

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
- 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 $\beta_0, \beta_1, \dots, \beta_p$ を用いた式:

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

- 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 $x_j^2, x_j x_k, x_j x_k x_l, \dots$
 - その他の非線形変換 $\log x_j, x_j^\alpha, \dots$
 - 全ての非線形関係ではない

回帰係数

- 線形回帰式:

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

- $\beta_0, \beta_1, \dots, \beta_p$: **回帰係数** (regression coefficients)
- β_0 : **定数項 / 切片** (constant term / 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

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

- $\epsilon_1, \dots, \epsilon_n$: **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): 観測データから $(\beta_0, \beta_1, \dots, \beta_p)$ 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 $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ を持つ回帰式の残差:

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

- 残差 $e_i(\beta)$ 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):

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

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:

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

- 残差平方和:

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

解の条件

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

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

演習

問題

- 残差平方和 $S(\beta)$ をベクトル β 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):

$$X^T X \beta = X^T y$$

- **Gram 行列** (Gram matrix): $X^T X$
 - $(p+1) \times (p+1)$ 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 $X^T X$ が **正則**
 - X の列ベクトルが独立 (後述)
- 正規方程式の解:

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

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
 - $X^T X$ 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**
- 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):

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

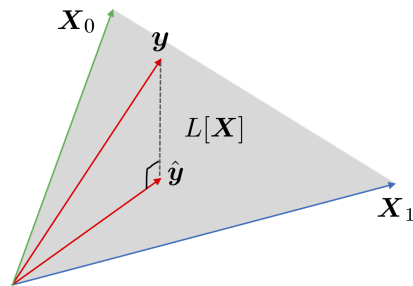


図 1: $n = 3, p + 1 = 2$ の場合の最小二乗法による推定

- 最小二乗推定量 \hat{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{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

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

線形回帰式と標本平均

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

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

- $\hat{\beta}$ が最小二乗推定量のとき以下が成立:

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

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

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

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

- 回帰式が標本平均を通る

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

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:

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

- 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:

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

平方和の分解

- 標本平均のベクトル: $\bar{\mathbf{y}} = \bar{y}\mathbf{1} = (\bar{y}, \bar{y}, \dots, \bar{y})^T$
- いろいろなばらつき
 - $S_y = (\mathbf{y} - \bar{\mathbf{y}})^T(\mathbf{y} - \bar{\mathbf{y}})$: 目的変数のばらつき
 - $S = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$: 残差のばらつき ($\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}$)
 - $S_r = (\hat{\mathbf{y}} - \bar{\mathbf{y}})^T(\hat{\mathbf{y}} - \bar{\mathbf{y}})$: 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

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

$$S_y = S + S_r$$

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

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

- 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
214	8	1	Sat	26.1	0.5	19.79	0	NE	2.6	1009.3	77	7.8
215	8	2	Sun	26.3	0.0	19.53	0	SSE	2.4	1011.0	75	5.5
216	8	3	Mon	27.2	0.0	24.73	0	SSE	2.4	1011.0	74	3.8
217	8	4	Tue	28.3	0.0	24.49	0	SSE	2.9	1012.2	77	4.3
218	8	5	Wed	29.1	0.0	24.93	0	S	2.9	1013.4	76	3.3
219	8	6	Thu	28.5	0.0	24.02	0	SSE	3.9	1010.5	79	7.8
220	8	7	Fri	29.5	0.0	22.58	0	S	3.4	1005.0	71	7.5

221	8	8	Sat	28.1	0.0	15.49	0	SE	2.7	1006.1	79	8.3
222	8	9	Sun	28.7	0.0	19.96	0	SSE	2.4	1006.9	77	9.5
223	8	10	Mon	30.5	0.0	20.26	0	SE	2.4	1010.3	73	10.0
224	8	11	Tue	31.7	0.0	25.50	0	S	4.0	1009.7	67	2.8
225	8	12	Wed	30.0	0.5	18.24	0	SSE	2.5	1009.0	79	6.8
226	8	13	Thu	29.4	21.5	19.01	0	N	2.2	1006.4	82	5.0
227	8	14	Fri	29.4	0.0	19.85	0	SE	2.8	1005.5	78	2.0

