回帰分析

回帰モデルの考え方と推定

村田 昇

講義の内容

- **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

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

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 $(\beta_0, \beta_1, \ldots, \beta_p)$ 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 $\boldsymbol{\beta}$

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

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

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

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

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

解の条件

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

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

演習

問題

- 残差平方和 $S(\boldsymbol{\beta})$ をベクトル $\boldsymbol{\beta}$ で微分して解の条件を求めなさい

正規方程式

正規方程式

- **正規方程式** (normal equation)

$$X^\mathsf{T} X\boldsymbol{\beta} = X^\mathsf{T}\boldsymbol{y}$$

- $X^\mathsf{T} X$: **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 $X^\mathsf{T} X$ が **正則**
  - X の列ベクトルが独立 (後述)
- 正規方程式の解

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

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

推定の幾何学的解釈

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

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

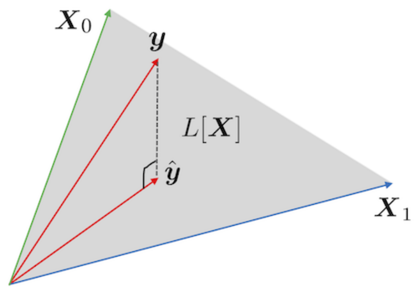

図 1: $n = 3$, $p + 1 = 2$ の場合の最小二乗法による推定

- 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)
  - $\hat{\boldsymbol{y}}$ 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

- $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^{\mathsf{T}}$: 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

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

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

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

演習

問題

- 最小二乗推定量について以下を示しなさい
  - 残差の標本平均が 0 となる

以下を示せばよい

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

ただし $\mathbf{1} = (1,\dots,1)^\top$ とする

- 回帰式が標本平均を通る

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

残差の分解

最小二乗推定量の残差

- 観測値と推定値 $\hat{\boldsymbol{\beta}}$ 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

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

平方和の分解

- $\bar{\boldsymbol{y}} = \bar{y}\mathbf{1} = (\bar{y},\bar{y},\dots,\bar{y})^\top$: 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

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

演習

問題

- 以下の関係式を示しなさい
  - あてはめ値と残差のベクトルが直交する

$$\hat{\boldsymbol{y}}^\top(\boldsymbol{y}-\hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}}^\top\hat{\boldsymbol{\epsilon}} = 0$$

  - 残差平方和の分解が成り立つ

$$S_y = S + S_r$$

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

解析の事例

データについて

- 気象庁より取得した東京の気候データ
  - 気象庁 https://www.data.jma.go.jp/gmd/risk/obsdl/index.php
  - データ https://noboru-murata.github.io/multivariate-analysis/data/tokyo_weather.csv

東京の 8 月の気候の分析

- 気候 (気温, 降雨, 日射, 降雪, 風向, 風速, 気圧, 湿度, 雲量)
  に関するデータ (の一部)

```
    month day day_of_week temp rain solar snow wdir wind  press humid cloud
213     8   1         Sun 28.7  0.0 26.58    0  SSE  3.2 1000.2    76   2.3
214     8   2         Mon 28.6  0.5 19.95    0   SE  3.4 1006.1    80   7.0
215     8   3         Tue 29.0  3.0 19.89    0    S  4.0 1009.9    80   6.3
216     8   4         Wed 29.5  0.0 26.52    0    S  3.0 1008.2    76   2.8
217     8   5         Thu 29.1  0.0 26.17    0  SSE  2.8 1005.1    74   5.8
218     8   6         Fri 29.1  0.0 24.82    0  SSE  2.9 1004.2    75   4.0
219     8   7         Sat 27.9  2.0 11.43    0   NE  2.5 1003.1    85   9.0
```

220	8	8	Sun	25.9	90.5	3.43	0	N	3.0	998.0	97	10.0
221	8	9	Mon	28.1	2.0	13.34	0	S	6.1	995.4	84	6.0
222	8	10	Tue	31.0	0.0	22.45	0	SSW	4.7	996.3	58	4.8
223	8	11	Wed	29.2	0.0	21.12	0	SE	2.9	1008.0	61	9.3
224	8	12	Thu	26.0	0.5	8.34	0	SSE	2.4	1008.8	84	9.5
225	8	13	Fri	22.5	20.5	4.36	0	NE	2.7	1008.0	97	10.0
226	8	14	Sat	22.3	77.0	2.76	0	N	2.7	1003.6	100	10.0

- 気温を説明する 4 つの線形回帰モデルを検討する
  - モデル 1: 気温 = F(気圧)
  - モデル 2: 気温 = F(気圧, 日射)
  - モデル 3: 気温 = F(気圧, 日射, 湿度)
  - モデル 4: 気温 = F(気圧, 日射, 雲量)
- 関連するデータの散布図

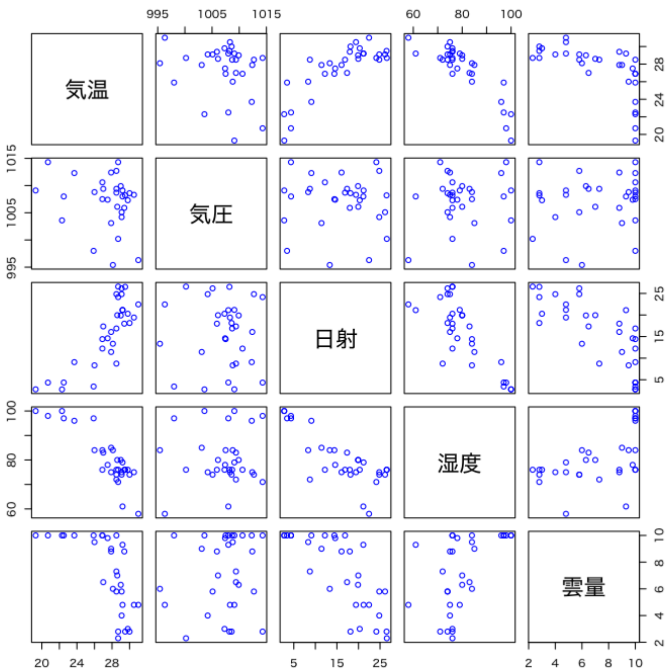

図 2: 散布図

- モデル 1 の推定結果

- モデル 2 の推定結果

- 観測値とあてはめ値の比較

- 決定係数・自由度調整済み決定係数
  - モデル 1: 気温 = F(気圧)

