应用概率统计 第十一卷
第三期 1995年8月

Chinese Journal of Applied Probability
and Statistics Vol. 11, No. 3, Aug. 1995

On Effects of Augmentation of data on Condition Index*

YANG HU
(*Chongqing Jiaotong Institute, Chongqing, 630074*)

WANG SONGGUI, ZHANG ZHONGZHAN
(*Beijing Polytechnic University, Beijing, 100022*)

Abstract

There are three approaches in the literature to remedy the multicollinearity in design matrix in a linear regression: augmentation of data, variable selection and alternative procedures to the ordinary least squares. In this note our emphasis is focused on the effect of augmentation of data on condition index. Our results show that when the additional data are properly chosen in practical possible situation, the condition index of design matrix can be reduced. the results obtained here are illustrated by Gaylor-Merrill data [1] which has been extensively discussed in the literature on regression optimal design.

Keywords: Condition number, Condition index, Multicollinearity.
AMS Subject Classification: 62J05.

§1. Introduction

In the recent years, the study on the multicollinearity among the independent variables in a linear regression has received much attention in the literature (see, for example, Mason, Gunst and Webster [2] , Kurmar[3], Dorsett, Gunst and Gartland[4], Belsely, Kuh and Welsch[5], Wetherill[6]). It is well known that severe multicollinearity among the independent variables could be very harmful to the statistical inference, in particular, to parameter estimation. Several remedy methods for multicollinearity in a linear regression has been proposed (see, Wetherill[6]). These include the augmentation of data , variable selection, and alternative procedures to ordinary least square estimator (e.g. ridge regression, principal component and latent root regression). However, augmenting experimental data within a region to maximize some criterion has been an interesting topic in the literature on regression optimal design, see, for example, Gaylor and Merrill[1], Dykstra[7], Box and Draper[8] and Evens[9]. In this note, our primary emphasis is on the effect of the

*Project supported partially by National Natural Science Foundation of China.
本文1992年2月26日收到, 1994年2月20日收到修改稿.

augmentation of data on multicollinearity.

Consider the usual multiple linear regression model

$$y = X\beta + e, \tag{1.1}$$

where y is an $n \times 1$ vector of observations on a response variable, X is an $n \times p$ matrix of n observations on p independent variables, e is an $n \times 1$ error vector with zero mean and covariance matrix $\sigma^2 I$, where I stands for the identical matrix. β and σ are unknown parameters. Throughout this paper we assume that X is centered and standardized

One of the popular measures for collinearity in X is the condition number of the matrix $X'X$, which is defined as

$$K(X'X) = \frac{\lambda_1(X'X)}{\lambda_p(X'X)}, \tag{1.2}$$

where $\lambda_1(X'X) \geq \cdots \geq \lambda_p(X'X)$ denote the order eigenvalues of $X'X$. The large values of $K(X'X)$ indicate the existence of collinearity in X. The condition number is closely related to the relative efficiency of the least square estimator(see, Yang Hu and Wang Song-gui[10]). The condition number also plays an important role in the study of the sensitivity of β (see, for example, Golub and Van Loan[11]).

Besides the condition number, another important measure for collinearity is the condition index which is a set of $p-1$ values

$$K_i(X'X) = \frac{\lambda_1(X'X)}{\lambda_i(X'X)} \quad i = 2, \cdots, p \tag{1.3}$$

The merit of the condition index is that the number of large values of the K indicates the number of the collinearities in X.

The results from the point of view of the eigenvalue analysis given by Wang, Tse and Chow[12] show that k additional data chosen appropriately can remove k collinearities from X. However it is possible that these additional data may be high leverage points. The diagonal element of the hat matrix $H = (h_{ij}) = X(X'X)^{-1}X', h_{ij}$denoted by h_i henceforth for the simplicity is a measure for the i-th row in X to be a high leverage case. Hoaglin and Welsch[13] suggested that the i-th row in X is called as a high leverage case if $h_i > 2p/n$. The high leverage case may make extreme influence on regression analysis, which is not expected in practice. Theorem 1 in the following section establishes an interesting relationship between the condition index and h_i when k additional rows are augmented to the design matrix. Some mild conditions on the augmentation data for decreasing the condition number are given in Theorems 2 and 3 of section 3. The results obtained in this note show the potentiality of removing multicollinearities. To illustrate our results, the numerical example given by Gaylor and Merrill [1] is discussed in section 4. This example has been extensively studied by Dyksbra[7], Box and Draper[8] and Evens[9] from the point of view of augmenting experimental data.

