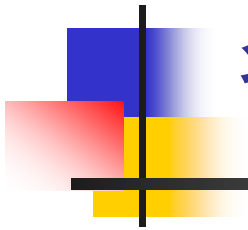

# 多変量解析　–数量化１類–

発表日：4月28日

発表者：加藤 友宏

# 数量化一類とは？

- 目的変数が量的変数で、
  説明変数が質的変数の場合の解析方法
- 質的変数（数値変数ではないもの）を
  数量化して分析する方法の一つ
- 実際のデータにおいて、説明変数に質的変数と量的変数が混在していることもある

# 解析の流れ

1. 質的変数をダミー変数に変換して、ダミー変数を量的変数と考えて重回帰モデルを想定。
2. 自由度調整済寄与率を求めて、得られた回帰式の性能を評価。
3. 説明変数の選択（変数選択）を行い、有効な変数を選択。
4. 残差とテコ比の検討を行い、得られた回帰式の妥当性を検討。
5. 得られた回帰式をより、将来得られるデータの値を予測。

質的変数をダミー変数に変換するところ以外は、重回帰分析

# 数量化一類の具体例

表は、大学卒業時の総合成績（量的変数）と線形代数の成績（質的変数）
サークル所属の有無（質的変数）のデータ

| サンプルNo. | 線形代数x1 | サークルx2 | 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 |

このデータから、
「総合成績は線形代数の成績およびサークルの所属の有無により予測できるか」
「どちらの変数のほうが説明力があるか」
などを検討する。

# ダミー変数の考え方（1）

「線形代数の成績」など質的な変数をアイテムと呼ぶ。
アイテムの中身（「優」、「良」、「可」）をカテゴリーと呼ぶ。

☆説明変数が1つの場合を考える。
（「線形代数の成績」が「総合成績」に関係するか）

- 「線形代数の成績」を量的変数に変換してみよう。

「優」　「良」　「可」
↓　　↓　　↓
1　　2　　3
というように割り当てるのはダメ。

3－2＝1と2－1＝1は同じとは言えない

# ダミー変数の考え方（2）

- 質的変数は、0と1だけの値を取る変数に変換

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

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

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

常に $x_{1(1)} + x_{1(2)} + x_{1(3)} = 1$ が成立するため

多重共線性が生じる

# ダミー変数の考え方（3）

- 変数のうち$x_{1(1)}$を削除し、重回帰モデルを想定

$$y_i = \beta_0 + \beta_{1(2)} x_{i1(2)} + \beta_{1(3)} x_{i1(3)} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

ここで、 $x_{1(2)} = x_{1(3)} = 0$ が「優」を意味する

$x_{1(2)}, x_{1(3)}$ をダミー変数と呼び、一般的に
「（質的変数のカテゴリー数）－1」個導入する

# ダミー変数を考慮した表

- 線形代数の部分をダミー変数に置き換えた表

| サンプルNo. | x1(2) | x1(3) | 総合成績y |
|---|---|---|---|
| 1 | 0 | 0 | 96 |
| 2 | 0 | 0 | 88 |
| 3 | 0 | 0 | 77 |
| 4 | 0 | 0 | 89 |
| 5 | 1 | 0 | 80 |
| 6 | 1 | 0 | 71 |
| 7 | 1 | 0 | 77 |
| 8 | 0 | 1 | 78 |
| 9 | 0 | 1 | 70 |
| 10 | 0 | 1 | 62 |

x1(2)＝x1(3)＝0が優、x1(2)＝1が良、x1(3)＝が可を示す。

# 予測値、残差平方和、重相関係数、寄与率、自由度調整済寄与率

予測値：$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_{1(2)} x_{i1(2)} + \hat{\beta}_{1(3)} x_{i1(3)}$

残差平方和：$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)$

重相関係数：$R = \dfrac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$

寄与率：$R^2 = \dfrac{S_R}{S_{yy}} \left( = \dfrac{S_{yy} - S_e}{S_{yy}} = 1 - \dfrac{S_e}{S_{yy}} \right)$

自由度調整済寄与率：$R^{*2} = 1 - \dfrac{S_e / (n-3)}{S_{yy} / (n-1)}$

※n-3＝n-(ダミー変数の個数)-1

# 説明変数の選択

- 一つの質的変数に対応する「ダミー変数の集まり」が、目的変数に効いているかを検討する

・定数項のみのモデル

$$MODEL0: y_i = \beta_0 + \varepsilon_i$$

・両方のダミー変数を含めたモデル

$$MODEL1: \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_{1(2)} x_{i1(2)} + \hat{\beta}_{1(3)} x_{i1(3)}$$

MODEL0が正しいときに、$F\left(\phi_T - \phi_{e(M1)}, \phi_{e(M1)}\right)$に従う

$$F_0 = \frac{(S_{yy} - S_{e(M1)}) / (\phi_T - \phi_{e(M1)})}{S_{e(M1)} / \phi_{e(M1)}}$$

**$F_0$値が大きければ、「線形代数の成績」を取り込む**

# 説明変数が2個以上の場合（1）

- 「線形代数の成績」と「サークルの所属の有無」で、「総合成績」にどんな関係があるかを考える。

目的変数：「総合成績ｙ」
説明変数：「線形代数の成績ｘ１」、「サークルｘ２」

「線形代数の成績」と「サークル」は質的変数なので、ダミー変数に変換する。

# 説明変数が2個以上の場合（2）

- 線形代数のダミー変数は、先ほどと同様

$$x_{1(2)} = \begin{cases} 1 & \text{良のとき} \\ 0 & \text{良でないとき} \end{cases} \qquad x_{1(3)} = \begin{cases} 1 & \text{可のとき} \\ 0 & \text{可でないとき} \end{cases}$$

- サークルに関しては、
  カテゴリーが2つなのでダミー変数は1つ

$$x_2 = \begin{cases} 1 & \text{所属のとき} \\ 0 & \text{無所属のとき} \end{cases}$$

これらのダミー変数を考慮した重回帰モデルは、

$$y_i = \beta_0 + \beta_{1(2)} x_{i1(2)} + \beta_{1(3)} x_{i1(3)} + \beta_2 x_{i2} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

# ダミー変数（3つ）を考慮した表

- 線形代数の部分とサークルの部分を
  ダミー変数に置き換えた表

| サンプルNo. | x1(2) | x1(3) | x2 | 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 |

x1(2)＝x1(3)＝0が優、x1(2)＝1が良、x1(3)＝が可、
x2＝1がサークル所属、x2＝0がサークル無所属を示す。

# 質的変数と量的変数が混在する場合

- 表は、先ほどの表に量的変数の「通学時間」を追加したものである。

| サンプルNo. | 線形代数x1 | サークルx2 | 通学時間x3 | 総合成績y |
|---|---|---|---|---|
| 1 | 優 | 所属 | 15 | 96 |
| 2 | 優 | 所属 | 85 | 88 |
| 3 | 優 | 無所属 | 78 | 77 |
| 4 | 優 | 無所属 | 15 | 89 |
| 5 | 良 | 所属 | 57 | 80 |
| 6 | 良 | 無所属 | 29 | 71 |
| 7 | 良 | 無所属 | 64 | 77 |
| 8 | 可 | 所属 | 22 | 78 |
| 9 | 可 | 所属 | 57 | 70 |
| 10 | 可 | 無所属 | 50 | 62 |

先ほどと同様にダミー変数を考慮した重回帰モデルは、

$$y_i = \beta_0 + \beta_{1(2)} x_{i1(2)} + \beta_{1(3)} x_{i1(3)} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$