多変量解析法（その2）

——外的基準のある場合——

高木 廣文　服部 芳明　柳井 晴夫

公 衆 衛 生

第47巻 第10号 別刷

1983年10月15日 発行

医学書院

講　座

多変量解析法（その2）

——外的基準のある場合——

高木 廣文*　服部 芳明**　柳井 晴夫***

■ はじめに

前回，多変量解析のうち外的基準のない場合の各手法の説明をした．ここでは外的基準のある場合について，いくつかの手法を紹介する．外的基準のある場合も，その基準変数が定量的な場合と定性的な場合に分けて考える必要がある．定量的な場合は，他の変数を用いて基準変数を予測することになる．たとえば高血圧症の患者の食塩摂取量を 10 g 以下にすると，どのくらい血圧は低くなるのか知りたいとする．患者のデータとして，年齢，現在の血圧値，他の食物摂取量など多数の変数を利用できるだろう．このような場合，重回帰分析を用いることができる[1]．さらに重回帰分析を応用して因果関係の推論をする場合には，パス解析法を用いることも可能である[2]．

基準変数が特定の疾患の有無のように定性的な場合，利用できる変数から各ケースが所属する群を識別することになる．そのような方法は判別分析とよばれる[1,3,7]．また判別分析と類似しているが，所属する群を識別するのではなく，その可能性を確率として表現する方法がある．それには多重ロジスティック・モデルを用いる方法が知られている[4]．

説明変数がある特性の有無のような定性的なもののみからなる場合，重回帰分析に対応して，数量化理論Ⅰ類があり，判別分析に対して数量化理論Ⅱ類がある[5]．しかし，ここでは数量化理論に関する説明は省き，重回帰分析と判別分析を中心に解説することにする．

* Takagi, Hirofumi：聖路加看護大学看護学部
** Hattori, Yoshiaki：日本原子力事業株式会社システム解析部
*** Yanai, Haruo：千葉大学文学部行動科学科
別刷請求先：(〒 104)東京都中央区明石町10-1
聖路加看護大学　看護学部　高木　廣文

■ 重回帰分析

重回帰分析は定量的な基準変数の値を，他のいくつかの説明変数を用いて予測するための手法である．また，基準変数に対する説明変数の寄与の程度を調べたりすることにも使われることがある．

ある疾患の重症度のような基準変数を Y，説明変数が X_1, X_2, …, X_p と p 個あるとする．ケース i のデータを y_i, x_{i1}, x_{i2}, …, x_{ip} と表わしたとき，y_i は p 個の説明変数を用いて，

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i \qquad (1)$$

となると仮定する．β_1, β_2, …, β_p は各変数にかける重みであり，偏回帰係数とよばれる．また β_0 は定数項，ε_i はケース i に関する誤差項で平均 0，分散一定の確率変数である．y_i の予測値 $\hat{y}_i$ は

$$\hat{y}_i=b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip} \qquad (2)$$

とかける．ただし，b_0, b_1, …, b_p は β_0, β_1, …, β_p の推定値とする．各係数の推定は以下のようにすればよい．

変数 x_j の平均偏差ベクトルを $\mathbf{x}_j$，予測値 $\hat{y}$ の平均偏差ベクトルを $\mathbf{y}$ とする．さらに，$\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$ とすると，$\mathbf{X}$ は $n\times p$ の大きさをもつ平均偏差データ行列となる（標本数を n とする）．偏回帰係数ベクトルを $\mathbf{b}=(b_1, b_2, \cdots, b_p)'$ とすると，（2）式の重回帰式は

$$\hat{\mathbf{y}}=\mathbf{Xb} \qquad (2)'$$

のようにかける．

実測値 y_i と予測値 $\hat{y}_i$ の差の 2 乗和が最小になるように，各偏回帰係数を求めると（最小二乗法という），

$$\mathbf{b}=C_{XX}^{-1}C_{Xy} \qquad (3)$$

のようになる．ただし，C_{XX} は p 個の説明変数間の共分散行列であり，C_{XX}^{-1} はその逆行列を示す（ただし，C_{XX} は正則で逆行例が存在するとする）．また，C_{Xy} は説明変数と基準変数の共分散ベクトルである．（3）式をその要素を用いて表わすと，

