응용통계연구 제 6권 2호
1993년 9월 pp. 421-433.

Influential Observations on Variable Selection in Linear Regression Model[1)]

Chi-Hoon Choi[2)], Ja-Heung Koo[2)], Jae June Lee[2)], Hongsuk Jorn[2)]

Abstract

Few observation can influence in model building procedure and can dominate the least squares fit of a selected model. An observation, however, may not have the same impact on all aspects of regression analysis. We introduce a statistic which measures the impact of individual cases on the overall goodness-of-fit statistics. We also propose an influence measure for variable selection problem. The property of uncorrelatedness between fitted values and residuals has been used to develop the influence measure. The performance of the measures are compared with other widely used influence measures by the analysis of real data.

1. Introduction

Regression analysis is widely used in data analysis and the development of empirical models. Meanwhile, it is well known that a few observations can influence in model building and can dominate the least squares fit of a selected model. Understanding the impact of those observations on the analysis procedures are essential to enhance the accuracy of the statistical analysis.

Various influence measures have been introduced in the context of selected regression models, based on the idea of case deletion, in general, see Cook and Weisberg(1982), Belsley et al.(1980), and Chatterjee and Hadi(1986). On the other hand, only few advice has been introduced so far for assessing influence when fitted model is chosen by a variable selection procedure. For example, Chatterjee and Hadi(1988) and Schall and Dunne (1990) studied the impact of simultaneously omitting a case and a variable from a full model, and Weisberg(1981) introduced a statistic to evaluate the contribution of each case to Mallow's C_p (Mallows, 1973). Leger and Altman(1993) introduced a statistics, namely unconditional Cook's distance, to assess the infuence of each case on the variable selection procedure.

In linear regression, the statistical quantities estimated from a least squares fit can be substantially affected by a few or even by a single observation. An observation, however, may not have the same impact on all aspects of regression analysis, as Chatterjee and Hadi(1986) start asking "Influence on what ?". In this paper, we introduce a statistic which measures influence of each case on overall goodness-of-fit in a selected model. The influence of each case is measured by the changes of the coefficient of determination (R^2) estimates when the case is deleted. In addition, we propose a statistic to assess the

1) This Paper was supported by Non DIRECTED RESEARCH FUND, Korea Research Foundation, 1991
2) Department of Statistics, Inha University, #253 Yonghyun-Dong, Namku, Inchon, 402-751

influence of individual observations on the process of variable selection. The proposed statistic has been developed to measure the impact of each case on the "all possible" regression procedure when the case is deleted.

We consider the linear regression model

$$y = X\beta + \varepsilon \tag{1.1}$$

where y is an n-vector of responses, X is an $n \times p$ full rank matrix of p independent variables possibly including one constant predictor, and ε is an n-vector of unobservable errors with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 I$. Throughout this paper, matrices and column vectors are denoted by uppercase and lowercase letters, respectively. Thus the i-th case is denoted by (y_i, x_i^t). The subscript notation (i) is used to indicate the deletion of the i-th observation, and the special character $\hat{}$ above any quantity is used to mean an estimator based on the method of least squares.

2. Influence Measure based on the Coefficient of Determination

Several methods have been introduced to identify observations with larger influence and evaluate the effects on the various aspects of the regression analysis, for example on the estimated parameters or predictions. All those measures have been constructed, based on residuals, the projection matrix, the volume of confidence ellipsoids, influence functions, and/or partial influence, see Chatterjee and Hadi(1986).

As is well-known, the coefficient of determination (R^2) is the maximum correlation between y and x. In finite samples, an observation deviating from the linear structure between y and x could change the estimated R^2 significantly. Lee(1990) pointed out this problem and proposed an influence measure which is based on R^2 estimates. We newly derive and introduce the statistic which measure the influence of each case on the goodness-of-fit statistic R^2 when the case is deleted.

