# 多変量解析

## 第8回　数量化1類

萩原•篠田
情報理工学部

# 数量化理論とは

順序尺度や名義尺度などの質的変数を
量的変数に変換して多変量解析を行う手法

| | | |
|---|---|---|
| 重回帰分析 | ←→ | 数量化1類 |
| 判別分析 | ←→ | 数量化2類 |
| 主成分分析 | ←→ | 数量化3類 |

# 数量化1類

量的変数 外的基準　　質的変数 アイテム

卒業時の総合成績は線形代数の成績とサークル所属の有無から予測可能か？

| サンプル No. | 線形代数 $x_1$ | サークル $x_2$ | 総合成績 y |
|---|---|---|---|
| 1 | 優 | 所属 | 96 |
| 2 | 優 | 所属 | 88 |
| 3 | 優 | 無所属 | 77 |
| 4 | 優 | 無所属 | 89 |
| 5 | 良 | 所属 | 80 |
| 6 | 良 | 無所属 | 71 |
| 7 | 良 | 無所属 | 77 |
| 8 | 可 | 所属 | 78 |
| 9 | 可 | 所属 | 70 |
| 10 | 可 | 無所属 | 62 |

keywords

質的変数、重回帰分析、ダミー変数、共線性、予測式、外的基準、カテゴリ数量、基準化

# 数量化1類

量的変数　外的基準　　質的変数　アイテム

卒業時の総合成績は線形代数の成績とサークル所属の有無から予測可能か？

ダミー変数導入で量的変数に変換→ 重回帰分析

回帰式：$\hat{y} = 83.0 - 10.0x_{11} - 19.0x_{12} + 9.0x_{21}$

| サンプル No. | 線形代数 $x_{11}$ | 線形代数 $x_{12}$ | サークル $x_{21}$ | 総合成績 y |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 96 |
| 2 | 0 | 0 | 1 | 88 |
| 3 | 0 | 0 | 0 | 77 |
| 4 | 0 | 0 | 0 | 89 |
| 5 | 1 | 0 | 1 | 80 |
| 6 | 1 | 0 | 0 | 71 |
| 7 | 1 | 0 | 0 | 77 |
| 8 | 0 | 1 | 1 | 78 |
| 9 | 0 | 1 | 1 | 70 |
| 10 | 0 | 1 | 0 | 62 |

keywords

質的変数、重回帰分析、ダミー変数、共線性、予測式、外的基準、カテゴリ数量、基準化

↓

予測式（回帰式）の定数や係数

# 数量化理論で用いる用語

外的基準：特性値（ex 重回帰分析での目的変数(従属変数)）
アイテム　：質的変数(ex アンケートでの質問項目)
カテゴリ　：アイテムの中身(ex アンケートでの回答)

アンケート調査票

項目1
あなたは野菜が好きですか
　1　はい　　2　いいえ

項目2
あなたはタンパク質が好きですか
　1　はい　　2　いいえ

項目3
あなたの体重は何Kgですか
　　＿＿＿＿＿Kg

アイテム　カテゴリ　外的基準

| 被験者No. | 野菜 | | タンパク質 | | 体重 |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | |
| 1 | レ | | | レ | 57 |
| 2 | レ | | レ | | 65 |
| 3 | | レ | レ | | 51 |
| 4 | レ | | レ | | 54 |
| 5 | | レ | レ | | 45 |
| 6 | レ | | | レ | 67 |

# 数量化1類

質的変数から定量的に測定される量的変数（外的基準）の予測や関係を調べる

Q1: 野菜やタンパク質の「好き／嫌い」から体重を予測する（予測式の導出）
Q2: 上記の予測式の精度は？（予測式の当てはまりの良さ）
Q3: 野菜とタンパク質のどちらが体重に影響を及ぼしているか？

アンケート調査票

項目1
あなたは野菜が好きですか
1 はい　2 いいえ

項目2
あなたはタンパク質が好きですか
1 はい　2 いいえ

項目3
あなたの体重は何Kgですか
＿＿＿＿＿Kg

アイテム　カテゴリ　外的基準

| 被験者No. | 野菜 | | タンパク質 | | 体重 |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | |
| 1 | レ | | | レ | 57 |
| 2 | レ | | レ | | 65 |
| 3 | | レ | レ | | 51 |
| 4 | レ | | レ | | 54 |
| 5 | | レ | レ | | 45 |
| 6 | レ | | | レ | 67 |