表 1　食道癌の訂正死亡率と食物供給量の相関係数

食道癌	(y)	1.000			
熱　量	(x_1)	0.107	1.000		
肉　類	(x_2)	0.259	0.766	1.000	
アルコール類	(x_3)	0.343	0.618	0.433	1.000
		y	x_1	x_2	x_3

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} S^2_{x_1} & S_{x_1x_2}\cdots S_{x_1x_p} \\ S_{x_2x_1} & S^2_{x_2}\cdots S_{x_2x_p} \\ \vdots & \vdots \\ S_{x_px_1} & S_{x_px_2}\cdots S^2_{x_p} \end{pmatrix}^{-1} \begin{pmatrix} S_{yx_1} \\ S_{yx_2} \\ \vdots \\ S_{yx_p} \end{pmatrix} \quad (4)$$

である．ただし，$S_{x_ix_j}$ などは変数間の共分散を表わすものとする．

定数項 b_0 は

$$b_0=\bar{y}-(b_1\bar{x}_1+b_2\bar{x}_2+\cdots+b_p\bar{x}_p) \quad (5)$$

により求めることができる．ただし，$\bar{y}$, $\bar{x}_1$, $\bar{x}_2$, $\cdots\bar{x}p$, は各変数の平均値である．

全ての変数を平均0，分散1に標準化した場合，共分散 $S_{x_ix_j}$ は相関係数 $r_{x_ix_j}$ と一致するので（4）式中の共分散を全て相関係数で置きかえればよい．この場合，各変数にかかる重みは標準偏回帰係数とよばれ変数 x_j について bs_j のように表わすと，

$$bs_j=S_{xj}\,b_j \quad (6)$$

の関係があり，各変数の標準偏差と偏回帰係数により求めることができる．また定数項は0になる．

〔例 1〕

表 1 は世界46カ国の男性の食道癌の訂正死亡率と，1人1日あたりの3つの食物供給量（熱量，肉類，アルコール類）の相関係数を示したものである[6]．食道癌の訂正死亡率を y，熱量を x_1，肉類を x_2，アルコール類を x_3 とし，重回帰式を求めてみよう．

標準偏回帰係数を bs_1, bs_2, bs_3 とすると，(4)式より

$$\begin{pmatrix} bs_1 \\ bs_2 \\ bs_3 \end{pmatrix} = \begin{pmatrix} 1.000 & 0.766 & 0.618 \\ 0.766 & 1.000 & 0.433 \\ 0.618 & 0.433 & 1.000 \end{pmatrix}^{-1} \begin{pmatrix} 0.107 \\ 0.259 \\ 0.343 \end{pmatrix}$$

$$\fallingdotseq \begin{pmatrix} 3.202 & -1.964 & -1.128 \\ -1.964 & 2.435 & 0.159 \\ -1.128 & 0.159 & 1.628 \end{pmatrix} \begin{pmatrix} 0.107 \\ 0.259 \\ 0.343 \end{pmatrix} = \begin{pmatrix} -0.553 \\ 0.475 \\ 0.479 \end{pmatrix}$$

となる．

この結果は，食道癌の訂正死亡率の重回帰式中で，肉類とアルコール類は正に（死亡率を高める），熱量は負に（死亡率を低める）寄与することを示している．しかし，標準偏回帰係数の符号から，個々の変数の寄与を考える場合，個別に判断を下すと誤るおそれがある．すなわち，熱量の係数が負となっているからといって，熱量の摂取を増加させれば食道癌の死亡が減ると考えるのは早計である．表1から明らかなように，説明変数間には相関があるのが普通である．そのため，ある変数の値を増加すれば，他の変数の値に必ず影響するはずである．分析に用いた説明変数の組合せが変われば，係数の値も符号も変わり得るので，変数間の相互関係をよく吟味してから結果の解釈は行うべきである．

ところで，基準変数 y と予測値 $\bar{y}$ の差，すなわち残差 e の平均 $\bar{e}$ と分散 S_e^2 はどのようになるであろうか．残差の平均 $\bar{e}$ は簡単な計算により0となることがわかる．残差の分散 S_e^2 は，

$$S_e^2=S_y^2(1-R^2_{y\cdot x}) \quad (7)$$

となる．ここで，S_y^2 は変数 y の分散，$R_{y\cdot x}$ は重相関係数であり，以下によって定義される．

$$R^2_{y\cdot x}=\frac{C'_{xy}\,C^{-1}_{xx}\,C_{xy}}{S_y^2}=\frac{1}{S_y^2}\sum_{i=1}^{p} S_{yx_i}\,b_i$$

$$=\sum_{i=1}^{p} r_{yx_i}\,bs_i \quad (8)$$

によって求められる．重相関係数は基準変数の実測値と予測値の相関係数であるが，（7）式のように，残差分散がどの程度になるかを示す目安となる．重相関係数の2乗 $R^2_{y\cdot x}$ は回帰により，基準変数の分散をどのくらい説明できるかを示す．このため多重決定係数とよばれることがある．

