

ON GENERALIZED COEFFICIENT OF DETERMINATION

柳井 晴夫 (東大医)

§1 はじめに

電子計算機の飛躍的な発展によつて、政治学、経済学、社会学、心理学、医学などのいわゆる行動科学 (behavioral science) と呼ばれる分野において、多変量解析の各種の技法を用いた人間の様々な行動現象に対する計量的な分析が数多く実施されるようになってきた。

ところで、周知のように多変量解析の手法には因子分析、重回帰分析、正準相関分析、判別分析、数量化理論など数多くあるが (Anderson (1958), 林他 (1970), 奥野他 (1971), 竹内・柳井 (1972)), これらの手法はいずれも異なった研究者によって開発されたもので、統一的立場にたつて、これらの手法の相互関連を明らかにしたものは見当らない。そこで、筆者は多変量解析の幾つかの手法を、線型空間からその空間に含まれる部分空間への射影子という観点から記述的 (descriptive) に統一する立場にたつて、二組の変数群の最大の関連の程度を示す一般化決定係数という新しい概念を導入し、多変量解析の各種の技

法のうち特に外的基準がある場合の技法の相互関連を明らかにすることを目指す。(竹内, 柳井 (1972), 本井 (1974))

§2 射影子についての性質

N 次元実ベクトル空間 V_N に含まれる p 個の 1 次独立なベクトル集合 $f_1, f_2, \dots, f_p$ を (N, p) 型行列

$$(1) \quad F = (f_1, f_2, \dots, f_p)$$

によって表わし, これらの p 個のベクトルによって生成される部分空間を W_F , V_N における W_F の直交補空間を $W_F^\perp$ とすれば, $W_F, W_F^\perp$ 上の射影子 $\pi_F, \pi_F^\perp$ は次のように表わされる。

$$(2) \quad \pi_F = F(F'F)^{-1}F' \quad \pi_F^\perp = I_N - \pi_F$$

(ただし $-$ は $AA^{-1} = A$ かつ $(A^{-1})' = A^{-1}$ を満たす一般化逆行列)

ここで, V_N に含まれるもう一つの部分空間を q 個のベクトル集合 $G = (g_1, g_2, \dots, g_q)$ で生成され W_G , この空間への射影子を π_G とするとき, 次の 6 つの関係が成立する。

$$(3) \quad W_F \perp W_G \Leftrightarrow \pi_F \pi_G = \pi_G \pi_F = 0$$

$$(4) \quad W_F \supset W_G \Leftrightarrow \pi_F \pi_G = \pi_G \pi_F = \pi_G$$

$$(5) \quad \|\pi_F y\|^2 \leq \|y\|^2 \quad \text{for any } y \in V_N$$

$$(6) \quad \pi_{F \cup G} = (F, G) \begin{pmatrix} F'F & F'G \\ G'F & G'G \end{pmatrix}^{-1} \begin{pmatrix} F' \\ G' \end{pmatrix}$$

$$= \pi_F + \pi_{G/F} \quad \left(\text{ただし } \pi_{G/F} = \pi_F^\perp G (G' \pi_F^\perp G)^{-1} G' \pi_F^\perp \right)$$

または $= \pi_G + \pi_{F/H} \quad \left(\pi_{F/H} = \pi_G^\perp F (F' \pi_G^\perp F)^{-1} F' \pi_G^\perp \right)$

(SASAKI 統計学 (1972) p.28)

(7) $\bar{W}_F \bar{W}_G$ のとき $\Pi_F \Pi_G = \Pi_F$ (6)を一般化連行列入定義

(8) $\bar{W}_F \bar{W}_G \sim \dim(\bar{W}_F \bar{W}_G) = \dim \bar{W}_F - \dim \bar{W}_G = r \sim \text{rank}(\Pi_G \tilde{F})$
 $= r$ とおきよければ (N, r) 型行列を $\tilde{F}$ とおくと,

$$\Pi_F = \Pi_G + \Pi_G \tilde{F} (\tilde{F}' \Pi_G \tilde{F})^{-1} \tilde{F}' \Pi_G$$