2.1 An Influence Measure

Assuming that both y and x_1 are random, and $z=(y : x_1)^t$ has a joint df F, a random predictor linear model can be constructed as follows:

$$y = \beta_0 + x_1^t \beta_1 + \varepsilon \tag{2.1}$$

where $\beta_0 + x_1^t \beta_1$ is the regression equation and the error ε, having $E(\varepsilon)=0$ and $var(\varepsilon)=\sigma_\varepsilon^2$, is uncorrelated with the components of x_1. If we put $\beta_0 = \mu_Y - \mu_{x_1}^t \beta_1$, the functional expression of the squared multiple correlation R^2 can be expressed as

$$R^2(F) = 1 - \frac{\sigma_\varepsilon^2(F)}{\sigma_{yy}(F)} \tag{2.2}$$

where $E_F(z) = (\mu_Y(F):\mu_{x_1}^t(F))^t$, $Var_F(\varepsilon) = \sigma_\varepsilon^2(F)$, and $Var_F(y) = \sigma_{yy}(F)$.

Motivated by the derivation in the Appendix, we define

$$dT_1(z;F) = \frac{(1-R^2(F))(y-\mu_y(F))^2-\varepsilon^2}{\sigma_{yy}(F)} \tag{2.3}$$

The estimators of statistical quantities obtained by least squares fit of (1.1) can be expressed as functionals evaluated at an empirical df, for example $\hat{F} = \sum_{i=1}^{n} \frac{\delta_{z_i}}{n}$ or $\hat{F}_{(i)} = \sum_{j \neq i} \frac{\delta_{z_j}}{n-1}$ where δ_z is the df which puts mass 1 at the point z. An influence measure, therefore, can be obtained by substituting a version of empirical df in (2.3). By setting $F=\hat{F}$ and taking $\lambda = \frac{-1}{n-1}$ in (A.2), we define an influence measure,

$$DT_1(z_i) = \gamma(i)\left[\frac{(1-\hat{R}^2)(y_i-\bar{y})^2 w(n) - \frac{e_i^2}{1-h_{ii}}}{\hat{\sigma}_{yy}}\right] \tag{2.4}$$

where $\gamma(i) = \frac{\hat{\sigma}_{yy}}{\hat{\sigma}_{yy(i)}}$, $w(n) = \frac{n}{n-1}$, e_i is the residual of i-th case, and h_{ii} is the i-th diagonal element of the hat matrix.

Different influence measures, besides $DT_1(z)$, can be developed by putting different versions of empirical df and λ in (A.2) and (2.3). Introduction of those measures, properties of them, and comparison with widely used influence measures such as Cook's distance (Cook, 1977) and DFFITS (Belsley et al., 1980) will be given in a separate paper later.

It is often useful to have a reference value to determine which values of an influence measure are actually "large". As suggested by Lee(1990), similar cut-offs can be used for the measure DT_1. That is, a case is identified as an influential case if absolute value of $DT_1(z)$ is larger than $k \times \sqrt{2}(1-\hat{R}^2)\frac{(n-1)}{(n-p)}$. For moderate to large sample size n, k=2 or 3 can be used.

In diagnostics, distribution theory and testing hypotheses can lead to incorrect conclusions if more than one outliers are present. The criterion for each measure, therefore, should be used as a rough guide to identify influential cases. The order based on each measure may be the more important information we can get from the data, and investigation on the few cases determined by the order would be a more reasonable approach.

2.2 Application to Real Data

The data set used to evaluate the performance of the influence measure is the Fuel data of Weisberg(1985). There are 50 observations and 4 predictors. The response is the 1972 fuel consumption in gallons per person. The predictors variables are Tax, Dlic, Inc, and Road. The full model and the model with minimum C_p using all possible regression are used in this evaluation.

For the full model and the subset model with minimum C_p, the values of the DT_1, Cook's distance, DFFITS, and *COVRATIO* (covariance ratio) of notable observations are given in Table 2.1 and Table 2.2 , respectively.

