Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 73(1) 2019 11–17

A Note on Effects on the Eigenstructure of a Data Matrix when Deleting a Subset of Observations

Ravindra Khattree
Department of Mathematics and Statistics, Co-Director, Center for Data Science and Big Data Analytics, and Participating Member, Center for Biomedical Research, Oakland University, Rochester, MI, 48309, USA

Received 01 July 2016; Revised 11 October 2018; Accepted 09 November 2018

SUMMARY

We provide a result which elegantly helps us identify influential observations in a data matrix based on the eigenstructure of a specific matrix which measures the effect of one or more influential observations. The theorem suggests that the corresponding statistic is easily computable. We illustrate its usefulness in data cleaning prior to modeling using a classical data of Graybill and Iyer and provide its implementation using a short SAS code. This approach is especially useful for large data, where model-free approach to identification of influential observations is a natural choice.

Keywords: Data cleaning, Data processing, Eigenstructure, Eigenvalues, Emphasis measure, Influential observations.

1. INTRODUCTION

In an article with almost the same title as this one, Wang and Nyquist (1991) explored what happens to the eigenstructure of the matrix $X'_{(-i)}X_{(-i)}$, compared to that of $X'X$ when an (row) observation is deleted from the $n \times p$ data matrix X. They attempted to develop a relationship between the two sets of eigenvalues and provided an approximation for the eigenvalues of $X'_{(-i)}X_{(-i)}$. They also showed that the eigenvalues of the two matrices follow certain interlacing inequalities, albeit under certain conditions. This was used to assess the impact on multicollinearity and determine the influence of a particular observation.

In this note, we address the same problem but with a different suggestion about the eigenstructure of a different matrix which is a function of the above two matrices. It turns out that this approach greatly simplifies the problem and provides a more elegant and more easily computed statistic. This is done through Theorem 1. Before we state it, we must set up the notations and motivate the problem.

Corresponding author: Ravindra Khattree
E-mail address: khattree@oakland.edu

2. THE FORMULATION AND THE RESULT

For a given data matrix X; let $A = X'X$ and let $B = X'_1X_1$ be the $p \times p$ matrix where X_1 is obtained by discarding (without loss of generality) last r rows of data matrix X. Also, we will assume that $r \leq p$ and that both A as well as B are of rank p.

How do we measure the collective influence of the r discarded data points? An appropriate approach will be to assess how different B is from A. If the data points were not influential then we expect B to resemble A in some meaningful sense - in our discussion in terms of their respective eigenstructures. In order to measure this distance between the two matrices, let U be the upper triangular square root matrix of B such that $B = U'U$ and define,

$$G = UA^{-1}U': \tag{1}$$

Clearly, if A and B are close to each other then G must resemble I_p; the identity matrix of order $p \times p$; and eigenvalues of G must be close to 1. The following theorem shows that the eigenstructure of G is pretty simple.

Theorem 1. *Let* $\delta_1 \geq \delta_2 \geq ... \geq \delta_p$ *be the ordered eigenvalues of* G. *Then* $\delta_j = 1$ *for* j = 1,2, .., $(p - r)$.

Proof. Let $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where the order of X_1 is $(n-r)\times p$ and the $r\times p$ matrix X_2 contains the last r row observations of X. Then for A and B defined earlier, it is easily seen that,

$$A^{-1} = (B + X'_2X_2)^{-1}$$
$$= B^{-1} - B^{-1}X'_2(I_r + X_2B^{-1}X'_2)^{-1}X_2B^{-1}$$

Thus from (1),

$$G = U[B^{-1} - B^{-1}X'_2(I_r + X_2B^{-1}X'_2)^{-1}X_2B^{-1}]U' = I_p - Z$$

where,

$$Z = U'^{-1}X'_2(I_r + X_2(X'_1X_1)^{-1}X'_2)^{-1}X_2U^{-1}$$
$$= U'^{-1}X'_2(I_r + X_2(U^{-1}U'^{-1})X'_2)^{-1}X_2U^{-1}$$
$$= W'(I_r + WW)^{-1}W,$$

and where $r\times p$ matrix $W = X_2U^{-1}$ is of rank r. Thus, the above matrix Z has the last $(p-r)$ eigenvalues as 0. Consequently, $G = I_p - Z$ has the first $(p-r)$ eigenvalues as 1.

