多変量解析　～数量化2類～

発表日：5月19日

発表者：鈴木　朋央

数量化2類とは？

手法		判別に用いる変数
・ 判別分析	→	量的変数
・ 数量化2類	→	質的変数

数量化2類の具体例

健常者･患者の症状のデータ

サンプルNo.	健常者･患者	吐き気 x1	頭痛 x2
1	健常者	無	少
2	健常者	少	無
3	健常者	無	無
4	健常者	無	無
5	健常者	無	無
6	患者	少	多
7	患者	多	無
8	患者	少	少
9	患者	少	多
10	患者	多	少

･その疾病にかかっているか否か

･変数の判別能力

･変数の判別精度

例：吐き気がなく、頭痛が多いなら？

解析の流れ

1. 健常者：母集団[1]　患者：母集団[2]
   質的変数　→　ダミー変数（量的変数）
   各母集団への距離：マハラノビスの距離の2乗
   → マハラノビスの距離の2乗値の小さい
   母集団へ判別する、判別方式の制定

2. 誤判別の確率、判別方式の精度の評価
3. 変数選択により、有用な変数の抽出
4. 属する母集団が不明なサンプルの判別

～変数が1個の場合の解析方法～

ダミー変数の適応 (1)

- 「吐き気」のような質的な変数：アイテム
- パラメータ「多」「少」「無」：カテゴリー

$$x_{1(2)} = \begin{cases} 1 & \text{少のとき} \\ 0 & \text{少でないとき} \end{cases}$$

$$x_{1(3)} = \begin{cases} 1 & \text{多のとき} \\ 0 & \text{多でないとき} \end{cases}$$

例：$x_{1(2)} = x_{1(3)} = 0$ が「無」を意味する

適応時の留意点

- 一般に、「(質的変数のカテゴリー数) －1」個導入する。　例：3－1＝2
- 変数が2つの場合の判別分析を「形式的」に適用する。

　　なぜ「形式的」にか？

ダミー変数は0と1の2値しかとらないため、正規分布に従わない。

⬇

誤判別の確率の計算は正確ではなく、参考程度にとどめる

ダミー変数の適応（2）

ダミー変数適応後の健常者･患者の症状のデータ

サンプルNo.	健常者･患者	x1(2)	x1(3)
1	健常者	0	0
2	健常者	1	0
3	健常者	0	0
4	健常者	0	0
5	健常者	0	0
6	患者	1	0
7	患者	0	1
8	患者	1	0
9	患者	1	0
10	患者	0	1

判別方式

・線形判別関数の推定式

$$\hat{z} = \left[\hat{\mu}^{[1]}_{1(2)} - \hat{\mu}^{[2]}_{1(2)}, \hat{\mu}^{[1]}_{1(3)} - \hat{\mu}^{[2]}_{1(3)}\right] \begin{bmatrix} \hat{\sigma}^{(2)(2)} & \hat{\sigma}^{(2)(3)} \\ \hat{\sigma}^{(2)(3)} & \hat{\sigma}^{(3)(3)} \end{bmatrix} \begin{bmatrix} x_{1(2)} - \hat{\bar{\mu}}_{1(2)} \\ x_{1(3)} - \hat{\bar{\mu}}_{1(3)} \end{bmatrix}$$

$$= \left[0.20 - 0.60, 0 - 0.40\right] \begin{bmatrix} 10.00 & 10.00 \\ 10.00 & 16.67 \end{bmatrix} \begin{bmatrix} x_{1(2)} - 0.40 \\ x_{1(3)} - 0.20 \end{bmatrix}$$

$$= 5.33 - 8.00\, x_{1(2)} - 10.67\, x_{1(3)}$$

$$= 5.33 + \begin{Bmatrix} 0 \\ -8.00 \\ -10.67 \end{Bmatrix}$$

・ダミー変数 $x_{1(2)}$ と $x_{1(3)}$ の値に対しての判別方式

$\hat{z} \geq 0 \Leftrightarrow \hat{D}^{[1]2} \leq \hat{D}^{[2]2} \Leftrightarrow$　母集団[1]（健常者）に属する

$\hat{z} < 0 \Leftrightarrow \hat{D}^{[1]2} > \hat{D}^{[2]2} \Leftrightarrow$　母集団[2]（患者）に属する

判別結果

ダミー変数の適応、およびスコアと判別結果

サンプルNo.	健常者･患者	x1(2)	x1(3)	スコア	判別結果
1	健常者	0	0	5.33	健常者
2	健常者	1	0	-2.67	患者
3	健常者	0	0	5.33	健常者
4	健常者	0	0	5.33	健常者
5	健常者	0	0	5.33	健常者
6	患者	1	0	-2.67	患者
7	患者	0	1	-5.34	患者
8	患者	1	0	-2.67	患者
9	患者	1	0	-2.67	患者
10	患者	0	1	-5.34	患者

