Step 3 we have no exact criterion to judge the observation to be influential. Usual regression diagnostics in general compare relative values of influence measures for each observation, or compare them with proper guidelines derived from various properties of the influence measures. We basically use the value of the minimum C_P as a guideline. However it is not always the best guideline for real data sets. In practice we may use a rough-and-ready procedure to pay attention to the observation with $C^T_{Pi} > C_{P*}$, where C_{P*} denotes the second minimum C_P value in all variable subsets, since it is clear that C^T_{Pi} is larger than the minimum C_P. Thus by comparing with the second minimum C_P value we can assess the influence of the i-th observation on the estimated PMSE. This rough-and-ready procedure may give an efficient guideline to assess the influence of individual observations in Step 3.

4. Illustrative examples

We give four examples to illustrate the effectiveness of the new influence measure given by (3.3). Applying our influence assessment procedure given in Section 3.2 to the following data sets we give some comments on the new influence measure. For all the examples we use Mallows' C_P criterion as the specified variable selection one in Step 1.

4.1. Artificial data

In this section we give two remarkable examples based on artificial data sets to show some basic properties of the new influence measure. The first example is the simple regression case and the second one is the multiple regression case.

4.1.1. Example 1

This data set is shown in Table 1. In simple regression there are $(n =)$ 6 observations. For the variable subset $P = \{X1\}$, $C_P = 2.000$ is the minimum. For the variable subset $P = \{\text{None}(X0)\}$, i.e. intercept only, $C_P = 3.458 (= C_{P*})$ is the second minimum.

In Table 2 we summarize the result of our influence assessment procedure given in Section 3.2. In Step 3 the observations No. 1, No. 5 and No. 6 can be

Table 1. Simple Regression Data.

No.	X	y
1	0.95	1.78
2	1.04	1.95
3	2.03	1.95
4	1.99	2.15
5	3.06	1.99
6	3.00	2.25

Table 2. Assessment of Influence.

No.	Subset P	C_{Pi}^T	D_i^u
1	$(X0)$	3.574	2.078
2	$X1$	2.202	0.101
3	$X1$	2.053	0.027
4	$X1$	2.259	0.129
5	$X1$	3.666	0.833
6	$(X0)$	3.581	2.098

regarded as influential. Applying the rough-and-ready procedure with the second minimum C_P, C_{P*}, in Step 3, we also get the same three influential observations.

For comparison we apply the Léger and Altman's influence assessment procedure with the usual C_P statistic as the variable selection criterion. As a result of the unconditional Cook's distance, D_i^u, given by (2.2) the observations No. 1, No. 5 and No. 6 may be regarded as influential since they have $D_i^u > 0.800$, whereas the other observations have $D_i^u < 0.130$ as in Table 2. However D_5^u for the observation No. 5 may not be regarded as influential in their procedure since it is smaller than half the values of D_1^u and D_6^u. Therefore, the observation No. 5 can be regarded as influential in our procedure, whereas it may not be so in their procedure.

Furthermore in the conditional Cook's distance, CD_{Pi}, given by (2.1) we have $CD_{P1} = 0.376$ and $CD_{P6} = 0.408$ for the optimal variable subset $P = \{X1\}$. The other observations have the same values as the unconditional Cook's distance D_i^u given in Table 2 in simple regression. Thus the observation No. 5 may be regarded as the most influential in the conditional Cook's distance and also in our procedure.

The observation No. 5 is the largest in CD_{Pi} and C_{Pi}^T. On the other hand, the observation No. 6 is so in D_i^u. Therefore we may consider that for the observation No. 5 the estimated squared distance between the true model and the predicted value in C_{Pi}^T is much larger than the squared distance between the predicted values based on the selected models with all observations and without the one out of them in D_i^u.

4.1.2. Example 2

This data set is shown in Table 3. There are $(n =)8$ observations with 2 predictor variables. For the variable subset $P = \{\text{None}(X0)\}$, i.e. intercept only,

Table 3. Multiple Regression Data.

