On Some collinearities with Some Observations in Linear Regression

- 선형회귀모형에서 다공선성을 은폐 혹은 확대하는 관찰치에 관한 식별 -

Kim, Seung Gu *

요 약

선형회귀모형에서 새로운 변수가 모형에 도입될때 몇몇 비정상적인 관찰치들은 변수들 간에 내재되어 있는 다공선성을 감추거나 혹은 오히려 더욱 크게 부풀림으로써 도입변수에 대한 해석을 매우 어렵게 만든다. 본고에서는 이러한 관찰치들을 식별할 수 있는 방법을 제안하였는데, 이와같은 식별법은 postulated model의 회귀계수추정치에 대한 도입변수의 섭동(perturbations)을 분해함으로써 가능하였다.

1. Introducton

The variable selection procedure is very important to regression analysis to find an optimized model. The main objective of these procedures is to find the parsimonious model which optimizes the fitness of data, and it is secondary to investigate the changes of the regression coefficients in the candidate model by entering variables. In many case, however, of social science the several explanatory variables are presented to explain a variation of the social phenomenon, the researchers want to examine the sensitivity of the estimates of regression coefficients caused by new explanatory varibles. This paper treats of the problems for the perturbations of coefficients estimates caused by several variables rather than the ploblems for the selection the *best* model. It's seemed that there are rarely studies in this ploblem except for Schall and Dunne(1990)'s works. In fact, this study starts with their main results.

It is well-known that once new variables enter into the given model, then several cases often intervene the variable selection procedures. There are many studies in this area to overcome the outliers' intervention; Belseley,et.al(1980) which might be the first introduces to the ploblems of the *simultaneous interpretations* about variables and cases, and Léger and Altman(1993) and Choi,et.al(1993) which introduce the measures to find the best model avoiding these outliers. Similarly, we can consider to the ploblems about the perturbation of the regression coefficients estimates in the given model, will be called *the postulated model.* It is possible to consider that when the coefficients estimates are highly perturbed by new varibles entered into the postulated model, this perturbation might almost be due to some observations although the influence on the postulated model by entered variables would be small without these observations. And conversely, these observations are able to make the perturbation of coefficients estimates be small although the influence, in fact, would be large without them.

---

* Full time lecturer, Dept. of Statistics, Sangji Univ.

접수 : 1994년 4월 20일
확정 : 1994년 5월 4일

This paper consider to the influence of new variables on the regression coefficients intervened by some observations. Consequently, this approach is useful tools to detect some observations which hide or induce the collinearity when the new variables enter into the postulated model; Section 2 introduces the definitions and basic concepts including Schall and Dunne's results. And in section 3, new statistic typed Cook's ditance is proposed and it is decomposed to several components which identify the structure of the proposed statistic. In section 4, two examples with two types of collinearity are presented.

2. Definitions and Basic Conceps

Consider the full-ranked linear model

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

,where $\boldsymbol{y}$ is an n vector of the dependents, $\boldsymbol{X}$ is a (n×p) matrix of predictors, $\boldsymbol{\beta}$ is a p vector unknown regression coefficients and vector $\boldsymbol{\varepsilon}$ has a covariance matrix $Cov(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. And we assume the model has a constant term. Model (2.1) has LS coefficients estimates $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y$ and the residuals sum of squares $SS_E = (n-p)\hat{\sigma}^2 = y'N_1y$,where $N_1 = I_n - X(X'X)^{-1}X'$. When an arbtrary variable $\boldsymbol{A}$ enter into the model (2.1), we call it the postulated model in the sense that those extra effects of variable $\boldsymbol{A}$ are a perturbation of the postulated model by analogy with the terminology of Cook(1986). Note that we assume the postulated model does not contain any observations which highly perturb it. The model included variable $\boldsymbol{A}$ is given by

