

回帰分析 - モデルの推定

数理科学統論J

(Press ? for help, n and p for next and previous slide)

村田 昇

2019.10.18

講義の予定

- **第1日: 回帰モデルの考え方と推定**
- 第2日: モデルの評価
- 第3日: 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 (regression analysis)

- ある変量を別の変量によって説明するための関係式を構成
- 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)
- 観測データ: n個の (y, x_1, \dots, x_p) の組

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

- y を x_1, \dots, x_p で説明するための関係式を構成:

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

一般には p変数関数 f を使う

線形回帰 (linear regression)

- 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 $\beta_0, \beta_1, \dots, \beta_p$ を用いた以下の式:

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

- 1次式の場合: **線形回帰 (linear regression)**
- 一般の場合: 非線形回帰 (nonlinear regression)
- 非線形な関係は新たな説明変数の導入で対応可能
 - 適切な多項式 $x_j^2, x_j x_k, x_j x_k x_l, \dots$
 - その他の非線形変換 $\log x_j, x_j^\alpha, \dots$

回帰係数

- 線形回帰式:

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

- $\beta_0, \beta_1, \dots, \beta_p$: **回帰係数 (regression coefficients)**
- β_0 : 定数項 (切片; constant term)
- 線形回帰分析: 未知の回帰係数をデータから決定

回帰の確率モデル

- 一般にデータは観測誤差などランダムな変動を含む
- 確率モデル: データのばらつきを表す項 ϵ_i を追加

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

- $\epsilon_1, \dots, \epsilon_n$: **誤差項/攪乱項 (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)**: 未知パラメータ $(\beta_0, \beta_1, \dots, \beta_p)$ を観測データから決定すること

回帰係数の推定

残差

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

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

- 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 (least squares)

- 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):

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

ベクトルの定義

- 目的変数, 誤差, 回帰係数のベクトル

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

行列・ベクトルによる表現

- 確率モデル:

$$y = X\beta + \epsilon$$

- 残差平方和:

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

解の条件

- 解 β では残差平方和の勾配は零ベクトル

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

- 成分 ($j = 0, 1, \dots, p$) ごとの条件式

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

但し $x_{i0} = 1$ ($i = 1, \dots, n$)

正規方程式 (normal equation)

- 条件を整理 (x_{ij} は行列 X の (i, j) 成分)

$$\sum_{i=1}^n x_{ij} \left(\sum_{k=0}^p x_{ik} \beta_k \right) = \sum_{i=1}^n x_{ij} y_i \quad (j = 0, 1, \dots, p)$$

- 正規方程式 (normal equation):**

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

- $X^{\top} X$: **Gram行列 (Gram matrix)**

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
- Gram 行列 $X^T X$ が正則ならば解が一意に決定
- 正規方程式の解

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

R: 関数 `lm()` の使い方

- ベクトルを用いる基本的な使い方:
 - ベクトル `y`: 目的変数 y
 - ベクトル `x1, ..., xp`: 説明変数 x_1, \dots, x_p
- データフレームを用いる方法: **(こちらの使い方を推奨)**
 - データフレーム `mydata`: 目的変数, 説明変数を含むデータ
 - 列名: y の変数名, x_1 の変数名, ..., x_p の変数名

ベクトルを渡す場合

```
lm(y ~ x1 + ... + xp)
```

データフレームを渡す場合

```
lm(yの変数名 ~ x1の変数名 + ... + xpの変数名, data = mydata)
```

演習: 回帰式の推定

- [04-lm.r](#) を確認してみよう

最小二乗推定量の性質

解と観測データの関係

- 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 (以下同値条件)
 - $X^T X$ が正則
 - $X^T X$ 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
- 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)**
 - 説明変数は多重共線性が強くないように選択すべき
 - 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

