ヘルスサイエンスのための統計科学 ●連載 8

判別分析と多重ロジスティックモデル

高木廣文

◉特定の疾患に対するリスク因子が多数ある場合の解析方法として判別分析と多重ロジスティックモデルを紹介する．実例を用いて各手法の概要と結果の解釈を中心に解説する．

*

キーワード：判別分析，多重ロジスティックモデル，リスク，疫学

循環器疾患の発症原因として加齢，高塩分摂取のみならず，肉類などの摂取による高コレステロールなど，いくつかのリスク因子が考えられるだろう．このように特定の疾患への罹患，または死亡などに対して複数個の原因が考えられる場合，それらの要因の影響をどのように評価すればよいのであろうか．

原因と考えられる変数間の相関関係を考慮し，このような目的に適した解析方法として，判別分析と多重ロジスティックモデルによる多変量解析がある．

いくつかのグループに対象者（ケース）が分類される場合，判別分析を用いることで，その分類に説明変数がどのように影響するかを解析することができる．

一方，多重ロジスティックモデルはアメリカのFramingham Studyではじめて用いられた方法であり，心疾患のリスク因子の評価に威力を発揮したことで知られている[1]．このため，多重ロジスティックモデルはリスク関数ともよばれている．

ここでは実例による分析結果を用いて，各方法の簡単な理論的な解説とともに，結果の解釈や注意すべき点などの説明を行う．

Discriminant analysis and multiple logistic model
Hirofumi TAKAGI：統計数理研究所

■判別分析

判別分析は基準変数が循環器疾患への罹患の有無のように質的であり，説明変数がコレステロール値などのように量的なデータの場合に用いられる多変量解析の一方法である．判別分析を用いるおもな目的として，

① 所属するグループが未知のケースの識別・分類

② 各説明変数のグループ識別への影響の程度を検討する

などがある．

さまざまな血液学的検査値などを用いて，新来患者の疾患の診断を行うという自動診断などは①を目的とした場合である．また，日常の各種の生活習慣が循環器疾患の発生にどのような影響を及ぼすかを検討したい，というような場合には②のような使用目的となる．

今回も都市，農村，漁村に在住する主婦の314人についての調査データ[2]を用いて分析方法の解説を行う．

いま，既応歴をもとにして，グループとして高血圧群と健常群を考えよう．この2つのグループ分けに影響する説明変数として，質問項目の回答から計算した22の生活習慣尺度得点と年齢，および肥満度（BMI，kg/m^2）を用いることにする．

この例のように，比較すべきグループが2つある場合，もっとも一般的に用いられる方法は各変数について「2群の母平均値の差の検定」を行うことである．すなわち，上記の説明変数それぞれについて，表1に示したようなt検定を行うのが普通である．

表1は，高血圧群と健常群について各変数の平均値，標準偏差，およびその検定結果を示したも

表1　グループ別平均値と標準偏差

変数名	高血圧群	健常群	t 値	p 値
1．肉・油脂	4.7(1.9)	4.7(2.0)	0.080	0.936
2．洋風の食事	4.6(1.8)	5.5(2.3)	3.158	0.002
3．高塩分	4.4(2.1)	4.3(1.9)	0.122	0.903
4．糖分	3.6(2.2)	4.0(2.0)	1.148	0.252
5．食事規則性	8.5(1.8)	8.4(1.9)	0.390	0.697
6．料理への進取性	6.8(2.6)	6.3(2.4)	1.319	0.188
7．娯楽	4.3(1.8)	4.5(2.0)	0.544	0.587
8．教養	5.8(2.2)	5.6(2.3)	0.550	0.583
9．社会奉仕	4.8(2.6)	5.0(3.1)	0.518	0.605
10．義理人情	9.3(2.0)	8.2(2.1)	3.647	0.000
11．経済型	7.7(1.8)	7.1(1.8)	2.081	0.038
12．伝統型	9.6(1.7)	8.7(1.9)	3.070	0.002
13．清潔	9.7(1.6)	9.9(1.7)	0.542	0.588
14．妻主導型	7.5(2.1)	7.3(2.4)	0.429	0.668
15．運動実施	3.1(2.4)	2.8(2.5)	0.882	0.379
16．疾病頻度	5.3(2.8)	2.8(2.3)	6.914	0.000
17．多愁訴	4.2(2.6)	3.7(2.7)	1.265	0.207
18．情緒不安定	4.2(2.8)	4.3(2.7)	0.191	0.849
19．外向性	5.3(1.8)	5.0(1.9)	1.358	0.176
20．共感性	8.3(1.7)	7.8(2.2)	1.337	0.182
21．自発性	7.1(2.3)	6.2(2.5)	2.142	0.017
22．遺伝的健康観	8.0(2.2)	6.5(2.6)	4.038	0.000
23．年齢	53.6(8.7)	47.2(8.7)	4.807	0.000
24．BMI	24.2(2.9)	22.8(2.9)	3.350	0.001