No.	$X1$	$X2$	y
1	0.00	2.00	2.00
2	0.95	3.73	2.00
3	1.05	0.27	2.02
4	1.98	0.00	3.00
5	2.01	4.00	2.10
6	3.00	3.72	2.20
7	3.02	0.25	2.25
8	3.98	2.00	2.30

Table 4. Assessment of Influence.

No.	Subset P		C^T_{Pi}	D^u_i
3		$X2$	2.583	4.340
4	$X1$		3.788	1.855
7		$X2$	1.954	2.678

$C_P = 1.683$ is the minimum in all variable subsets. For the variable subset $P = \{X2\}$, $C_P = 1.832(= C_{P*})$ is the second minimum. For the variable subset $P = \{X1\}$, $C_P = 2.822$ is the third minimum. For the variable subset $P = \{X1, X2\}$, $C_P = 3.000$ is the maximum.

We summarize the main result of the sensitivity analysis with (3.3) through our influence assessment procedure in Steps 1 and 2 as in Table 4. The other observations except the ones listed in it have the C^T_{Pi}'s between 1.683 and 1.694 for the variable subset $P = \{\text{None}(X0)\}$. So we can regard the observations No. 3, No. 4 and No. 7 as influential in Step 3. Applying the rough-and-ready procedure with the second minimum C_P in Step 3, we also get the same three influential observations.

For comparison we apply the Léger and Altman's influence assessment procedure in the same way as in Section 4.1.1. In the unconditional Cook's distance the observations No. 3, No. 4 and No. 7 may be regarded as influential since they have $D^u_i > 1.855$, whereas the other observations have $D^u_i < 0.090$. In the conditional Cook's distance we have $CD_{P3} = 0.009$, $CD_{P4} = 0.110$ and $CD_{P7} = 0.000$ for the optimal variable subset $P = \{\text{None}(X0)\}$. The other observations have $CD_{Pi} < 0.010$. Thus the observation No. 4 only may be regarded as influential in the conditional Cook's distance.

The observation No. 3 is the largest in D^u_i. The observation No. 4 is the largest in CD_{Pi} and C^T_{Pi}. On the other hand, the observation No. 7 is the smallest in CD_{Pi}, whereas it is the second largest in D^u_i and is the third in C^T_{Pi}. In addition we may say that the observation No. 7 is regarded as influential in D^u_i, although it can not be always regarded as large in C^T_{Pi}. Then the three influence measures may assess the influence of individual observations from the different viewpoints of the squared distance based on the predicted value as discussed in Section 4.1.1.

4.2. Real data

We give two well-known real data sets to exemplify the effectiveness of the new influence measure.

4.2.1. Longley data

This data set as in Longley (1967) is often used in regression analysis. There are (n =)16 observations with 6 predictor variables, where $X1$ is the implicit price deflator for the gross national product, $X2$ is the gross national product, $X3$ is the unemployment, $X4$ is the size of armed forces, $X5$ is the noninstitutional population 14 years of age and over, and $X6$ is the calendar year. The response variable y is the total derived employment. For the variable subset $P = \{X2, X3, X4, X6\}$, $C_P = 3.240$ is the minimum in all variable subsets. For the variable subset $P = \{X3, X4, X5, X6\}$, $C_P = 4.606 (= C_{P*})$ is the second minimum. For the variable subset $P = \{X3, X4, X6\}$, we have $C_P = 6.239$, which is the sixth minimum.

We summarize the result of our influence assessment procedure in Steps 1 and 2 as in Table 5. From it the observations No. 5, No. 10 and No. 16 can be regarded as influential in Step 3 since they have larger values than the other observations. Applying the rough-and-ready procedure with the second minimum C_P in Step 3, we also detect the same influential observations No. 5, No. 10 and No. 16 because C^T_{P5}, C^T_{P10} and C^T_{P16} have larger values than C_{P*}.