重相関係数は変数の数が増加するにつれ増加し，変数の数が標本数より1だけ少ない数に達すると，どのような場合であれ必ず1になる．そのため，標本数 n と説明変数の数 p を用いて，自由度調整ずみ重相関係数 $\hat{R}_{y\cdot x}$, すなわち

$$\hat{R}^2_{y\cdot x}=1-\frac{n-1}{n-p-1}(1-R^2_{y\cdot x}) \quad (9)$$

を求めることがある．また，

$$\hat{\hat{R}}^2_{y\cdot x}=1-\frac{(n-1)(n+p+1)}{(n+1)(n-p-1)}\cdot(1-R^2_{y\cdot x}) \quad (10)$$

を用いることもある．$\hat{\hat{R}}_{y\cdot x}$ は自由度再調整ずみ重相関係数とよばれる．$\hat{R}_{y\cdot x}$, $\hat{\hat{R}}_{y\cdot x}$ は変数の数が増加しても，ある一定の数を越えると数値の増加は止み，逆に減少するようになる．そのため，重回帰分析にどの変数を含めればよいかの基準として用いることができる．

〔例 2〕

例1の結果を用いて，重相関係数を求めてみよう．

(8) 式に数値を代入すると，

$$R^2_{y\cdot x}=0.107\times(-0.553)+0.259\times0.475$$
$$+0.343\times0.479\fallingdotseq0.228$$

表 2　前進選択法による変数選択

順位	変　数	F値	$R^2_{y\cdot x}$	$\hat{R}^2_{y\cdot x}$	$\hat{\hat{R}}^2_{y\cdot x}$
1	アルコール類	5.87	0.118	0.098	0.078
2	熱　量	0.87	0.135	0.095	0.057
3	肉　類	5.03	0.228	0.173	0.120

となる．$R_{y\cdot x}$ は 0.478 となり，(9)，(10) 式より

$$\hat{R}^2_{y\cdot x}=1-\frac{46-1}{46-3-1}(1-0.228)\fallingdotseq 0.173$$

$$\hat{\hat{R}}^2_{y\cdot x}=1=\frac{(46-1)(46+3+1)}{(46+1)(46-3-1)}\cdot(1-0.228)\fallingdotseq 0.120$$

となる．多重決定係数は 0.228 なので，3つの食物供給量により食道癌の訂正死亡率が約 23% 説明できることになる．逆に，残りの部分は他の要因によるものである．

■ 変数選択の方法

重回帰分析の予測の効率を高めるために，変数のうち重要性の高いものを選ぶ必要がある．一般には，一定の基準を設定しそれを満たす変数のみを重回帰式に採用するという方式をとる．通常よく用いられる方法は，変数が基準を満たさなくなるまで，1つずつ変数を増加していく手法と，逆に全ての変数のうち基準を満たさないものを1つずつ減らしていく手法である．前者は変数増加法（前進選択法），後者は変数減少法（後進選択法）とよばれる．

基準変数を y，説明変数を $x_1, x_2, \cdots, x_p$，標本数を n とする．すでに k 個の変数 $X_k=(x_1, x_2, \cdots, x_k)$ が重回帰式に含まれており，その重相関係数が $R_{y\cdot X_k}$ であるとする．変数 x_j を追加するか否かの基準として，

$$F_j=\frac{(n-k-2)(R^2_{y\cdot X_k x_j}-R^2_{y\cdot X_k})}{1-R^2_{y\cdot X_k x_j}} \qquad (11)$$

を求め，F_j が最大となる変数を選べばよい．逆に変数を取り除くには，X_k から変数 x_j を減した場合の重相関係数を $R_{y\cdot X_{k(x_j)}}$ とすると，

$$F_j=\frac{(n-k-1)(R^2_{y\cdot X_k}-R^2_{y\cdot X_{k(x_j)}})}{1-R^2_{y\cdot X_k}} \qquad (12)$$