誤判別の確率（1）

判別表

データ結果	判別結果		計
	健常者	患者	
健常者	4	1	5
患者	0	5	5
計	4	6	10

・本来の健常者を患者と誤判別した割合：1/5 = 0.20

・本来の患者を健常者と誤判別した割合：0/5 = 0

誤判別の確率（2）

・判別効率の推定値

$$\hat{D}^2_{x_{1(2)}x_{1(3)}}([1],[2]) = \left(\hat{\mu}^{[1]} - \hat{\mu}^{[2]}\right)' \hat{\Sigma}^{-1} \left(\hat{\mu}^{[1]} - \hat{\mu}^{[2]}\right) = \hat{\delta}' \hat{\Sigma}^{-1} \hat{\delta}$$

$$= [0.20 - 0.60, 0 - 0.40] \begin{bmatrix} 10.00 & 10.00 \\ 10.00 & 16.67 \end{bmatrix} \begin{bmatrix} 0.20 - 0.60 \\ 0 - 0.40 \end{bmatrix}$$

$$= 7.468$$

・誤判別の確率　（数量化2類では正確でないが、正規分布に基づく方法を判別結果に適用する）

$$\Pr\left(u > \sqrt{\hat{\delta} \hat{\Sigma}^{-1} \hat{\delta}} / 2\right) = \Pr\left(u > \sqrt{7.468} / 2\right) = \Pr(u > 1.37)$$

$$= 0.0853$$

本来の健常者・患者を誤判別する確率：0.0853

変数選択

・判別分析 ⟶ 変数を1つずつ取り込む

・数量化2類 ⟶ ダミー変数を1組とする

$$F_0 = \frac{\left(n^{[1]} + n^{[2]} - p - r - 1\right) n^{[1]} n^{[2]} \left\{ \hat{D}^2_{x(p+r)}\left([1],[2]\right) - \hat{D}^2_{x(p)}\left([1],[2]\right) \right\}}{r\left\{ \left(n^{[1]} + n^{[2]} - 2\right)\left(n^{[1]} + n^{[2]}\right) + n^{[1]} n^{[2]} \hat{D}^2_{x(p)}\left([1],[2]\right) \right\}}$$

$$= \frac{\left(5 + 5 - 0 - 2 - 1\right) 5 \times 5 \left\{ 7.468 - 0 \right\}}{2\left\{ \left(5 + 5 - 2\right)\left(5 + 5\right) + 5 \times 5 \times 0 \right\}}$$

$$= 8.168$$

p：変数追加前のダミー変数の個数

r：追加するダミー変数の個数

変数選択の目安の値である2を超えているので、「吐き気」は判別に寄与していると考えられる。

不明なサンプルの判別

どちらの母集団に属するか不明なサンプルについて、その変数の値と得られた判別方式に基づいて、サンプルの判別を行う。

例：吐き気が無い人はどちらの母集団に属するか

⬇

得られた判別方式から、$\hat{z} = 5.33 > 0$ となるので、健常者と判別できる。

変数が2個以上の場合の解析方法

ダミー変数の適応～複数時（1）

・吐き気 x1

$$x_{1(2)} = \begin{cases} 1 & \text{少のとき} \\ 0 & \text{少でないとき} \end{cases}$$

$$x_{1(3)} = \begin{cases} 1 & \text{多のとき} \\ 0 & \text{多でないとき} \end{cases}$$

・頭痛 x2

$$x_{2(2)} = \begin{cases} 1 & \text{少のとき} \\ 0 & \text{少でないとき} \end{cases}$$

$$x_{2(3)} = \begin{cases} 1 & \text{多のとき} \\ 0 & \text{多でないとき} \end{cases}$$

頭痛x2も同様に、$x_{2(2)} = x_{2(3)} = 0$　が「無」を意味する

ダミー変数の適応～複数時（2）

複数のダミー変数適応後の健常者･患者の症状のデータ

サンプルNo.	健常者･患者	x1(2)	x1(3)	x2(2)	x2(3)
1	健常者	0	0	1	0
2	健常者	1	0	0	0
3	健常者	0	0	0	0
4	健常者	0	0	0	0
5	健常者	0	0	0	0
6	患者	1	0	0	1
7	患者	0	1	0	0
8	患者	1	0	1	0
9	患者	1	0	0	1
10	患者	0	1	1	0

判別方式

・線形判別関数の推定式

$$\hat{z} = \left(\hat{\mu}^{[1]} - \hat{\mu}^{[2]}\right)' \hat{\Sigma}^{-1}\left(x - \hat{\bar{\mu}}\right)$$

$$= 12.80 - 9.60x_{1(2)} - 20.80x_{1(3)} - 6.40x_{2(2)} - 14.40x_{2(3)}$$

$$= 12.80 + \begin{Bmatrix} 0 & (\text{吐き気が無}) \\ -9.60 & (\text{吐き気が少}) \\ -20.80 & (\text{吐き気が多}) \end{Bmatrix} + \begin{Bmatrix} 0 & (\text{頭痛が無}) \\ -6.40 & (\text{頭痛が少}) \\ -14.40 & (\text{頭痛が多}) \end{Bmatrix}$$