（　）内は標準偏差．t 値と p 値は2群の母平均値の差の検定．
標本数は，妻主導型と BMI を除き，高血圧群 53 例，健常群 252 例．
妻主導型では高血圧群 46 例，健常群 236 例，BMI では各 52 例，249 例．

のである．2群の母平均値の差の検定は t 分布を仮定して行われる．表中の t 値が検定のための統計量であり，それをもとに p 値（有意確率）を計算する．その p 値が小さいほど，「2群の母平均値は等しい」という帰無仮説は考えにくくなる．表中で灰色の帯を付した変数は p 値が5％以上のものである．したがって，2群間で母平均値に差があると考えられる変数は，洋風の食事，義理人情，経済型，伝統型，疾病頻度，自発性，遺伝的健康観，年齢，BMI の9変数であった．

多変量解析を用いない場合には，上記の結果から有意差のあった変数の平均値をみて高血圧群の特徴を検討することになる．ところが，年齢をみると高血圧群で6.4歳も平均値が高い．2グループでの年齢差が他の変数での差の原因になっている可能性があるだろう．このような場合，年齢などの影響を除いて，あるいは変数間の相関を考慮した分析を行う必要があるだろう．そのためには多変量解析の一手法である判別分析を行えばよい．

判別分析では判別のための「判別関数」として重回帰分析と同様に，

判別得点＝（重み1）×（説明変数1）
　　＋（重み2）×（基準変数2）＋……
　　＋（重みP）×（説明変数P）
　　＋定数

のように，各説明変数に適当な重みをかけた式を考える．

各グループをうまく判別するためには，このようにして求めた判別得点について，2グループでの平均的距離ができるだけ大きくなるように各変数の重みを決めればよい．この重みは「判別係数」

表2　説明変数の判別係数

変数名	判別係数	F値	p値
2．洋風の食事	−0.167116	3.652	0.057
6．料理への進取性	0.197133	5.686	0.018
9．社会奉仕	−0.170790	5.377	0.021
10．義理人情	0.132480	2.111	0.147
16．疾病頻度	0.442502	25.857	0.000
17．多愁訴	−0.154755	3.584	0.059
21．自発性	0.166040	3.687	0.056
22．遺伝的健康観	0.193965	6.980	0.009
23．年齢	0.045647	4.435	0.036
24．BMI	0.098424	2.261	0.134
定数	−8.31679		

とよばれている．

重回帰分析の場合と同様に，説明変数の影響の有無は判別係数の検定によって調べることができる．さらに，2グループの判別分析は重回帰分析と同様な考え方で，説明変数の選択が行える．すなわち，帰無仮説「母判別係数＝0」の検定のためのF値の小さな変数を除いて，ある基準値以上の変数のみを残せばよい．

ここでは全説明変数を用いた判別分析から，F値のもっとも小さな変数を1つずつ除き，F値が2以上の変数のみを残した．

表2に示したように，変数減少法によって10個の説明変数が判別分析に有用なものとして残された．表1に示した検定結果では単独の変数では有意でなかった変数のうち，料理への進取性，社会奉仕，多愁訴の3変数が選択されている．逆に，単独では有意差があるものの，他の変数との関係から判別への寄与が小さいとされ，選択されなかった変数もある．すなわち，経済型と伝統型の2尺度は選択されなかった．このように，一変量解析の結果と，多変量解析の結果はかならずしも一致するものではない．なお，判別分析の場合も重回帰分析と同様に変数選択はp値を5％とせずにF値が2以上で行っており，通常の検定基準に比べるとかなり甘くなっている．この例ではp値が15％程度になる．

判別得点は表2の各判別係数を変数のデータにかけた和を計算し，定数を加えればよい．求めた判別得点が正ならば高血圧群，負ならば健常群にそのケースとして判別される．したがって，判別係数の正負によって，その説明変数のもつ意味が逆になる．表2の結果から疾病頻度の得点が大きいほど高血圧群の可能性が高くなる．同様に，料理の進取性，義理人情，自発性，遺伝的健康観，年齢，BMIなど，いずれも値が大きいほど高血圧群の可能性が高くなる．ただし，ここで注意すべき点は，これらの説明変数がかならずしも高血圧の原因ではないという点である．なぜならば，この調査はある一時点での断面調査なので，因果関係には言及できないからである．疾病頻度尺度は病気がち傾向を示すものであるが，これはどちらかといえば，高血圧症になった結果と考えるべきであろう．