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\,\boldsymbol{\mu} + \boldsymbol{\varepsilon} \\ &= [\boldsymbol{X}:\boldsymbol{A}]\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\lambda}\end{bmatrix} + \boldsymbol{\varepsilon} \end{aligned} \tag{2.2}$$

,where the augmented variable $Z = [X: A]$ and $\mu' = [\beta': \lambda']$ and A is (n×k) matrix. The model (2.2) also has the LS coefficients estimates $\tilde{\mu} = (Z'Z)^{-1}Z'y$, which is able to be expressed as

$$\tilde{\mu} = \begin{bmatrix}\tilde{\beta}\\ \tilde{\lambda}\end{bmatrix} = \begin{matrix}(X'N_2\,X)^{-1}X'N_2y \\ \\ (A'N_1\,A)^{-1}A'N_1y\end{matrix}$$

,where $N_2 = I_n - A(A'A)^{-1}A'$. Under the model (2.2), Schall and Dunne present a Cook's distant typted-diagnostic measure to assess an influence on $\hat{\beta}$ by the variable A, such as

$$SD_A = (\hat{\beta} - \tilde{\beta})'X'X\,(\hat{\beta} - \tilde{\beta})/p\hat{\sigma}^2 \tag{2.3}$$

If $\hat{\beta}$ will be highly perturbed by A , and so changed to $\tilde{\beta}$ be large in magnitude, then SD_A will be large,which implies that the variable A must be the influence variable on model (2.1). It is possible, similarly, to explain it in small scaled changes between $\hat{\beta}$ and $\tilde{\beta}$. The quantity (2.3) are presented as

a decomposed form with two major terms apart from a constant factor (n-p)/p ,

$$SD_A = [ESS/SS_E][D/ESS - 1][(n-p)/p] \qquad (2.4)$$

,where $ESS = \boldsymbol{\lambda}'\boldsymbol{A}'\boldsymbol{N}_1\boldsymbol{A}\boldsymbol{\lambda}$ is the extra sum of squares due to $\boldsymbol{A}$ given $\boldsymbol{X}$ and $D = \boldsymbol{\lambda}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{\lambda}$ is associated with the variance inflation in variable $[\boldsymbol{X}: \boldsymbol{A}]$.

The first factor factor of the right hand side of (2.4) is known as a monotonic increasing function of the F-statistic under the null hypothesis $\boldsymbol{\lambda} = \boldsymbol{0}$. And the second factor is a monotonic function of the variance inflation factor associated with the variable $\boldsymbol{A}$ in the model (2.2). Also, for k=1, Schall and Dunne introduce the quantity $h_{(A)} = 1 - \boldsymbol{A}'\boldsymbol{N}_1\boldsymbol{A}/\boldsymbol{A}'\boldsymbol{A}$ as the leverage of the variable, expanding the idea of the leverage an observation introduced first by Hoaglin and Welsch(1978). The quantity $h_{(A)}$ is the potential influence of the variable $\boldsymbol{A}$ because it does contain the variable $\boldsymbol{A}$ and $\boldsymbol{X}$ but not $\boldsymbol{y}$. Note that the second factor of (2.4) is a monotonic increasing function of $h_{(A)}$ and if $\boldsymbol{A}$ is an high leveraged variable, the power of F-test decrease rapidly and then the ESS tends to be small,which implies that the high leveraged variable $\boldsymbol{A}$ makes the F-test be useless.

It is possible for some observations to be an outlying cases on the space $[\boldsymbol{y}: \boldsymbol{X}: \boldsymbol{A}]$ but not on $[\boldsymbol{y}: \boldsymbol{X}]$. Here, we assume that there are no outlying observations such that they highly fluctuate $\hat{\boldsymbol{\beta}}$ in model (2.1). So that, we call this assumption to *the postulated assumption*. Consider a possibilty of which some observations cause the variable $\boldsymbol{A}$ to the value of SD_A be large however these observations might never be indicated to be highly influential in postulated model. Indeed, the magnitude of SD_A senstively depends on such an observations since the first factor of (2.4) is a function of the LS estimates $\hat{\boldsymbol{\beta}}$ which is very sensitive to an extreme outliers and the second factor also affected by the high leveraged points on $[\boldsymbol{X}: \boldsymbol{A}]$. It is well known that small changes of data induce large perturbation of the regression coefficients when the degree of multicollinearity is high. Consequently, existence of such an observations inhibit interpreting SD_A.