# 数量化1類の予測式（アイテムと外的基準の1次式）

ダミー変数 $x_{ij}$ を導入

$$x_{ij} = \begin{cases} 1 & \cdots \text{アイテム } i \text{ のカテゴリ } j \text{に反応したとき} \\ 0 & \cdots \text{その他} \end{cases}$$

| アイテム1　野菜 | | アイテム2　タンパク質 | |
|---|---|---|---|
| カテゴリ1<br>はい | カテゴリ2<br>いいえ | カテゴリ1<br>はい | カテゴリ2<br>いいえ |
| ⬆ | ⬆ | ⬆ | ⬆ |
| $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ |

予測式：$Y = b_{11}x_{11} + b_{12}x_{12} + b_{21}x_{21} + b_{22}x_{22} + b_0$　（$b_{ij}$：カテゴリ数量）

$x_{11}+x_{12} = 1$, $x_{21}+x_{22} = 1$ の関係式が成立している

→ 多重共線性（他のカテゴリから残りのカテゴリが予測可能）

$$Y = (b_{12} - b_{11})x_{12} + (b_{22} - b_{21})x_{22} + b_0 + b_{11} + b_{21}$$

$$\boldsymbol{Y = c_{12}x_{12} + c_{22}x_{22} + c_0}$$

を求めることにする　ただし $\begin{cases} c_{12} = b_{12} - b_{11} \\ c_{22} = b_{22} - b_{21} \\ c_0 = b_0 + b_{11} + b_{21} \end{cases}$

カテゴリ数量 $c_{ij}$と定数項$c_0$の決定方法　　$\boldsymbol{Y = c_{12}x_{12} + c_{22}x_{22} + c_0}$

外的基準 $y$の最適予測　→　$y$ と予測値 $Y$ との差をできるだけ小さくする

「$Q =$ (外的基準 $y -$ 予測値 $Y$)の2乗和」を最小にする $c_{ij}$と$c_0$を求める

→ 最小二乗法

| 被験者 No. | 外的基準 $y$ | アイテム1 | | アイテム2 | | 予測値 $Y$ | 誤差 $y - Y$ |
|---|---|---|---|---|---|---|---|
| | | $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ | | |
| 1 | 57 | 1 | 0 | 0 | 1 | $c_{22} + c_0$ | $57 - c_{22} + c_0$ |
| 2 | 65 | 1 | 0 | 1 | 0 | $c_0$ | $65 - c_0$ |
| 3 | 51 | 0 | 1 | 1 | 0 | $c_{12} + c_0$ | $51 - c_{12} - c_0$ |
| 4 | 54 | 1 | 0 | 1 | 0 | $c_0$ | $54 - c_0$ |
| 5 | 45 | 0 | 1 | 1 | 0 | $c_{12} + c_0$ | $45 - c_{12} - c_0$ |
| 6 | 67 | 1 | 0 | 0 | 1 | $c_{22} + c_0$ | $67 - c_{22} - c_0$ |

$Q = \Sigma$ (外的基準 $y -$ 予測値 $Y$) $^2$

$$= (57 - c_{22} - c_0)^2 + (65 - c_0)^2 + (51 - c_{12} - c_0)^2 + (54 - c_0)^2 + (45 - c_{12} - c_0)^2 + (67 - c_{22} - c_0)^2$$

$Q$が最小 → $c_{12}, c_{22}, c_0$による$Q$の偏微分$=0$

$$\begin{cases} \dfrac{\partial Q}{\partial c_{12}} = -2(51 - c_{12} - c_0) - 2(45 - c_{12} - c_0) = 0 \\ \dfrac{\partial Q}{\partial c_{22}} = -2(57 - c_{22} - c_0) - 2(67 - c_{22} - c_0) = 0 \\ \dfrac{\partial Q}{\partial c_0} = -2(57 - c_{22} - c_0) - 2(65 - c_0) - 2(51 - c_{12} - c_0) \\ \qquad -2(54 - c_0) - 2(45 - c_{12} - c_0) - 2(67 - c_{22} - c_0) = 0 \end{cases}$$

発展課題：
分散共分散行列を用いて予測式を求めよ.

