

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 $\beta_0, \beta_1, \dots, \beta_p$ 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ことに注意

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
 - 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

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

- $\epsilon_1, \dots, \epsilon_n$ ：**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 $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ 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)

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

– $n \times (p+1)$ 行列

ベクトルの定義

- 目的変数, 誤差, 回帰係数のベクトル

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

– $\mathbf{y}, \boldsymbol{\epsilon}$ 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

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

- 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

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

演習

問題

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

解答例

- 残差平方和を展開しておく

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y} - X\boldsymbol{\beta})^\top (\mathbf{y} - X\boldsymbol{\beta}) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top X\boldsymbol{\beta} - (X\boldsymbol{\beta})^\top \mathbf{y} + (X\boldsymbol{\beta})^\top X\boldsymbol{\beta} \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top X\boldsymbol{\beta} - \boldsymbol{\beta}^\top X^\top \mathbf{y} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta} \end{aligned}$$

- ベクトルによる微分を行うと以下ようになる

$$\begin{aligned} \frac{\partial S}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) &= -(\mathbf{y}^\top X)^\top - X^\top \mathbf{y} + (X^\top X + (X^\top X)^\top)\boldsymbol{\beta} \\ &= -2X^\top \mathbf{y} + 2X^\top X\boldsymbol{\beta} \end{aligned}$$

- したがって $\boldsymbol{\beta}$ の満たす条件は以下となる

$$\begin{aligned} -2X^\top \mathbf{y} + 2X^\top X\boldsymbol{\beta} &= 0 \quad \text{より} \\ X^\top X\boldsymbol{\beta} &= X^\top \mathbf{y} \end{aligned}$$

補足

- 成分ごとの計算は以下ようになる

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

ただし, $x_{i0} = 1$ ($i = 1, \dots, n$), $j = 0, 1, \dots, p$

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

x_{ij} は行列 X の (i, j) 成分であることに注意

正規方程式

正規方程式

- 正規方程式 (normal equation)

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

- $X^\top X$: **Gram 行列** (Gram matrix)
 - $(p+1) \times (p+1)$ 行列 (正方行列)
 - 半正定値行列 (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)

$$\hat{y} = X\hat{\beta} = \hat{\beta}_0 X_{0(\text{切片})} + \cdots + \hat{\beta}_p X_{p(\text{第 } p \text{ 成分})}$$

- 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{y} は y の $L[X]$ への直交射影
 - **残差** (residuals) $\hat{\epsilon} = y - \hat{y}$ はあてはめ値 \hat{y} に直交

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

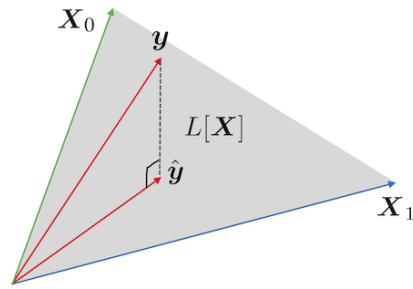


Figure 1: $n = 3, p + 1 = 2$ の場合の最小二乗法による推定

線形回帰式と標本平均

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

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

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

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

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
- 目的変数や残差のベクトルについて以下を示せばよい

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

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

- 回帰式が標本平均を通る

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

解答例

- 残差の表現を整理する

$$\begin{aligned} \hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \end{aligned}$$

- 左から \mathbf{X}^\top を乗じる

$$X^T y - X^T X (X^T X)^{-1} X^T y = X^T y - X^T y = \mathbf{0}$$

- 行列 X の 1 列目が $\mathbf{1}$ であることより明らか
- 説明変数の標本平均をデザイン行列で表す

$$\mathbf{1}^T X = n(1, \bar{x}^T)$$

- したがって以下が成立する

$$\begin{aligned} n(1, \bar{x}^T) \hat{\beta} &= \mathbf{1}^T X \hat{\beta} \\ &= \mathbf{1}^T \hat{y} = \mathbf{1}^T y \\ &= n\bar{y} \end{aligned}$$

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

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

- 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

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

平方和の分解

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

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

$$S_y = S + S_r$$

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

$$S_y = S + S_r$$

解答例

- 残差の表現を整理する

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}\end{aligned}$$

- 左から $\hat{\mathbf{y}}$ を乗じる

$$\begin{aligned}\hat{\mathbf{y}}^T \hat{\boldsymbol{\epsilon}} &= \hat{\boldsymbol{\beta}}^T \mathbf{X}^T (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y} \\ &= \hat{\boldsymbol{\beta}}^T (\mathbf{X}^T - \mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y} \\ &= \hat{\boldsymbol{\beta}}^T (\mathbf{X}^T - \mathbf{X}^T) \mathbf{y} = 0\end{aligned}$$

