回帰分析

回帰モデルの考え方と推定

村田 昇

講義概要

- **第 1 回 : 回帰モデルの考え方と推定**
- 第 2 回 : モデルの評価
- 第 3 回 : モデルによる予測と発展的なモデル

回帰分析の考え方

回帰分析

- ある変量を別の変量で説明する関係式を構成する
- 関係式 : **回帰式** (regression equation)
  - 説明される側 : **目的変数**, 被説明変数, 従属変数, 応答変数
  - 説明する側 : **説明変数**, 独立変数, 共変量
- 説明変数の数による分類
  - 一つの場合 : **単回帰** (simple regression)
  - 複数の場合 : **重回帰** (multiple regression)

一般の回帰の枠組

- **説明変数** : $x_1, \dots, x_p$ (p 次元)
- **目的変数** : y (1 次元)
- **回帰式** : y を $x_1, \dots, x_p$ で説明するための関係式

$$y = f(x_1, \dots, x_p)$$

- 観測データ : n 個の $(y, x_1, \dots, x_p)$ の組

$$\{(y_i, x_{i1}, \dots, x_{ip})\}_{i=1}^n$$

線形回帰

- 任意の f では一般的すぎて分析に不向き
- f として **1 次関数** を考える

  ある定数 $\beta_0, \beta_1, \ldots, \beta_p$ を用いた式 :

  $$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - 1 次関数の場合 : **線形回帰** (linear regression)
  - 一般の場合 : 非線形回帰 (nonlinear regression)
- 非線形関係は新たな説明変数の導入で対応可能
  - 適切な多項式 : $x_j^2, x_j x_k, x_j x_k x_l, \ldots$
  - その他の非線形変換 : $\log x_j, x_j^\alpha, \ldots$
  - 全ての非線形関係ではないことに注意

回帰係数

- 線形回帰式

  $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - $\beta_0, \beta_1, \ldots, \beta_p$: **回帰係数** (regression coefficients)
  - β_0 : **定数項 / 切片** (constant term / intersection)
- 線形回帰分析 (linear regression analysis)
  - 未知の回帰係数をデータから決定する分析方法
  - 決定された回帰係数の統計的な性質を診断

回帰の確率モデル

- 回帰式の不確定性
  - データは一般に観測誤差などランダムな変動を含む
  - 回帰式がそのまま成立することは期待できない
- 確率モデル : データのばらつきを表す項 ϵ_i を追加

  $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad (i = 1, \ldots, n)$$

  - $\epsilon_1, \ldots, \epsilon_n$: **誤差項 / 撹乱項** (error / disturbance term)
    - 誤差項は独立な確率変数と仮定
    - 多くの場合，平均 0，分散 σ^2 の正規分布を仮定
- **推定** (estimation) : 観測データから回帰係数を決定

回帰係数の推定

残差

- **残差** (residual) : 回帰式で説明できない変動
- 回帰係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\mathsf{T}$ を持つ回帰式の残差

  $$e_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \quad (i = 1, \ldots, n)$$

- 残差 $e_i(\boldsymbol{\beta})$ の絶対値が小さいほど当てはまりがよい

最小二乗法

- **残差平方和** (residual sum of squares)

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2$$

- **最小二乗推定量** (least squares estimator)

  残差平方和 $S(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^\top = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

行列の定義

- デザイン行列 (design matrix)

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

  - $n \times (p+1)$ 行列

ベクトルの定義

- 目的変数，誤差，回帰係数のベクトル

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

  - $\boldsymbol{y}, \boldsymbol{\epsilon}$ は n 次元ベクトル
  - $\boldsymbol{\beta}$ は p+1 次元ベクトル

行列・ベクトルによる表現

- 確率モデル

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- 残差平方和

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

解の条件

- 解 $\boldsymbol{\beta}$ では残差平方和の勾配は零ベクトル

$$\nabla S(\boldsymbol{\beta}) = \left(\frac{\partial S}{\partial \beta_0}(\boldsymbol{\beta}), \frac{\partial S}{\partial \beta_1}(\boldsymbol{\beta}), \dots, \frac{\partial S}{\partial \beta_p}(\boldsymbol{\beta})\right)^\top = \mathbf{0}$$

- 成分 $(j = 0, 1, \dots, p)$ ごとの条件式

$$\frac{\partial S}{\partial \beta_j}(\boldsymbol{\beta}) = -2\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{p}\beta_k x_{ik}\right)x_{ij} = 0$$