The analysis based on the influence measures leads to the following findings:

Table 2.1 :Fuel Data

Influence Measures in the Full Model (* : Influential case)

case	DT_1	Cook	DFFITS	COVRATIO	r_i	h_{ii}
Hawaii	-7.73**	1.47*	-3.30*	0.21*	-4.77	0.32
Wyoming	-0.12	0.22	1.20*	0.28*	3.87	0.09
Alaska	-3.37**	0.28	1.25*	0.68*	2.59	0.19
New York	2.20*	0.01	-0.24	1.47*	-0.40	0.25
South Dakota	2.15*	0.13	0.81*	0.97	1.74	0.18
Other cases Influential			NV	CN, IL TX		

Table 2.2 : Fuel Data

Influence Measures in the Minimum C_p Model (* : Influential case)

case	DT_1	Cook	DFFITS	COVRATIO	r_i	h_{ii}
Hawaii	-4.41**	0.33	-1.12*	0.52*	-3.75	0.08
Wyoming	-3.09**	0.34	1.15*	0.51*	3.81	0.08
Alaska	0.37	0.37	1.11*	0.88	2.47	0.17
New York	2.37*	0.00	-0.07	1.31*	-0.14	0.19
South Dakota	2.27*	0.20	0.79*	1.06	1.76	0.17
Other cases Influential			NV			

(1) when the full model is fitted, the DT_1 identifies Hawaii and Wyoming as influential cases, using k=3 (marked by **) in the criterion. For k=2 (marked by *), New York and South Dakota also pass through the cutoff value. When the subset model with minimum C_p is fitted, same observations are detected as the full model.

(2) The Cook's distance identifies Hawaii only, when the full model is fitted. For the minimum C_p model, no case is detected as influential case. The *DFFITS* identified same 5 states including Hawaii, Wyoming, and Alaska in both models. By using the *COVRATIO*, Hawaii, Wyoming, and New York states are detected in both models and 4 more states are identified for the full model. From Table 2.1 and 2.2, we note that Hawaii and Wyoming have large studentized residuals r_i in both models, but the h_{ii}'s are quite different between those two models.

(3) Each measure identified different observations as influential cases. The Cook's distance detects cases conservatively. We think the *COVRATIO* and DT_1, in some degree, detect cases in a similar way.

3. Influence in Variable Selection

A number of measures have been proposed to identify observations with large influence on various aspects of regression analysis, but almost all the measures have been developed in the context of a selected model. Though some measures have been used to assess influences on variable selection(see, Weisberg(1981), Chatterjee and Hadi(1988), and Schall and Dunne(1990)), none of those approaches actually addresses directly the model selection aspect of the problem.

Leger and Altman(1993) proposed an influence measure based on distance between predicted values estimated from the full data set and those computed from the data with the i-th case omitted. For each data set, minimum C_p model is chosen by all possible regression algorithm, and is used to predict response y. For the variable selection problem, their method measures influence on the predicted values, rather than on selected variables. Their statistic is an extended version of Cook's distance for variable selection problem.

3.1 The Proposed Measure for Variable Selection

In linear regression, many model selection techniques and criteria have been introduced in the literature, for example Hocking(1976) and Thompson(1978a,b). In this paper, we restrict our attention on "all possible regression" algorithm with minimum C_p as the criterion for automatic variable selection, and propose a statistic which assesses the influence of each case on variable selection procedure.

For the clear discussion that follows, the subscript " (i)" is used to denote models and estimates bases on $Z_{(i)}$, and to denote vectors or matrices with the i-th row deleted. The superscript "s", and "i" denote the set of selected variables based on the full data set Z and the data set without i-th case, $Z_{(i)}$.

In addition to these notation, the subscript "-i" is to be used to denote estimated responses when parameters are estimated using $Z_{(i)}$ and responses y are predicted for every case except case i. For example, if minimum C_p model is selected using the full data set, but the parameters are estimated using $Z_{(i)}$, then the predicted values for all cases are denoted and expressed by $\hat{y}^s_{(i)} = X^s \hat{\beta}^s_{(i)}$, while $\hat{y}^s_{-i} = X^s_{(i)} \hat{\beta}^s_{(i)}$ denotes predicted values with the i-th case omitted in the stage of estimation as well as prediction.