§2. High leverage cases and the condition index

Denote by $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ the eigenvalues of A. For the simplicity, $\lambda_1 \geq \cdots \geq \lambda_n$ will be reserved for the eigenvalues of $X'X$. Let $\phi_1, \cdots, \phi_p$ be the orthonormal eigenvectors of $X'X$.

If we take k additional data as

$$X_0 = \begin{pmatrix} x'_{n+1} \\ \vdots \\ x'_{n+k} \end{pmatrix}, \tag{2.1}$$

where $x_{n+1} = c_1\phi_{p-k+1}, \cdots, X_{n+k} = c_k\phi_p$, then the unordered eigenvectors of X'_aX_a, where $X_a = (X'\vdots X'_0)'$ are $\lambda_1, \cdots, \lambda_{p-k}, \lambda_{p-k+1} + c_1^2, \cdots, \lambda_p + c_k^2$. (Wang et al.[12]). It is easy to see that for appropriately chosen $c_1, \cdots, c_k$, we have

$$\begin{aligned} &\lambda_i(X'_aX_a) = \lambda_i, && i \leq p-k, \\ &\lambda_i(X'_aX_a) = \lambda_i + c^2_{i-(p-k)}, && i > p-k. \end{aligned} \tag{2.2}$$

Furthermore, we can prove the following theorem,which establishes a relationship between the condition index of X'_aX_a and $X'X$.

Denote

$$I_{(a,b]}(i) = \begin{cases} 1, & i \in (a,b], \\ 0, & i \notin (a,b]. \end{cases}$$

Theorem 1 For the X_0 and $c_r, r = 1, \cdots, k$ defined by (2.1) and (2.2)

$$K_i(X'_aX_a) = K_i(X'X)[1 - h^{(a)}_{i+n+k-p}I_{(p-k,p]}(i)], \quad i = 2, \cdots, p,$$

where $h_i^{(a)}$ is the diagonal element of the hat matrix $H_a = X_a(X'_aX_a)^{-1}X'_a$.

Proof It is easy to see that we need to prove the theorem for $i > p-k$ only. In this case, it follows from (2.2) that

$$K_i(X'X) = \frac{\lambda_1}{\lambda_i} = \frac{\lambda_1(X'_aX_a)}{\lambda_i(X'_aX_a) - c^2_{i-(p-k)}} = K_i(X'_aX_a)[1 - \lambda_i^{-1}(X'_aX_a)c^2_{i-(p-k)}]^{-1},$$

i.e.

$$K_i(X'_aX_a) = K_i(X'X)\left(1 - \frac{c^2_{i-(p-k)}}{\lambda_i(X'_aX_a)}\right).$$

Thus it is sufficient to show

$$h^a_{i+n+k-p} = \frac{c^2_{i-(p-k)}}{\lambda_i(X'_aX_a)}, \quad i > p-k. \tag{2.3}$$

To do so, let

$$X = P\Lambda^{1/2}Q' \tag{2.4}$$

be the singular value decomposition of X, where P is an $n \times n$ orthogonal matrix, Q is a $p \times p$ orthogonal matrix, $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$. Denote $C = diag(c_1, \cdots, c_k)$ and partition P and Λ as

$$P = (P_1 \vdots P_2), \quad P : n \times (p-k), \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}, \quad \Lambda_1 : (p-k) \times (p-k).$$

Thus the argumented design matrix admits the representation

$$X_a = \begin{pmatrix} P_1 \Lambda_1^{1/2} & P_2 \Lambda_2^{1/2} \\ 0 & C \end{pmatrix} Q'.$$

Consequently

$$H_a = \begin{pmatrix} P_1 P_1' + P_2 \Lambda_2^{1/2} (\Lambda_2 + C^2)^{-1} \Lambda_2^{1/2} P_2' & * \\ * & C(\Lambda_2 + C^2)^{-1} C \end{pmatrix}. \tag{2.5}$$

It can be verified that the south-east corner of (2.5) gives (2.3). The proof is completed.

