The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 267-272

A Role of Local Influence in Selecting Regressors

Myung Geun Kim[1)]

Abstract

A procedure for selecting regressors in the linear regression model is suggested using local influence approach. Under an appropriate perturbation scheme, the effect of perturbation of regressors on the profile log-likelihood displacement is assessed for variable selection. A numerical example is provided for illustration.

Keywords : F-test; Likelihood displacement; Local influence; Regression.

1. Introduction

Variable selection or subset selection problem has been studied in various statistical models, especially in the linear regression model. Numerous selection procedures based on diverse selection criteria have been suggested. Miller (2001) gave a comprehensive treatment of variable selection methods in regression. From Chapter 3 of Miller's book we can see that most selection methods do not guarantee a finding of the best-fitting subsets, or that some methods need enormous computation even when the number of regressors becomes a little large. Also, many selection methods do not consider all of regressors simultaneously at each selection stage. and this may be one reason why most selection methods do not guarantee a finding of the best-fitting subsets.

In this work a procedure for selecting a regressor at each step is suggested. The procedure starts with the location model. At each step, we find the direction vector corresponding to the largest curvature of the profile log-likelihood displacement surface in view of Cook (1986) and select a regressor associated with the element of the above direction vector that has the largest absolute value.

If an F-test of the hypothesis that the regression coefficients for regressors not selected so far are zero is not significant, then the selection procedure stops, and otherwise, we proceed further. This selection method considers all of regressors simultaneously at each step and it needs far less computation compared with the existing selection methods.

1) Professor, Department of Applied Statistics, Seowon University, Cheongju 361-742, Korea.
E-mail : mgkim@seowon.ac.kr

In Section 2 we find a formula for selecting regressors using the local influence approach under an appropriate perturbation scheme. In Section 3 a procedure for selecting regressors is suggested. An illustrative example is provided in Section 4.

2. Profile Log-likelihood Displacement

Consider the linear regression model

$$y = X\beta + \epsilon$$

where y is an $n\times 1$ vector of response variables, X is an $n\times p$ matrix of fixed regressors, $\beta=(\beta_0,\beta_1,\dots,\beta_{p-1})^T$ is a $p\times 1$ vector of unknown regression coefficients that includes an intercept term as its first element, and ϵ is an $n\times 1$ vector of random errors that are independent and identically distributed as a normal distribution with mean 0 and variance σ^2. The least squares estimator of β is $\hat\beta=(X^TX)^{-1}X^Ty$ and an unbiased estimator of σ^2 is $\hat\sigma^2=(y-X\hat\beta)^T(y-X\hat\beta)/(n-p)$. The regression sum of squares for the full model is $SSR_p=\hat\beta^TX^TX\hat\beta$.

We will consider a situation in which for convenience the first k (k may be zero) regressors together with the intercept term (X_1 in (1)) have been already included in the regression model and further regressors among the remaining regressors (X_2 in (1)) will be selected. To this end, we partition

$$\beta=\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix},\quad \hat\beta=\begin{bmatrix}\hat\beta_1\\ \hat\beta_2\end{bmatrix},\quad X=[X_1\ \ X_2],$$

where β_1 is a $(k+1)\times 1$ vector of the first k+1 elements of β and X is partitioned according to the partition of β. Then we can rewrite the regression model as

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{1}$$

Let $L(\beta_1,\beta_2,\sigma^2)$ be the log-likelihood function for the model parameters and $\hat\sigma^2_M=(n-p)\hat\sigma^2/n$. The maximized log-likelihood becomes $L(\hat\beta_1,\hat\beta_2,\hat\sigma^2_M)=-(n/2)[\log(\hat\sigma^2_M)+1]$. For each fixed β_2 and σ^2, $L(\beta_1,\beta_2,\sigma^2)$ is maximized when β_1 is a function of β_2 given by $g(\beta_2)=(X_1^TX_1)^{-1}X_1^T(y-X_2\beta_2)$. Hence the profile log-likelihood function for β_2 and σ^2 becomes $L(g(\beta_2),\beta_2,\sigma^2)$.