For comparison we apply the Léger and Altman's influence assessment procedure with the usual C_P statistic as the variable selection criterion. In the unconditional Cook's distance, the observations No. 5, No. 10 and No. 16 may be regarded as influential since they have $D^u_5 = 0.387$, $D^u_{10} = 1.199$ and $D^u_{16} = 1.352$, respectively. The other observations have $D^u_i < 0.172$. In the conditional Cook's distance the observations No. 4, No. 5, No. 6, No. 10, No. 15 and No. 16 may be regarded as influential since they have $CD_{P4} = 0.178$, $CD_{P5} = 0.461$, $CD_{P6} = 0.109$, $CD_{P10} = 0.361$, $CD_{P15} = 0.205$ and $CD_{P16} = 0.303$ for the optimal variable subset $P = \{X2, X3, X4, X6\}$, respectively. The other observations have $CD_{Pi} < 0.075$ for the same variable subset. In particular we may pay attention to the observations No. 5, No. 10 and No. 16. Therefore the observations

Table 5. Assessment of Influence.

No.	Subset P	C^T_{Pi}	D^u_i	No.	Subset P	C^T_{Pi}	D^u_i
1	$X2, X3, X4, X6$	3.554	0.063	11	$X2, X3, X4, X6$	3.240	0.000
2	$X2, X3, X4, X6$	3.282	0.008	12	$X2, X3, X4, X6$	3.241	0.000
3	$X2, X3, X4, X6$	3.254	0.003	13	$X2, X3, X4, X6$	3.261	0.004
4	$X2, X3, X4, X6$	3.988	0.150	14	$X2, X3, X4, X6$	3.258	0.004
5	$X2, X3, X4, X6$	5.177	0.387	15	$X2, X3, X4, X6$	4.101	0.172
6	$X2, X3, X4, X6$	3.696	0.091	16	$X3, X4, X6$	8.001	1.352
7	$X2, X3, X4, X6$	3.507	0.054				
8	$X2, X3, X4, X6$	3.269	0.006				
9	$X2, X3, X4, X6$	3.240	0.000				
10	$X3, X4, X6$	7.232	1.199				

No. 5, No. 10 and No. 16 may be also regarded as influential in their procedure. We basically get the same conclusion from both their and our procedures.

4.2.2. Fuel data

We utilize the same data set as in Léger and Altman (1993). The data set is called the Fuel Data by Weisberg (1985, pp. 36–37 and 126). There are (n =)50 observations with 4 predictor variables, where $X1$ (TAX) is the motor fuel tax rate, $X2$(DLIC) is the percent of population with driver's licenses, $X3$(INC) is the average income, and $X4$(ROAD) is the miles of federal-aid primary highways. The response variable y is the motor fuel consumption. For the variable subset $P = \{X2, X3\}$, $C_P = 2.517$ is the minimum in all variable subsets. For the variable subset $P = \{X2, X3, X4\}$, $C_P = 3.313$ ($= C_{P*}$) is the second minimum. For the variable subset $P = \{X1, X2, X3\}$, $C_P = 3.526$ is the third minimum.

Following our influence assessment procedure given in Section 3.2 we summarize the main result of the sensitivity analysis with (3.3) as in Table 6. For the variable subset $P = \{X2, X3\}$, the C^T_{Pi}'s have the values between 2.517 and 2.663 except for the five observations listed in it.

From Table 6 we can regard the observations No. 19, No. 40, No. 45, No. 49 and No. 50 as influential in Step 3. Applying the rough-and-ready procedure with the second minimum C_P in Step 3, we can reduce these five observations to the ones No. 40, No. 49 and No. 50.

For comparison we apply the Léger and Altman's influence assessment procedure in the same way as in Section 4.2.1. In the unconditional Cook's distance the observations No. 40, No. 49 and No. 50 may be regarded as influential since they have $D^u_{40} = 0.876$, $D^u_{49} = 1.066$ and $D^u_{50} = 2.403$, respectively. The other observations have $D^u_i < 0.200$. In the conditional Cook's distance the observations No. 19, No. 40, No. 45, No. 49 and No. 50 may be regarded as influential since they have $CD_{P19} = 0.201$, $CD_{P40} = 0.344$, $CD_{P45} = 0.164$, $CD_{P49} = 0.367$ and $CD_{P50} = 0.325$ for the optimal variable subset $P = \{X2, X3\}$, respectively. The other observations have $CD_{Pi} < 0.050$ for the same variable subset. In particular we may pay attention to the observations No. 40, No. 49 and No. 50. So in their procedure the above three observations may be regarded as influential.