整理して $\begin{cases} 48 - c_{12} - c_0 = 0 \\ 62 - c_{22} - c_0 = 0 \\ 339 - 2c_{12} - 2c_{22} - c_0 = 0 \end{cases}$ これを解いて $\begin{cases} c_{12} = -11.5 \\ c_{22} = 2.5 \\ c_0 = 59.5 \end{cases}$

予測式

$$Y = -11.5x_{12} + 2.5x_{22} + 59.5$$
$$= b_{11}x_{11} + b_{12}x_{12} + b_{21}x_{21} + b_{22}x_{22} + b_0$$

$$\begin{cases} c_{12} = b_{12} - b_{11} \\ c_{22} = b_{22} - b_{21} \\ c_0 = b_0 + b_{11} + b_{21} \end{cases}$$

どのようにして$b_{ij}, b_0$の式に変換するか？

→ 基準化（アイテム内のカテゴリ数量の平均が0となるようにする）

## 基準化（アイテム内のカテゴリ数量の平均が0となるようにする）

| 被験者 No. | 外的基準 $y$ | アイテム1 | | アイテム2 | | 予測値 $Y$ |
|---|---|---|---|---|---|---|
| | | $x_{11}$ | $x_{12}$ | $x_{21}$ | $x_{22}$ | |
| 1 | 57 | 1 | 0 | 0 | 1 | $c_{22}+c_0=62$ |
| 2 | 65 | 1 | 0 | 1 | 0 | $c_0=59.5$ |
| 3 | 51 | 0 | 1 | 1 | 0 | $c_{12}+c_0=48$ |
| 4 | 54 | 1 | 0 | 1 | 0 | $c_0=59.5$ |
| 5 | 45 | 0 | 1 | 1 | 0 | $c_{12}+c_0=48$ |
| 6 | 67 | 1 | 0 | 0 | 1 | $c_{22}+c_0=62$ |
| 平均 | 56.5 | 4/6 | 2/6 | 4/6 | 2/6 | 56.5 |

アイテム内のカテゴリ数量の平均が0

$$\begin{cases} \bar{x}_{11}b_{11}+\bar{x}_{12}b_{12}=\dfrac{4}{6}b_{11}+\dfrac{2}{6}b_{12}=0 \\ \bar{x}_{21}b_{21}+\bar{x}_{22}b_{22}=\dfrac{4}{6}b_{21}+\dfrac{2}{6}b_{22}=0 \end{cases}$$

$b_{ij}$ , $b_0$ と $c_{ij}$ , $c_0$ の関係式

$$\begin{cases} b_{12}-b_{11}=c_{12}=-11.5 \\ b_{22}-b_{21}=c_{22}=2.5 \\ b_0+b_{11}+b_{21}=c_0=59.5 \end{cases}$$

上記の5式を解いて $b_{ij}$ , $b_0$ を得る

$$\boldsymbol{Y=3.833x_{11}-7.667x_{12}-0.833x_{21}+1.667x_{22}+56.5}$$

$$(Y=-11.5x_{12}+2.5x_{22}+59.5,\ x_{11}+x_{12}=1,\ x_{21}+x_{22}=1)$$

カテゴリ数量の範囲の大きいアイテムほど予測値に大きな影響を与える

➡ 範囲 ＝ アイテムが外的基準に及ぼす影響の大きさ

アイテム「野菜」の範囲　　$=b_{11}-b_{12}=3.833-(-7.667)=11.5$

アイテム「タンパク質」の範囲 $=b_{21}-b_{22}=1.667-(-0.833)=2.5$

➡ 体重に及ぼす影響：タンパク質の好き嫌い ＜ 野菜の好き嫌い

予測式がどれだけ良く外的基準を推定できているか（当てはまりの良さ）

- 重相関係数$R$＝実測値と予測値の相関係数 $=\dfrac{Cov(y,Y)}{\sqrt{Var(y)Var(Y)}}=\dfrac{\sum(y_i-\bar{y})(Y_i-\bar{Y})}{\sqrt{\sum(y_i-\bar{y})^2\sum(Y_i-\bar{Y})^2}}$

