数量化 II 類はダミー変数を用いた判別分析である

青木繁伸

2005 年 10 月 17 日

目次

表目次

1　はじめに

前報で、数量化 I 類とダミー変数を用いた重回帰分析が全く同じものであることを示した。

重回帰分析と判別分析はこれまた全く同じものであることも周知の事実であることから、判別分析と数量化 II 類が同じものであることは容易に推察できる。

この文書は、簡単な例を取り上げて、数量化 II 類がダミー変数を用いた判別分析に他ならないことを示すために作成する。

2　データおよび方法

2.1　解析に使用するデータ

表 1 に示すような、4 変数、15 ケースからなる架空データを使用する。3 個の独立変数 (アイテム変数: X_1, X_2, X_3) は 3 個のカテゴリーを持つ変数であり、それぞれ 1 から 3 までの整数値でコード化されている。群変数 (Y) は 1 または 2 の整数値を持つ群変数である。

2.2　解析法

判別分析および数量化 II 類を行う統計解析プログラムは標準的なものでよいが、ここでは NAP [3] を用いた。判別分析プログラムを適用するための準備としては JGAWK [1] などを用い、判別分析プログラムからの出力結果から数量化 II 類の解析プログラムが出力する統計量を導くためには Lotus 1–2–3 を用いた。

3　既存の統計解析プログラムによる解析結果

3.1　数量化 II 類による解析結果

数量化 II 類を用いた解析結果は以下のようになる。

群変数	カテゴリー数	独立変数	カテゴリー数
Y	2	X_1	3
		X_2	3
		X_3	3

★ 正規化カテゴリースコア

アイテム	カテゴリー	カテゴリースコア
X_1	1	–0.0437217
	2	0.940017
	3	–1.19142
X_2	1	0.207678
	2	–0.120235
	3	–0.284191
X_3	1	–1.31165
	2	0.819782
	3	1.31165

	重心
第 1 群	–1.016530
第 2 群	0.508265
分割点	–0.254133
決定係数	0.516667

★ 偏相関係数

アイテム	第 1 軸
X_1	0.52004
X_2	0.10701
X_3	0.49217

★ サンプルスコア

ケース	R		P	第 1 軸
1	2		2	0.65583
2	2	##	1	–0.49187
3	2		2	0.65583
4	2	##	1	–1.14770
5	2		2	1.47561
6	2		2	0.98374
7	2		2	1.63956
8	1		1	–1.14770
9	1		1	–0.49187
10	2		2	0.98374
11	1		1	–1.14770
12	1		1	–1.14770
13	2		2	–0.16396
14	1		1	–1.14770
15	2		2	0.49187

★ 判別結果

	判別された群		
実際の群	第 1 群	第 2 群	合計
第 1 群	5	0	5
%	(100.0)	(0.0)	(100.0)
第 2 群	2	8	10
%	(20.0)	(80.0)	(100.0)

正判別率 = 86.67%

3.2 判別分析プログラムによる解析

3.2.1 判別分析プログラムを適用するための準備

まず最初に、アイテム変数をダミー変数に展開する。あるアイテム変数の持つ情報をダミー変数で表現するとき、アイテム変数が k 個のカテゴリーを持つ場合には、0 か 1 かのいずれかを持つ二値データ k 個のダミー変数に展開される。例えば、あるアイテム変数が i という値を持つ場合、i 番目のダミー変数は値 1 を持ち、残りのダミー変数は値 0 を持つ。

表 1 に示したデータ中の 3 つのアイテム変数のデータは、表 2 のように、延べ 9 個のダミー変数 ($D1_1, \ldots, D3_3$) に展開される

しかし、このダミー変数は冗長な情報を持つ。例えば、$k-1$ 個のダミー変数が 0 であるとき、残りの 1 個

表 1　判別分析に用いられるアイテム変数データ例

ケース	X_1	X_2	X_3	Y
1	1	2	2	2
2	3	2	2	2
3	1	2	2	2
4	1	1	1	2
5	2	3	2	2
6	1	3	3	2
7	2	2	2	2
8	1	1	1	1
9	3	2	2	1
10	1	1	2	2
11	1	1	1	1
12	1	1	1	1
13	2	1	1	2
14	1	1	1	1
15	1	3	2	2

