数量化 I 類はダミー変数を用いた重回帰分析である

青木繁伸

2005 年 10 月 17 日

目次

表目次

1　はじめに

アイテム変数[*1]を用いて外的基準変数 (従属変数) を予測する多変量統計手法として数量化 I 類が提唱されている [2]。しかし、これを用いた論文を欧文誌に投稿しようとしたときに参考文献をあげようとして、はたと困惑したという話をよく聞く。しかし、適当な参考文献を付けても、欧文誌の査読者にとっては、数量化理論はなじみのない解析手法としか判断されないであろう。数量化 I 類と同じ解析を行う手法として彼らに通用するのは、“ダミー変数を用いた重回帰分析” なのである。

この文書は、簡単な例を取り上げて、数量化 I 類がダミー変数を用いた重回帰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。従属変数 (Y) は連続変数である。

2.2　解析法

重回帰分析および数量化 I 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。重回帰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、重回帰分析プログラムからの出力結果から数量化 I 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 I 類による解析結果

数量化 I 類を用いた解析結果は以下のようになる。

従属変数	独立変数	カテゴリー数
Y	X_1	3
	X_2	3
	X_3	3

重相関係数 = 0.68503

決定係数（重相関係数の二乗）= 0.46927

★ 相関係数行列

Y	1.00000			
X_1	0.19150	1.00000		
X_2	0.47051	–0.22261	1.00000	
X_3	0.58249	–0.07030	0.47866	1.00000
	Y	X_1	X_2	X_3

[*1] アイテム変数の持つカテゴリーと区別しにくい場合があるので、数量化理論ではアイテム変数と呼ばれることが多い。

★ 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	偏相関係数
X_1	1	–67.3037	0.36641
	2	450.0296	
	3	–338.5259	
X_2	1	–168.8741	0.35325
	2	372.3481	
	3	–226.5407	
X_3	1	–450.4444	0.47610
	2	281.1111	
	3	453.7778	
定数項		6119.067	

★ 予測値および残差

ケース	観察値	予測値	残差
1	6837.000	6705.222	131.7778
2	7397.000	6434.000	963.0000
3	7195.000	6705.222	489.7778
4	6710.000	5432.444	1277.556
5	6670.000	6623.667	46.33333
6	6279.000	6279.000	6.366463e–12
7	6601.000	7222.556	–621.5556
8	4929.000	5432.444	–503.4444
9	5471.000	6434.000	–963.0000
10	6164.000	6164.000	8.185452e–12
11	5095.000	5432.444	–337.4444
12	4766.000	5432.444	–666.4444
13	6525.000	5949.778	575.2222
14	5087.000	5432.444	–345.4444
15	6060.000	6106.333	–46.33333

3.2 重回帰分析プログラムによる解析

3.2.1 重回帰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[*2]。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。

実際に重回帰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。

[*2] 2 個のカテゴリーを持つ変数 (例えば性別) は、冗長性のない 1 個のダミー変数に変換される。このことは、2 個のカテゴリーを持つ変数は、重回帰分析にそのまま使えることを意味する。

表 1　重回帰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	6837
2	3	2	2	7397
3	1	2	2	7195
4	1	1	1	6710
5	2	3	2	6670
6	1	3	3	6279
7	2	2	2	6601
8	1	1	1	4929
9	3	2	2	5471
10	1	1	2	6164
11	1	1	1	5095
12	1	1	1	4766
13	2	1	1	6525
14	1	1	1	5087
15	1	3	2	6060

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

3.2.2 重回帰分析プログラムによる解析結果

ダミー変数を用いた重回帰分析の結果は以下のようになる。

	平均値	不偏分散	標準偏差
Y	6119.0666667	726281.78095	852.22167360
$D1_2$	0.20000000000	0.17142857143	0.41403933561
$D1_3$	0.13333333333	0.12380952381	0.35186577527
$D2_2$	0.33333333333	0.23809523810	0.48795003647
$D2_3$	0.20000000000	0.17142857143	0.41403933561
$D3_2$	0.53333333333	0.26666666667	0.51639777949
$D3_3$	0.066666666667	0.066666666667	0.25819888975