The above theorem clearly indicates that the effect of deleting r observations is concentrated only on last r eigenvalues of G. This effect can be measured by some appropriate and meaningful function of these eigenvalues such as $\sum_{i=p-r+}$ or $\prod_{i=p-r+1}^{p}\delta_i$ or the smallest eigenvalue δ_p. Of special practical interest is the case of $r = 1$, that is, when deleting one observation at a time to assess if the particular observation is influential. - the case considered by Wang and Nyquist (1991). The following theorem for this special case indicates that in this special case, this measure can be computed rather easily.

Theorem 2. *Consider the matrix* G *when* $X_2 = x'$ is the n^{th} *row of matrix* X. *Then the smallest eigenvalue of* G *is given by*

$$\delta_p = 1 - x'(X'X)^{-1}x. \qquad (2)$$

Proof. In view of Theorem 1, it follows that only one of the eigenvalues of G matrix is not equal to 1 and thus

$$\delta_p = tr(G) - (p - 1).$$

But,

$$tr(G) = \text{tr}[U(X'X)^{-1}U']$$
$$= tr[(X'X)^{-1}U'U]$$
$$= tr[(X'X)^{-1}B] = tr[(X'X)^{-1}(X'X - xx')]$$
$$= \text{tr}(I_p) - x'(X'X)^{-1}x.$$

Thus, $\delta_p = tr(G) - p + 1 = 1 - x'(X'X)^{-1}x$.

Therefore, the influence of an observation can be measured by $x'(X'X)^{-1}x$. However, while the mathematical expression is identical in appearance, this should not be confused with what Cook and Weisberg (1982) call leverage or potential. Specifically, our data matrix may not be the model matrix in that the model matrix will usually contain a constant column corresponding to intercept and may further contain columns corresponding to other terms such as polynomial or cross-products. Leverage is computed for a given model and hence in this case, the matrix X must be replaced by the augmented matrix $[X : X^*]$ where X^* contains the columns corresponding to polynomial, cross products, intercepts or other similar terms, which are integral part of the model assumed. Our object of interest here is the matrix of raw data X without any consideration of model. Thus, to make the conceptual distinction between the two clearer, we may alternatively term our $x'(X'X)^{-1}x$ as *Emphasis* of the observation x, while leverage is defined as $[x' : x'^*]([X : X^*]'[X : X^*])^{-1}[x' : x'^*]$, with $[x' : x'^*]$ defined similarly. More explicitly, the leverage is a function of data as well as assumed model, while emphasis is defined as an exclusive function of data matrix of explanatory variables.

Clearly, smaller the value of δ_p, larger is the emphasis, $x'(X'X)^{-1}x$. Khattree (2015) explores, among other things, the use of G matrix in various applications of identification of influential observations and in a variety of data sets.

In the general context of deleting r observations, it is appropriate to use a meaningful function of last r eigenvalues of the G matrix. One such measure could be based on $tr(G) = \delta_{p-r+1} + \delta_{p-r+2} + ... + \delta_p + (p-r)$. Also $tr(G)$ can be expressed as $tr(G) = tr[X_1(X'X)^{-1}X'_1] = p - tr[X_2(X'X)^{-1}X'_2]$. Thus, a measure of influence in this case is simply $tr[X_2(X'X)^{-1}X'_2]$. We may as well term the matrix $[X_2(X'X)^{-1}X'_2]$ as the emphasis matrix.

