Journal of the Korean
Statistical Society
Vol. 23, No. 2, 1994

Variable Selection Criteria in Regression[†]

Choongrak Kim[1]

ABSTRACT

In this paper we propose a variable selection criterion minimizing influence curve in regression, and compare it with other criteria such as C_p (Mallows 1973) and adjusted coefficient of determination. Examples and extension to the generalized linear models are given.

KEYWORDS: Deviance, Influence curve, Mallows C_p.

1. INTRODUCTION

Suppose we wish to establish a linear regression model for a response Y in terms of predictor variables $X_1, \ldots, X_k$ which are all possible candidate variables. We call a regression model based on all candidate variables as the full model. For parsimonious modelling, we choose only part of k variables, and call this model as the current model defined as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} X_{p-1i} + \epsilon_i, \; i = 1, \ldots, n \qquad (1.1)$$

[†]This research was supported by the Ministry of Education (BSRI-93-116)
[1]Department of Statistics, Pusan National University, Pusan, 609-735, Korea.

where ϵ_i's are iid with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. To check the adequacy of the current model, many criteria were suggested. Among them, Mallows'C_p

(Mallows 1973) defined as

$$C_p = \frac{RSS_p}{s^2} - (n - 2p) \tag{1.2}$$

is most widely used. Here, RSS_p is the residual sum of squares under the current model and s^2 is the residual mean square from the full model. C_p is a criterion minimizing the mean squared errors for the current model, and closely related with the adjusted coefficient of determination

$$R^2_{ap} = 1 - (n-1)(1 - R^2_p)/(n-p)$$

where R^2_p is the coefficient of determination.

Note that C_p is very sensitive to outliers since $RSS_p = \sum e_i^2$ where $e_i = y_i - \hat{y}_i$ and $\hat{y}_i$ is the fitted value of y_i under (1.1). Undesirable features of C_p are discussed in detail by Miller (1990). To be more specific, let the jth observation have large residual but not influential to $\hat{\beta}$. This situation occurs when the jth observation is located around the middle of X-space. Then, the leverage of x_j is quite small, and the influence of jth observation on $\hat{\beta}$ is negligible even though e_j is large. Therefore, an adequate model with one or few outliers which are not influential may be regarded as inappropriate if C_p is used. This situation will be numerically explained in Section 3.2 throgh an artificial data set. In this paper, we propose a variable selection criterion K_p, defined in Section 2, which is aimed to minimize the overall influence. In Section 2, K_p is defined and justified. Also, reference value is given. Examples are given in Section 3, and extension of K_p to generalized linear models with an example is in Section 4.

2. A CRITERION MINIMIZING INFLUENCE

Consider a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the infinitesimal perturbation approach is obtained by specifying $\epsilon_j \sim N(0, \sigma^2/w_j)$, where

$$w_j = \begin{cases} w, & \text{if } j = i \\ 1, & \text{otherwise} \end{cases}$$

and $0 \leq w \leq 1$. Under the specification, the weighted least squares estimator is $\hat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ with $\mathbf{W} = diag\{w_i\}$. Pregibon (1981) showed that

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_w = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i(1-w)e_i}{\{1-(1-w)h_{ii}\}}$$

where h_{ii} is the ith diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The effect of infinitesimal perturbations of the variance of the ith data point is obtained by differntiation of $\hat{\boldsymbol{\beta}}_w$, i.e.,

$$\Delta\hat{\boldsymbol{\beta}}_w = \frac{\partial}{\partial w}\hat{\boldsymbol{\beta}}_w = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{\{1-(1-w)h_{ii}\}^2}.$$

Evaluation at $w = 1$, $\Delta\hat{\boldsymbol{\beta}}_1$, describes local changes in $\hat{\boldsymbol{\beta}}_w$ at the usual least squares solution. This function is termed the influence curve of $\hat{\boldsymbol{\beta}}$ in the literature of robust and resistant estimation. One of scalar versions of $\Delta\hat{\boldsymbol{\beta}}_1$ can be

$$\begin{aligned} C_i &\equiv \Delta\hat{\boldsymbol{\beta}}_1'\mathbf{X}'\mathbf{X}\Delta\hat{\boldsymbol{\beta}}_1/p\sigma^2 \\ &= e_i^2 h_{ii}/p\sigma^2. \end{aligned}$$