★ 相関係数行列

Y	1.00000						
$D1_2$	0.29126	1.00000					
$D1_3$	0.15003	–0.19612	1.00000				
$D2_2$	0.49910	–0.00000	0.55470	1.00000			
$D2_3$	0.13194	0.16667	–0.19612	–0.35355	1.00000		
$D3_2$	0.55873	0.13363	0.36690	0.66144	0.13363	1.00000	
$D3_3$	0.05192	–0.13363	–0.10483	–0.18898	0.53452	–0.28571	1.00000
	Y	$D1_2$	$D1_3$	$D2_2$	$D2_3$	$D3_2$	$D3_3$

★ 偏回帰係数など

	偏回帰係数	標準誤差	t 値	p 値	標準化偏回帰係数
$D1_2$	517.3333	580.7562	0.8907926	0.39904	0.2513388
$D1_3$	–271.2222	774.3416	0.3502617	0.73519	–0.1119824
$D2_2$	541.2222	967.9270	0.5591560	0.59136	0.3098835
$D2_3$	–57.66667	1046.973	0.0550794	0.95743	–0.02801650
$D3_2$	731.5556	892.3846	0.8197761	0.43608	0.4432810
$D3_3$	904.2222	1395.964	0.6477403	0.53530	0.2739536
定数項	5432.444	348.9911	15.5661426	0.00000	

t 値の自由度 = 8

★ 分散分析表

要因	平方和	自由度	平均平方	F 値	p 値
回帰	4771500.	6	795250.1	1.178925	0.40299
残差	5396444.	8	674555.6		
全体	1.016794e+07	14			

重相関係数 = 0.68503

決定係数（重相関係数の二乗） = 0.46927

自由度調整済み重相関係数の二乗 = 0.07122

★ 予測値、残差および標準化残差

ケース	従属変数	予測値	残差	標準化残差
1	6837.000	6705.222	131.7778	0.2017304
2	7397.000	6434.000	963.0000	1.612260
3	7195.000	6705.222	489.7778	0.7497702
4	6710.000	5432.444	1277.556	1.710442
5	6670.000	6623.667	46.33333	0.08786562
6	6279.000	6279.000	–9.094947e–13	–4.439341e–15
7	6601.000	7222.556	–621.5556	–1.095709
8	4929.000	5432.444	–503.4444	–0.6740315
9	5471.000	6434.000	–963.0000	–1.612260
10	6164.000	6164.000	–1.818989e–12	–8.878683e–15
11	5095.000	5432.444	–337.4444	–0.4517841
12	4766.000	5432.444	–666.4444	–0.8922624
13	6525.000	5949.778	575.2222	0.9750640
14	5087.000	5432.444	–345.4444	–0.4624948
15	6060.000	6106.333	–46.33333	–0.08786562

4　重回帰分析および数量化 I 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
  - 重相関係数および決定係数は全く同じである。
  - 予測値および残差については全く同じである。
- 数量化 I 類の解析結果からのみ得られる情報
  - アイテム変数と従属変数間の相関係数
  - 各アイテム変数と従属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
- 重回帰分析の解析結果からのみ得られる情報
  - 各ダミー変数の平均値 (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)
  - 各ダミー変数と従属変数間の相関係数
  - 各ダミー変数に対する偏回帰係数、有意検定、標準化偏回帰係数
  - 回帰の分散分析表
  - 自由度調整済みの重相関係数の二乗

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、重回帰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 I 類の解析結果は、重回帰分析から得られる情報に包含されている。

5　重回帰分析の結果から数量化 I 類による結果を導く方法

5.1　偏回帰係数と正規化カテゴリースコアの関係

数量化 I 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偏回帰係数に他ならない。

重回帰分析の解析結果は、冗長性を除いたダミー変数に対する偏回帰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偏回帰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 という偏回帰係数を持つことになっている。従って、表 4 の、偏回帰係数の空欄 ($D1_1$, $D2_1$, $D3_1$) 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 I 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 であるケース数と偏回帰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、$0\times10+517.333\times3-271.222\times2 = 1552.000-542.444 = 1009.556$ が合計であるので、正規化しないときの平均値は $1009.556/15 = 67.304$ である。そこで、偏回帰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、$0-67.304 = -67.304$, $517.333-67.304 = 450.030$, ... のように変換したものを表 4 の、最右欄に書いた。これは、数量化 I 類での正規化カテゴリースコアに一致する。

また、正規化するために偏回帰係数を調整すると、重回帰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、$67.304 + 168.874 + 450.444 = 686.622$ 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、$5432.444 + 686.622 = 6119.067$ が、正規化された重回帰式の定数項になる。これは当然ながら数量化 I 類の解析結果の定数項に一致する。