The determinant of G is another such measure. However, being a product of eigenvalues which are less than 1, its value becomes very small as r increases. Alternatively, one may argue that in the ideal situation of absolutely no influence, all eigenvalues and hence (equivalently) the smallest one must equal 1. Thus the smallest eigenvalue or its departure from unity can be taken as an appropriate index of collective influence. However, this quantity essentially measures influence only in the direction of only one principal axis.

3. AN ILLUSTRATION: MOUNTAIN DATA

We take an example of Graybill and Iyer (1994, Table 5.6.1 page 400) where seven variables (X_1 = length at birth, X_2; X_3 = mother and father's heights at age 18, X_4; X_5 = maternal grandmother and grandfather's heights at age 18, X_6; X_7 = paternal grandmother and grandfather's heights at age 18) are measured for 20 individuals who had lived in mountain isolation for several generations. Graybill and Iyer were interested in modeling the person's height at age 18 as a function of these variables. Our interest here is in finding influential observations through the changes in the eigenstructure of the X′X matrix where X is a 20 × 7 matrix consisting of 20 observations on seven variables specified above. Data are given as part of Table 1. The column titled "Deleted Obs." indicates the observation number of the original data as given by these authors, which we deleted in the analysis that follows. We calculate three measures based on $\boldsymbol{G}$ matrix, namely,

$$I_1 = [tr(\mathrm{G}) - (p - r)]/r,$$

$$I_2 = [det(\mathrm{G})]^{\frac{1}{r}},$$

and

$$I_3 = \text{smallest eigenvalue of G} = \delta_p.$$

By definition, for complete data, when $r = 0$, $I_1 = I_2 = I_3 = 1$, as all eigenvalues are equal to 1. All three measures are between 0 and 1. Table 1 presents the smallest eigenvalue when $r = 1$ observations are deleted. Clearly the above three measures are all identical in this case. Five smallest values correspond to observation numbers 17, 1, 5, 12, 16 (in that order). Changes in the values of this eigenvalue thereafter are gradual without any appreciable jump. In Table 2 we present the above measures when two observations are deleted simultaneously. Only the first eight cases corresponding smallest eight values of I_1 out of

Table 1. Raw Data And Measure of Influence When One Observation Is Deleted

Sr. No.	Deleted Obs.	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	Smallest Evalue ($I_1 = I_2 = I_3$)
1	17	21.3	66.1	65.4	64.8	68.4	66.4	70.8	0.37140
2	1	19.7	60.5	70.3	65.7	69.3	65.7	67.3	0.43952
3	5	19.7	65.1	65.1	65.5	65.5	61.8	70.9	0.48553
4	12	18.3	63.1	65.2	65.4	66.6	61.7	64.0	0.49003
5	16	19.6	63.5	65.2	63.9	70.0	64.2	64.5	0.54089
6	7	19.8	64.3	67.9	62.4	71.4	63.4	69.4	0.63094
7	6	19.6	65.2	71.1	63.5	66.2	67.3	68.6	0.65730
8	11	18.9	63.3	70.4	63.7	68.2	66.2	68.5	0.66288
9	18	20.1	64.8	70.2	65.3	65.5	63.7	66.9	0.67606
10	10	19.9	63.4	70.3	65.9	69.0	63.7	65.1	0.68173
11	4	19.4	63.4	71.9	60.7	68.0	64.9	67.1	0.68612
12	3	19.4	65.4	65.8	66.2	68.8	64.0	69.4	0.69890
13	13	20.3	64.9	68.8	65.2	70.2	62.4	67.0	0.70645
14	8	19.7	65.3	68.8	61.5	66.0	62.4	67.7	0.71042
15	19	20.2	62.6	68.6	63.7	69.8	66.7	68.0	0.72217
16	14	19.7	63.5	70.3	63.1	64.4	65.1	67.0	0.72642
17	2	19.6	64.9	70.4	62.6	69.6	64.6	66.4	0.75913
18	20	19.2	62.2	67.3	63.6	70.9	63.6	66.7	0.76900
19	9	19.7	64.5	68.7	63.9	68.8	62.3	68.8	0.78393
20	15	19.9	62.0	65.5	64.1	67.7	62.1	66.5	0.80120

