数理解析研究所講究録
第231巻 1975年 121-130

多変量解析における変数の選択について

岡山大 工　大崎 紘一

1. 緒言　種々の状況のもとで、多次元のデータが測定される。このデータをもとに、次におこると予想される状況を予測するため、あるいは他の状況との比較をするために、多変量解析に含まれる各種の方法が用いられる。ところで、多次元のデータのなかにも、目的に対してきわめて有効なものもあるし、反対にほとんど有効でないものもある。

そこで、多変量解析のうちで、重回帰分析、2つの多次元正規分布の平均値の比較、そして線形判別関数を利用する際の変数の重要性について統一的に考えることにする。

2 変数の効果

1)重回帰分析　回帰方程式における独立変数の効果は、回帰直線からの偏差平方和に含まれている。

$$S_y^2=\{\sum_{i=1}^{n}(y_i-\bar{y})^2-d'S^{-1}d\}/(n-k-1) \quad \cdots\cdots\cdots (1)$$

d : $k\times 1$ vector,　S : $k\times k$ matrix.

ロ) 2つの多次元正規分布の平均値間の距離　等分散であるという仮定の下で，変数の効果は，次に示す距離に含まれる。

$$D = d' S^{-1} d \quad \cdots\cdots\cdots (2)$$

d ; $k \times 1$ 平均値差ベクトル，S ; $k \times k$ 分散又分散行列

ここで，D は，マハラノビスの距離である。

ハ) 線形判別関数　2群を判別するための判別関数における変数の効果は，誤まって判別する確率に含まれる。

$$P = Pr\{ z > \sqrt{D/2} \} \quad \cdots\cdots\cdots (3)$$

ただし，$Pr\{z > \alpha\}$ は，$N(0,1)$ による確率

(1)，(2)，(3) いずれにおいても，k 個の変数の効果は，全体として，$d' S^{-1} d$ に含まれている。そこで，各変数の効果あるいは変数の組の効果を $d' S^{-1} d$ から分離し，これらの方法を使用する際に有効な変数を選ぶ方法を次に述べる。

3　変数の選択方法

$d = (d_1, d_2, \cdots\cdots, d_k)'$，$S = (s_{ij})$ とおけば，

(1) では，
$$\begin{cases} d_i = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(y_j - \bar{y}) \\ s_{ij} = \sum_{\ell=1}^{n} (x_{i\ell} - \bar{x}_i)(x_{j\ell} - \bar{x}_j) \end{cases}$$

そして，(2)，(3) では，

$$d_i = \bar{x}_{1i} - \bar{x}_{2i}$$

$$s_{ij}=\sum_{p=1}^{2}\sum_{l=1}^{n_p}(x_{pil}-\bar{x}_{pi})(x_{pjl}-\bar{x}_{pj})/(n_1+n_2-2)$$

である。さらに、 $r_{ij}=s_{ij}/\sqrt{s_{ii}s_{jj}}$, $\delta_i=d_i/\sqrt{s_{ii}}$

とおき、 $R=(r_{ij})$, $\delta=(\delta_1,\delta_2,\cdots\cdots,\delta_k)'$ とおけば、

$$d's^{-1}d=\delta'R^{-1}\delta=\delta_{i_1}^2+(\delta_{i_2}-r_{i_1i_2}\delta_{i_1})^2/(1-r_{i_1,i_2}^2)+\cdots\cdots$$

$$+a_{i_ji_j}^{j}\Big(\delta_{i_j}+\sum_{l=1}^{j-1}\delta_{i_l}a_{i_ji_l}^{j}/a_{i_ji_j}^{j}\Big)^2+\cdots\cdots \quad \cdots\cdots(4)$$

$$+a_{i_ki_k}^{k}\Big(\delta_{i_k}+\sum_{l=1}^{k-1}\delta_{i_l}a_{i_ki_l}^{k}/a_{i_ki_k}^{k}\Big)^2$$

但し、 $a_{i_ji_l}^{j}$ は、変数 x_{i_1} , $x_{i_2}\cdots\cdots x_{i_j}$ の相関行列の逆行列の j 行 l 列の要素

(4)の性質を用いて、変数の選択方法を定式化する。重回帰分析において、N. Draper[1]は、変数の選択方法の名前として、Forward Selection Procedure, Backward Elimination Procedure, そして、All Possible Selection Procedure等を使用しているので、ここでもそれらの名前を用いることにする。

解析に用いられる k 個の変数の集合を

$$C=\{x_1,x_2,\cdots\cdots x_k\}$$

で示す。また、 $D(i_1,i_2,\cdots\cdots,i_m)$ を m 個の変数 $x_{i_1},x_{i_2},\cdots\cdots,x_{i_m}$ から計算した $d's^{-1}d$ の値とする。さらに $E(m,i_l)$ を m 個の変数 $x_{i_1},x_{i_2},\cdots\cdots x_{i_m}$ から変数 x_{i_l} を除き、残り $m-1$ 個の変数から計算した $d's^{-1}d$ の値とする。