上記の性質を用いて、次に一般化決定係数の基本となる定理を示そう。(柳井(1979))

定理1 $d_{F,G}^2 = \text{tr}(\Pi_F \Pi_G) \leq r$

ただし $r = \text{Min}(\text{rank} \Pi_F, \text{rank} \Pi_G) = \text{Min}(p, r)$

(証明) $y = \Pi_G x$ とおくと (5) を用いると $\|\Pi_F \Pi_G x\|^2 \leq \|\Pi_G x\|^2$
 となり $x'(\Pi_G \Pi_F \Pi_G - \Pi_G)x \leq 0$ x は任意の N -次元ベ
 クトルであるから $-(\Pi_G \Pi_F \Pi_G - \Pi_G) = \Pi_G - \Pi_G \Pi_F \Pi_G$ は非負
 定値行列となる。したがって $\text{tr}(\Pi_G) \geq \text{tr}(\Pi_G \Pi_F \Pi_G) = \text{tr}(\Pi_G \Pi_F)$
 同様に $\text{tr}(\Pi_F) \geq \text{tr}(\Pi_G \Pi_F)$ さらに $\text{rank} \Pi_F = \text{tr}(\Pi_F)$
 より定理が証明される。(証明終)

この定理で $F=(f)$, $G=(g)$ とおくとこれは Schwarz の不等式に一致する。よって、上記の定理は Schwarz の不等式を多次元の場合に一般化したものになり、2113 ことがわかる。

93 一般化決定係数による 44 変量解析の各種技法の統一
 的表現

二組の変数群の測定値を $F=(f_{ij})$, $G=(g_{ij})$ と表わ

した場合、各列ベクトルに含まれる成分の和が0になるように変換するには、ベクトル $\mathbf{1}_N' = \overbrace{(1, 1, \dots, 1)}^{N\text{個の1}}$ で生成される部分空間 $\mathcal{V}_N$ への射影子を $\pi_N = \mathbf{1}_N (\mathbf{1}_N' \mathbf{1}_N)^{-1} \mathbf{1}_N' = \frac{1}{N} \mathbf{1}_N \mathbf{1}_N'$, その直交補空間 $\mathcal{V}_N^\perp$ への射影子を $\pi_N^\perp = \mathbf{I}_N - \pi_N$ と定義したとき、

$$(9) \quad \tilde{F} = \pi_N^\perp F \quad \tilde{G} = \pi_N^\perp G$$

と基準化すればよい。このとき、

$$(10) \quad D_{F, G}^2 = d_{\tilde{F}, \tilde{G}}^2 / n$$

と定義すると定理1より $0 \leq D_{F, G}^2 \leq 1$ となる。ここで

(10)式は F と G に含まれる二組の変数群全体の相互関連の強さを示す指標となるもので、一般化決定係数 (generalized coefficient of determination) と呼ぶ。 (FAP#(1994))

多変量解析の各種の技法は F, G に含まれる変数が (a) 間隔尺度か名義尺度か (b) 変数の個数が二つ以上であるか一つであるか、という二つの組み合わせにより、2分類することが可能である。(表参照)

3.1 ともに間隔尺度の場合

ベクトルは変数が一つで、行列は二つ以上の変数があることを意味するものである。一例として、重相関係数の平方は、

$$\begin{aligned} R_{x, y}^2 &= \text{tr}(\pi_x \pi_y) = \text{tr}(\pi_x y (y' y)^{-1} y') = y' \pi_x y (y' y)^{-1} \\ &= \|\pi_x y\|^2 / \|y\|^2 = C_{yx} C_{xx}^{-1} C_{xy} / \sigma_y^2 = R_{yx} R_{xx}^{-1} R_{xy} \end{aligned}$$