を求め，F_j が最小となる変数を除けばよい．

通常は前進選択法の場合，(11) 式の F 値が1もしくは2以上の変数があれば重回帰式に追加し，なければ変数選択を終了する．後進選択の場合には，(12) 式の F 値が1もしくは2未満の変数があれば重回帰式より取り除き，なければ変数選択を終了する．この他に変数選択の基準として PSS[1]，AIC[8] などがあるが，ここでは割愛する．

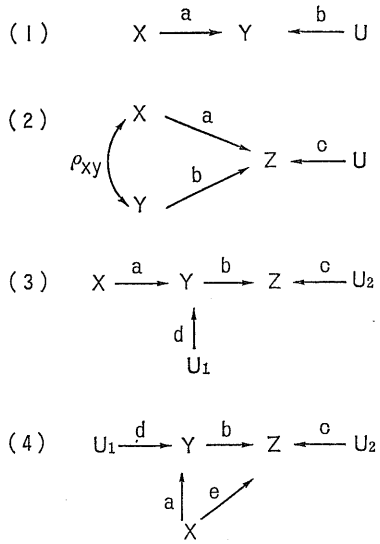

図 1　いくつかのパス・モデルの例

〔例 3〕

表 2 は表 1 の相関係数行列をもとに，食道癌の訂正死亡率の予測式に含める変数を前進選択した結果である．

変数選択の順位は，アルコール類，熱量，肉類の順であった．2番目の熱量の F 値は1以下であり，前述したように，通常はこの段階で変数選択を止めることになる．しかし，これにかまわずに選択を続けると，次の肉類では F 値は 5.03 になり，自由度調整ずみ重相関係数の値も増加する．このように，実際のデータでは F 値が基準以下になったからといって，必ずしも変数選択を止める必要はない．とくに標本数が小さい場合，変数選択の基準となる F 値が，変数の追加や削除により大きく変化することがあり，上記の例のようなことが起りやすい．また可能ならば，すべての変数の組合せについて自由度（再）調整ずみ重相関係数を求め，その値が最大となる変数の組合せを求めることも考えられる．ただこの方法では説明変数の数が増加するに従い，その計算量は膨大なものとなる．

■ パス解析

統計学では通常，変数間の相関を問題にし，その変数間の因果関係を扱うことはない．しかし，因果関係が存在すれば必ず相関関係があるので，因果関係を相関係数などをもとに推論する場合もある．この目的に対して，パス解析という方法を用いることができる．

いま2つの変数X，Yがあり，XはYの原因であるとしよう．このような場合，X→Yのように，XからYに向かって矢印を引き，XがYに影響を及ぼしていることを示すようにする．この矢印はパスとよばれる．しかし，

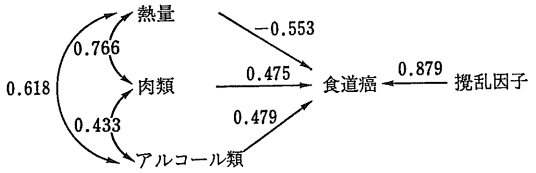

図 2　重回帰分析のパス・モデルによる表示

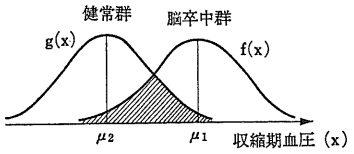

図 3　1変数による判別

XによってYの変動が全て説明されるわけではないのが普通である．そこで，X以外の他の要因による影響も考える必要があるだろう．図1-(1)をみよう．これはYに対して変数Xが影響し，その他にUという未知の要因も関与するという状態を示したものである．Uは攪乱因子（誤差因子）ともよばれるものである．各変数の結果に及ぼす影響の強さは，各パスの上にa，bのように記すことにする．これらはパス係数とよばれる．

パス解析では全ての変数は平均0，分散1に標準化して用いられる．このため，図1-(1)のような場合には，パス係数aはXとYの相関係数に一致する．しかし，変数が多数ある場合，この関係は必ずしも成り立つものではない．

図1-(2)は，原因と考えられる変数がX，Yと2つあり，これがZに影響しているという状態を示すパス・モデルである．この図では，XとYは両端に矢じりのついた曲線で結ばれているが，これは2つの変数間に相関があり，その大きさがρ_{XY}であることを示している．