A perturbation scheme is characterized by a point $w=(w_1,\dots,w_q)^T$ imposed on X_2, where $q=p-k-1$. We can express the perturbation vector as $w=1_q+ad$, where 1_q is a $q\times 1$ vector with all elements equal to one, a scalar a is the magnitude of the perturbation w along the direction d of unit length. Let W be a diagonal matrix of order q with $w_1,\ldots,w_q$ as its diagonal elements. We consider a perturbed model

$$y = X_1\beta_1 + X_2 W\beta_2 + \epsilon. \tag{2}$$

When all the w_i are set equal to one, that is $a=0$, the perturbed model (2) reduces to the unperturbed model (1). Under this perturbed model, the maximum likelihood estimators of β_2 and σ^2 are given by

$$\hat{\beta}_2(w) = W^{-1}\hat{\beta}_2 \quad \text{and} \quad \hat{\sigma}^2_M(w) = \hat{\sigma}^2_M.$$

As suggested by Cook (1986, p. 137), we can assess the effect of perturbation of regressors X_2 not yet selected on the profile log-likelihood displacement

$$LD(a) = 2[L(\hat{\beta}_1,\hat{\beta}_2,\hat{\sigma}^2_M) - L(g(\hat{\beta}_2(w)),\hat{\beta}_2(w),\hat{\sigma}^2_M(w))].$$

As a varies over R^1, the vector $(a, LD(a))$ forms a curve in the plane. The direction vector $d_{\max}$ associated with the largest curvature of the curve at $a=0$ provides information about regressors that cause a great change in the profile log-likelihood displacement. A regressor in X_2 corresponding to the largest absolute element of $d_{\max}$ is most influential and therefore it will be included in the model.

Next, we will find a Taylor series expansion of $LD(a)$ about $a=0$ to get curvatures and direction vectors. A Taylor series expansion of W^{-1} about $a=0$ is

$$W^{-1} = 1_q - aD(d) + a^2D^2(d) + o(a^2),$$

where $D(d)$ is a diagonal matrix whose diagonal elements are the elements of d. Then we have

$$y - X_2W^{-1}\hat{\beta}_2 = y - X_2\hat{\beta}_2 + aX_2D(d)\hat{\beta}_2 - a^2X_2D^2(d)\hat{\beta}_2 + o(a^2).$$

Since $\hat{\beta}_2 = [X_2^T(I-H_1)X_2]^{-1}X_2^T(I-H_1)y$ where $H_1 = X_1(X_1^TX_1)^{-1}X_1^T$, it is obvious that $X_2^T(I-H_1)(y-X_2\hat{\beta}_2)=0$. Hence a Taylor series expansion of $L(g(\hat{\beta}_2(w)),\hat{\beta}_2(w),\hat{\sigma}^2_M(w))$ about $a=0$ is easily given by

$$L(g(\hat{\beta}_2(w)),\hat{\beta}_2(w),\hat{\sigma}^2_M(w)) = -\frac{n}{2}[\log(\hat{\sigma}^2_M)+1] - \frac{a^2}{2}d^TT_qd + o(a^2),$$

where T_q is the $q\times q$ matrix defined by

$$T_q = \frac{n}{n-p}\frac{D(\hat{\beta}_2)[X_2^T(I-H_1)X_2]D(\hat{\beta}_2)}{\hat{\sigma}^2}.$$

Finally we get

$$LD(a) = a^2d^TT_qd + o(a^2).$$

Thus the curvatures and direction vectors are just the eigenvalues and eigenvectors of $2T_q$, respectively.

3. A Selection Procedure

A test of the hypothesis $H_0 : \beta_2 = 0$ can be performed by using an F-statistic given by

$$F_q = \frac{\hat{\beta}_2^T [\boldsymbol{X}_2^T (\boldsymbol{I} - \boldsymbol{H}_1) \boldsymbol{X}_2] \hat{\beta}_2}{q\hat{\sigma}^2}.$$

When H_0 is true, F_q is distributed as an F-distribution with q and $n-p$ degrees of freedom.

All regressions in what follows are assumed to include the intercept term. The regression sum of squares for the reduced model defined by $\boldsymbol{X}_1$ only is denoted by SSR_k. It is understood that SSR_0 corresponds to the location model. The loss in the regression sum of squares due to the use of this reduced model is computed as

$$SSR_p - SSR_k = q\hat{\sigma}^2 F_q.$$

This equation means that when the model is defined by $\boldsymbol{X}_1$ only, the loss in the regression sum of squares is absorbed entirely into the test statistic of the hypothesis $H_0 : \beta_2 = 0$. If this test is not significant, further inclusion of some of the remaining regressors does not improve the regression fit so much. Otherwise, we need to find regressors for further inclusion.

