データ解析（第1回）

静岡大学システム工学科

安藤 和敏

2004.10.06

本講義の内容について

多変量解析手法のうちで主なものに，

- 回帰分析法
- 判別分析法
- 主成分分析法
- 因子分析法

などがある．本講義においては，回帰分析法，判別分析法，主成分分析について学ぶ予定である．

多変量データとは何か?

データNo.	変数(変量)			
	x_1	x_2	$\cdots$	x_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

回帰分析法

浜松駅周辺の中古マンションのデータ

データNo.	広さ x_1 (m^2)	築年数 x_2 (年)	価格 y (千万円)
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

このデータについて，以下の事を知りたいとする．

1. 価格は広さと築年数によって，予測できるか．
2. 予測できるとすれば，その精度はどれくらいか．
3. 同じ地区で広さ$\mathbf{70}m^2$，築年数**10**年，価格**5.8**千万円のマンションを提示された．この価格は妥当か．

回帰分析法によって，以下の事が分かる．

1. 価格と広さと築年数は以下の関係にあると推定される．

$$y = 1.02 + 0.0668x_1 - 0.0808x_2$$

2. 寄与率は **0.933** で上式の精度は十分高い．

3. $x_1 = 70$，$x_2 = 10$ を代入すると，$y = 4.89$ となるので，
5.8 千万円は相場より高い．

判別分析とは

健常者・患者の検査値のデータ

サンプルNo.	健常者・患者	検査値1 x_1	検査値2 x_2
1	健常者	50	15.5
2	健常者	69	18.4
3	健常者	93	26.4
4	健常者	76	22.9
5	健常者	88	18.6
6	患者	43	16.9
7	患者	56	21.6
8	患者	38	12.2
9	患者	21	16.0
10	患者	25	10.5

このデータに基づいて知りたいことは以下の通りである.

1. 疾病にかかっているか否かを検査値1と検査値2から判別できるか.
2. 判別できるとすれば，その精度はどれくらいか.
3. 例えば，$x_1 = 70, x_2 = 19.0$ならどのように判別されるか.

判別分析によって以下のことがわかる.

1. 判別式$\hat{z} = -8.843 + 0.158x_1$が求まって，$\hat{z} \geq 0$ならば健常者，$\hat{z} < 0$なら患者と判別する.
2. 本当は健常者なのに患者と誤判別する確率は**0.1075**，本当は患者なのに健常者と誤判別する確率も**0.1075**.
3. $\hat{z} = -8.843 + 0.158x_1$に$x_1 = 70$を代入すると$\hat{z} \geq 0$となるので，健常者と判別される.

主成分分析法とは

試験の成績のデータ

生徒No.	国語 x_1	英語 x_2	数学 x_3	理科 x_4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91

1. 主成分の構成により低い次元でデータを解釈できないか.

2. それぞれの主成分の説明力はどれくらいか.

3. 科目や生徒の特徴付け及び分類をどのようにできるか.

1. 主要な主成分として第1主成分z_1と第2主成分z_2を得る.

$$z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$
$$z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$

ここで，u_jはx_jを標準化したものである.

2. z_1の寄与率は **0.680** で，z_2の寄与率は **0.306**である．第2主成分までの累積寄与率は **0.680＋0.306＝0.986**である.

3. 係数の値より，z_1は「総合的学力」を，z_2は「理系と文系の学力の違い」を表すと解釈できる.

教科書

永田靖，棟近雅彦：多変量解析法入門．サイエンス社，2001年．

参考書

田中豊，脇本和昌：多変量統計解析法．現代数学社，1983年．

Webページ

http://coconut.sys.eng.shizuoka.ac.jp/data/

第２章 統計的方法の基礎知識

2.1 データのまとめ方

(1) 1つの量的変数の場合

変数xに関するn個のデータ

$$x_1, x_2, \cdots, x_n$$

が与えられているとする．これらのデータに対していくつかの基本統計量が定義される．

平均

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

平方和

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \tag{2.2}$$

分散

$$V_x = \frac{S_{xx}}{n-1} \tag{2.3}$$

標準偏差

$$s_x = \sqrt{V_x} \tag{2.4}$$

範囲

$$R_x = x_{\mathsf{max}} - x_{\mathsf{min}} \text{ (最大値 } - \text{ 最小値)} \tag{2.5}$$

標準化 以下の計算を標準化と呼ぶ.

$$u_i = \frac{x_i - \bar{x}}{s_x} \quad (i = 1, \cdots, n) \tag{2.6}$$

標準化されたデータ $u_1, \cdots, u_n$ の平均は0，分散は1である:

$$\bar{u} = 0, \quad V_u = 1 \tag{2.7}$$