In the least squares regression problem (1.1), it is well known that the predicted values and the residuals are uncorrelated. That is, from the least squares fit of the model, the following property

$$COV(\hat{y}, e) = 0 \tag{3.1}$$

is satisfied. Meanwhile, if an observation is influential on variable selection procedure, the models selected from the full set of data and the data set without the case (denoted by $Z_{(i)}$) would be different from each other. Thus, the discrepancy between the two selected models could be measured by the magnitude of the covariance between the

predicted values from one model and the residual from the other.

We now define a statistic which measures the influence on variable selection procedure as follows:

$$CS_i = \frac{(y_{(i)} - \hat{y}^s_{-i})'(\hat{y}^i_{-i} - \tilde{y}^i_{-i})}{\sqrt{(y_{(i)} - \hat{y}^s_{-i})'(y_{(i)} - \hat{y}^s_{-i})}\sqrt{(\hat{y}^i_{-i} - \tilde{y}^i_{-i})'(\hat{y}^i_{-i} - \tilde{y}^i_{-i})}} = \frac{(y_{(i)} - \hat{y}^s_{-i})'(y^i_{-i} - \tilde{y}^i_{-i})}{SS(y_{(i)} - \hat{y}^s_{-i})\,SS(\hat{y}^i_{-i} - \tilde{y}^i_{-i})} \tag{3.2}$$

where $\hat{y}^s_{-i} = X^s_{(i)}\,\hat{\beta}^s_{(i)}$, $\hat{y}^i_{-i} = X^i_{(i)}\,\hat{\beta}^i_{(i)}$, and $\tilde{y}^i_{-i}$ is the mean of the components in $\hat{y}^i_{-i}$. That is, $\tilde{y}^i_{-i} = \frac{\sum_{j \neq i} \hat{y}^i_{-i,j}}{n-1}$.

In the defining equation of CS_i, $SS(y_{(i)} - \hat{y}^s_{-i})$ and $SS(\hat{y}^i_{-i} - \tilde{y}^i_{-i})$ are the sum of squares of the components of vector $y_{(i)} - \hat{y}^s_{-i}$ and $\hat{y}^i_{-i} - \tilde{y}^{(i)}_{-i}$, respectively.

Actually, the CS_i measures the correlation between the residuals from the model selected using Z and the fitted values from the model selected using $Z_{(i)}$. Though different data sets are used in variable selection, same reduced data set $Z_{(i)}$ is used to estimate parameters and predict responses. To measure the influence of individual points on variable selection procedure, the discrepancy between two selected models is evaluated in the reduced space of observations.

In practice, the influence measure CS_i is computed as follows:

step 1. Using " all possible regression " algorithm with minimum C_p criterion, develop a model based on the full data set Z, and estimate the parameters of the model using the reduced data set in which the i-th case is omitted. Using the selected model, compute the vector of fitted values $\hat{y}^s_{-i} = X^s_{(i)}\,\hat{\beta}^s_{(i)}$ and the residuals $y_{(i)} - \hat{y}^s_{-i}$ for all the cases except the i-th case.

step 2. Using the same selection method, develop a model based on the reduced data set $Z_{(i)}$ and estimate the parameters of the selected model using $Z_{(i)}$. Using the model and the estimated parameters, compute the vector of fitted values ($\hat{y}^i_{-i} = X^i_{(i)}\,\hat{\beta}^i_{(i)}$) except the i-th case.

step 3. Using the residuals obtained in step 1 and the fitted values in step 2, compute the statistic CS_i which is the correlation between those two values.