The reason for choosing $w = 1$ is geometrically justified by Cook(1986). If the assumed model is good it will be robust to the infinitesimal perturbation of each data point, i.e., $\sum C_i$ will be small. Therefore, we suggest to use $\sum_{i=1}^n C_i$ as dimensionality selection in regression. Since σ^2 is usually unknown we replace it by s^2, residual mean squares from the full model. To be useful $\sum C_i$ can be adjusted as

$$K_p = \frac{n}{p}\frac{\sum e_i^2 h_{ii}}{s^2} - (n-2p)$$

in view of C_p. Note that C_p is compared with p since $E(C_p) \simeq p$ if an equation with $p-1$ parameters is adequate, i.e., we choose a regression equation with $C_p \leq p$. Reference value for K_p is similarly obtained. By letting $h_{ii} \simeq p/n$, average of leverages, we have $E(\sum e_i^2 h_{ii}) \simeq p(n-p)\sigma^2/n$ if the model is adequate. Therefore, we choose a regression equation with $K_p \leq p$. Note that K_p is a weighted version of C_p with weight h_{ii}.

3. EXAMPLES

3.1 Hald Data.

For the Hald data (see Draper and Smith (1981) for details) we have $n = 13$ and $s^2 = 5.983$ from the full model fitted to all four predictor variables. C_p and K_p are listed for $p = 2, 3, 4$ in Table 1. We see that the fitted model with X_1 and X_2 is preferred over all others by both C_p and K_p.

Table 1. C_p and K_p in Hald Data

p	Variables	C_p	Variables	K_p
	4	138.7	2	112.7
2	2	142.5	4	116.0
	1	202.5	1	161.8
	3	315.2	3	274.3
	1,2	2.7	1,2	1.0
3	1,4	5.5	1,4	3.4
	3,4	22.4	3,4	20.3
	2,3	62.4	2,3	58.9
	1,2,4	3.0	1,2,4	1.0
4	1,2,3	3.0	1,2,3	1.3
	1,3,4	3.5	1,3,4	1.5
	2,3,4	7.3	2,3,4	4.9

3.2 Artificial Data.

For $X_1 = 1(1)20$, we generate 20 ϵ_i's from $N(0, .25)$, and set $y_i = i + \epsilon_i$, $i = 1, \ldots, 20$. Also, we create two irrelevant variables X_2 and X_3 from $Unif(0, 1)$ and $B(20, .5)$, respectively. X_1, X_2, X_3 and Y are listed in Table 2. Here, we have $n = 20$, and $s = .19$ from the full model. We apply C_p and K_p to this artificial data set, and see whether X_1 is successfully chosen. As shown in Table 3, K_p is quite successful, however, C_p is not. C_p for X_1 is 8.6, and one will hesitate to choose X_1 if he or she relies on C_p. Instead, he or she will choose X_1 and X_3. Our question is that why K_p choose X_1 successfully and C_p do not. One possible answer is given in Figure 1 containing plot of $e_i = y_i - \hat{y}_i$ and h_{ii} from $Y = \beta_0 + \beta_1 X_1 + \epsilon$. Residuals with low leverages are quite large and residuals with high leverages are quite small. If the relation between residuals and leverages is the opposite pattern, the result will be converse. Note that a small change of x-values away from the middle (i.e., x-values with high leverages) will change $\hat{\beta}$ significantly. This is why most influence measures are increasing functions of both the residual and leverage. In fact, the effect of leverage is inserted to K_p but not C_p.