推定の幾何学的解釈

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

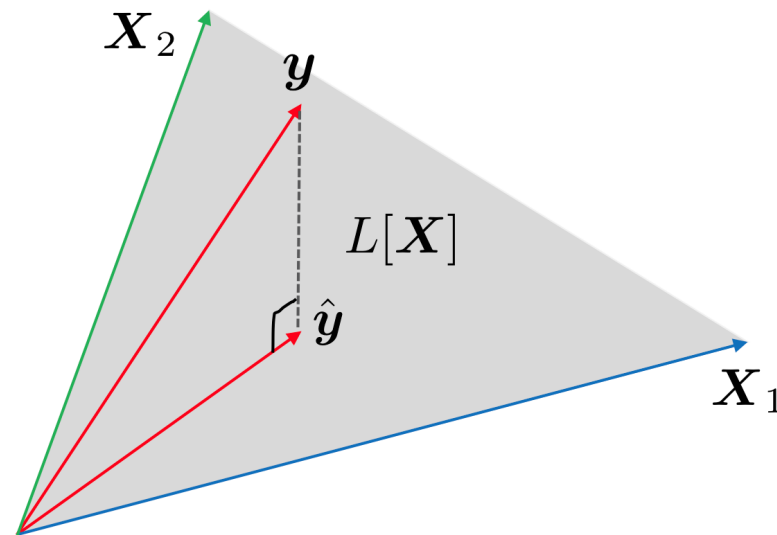
$$\hat{y} = X\hat{\beta} = \hat{\beta}_0 X_{\text{第0列}} + \cdots + \hat{\beta}_p X_{\text{第p列}}$$

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

$$\hat{e} \cdot \hat{y} = 0$$

- 幾何学的な考察からも一意に決まる

推定の幾何学的解釈



$n = 3, p + 1 = 2$ の場合の最小二乗法による推定

線形回帰式と標本平均

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

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

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

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

- 以下の関係から簡単に示すことができる:

$$\mathbf{1} \cdot (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{1} \cdot \hat{\boldsymbol{\epsilon}} = 0$$

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
- 情報を取り出すための関数が用意されている

```
## 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
model.matrix(lmの出力) # デザイン行列
```

演習: 最小二乗推定量の性質

- [04-lse.r](#) を確認してみよう

残差の性質

残差

- 観測値と推定値 $\hat{\beta}$ による予測値の差:

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

- 誤差項 $\epsilon_1, \dots, \epsilon_n$ 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無いほど良い
- 残差ベクトル

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

ばらつきの分解

- いろいろなばらつき

- $\bar{y} = \bar{y}\mathbf{1} = (\bar{y}, \bar{y}, \dots, \bar{y})^\top$: 標本平均のベクトル
- $S_y = (\mathbf{y} - \bar{y})^\top (\mathbf{y} - \bar{y})$: 目的変数のばらつき
- $S = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$: 残差のばらつき ($\hat{\mathbf{e}}^\top \hat{\mathbf{e}}$)
- $S_r = (\hat{\mathbf{y}} - \bar{y})^\top (\hat{\mathbf{y}} - \bar{y})$: あてはめ値(回帰)のばらつき

- 3つのばらつきの関係

$$(\mathbf{y} - \bar{y})^\top (\mathbf{y} - \bar{y}) = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{y})^\top (\hat{\mathbf{y}} - \bar{y})$$

$$S_y = S + S_r$$

分解に用いる残差の性質


- 証明には以下の関係を使う

$$y - \bar{y} = y - \hat{y} + \hat{y} - \bar{y}$$

$$\hat{y} \cdot (y - \hat{y}) = \hat{y} \cdot \hat{\epsilon} = 0$$

$$\mathbf{1} \cdot (y - \hat{y}) = \mathbf{1} \cdot \hat{\epsilon} = 0$$

演習: 残差の性質

- 04-resid.r  を確認してみよう

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

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

- 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

演習: 決定係数

- [04-rsq.r](#) を確認してみよう

演習

- 以下のデータで回帰分析を行ってみよう
 - `datasets::airquality`
 - `datasets::LifeCycleSavings`