5.2 予測値の計算

予測値は各ダミー変数に割り当てられた偏回帰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 I 類のときの予測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重回帰式 (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。

これを全てのアイテムについて行い、合計値を求め、定数項を加えたものが予測値である (表 6)。アイテム変数ごとのカテゴリースコアを全ケースについて合計したものは 0 になる (そうなるように正規化したわけである)。

5.3 アイテム変数と従属変数の相関関係

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従属変数間の相関関係は論じることができない。

数量化 I 類や重回帰分析で各カテゴリーに割り付けられた数値 (正規化カテゴリースコア、偏回帰係数) 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従属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、従属変数および予測値について相関係数行列を求める。

以下のような手順を踏めば、Lotus 1–2–3 で処理することもできる。まず、各アイテムごとのサンプルスコア、従属変数および予測値を、平均値 0、標準偏差 1 に標準化する (表 7)。

この正規化データ行列を X とすると、XX^T は相関係数行列 R になる (表 8)。これは、数量化 I 類において出力される “アイテム変数と従属変数相互間の相関係数行列” である。

次に、“アイテム変数と従属変数の間の偏相関係数” を求める。表 8 の相関係数行列の逆行列を求めたものを、表 9 に示す。従属変数 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 9 の最右欄に書かれている。これは、数量化 I 類の解析結果で得られる “アイテム変数と従属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 I 類は、ダミー変数を使用した重回帰分析に他ならないことが示された。では、なぜ数量化 I 類が特に開発されたかという疑問が残る。統計解析プログラムが出力すべき情報の種類は決められているわけではないので、重回帰分析を行う解析プログラムが数量化 I 類が出力する情報を追加出力するのは、何の問題もない。しかし、従来の不偏的な解析プログラムはこれらの情報を出力しないものが多い。それを補完する意味で数量化 I 類を行うプログラムで必要な情報を出力するように設計したというのが実際であろう。また、ダミー変数の持つ性質を巧く利用し計算処理が簡単になるように体系づけたのが、“数量化理論” と呼ばれる所以ではないだろうか。

従来の多くの重回帰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重回帰分析のプログラムで自動的に行ってくれれば、数量化 I 類のプログラムは特に必要ではないということになる。

重回帰分析プログラムは、数量化 I 類のプログラムが不得意とする部分が容易になるということが挙げられる。

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　重回帰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	6837
2	0	1	1	0	1	0	7397
3	0	0	1	0	1	0	7195
4	0	0	0	0	0	0	6710
5	1	0	0	1	1	0	6670
6	0	0	0	1	0	1	6279
7	1	0	1	0	1	0	6601
8	0	0	0	0	0	0	4929
9	0	1	1	0	1	0	5471
10	0	0	0	0	1	0	6164
11	0	0	0	0	0	0	5095
12	0	0	0	0	0	0	4766
13	1	0	0	0	0	0	6525
14	0	0	0	0	0	0	5087
15	0	0	0	1	1	0	6060
反応数合計	3	2	5	3	8	1	

表4　ダミー変数を用いた重回帰分析の結果から正規化カテゴリースコアを求める

	偏回帰係数	反応数合計	合計得点	アイテム変数ごとの合計	平均値	偏回帰係数の正規化
$D1_1$	(0)	10				–67.304
$D1_2$	517.333	3	1552.000			450.030
$D1_3$	–271.222	2	–542.444	1009.556	67.304	–338.526
$D2_1$	(0)	7				–168.874
$D2_2$	541.222	5	2706.111			372.348
$D2_3$	–57.667	3	–173.000	2533.111	168.874	–226.541
$D3_1$	(0)	6				–450.444
$D3_2$	731.556	8	5852.444			281.111
$D3_3$	904.222	1	904.222	6756.667	450.444	453.778
定数項	5432.444					6119.067