表3　みかけの的中率

実際	判別関数による予測		計
	高血圧	健常	
高血圧群	38	14	52
健常群	46	203	249

同様に，健康は遺伝で決まるという遺伝的健康観尺度の得点が高くなるのも，病気になった結果として，そのような考えを抱くようになったのかもしれない．社会奉仕や，洋風の食事，多愁訴が低いと高血圧群になる傾向が高いというのはやや説明が難しい．やはり結果として社会奉仕が困難になり，また洋風の食事を控えるような傾向があるのかもしれない．

多愁訴は身体が痛いなどの訴えの多さについての尺度であり，普通に考えれば高血圧群で高いはずである．事実，表1の結果ではそのようになっている．しかし，判別分析の結果は符号が逆になっている．これは多愁訴が疾病頻度と相関係数0.522と高い相関があるためである．すなわち，重回帰分析の場合と同様に，説明変数間に，きわめて高い相関が認められる場合，一方の判別係数が正であれば，他方の変数については負の値になることが多い．実際，表2の結果では疾病頻度は正の判別係数で，かつ最大のF値となっていることがわかる．

求められた判別関数の妥当性は，それを用いた場合，実際にグループの判別がうまくいくかにかかっている．

表3は分析に用いた対象者について判別関数を用いた場合に予測されるグループと実際のグループとの一致の程度を調べるために作成したクロス

表4　多重ロジスティックモデル

変数名	係数（標準誤差）	F値	p値
2．洋風の食事	−0.1575954(0.09788813)	2.592	0.108
6．料理への進取性	0.2027369(0.08867177)	5.228	0.023
9．社会奉仕	−0.1727559(0.07831431)	4.866	0.028
10．義理人情	0.1213648(0.09897231)	1.504	0.221
16．疾病頻度	0.3291272(0.08515986)	14.937	0.000
17．多愁訴	−0.1192843(0.08228413)	2.102	0.148
21．自発性	0.1341697(0.09486956)	2.000	0.158
22．遺伝的健康観	0.1996079(0.07545904)	6.997	0.009
23．年齢	0.0535390(0.02343440)	5.220	0.023
24．BMI	0.0963195(0.06652966)	2.096	0.149
定数	−10.45181　(2.302617　)	20.603	0.000

表である．高血圧群では52例中38例（73.1%）が正しく判別された．また，健常群では249例中203例(81.5%)が正しく判別され，全体の的中率は80.1%であった．このように，分析に用いた対象者について，判別関数で判別の的中率を調べる方法は「みかけの的中率」などとよばれる．

より厳しい基準で判別関数の有効性を確認するには，分析に用いなかった新規の患者などに，その判別関数を適用してみるとよい．すなわち，判別関数による分類と専門医などの診断が一致するかどうかを調べればよい．的中率が高ければ高いほど，その判別関数は妥当性も高いといえる．

■多重ロジスティックモデル

判別分析では判別得点の正負により高血圧群と健常群にグループ分けを行った．それでは求めた判別得点が5の場合には高血圧症である可能性はどのくらいの大きさなのであろうか．また，判別得点が1ならばどうなのであろうか．肥満度(BMI)が1大きくなると，どのくらい高血圧になるリスクが高くなるのであろうか．

このような疑問に答えるには多重ロジスティックモデルを用いた分析を行えばよい．

ある疾患の発生率と非発生率（＝1−発生率）の比はオッズとよばれるが，オッズの対数をとった値（対数オッズ）に対して，

$$\log\left[\frac{\text{発生率}}{\text{非発生率}}\right]=(\text{重み}1)\times(\text{説明変数}1)+(\text{重み}2)\times(\text{基準変数}2)+\cdots\cdots+(\text{重み}P)\times(\text{説明変数}P)+\text{定数}$$

のような式を考える．この式の右側は判別分析での判別得点を求める式と同じ形式であることがわかるであろう．

発生率について式を書き直すと，

$$\text{発生率}=\frac{1}{1+\exp\left[-\text{判別得点}\right]}$$

となる．

判別得点を横軸に，発生率を縦軸にとり，図示するとS字状の曲線が得られる．すなわち，判別得点が負で大きい場合には発生率は0に近く，得点が0の場合には発生率は0.5となり，得点が正に大きくなるにつれ，発生率も1に限りなく近づいていく．この曲線はロジスティック曲線とよばれ，上記のモデルは多重ロジスティックモデルとよばれる．

実際の計算では各説明変数の重みを求める必要がある．判別分析の結果をそのまま用いることもできるが，ここでは最尤法とよばれる方法で求めた結果を表4に示した．

表4の結果を表3と比べると大筋では大差のない結果といえる．すなわち，判別分析で求めた係数も，最尤法で求めた係数も，大差のない値になっていた．ただし，義理人情のF値が2.111から1.504と2未満になり，説明変数としてモデルに含めるには不適当ということになってしまった．