- 以下の関係を用いて展開すればよい

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

$$\text{ただし } \bar{\mathbf{y}} = \bar{y} \mathbf{1}$$

- このとき以下の項は 0 になる

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

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

実データによる例

- 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

気温に影響を与える要因の分析

- データの概要
- 気温を説明する 5 種類の線形回帰モデルを検討
 - モデル 1 : 気温 = F(気圧)
 - モデル 2 : 気温 = F(日射)
 - モデル 3 : 気温 = F(気圧, 日射)
 - モデル 4 : 気温 = F(気圧, 日射, 湿度)
 - モデル 5 : 気温 = F(気圧, 日射, 雲量)

分析の視覚化

- 関連するデータの散布図
- モデル 1 の推定結果
- モデル 2 の推定結果
- モデル 3 の推定結果
- 観測値とあてはめ値の比較

モデルの比較

- 決定係数 (R^2 , Adjusted R^2)

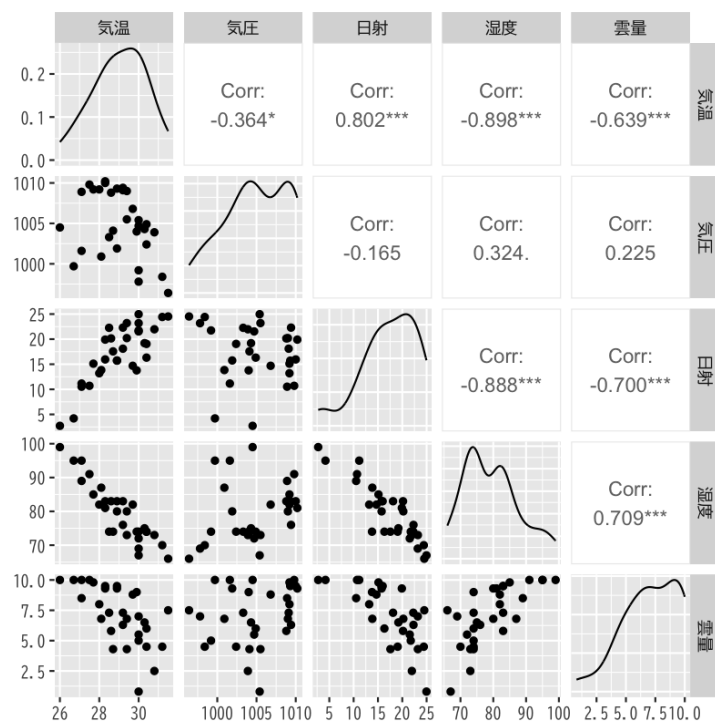


Figure 2: 散布図

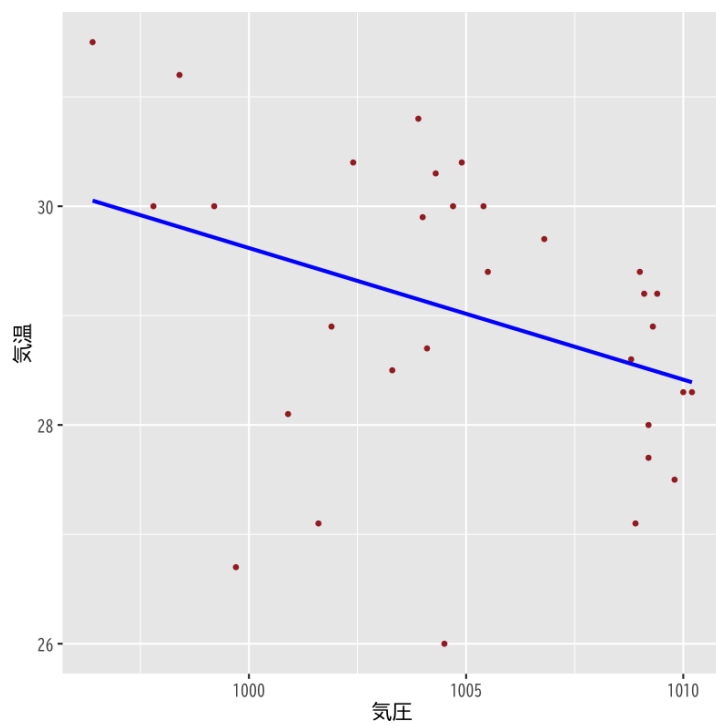


Figure 3: モデル 1

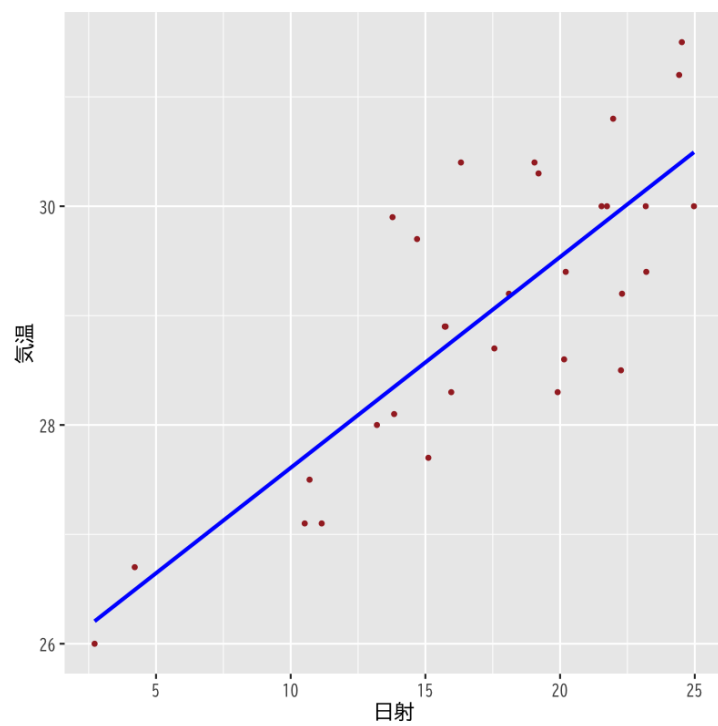


Figure 4: モデル 2

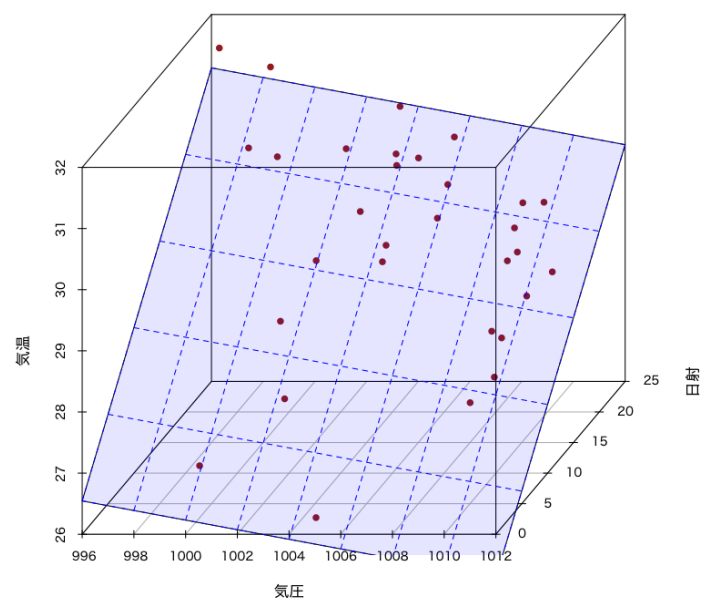