図1-(2)のモデルを式にかき表わすと，

$$Z=aX+bY+cU \qquad (13)$$

となる．上式は重回帰式と異なり攪乱因子Uを含んでおり，構造式とよばれる．(13)式から各変数間の相関をもとにパス係数を求めることができる．しかし，通常はより簡単な方法でパス係数を求めることができる．

ある2変数間の相関係数は，一方の変数から，他方の変数に到るすべての可能な道順でのパス係数の積和に一致する．ただし，パスの矢の方向が変数に向いていなければ，その変数に達することはできない．たとえば，XとYはUに到るパスは存在しないので，Uとの相関は0である．XとZの相関ρ_{XZ}を考えてみよう．XからZに到るパスは，直接Zに向かうものと，Yを経由してZに向かう2つがある，直接Zに達するパスのパス係数はa，Yを経由してZに向かうパス係数の積は$\rho_{XY}\cdot b$となるので，

$$\rho_{XZ}=a+\rho_{XY}b \qquad (14)$$

となる．同様に$\rho_{YZ}=\rho_{XY}a+b$の関係が成り立つ．これらの2式から，a，bを求めることができるが，この場合はZを基準変数，X，Yを説明変数とした重回帰分析による標準偏回帰係数に，aとbは完全に一致する．またcは

$$c^2=1-(a\rho_{XZ}+b\rho_{YZ}) \qquad (15)$$

となり，1から多重決定係数を引いたものとなる．実際の計算では，母相関係数ρ_{XZ}などのかわりに標本相関係数r_{XZ}などを用いればよい．

図1-(3)は，XがYの原因になり，YがZの原因になるという連鎖モデルを示している．このような場合，各パス係数と相関係数の間には，

$$a=\rho_{XY}b=\rho_{YZ} \qquad (16)$$

の関係が成り立つが，さらに

$$\rho_{XZ}=ab=\rho_{XY}\rho_{YZ} \qquad (17)$$

が成り立つ必要がある．(17)式のような条件は制約条件とよばれるが，かりに条件が成り立たねばモデルを変える必要が生じる．

図1-(4)は，図1-(2)と似ているが，XとYの間にXがYの原因でもあるという場合を示している．この場合，XとZの間には，XからZに直接影響を及ぼすパスと，Yを経由するものとに分けて考えることができる．すなわち，

$$\rho_{XZ}=e+ab \qquad (18)$$

とかける．eは直接効果，abは間接効果とよばれるものである．

〔例4〕

例1で求めた重回帰分析の結果を，パス・モデルで表わしてみよう．

図2がその結果である．このように視覚的に変数相互の関係を把握することができ，かつ必要な数値を全て含むことが，パス解析の利点である．しかし，モデルを現実に適応する場合，結論を急ぐあまりパス解析の結果を直接因果関係に結びつけるべきではない．どのような分析であれ，相関関係をもとに行われているものは，直接因果関係を立証することは不可能であり，研究の仮説設定のために用いるべきであろう．

■ 判別分析

判別分析は，基準変数が特定の疾患の有無のように定性的な場合に用いられる手法である．標本の多変量データが与えられているとき，その標本がどのようなグルー

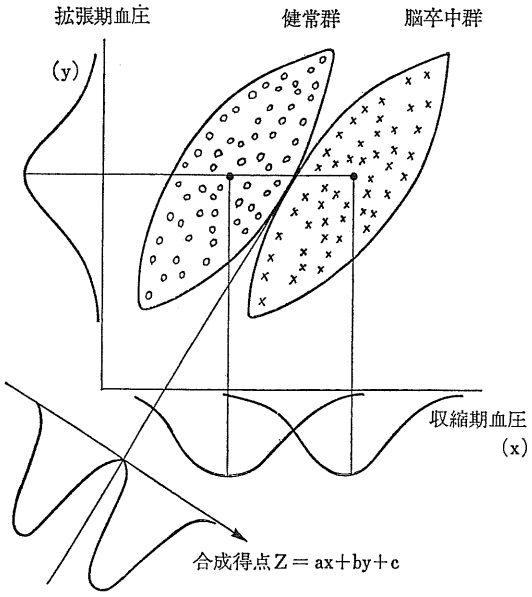

図 4　2変数を用いた判別

プ（患者か健常人かなど）に属するかを識別するための手法といえる．

収縮期血圧のデータから，個人が脳卒中群になるか健常群になるかを判別する方法を考えてみよう．図3は2群での収縮期血圧の分布を示したものである（架空例）．収縮期血圧をxとし，脳卒中群と健常群はそれぞれ $f(x)$ $g(x)$ に従って分布しているとする．どちらの群に属するのか不明のケースからデータ x_1 を得たとする．このような場合，ケースをどちらの群に分類すれば誤りが少なくなるだろうか．それには2群それぞれに属する確からしさ，すなわち次式で定義される尤度