3. Infuential Variables caused by Observations

It is natural to investigate the fluctuations of SD_A caused by some observatons such as

$$C_{A(I)} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)})/p\hat{\sigma}^2 \qquad (3.1)$$

,where index set $I = \{i_1, i_2, \cdots, i_m; 1 \le m \le n-p\}$ and under_subsdript (I) indicates a popular deletion rule. And when the number of elements is one, we use 'i', i=1,$\cdots$,n, instead of I. For convenience of expression, we will call the index set of I to 'the observations I' simply.

In order to assess the exact fluctuations of SD_A caused by observations I ,we might have to investigate the $SD_{A(I)}$ which is a deletion version of (2.3). And $SD_{A(I)}$ is given by replacing the terms $\hat{\boldsymbol{\beta}}$, $\boldsymbol{X}'\boldsymbol{X}$ and $\hat{\sigma}^2$ with the terms $\hat{\boldsymbol{\beta}}_{(I)}$, $\boldsymbol{X}'_{(I)}\boldsymbol{X}_{(I)}$ and $\hat{\sigma}^2_{(I)}$ respectively in formular (3.1). As for our experiences, however, it seems to be heavily burdensom that $SD_{A(I)}$ is struturized to several factors like the formular (2.4). But under the postulated assumption described in section 2, each of three terms approximately closes to each of the deletetion versions respectively. Accordingly, it is reasonable to use $C_{A(I)}$ instead of $SD_{A(I)}$ to measure the fluctuations SD_A caused by observations I. It implies that $C_{A(I)}$ assess the influence of observations I on $\hat{\boldsymbol{\beta}}$ along with the included variable $\boldsymbol{A}$,

that is to say, it measure how much the influence of the variable A on the postulated model is due to observations I.

The difference $|\hat{\beta} - \hat{\beta}_{(I)}|$ says that how much observations I fluctuate the $\hat{\beta}$ which is play an important role to explain the influence of the variable A on the postulated model. If observations I do little influence on $\hat{\beta}$, then $\hat{\beta}_{(I)} \simeq \hat{\beta}$ and $C_{A(I)} \simeq SD_A$. On the other hands, if the difference $|\hat{\beta} - \hat{\beta}_{(I)}|$ is large, then $|C_{A(I)} - SD_A|$ is also large, which implies that observatios I has the large the influence on $\hat{\beta}$ and a large portion of the influence of the variable A on the postulated model is due to the observations I. Here,

$$X\hat{\beta} - X\hat{\beta}_{(I)} = X\hat{\beta} - X\hat{\beta} + Xd_I \tag{3.2}$$

,where $d_I = \hat{\beta} - \hat{\beta}_{(I)}$ which is easily calulated by one-procedure updating formular(*see* appendix). By some algebric operations,

$$X\hat{\beta} - X\hat{\beta} = (I_n - N_I)A\hat{\lambda} \tag{3.3}$$

and $Xd_I = (I_n - N_I)Xd_I$. Hence, equation (3.2) becomes

$$X\hat{\beta} - X\hat{\beta}_{(I)} = (I_n - N_I)(A\hat{\lambda} + Xd_I) \tag{3.4}$$