と表現される。

3.2 一方が名義尺度の場合

N 人の被験者が m 個の水準をもつ因子 A によつて割り付けられるとき $G_A = (g_1, g_2, \dots, g_m)$ の成分がすべて1か0で $\|g_i\|^2 = (g_i, g_i) = N_i$, $(g_i, g_j) = 0$ ($i \neq j$) さらに

$$(1) \quad G_A \mathbb{I}_m = \mathbb{I}_N$$

の性質を満たすとき G_A はタミー変数行列とよばれる。 G_A を (9) 式によつて $\tilde{G}_A = \pi_M^\perp G_A$ と基準化すると $\text{rank } \tilde{G}_A = \text{rank } G_A - 1$ となるので、 G_A から任意に1つの列ベクトルを取り除いた行列を G_A とすると、 $\tilde{G}_A$ と $\tilde{G}_A$ は同一の部分空間 $W_{A/M} (= W_A \cap V_M^\perp)$ を生成する。よつて、 $W_{A/M}$ の射影子は (8) の性質から

$$(12) \quad \pi_A = \pi_{A/M} = \pi_{A/M} - \pi_M = \pi_A - \pi_M = \pi_M^\perp \pi_A = \pi_A \pi_M^\perp$$

となる。さらに、次の関係式も成立する。

$$(13) \quad \sum_i \text{tr}(\pi_i \pi_x) = \frac{1}{N} \sum_{i,j} \text{tr}(\pi_i \pi_j \pi_x) = \text{tr}(\bar{G}_{xx}^{-1}(A))$$

3.3 ともに名義尺度の場合

(12) の関係式を用いると、クロス集計における一般化決定係数は

$$D_{A,B}^2 = \text{tr}(\pi_A \pi_B) / r = \text{tr}(\pi_A \pi_M^\perp \pi_B \pi_M^\perp) / r = \text{tr}(SS') / r$$

となる。(ただし $S' = (G_A' G_A)^{-\frac{1}{2}} (G_A' \pi_M^\perp G_B) (G_B' G_B)^{-\frac{1}{2}}$)

こゝで $s_{ij} = (N_{ij} - \frac{1}{N} N_{i.} N_{.j}) / \sqrt{N_{i.} N_{.j}}$ であり $\text{tr}(SS') / r$ が分割表の検定におけるカイニ乗統計量の N 分の $r-1$ になる。

3.4 X の変数群の影嚮を取り除く場合

X, Y より Z の影嚮を取り除いた $x_z^\perp = \pi_z^\perp x$, $y_z^\perp = \pi_z^\perp y$ の

4.1 判別分析における決定係数と2クラス間の距離

判別すべき群が二つで、二群の母分散共分散行列が等しいと仮定できる場合、尤度比によるFisherの判別関数が導かれるが、この方法は記述的には一群に $-\frac{N_2}{N_1+N_2}$ 、二群に $\frac{+N_1}{N_1+N_2}$ を基準変数の測定値として与えれば、重回帰分析に形式的に完全に一致する。ここで、上記の場合に得られる一般化決定係数は $D_c^2 = \text{tr}(\pi \hat{\pi}_2 \pi_x) = \frac{N_1 N_2}{N^2} (\bar{x}_1 - \bar{x}_2)' C_{xx}^{-1} (\bar{x}_1 - \bar{x}_2)$ となる。なお、二群間の2クラス間の距離の決定値は $D_H^2 = (\bar{x}_1 - \bar{x}_2)' C_{eA}^{-1} (\bar{x}_1 - \bar{x}_2) \frac{N_1 N_2 - 2}{N}$ となるから、 D_c^2 と D_H^2 の間には次の関係が成立する。(C_{eA}は総分散)

$$(16) \quad D_H^2 = \frac{D_c^2 (N_1 + N_2 - 2)}{(1 - D_c^2) N}$$

(証明) 重回帰分析における行列方程式を $C_A a = \lambda C_{xx} a$ とおくと、群が二つの場合 $C_A = \frac{N_1 N_2}{N^2} (\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)'$ となる。したがって、