$$L_1=f(x_i),\ L_2=g(x_i) \qquad (19)$$

を求め，$L_1>L_2$ ならば群1（脳卒中群）に，$L_1<L_2$ ならば群2（健常群）に判別すればよい．ただし，2群の事前確率を考慮しないとする．もしも，分布形が等しく，たんに平均値が μ_1，μ_2 と異なるだけであれば，平均値からの差が小さい方に判別すればよいことになる．しかし，1変数のみによる判別ではそれほど高い正確さで，所属する群を識別することはできない．図3の斜線部は上記の方法によっては誤って判別してしまう部分を示している．誤判別率を低くするには，変数の数を増やすことも1つの方法であろう．

図4は2変数を用いた場合の判別の方法を示したものである．図のようにかりに拡張期血圧の平均値に差がなくとも，2つの変数から適当な合成得点を求めることで，判別がほぼ完全に行えることを示したものである．これは架空例であるが，同様の考え方は変数がp個の場合にも成り立つ．すなわち，変数を $X_1, X_2, \cdots, X_p$ とすると，合成得点Zを

$$z_i=b_0+b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip} \qquad (20)$$

のように各ケースについて求め，これをもとに判別すればよい．上式は判別式，または判別関数などとよばれ，z_i は判別得点，各変数にかける重み b_j は判別係数とよばれる．変数が標準化されている場合は b_j は標準判別係数とよばれ，重回帰分析での標準偏回帰係数と同様の意味をもつ．また b_0 は定数項である．また，p個の変数がp次元多変量正規分布に従い，2群の母共分散行列が等しい場合，(20) 式を導くことができ，Fisher の線型判別関数とよばれる．変数 x_j の各群の平均値を μ_{j1}，μ_{j2} とかくことにする．平均値ベクトルを $\mu_1=(\mu_{11}, \mu_{21}, \cdots, \mu_{p1})'$ のように定義する．母共分散行列 Σ が正則で逆行列が存在すれば，判別係数ベクトル $\mathbf{b}$ は，

$$\mathbf{b}=\Sigma^{-1}(\mu_1-\mu_2) \qquad (21)$$

により求められる．実際の計算には母共分散行列を推定する必要がある．2群の標本数を n_1，n_2 とし，標本共分散行列を $C_{XX(1)}$，$C_{XX(2)}$ とした場合，

$$\hat{\Sigma}=\frac{1}{n_1+n_2-2}(n_1C_{XX(1)}+n_2C_{XX(2)}) \qquad (22)$$

により推定すればよい．また定数項 b_0 は，

$$\hat{b}_0=-\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)'\hat{\mathbf{b}}=-\frac{1}{2}\sum_{i=1}^{p}(\hat{\mu}_{i1}+\hat{\mu}_{i2})\hat{b}_i \qquad (23)$$

により推定できる．ただし，$\hat{\mu}_i$ は μ_i の，$\hat{\mathbf{b}}$ は $\mathbf{b}$ の推定値よりなるベクトルとする．

求めた判別得点が正ならば群1，負ならば群2に各ケースを判別すればよい．もし，2群の母集団の事前確率 p_1，p_2 ($p_1+p_2=1$) の大きさが推定できるのなら，判別得点が $\ln(p_1/p_2)$ より大ならば群1に，小ならば群2とすればよい．

表 3　2群での各変数の平均値（標準偏差）

変　数		I群	N群
空腹時血糖値	(mg/dl)	202.1(75.1)	147.7(51.1)
発症年齢	(歳)	41.6(13.6)	49.7(11.5)
血清尿素窒素	(mg/dl)	17.6(5.1)	16.0(5.7)