And the formular (3.1) is induced as

$$\begin{aligned} C_{A(I)} &= (A\hat{\lambda} + Xd_I)'(I_n - N_I)'(I_n - N_I)(A\hat{\lambda} + Xd_I)/pMS_E \\ &= (A\hat{\lambda} + Xd_I)'(I_n - N_I)'(A\hat{\lambda} + Xd_I)/pMS_E \\ &= [(A\hat{\lambda} + Xd_I)'(A\hat{\lambda} + Xd_I) - (A\hat{\lambda} + Xd_I)'N_I(A\hat{\lambda} + Xd_I)]/pMS_E \\ &= (D_I - ESS_I)/pMS_E \end{aligned}$$

, where $D_I = (A\hat{\lambda} + Xd_I)'(A\hat{\lambda} + Xd_I)$, $ESS_I = (A\hat{\lambda} + Xd_I)'N_I(A\hat{\lambda} + Xd_I)$ and MS_E is a mean of residuars sum of squares of model (2.1). Note that

$$\begin{aligned} ESS_I &= (A\hat{\lambda})'N_I(A\hat{\lambda}) + (Xd_I)'N_I(Xd_I) + 2(Xd_I)'N_I(A\hat{\lambda}) \\ &= (A\hat{\lambda})'N_I(A\hat{\lambda}) \end{aligned}$$

, since $(Xd_I)'N_I = 0$. Hence, $ESS_I = ESS$. Accordingly, formular (3.4) is become to

$$C_{A(I)} = [ESS/SS_E][D_I/ESS - 1][(n-p)/p] \tag{3.5}$$

, which is exactly same as SD_A except for the second factor. Remember that the first factor of (3.5) is a monotonic increasing funtion of partial F-test with respect to ESS. And the second factor is a monotonic increasing funtion of leverage $h_{(A)}$ which is inversely related to the power of F-test and is also a monotonic function of the variance inflation factor that indicates the multicollinearity associated with the varible A. Accordingly, it is possible to interprete, under the postulated assumption, when $C_{A(I)} > SD_A$ has been observed, the power of partial F-test used full data set was reduced as much as the quantity corresponding to increaments of D up to D_I. This says that

the second factor of $C_{A(I)}$ measures how seriously the observations I cause the F-test to be small and prevent the variable A from entering the postulated model undel given IN-F-*value* . In this case , it can reveal *the intrinsic collinearity* which was hidden by the observations I. On the other hand, when $D_I < D$ and so $C_{A(I)} < SD_A$, the second factor of $C_{A(I)}$ mesures how much the observations I make the power of the partial F-test be better corresponding to the magnitued of changes D to D . Also, in this case, it reveals *the outliers_induced collinearity*, which was introduced by Mason and Gunst(1985).

Because

$$D_I = (A\boldsymbol{\lambda} + X\boldsymbol{d}_I)'(A\boldsymbol{\lambda} + X\boldsymbol{d}_I)$$
$$= D + \boldsymbol{d}_I'X'X\boldsymbol{d}_I + 2\boldsymbol{d}_I'X'A\boldsymbol{\lambda} ,$$

$C_{A(I)}$ is devided by three terms such that

$$C_{A(I)} = SD_A + C_{AI} + 2C_{A2} \qquad (3.6)$$

, where $C_{AI} = \boldsymbol{d}_I'X'X\boldsymbol{d}_I / pMS_E$ and $C_{A2} = \boldsymbol{d}_I'X'A\boldsymbol{\lambda} / pMS_E$. Note that the perturbation term caused by observations I are the last two terms, C_{AI} and C_{A2} in (3.6). Since C_{AI} is always greater than zero, the source that $C_{A(I)} < SD_A$ is due to C_{A2} which must be have negative sign . And, in this case, absolute value of C_{A2} is greater than $C_{AI}/2$. When the observations I little influence on $\boldsymbol{\beta}$, then $\boldsymbol{d}_I \approx \mathbf{0}$ and so $C_{A(I)} \approx SD_A$. And if variable A has little contribute to the model (2.1) , that is $\boldsymbol{\lambda} \approx \mathbf{0}$ and so $SD_A \approx \mathbf{0}$ and $C_{A2} \approx \mathbf{0}$, then we have $C_{A(I)} \approx C_{AI}$ regardless to the collinearity associated with the variable A . And if the variable A is almost orthogonal to X , then since $SD_A \approx \mathbf{0}$, $C_{A2} \approx \mathbf{0}$ and $\boldsymbol{\beta} \approx \boldsymbol{\beta}$ regardless to the contribution to the model, $C_{A(I)} \approx C_{AI}$ which is a simple Cook's statistic in model (2.1).