表 2　アイテム変数をダミー変数に展開した結果

ケース	$D1_1$	$D1_2$	$D1_3$	$D2_1$	$D2_2$	$D2_3$	$D3_1$	$D3_2$	$D3_3$
1	1	0	0	0	1	0	0	1	0
2	0	0	1	0	1	0	0	1	0
3	1	0	0	0	1	0	0	1	0
4	1	0	0	1	0	0	1	0	0
5	0	1	0	0	0	1	0	1	0
6	1	0	0	0	0	1	0	0	1
7	0	1	0	0	1	0	0	1	0
8	1	0	0	1	0	0	1	0	0
9	0	0	1	0	1	0	0	1	0
10	1	0	0	1	0	0	0	1	0
11	1	0	0	1	0	0	1	0	0
12	1	0	0	1	0	0	1	0	0
13	0	1	0	1	0	0	1	0	0
14	1	0	0	1	0	0	1	0	0
15	1	0	0	0	0	1	0	1	0

のダミー変数は必ず 1 である。そこで、多変量解析においては、各アイテム変数に対応する複数のダミー変数のうちの 1 つを除いて解析に使用する。どのダミー変数を除いてもよいが、数量化理論の解析プログラムにおいては最初のダミー変数を除くのが通例なので、ここでもそれに従うことにする。

実際に判別分析プログラムで使われるのは、表 3 のように加工されたデータ行列である。

3.2.2 判別分析プログラムによる解析結果

ダミー変数を用いた判別分析の結果は以下のようになる。

群を表わす変数　Y 第 1 群が取る値 = 1, 第 2 群が取る値 = 2

独立変数　$D1_2, D1_3, D2_2, D2_3, D3_2, D3_3$

★ 平均値

	全体	第 1 群	第 2 群
	15 ケース	5 ケース	10 ケース
$D1_2$	0.2000000	0.000000	0.3000000
$D1_3$	0.1333333	0.2000000	0.1000000
$D2_2$	0.3333333	0.2000000	0.4000000
$D2_3$	0.2000000	0.000000	0.3000000
$D3_2$	0.5333333	0.2000000	0.7000000
$D3_3$	0.06666667	0.000000	0.1000000

★ プールされた群内相関係数行列

$D1_2$	1.00000					
$D1_3$	–0.15878	1.00000				
$D2_2$	–0.07715	0.60025	1.00000			
$D2_3$	0.04762	–0.15878	–0.46291	1.00000		
$D3_2$	–0.04052	0.49542	0.65653	–0.04052	1.00000	
$D3_3$	–0.21822	–0.08085	–0.23570	0.50918	–0.43329	1.00000
	$D1_2$	$D1_3$	$D2_2$	$D2_3$	$D3_2$	$D3_3$

★ 分類関数

	第 1 群	第 2 群	偏 F 値	p 値
$D1_2$	–0.67241	–6.0517	1.1034	0.32420
$D1_3$	–2.4655	3.8103	0.84483	0.38490
$D2_2$	0.22414	2.0172	0.044138	0.83885
$D2_3$	0.33621	3.0259	0.084881	0.77820
$D3_2$	–1.4569	–13.112	2.1939	0.17683
$D3_3$	–1.7931	–16.138	1.3581	0.27744
定数項	0.36983	5.256		

上の表中の偏 F 値の自由度 = (1, 8)

Wilks の $\Lambda = 0.4833333$

等価な F 値 = 1.425287　自由度 = (6, 8.00)　p 値 = 0.31304

★ 判別関数

● 第 1 群と第 2 群の判別

マハラノビスの汎距離 = 2.04180

理論的誤分類率 = 0.15365

	判別係数	標準化判別係数
$D1_2$	–2.68966	–1.07586
$D1_3$	3.13793	1.06669
$D2_2$	0.89655	0.42264
$D2_3$	1.34483	0.53793
$D3_2$	–5.82759	–2.90731
$D3_3$	–7.17241	–1.78911
定数項	2.44310	