$$\lambda = \text{tr}(C_{xx}^{-1} C_A) = D_c^2 \quad \text{一方} \quad C_{xx} = C_A + C_{eA} \quad \text{より}$$

$$C_A a = \frac{\lambda}{1-\lambda} C_{eA} a \quad \text{--- ②} \quad \text{これより} \quad \frac{\lambda}{1-\lambda} = \text{tr}(C_{eA}^{-1} C_A) = D_H^2 \times \frac{N}{N_1 N_2 - 2}$$

①と②より (16) が導かれる。

(証明終)

上記の関係式より、一般化決定係数 D_c^2 を用いることは、2群の判別分析における変数選択は重回帰分析における変数選択と全く同様である。しかも同一プログラムを用いて行うことが可能である。

$$(例) \quad N_1 = N_2 = 50 \quad D_H^2 = 7.24 \quad (\text{理論的正誤率 } 91.1\%) \quad \text{のとき} \quad D_c = 0.803$$

$$D_H^2 = 6.67 \quad \text{のとき} \quad D_c = 0.791$$

4.2 変数選択と一般化決定係数

Y を基準変数, 説明変数が $X = (X_1, X_2)$ と分離されるとき,

$$\begin{aligned} d_{X \cdot Y}^2 &= \text{tr}(\Pi_X \Pi_Y) = \text{tr}\{(\Pi_{X_1} + \Pi_{X_2/X_1}) \Pi_Y\} = \text{tr}(\Pi_{X_1} \Pi_Y) + \text{tr}(\Pi_{X_2/X_1} \Pi_Y) \\ &= d_{X_1 \cdot Y}^2 + d_{X_2/X_1 \cdot Y}^2 \end{aligned}$$

となる。ここで上式の X = 項は、 X_1 に対し X_2 を追加したことによる増加する決定係数の増分を示すもので、この値が大きくなるような変数を選択すればよい。ここで、 $Y = (y)$, または $Y = (\hat{y}_{12})$ のとき次式が得られる。

$$(17) \quad d_{X_2/X_1 \cdot Y}^2 = (C_{yX_2} - C_{yX_1} C_{X_1X_1}^{-1} C_{X_1X_2}) (C_{X_2X_2} - C_{X_2X_1} C_{X_1X_1}^{-1} C_{X_1X_2})^{-1} (C_{X_2Y} - C_{X_2X_1} C_{X_1X_1}^{-1} C_{X_1Y}) / dy$$

$$(18) \quad d_{X_2/X_1 \cdot \hat{y}_{12}}^2 = (m_{1 \cdot X_2/X_1} - m_{2 \cdot X_2/X_1}) (C_{X_2X_2} - C_{X_2X_1} C_{X_1X_1}^{-1} C_{X_1X_2})^{-1} (m_{1 \cdot X_2/X_1} - m_{2 \cdot X_2/X_1})$$

$$\text{ただし } m_{ij \cdot X_2/X_1} = \bar{X}_{2j} - C_{X_2X_1} C_{X_1X_1}^{-1} \bar{X}_{1j}$$

4.3 多変量解析の各種手法の極値条件

$$V_{xy}^2 = \text{tr}(\Pi_X \Pi_Y) \text{ と表わされることから, } F_X = XW, F_Y = YV$$

とおくことにすると、 $\text{tr}(\Pi_{F_X} \Pi_{F_Y}) \rightarrow \max$ として正準相関分析の固有方程式を導くことができる。その他、 $F_X = XW, G_Y = YV$ と重みベクトルを別次元に拡張することによって、次のような極値条件を導くことができる。

$$(a) \text{ 重回帰分析 } \text{tr}(\Pi_{F_X} \Pi_Y) \quad (b) \text{ 主成分分析 } \text{tr}(X' \Pi_{F_X} X)$$