Table 2. Various Measures of Influence When Two Observations Are Deleted

Sr. No.	Deleted Obs.		Trace I_1	Determinant I_2	Smallest Evalue I_3
1	1	17	0.40546	0.39672	0.32172
2	5	17	0.42846	0.40497	0.28854
3	12	17	0.43071	0.39983	0.27057
4	16	17	0.45614	0.39751	0.23243
5	1	5	0.46253	0.46151	0.43195
6	1	12	0.46477	0.46200	0.41411
7	5	12	0.48778	0.48205	0.41326
8	1	16	0.49020	0.47635	0.37450

Table 3. Various Measures of Influence When Three Observations Are Deleted

Sr. No.	Deleted Obs.			Trace I_1	Determinant I_2	Smallest Evalue I_3
1	1	5	17	0.43215	0.41111	0.27055
2	1	12	17	0.43365	0.40753	0.26118
3	5	12	17	0.44898	0.41290	0.24847
4	1	16	17	0.45060	0.38924	0.17568
5	5	16	17	0.46594	0.39991	0.23234
6	12	16	17	0.46744	0.39432	0.23237
7	1	5	12	0.47169	0.46590	0.40245
8	1	7	17	0.48062	0.46183	0.31652

Table 4. Various Measures of Influence When Four Observations Are Deleted

Sr. No.	Deleted Obs.				Trace I_1	Determinant I_2	Smallest Evalue I_3
1	1	5	12	17	0.44662	0.41546	0.23988
2	1	5	16	17	0.45933	0.39217	0.17364
3	1	12	16	17	0.46046	0.38653	0.17171
4	5	12	16	17	0.47196	0.39673	0.23233
5	1	5	7	17	0.48185	0.45378	0.25792
6	1	7	12	17	0.48297	0.44657	0.23879
7	1	5	6	17	0.48844	0.45701	0.25245
8	1	5	12	16	0.48899	0.43391	0.20803

Table 5. Various Measures of Influence When Five Observations Are Deleted

Sr.No.	Deleted Obs.					Trace I_1	Determinant I_2	Smallest Evalue I_3
1	1	5	12	16	17	0.46547	0.38914	0.16780
2	1	5	7	12	17	0.48348	0.44328	0.21488
3	1	5	6	12	17	0.48875	0.45064	0.23182
4	1	5	11	12	17	0.48987	0.42833	0.17078
5	1	5	12	17	18	0.49251	0.45255	0.23939
6	1	5	10	12	17	0.49364	0.42552	0.15227
7	1	5	7	16	17	0.49365	0.42827	0.17100
8	1	4	5	12	17	0.49452	0.44679	0.22697

$^{20}C_2 = 190$ are presented here. It is evident that two of the above five observations are involved in all eight scenarios and observation numbers 17 and 1 appear to be the most significant players. Although the measures I_2 and I_3 do not organize these eight scenarios in the same order, their values are also towards the lower spectrum. Similar phenomenon is observed when three observations are deleted. Among the lowest eight scenarios (with respect to I_1); seven consist of three observations out of the set of {17; 1; 5; 12; 16}: Again influences of observation number 17 and 1 are evident as most cases correspond to their deletion as part of it.

The eight scenarios of Table 4 contain all $^5C_4 = 5$ cases of four observation deletions, when these four observations are chosen from the set {17; 1; 5; 12; 16}: This again confirms what we have seen in Tables 1-3. Finally, in Table 5, five observation deletions are considered and true to our conviction so far, the smallest I_1 corresponds to the deletion of the above set of five and this value is much smaller than other values of I_1: The value of I_2 is among the smaller values as well. However, I_3 for this set is not small. As one expects, I_3 being based on a single eigenvalue, may not be the most effective measure. After all, it represents influence only in a certain principal axis. We recommend to use all three measures to assess the influence of the subset of observations.