An important special case $i = p$ in Theorem 1 gives a following relationship of the condition number

$$K(X_a' X_a) = K(X'X)(1 - h_{n+k}^{(a)}),$$

from which we know that the condition number $K(X_a' X_a)$ decrease as $h_{n+k}^{(a)}$ increase. Generally, however, we do not expect that any of the k additional data is a high leverage case. The expression of the new hat matrix H_a(2.5) may be rewritten as

$$H_a = \begin{pmatrix} PP' + P_2 [\Lambda_2^{1/2} (\Lambda_2 + C^2)^{-1} \Lambda_2^{1/2} - I] P_2' & * \\ * & C(\Lambda_2 + C^2)^{-1} C \end{pmatrix} \tag{2.6}$$

$$= \begin{pmatrix} H + P_2 [(I + \Lambda_2^{-1/2} C^2 \Lambda_2^{-1/2})^{-1} - I] P_2' & * \\ * & C(\Lambda_2 + C^2)^{-1} C \end{pmatrix}. \tag{2.7}$$

In the last equality we use $H = PP'$ which follows from (2.4). Since the matrix

$$P_2 [(I + \Lambda_2^{-1/2} C^2 \Lambda_2^{-1/2})^{-1} - I] P_2'$$

is a negative definite, thus the first n diagonal elements of H_a are all not greater than those of H. This results shows that the case which is not high leverage in the original regression model will not be high leverage in the augmented model. However, in order to avoid that the k additional data are high leverage cases, according to the following high leverage criterion suggested by Haglin and Welsch [13] the i-th case in X is said to be 'high leverage' one if the corresponding diagonal element h_i of H is greater then $2p/n$ (also see Cook and Weisberg, [14]). We require

$$C(\Lambda_2 + C^2)^{-1} C \leq \frac{2p}{n+k} I, \tag{2.8}$$

where $A \leq B$ stands for the Löwner ordering, i.e., A and B are symmetric s.t. $B - A$ is nonnegative definite. (2.8) is equivalent to

$$c_i \leq \left(\frac{2p}{n+k-2p} \lambda_{i+p-k} \right)^{1/2}, \quad i = 1, \cdots, k.$$

Thus we have the following corollary.

Corollary 1 If we take

$$c_i = \left(\frac{2p}{n+k-2p}\lambda_{i+p-k}\right)^{1/2}, \quad i = 1, \cdots, k \tag{2.9}$$

in (2.4), then (1) The diagonal elements of H_a

$$h_i^{(a)} \begin{cases} \leq h_i, & i \leq n, \\ = \dfrac{2p}{n+k}, & i > n, \end{cases} \tag{2.10}$$

(2) $$\lambda_i(X_a'X_a) = \left[1 + \frac{n+k}{n+k-2p} I_{(p-k,p]}(i)\right] \lambda_i, \tag{2.11}$$

(3) $$K_i(X_a'X_a) = K_i(X'X)\left[1 - \frac{2p}{n+k} I_{(p-k,p]}(i)\right]. \tag{2.12}$$

The above discussion indicates that if the k additional data are chosen by (2.1) and (2.2), then to avoid any of the k data to be high leverage case, the best choice of the $c_i, i = 1, \cdots, k$ in (2.1) for decreasing the condition index is (2.9). In this case, the condition index follows relation (2.11) and some severe multicollinearities in X may not be removed.

Walker[15] also studied relationship between the condition index and the diagonal element of the hat matrix. However he consider a rare situation where a deleted case $x_i = C\phi_j$, given X.

§3. Influence of augmented data on the conditional number

In the previous section we assume that the k additional data satisfy the strong condition (2.1). In this section, however, we consider a general case in which the additional data may be either arbitrary or restricted by using some mild conditions. Thus in the following discussion, we will adopt the notation $\tilde{X}$ to denote the k additional data and $X_a = (X'\vdots\tilde{X}')'$.