表5　正規化カテゴリースコアからサンプルスコアを求める

ケース		$D1_1$		$D1_2$		$D1_3$	X'_1
1	1	–67.304	0	0	0	0	–67.3037
2	0	0	0	0	1	–338.526	–338.5259
3	1	–67.304	0	0	0	0	–67.3037
4	1	–67.304	0	0	0	0	–67.3037
5	0	0	1	450.030	0	0	450.02963
6	1	–67.304	0	0	0	0	–67.3037
7	0	0	1	450.030	0	0	450.02963
8	1	–67.304	0	0	0	0	–67.3037
9	0	0	0	0	1	–338.526	–338.5259
10	1	–67.304	0	0	0	0	–67.3037
11	1	–67.304	0	0	0	0	–67.3037
12	1	–67.304	0	0	0	0	–67.3037
13	0	0	1	450.030	0	0	450.02963
14	1	–67.304	0	0	0	0	–67.3037
15	1	–67.304	0	0	0	0	–67.3037

表6　サンプルスコア、観察値および残差

	アイテム変数ごとのスコア			予測値	従属変数	残差
ケース	X'_1	X'_2	X'_3	$\hat{Y}$	Y	$Y-\hat{Y}$
1	–67.304	372.348	281.111	6705.222	6837	131.778
2	–338.526	372.348	281.111	6434.000	7397	963.000
3	–67.304	372.348	281.111	6705.222	7195	489.778
4	–67.304	–168.874	–450.444	5432.444	6710	1277.556
5	450.030	–226.541	281.111	6623.667	6670	46.333
6	–67.304	–226.541	453.778	6279.000	6279	0.000
7	450.030	372.348	281.111	7222.556	6601	–621.556
8	–67.304	–168.874	–450.444	5432.444	4929	–503.444
9	–338.526	372.348	281.111	6434.000	5471	–963.000
10	–67.304	–168.874	281.111	6164.000	6164	0.000
11	–67.304	–168.874	–450.444	5432.444	5095	–337.444
12	–67.304	–168.874	–450.444	5432.444	4766	–666.444
13	450.030	–168.874	–450.444	5949.778	6525	575.222
14	–67.304	–168.874	–450.444	5432.444	5087	–345.444
15	–67.304	–226.541	281.111	6106.333	6060	–46.333
合計	0.000	0.000	0.000	6119.067	6119.067	
標準偏差	242.498	264.173	370.180	564.004	823.324	

表 7　アイテムスコア、従属変数および予測値の正規化データ行列

ケース	X'_1	X'_2	X'_3	Y	$\hat{Y}$
1	–0.27754	1.40949	0.75939	0.87199	1.03928
2	–1.39600	1.40949	0.75939	1.55216	0.55839
3	–0.27754	1.40949	0.75939	1.30682	1.03928
4	–0.27754	–0.63926	–1.21682	0.71774	–1.21741
5	1.85581	–0.85755	0.75939	0.66916	0.89468
6	–0.27754	–0.85755	1.22583	0.19425	0.28357
7	1.85581	1.40949	0.75939	0.58535	1.95653
8	–0.27754	–0.63926	–1.21682	–1.44544	–1.21741
9	–1.39600	1.40949	0.75939	–0.78713	0.55839
10	–0.27754	–0.63926	0.75939	0.05458	0.07967
11	–0.27754	–0.63926	–1.21682	–1.24382	–1.21741
12	–0.27754	–0.63926	–1.21682	–1.64342	–1.21741
13	1.85581	–0.63926	–1.21682	0.49304	–0.30016
14	–0.27754	–0.63926	–1.21682	–1.25354	–1.21741
15	–0.27754	–0.85755	0.75939	–0.07174	–0.02258
合計	0.00000	0.00000	0.00000	0.00000	0.00000
標準偏差	1.00000	1.00000	1.00000	1.00000	1.00000

表 8　アイテム変数ごとのサンプルスコア、従属変数および予測値間の相関係数行列

	X'_1	X'_2	X'_3	Y	$\hat{Y}$
X'_1	1.00000	–0.22261	–0.07030	0.19150	0.27955
X'_2	–0.22261	1.00000	0.47866	0.47051	0.68684
X'_3	–0.07030	0.47866	1.00000	0.58249	0.85032
Y	0.19150	0.47051	0.58249	1.00000	0.68503
$\hat{Y}$	0.27955	0.68684	0.85032	0.68503	1.00000

表 9　アイテムスコア、従属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	1.21749	0.43643	0.19995	–0.55496	0.36641
X'_2	0.43643	1.55452	–0.36125	–0.60456	0.35325
X'_3	0.19995	–0.36125	1.68044	–0.84716	0.47610
Y	–0.55496	–0.60456	–0.84716	1.88419	