しかし，表4の説明変数の係数をもとにした解釈を行うだけでは，多重ロジスティックモデルを用いる意味は薄い．説明変数のリスクを評価するためには，表4に示した各説明変数の係数と標準誤差の値からオッズ比とその信頼区間を求めればよい．オッズ比はリスク比の推定値として用いられる．表5では各変数の1単位の増加によって，どれだけリスク（この場合にはオッズ比）が増え

表5　各変数の1単位当りの推定オッズ比と信頼区間

変数名	オッズ比	95%信頼区間（下限，上限）	99%信頼区間（下限，上限）
2．洋風の食事	0.854	(0.705, 1.035)	(0.664, 1.099)
6．料理への進取性	1.225	(1.029, 1.457)	(0.975, 1.539)
9．社会奉仕	0.841	(0.722, 0.981)	(0.688, 1.029)
10．義理人情	1.129	(0.930, 1.371)	(0.875, 1.457)
16．疾病頻度	1.390	(1.176, 1.642)	(1.116, 1.731)
17．多愁訴	0.888	(0.755, 1.043)	(0.718, 1.097)
21．自発性	1.144	(0.950, 1.377)	(0.896, 1.460)
22．遺伝的健康観	1.221	(1.053, 1.416)	(1.005, 1.483)
23．年齢	1.055	(1.008, 1.105)	(0.993, 1.121)
24．BMI	1.101	(0.966, 1.254)	(0.928, 1.307)

るかを示すものである．

表5は各変数の1単位当りのオッズ比と信頼区間を示している．オッズ比の信頼区間は95%と99%の両方を示した．この信頼区間に1が含まれない場合には，より積極的にその変数が高血圧症に関係すると考えてよいことになる．

洋風の食事，社会奉仕，多愁訴の各尺度得点が高い場合には高血圧症の罹患は低く，他の変数ではその値が大きくなると高血圧症の罹患率が高くなることがわかる．ただし，95%信頼区間でみると，料理への進取性，疾病頻度，遺伝的健康観，年齢の4変数がすべて1を越えており，逆に社会奉仕が1未満となっている．ここで，注意すべきことは，表5に示したオッズ比は，他の変数が一定，すなわち変化しないと仮定した場合に，当該の変数を1単位だけ変えれば，どのようなオッズ比になるかを計算したものである．

実際には疾病頻度と多愁訴の間のように，変数間にはなんらかの相関関係があることが多い．そのような場合には，一方を変えると，他方も変化するので，現実問題への適応や解釈は注意が必要である．

ところで，他の変数が変化せずに，年齢が5歳増えた場合，高血圧症の罹患率についてのオッズ比はどのくらいになるであろうか．実際にはオッズ比は5倍ではなく5乗になる．すなわち，5歳年齢が増えると，$1.055^5=1.307$，となり，約1.3倍のリスク増加と推定される．いくつかの変数の組合せについても同様な計算により罹患のリスクが何倍になるか推定できる．たとえば，自発性得点が2増え，年齢が5歳増え，BMIが3増えた場合，他の変数が変わらなければ，

$1.144^2\times1.055^5\times1.101^3=2.283$

となり，約2.3倍にリスクが増えることになる．

■おわりに

特定の疾患への罹患や死亡と，どのような変数がどのように関係しているかを解析するための手法として判別分析と多重ロジスティックモデルによる多変量解析を紹介した．判別分析は，どちらかといえばケースの識別，判別に説明変数がどのくらいの強さで関係しているのかを検討するための方法といえよう．これに対して多重ロジスティックモデルを用いる場合には各説明変数についてオッズ比を求め，疾患の発生などにどのようなリスクがあるのかを検討する目的があるといえるであろう．両者はよく似た結果を与えるが，医学領域では多重ロジスティックモデルの果たす役割が大きいのではないかと考えられる．今後，現在以上にこれらの方法がさまざまな場で用いられることが期待される．

なお今回も実際の分析はHALBAUを用いて行った[3,4]．また，今回紹介した判別分析，多重ロジスティックモデルの数理面に興味ある読者は成書[4]を参照していただきたい．

文献

1) Truett, J. et al.：A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.* **20**：511-524, 1967.
2) 高木廣文：健康と習慣に関する重回帰分析．医学のあゆみ．**174**：217-222，1995.
3) 高木廣文：HALBAU-4マニュアルIII．多変量解析．現代数学社，1994.
4) 柳井晴夫，高木廣文（編著）：多変量解析ハンドブック．現代数学社，1986.