Theorem 2 (1)

$$\lambda_i(X_a'X_a) = \lambda_i + t_i\lambda_1(\tilde{X}'\tilde{X}), \quad i = 1, \cdots, p, \tag{3.1}$$

where $0 \leq t_i \leq 1, i = 1, \cdots, p, \sum_{i=1}^p t_i = \mathrm{Tr}(\tilde{X}'\tilde{X})/\lambda_1(\tilde{X}'\tilde{X})$.

(2) If

$$\lambda_1(\tilde{X}'\tilde{X}) \leq [\lambda_p(X_a'X_a) - \lambda_p]K(X'X), \tag{3.2}$$

then $K(X_a'X_a) \leq K(X'X)$.

Proof (1) Using Weyl theorem (cf. Horn and Johnson, [16], p.184) in the following equality

$$X_a'X_a = X'X + \tilde{X}'\tilde{X}$$

yields

$$\lambda_i(X_a'X_a) \leq \lambda_i + \lambda_1(\tilde{X}'\tilde{X}), \quad i = 1, \cdots, p. \tag{3.3}$$

It is obvious that

$$\lambda_i(X_a'X_a) \geq \lambda_i, \quad i = 1, \cdots, p. \tag{3.4}$$

From (3.3) and (3.4) it follows that

$$\lambda_i(X_a'X_a) = \lambda_i + t_i\lambda_1(\tilde{X}'\tilde{X}), \text{ for some } 0 \leq t_i \leq 1. \tag{3.5}$$

Since $tr(X_a'X_a) = tr(X'X) + tr(\tilde{X}'\tilde{X})$, where trA stands for the trace of A, (3.1) is straightforward from (3.5).

(2) From (3.5) with $i = 1$ we have

$$K(X_a'X_a) = \frac{\lambda_1(X_a'X_a)}{\lambda_p(X_a'X_a)} = \frac{\lambda_p}{\lambda_p(X_a'X_a)}\left[K(X'X) + t_1\frac{\lambda_1(\tilde{X}'\tilde{X})}{\lambda_p}\right], \tag{3.6}$$

by using $t \leq 1$ and (3.2). The proof is completed.

As mentioned above, the additional data in many practical situations may not exactly satisfy the condition (2.1). They may be represented as

$$\tilde{X} = X_0 + E, \tag{3.7}$$

where X is the same as that in section 2. E is a $k \times p$ deviation matrix of $\tilde{X}$ from X_0. Our next theorem shows that if E is not very large in some sense, then the additional data will lead up to decrease the condition number. Denote

$$X_a = \begin{pmatrix} X \\ \tilde{X} \end{pmatrix}, \qquad d = \left(\frac{c_k}{c}\right)^2\left(1 + \frac{1}{K(X'X)}\right)^{-1},$$

where $c = \max\{|c_i|\} = \|X_0\|$, the spectral norm of X_0.

Theorem 3 Assume

$$\|E\| \leq c\left(\sqrt{1+d} - 1\right). \tag{3.8}$$

Then

$$K(X_a'X_a) \leq K(X'X).$$

Proof Since

$$X_a'X_a = X'X + (X_0 + E)'(X_0 + E) = X'X + X_0'X_0 + E'E + E'X_0 + X_0'E,$$

it follows from Weyl Theorem ([12]) that

$$\begin{aligned} &|\lambda_i(X_a'X_a) - \lambda_i(X'X + X_0'X_0)| \\ &\quad \leq \max\{\lambda_1(E'E + E'X_0 + X_0'E), |\lambda_p(E'E + E'X_0 + X_0'E|\} \\ &\quad = \|E'E + E'X_0 + X_0'E\| \leq \|E\|(\|E\| + 2\|X_0\|). \end{aligned}$$

Hence

$$\lambda_i(X_a'X_a) = \lambda_i(X'X + X_0'X_0) + w_i\|E\|(\|E\| + 2\|X_0\|), \quad |w_i| \le 1, i = 1, \cdots, p,$$
$$= \begin{cases} \lambda_i + w_i\|E\|(\|E\| + 2\|X_0\|), & i \le p-k, \\ \lambda_i + c_{i-p+k}^2 + w_i\|E\|(\|E\| + 2\|X_0\|), & i > p-k, \end{cases}$$