Before we state a procedure for electing regressors, we note that the curvature associated with the direction vector $\boldsymbol{d}_0 = 1_q / \sqrt{q}$ is computed as $2\boldsymbol{d}_0^T \boldsymbol{T}_q \boldsymbol{d}_0 = [2n/(n-p)]F_q$. In this case the regressors in $\boldsymbol{X}_2$ are equally important for further selection.

A procedure for selecting one regressor at each step is suggested as follows.

Step 1. Start with F_{p-1} and $2\boldsymbol{T}_{p-1}$ associated with the location model.

Step 2. Assume that we have previously selected $k (k = 0, \ldots, p-1)$ regressors. Stop if the value of the statistic $F_q (q = p-k-1)$ is not significant. Otherwise, find the eigenvector $\boldsymbol{d}_{\max}$ associated with the largest eigenvalue of $2\boldsymbol{T}_q$ and then select the regressor corresponding to the largest absolute element of $\boldsymbol{d}_{\max}$ for further inclusion.

Step 3. Repeat Step 2 until the hypothesis H_0 is not rejected.

At each selection step we may select, if any, more than one regressors corresponding to relatively large absolute elements of $\boldsymbol{d}_{\max}$ compared with the other elements, which needs further investigation for practical use.

4. A Numerical Example

We will use the Hald cement data to illustrate the selection procedure in Section 3. The Hald cement data set (Draper and Smith, 1981) consists of 13 observations with four regressors. With these data Montgomery and Peck (1982) illustrated various selection methods. Recently, Ronchetti and Staudte (1994) and Sommer and Huggins (1996) analyzed these data. Their general conclusion is that the first two regressors are enough for the best fitting.

Some computational results for this data set are included in <Table 1>. At each selection step, <Table 1> includes the largest eigenvalue and the corresponding eigenvector (d_1 to d_4) of $2\boldsymbol{T}_q$, and a selected variable number (SV).

At step 1, the value of F-statistic for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is 111.48 and its p-value is 0. Hence the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is rejected at any significance level and then we compute $\boldsymbol{T}_q$ with $\boldsymbol{X}_2$ including all four regressors. The maximum eigenvalue of $2\boldsymbol{T}_q$ is 619.4 and its associated eigenvector has components d_1 to d_4. The largest absolute component is 0.834 (d_1) and thus the first regressor is selected to be included in the model.

At step 2, we perform a test of $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, and the corresponding F-statistic has 67.85 and its p-value is 0. Thus we reject $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ at any significance level and compute $\boldsymbol{T}_q$ with $\boldsymbol{X}_2$ including the last three regressors. The maximum eigenvalue of $2\boldsymbol{T}_q$ is 423.2 and its associated eigenvector has components d_2 to d_4. The largest absolute component is 0.959 (d_2) and thus the second regressor is selected to be included in the model.

At the final step, the value of the F-statistic for testing the hypothesis that the last two regression coefficients are zero is 0.84 and its p-value is 0.47. Thus we stop the selection process at reasonable significance levels and the first two regressors are selected to be included in the regression model.

<Table 1> Computational results for the Hald cement data

k	eigenvalue	d_1	d_2	d_3	d_4	SV
1	619.4	0.834	0.526	-0.047	0.160	1
2	423.2		0.959	0.005	0.283	2

References

[1] Cook, R.D. (1986). Assessment of local influence (with discussions). *Journal of*

the Royal Statistical Society, Series B, Vol. 48, 133-169.

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

[3] Miller, A.J. (2001). *Subset Selection in Regression*, 2nd ed., Chapman & Hall, New York.

[4] Montgomery, D.C. and Peck, E.A. (1982). Introduction to Linear Regression Analysis, Wiley, New York.

[5] Ronchetti, E. and Staudte, R.G. (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, Vol. 89, 550-559.

[6] Sommer, S. and Huggins, R.M. (1996). Variable selection using Wald test and a robust C_p. *Applied Statistics*, Vol. 45, 15-29.

[Received March 2006, Accepted May 2006]