ただし $x_{i0} = 1 \ (i = 1, \dots, n)$

正規方程式

正規方程式

- **正規方程式** (normal equation)

$$X^\top X\boldsymbol{\beta} = X^\top \boldsymbol{y}$$

- $X^\top X$: **Gram 行列** (Gram matrix)
  - $(p+1) \times (p+1)$ 行列 (正方行列)
  - 正定対称行列 (固有値が非負)

正規方程式の解

- 正規方程式の基本的な性質
  - 正規方程式は必ず解をもつ (一意に決まらない場合もある)
  - 正規方程式の解は最小二乗推定量であるための必要条件
- 解の一意性の条件
  - Gram 行列 $X^\top X$ が **正則**
  - X の列ベクトルが独立 (後述)
- 正規方程式の解

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top \boldsymbol{y}$$

実習

R：関数 `lm()` による推定

- 線形モデルの推定

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
#' formula: 目的変数名 ~ 説明変数名 (複数ある場合は + で並べる)
#' data: 目的変数, 説明変数を含むデータフレーム
#' subset: 推定に用いるデータフレームの部分集合を指定 (指定しなければ全て)
```

```r
#' na.action: 欠損の扱いを指定 (既定値は option("na.action") で設定された処理)
#' model,x,y,qr: 返値に model.frame,model.matrix, 目的変数,QR 分解を含むか指定
```

データセットの準備

- 回帰分析では以下のデータセットを使用する
  - https://www.statlearning.com/s/Advertising.csv

    広告費 (TV,radio,newspapers) と売上の関係を調べたもの

    “Datasets in this presentation are taken from ”An Introduction to Statistical Learning, with applications in R“ (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani ”

    https://www.statlearning.com
  - tokyo_weather.csv

    気象庁より取得した東京の気候データを回帰分析用に整理したもの

    https://www.data.jma.go.jp/gmd/risk/obsdl/index.php

練習問題

- 前掲のデータセットを用いて回帰式を構成しなさい
  - 東京の 8 月の気候データ

    formula = temp ~ solar + press
  - 広告費と売上データ

    formula = sales ~ TV

    formula = sales ~ radio

    formula = sales ~ TV + radio

最小二乗推定量の性質

解析の上での良い条件

- 最小二乗推定量がただ一つだけ存在する条件
  - $X^{\mathsf{T}}X$ が正則
  - $X^{\mathsf{T}}X$ の階数が p+1
  - X の階数が p+1
  - X の列ベクトルが **1 次独立**

  これらは同値条件

解析の上での良くない条件

- 説明変数が 1 次従属 : **多重共線性** (multicollinearity)
- 多重共線性が強くならないように説明変数を選択
  - X の列 (説明変数) の独立性を担保する
  - 説明変数が互いに異なる情報をもつように選ぶ
  - 似た性質をもつ説明変数の重複は避ける

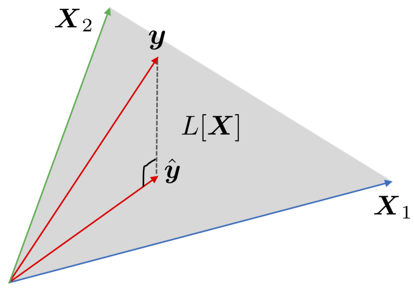

図 1: $n = 3$, $p + 1 = 2$ の場合の最小二乗法による推定

推定の幾何学的解釈

- **あてはめ値 / 予測値** (fitted values / predicted values)

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = \hat{\beta}_0 X_{\text{第 0 列}} + \cdots + \hat{\beta}_p X_{\text{第 p 列}}$$

- 最小二乗推定量 $\hat{\boldsymbol{y}}$ の幾何学的性質
  - $L[X]$: X の列ベクトルが張る $\mathbb{R}^n$ の線形部分空間
  - X の階数が p+1 ならば $L[X]$ の次元は p+1 (解の一意性)
  - $\hat{\boldsymbol{y}}$ は $\boldsymbol{y}$ の $L[X]$ への直交射影
  - **残差** (residuals) $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ はあてはめ値 $\hat{\boldsymbol{y}}$ に直交

$$\hat{\boldsymbol{\epsilon}} \cdot \hat{\boldsymbol{y}} = 0$$

線形回帰式と標本平均

- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$: i 番目の観測データの説明変数
- 説明変数および目的変数の標本平均

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i,$$

- $\hat{\boldsymbol{\beta}}$ が最小二乗推定量のとき以下が成立

$$\bar{y} = (1, \bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\beta}}$$

実習

R：推定結果からの情報の取得

- 関数 `lm()` の出力には様々な情報が含まれる