from which we obtain

$$K(X_a'X_a) \le \frac{\lambda_1 + \|E\|(\|E\| + 2\|X_0\|)}{\lambda_p + c_k^2 - \|E\|(\|E\| + 2\|X_0\|)}$$
$$= \frac{\lambda_1}{\lambda_p} \frac{1 + \lambda_1^{-1}\|E\|(\|E\| + 2\|X_0\|)}{1 + \lambda_p^{-1}(c_k^2 - \|E\|(\|E\| + 2\|X_0\|)}$$
$$= K(X'X)\left(1 + \frac{S}{\lambda_1\lambda_p\{1 + \lambda_p[c_k^2 - \|E\|(\|E\| + 2\|X_0\|)]\}}\right),$$

where $S = \lambda_p(\|E\|^2 + 2\|X_0\| \cdot \|E\|) - \lambda_1[c_k^2 - (\|E\|^2 + 2\|X_0\| \cdot \|E\|)]$. Obviously, it is sufficient to show $S \le 0$. In fact, S can be decomposed as $S = [\lambda_1 + \lambda_p]S_1S_2$, where

$$S_1 = \|X_0\| + \|E\| + \sqrt{\|X_0\|^2 + c_k^2\lambda_1(\lambda_1 + \lambda_p)^{-1}} \ge 0,$$
$$S_2 = \|X_0\| + \|E\| - \sqrt{\|X_0\|^2 + c_k^2\lambda_1(\lambda_1 + \lambda_p)^{-1}} \le c + \|E\| - c\sqrt{1+d} \le 0,$$

which follows from (3.8). The proof is completed.

Condition (3.8) shows that when the deviation E of additional data from X is not very large, the augmentation of data will lead to reduction of condition number of the design matrix. The upper bound of the deviation, measured by $\|E\|$, depends on c_i and the condition number $K(X'X)$ of original design matrix. The larger the $K(X'X)$ is , the larger the upper bound of the deviation is.

In our discussion, the assumption $\mathrm{Cov}(e) = \sigma^2 I$, i.e., the random errors are uncorrelated and have the same variance is adapted, the results obtained here, however, can be easily extended to the general case $\mathrm{Cov}(e) = \sigma^2\Sigma$, where Σ is known positive definite matrix.

§4. An example

To demonstrate the feasibility of the conditions of Theorem 2 and 3, we shall discuss the data set with 20 cases and 3 independent variables presented by Gaylor and Merrill[1] and later considered by many authors (see, for example, [7], [8], and [9]). The data follows the linear regression model (1.1). The original data of independent variables is reprinted in Table 1.

The experimental region of interest is $-7 \le x_1 \le 6, -7 \le x_2 \le 7, -5 \le x_3 \le 12$. Gaylor and Merrill [1] pointed out that the candidates of points for augmenting experimental data to maximize $|X_a'X_a|$ (i.e. D-optimal) are the corners of the experimental region. In the following we shall only consider the two kind of corner points. The first is the corners of the experimental region. The second is called corners of data set, such as

Table 1 Original data: $n = 20, p = 3$

Run	x_1	x_2	x_3	Run	x_1	x_2	x_3
1	−6.389	−5.330	6.0437	11	−1.593	−3.957	0.1896
2	−6.179	−5.549	9.0819	12	1.338	−2.613	−0.3136
3	−4.533	−5.717	7.6283	13	−0.787	−2.487	−2.7032
4	−5.293	−6.492	7.8131	14	−1.649	−1.077	−2.1917
5	−4.004	−6.464	3.0976	15	2.075	1.719	−2.2917
6	−2.631	−5.320	5.4978	16	2.224	0.946	−2.8516
7	−3.012	−4.080	1.0688	17	2.382	3.879	−4.2335
8	−2.864	−4.583	4.5822	18	3.350	3.510	−4.8033
9	−0.979	−2.887	1.0250	19	3.384	6.383	−2.5554
10	−0.420	−4.094	2.2669	20	5.984	6.499	−2.1206

$(-6.389, 6.499, 9.0819)$, which are constructed by maximums and minimums of x_i in the data set.