We now consider a reference value of the proposed statistics CS_i to identify influential observations on variable selection procedure. Since the CS_i computes the correlation

between $(y_{(i)} - \hat{y}^{s}_{-i})$ and $(y^{i}_{-i} - \tilde{y}^{i}_{-i})$ and the correlation is supposed to be "zero" when the models developed from the full set of data Z and the data set without the case $Z_{(i)}$, the testing peocedure of $H_0 : \rho = 0$ can be used to determine a reference value of the statistics CS_i. We thus determine an individual case z_i as an influential observation on variable selection procedure if

$$|t_i| = \left| \frac{CS_i \sqrt{(n-1)-2}}{\sqrt{1-CS_i^2}} \right| \tag{3.3}$$

is larger than $t_{(n-1)-2}(\alpha)$ for a certain significance level α. Since the component of the two statistics $(y_{(i)} - \hat{y}^{s}_{-i})$ and $(y^{i}_{-i} - \tilde{y}^{i}_{-i})$ are not independent, the t-distribution is not the correct sampling distribution of the statistic CS_i. Therefore, the reference values from (3.3) can be used just as a rough guide to identify influential observations. Based on our experience, we suggest using $\alpha = 0.25$ for our two-sided hypothesis as is used for the Cook's distance ($\alpha = 0.5$) for his one-sided problem.

Leger and Altman (1993) also considered same problem and developed a statistic to measure the influence of individual cases on variable selection procedure. their measure was defined by

$$D_i^u = \frac{(\hat{y}^{s} - \hat{y}^{(i)}_{(i)})' (\hat{y}^{s} - \hat{y}^{(i)}_{(i)})}{q^s MSE^F}$$

$$= \frac{(X^s \hat{\beta}^{s} - X^{(i)} \hat{\beta}^{(i)}_{(i)})' (X^s \hat{\beta}^{s} - X^{(i)} \hat{\beta}^{(i)}_{(i)})}{q^s MSE^F}$$

where q is the number of parameters in the minimum C_p model selected from the full data set and superscript "F" denotes the full model. The statistics D_i^u measure distance between the predicted values computed from the full data set and from the data set without i-th case. The minimum C_p models from the two data sets respectively are used to predict responses y.

The performance of the proposed measure and the unconditional Cook's distance by Leger and Altman will be compared and discussed by analyzing real data sets in the following section.

3.2 Examples

In this section two data sets are analyzed to evaluate the performance of the proposed statistic CS_i, and to compare the results from the statistic CS_i with those from the unconditional Cook's distance D_i^u.

Berkeley Boy's Data:

The first data set is the data set for 26 boys from the Berkeley Guidance Study (Weisberg, 1985). The response variable is somatotype, a measure of fatness based on a seven-point scale, at age 18. The predictor variables are weight and height at ages 2, 9, and 18 and leg circumference and a measure of strength at ages 9 and 18, namely X_1, X_2, $\cdots$,X_{10}. Using the full data set, the predictors X_1, X_7, X_{10} (weight at 2, weight at 18, strength at age 18) are selected from the "all possible regression" procedure with minimum C_p criterion.

Table 3.1: Berkeley Boy's Data
Influence Measures (values in CS_i are multiplied by 10)

Case	ID	Selected variables	CS_i	D_i^y	Case	ID	Selected variables	CS_i	D_i^y
1	201	1,7,10	0.000	0.020	14	215	1,7,10	0.000	0.014
2	202	1,4,7,10	0.780	1.817	15	216	1,6,9	1.872	0.339
3	203	1,4,7,10	0.654	0.048	16	217	1,7,10	0.000	1.194
4	204	1,7,10	0.000	0.002	17	218	1,6,9	1.730	0.019
5	205	1,4,7,10	0.756	0.011	18	219	1,4,5,7,8,10	1.470	1.730
6	206	1,7,10	0.000	0.034	19	221	1,7,10	1.222	2.536
7	207	1,4,7,8,10	1.311	2.974	20	222	1,3,6,9	0.000	0.036
8	209	1,7,10	0.000	1.691	21	223	1,7,10	0.000	0.017
9	210	1,6,9	1.551	0.218	22	224	1,7,10	0.000	0.004
10	211	1,7,10	0.000	0.018	23	225	1,7,10	0.000	0.007
11	212	1,4,7,8,10	1.419	1.262	24	226	1,7,10	0.000	2.923
12	213	1,7,10	0.000	0.001	25	227	1,7,10	0.000	0.002
13	214	1,7,10	0.000	0.007	26	228	1,6,9	1.721	0.025

Table 3.1 contains the values of CS_i and D_i^y computed from the reduced data set in which the i-th subject is deleted from the full data set. In addition to those measures, Table 3.1 contains the variables selected from the selection procedure using the reduced data sets where a single case is deleted one at a time. If the selected variables from the reduced data sets are different from the variables selected from the full data set, and if the discrepancy between those two models are significant, then the deleted case could be identified as an influential observation for variable selection procedure. The computed values of CS_i and TS_i for each individual case are shown in Figure 3.1, with the line of cutoff value $t_{23}(0.25) = 0.685$.