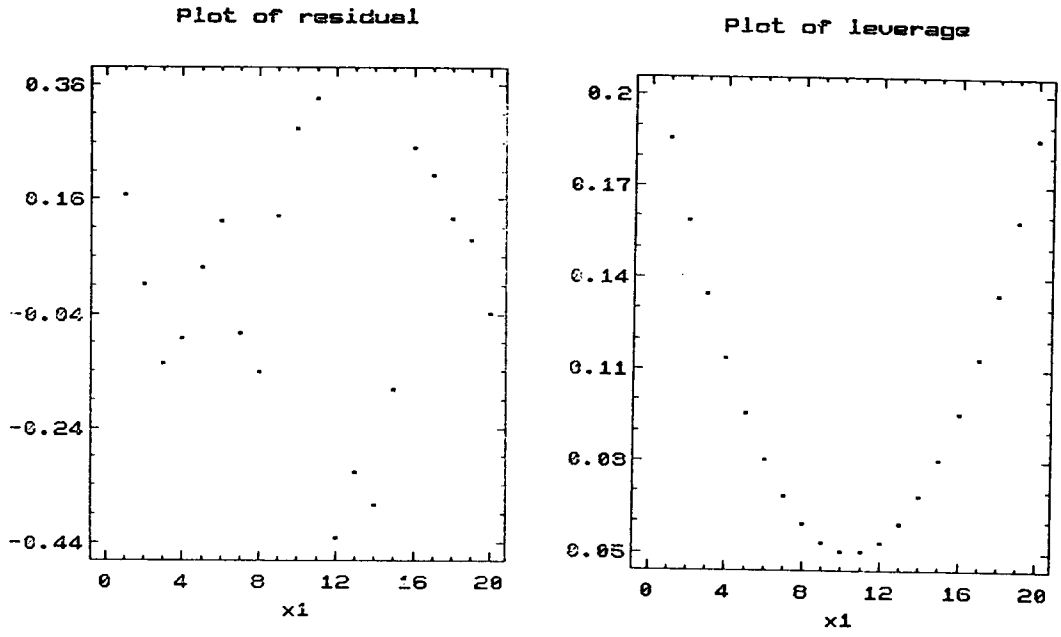
Table 2. Artificial Data

X_1	X_2	X_3	Y
1	0.35	11	1.15
2	0.01	7	2.00
3	0.73	8	2.87
4	0.21	11	3.92
5	0.75	13	5.05
6	0.13	12	6.14
7	0.61	9	6.95
8	0.90	8	7.89
9	0.93	9	9.17
10	0.93	13	10.33
11	0.43	14	11.39
12	0.25	8	11.63
13	0.66	6	12.75
14	0.57	9	13.70
15	0.81	11	14.91
16	0.20	11	16.34
17	0.17	9	17.30
18	0.69	10	18.23
19	0.13	12	19.20
20	0.78	15	20.08

Table 3. R^2, R^2_{ad}, C_p, and K_p in Artificial Data

Variables	R^2	R^2_{ad}	C_p	K_p
X_1	99.9	99.9	8.5	1.64
X_3	6.0	0.8	20000	16421.76
X_2	0.5	0.0	20000	19005.16
X_1, X_3	99.9	99.9	2.6	1.91
X_1, X_2	99.9	99.9	10.1	3.23
X_2, X_3	6.3	0.0	20000	16213.20
X_1, X_2, X_3	99.9	99.9	4.0	2.47

Figure 1. Residuals and leverages in Artificial Data



4. EXTENSION TO GLMS

4.1 Deviance, Pearson χ^2, and K_p.

In the generalized linear models (McCullagh and Nelder 1989), deviance and Pearson χ^2 are often used as goodness-of-fit measure. Suppose that $y_1, \ldots, y_n$ are independent observations having a density in the exponential family with the form

$$f(y_i; \theta_i, \phi) = \exp\Big\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i, \phi)\Big\}$$

for some functions $a_i(\cdot), b(\cdot)$, and $c(\cdot)$. Let $g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, where $\mu_i = E(Y_i)$ and g is a link function. Deviance is defined as

$$D = 2\sum\Big\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\Big\}\Big/a_i(\phi)$$

where $\tilde{\theta}_i$ and $\hat{\theta}_i$ are estimates of θ_i under the maximal model and the current model, respectively. The other measure of discrepancy is the Pearson χ^2 statistic, which takes the form

$$\chi^2 = \sum(y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i)$$

where $V(\hat{\mu}_i)$ is the estimated variance function for the distribution concerned. Both the deviance and the Pearson χ^2 are compared with χ^2_p distribution.