1) Forward Selection Procedure (FSP)

$$\max_{i_m} D(I_1, I_2, \cdots, I_{m-1}, i_m),\ x_{i_m} \in C - \{x_{I_1}, x_{I_2}, \cdots, x_{I_{m-1}}\}$$

を満たす変数を x_{I_m} とする。

$m = 1, 2, \cdots, k$

ロ) Backward Elimination Procedure (BEP)

$$\max_{i_\ell} E(m, i_\ell),\ x_{i_\ell} \in C - \{x_{J_k}, x_{J_{k-1}}, \cdots, x_{J_{k-m-1}}\}$$

を満たす変数を x_{J_m} とする。

$m = k, k-1, \cdots, 2, 1$

ハ) All Possible Selection Procedure (APSP)

$$\max_{(i_1, i_2, \cdots i_m)} D(i_1, i_2, \cdots, i_m)$$

$$\{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\} \subset C \quad (i_j \neq i_\ell)$$

を満たす変数の組を $\{x_{K_1^m}, x_{K_2^m}, \cdots, x_{K_m^m}\}$ とする。

$m = 1, 2, \cdots, k$

FSP と APSP では，効率の高い変数から順に選ばれるが，BEP では，効率の低い変数から順に選ばれることになる。それゆえ，BEP の結果を逆にして $x_{J_1}, x_{J_2}, \cdots, x_{J_m}, \cdots x_{J_k}$ として用いれば，FSP あるいは APSP の結果と比較するのに便利である。これら3つの選択方法の間には，次の関係が成立す

3.

定　理　1

All Possible Selection Procedure において

$$\{x_{K_1^i}, x_{K_2^i}, \cdots\cdots, x_{K_i^i}\} \subset \{x_{K_1^{i+1}}, x_{K_2^{i+1}}, \cdots\cdots, x_{K_{i+1}^{i+1}}\} \cdots\cdots (5)$$

$$i = 1, 2, \cdots\cdots, k-1$$

が成立するならば，Forward Selection Procedure と Backward Elimination Procedure の結果は一致する。

すなわち，$x_{I_i} = x_{J_i}$，$i = 1, 2, \cdots\cdots, k$

証明

1) APSP と FSP との一致

$i = 1$ では，APSP と FSP の性質から $x_{I_1} = x_{K_1}$。

$i = 2$ では，(5) と $i = 1$ の結果より，

$$x_{I_1} = x_{K_1^1} = x_{K_1^2}$$

としても一般性を失なわない。そして，

APSP は，

$$\max_{(i_1, i_2)} D(i_1, i_2) = \max_{(I_1, i_2)} D(I_1, i_2)$$

となる。よって，

$$x_{I_2} = x_{K_2^2}$$

帰納法によって，$\{x_{I_1}, x_{I_2}, \cdots\cdots, x_{I_m}\} = \{x_{K_1^m}, x_{K_2^m}, \cdots\cdots, x_{K_m^m}\}$

となる。そして，$x_{K_j^{m+1}} = x_{K_j^m} = x_{I_j}$，$j = 1, 2, \cdots\cdots m-1$

とおけば，$x_{I_m} = x_{K_m^m}$ となる。

2) APSP と BEP および FSP との一致

$i=k-1$ では，APSP と BEP の性質から

$$\max_{(i_1,\cdots i_{k-1})} D(i_1, i_2, \cdots, i_{k-1}) = \max E(k, i_k)$$

$$\{x_{i_1}, x_{i_2}, \cdots x_{i_{k-1}}\} \subset C^{i_k}, \quad \{x_{i_k}\} \subset C$$

であり，$\{x_{K_1^{k-1}}, x_{K_2^{k-1}}, \cdots, x_{K_{k-1}^{k-1}}\} = C - \{x_{J_k}\}$

となる。

APSP と FSP との一致から

$$x_{I_k} = x_{J_k} .$$

以下帰納法によって，$\{x_{K_1^{i+1}}, x_{K_2^{i+1}} \cdots, x_{K_{i+1}^{i+1}}\}$

$$= C - \{x_{J_k}, \cdots, x_{J_{k-i-1}}\}$$

の関係から

$$x_{I_i} = x_{J_i}$$

が成立する。

定　理　2

Forward Selection Procedure において，

$$D(I_1, I_2, \cdots, I_m) \geq D(i_1, i_2, \cdots i_{k-1}) \qquad \cdots (6)$$

$$\{x_{i_1}, x_{i_2}, \cdots, x_{i_{k-1}}\} = C - \{x_{I_\ell}\}$$

$$\ell = 1, 2, \cdots, m-1$$

ならば，$\{x_{I_1}, x_{I_2}, \cdots, x_{I_m}\} = \{x_{K_1^m}, x_{K_2^m}, \cdots x_{K_m^m}\}$

$$i = 2, 3, \cdots k-1$$

が成立する。

証　明

$C-\{x_{I_\ell}\}=G_\ell$, $\ell=1,2,\cdots m-1$ とおく。

また，$G_1 \wedge G_2 \wedge \cdots \wedge G_{m-1}=G_0$ とおく。

G_0 に含まれる変数の数は，$n-i+1$ 個となる。

$$C-G_0=\{x_{I_1}, x_{I_2}, \cdots x_{I_{m-1}}\} \quad \cdots\cdots\cdots (7)$$

となる。

${}_kC_m$ は，G_0 と $C-G_0$ とに含まれる変数を区別すると，

$${}_kC_m=\sum_{j=0}^{m-1} {}_{k-m+1}C_{m-j} \times {}_{m-1}C_j \quad \cdots\cdots\cdots (8)$$