It must be pointed out that Graybill and Iyer's analysis identifies observations with number 5, 12, 16, 17 as potentially influential but misses out the observation number 1, which appears to be the second most influential observation after number 17. This again reinforces the importance of assessing the influence of a subset of observations rather than doing that one observation at a time and also the value of the methodology described above.

We have included, as an appendix, a short SAS code which performs an exhaustive search for the sets of most influential five observations. In fact, this code was used to generate our Table 5. It must be pointed out that even for moderately sized data, the number of subsets of observations of size $r > 1$ can be prohibitively large. For example, with $n = 20$ and $r = 5$; the total number of subsets is $^{20}C_5 = 15504$. Thus instead of an exhaustive search for $r > 1$; (even though our searches for the above example were exhaustive) one should first identify and narrow down potential

influential observations by calculating their individual (r = 1) influence. Clearly, in view of Theorem 2, for r = 1 case, the computations (of emphasis) are especially simple.

However, it turns out that computational and algorithmic burden can be greatly simplified and in view of this simplification, we can develop some guidelines as to how we can identify influential observations. Specifically, let G be as defined in (1) for the last r observations and let us define G_i as the matrix in (1) where only i^{th} observation (that is, row) from matrix X; $i = n - r + 1, \ldots, n$, say, x'_i is deleted. Also let the corresponding smallest eigenvalue of G_i be $\delta_{p,i}$. Accordingly we have,

$$r \cdot I_1 = tr(G) - (p-r) = p - tr[X_2(X'X)^{-1}X'_2] - (p-r)$$

$$= r - tr[X_2(X'X)^{-1}X'_2] = r - \sum_{i=n-r+1}^{n} x'_i(X'_iX)^{-1}x_i$$

$$= r - \sum_{i=n-r+1}^{n}(1-\delta_{p,i}) = \sum_{i=n-r+1}^{n}\delta_{p,i} \qquad (3)$$

Thus I_1 is nothing but the average of smallest eigenvalues corresponding to the cases when observations are deleted, one at a time. Thus, I_1 is linear in the sense that for two disjoint sets of observations A_1 and A_2 of sizes r_1 ans r_2 respectively, $(r_1 + r_2).I^{A_1 \cup A_2} = r_1.I_1^{A_1} + r_2.I_1^{A_2}$. Of course, the choice of last r observations in all the above is taken to mean any r observations since the rows of the data can always be rearranged. What this means is that I_1 can be easily and more directly calculated, for the deletion of any set of r observations, as the average of corresponding smallest eigenvalues and hence I_1 values reported in Tables 2-5 can be derived directly from Table 1.

Above simplifies the identification of influential observations greatly. As asked by a referee, it also helps us come up with a strategy to wisely consider various measures to identify the influential observations. Clearly $r \leq p$. Thus, upon generation of Table 1, as the first step, we choose the (at most) p observations corresponding to smallest eigenvalues using the measure I_3 (Which would be identical to other measures as well since initially r = 1; Further, in view of (2), only one matrix inversion of a $p \times p$ matrix is required for this calculation). This forms our initial candidate set for influence, observations from which may now be evaluated using other measures. For example, for our Mountain data set, this set consists of observation numbers {17,1,5,12,16,7,6}. A total of $2^7 - 1 = 127$ values of I_1 corresponding to various deletion choices should be calculated. As a second step, we may choose a certain number of deletion possibilities out of these 127, using the measure I_1, which themselves can directly be calculated using (3). As the final step, for these narrowed down choices, I_2 and I_3 can now be calculated and looking at all three measures simultaneously, a decision about which observations are collectively influential can be made. This reduction of effort in examining these measures makes the task much more efficient and manageable. This is especially important when n and/or p are very large resulting in a very large number of deletion possibilities.