4. EXAMPLES

In this secton two situatios are considered. The first is a case when the intrinsic collinearity hidden by an observation exists and the second is a cases when the outlier_induced collinearity by an observation exists. To investigate how statistics C_{AI} , C_{A2} and $C_{A(I)}$ detect to such an observation, a well known data set is prepared, which is presented by Hald(1952)(*see* Montgomery and Peck(1982),page 257). This data set includes four independent variables X_1, X_2, X_3 and X_4.

Table 1. correlations for Hald data

	X_1	X_2	X_3	X_4	y
X_1	1.0				
X_2	0.23	1.0			
X_3	-0.82	-0.14	1.0		
X_4	-0.25	-0.97	0.03	1.0	
y	0.73	0.82	-0.54	-0.54	1.0

I. A case of the intrinsic collinearity

It is known that many criteria, for instance R^2, *Mallow's Cp* and SD_A , indicates that a model

with the variable X_1 and X_2 to be adiquate. Now, we consider the influence of X_4 on the postulated model with the independent variables X_1 and X_2. The correlations between X_2 and X_4 is very high as much as a ploblem of collinearity occurs. To pretend that the collinearity is hidden by an observation, X_4 is modified such that x_{41}=60 is replaced with x_{41}=200, and let the modified X_4 be Q_4. Then the correlation coefficiens of X_2 and Q_4 is -0.66, so that we can say that observation x_1 = [7 26 200] hides the intrinsic collinearity. Indeed, the real partial F-value of X_4 is 1.86 is small to reject null hypothesis but for the variable Q_4 ,the partial F-value equals 68.57 which sufficient to accept to include the postulated model. And $C_{Q4(1)}$ = 49.81 > 0.49 = SD_{Q4} , and so it is suspected that there exists the collinearity hidden by observation 1.

Table 2. C_{A1} , $2C_{A2}$ **and** $C_{A(i)}$, **where** A = Q_4 **On** X_1 and X_2

obs.i statitics	1	3	5	7	9	11	13
C_{A1}	40.62	0.03	0.01	0.07	0.05	0.25	0.16
$2C_{A2}$	8.70	0.12	0.08	-0.13	-0.13	-0.67	-0.17
$C_{A(i)}$	49.81	0.63	0.58	0.42	0.41	0.12	0.48

From Table 2, it is apparent that observation 1 highly perturb the postulated model when variable Q_4 comes into the model. The values of C_{A1} , C_{A2} and $C_{A(i)}$ for observation 1 are distinctively larger than others. And last note that SD_{X4} = 47.94 closed to $C_{Q4(1)}$ = 49.81 ,which implies that statistic $C_{A(i)}$ almost measures the original influence of a variable.

II. A case of the outlier_induced collinearity

Here, we consider the influence of X_4 on the postulated model with a independent variable X_3. To pretend that the collinearity is induced by observation 1, [y_1 x_{31} x_{41}]=[78.5 6 60] is modified to [0 100 200], and let's say y,X_3 and X_4 to u, Q_3 and Q_4 respectively. The variable X_3 and X_4 are little correlated each other. But the correlation coefficients between Q_3 and Q_4 is 0.95 , this make it induce the collinearity. While the real partial F-value used original data set is 100.36 , the partial F-value used modified data set is 5.88. Note that the correlation coefficients for (u,Q_3) and (u,Q_4) are -0.94 and -0.95 repectively. So we can say that the partial F-value is apparently decreased becaused of the collinearity induced observation 1. Under the postulated model with independent variable Q_3, $C_{Q4(1)}$ = 17.58 < 30.65 = SD_{Q4} , which implies that observation 1 is able to be suspected to an observation that induces the collinearity.