となる。

${}_{k-m+1}C_{m-j} \times {}_{m-1}C_j$ は，G_0 から $m-j$ 個，$C-G_0$ から j 個の変数を選び合わせて m 個の変数の組を作る場合の組み合せである。

ところで，(7) より，$C-G_0$ の任意の部分集合は，G_ℓ の少なくとも一つに含まれる。以上は，$j=1,2,\cdots,m-2$ において成立する。

${}_{k-m+1}C_1 \times {}_{m-1}C_{m-1}$ は，FSP における組み合せである。

また，(4) より，$D(i_1, i_2, \cdots, i_p) \geq D(j_1, j_2, \cdots, j_q)$

$\{x_{j_1}, x_{j_2} \cdots x_{j_q}\} \subset \{x_{i_1}, x_{i_2}, \cdots, x_{i_p}\}$

$q \leq p$

が成立する。

よって，(6)，(8)，(9)より，FSPとAPSPとは一致する。

定　理　3

Backward Elimination Procedureにおいて，

$$E(k-m, J_{k-m}) \geq D(J_k, J_{k-1}, \cdots, J_{k-m}, j_1, j_2, \cdots j_{k-m-2})$$

$$\{x_{j1}, x_{j2}, \cdots x_{j_{k-m-2}}\} \subset C-\{x_{J_k}, x_{J_{k-1}}, \cdots, x_{J_{k-m}}\}$$

ならば，$C-\{x_{J_k}, \cdots, x_{J_{k-m}}\} = \{x_{K_1^{k-m-1}}, x_{K_2^{k-m-1}}, \cdots\cdots, x_{K_{k-m-1}^{k-m-1}}\}$

$$m = 0, 1, 2, \cdots, k-2$$

証明は，定理2とほぼ同様に行なうことができる。

4. APSPの計算アルゴリズム

APSPにおいて，すべての組み合せについて計算を行なうと，2^k-1回行なわなければならない。それゆえ，工藤，重水[2),3)]等は，計算時間を短かくするためのAPSPの計算アルゴリズムについて述べている。

ここでは，中間結果を蓄える領域を2箇所にし，組み合せを作る回数を少なくした計算アルゴリズムについて述べる。

k個の変数のうちで，m個の変数を用いる場合に，残りの$k-m$個の変数についても同時に考えることにする。m個の

変数を $\{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\}$ とおくと，残りの $k-m$ 個の変数は

$$\{x_{j_1}, x_{j_2}, \cdots, x_{j_{k-m}}\} = C - \{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\} \quad \cdots\cdots (6)$$

となる。(6) において，右辺の m 個の変数を k 個のうちから任意に選ぶ選び方は，${}_kC_m$ 通りある。右辺の m 個の変数に対応する δ，R の要素を一箇所に，残りの $k-m$ 個のものを他の箇所に集める。そして

$$\max_{(i_1, i_2, \cdots i_m)} D(i_1, i_2, \cdots i_m), \quad \{x_{i_1}, x_{i_2}, \cdots, x_{i_m}\} \subset C$$

$$\max_{(j_1, j_2, \cdots j_{k-m})} D(j_1, j_2, \cdots, j_{k-m}), \{x_{j_1}, x_{j_2}, \cdots, x_{j_{k-m}}\} = C - \{x_{i_1}, x_{i_2}, \cdots x_{i_m}\}$$

を考える。二項定理より、${}_kC_m = {}_kC_{k-m}$ となることから，${}_kC_m$ 回の変数の組み合せから

$$\{x_{K_1^m}, x_{K_2^m}, \cdots, x_{K_m^m}\}, \{x_{K_1^{k-m}}, x_{K_2^{k-m}}, \cdots\cdots, x_{K_{k-m}^{k-m}}\}$$
$$m = 1, 2, \cdots\cdots, k'$$

を同時に求めることができる。但し，k が奇数の場合は，$k' = (k-1)/2$，偶数の場合は，$k' = k/2$ である。

参考文献

1). Draper, N. and Smith, H. (1966), Applied Regression Analysis, John Wiley Sons, Inc.

2) Kudō, A. (1963). Mahalanobis measure as a criterion for the selection of variables. Memoirs of the Faculty of Science, Kyushu University, Series A.

17, 63–75

3) Tarumi, T and Kudô, A. (1974). An algorithm related to all possible regression and discriminant analysis, Journ. Japan Statist. Soc. 4.2, 47–56