For the simplicity, we will denote by 1 and -1 the maximum and the minimum of the value of each independent variable in the data set respectively. Thus the corners of the data set are the following:

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ (-1,-1,-1) & (1,-1,-1) & (-1,1,-1) & (1,1,-1) \\ 5 & 6 & 7 & 8 \\ (-1,-1,1) & (1,-1,1) & (-1,1,1) & (1,1,1) \end{array}$$

The eigenvalues of $X'X$ are 2.716011, 0.204268 and 0.079721, and the condition number of $X'X$ is 34.068913.

1. The Condition (3.2) Some results about Theorem 2 are listed in Table 2. For $k = 1$, there are 5 corner points satisfy the condition (3.2) and so the condition number decreased, among which the point 6 is also D-optimal [7]. For $k = 2$, 36 combinations among 28 combinations satisfy the condition (3.2) and the minimum of the condition number $K(X_a'X_a)$ is 4.247848 which is corresponding to (6, 7), the D-optimal [7].

If we add 3 runs to experimental data , we have 97 candidates in all 120 satisfying the condition (3.2), the minimum of condition number $K(X_a'X_a)$ is 3.785424, which is corresponding to (6, 6, 7). Although the (6, 6, 7) is not D-optimal, its condition number 3.785424 is very close to the condition number 3.932281 of the D-optimal augmentation (6, 6, 8).

Furthermore, our computing experience shows that the closer $\lambda_1(\tilde{X}'\tilde{X})$ is to the bound, the less the condition number decrease, and in a large neighbourhood of a D-optimal point, the bound is about twenty times of $\lambda_1(\tilde{X}'\tilde{X})$ if $k = 2$ or $k = 3$, so the condition (3.2) is very easy to satisfy , and does not has any essential limitation to useful candidates.

All above conclusions also hold for the corners of experimental region.

2. The Condition (3.8) At first we consider the corners of the data set . Because we assume that X_0 in (3.7) satisfies (2.2), so if $k = 1, c_1^2 \leq \lambda_2 - \lambda_3$, and if $k = 2, c_1$ and c_2

Table 2 Some results for Theorem 2

k	$\tilde{X}$	$\lambda_1(\tilde{X}'\tilde{X})$	bound*	$K(X_a'X_a)$	Δ^{**}	D-optimality
1	(2)	0.396814	2.295782	19.139748	14.929165	
	(3)	0.498297	4.191665	13.711267	20.357646	
	(6)	0.616223	2.938895	16.711920	17.356993	√
2	(1,3)	0.573915	8.985856	8.211359	25.857554	
	(2,6)	0.654269	11.347485	6.936407	27.132506	
	(3,7)	0.836458	11.201538	7.012513	27.056400	
	(3,8)	0.829669	12.442835	6.379636	27.689277	
	(6,7)	0.719179	20.324554	4.247848	29.821065	√
3	(1,3,7)	0.849803	20.359003	4.274277	29.794636	
	(1,6,7)	0.877293	22.780813	3.870490	30.198423	
	(2,3,8)	0.831496	22.700220	3.902745	30.166168	
	(2,6,7)	1.091953	20.667645	4.373322	29.695591	
	(3,3,8)	1.110215	19.373817	4.498054	29.570859	
	(3,6,7)	1.107467	22.697376	3.946184	30.122729	
	(3,7,8)	1.126405	23.271837	3.799649	30.269264	
	(6,6,7)	1.233046	23.946129	3.785424	30.283489	
	(6,7,8)	1.282491	22.364878	3.932281	30.136632	√

* The bound means the right hand side of (3.2), i.e., $[\lambda_p(X_a'X_a) - \lambda_p]K(X'X)$.
** $\Delta = K(X'X) - K(X_a'X_a)$.

satisfy $c_1^2 \le \lambda_1 - \lambda_2$, and $c_2^2 \le c_1^2 + \lambda_2 - \lambda_3$. These relationships put some limitations on $\tilde{X}$ indeed, but the limitations are not severe for the D-optimal augmentations. In fact, after data are centered and standardized by data mean and data deviation, the projections of corner points on ϕ_i are not large. In our example, many good augmentations including D-optimal points satisfy (2.2) and (3.8). Some details are given in Table 3.