```r
#' lmの出力を引数とする関数の例
coef(lmの出力)          # 推定された回帰係数
fitted(lmの出力)        # あてはめ値
resid(lmの出力)         # 残差
model.frame(lmの出力)   # modelに必要な変数の抽出 (データフレーム)
model.matrix(lmの出力)  # デザイン行列
```

R：行列とベクトル

- データフレーム以外の重要なデータ構造
  - ベクトル (vector) : 1 次元の配列
  - 行列 (matrix) : 2 次元の同じデータ型の配列
- 必要であれば明示的に変換できる

```r
as.vector(データフレーム[列名]) # base::as.vector()
as_vector(データフレーム[列名]) # purrr::as_vector()
as.matrix(データフレーム) # base::as.matrix()
#' データフレームを行列に変換する場合は全て同じデータ型でなくてはならない
```

R：行列とベクトルの計算

- $X^{\mathsf{T}}Y$ および $X^{\mathsf{T}}X$ の計算

```r
crossprod(X, Y) # cross product の略
#' X: 行列 (またはベクトル)
#' Y: 行列 (またはベクトル)
crossprod(X) # 同じものを掛ける場合は引数は1つで良い
```

- 行列 A, B の積 AB およびベクトル a, b の内積 $a \cdot b$

```r
A %*% B # 行列の大きさは適切である必要がある
a %*% b # ベクトルは同じ長さである必要がある
```

- 正方行列 A の逆行列 A^{-1}

```r
solve(A) # 他にもいくつか関数はある
```

練習問題

- 前問の推定結果を用いて最小二乗推定量の性質を確認しなさい
  - 推定された係数が正規方程式の解

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\boldsymbol{y}$$

    となること
  - あてはめ値と残差が直交すること
  - 回帰式が標本平均を通ること

残差の分解

最小二乗推定量の残差

- 観測値と推定値 $\hat{\beta}$ による予測値の差

$$\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}) \quad (i = 1, \ldots, n)$$

  - 誤差項 $\epsilon_1, \ldots, \epsilon_n$ の推定値
  - 全てができるだけ小さいほど良い
  - 予測値とは独立に偏りがないほど良い

- 残差ベクトル

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n)^\top$$

平方和の分解

- $\bar{\boldsymbol{y}} = \bar{y}\boldsymbol{1} = (\bar{y}, \bar{y}, \ldots, \bar{y})^\top$: 標本平均のベクトル
- いろいろなばらつき
  - $S_y = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\top(\boldsymbol{y} - \bar{\boldsymbol{y}})$: 目的変数のばらつき
  - $S \ \ = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$: 残差のばらつき ($\hat{\boldsymbol{\epsilon}}^\top\hat{\boldsymbol{\epsilon}}$)
  - $S_r = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$: あてはめ値 (回帰) のばらつき
- 3 つのばらつき (平方和) の関係

$$\begin{aligned}
&(\boldsymbol{y} - \bar{\boldsymbol{y}})^\top(\boldsymbol{y} - \bar{\boldsymbol{y}}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) \\
&S_y = S + S_r
\end{aligned}$$

実習

練習問題

- 前問の結果を用いて残差の性質を確認しなさい
  - 以下の分解が成り立つこと

$$\begin{aligned}
&(\boldsymbol{y} - \bar{\boldsymbol{y}})^\top(\boldsymbol{y} - \bar{\boldsymbol{y}}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) \\
&S_y = S + S_r
\end{aligned}$$

決定係数

回帰式の寄与

- ばらつきの分解

$$S_y \ (\text{目的変数}) = S \ (\text{残差}) + S_r \ (\text{あてはめ値})$$

- 回帰式で説明できるばらつきの比率

$$(\text{回帰式の寄与率}) = \frac{S_r}{S_y} = 1 - \frac{S}{S_y}$$

- 回帰式のあてはまり具合を評価する代表的な指標

決定係数 (R^2 値)

- **決定係数** (R-squared)

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- **自由度調整済み決定係数** (adjusted R-squared)

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

  - 不偏分散で補正している

実習

練習問題

- 決定係数を用いてモデルの比較を行いなさい
  - 東京の 8 月の気候データ

    temp ~ solar

    temp ~ solar + press

    temp ~ solar + press + cloud
  - 広告費と売上データ

    sales ~ TV

    sales ~ radio

    sales ~ TV + radio

次回の予定

- 第 1 回 : 回帰モデルの考え方と推定
- **第 2 回 : モデルの評価**
- 第 3 回 : モデルによる予測と発展的なモデル