- 決定係数$R^2$＝$\dfrac{\text{予測値の偏差平方和}}{\text{実測値の偏差平方和}}=\dfrac{\text{予測値の分散}}{\text{実測値の分散}}$＝重相関係数$^2$

| 被験者 No. | 野菜 | | タンパク質 | | 体重 実測値 | 体重 予測値 | $(y_i-\bar{y})^2$ | $(Y_i-\bar{Y})^2$ | $(y_i-\bar{y})(Y_i-\bar{Y})$ |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | | | | | |
| 1 | 1 | 0 | 0 | 1 | 57 | 62 | 0.25 | 30.25 | 2.75 |
| 2 | 1 | 0 | 1 | 0 | 65 | 59.5 | 72.25 | 9 | 25.5 |
| 3 | 0 | 1 | 1 | 0 | 51 | 48 | 30.25 | 72.25 | 46.75 |
| 4 | 1 | 0 | 1 | 0 | 54 | 59.5 | 6.25 | 9 | -7.5 |
| 5 | 0 | 1 | 1 | 0 | 45 | 48 | 132.25 | 72.25 | 97.75 |
| 6 | 1 | 0 | 0 | 1 | 67 | 62 | 110.25 | 30.25 | 57.75 |
| 和 | 4 | 2 | 4 | 2 | 339 | 339 | 351.5 | 223 | 223 |
| 平均 | 2/3 | 1/3 | 2/3 | 1/3 | 56.5 | 56.5 | | | |
| 分散/共分散 | | | | | | | 70.3 | 44.6 | 44.6 |

$R=0.797$

$R^2=0.634$

## ダミー変数により量的変数に変換後，
## 重回帰分析と同じく分散共分散行列を用いて回帰式を得ることができる

重回帰式 $Y = a_0 + a_1 x_1 + a_2 x_2$ を分散共分散行列を用いて求める方法

データ(説明変数$x_1, x_2$, 目的変数$y$ )による分散共分散行列と
重回帰式(説明変数$x_1, x_2$, 重回帰式$Y$ )による分散共分散行列を比較する

| | $x_1$ | $x_2$ | $y$ |
|---|---|---|---|
| $x_1$ | 分散 | 共分散 | 共分散 |
| $x_2$ | 共分散 | 分散 | 共分散 |
| $y$ | 共分散 | 共分散 | 分散 |

分散(Variance)　Var

共分散(Covariance)　Cov

重回帰式による分散共分散行列

| | $x_1$ | $x_2$ | $Y$ |
|---|---|---|---|
| $x_1$ | $\mathrm{Var}(x_1)$ | $\mathrm{Cov}(x_1, x_2)$ | $\mathrm{Cov}(x_1, Y)$ |
| $x_2$ | $\mathrm{Cov}(x_1, x_2)$ | $\mathrm{Var}(x_2)$ | $\mathrm{Cov}(x_2, Y)$ |
| $Y$ | $\mathrm{Cov}(x_1, Y)$ | $\mathrm{Cov}(x_2, Y)$ | $\mathrm{Var}(Y)$ |

$=$

データによる分散共分散行列

| | $x_1$ | $x_2$ | $y$ |
|---|---|---|---|
| $x_1$ | $\mathrm{Var}(x_1)$ | $\mathrm{Cov}(x_1, x_2)$ | $\mathrm{Cov}(x_1, y)$ |
| $x_2$ | $\mathrm{Cov}(x_1, x_2)$ | $\mathrm{Var}(x_2)$ | $\mathrm{Cov}(x_2, y)$ |
| $y$ | $\mathrm{Cov}(x_1, y)$ | $\mathrm{Cov}(x_2, y)$ | $\mathrm{Var}(y)$ |

重回帰式 $Y = a_0 + a_1 x_1 + a_2 x_2$ を求める方法　　□ の部分を比較してみる

重回帰式による分散共分散行列

$$\begin{array}{c} \\ x_1 \\ x_2 \\ Y \end{array}\begin{array}{c} \begin{array}{ccc} x_1 & x_2 & Y \end{array} \\ \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2) & \mathrm{Cov}(x_1,Y) \\ \mathrm{Cov}(x_1,x_2) & \mathrm{Var}(x_2) & \mathrm{Cov}(x_2,Y) \\ \mathrm{Cov}(x_1,Y) & \mathrm{Cov}(x_2,Y) & \mathrm{Var}(Y) \end{bmatrix} \end{array} =$$

データによる分散共分散行列

$$\begin{array}{c} \\ x_1 \\ x_2 \\ y \end{array}\begin{array}{c} \begin{array}{ccc} x_1 & x_2 & y \end{array} \\ \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2) & \mathrm{Cov}(x_1,y) \\ \mathrm{Cov}(x_1,x_2) & \mathrm{Var}(x_2) & \mathrm{Cov}(x_2,y) \\ \mathrm{Cov}(x_1,y) & \mathrm{Cov}(x_2,y) & \mathrm{Var}(y) \end{bmatrix} \end{array}$$