In Table 6, we illustrate the above approach for our candidate set consisting of, for the sake of brevity, only four observations namely, {17,1,5,12} corresponding the lowest values of I_3 (Doing so for a set of p = 7 observations will result in $2^7 - 1 = 127$ possibilities, which are too cumbersome to present/interpret in a table and too distracting for the main discussion with little extra gain in insight; However, calculations similar to those done in Tables 1-5 for $r = p = 7$ deleted observations (table not shown here) show that in the twelve most suspect sets of seven influential points, these four always appear). A total of $2^4 - 1 = 15$ deletion possibilities are then explored by computing I_1 and I_2. Table 6 is arranged in the ordered values of I_1 from smallest to largest. In fact, same was also done by using I_2 as well and it resulted in more or less the same order with rankings occasionally interchanged by one place but with very insignificant differences between the actual values. The observations numbered as 17 and 1 seem to stand out from the remaining as influential. We may remark that we do observe certain other choices which give smaller values of I_3 but as we pointed out earlier, it represents influence only in the direction of a certain principal axis and hence should not be overemphasized in interpretations.

One of the referees asked if there is a way to define a minimum cut-off point for the smallest eigenvalues. Since the approach is exploratory in nature with no distributional assumptions made whatsoever, the answer, at least in strict sense, is no. However, as pointed out earlier, for r = 1 the emphasis measure $x'(X'X)^{-1}x = 1 - \delta_p$ is similar in expression (although not in interpretation) to the leverage

Table 6. Various Measures of Influence (Ordered by I_1)

Sr. No.	Deleted Obs.	Trace I_1	Determinant I_2	Smallest Evalue I_3
1	{17}	0.37140	0.37140	0.37140
2	{1 17}	0.40546	0.39672	0.32172
3	{5 17}	0.42846	0.40497	0.28854
4	{12 17}	0.43071	0.39983	0.27057
5	{1 5 17}	0.43215	0.41111	0.27055
6	{1 12 17}	0.43365	0.40753	0.26118
7	{1}	0.43952	0.43952	0.43952
8	{1 5 12 17}	0.44662	0.41546	0.23988
9	{5 12 17}	0.44898	0.41290	0.24847
10	{1 5}	0.46253	0.46151	0.43195
11	{1 12}	0.46477	0.46200	0.41411
12	{1 5 12}	0.47169	0.46590	0.40245
13	{5}	0.48553	0.48553	0.48553
14	{5 12}	0.48778	0.48205	0.41326
15	{12}	0.49003	0.49003	0.49003

value. Thus, following the conventional wisdom (Kutner, Nachtsheim and Neter (2004), p. 398) we recommend that for $r = 1$ the smallest eigenvalue less than $\left(1-\frac{2p}{n}\right)$ should be viewed as an indication of possible influence. For our data, this value is 0.30. Another recommendation especially for small datasets (such as ours) is to view moderate influence if δ_p is between 0.5 and 0.8 and strong influence if δ_p is less than 0.5 (In fact, this was our another rationale for choosing {17,1,5,12} as our initial candidate set in the previous paragraph). Of course, as illustrated earlier, the existence of gap between values is another feature one should observe in identifying the influence (and we do observe such a gap after candidate set {17,1} for our analysis. See Table 6).

Often for large datasets, data cleaning and processing is an essential step before any modeling issue is undertaken. This step must necessarily be free from any distributional or model assumptions. Our approach provides an efficient and effective way to do so. However, one should realize that since an essential requirement is that r be less than or equal to p, influential observations must be identified in several iterations - at most p at a time.

ACKNOWLEDGMENT

I wish to thank two referees for their helpful suggestions resulting in the improvements in the original draft.

REFERENCES

Cook, R.D. and Weisberg, S. (1982). "Residuals and influence in regression." London: Chapman and Hall.

Graybill, F.A. and Iyer, H.K. (1994). "Regression analysis, concepts and applications." Belmont, CA: Duxbury Press.

Khattree, R. (2015). "Antieigenvalues, multicollinearity and influence: A revisit to regression diagnostics." Preprint.

Kutner M.H., Nachtsheim, C.J. and Neter, J. (2004). " Applied linear regression model." New York, NY: McGraw Hill-Irwin.