The order of the three observations to be regarded as influential in their procedure is different from the one in our procedure. The most influential observation is No. 50 in D^u_i and C^T_{Pi} and is No. 49 in CD_{Pi}. The second one is No. 40 in C^T_{Pi}, No. 49 in D^u_i, and No. 50 in CD_{Pi}. The third one is No. 49 in C^T_{Pi} and is No. 40 in D^u_i and CD_{Pi}. Therefore our procedure using the new

Table 6. Assessment of Influence.

No.	Subset P	C^T_{Pi}	D^u_i
50	$X1, X2, X3$	6.968	2.403
40	$X2, X3, X4$	4.654	0.876
49	$X2, X3, X4$	4.394	1.066
19	$X2, X3$	3.113	0.199
45	$X2, X3$	3.005	0.163

influence measure may give the different influence assessment than theirs from the different viewpoints of the prediction as in Section 4.1.

5. Concluding remarks

We propose a new influence measure based on the estimated PMSE. The new influence measure has an interesting expression. Surprisingly, the new influence measure consists of Cook's distance and Mallows' C_P statistic. We may consider that apart from the constant terms Cook's distance measures the influence of each observation on the estimated PMSE. A major advantage is that the new influence measure enables us to assess the influence of individual observations on the estimated PMSE in variable selection problems. From another viewpoint, employing the new influence measure based on the estimated PMSE we can investigate the influence of individual observations on the distance from the predicted value to the true model.

The illustrative examples in Section 4 show the effectiveness of the new influence measure. Our influence assessment procedure using the new influence measure may assess the influence of individual observations from viewpoints different from the Léger and Altman's (1993) one. However if an observation is regarded as influential in either influence assessment procedure, then it may have a large influence on the squared distance employed in the procedure. Therefore both influence assessment procedures may be important to assess the influence of individual observations in the variable selection problem.

Appendix: Derivation of (3.3)

In the similar way to the derivation of Mallows' C_P statistic given by (2.3) we can derive (3.3) as follows: From (3.2) we have

$$\begin{aligned}\sum_{j}^{n} E(\hat{y}_{Pj(i)} - \theta_j)^2 &= \sum_{j\neq i}^{n} V(\hat{y}_{Pj(i)}) + V(\hat{y}_{Pi(i)}) \\ &\quad + \sum_{j\neq i}^{n} [E(\hat{y}_{Pj(i)}) - \theta_j]^2 + [E(\hat{y}_{Pi(i)}) - \theta_i]^2.\end{aligned}$$

Since we have

$$V(\hat{y}_{Pj(i)}) = \sigma^2 \boldsymbol{x}_{Pj} (\boldsymbol{X}'_{P(i)} \boldsymbol{X}_{P(i)})^{-1} \boldsymbol{x}'_{Pj},$$

and

$$V(\hat{y}_{Pi(i)}) = \sigma^2 \boldsymbol{x}_{Pi} (\boldsymbol{X}'_{P(i)} \boldsymbol{X}_{P(i)})^{-1} \boldsymbol{x}'_{Pi},$$

we get

$$\begin{aligned}\sum_{j}^{n} V(\hat{y}_{Pj(i)}) &= \mathit{trace}\, \sigma^2 \boldsymbol{X}_P (\boldsymbol{X}'_{P(i)} \boldsymbol{X}_{P(i)})^{-1} \boldsymbol{X}'_P \\ &= \sigma^2\, \mathit{trace} \left[\boldsymbol{P} + \frac{1}{1-p_{ii}} \boldsymbol{X}_P (\boldsymbol{X}'_P \boldsymbol{X}_P)^{-1} \boldsymbol{x}'_{Pi} \boldsymbol{x}_{Pi} (\boldsymbol{X}'_P \boldsymbol{X}_P)^{-1} \boldsymbol{X}'_P \right] \\ &= \sigma^2 \left(p + \frac{p_{ii}}{1-p_{ii}} \right).\end{aligned}$$