★ 各ケースの判別結果

ケース	R		P	平方距離 1	平方距離 2	判別値
1	2		2	8.83(0.183)	3.86(0.696)	–2.48793
2	2	##	1	5.92(0.433)	7.22(0.301)	0.65000
3	2		2	8.83(0.183)	3.86(0.696)	–2.48793
4	2	##	1	0.37(0.999)	5.26(0.511)	2.44310
5	2		2	16.51(0.011)	7.05(0.316)	–4.72931
6	2		2	18.47(0.005)	11.70(0.069)	–3.38448
7	2		2	16.68(0.011)	6.32(0.388)	–5.17759
8	1		1	0.37(0.999)	5.26(0.511)	2.44310
9	1		1	5.92(0.433)	7.22(0.301)	0.65000
10	2		2	18.47(0.005)	11.70(0.069)	–3.38448
11	1		1	0.37(0.999)	5.26(0.511)	2.44310
12	1		1	0.37(0.999)	5.26(0.511)	2.44310
13	2		2	7.09(0.312)	6.60(0.359)	–0.24655
14	1		1	0.37(0.999)	5.26(0.511)	2.44310
15	2		2	11.13(0.084)	7.05(0.316)	–2.03966

注：‘平方距離’は各群の重心までのマハラノビスの平方距離カッコ内は各群に属する確率

★ 判別結果

	判別された群		
実際の群	第 1 群	第 2 群	合計
第 1 群	5	0	5
%	(100.0)	(0.0)	(100.0)
第 2 群	2	8	10
%	(20.0)	(80.0)	(100.0)

正判別率 = 86.67 %

4 判別分析および数量化 II 類の解析結果の比較

3.1 節の結果と 3.2.2 節の結果を比較すると以下のような点が挙げられる。

- 類似点
  - それぞれの方法による判別値は異なるが、両者の相関係数は 1 である
  - 判別結果は全く同じである

- 数量化 II 類の解析結果からのみ得られる情報
  - 群変数とアイテム変数間の偏相関係数
  - 各カテゴリーに与えられたカテゴリースコア
- 判別分析の解析結果からのみ得られる情報
  - 各ダミー変数の各群および全体での平均値 (ダミー変数が 0/1 型の二値データであることから、この平均値は各カテゴリーに反応したものの比率である)
  - ダミー変数間の群内相関係数
  - 分類関数、偏 F 値、Wilks の Λ、マハラノビスの汎距離、理論的誤分類率

この比較からは、それぞれの解析はそれなりの情報を与えていることがわかるが、総じていえば、判別分析から得られる情報量の方が多い。実際、以下に示すように数量化 II 類の解析結果は、判別分析から得られる情報に包含されている。

5 判別分析の結果から数量化 II 類による結果を導く方法

5.1 判別係数と正規化カテゴリースコアの関係

数量化 II 類は延べ 9 個のカテゴリーに対して正規化カテゴリースコアを出力する。これは、表 2 に示したような (冗長な) ダミー変数に対する判別係数に他ならない。

判別分析の解析結果は、冗長性を除いたダミー変数に対する判別係数しか与えないので、冗長であるとして除かれたダミー変数を含めた場合の判別係数を導く必要がある。実際的には、除かれたダミー変数は 0 という判別係数を持つことになっている。従って、表 4 の、判別係数の空欄 ($D1_1$, $D2_1$, $D3_1$) は、0 という数値があることになる。

そのままでもよいのであるが、アイテム変数ごとに全ケースの平均値が 0 になるように調整したものが数量化 II 類における “正規化カテゴリースコア” になる。

正規化処理は簡単である。まず、各ダミー変数が 1 であるケース数と判別係数を掛けそれをアイテム変数ごとに合計する。表 4 の 1 番目のアイテム変数についていえば、$0\times10-2.68966\times3+3.13793\times2=-8.06898+6.27586=-1.79312$ が合計であるので、正規化しないときの平均値は $-1.79312/15=-0.11954$ である。そこで、判別係数からこの値を差し引いた値を各ダミー変数への重みとし、$0+0.11954=0.11954$, $-2.68966+0.11954=-2.57012$, ... のように変換したものを表 4 の、“原点移動” と表示した欄に書いた。

さらに、後述するように、全ケースの判別値が、平均 0、標準偏差 1 になるように正規化したものを、表 4 の最右欄に示した。これは、数量化 II 類での正規化カテゴリースコアに一致する。

