多変量解析
第 1 回: 多変量解析の概要

藤本 衡

2018 年 4 月 11 日

本日の内容

- 講義の進め方
- 多変量とは？
- 変量の関係と得られうる性質
- 多変量解析の手法を眺めてみる

講義の進め方

- 教科書無し、資料配布
- 評価：中間 30%, 期末 50%, 平常点 20%
  - 期末が 100 点満点中 80 点以上なら中間・平常が駄目でも評価します
  - 出席は取りますが親御さんを安心させるためだけです
  - 平常点は宿題・小テストなどでつけます

参考書

- 資料作成のネタ本 (朝倉からは何ももらってないですよ！)
  - 柳井晴夫, 「多変量データ解析法」, 朝倉書店, 1994.
  - 水野欽司, 「多変量データ解析講義」, 朝倉書店, 1996.
  - 圓川隆夫, 「多変量のデータ解析」, 朝倉書店, 1998.
  - 長畑秀和, 「R で学ぶ多変量解析」, 朝倉書店, 2017.
- 自分で探すとき
  - 「多変量」＋「解析」で検索
  - 特定の手法については、その分析手法で検索してもよい
  - タイトルに「医療」とか「建設」とかあると例題が偏りがち
  - SPSS で説明している本もある (大学環境なら SPSS 利用可能？)

変量 (variate) とデータ (datum)

- **変量**: variate
  - **個別の対象ごとに異なる (可能性がある) 数量**
  - **たとえば身長や体重は個々人が持つ変量と言える**
  - **大文字で書くことが多い (身長を X, 体重を Y のように)**
  - **個体 i の身長・体重は X_i, Y_i のように書く**
  - **昔は「変数」と明確に区別していたが最近はそうでもない**
- **データ**: datum
  - **変量を実際に測定した結果**
  - **測定値** (measured value), **観測値** (observed value), **実現値** (realized value) **とも**
  - data **は複数形 (データはたくさんあるので普通複数形を使う)**

駆け足で統計のおさらい

- たくさんの個体
- 1 つの変量がどう散らばっているか？
  - 度数分布表、ヒストグラム、累積分布
  - 平均値、中央値、最頻値
  - 分散、標準偏差
- 2 つの変量 がどう散らばっているか？
  - 共分散、相関係数、散布図

多変量 (multivariate) とは？

- 個体 i の身長・体重 X_i, Y_i は「2 変量」(bivariate)
- 個体から得られる 2 以上の変量をまとめて扱う時に用いる語
- i さんの身長と j さんの身長を比較する時は「多変量」とは呼ばない

多変量を扱う意味

- 身長と体重を独立な数量として見た場合
  - それぞれに平均や分散を取っておしまい
- 身長と体重に何らかの依存 (相関) があると仮定した場合
  - 例：Y_i は X_i の 1 次関数として表現できると仮定
    $\Longrightarrow Y_i = aX_i + b$ が成り立つ（数理モデル）
  - a, b はどんな値だろう？→最小二乗法
- 変量間の相互関係を知る／活用することが目的
- できるだけ表現を簡略化 (元の変量よりも数を減らす) ことが望ましい

多変量の関係を表す数理モデル

- モデルの妥当性に関する議論
  - 身長と体重は本当に関係があるのか？独立では無いのか？
  - 1 次比例なのか？ ($\Longrightarrow$ 多分 3 次比例だよね…)
- 求めたい性質、関係に適したモデルと解析手法
  - どの変量が強く関係しているかわからないのに比例関係を仮定してよいのか？

多変量解析の性質と分類

- **量的変数と質的変数**
  - **質的変数**: Yes/No, **男女**, **所属組織など**, **名義・順序尺度**
  - **質的変数に対する解析では便宜的に数値化する必要がある (数量化理論)**
- **目的変数** (criterion)
  - **他の変数 (説明変数, 独立変数) を用いて表現したい対象**
  - **外的基準 (変数), 従属変数とも呼ばれる**
  - **目的変数が無いモデルもある**
- **潜在変数** (latent variable)
  - **実際には観測が困難だが, モデルに入れておくと解析に有利な変数**

これから学ぶ解析手法

1. 回帰分析: (線形) 単回帰, (線形) 重回帰, 非線形回帰
   - 量的目的変数あり
2. 主成分分析
   - 目的変数なし: 量的独立変数間の関係を求める
3. 因子分析
   - 量的目的変数あり, 潜在変数あり: 量的独立変数の影響を求める
4. 判別分析
   - 質的目的変数あり: 個体の分類
5. ベイズ統計とトピックモデル