Table 3 Some results for Theorem 3

corner	k	$\tilde{X}$	$\|E\|$	bound*	$K(X_a'X_a)$	Δ^{**}	D-optimality
data corner	1	(3)	0.053236	0.241754	13.774960	20.293953	
		(6)	0.089454	0.241685	15.529750	18.539153	√
	2	(3,8)	0.156968	0.223110	8.055696	26.013217	
		(6,7)	0.229746	0.241685	5.462787	28.606126	√
		(6,8)	0.140117	0.222998	10.364929	23.703984	
region corner	1	(3)	0.057354	0.268334	13.711267	20.357646	
		(6)	0.193932	0.262597	16.711920	17.356993	√
	2	(3,8)	0.131533	0.228608	6.379636	27.689377	
		(6,8)	0.213418	0.219958	9.605017	24.463896	

* The bound stands for the right hand side of (3.8), i.e., $c(\sqrt{1+d}-1)$.
** $\Delta = K(X'X) - K(X_a'X_a)$.

In the case of the corner of the experimental region, the D-optimal augmentation (6.7) in case $k = 2$ does not follow (3.8). This is the only exception, but the value of right

hand side of (3.8) almost equal to the left at this point.

In summary, this example demonstrates the potentiality of the data augmentation to remedy the multicollinearity.

Acknowledgement. The authors are indebted to a referee for valuable comments which led to improvements in the paper.

References

[1] Gaylor, D. W. and Merrill, J. A., Augmenting existing data in multiple regression, Technometrics **10**(1968), 73–81.
[2] Mason, R. L., Gunst, R. F, and Webster, J. T., Regression analysis and problems of multicollinearity, Commu. Statist. **4**(1975), 277–292.
[3] Kurmar, T. K., Multicollinearity in regression analysis, Rev. Econom. Statist. **57**(1975), 365–366.
[4] Dorsett, D., Gunst, R. F., and Gartland, Jr. E. C., Multicollinear effects of weighted least squares regression, Statist. Proba. Lett. **1**(1983), 207–211.
[5] Belsley, D. A., Kuh, E., and Welsch, R. E., *Regression Diagnostice: Identifying Influential Data and Sources of Collinearity*, Wiley, New Yok, 1980.
[6] Wetherill, G. B., *Regression Analysis with Applications*, Chapman and Hall, New Yok, 1986.
[7] Dykstra, O., The augmentation of experimental data to maximize $|X'X|$, Technometrics **13**(1971), 682–688.
[8] Box, M. J. and Draper, N. R., Factorial designs, the $|X'X|$ criterion, and some related matters. Technometrics **13**(1971), 731–742.
[9] Evans, J. W., Computer Augmentation of Experimental Designs to Maximize $|X'X|$, Technometrics **21**(1979), 321–330.
[10] Yang, H., Wang, S. G., Condition Number, Spectral Norm and Measure of Inefficiency, Chinese J. of Appl. Prob. and Stat. **7**(1991), 337–343.
[11] Golub, G., Van Loan, C. F., *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1983.
[12] Wang, S. G., Tse, S. K., Chow, S.C., On the Measures of Multicollinearity in Least Squares Regression, Statist. Prob. Lett. **9**(1990), 347–355.
[13] Haglin, D. C. , Welsch, R., The Hat Matrix in Regression and ANOVA, Amer. Statistican **32**(1978), 17–22.
[14] Cook, R. D., Weisberg, S., *Residuals and Influence in Regression*, Chapman and Hall, London, 1989.
[15] Walker, E., Detection of Collinearity-influential Observations, Commu.Statist. **18**(1989), 1675–1690.
[16] Horn, R. A., Johnson, C. R., *Matrix Analysis*, Cambridge Univ. Press, London, 1985.

追加数据对条件指数的影响

杨 虎　　王松桂　张忠占

(重庆交通学院, 重庆)　(北京工业大学, 北京)

克服线性回归模型中设计矩阵的复共线性, 主要有三种方法:追加数据, 自变量选择和非最小二乘法, 本文研究追加数据在减少条件数中的作用, 我们的研究表明, 在可能情况下适当地选择追加数据, 设计矩阵的条件指数可以被减少. 我们用在回归最优设计中广泛被研究过的Gaylor-Merrill 数据说明了本文理论结果的实用意义.

关键词: 条件数, 条件指标, 多重共线性.

学科分类号: 212.1.