$$(c) \text{ 変形主成分分析 } \text{tr}(Y' \Pi_{F_X} Y) \quad (d) \text{ 重判別分析 } \text{tr}(\Pi_{F_X} \Pi_A)$$

$$(e) \text{ 正準相関分析 } \text{tr}(\Pi_{F_X} \Pi_{G_Y}) \text{ あるいは } \text{tr}(\Pi_{F_X} \Pi_Y)$$

$$(f) \text{ 一般正準相関分析 } \text{tr}(\Pi_{F_X} \Pi_{G_Y}) + \text{tr}(\Pi_{F_X} \Pi_{H_2}) + \text{tr}(\Pi_{G_Y} \Pi_{H_2})$$

(一般正準相関分析は重判別分析の二上の場合の一解(七上))

表 一般化決定係数による多変量解析各種技法の分類

	分析手法	説明変数		基準変数		一般化決定係数 (第5章)
		名義	間隔	名義	間隔	
ともに 間隔	単回帰分析	/	$\mathcal{R}^{(1)}$	/	$\mathcal{Y}$	$t_r(\pi_x \pi_y) = r_{xy}^2$ 単相関係数
	重回帰分析	/	X	/	$\mathcal{Y}$	$t_r(\pi_x \pi_y) = R_{x \cdot y}^2$ 重相関係数
	正準相関分析	/	X	/	$\mathcal{Y}$	$t_r(\pi_x \pi_y)/r = R_{x \cdot y}^2$ 正準相関係数
一方が 名義尺度 の場合	分散分析	$\tilde{G}_A$	/	/	$\mathcal{Y}$	$t_r(\pi_A \pi_y) = \sigma_y^2 / \sigma_y^2$ 相関比
	共分散分析	$\tilde{G}_A$	X	/	$\mathcal{Y}$	$t_r(\pi_A \pi_y)$ 距離
	判別分析(1)	/	X	$\tilde{g}_i$	/	$t_r(\pi_{\tilde{g}_i} \pi_x) = \frac{N_i}{N - N_i} (\bar{x}_i - \bar{x})' C_{xx}^{-1} (\bar{x}_i - \bar{x})$
	判別分析(2)	/	X	$\tilde{h}_{ij}^{(2)}$	/	$t_r(\pi_{\tilde{h}_{ij}} \pi_x) = \frac{N_i N_j}{N(N_i + N_j)} (\bar{x}_i - \bar{x}_j)' C_{xx}^{-1} (\bar{x}_i - \bar{x}_j)$
	重判別分析	/	X	$\tilde{G}_A$	/	$t_r(\pi_x \pi_A)/r = t_r(C_{xx}^{-1} C_A)/r$
	数量化一類	$\tilde{G}_{(m)}^{(1)}$	/	/	$\mathcal{Y}$	$t_r(\pi_{\tilde{G}_{(m)}} \pi_y)$ } 相関比
ともに 名義	数量化二類	$\tilde{G}_{(m)}^{(2)}$	/	$\tilde{G}_A$	/	$t_r(\pi_{\tilde{G}_{(m)}} \pi_A)/r$ }
	ノルム集計	$\tilde{G}_A$	/	$\tilde{G}_B$	/	$t_r(\pi_A \pi_B)/r = \chi_{AB}^2 / Nv$ (ノルム集計)
	数量化三類 (対称ノルム集計)	$\tilde{G}_{(m)}^{(3)}$	/	$\tilde{G}_{AB}$	/	$t_r(\pi_{\tilde{G}_1} \pi_{\tilde{G}_2})/r$ (")
Xの 影響を 取り除く 場合	偏相関分析	/	$\pi_{1/2}$	/	$\pi_{2/1}$	$t_r(\pi_{1/2} \pi_{2/1})$ 偏相関係数
	偏正準相関分析	/	$\pi_{1/2} X$	/	$\pi_{2/1} Y$	$t_r(\pi_{1/2} \pi_{2/1})/r$ 偏正準相関係数
	偏ノルム集計	$\pi_{1/2} \tilde{G}_A$	/	$\pi_{2/1} \tilde{G}_B$	/	$t_r(\pi_{1/2} \pi_{2/1})/r$ 偏関連係数