From the Figure 3.1 and Table 3.1, cases 210, 212, 216, 218, 219 and 228 are influential for variable selection according to the proposed statistic CS_i or TS_i . The TS_i values of cases 207 and 222 are slightly less than the cutoff criterion 0.685. The D_i^y proposed by Leger and Altman identified all those 8 data points as influential cases on variable selection procedure.

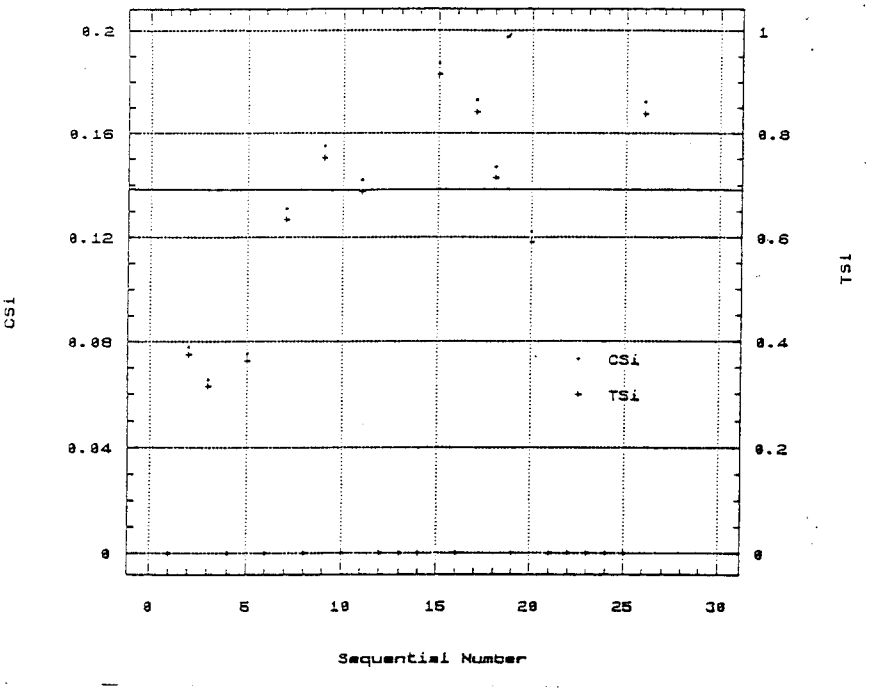

Figure 3.1 CS_i Values for the Berkeley Boy's Data

Note that the selected variables for the individual cases 210, 216, 218, 222, 228 contains X_1, X_3 and/or X_6, X_9. The variables selected for the cases 207, 212, and 219 contains variables X_4 and X_8 in addition to the variables X_1, X_7, X_{10} of the minimum C_p model chosen from the full data set. The correlation between X_4 and X_8 are large, and therefore the problem of multi-collinearity would be involved in these cases.

Fuel Data:

The second data set is the Fuel data which was analyzed in section 2.2. Using "all possible regression" procedure with minimum C_p criterion as the model selector, the model chosen using all 50 observations is Dlic and Inc, namely X_2 and X_3. For all but three states (Hawaii, Wyoming, and Alaska), the selected model from the selection procedure is the same as the model chosen using all 50 cases.