$\mathrm{Cov}(x, ax + by + c) = a\mathrm{Var}(x) + b\mathrm{Cov}(x,y)$ より

$$\mathrm{Cov}(x_1, Y) = \mathrm{Cov}(x_1, a_0 + a_1 x_1 + a_2 x_2) = a_1 \mathrm{Var}(x_1) + a_2 \mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_1, y)$$
$$\mathrm{Cov}(x_2, Y) = \mathrm{Cov}(x_2, a_0 + a_1 x_1 + a_2 x_2) = a_1 \mathrm{Cov}(x_1, x_2) + a_2 \mathrm{Var}(x_2) = \mathrm{Cov}(x_2, y)$$

$$\begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2) \\ \mathrm{Cov}(x_1,x_2) & \mathrm{Var}(x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(x_1,y) \\ \mathrm{Cov}(x_2,y) \end{bmatrix} \Rightarrow \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2) \\ \mathrm{Cov}(x_1,x_2) & \mathrm{Var}(x_2) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(x_1,y) \\ \mathrm{Cov}(x_2,y) \end{bmatrix}$$

説明変数が$p$個の場合

目的変数を$y$、説明変数を$x_1, x_2 \cdots, x_p$とすると

| | 変数 | | | | |
|---|---|---|---|---|---|
| | $y$ | $x_1$ | $x_2$ | $\cdots$ | $x_p$ |
| 個体 1 | $y_1$ | $x_{11}$ | $x_{21}$ | $\cdots$ | $x_{p1}$ |
| 2 | $y_2$ | $x_{12}$ | $x_{22}$ | $\cdots$ | $x_{p2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| $n$ | $y_n$ | $x_{1n}$ | $x_{2n}$ | $\cdots$ | $x_{pn}$ |
| 平均 | $\bar{y}$ | $\bar{x}_1$ | $\bar{x}_2$ | $\cdots$ | $\bar{x}_p$ |

分散共分散行列

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_p \\ y \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & \cdots & x_p & y \end{array} \\ \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} & s_{1y} \\ s_{12} & s_2^2 & \cdots & s_{2p} & s_{2y} \\ \vdots & \vdots & & \vdots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 & s_{py} \\ s_{1y} & s_{2y} & \cdots & s_{py} & s_y^2 \end{bmatrix} \end{array}$$

重回帰式 $Y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p$　　$Y_i = a_0 + a_1x_{1i} + a_2x_{2i} + \cdots + a_px_{pi}$

回帰係数 $a_1, a_2, \cdots, a_p$ と定数項 $a_0$ は次の連立方程式からもとまる

$$\begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{12} & \cdots & s_p^2 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} s_{1y} \\ s_{2y} \\ \vdots \\ s_{py} \end{bmatrix}$$

$$\Rightarrow \quad \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{12} & \cdots & s_p^2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} s_{1y} \\ s_{2y} \\ \vdots \\ s_{py} \end{bmatrix}$$

$\bar{y} = a_0 + a_1\bar{x}_1 + a_2\bar{x}_2 + \cdots + a_p\bar{x}_p$

$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - \cdots - a_p\bar{x}_p$

# 数量化1類

① 数量化1類とは何か。
② アイテムと外的基準の関係式を1次式で表現せよ。
　予測式：Y、ダミー変数：$x_{ij}$、カテゴリ数量：$b_{ij}$、定数項：$b_0$を用いよ。
③ カテゴリ数量$b_{ij}$と定数項$b_0$の決定方法を説明せよ（ダミー変数間の共線性考慮）。
④ カテゴリー数量の基準化を行い予測式を求めよ。
⑤ アイテムの外的基準に与える影響の大きさを検討せよ。
⑥ 予測式の性能（当てはまりの良さ）を検討せよ。

アンケート調査票

項目1
あなたは野菜が好きですか
　1　はい　　2　いいえ

項目2
あなたはタンパク質が好きですか
　1　はい　　2　いいえ

項目3
あなたの体重は何Kgですか
　　＿＿＿＿＿Kg

アイテム　カテゴリ　外的基準

| 被験者 No. | 野菜 | | タンパク質 | | 体重 |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | |
| 1 | レ | | | レ | 57 |
| 2 | レ | | レ | | 65 |
| 3 | | レ | レ | | 51 |
| 4 | レ | | レ | | 54 |
| 5 | | レ | レ | | 45 |
| 6 | レ | | | レ | 67 |