Wang, S.G. and Nyquist, H. (1991). "Effects on the eigenstructure of a data matrix when deleting an observation." Computational Statistics and Data Analysis, **11**, 179-188.

Appendix: A Representative SAS Code for Detecting Influential Observations

Here we present a short SAS program to detect a set of five most influential observations for the data set used in Section 3. The eight most influential sets of five observations are printed here. Program can be easily modified for any other data set and for any other number of influential observations.

```
options nosource nonotes ;
proc datasets; delete myresults;;
run;
data mountain;
input y x1-x7;datalines;
 67.2  19.7  60.5  70.3  65.7  69.3  65.7  67.3
 69.1  19.6  64.9  70.4  62.6  69.6  64.6  66.4
 67.0  19.4  65.4  65.8  66.2  68.8  64.0  69.4
 72.4  19.4  63.4  71.9  60.7  68.0  64.9  67.1
 63.6  19.7  65.1  65.1  65.5  65.5  61.8  70.9
 72.7  19.6  65.2  71.1  63.5  66.2  67.3  68.6
 68.5  19.8  64.3  67.9  62.4  71.4  63.4  69.4
 69.7  19.7  65.3  68.8  61.5  66.0  62.4  67.7
 68.4  19.7  64.5  68.7  63.9  68.8  62.3  68.8
 70.4  19.9  63.4  70.3  65.9  69.0  63.7  65.1
 67.5  18.9  63.3  70.4  63.7  68.2  66.2  68.5
 73.3  18.3  63.1  65.2  65.4  66.6  61.7  64.0
 70.0  20.3  64.9  68.8  65.2  70.2  62.4  67.0
 69.8  19.7  63.5  70.3  63.1  64.4  65.1  67.0
 63.6  19.9  62.0  65.5  64.1  67.7  62.1  66.5
 64.3  19.6  63.5  65.2  63.9  70.0  64.2  64.5
 68.5  21.3  66.1  65.4  64.8  68.4  66.4  70.8
 70.5  20.1  64.8  70.2  65.3  65.5  63.7  66.9
 68.1  20.2  62.6  68.6  63.7  69.8  66.7  68.0
 66.1  19.2  62.2  67.3  63.6  70.9  63.6  66.7
;
```

```
data mountain; set mountain; sr = _n_;drop y; run;
%macro tryout;
%do i = 1 %to 20;
%do j = &i+1 %to 20;
%do k = &j+1 %to 20;
%do m = &k+1 %to 20;
%do s = &m+1 %to 20;
data mountain2; set mountain;
if (sr ˜in (&i &j &k &m &s));
run;

proc iml;
n = 20;
p = 7;
r =5;
use mountain;
read all var {x1 x2 x3 x4 x5 x6 x7} into X;

use mountain2;
read all var {x1 x2 x3 x4 x5 x6 x7} into T_1; A = (X)‘*X;
Ainv = inv(A);
B_1 = (T_1)‘*T_1;
U_1 =root(B_1);
G_1 = U_1*Ainv*(U_1)‘;
eigenv_1 = eigval(G_1);
sum = (trace(G_1) -(p-r))/r;;
deter = (det(G_1))**(1/r);
e5 = eigenv_1[7];

result =J(1,8,1);
result[4] =sum;
result[5] = deter;
result[6] = e5;
result[1] =&i;
result[2] =&j;
result[3] =&k;
result[7]  =&m;
result[8] =&s;
create mydata from result;append from result;close mydata;
quit;

proc datasets nolist;
      append base=myresults data=mydata force; run;
%end;
%end;
%end;
%end;
%end;
run;
%mend tryout;
%tryout;

title “Check the Set of 5 Observations”;
data myresults; set myresults;
sum = col4;deter = col5; smallest = col6; drop col4 col5 col6;
run;

proc sort data = myresults; by sum;run;
proc print data = myresults(obs = 8);
run;
```