また、正規化するために判別係数を調整すると、判別式の定数項も調整しなければならない。各アイテムごとの調整の総和は、$-0.11954+0.56782-3.58621=-3.13793$ であり、これをキャンセルするために当初の定数項にこれを加え、$2.44310-3.13793=-0.69483$ が、原点移動された判別式の定数項になる。さらに各ケースの判別値が平均値 0、標準偏差 1 になるように正規化すると、定数項は 0 になる。

5.2 判別値の計算

判別値は各ダミー変数に割り当てられた判別係数の合計値である (ダミー変数が 0/1 型の二値データであるため計算が簡単になる)。ここで、各ケースの各アイテム変数ごとのスコアを考える。各ケースのアイテムごとのスコアは、表 5 のようにして求められる。なお、ここで示したのは冗長なダミー変数を除かない場合であり、数量化 II 類のときの判別値の求め方の説明でもあるが、冗長なダミー変数を除いた判別式 (3.2.2 節) を用いても同じになることはすでに 5.1 節において示された。

5.3 アイテム変数と群変数の相関関係

アイテム変数の持つ値は、例に挙げたデータでは 1, 2, 3 という数値である。しかし、これらの数値は本質的には名義尺度 (順序尺度の場合もある) である。従って、表 1 のようなデータからはアイテム変数と群変数間の相関関係は論じることができない。

数量化 II 類や判別分析で各アイテムに割り付けられた数値 (正規化カテゴリースコア、判別係数) は、アイテム変数を間隔尺度に格上げするための重みである。従って、この数値に基づくスコアは群変数との相関関係を論じるために使用できる。

表 6 におけるアイテム変数ごとのサンプルスコア、群変数および判別値について相関係数行列を求める。

このデータ行列から相関係数行列 表 7 を求める。

次に、“アイテム変数と群変数間の偏相関係数” を求める。表 7 の相関係数行列の逆行列を求めたものを、表 8 に示す。群変数 Y と、アイテム変数 X_i との偏相関係数 r_{Y,X_i} は、R^{Y,X_i} などを R の逆行列 R^{-1} の (Y, X_i) 要素としたとき、

$$r_{Y,X_i} = \frac{-R^{Y,X_i}}{\sqrt{R^{Y,Y}R^{X_i,X_i}}}$$

によって計算される。このようにして得られた偏相関係数が、表 8 の最右欄に書かれている。これは、数量化 II 類の解析結果で得られる “アイテム変数と群変数の間の偏相関係数” に一致する。

6 考察

以上のように、アイテム変数を対象とする数量化 II 類は、ダミー変数を使用した判別分析に他ならないことが示された。

従来の多くの判別分析プログラムが出力するしないに関わらず、この文書で述べたように、補助的な解析処理を行えば必要な情報は全て得られる。これらの補助的な解析を判別分析のプログラムで自動的に行ってくれれば、数量化 II 類のプログラムは特に必要ではないということになる。

参考文献

[1] A. V. Aho, B. W. Kernighan, P. J. Weinberger; 足立高徳 訳: プログラミング言語 AWK, トッパン

[2] C. Hayasi(1952): On the prediction of phenomena from qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, **3**, 69-98.

[3] 青木繁伸 (1989): 医学統計解析リファレンスマニュアル、医学書院 (東京)

表 3　判別分析に用いられるダミー変数データ例

ケース	$D1_2$	$D1_3$	$D2_2$	$D2_3$	$D3_2$	$D3_3$	Y
1	0	0	1	0	1	0	2
2	0	1	1	0	1	0	2
3	0	0	1	0	1	0	2
4	0	0	0	0	0	0	2
5	1	0	0	1	1	0	2
6	0	0	0	1	0	1	2
7	1	0	1	0	1	0	2
8	0	0	0	0	0	0	1
9	0	1	1	0	1	0	1
10	0	0	0	0	1	0	2
11	0	0	0	0	0	0	1
12	0	0	0	0	0	0	1
13	1	0	0	0	0	0	2
14	0	0	0	0	0	0	1
15	0	0	0	1	1	0	2
反応数合計	3	2	5	3	8	1	