Table 3. C_{A1} , $2C_{A2}$ **and** $C_{A(i)}$, **where** A = Q_4 **On** Q_3

obs.i statitics	1	3	5	7	9	11	13
C_{A1}	22.64	0.00	0.02	0.08	0.18	0.36	0.05
$2C_{A2}$	-35.75	-0.19	-1.50	3.21	4.64	6.62	-2.24
$C_{A(i)}$	17.58	30.52	29.20	33.98	35.50	37.66	28.49

All of $C_{A(i)}$ are apploxmately same as SD_{Q4} = 30.65 except for the observation 1. And the values of C_{A1} , C_{A2} and $C_{A(i)}$ are distinctively larger in magnitude than others in Table 3, which implies that

the observation 1 highly perturbe the coefficients estimates when the variable Q_4 comes into the postulated model with the independent variable Q_3. Also, note that the real influence of X_4 in original data set, SD_{X4} = 17.43 is very closed to $C_{Q4(1)}$ = 17.58. This means that statistic $C_{A(i)}$ almost measures the original influence of a variable.

REFERENCES

1. Besley,D.A,E.Kuh and R.E.Welsch(1980),*Regression Diagnostic*:Identifing Influential Data and Source of Collinearity,Wiley, NewYork.
2. Choi,C.H,J.H.Koo,J.J.Lee and H.Jorn(1993),"Influential Observations on Vriable Selection in Linear Regression model",The Korean Statistical Society,Vol.**6**,No.2,421.
3. Haglin,P.H and S.Weisberg(1982),"The hat Matrix in Regression and ANOVA",The American Statistician,**32**,17-22.
4. Léger,C and N.Altman(1993),"Assessing Influence in Variable Selection Problems",JASA,Vol.**88**, NO.422,547-556.
5. Mason,R.L and R.F.Gunst(1985),"Outier-induced Collinearities",Technometrics,**27**,401-407.
6. Montgomery,D.C and E.A.Peck(1982),"*Introduction to linear Regression Analysis*",John Wiely:NewYork.
7. Schall,R and T.T.Dunne(1990),"Influential Variables in Linear Regression",Technometrics,**32**, 323-330.

APPENDIX

For the variable $Z = [X:A]_{n\times(p+k)}$,well-known updating formular for the LS coefficients estimates is given by

$$\tilde{\mu} = \tilde{\mu}_{(I)} + (Z'Z)^{-1}Z_I'(I - V_I)^{-1}(y_I - Z_I\tilde{\mu}) \qquad \text{(a.1)}$$

, where y_I and Z_I are the subvector of y indexed I and the submatrix of Z indexed I respectively and $V_I = Z_I(Z'Z)^{-1}Z_I'$. Let $L = (Z'Z)^{-1}Z_I'(I - V_I)^{-1}(y_I - Z_I\tilde{\mu}) = [L_1' , L_2']$, where L_1 and L_2 are the first (p×1) subvector and the last (k×1) subvector of L . Note that

$$\tilde{\mu} - \tilde{\mu}_{(I)} = [d_I' , g_I'] = [L_1' , L_2']$$

,where $d_I = \beta - \beta_{(I)}$ and $g_I = \lambda - \lambda_{(I)}$. So that $d_I = L_1$. And for I =i; i=1,$\cdots$,n, $d_i = L_1$, where L_1 is the first (p×1) subvector of

$$L = (Z'Z)^{-1}z_i'(y_i - z_i\tilde{\mu})/(1 - v_{ii}) \qquad \text{(a.2)}$$

,where v_{ii} is the i*th* diagonal elements of the hat matrix $V = Z(Z'Z)^{-1}Z'$.