I群 257例，N群 261例

表 4　3変数の共分散行列

空腹時血糖値	(x_1)	4130		
発症年齢	(x_2)	−18	159	
血清尿素窒素	(x_3)	− 3	6	29
		x_1	x_2	x_3

判別式に各群の変数の平均値を代入し，各群の判別得点の差 D^2_M を求めてみよう．

$$D^2_M=(\mu_1-\mu_2)'\sum^{-1}(\mu_1-\mu_2) \qquad (24)$$

となることがわかる．D^2_M は2つの母集団の平均値間の距離の平方を示し，マハラノビスの汎距離とよばれるものである．

ところで，2群の判別分析は基準変数hの値として，群1では $1/n_1$，群2では $-1/n_2$ を各ケースに与えた場合の重回帰分析に対応する．そこで基準変数hと説明変数を用いて，重相関係数 $R_{h.x}$ を考えることができる．

$$D^2_M=\frac{(n_1+n_2-2)(n_1+n_2)}{n_1n_2}\cdot\frac{R^2_{h.x}}{1-R^2_{h.x}} \qquad (25)$$

の関係が成り立つことがわかっている．上記の関係から，$R^2_{h.x}$ が1に近づくにつれ D^2_M は無限に大きくなり，かつ $R^2_{h.x}$ の増大とともに D^2_M は単調に増加する．従って，重回帰分析の場合と全く同一に変数選択を行うことができる．また，(25) 式の関係から(9)，(10)式を用いて自由度調整ずみマハラノビスの汎距離，自由度再調整ずみマハラノビスの汎距離も求めることができる[1]．

〔例5〕

糖尿病患者を初診時以降5年間の治療法により，インスリン治療群（I群257例），と非インスリン治療群（N群261例）に分けた．**表3**に示した3変数は初診時での各群の平均値と標準偏差である[7]．初診時でのこれら3変数のデータを用いて，I群とN群の判別分析を行ってみよう．ただし，3変数間の共分散行列は**表4**のようになっていると仮定する．

(21) 式より判別係数は

$$\begin{pmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{b}_3\end{pmatrix}=\begin{pmatrix}4130 & -18 & -3\\ -18 & 159 & 6\\ -3 & 6 & 29\end{pmatrix}^{-1}\begin{pmatrix}202.1-147.7\\ 41.6-49.7\\ 17.6-16.0\end{pmatrix}$$

$$=\begin{pmatrix}2.423\text{E}-4 & 2.669\text{E}-5 & 1.954\text{E}-5\\ 2.669\text{E}-5 & 6.342\text{E}-3 & -1.309\text{E}-3\\ 1.954\text{E}-5 & -1.309\text{E}-3 & 3.475\text{E}-2\end{pmatrix}\begin{pmatrix}54.4\\ -8.1\\ 1.6\end{pmatrix}$$

$$=\begin{pmatrix}0.0130\\ -0.0520\\ 0.0662\end{pmatrix}\quad(\text{“E}-4\text{”}\ \text{などは “}\times10^{-4}\text{” を示す})$$

となる．また定数項 $\hat{b}_0$ は (23) 式より

$$\hat{b}_0=-\frac{1}{2}\{(202.1+147.7)\times0.0130-(41.6+49.7)\times0.0520+(17.6+16.0)\times0.0662\}\fallingdotseq-1.012$$

となる．判別式は

$$Z=0.013\times(\text{空腹時血糖値})-0.052\times(\text{発症年齢})+0.0662\times(\text{血清尿素窒素})-1.012$$

となる．もしも，ある糖尿病患者の初診時のデータが空腹時血糖値 170 mg/dl，発症年齢 40 歳，血清尿素窒素 18 mg/dl であれば，上式に数値を代入すると，

$$Z=0.013\times170-0.052\times40+0.0662\times18-1.012\fallingdotseq0.31$$

となり，$Z>0$ なのでI群と判別される．

(24) 式からマハラノビスの汎距離 D^2_M を求めると，$D^2_M\fallingdotseq1.234$ となり，重相関係数 $R_{h.x}$ は 0.486 となる．これらの値はまだ十分大きくないので，理論的に予測される判別の正確さ（正診率）はそれほど高くない（正規分布を仮定して求めると 71.1%）．このためさらに判別式に他の変数を追加した方がよいだろう．