注1) $\mathcal{X}, \mathcal{Y}$ は全2成分の平均が0にならうに基準にされる。注2) $h_{ij} = g_i/N_i - g_j/N_j$ $\tilde{h}_{ij} = \pi_{1/2} h_{ij}$ 注3) $G_{(m)} = (G_1, G_2, \dots, G_m)$ G_j は1次元の1次元列注4) $C_{xx} = \frac{1}{N} X'X$ (共分散行列) $C_A = \frac{1}{N} X' \pi_A X$ (組別共分散行列)

注5) Dharmoto & Endo (1973) による。

3.5 一般化決定係数の分布

二変数群 X, Y が母共分散行列 $\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$ をもつ正規母集団からの標本であるとき, $\Sigma_{xy} = 0$ の場合の標本正準相関係数の分布は次式となる。(X が p 個, Y が q 個の変数を含む $q \leq p$)

$$(11) f(v_1, v_2, \dots, v_q) = \frac{\pi^{pq/2} \prod_{i=1}^q \Gamma(\frac{1}{2}(N-i+1))}{\prod_{i=1}^q \Gamma(\frac{1}{2}(N-p-i+1)) \Gamma(\frac{1}{2}(q-i+1)) \Gamma(\frac{1}{2}(p-i+1))} \times \prod_{i=1}^q \frac{v_i^{p-i-1} (1-v_i)^{q-i-1}}{(1-v_i)^{p-i-1}} \prod_{i,j=1}^q (v_i v_j - v_{ij}^2)^{-\frac{1}{2}}$$

定理2 $\tau(q, N) = v_1^2 + v_2^2 + \dots + v_q^2$ とおくと, $N \rightarrow \infty$ の

とき, τ は自由度 pq の χ^2 分布にしたがう。

(略証1) $N \rightarrow \infty$ のとき (11) は $\Sigma = I_N$ をもつ標本共分散行列の固有値の分布に一致し, さらにこの分布における固有値の和が χ^2 分布に従う。

(略証2) $\tau(q, N)$ の標本分布の特性関数は

$$\varphi(t) = \sum_{k=0}^q \frac{(it)^k}{k!} \sum_{k=0}^q \frac{C(p)_k}{C(q)_k (N-q)_k} C_k(I_q) \quad \text{となり, この分布の漸近}$$

分布 $(N \rightarrow \infty)$ は自由度 pq の χ^2 分布の特性関数となる。

(Pillai (1956), James (1964), Sugiyama-Fujikoshi (1968) を参照)

上記の定理は 3.3 に対応するもの2, 分布論的にも一般化決定

係数 $\tau(q, N)$ の分布の特別な場合として, 相関比, 重相関(係数)の標本分布を取扱ふことかできる。

REFERENCE

- Pillai, K.C.S. (1956) Some results useful in multivariate analysis A.M.S. 27
 Anderson, T (1958) Introduction to Multivariate Analysis
 James, A.T. (1964) Distribution of matrix variates and latent roots derived from normal samples, A.M.S. 35, 475-501
 Sugiura, N & Fujikoshi, Y (1968) Asymptotic expansions of the non-null distributions of the likelihood ratio criteria for multivariate linear hypothesis and independence, z A.M.S. 40, 942-952
 林和久他 (1970) 統計学教理と小情報処理 産業図書
 奥野忠一他 (1971) 多変量解析法 日科社
 山内隆, 柳井山壽夫 (1972) 多変量解析の基礎 東洋館
 Okamoto, M & Endo, H (1973) Basic Properties of categorical canonical correlation analysis, J. of Japan Statist. Soc.
 柳井山壽夫 (1974) 一般化決定係数に対する多変量解析の基礎と統計的表現の相関性