Defining $\eta_{j(i)} = E(\hat{y}_{Pj(i)})$ and $\eta_{i(i)} = E(\hat{y}_{Pi(i)})$, we get the following $n \times 1$ vector,

$$\begin{aligned}\boldsymbol{\eta}_{(i)} &= E(\hat{\boldsymbol{y}}_{P(i)}) = E(\boldsymbol{X}_P\hat{\boldsymbol{\beta}}_{P(i)})\\ &= E(\boldsymbol{X}_P\hat{\boldsymbol{\beta}}_P) - \frac{E(e_{Pi})}{1-p_{ii}}\boldsymbol{X}_P(\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{x}'_{Pi}\\ &= \boldsymbol{\eta} - \frac{E(y_i)-E(\hat{y}_{Pi})}{1-p_{ii}}\boldsymbol{X}_P(\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{x}'_{Pi}\\ &= \boldsymbol{\eta} + \frac{\eta_i-\theta_i}{1-p_{ii}}\boldsymbol{X}_P(\boldsymbol{X}'_P\boldsymbol{X}_P)^{-1}\boldsymbol{x}'_{Pi},\end{aligned}$$

where y_i and η_i are the i-th elements of $\boldsymbol{y}$ and $\boldsymbol{\eta}$ respectively. Then from $\boldsymbol{X}'_P(\boldsymbol{\eta}-\boldsymbol{\theta}) = -\boldsymbol{X}'_P(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\theta} = \boldsymbol{0}$ we have

$$\begin{aligned}\sum_j^n[E(\hat{y}_{Pj(i)})-\theta_j]^2 &= \sum_{j\neq i}^n(\eta_{j(i)}-\theta_j)^2 + (\eta_{i(i)}-\theta_i)^2\\ &= (\boldsymbol{\eta}_{(i)}-\boldsymbol{\theta})'(\boldsymbol{\eta}_{(i)}-\boldsymbol{\theta})\\ &= SSB_P + \frac{p_{ii}}{(1-p_{ii})^2}(\eta_i-\theta_i)^2.\end{aligned}$$

Therefore we can reduce (3.1) to

$$\Gamma_{P(i)} = \frac{SSB_P + \dfrac{p_{ii}}{(1-p_{ii})^2}(\eta_i-\theta_i)^2}{\sigma^2} + p + \frac{p_{ii}}{1-p_{ii}}.$$

Following the derivation of the C_P statistic given by (2.3) we substitute the unbiased estimators for the parameters. From $E(RSS_P) = (n-p)\sigma^2 + SSB_P$ and $E(e_{Pi}^2) = (1-p_{ii})\sigma^2 + (\eta_i-\theta_i)^2$ we replace SSB_P by $RSS_P - (n-p)\sigma^2$ and $(\eta_i-\theta_i)^2$ by $e_{Pi}^2 - (1-p_{ii})\sigma^2$ respectively. Then we get

$$\text{(A.1)}\qquad \hat{\Gamma}_{P(i)} = \frac{RSS_P}{\sigma^2} + 2p - n + p\frac{\hat{\sigma}_P^2}{\sigma^2}\cdot\frac{1}{p}\left(\frac{e_{Pi}}{\hat{\sigma}_P\sqrt{1-p_{ii}}}\right)^2\frac{p_{ii}}{1-p_{ii}}.$$

Setting $\sigma^2 = \hat{\sigma}^2$ in (A.1) we have (3.3) as an estimator of (3.1).

Acknowledgements

The author is grateful to the referee for insightful comments on the early manuscript.

References

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.

Cook, R. D. (1977). Detection of influential observations in linear regression, *Technometrics*, **19**, 15–18.

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

Gupta, S. S. and Huang, D.-Y. (1996). On detecting influential data and selecting regression variables, *Journal of Statistical Planning and Inference*, **53**, 421–435.

Léger, C. and Altman, N. (1993). Assessing influence in variable selection problems, *Journal of the American Statistical Association*, **88**, 547–556.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user, *Journal of the American Statistical Association*, **62**, 819–841.

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

Takeuchi, H. (1994). Sensitivity analysis with an extension of Cook's distance in ridge regression, *Journal of the Japan Statistical Society*, **24**, 221–236.

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

Weisberg, S. (1985). *Applied Linear Regression*, 2nd edition, Wiley, New York.