    [1] "R2: 0.0483 ; adj. R2: 0.0155"
  - モデル 2: 気温 = F(気圧, 日射)

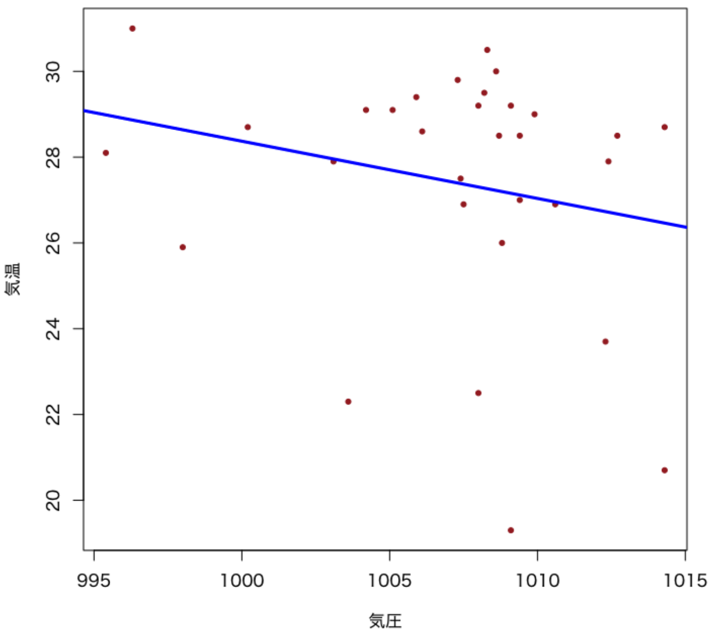

図 3: モデル 1

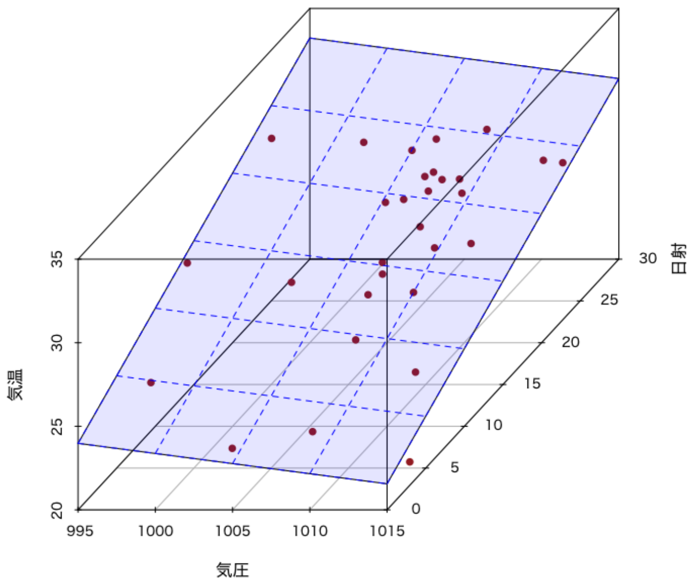

図 4: モデル 2

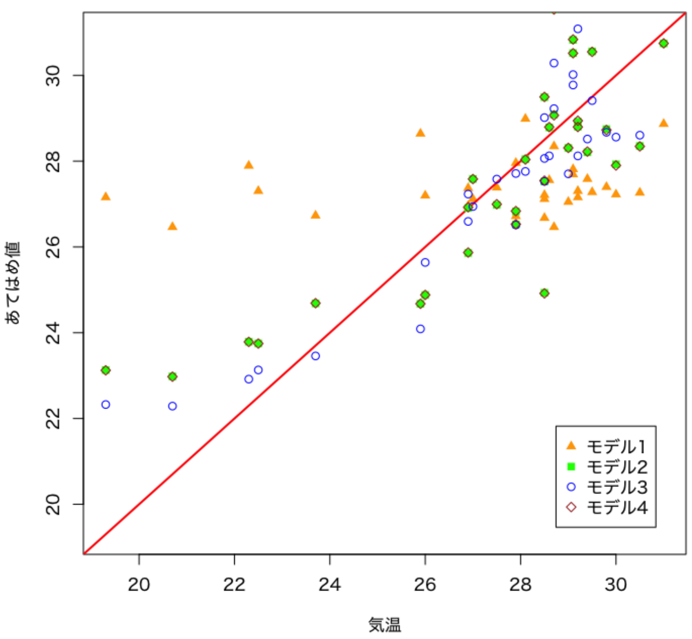

図 5: モデルの比較

```
[1] "R2: 0.703 ; adj. R2: 0.681"
```

- モデル 3: 気温 = F(気圧, 日射, 湿度)

```
[1] "R2: 0.83 ; adj. R2: 0.811"
```

- モデル 4: 気温 = F(気圧, 日射, 雲量)

```
[1] "R2: 0.703 ; adj. R2: 0.67"
```

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