・ダミー変数 $x_{1(2)}, x_{1(3)}, x_{2(2)}, x_{2(3)}$の値に対しての判別方式

$\hat{z} \geq 0 \Leftrightarrow \hat{D}^{[1]2} \leq \hat{D}^{[2]2} \Leftrightarrow$ 母集団[1]（健常者）に属する

$\hat{z} < 0 \Leftrightarrow \hat{D}^{[1]2} > \hat{D}^{[2]2} \Leftrightarrow$ 母集団[2]（患者）に属する

判別結果

ダミー変数の適応、およびスコアと判別結果

サンプルNo.	健常者･患者	x1(2)	x1(3)	x2(2)	x2(3)	スコア	判別結果
1	健常者	0	0	1	0	6.4	健常者
2	健常者	1	0	0	0	3.2	健常者
3	健常者	0	0	0	0	12.8	健常者
4	健常者	0	0	0	0	12.8	健常者
5	健常者	0	0	0	0	12.8	健常者
6	患者	1	0	0	1	-11.2	患者
7	患者	0	1	0	0	-8	患者
8	患者	1	0	1	0	-3.2	患者
9	患者	1	0	0	1	-11.2	患者
10	患者	0	1	1	0	-14.4	患者

誤判別の確率（1）

判別表

データ結果	判別結果		計
	健常者	患者	
健常者	5	0	5
患者	0	5	5
計	5	5	10

・本来の健常者を患者と誤判別した割合：0/5 = 0

・本来の患者を健常者と誤判別した割合：0/5 = 0

誤判別の確率（2）

・判別効率の推定値

$$\hat{D}^2_{x_{1(2)}x_{1(3)}}([1],[2]) = \left(\hat{\mu}^{[1]} - \hat{\mu}^{[2]}\right)' \hat{\Sigma}^{-1}\left(\hat{\mu}^{[1]} - \hat{\mu}^{[2]}\right) = \hat{\delta}' \hat{\Sigma}^{-1} \hat{\delta}$$

$$= 19.20$$

・誤判別の確率　（数量化2類では正確でないが、正規分布に基づく方法を判別結果に適用する）

$$\Pr\left(u > \sqrt{\hat{\delta}\hat{\Sigma}^{-1}\hat{\delta}}/2\right) = \Pr\left(u > \sqrt{19.20}/2\right) = \Pr(u > 2.19)$$

$$= 0.0143$$

本来の健常者・患者を誤判別する確率：0.0143

変数選択

- 「吐き気」、「頭痛」のどちらの変数を取り入れるのか？

  吐き気：F_0 =8.168　　頭痛：F_0 =2.466

↓ 「吐き気」を取り入れる

- 「頭痛」を取り込む価値があるかどうか

  吐き気を取り込んだ状態での頭痛：F_0 =2.746

↓

「頭痛」も判別方式に取り込む意味がある

不明なサンプルの判別

$$\hat{z} = 12.80 + \begin{Bmatrix} 0 & (\text{吐き気が無}) \\ -9.60 & (\text{吐き気が少}) \\ -20.80 & (\text{吐き気が多}) \end{Bmatrix} + \begin{Bmatrix} 0 & (\text{頭痛が無}) \\ -6.40 & (\text{頭痛が少}) \\ -14.40 & (\text{頭痛が多}) \end{Bmatrix}$$

例：吐き気が無く、頭痛が多い人が属する母集団は？

⬇

得られた判別方式から、$\hat{z} = -1.60 > 0$ となるので、健常者と判別できる。

～変数に量的変数と質的変数が混在する場合～

変数の混在例

変数が混在している場合の健常者･患者の症状データ

サンプルNo.	健常者･患者	吐き気 x1	頭痛 x2	検査値1 x3	検査値2 x4
1	健常者	無	少	50	15.5
2	健常者	少	無	69	18.4
3	健常者	無	無	93	26.4
4	健常者	無	無	76	22.9
5	健常者	無	無	88	18.6
6	患者	少	多	43	16.9
7	患者	多	無	56	21.6
8	患者	少	少	38	12.2
9	患者	少	多	21	16
10	患者	多	少	25	10.5

解析の流れ

ダミー変数を適応後、これまでとまったく同様の解析方法を用いる。

サンプルNo.	健常者･患者	x1(2)	x1(3)	x2(2)	x2(3)	検査値1 x3	検査値2 x4
1	健常者	0	0	1	0	50	15.5
2	健常者	1	0	0	0	69	18.4
3	健常者	0	0	0	0	93	26.4
4	健常者	0	0	0	0	76	22.9
5	健常者	0	0	0	0	88	18.6
6	患者	1	0	0	1	43	16.9
7	患者	0	1	0	0	56	21.6
8	患者	1	0	1	0	38	12.2
9	患者	1	0	0	1	21	16
10	患者	0	1	1	0	25	10.5