Table 3.2 Summarizes the values of CS_i, D_i^y and the variables selected for the cases Hawaii, Wyoming, and Alaska are deleted respectively from the data set. Also, the TS_i values are given in the table. Since the cutoff value is $t_{(50-1)-2}(0.25) = 0.6794$, only Hawaii is influential according to the CS_i or TS_i. The D_i^y identified Hawaii and Alaska as

influential cases on variable selection. It can be noted that the D_i^y of Wyoming, 0.8756, is slightly less than the cutoff value(1.0) suggested by Leger and Altman. From Table 3.2, we note that the TS_i values of Wyoming and Alaska are less than the cutoffs, 0.6794, but those values are far larger than those of other states. Though CS_i identifies cases somewhat conservatively than the D_i^y does, the CS_i values of those three states in magnitude are distinguishable from those of other states.

Table 3.2: Fuel Data: Influence Measures

ID	State	Variables	CS_i	TS_i	D_i^y
50	Hawaii	X_1, X_2, X_3	0.1216	0.8895	2.4023
40	Wyoming	X_2, X_3, X_4	0.0475	0.3337	0.8756
49	Alaska	X_2, X_3, X_4	0.0438	0.3140	1.0665

4. Summary and Conclusions

In this paper we have introduced an influence measure for goodness-of-fit and a measure of influence for variable selection. The former identifies observations having large influence on overall goodness-of -fit in a selected or tentative model, and thus is a kind of conditional influence measure. Meanwhile, the latter identifies observations having large influence on variable selection procedure. No tentative model is assumed for this measure.

The $DT_1(\cdot)$ measures influence of an individual observation on overall goodness-of-fit directly for the assumed model, and thus almost no measure introduced so far would entertain the same aspect of regression analysis as DT_1 does. As shown in Table 2.1 and 2.2, the cases identified by DT_1 would be somehow different from the cases detected by other measures.

In this study, the statistic $DT_1(\cdot)$ has been investigated whether it can be used for the variable selection problem. We considered such a statistics, say $DT_2(\cdot)$, as

$$DT_2(\cdot) = (n-1)[\ \hat{R}^{2s} - \hat{R}_{(i)}^{2i}]$$

where $\hat{R}^{2s}$ and $\hat{R}_{(i)}^{2i}$ are the estimated R^2 by fitting miminum C_p model selected from the full data set and estimating it using all the data points and by fitting minimum C_p model from the reduced data set and estimating it by reduced data set, respectively. As noted by Leger and Altman(1993, page 547), the problem of multicollinearity may lead to large fluctuation of the regression coefficients, so that many different models may have very similar fit. Therefore, direct use of R^2 in developing an influence measure for model selection may not result in satisfactory results, as we confirmed from the data analysis.

The statistic CS_i has been developed by using the property $COV(e, \hat{y}) = 0$, and the TS_i has been considered to obtain a cutoff criterion by using the testing problem, $H_0 : \rho = 0$. Since the dependence among the components of the vectors which consist of TS_i would not be negligible, the cutoff value ($\alpha = 0.25$) suggested in section 3.1 would be somewhat too conservative to identify influential observations. Ordering of cases based on the magnitudes of TS_i could be a reasonable approach to assess the impact of individual cases.

Computing CS_i and D_i^y requires very intensive computational works. For the data set of size n, all possible regression algorithm should be executed $n + 1$ times. Though modern computing environments make it feasible, it is needed to introduce a statistic which execute all possible regression once, but still can measure the influence of individual cases on variable selection procedure.

References

[1] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

[2] Chatterjee, S., and Hadi, A. S. (1986), "Influential Obsevations, High Leverage Points, and Outliers in Linear Regression(with discussion)," *Statistical Science*, **1**, 379-416.

[3] Chatterjee, S., and Hadi, A. S. (1988), "Impact of Simultaneous Omission of a Variable and an Observation on a Linear Regression Equation," *Computational Statistics and Data Analysis*, **6**, 129-144.

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

[5] Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, **19**, 15-18.

[6] Hample, F. R. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, **69** ,383-393.

[7] Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, **32**, 1-49.

[8] Lee, J. J. (1990), *A Study on Influential Observations in Linear Regression and Time series*, unpublished Ph.D. dissertation, Department of Statistics, University of Wisconsin, Madison.