- 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

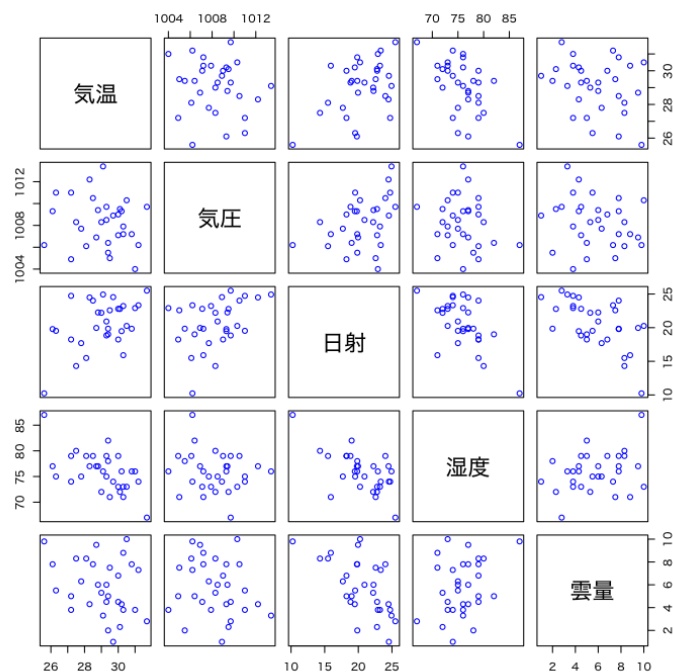


図 2: 散布図

- 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: 0.0169 ; adj. R2: -0.017"
```
 - モデル2: 気温 = F(気圧, 日射)

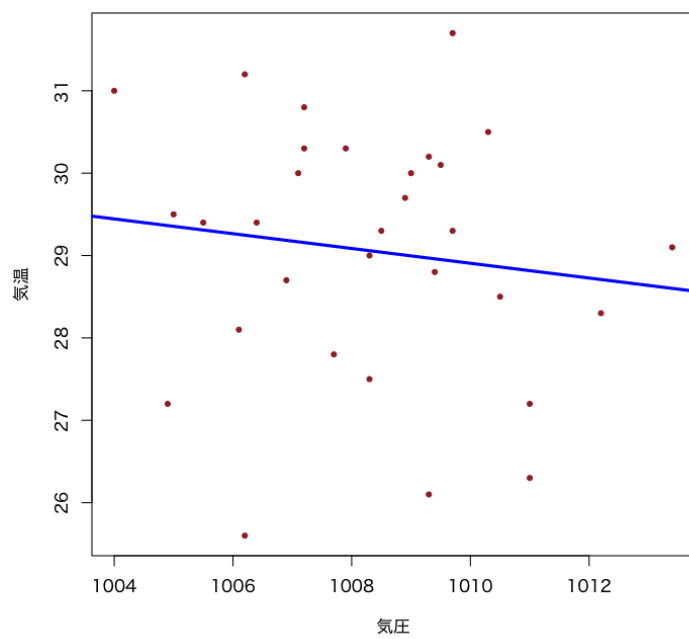


図 3: モデル 1

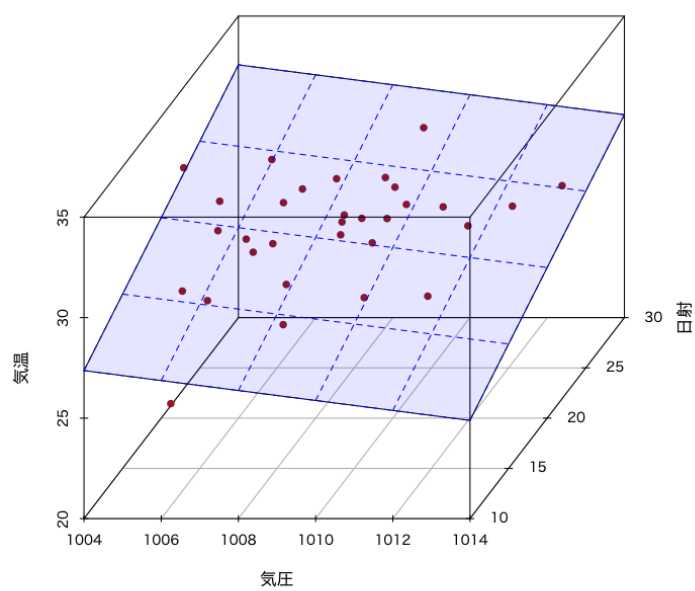


図 4: モデル 2

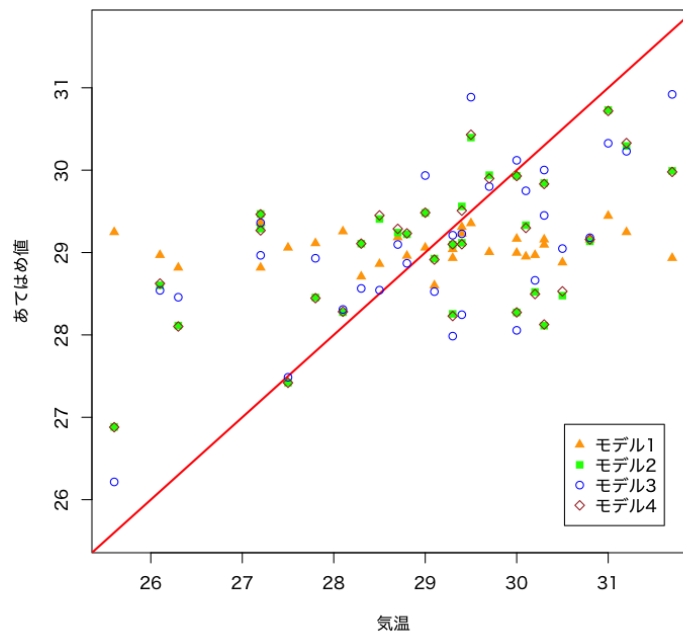


図 5: モデルの比較

- [1] "R2: 0.32 ; adj. R2: 0.271"
- モデル 3: 気温 = F(気圧, 日射, 湿度)
- [1] "R2: 0.422 ; adj. R2: 0.358"
- モデル 4: 気温 = F(気圧, 日射, 雲量)
- [1] "R2: 0.32 ; adj. R2: 0.245"

次週の予定

- 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