Based on the same arguments as in the linear model (see Section 2), we can easily define K_p in the generalized linear models as

$$K_p = n\sum r_i^2 h_{ii}^*/p$$

where $r_i = (y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$ and h_{ii}^* is the i-th diagonal element of $\mathbf{H}^* = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$. Therefore, K_p is a weighted version of the Pearson χ^2 with weight h_{ii}^*. We will use χ^2_p as a reference distribution to K_p.

4.2 Tuberculin response Data.

Fisher (1949) published some data consisting of 16 measurements of tuberculin response to four treatments. These were applied in a Latin square design so that the effects were not confounded with type of cow and site. Following the structure of design and notations of Baker and Nelder (1978, Appendix

D.3), we use the log-linear model and for simplicity let X_1 = A, X_2 = B, X_3 = cow, X_4 = site. We compute D, χ^2, and K_p with the refrence values $\chi^2_{df}(\alpha)$ in Table 4, and see that all the three measures give the same result ; X_2, X_3, and X_4.

Table 4. Goodness of fit measures D, χ^2, K_p

model	d.f.	deviance (D)	Pearson (χ^2)	K_p	$\chi^2_{df}(.05)$
null (1)	15	265.30	278.73	278.73	25.00
X_1	14	265.28	278.69	278.69	23.68
X_2	14	203.09	210.85	210.84	23.68
X_1, X_2	13	203.07	210.79	212.72	22.36
X_4	12	232.38	238.09	228.60	21.03
X_3	12	91.76	92.43	98.51	21.03
X_1, X_4	11	232.37	238.14	238.66	19.68
X_1, X_3	11	91.74	92.39	98.16	19.68
X_2, X_4	11	170.17	177.60	174.73	19.68
X_2, X_3	11	29.55	29.47	30.78	19.68
X_1, X_2, X_4	10	170.15	177.45	175.18	18.31
X_1, X_2, X_3	10	29.53	29.48	30.68	18.31
X_3, X_4	9	58.84	58.78	55.27	16.92
X_1, X_3, X_4	8	58.82	58.74	56.01	15.51
X_2, X_3, X_4	8	1.41	1.42	1.35	15.51
X_1, X_2, X_3, X_4	7	1.40	1.41	1.30	14.07

5. CONCLUDING REMARKS AND FUTURE RESEARCH

Many statistical packages use C_p as a variable selection criterion. As Miller (1990) indicated, C_p has undesirable features. In this paper, we suggest a variable selection criterion K_p which is aimed to minimize sum of influence of each observation. Based on limited experiences, K_p works well and is better than C_p in some circumstances.

As noted by two referees, systematic comparison of K_p with C_p should be done. In this study, a careful choice of appropriate models is necessary. Also,

consideration of the weighted version of C_p is desirable. In the GLMs, the robust version of the AIC (Akaike Information Criterion) can be a good substitute for K_p. These problems deserve future research. Also, the performance of K_p in the sense of model parsimony should be studied because C_p tends to select unnecessarily large model.

REFERENCES

(1) Baker, R.J. and Nelder, J.A. (1978). *The GLIM System* (Release 3) ; *Generalized Linear Interactive Modelling*, Numerical Algorithms Group, Oxford.

(2) Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*, Ser. **B, 48, 133–169.**

(3) Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*, (2nd ed.) New York : Wiley.

(4) Fisher, R.A. (1949). A biblogical assay of tuberculins. *Biometrics*, **5, 300–316.**

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

(6) McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, (2nd ed.) New York : Chapman and Hall.

(7) Miller, A.J. (1990). *Subset Selection in Regression*. London : Chapman and Hall.

(8) Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9, 705–724.**