[9] Leger, C. and Altman, N. (1993), "Assessing Influence in Variable Selection Problems," *Journal of the American Statistical Association*, **88**, 547-556.

[10] Mallows, C. L. (1973), "Some comments on C_p," *Technometrics*, **15**, 661-676.

[11] Schall, R. and Dunne, T. T. (1990), "Influential Variables in Linear Regression," *Technometrics*, **32**, 323-330.

[12] Serfling, J. S. (1980), *Approximation Theorems of Mathematical Statistics*, New York : John Wiley.

[13] Thompson, M. L. (1978a), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review*, **46**, 1-19.

[14] Thompson, M. L. (1978b), "Selection of Variables in Multiple Regression: Part II. Chosen Procedures, Computations, and Examples," *International Statistical Review*, **46**, 129-146.

[15] Weisberg, S. (1981), "A Statistic for Allocating C_p to Individual Cases," *Technometrics*, **23**, 27-31.

[16] Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), New York : John Wiley.

Appendix

Derivation of equation (2.3)

Following Hample (1974), define

$$d_1T(z:F) = \lim_{\lambda \to 0+} \frac{R^2(F_\lambda) - R^2(F)}{\lambda} \tag{A.1}$$

where

$$F_\lambda = (1-\lambda)F + \lambda\delta_z \tag{A.2}$$

and δ_z is the df which puts mass 1 at the point z. It is easy to show that

$$\frac{d}{d\lambda}\sigma_{yy}(F_\lambda)|_{\lambda=0} = (y-\mu_y(F))^2 - \sigma_y^2(F) \tag{A.3}$$

$$\frac{d}{d\lambda}\sigma_\varepsilon^2(F_\lambda)|_{\lambda=0} = \varepsilon^2 - \sigma_\varepsilon^2(F) \tag{A.4}$$

where $\sigma_{yy}(F_\lambda)$ and $\sigma_\varepsilon^2(F_\lambda)$ are variances of Y and ε evaluated at df F_λ defined in (A.2), respectively(Serfling (1980), section 6.2.1). By subtituting (A.3)and (A.4) in the following equation, we have

$$\begin{aligned} dT_1(z:F) &= \frac{d}{d\lambda}R^2(F_\lambda)|_{\lambda=0} \\ &= \left.\frac{[\frac{d}{d\lambda}\sigma_{yy}(F_\lambda)]\sigma_\varepsilon^2(F_\lambda) - [\frac{d}{d\lambda}\sigma_\varepsilon^2(F_\lambda)]\sigma_{yy}(F_\lambda)}{\sigma_{yy}^2(F_\lambda)}\right|_{\lambda=0} \\ &= \frac{(1-R^2(F))(y-\mu_y(F))^2 - \varepsilon^2}{\sigma_{yy}(F)} \end{aligned} \tag{A.5}$$

선형회귀모형에서 변수 선택에 영향을 미치는 관측점에 관한 연구[1)]

최지훈[2)], 구자홍[2)], 이재준[2)], 전홍석[2)]

요 약

회귀분석에서 몇개의 관측치가 모형선택과정이나 최소제곱방법에의한 모형의 적합에서 지대한 영향을 끼칠 수 있다. 그러나 그러한 관측치가 회귀분석의 모든 면에 같은 정도의 영향을 끼치는 것은 아니다. 본 논문에서는 개개의 관측치가 적합성의 총체적 측도에 미치는 영향을 측정할 수 있는 통계량을 소개하였다. 또한, 개개의 관측치가 변수선택 과정에 미치는 영향을 측정할 수 있는 영향측도를 제시하였다. 이 측도는 자료를 모형에 적합시켜 구해진 잔차와 적합치 사이에 만족되는 비상관성의 성질을 이용하여 구해진 것이다. 마지막으로, 본 논문에서 소개된 통계량들과 가장 많이 이용되고 있는 영향측도들을 실제자료의 분석을 통하여 그 성질과 효용성을 비교하였다.

1) 이 논문은 1991년도 교육부지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음
2) (402-751) 인천직학시 남구 용현동 253, 인하대학교 이과대학 통계학과