表 4　ダミー変数を用いた判別分析の結果から正規化カテゴリースコアを求める

	判別係数	反応数	合計	カテゴリ合計	原点移動	正規化
$D1_1$	0.00000	10	0.00000		0.11954	0.04372
$D1_2$	–2.68966	3	–8.06898	–0.11954	–2.57012	–0.94002
$D1_3$	3.13793	2	6.27586		3.25747	1.19142
$D2_1$	0.00000	7	0.00000		–0.56782	–0.20768
$D2_2$	0.89655	5	4.48275	0.56782	0.32873	0.12023
$D2_3$	1.34483	3	4.03449		0.77701	0.28419
$D3_1$	0.00000	6	0.00000		3.58621	1.31165
$D3_2$	–5.82759	8	–46.62072	–3.58621	–2.24138	–0.81978
$D3_3$	–7.17241	1	–7.17241		–3.58620	–1.31165
定数項	2.44310	15	36.64650		–0.69483	

表5　各ケースのカテゴリごとのスコアの求め方

ケース	$D1_2$	$D1_3$	$D2_2$	$D2_3$	$D3_2$	$D3_3$	判別係数	行列積	定数項付加	正規化
1	0	0	1	0	1	0	–2.68966	–4.93104	–2.48794	–0.65583
2	0	1	1	0	1	0	3.13793	–1.79311	0.64999	0.49187
3	0	0	1	0	1	0	0.89655	–4.93104	–2.48794	–0.65583
4	0	0	0	0	0	0	1.34483	0.00000	2.44310	1.14770
5	1	0	0	1	1	0	–5.82759	–7.17242	–4.72932	–1.47561
6	0	0	0	1	0	1	–7.17241	–5.82758	–3.38448	–0.98373
7	1	0	1	0	1	0	2.44310	–7.62070	–5.17760	–1.63957
8	0	0	0	0	0	0		0.00000	2.44310	1.14770
9	0	1	1	0	1	0		–1.79311	0.64999	0.49187
10	0	0	0	0	1	0		–5.82759	–3.38449	–0.98374
11	0	0	0	0	0	0		0.00000	2.44310	1.14770
12	0	0	0	0	0	0		0.00000	2.44310	1.14770
13	1	0	0	0	0	0		–2.68966	–0.24656	0.16396
14	0	0	0	0	0	0		0.00000	2.44310	1.14770
15	0	0	0	1	1	0		–4.48276	–2.03966	–0.49187
								平均値	–0.69483	
								標準偏差	2.73412	

表6　各ケースのカテゴリごとのスコア

ケース	X_1	X_2	X_3	Y
1	0.00000	0.89655	–5.82759	2
2	3.13793	0.89655	–5.82759	2
3	0.00000	0.89655	–5.82759	2
4	0.00000	0.00000	0.00000	2
5	–2.68966	1.34483	–5.82759	2
6	0.00000	1.34483	–7.17241	2
7	–2.68966	0.89655	–5.82759	2
8	0.00000	0.00000	0.00000	1
9	3.13793	0.89655	–5.82759	1
10	0.00000	0.00000	–5.82759	2
11	0.00000	0.00000	0.00000	1
12	0.00000	0.00000	0.00000	1
13	–2.68966	0.00000	0.00000	2
14	0.00000	0.00000	0.00000	1
15	0.00000	1.34483	–5.82759	2
平均値	–0.11954	0.56782	–3.58621	1.66667
標準偏差	1.65694	0.55429	2.94637	0.47140

表 7　相関係数行列

	X_1	X_2	X_3	Y
X_1	1.00000	0.04473	–0.15916	–0.31884
X_2	0.04473	1.00000	–0.85312	0.49562
X_3	–0.15916	–0.85312	1.00000	–0.58095
Y	–0.31884	0.49562	–0.58095	1.00000

表 8　アイテムスコア、群変数間の相関係数行列の逆行列と偏相関係数

	X_1	X_2	X_3	Y	偏相関係数
X_1	1.45168	0.48562	1.16893	0.90126	0.52004
X_2	0.48562	3.83649	3.52544	0.30149	0.10701
X_3	1.16893	3.52544	5.12472	1.60261	0.49217
Y	0.90126	0.30149	1.60261	2.06897	