この他，判別分析は広く疾病の計量診断に用いられているが，それについては次回で説明する．

■ 多重ロジスティック・モデル

判別分析とよく似た方法であるが，所属する群を定性的に識別するのでなく，ある群に所属する確率を定量的に求める方法である．

2群のうち1群である確率をp，群2の確率をq（$p+q=1$）としよう．p個の変数 $x_1, x_2, \cdots, x_p$ を用いて，ある標本の群1に属する確率 p_i（例えばある疾患の危険度）を，

$$\log\frac{p_i}{q_i}=b_1x_{i1}+b_2x_{i2}+\ +b_px_{ip}+c \qquad (25)$$

とする．左辺は p_i のロジットとよばれるが，上式はロジットがp個の変数に関して線型であるとしたものである．p_i について整理すると

$$p_i=\frac{1}{1+e^{-z_i}} \qquad (26)$$

となる．ただし，

$$z_i=b_1x_{i1}+b_2x_{i2}+\cdots+b_px_{ip}+c \qquad (27)$$

である．(26) 式はロジスティック関数とよばれるが，(27) 式のようにロジットが多変数の合成得点よりなるので，多重ロジスティック関数とよばれる．(26) 式は特定の疾患に対する危険度（リスク）の算出によく用いられるため，リスク関数，用いられた各変数はリスク・ファクター（危険因子）とよばれる．(27) 式はp個の変数がp次元多変量正規分布に従い，2群の母共分散行列が等しい場合，すでに述べた判別関数と完全に一致する．その場合には，(27) 式の各変数の係数には判別係数を用いればよいことになる．しかし一般には，Walker-Duncan 法などの最尤法を用いることが多い[4]．変数選択の方法は，判別分析の場合と同様に行っても大きく誤ることはないが，赤池による情報量規準 AIC を用いることもできる[8]．

〔例 6〕

例 5 で求めた判別関数を用いて，I 群になるリスクを求めてみよう．いまある糖尿病患者の初診時のデータは，空腹時血糖値 190 mg/dl，発症年齢 42 歳，血清尿素窒素 17 mg/dl であったとする．**例 5** の結果を用いて，

$$Z=0.013\times190-0.052\times42+0.0662\times17-1.012=0.3994$$

となる．(26) 式に代入すると，

$$p=\frac{1}{1+e^{-0.3994}}\fallingdotseq0.599$$

となり，I 群である可能性は 59.9% である．

■ その他の方法

今回説明していない方法として，説明変数が定性的な場合，重回帰分析に対して数量化理論 1 類があり，また判別分析に対しては数量化理論 2 類がある．また重回帰分析を特別な場合に含む，基準変数が多数ある場合に用いる正準相関分析がある．判別分析において判別する群が 3 群以上の場合，重判別分析を用いることができる．また特定の疾患等の死亡率や生存率の時間的推移と説明変数を用いる分析として，Cox の生命表分析を用いることもできる[9]．この手法の応用例として，比例ハザード・モデルを用いて，長期低線量放射線の健康影響を調べた報告がある[10]．

上記の各手法の詳細は紙面の都合で説明を省略するが，近年医学分野でも徐々に適用されるようになってきた．興味のある読者は参考文献を参照してもらいたい[1,3,4,9]．

参考文献

1) 柳井晴夫，高根芳雄：多変量解析法，朝倉書店(東京)，1977.
2) Blalock, H.M. Jr. : Causal inferences in nonexperimental research, the Univ. of North Carolina, Chapel Hill, 1964.
3) 安田三郎：社会統計学，丸善(東京)，1969.
4) Walker, S.H. and Duncan, D.B. : Estimation of the probability of an event as a function of several independent variables, Biometrika, **54**；167-179, 1967.
5) 林知己夫：数量化の方法，東洋経済(東京)，1974.
6) 高木廣文：疫学研究における因果モデル作成上の新しい手法の開発とその応用について，聖路加看護大学紀要，**8**；11-19, 1982.
7) 高木廣文，他：多変量解析による患者の初診時臨床像と治療法に関する研究，日本公衛誌，**26** (3)；125-134, 1979.
8) 赤池弘次：情報量規準 AIC とは何か，その意味と将来への展望，数理科学，**153**；5-10, 1976.
9) Cox, D.R. : Regression models and life-tables, J.R. Stat. Soc., **B 34**；187-220, 1972.
10) 服部芳明，前田和甫：比例ハザード・モデルを用いた長期低線量放射線被曝による健康影響の評価，保健物理，17巻，127-136, 1982.