Figure 5: モデル 3

| 日付 | 気温 | 降雨 | 日射 | 降雪 | 風向 | 風速 | 気圧 | 湿度 | 雲量 |
|------------|------|------|-------|----|-----|-----|--------|----|------|
| 2024-08-01 | 28.5 | 0.0 | 22.26 | 0 | NE | 2.4 | 1003.3 | 74 | 7.3 |
| 2024-08-02 | 28.7 | 0.0 | 17.56 | 0 | SSE | 2.6 | 1004.1 | 74 | 4.3 |
| 2024-08-03 | 29.4 | 0.0 | 23.20 | 0 | SSE | 2.6 | 1005.5 | 73 | 4.3 |
| 2024-08-04 | 30.0 | 0.0 | 24.97 | 0 | SSE | 2.5 | 1005.4 | 67 | 0.8 |
| 2024-08-05 | 30.0 | 0.0 | 21.54 | 0 | SSE | 2.6 | 1004.7 | 72 | 5.5 |
| 2024-08-06 | 29.9 | 0.0 | 13.78 | 0 | SE | 2.3 | 1004.0 | 74 | 9.0 |
| 2024-08-07 | 28.9 | 76.5 | 15.75 | 0 | NNE | 2.6 | 1001.9 | 80 | 9.3 |
| 2024-08-08 | 28.1 | 0.0 | 13.84 | 0 | NW | 2.2 | 1000.9 | 87 | 6.8 |
| 2024-08-09 | 30.0 | 0.0 | 21.74 | 0 | NE | 2.7 | 999.2 | 74 | 5.0 |
| 2024-08-10 | 30.0 | 0.0 | 23.18 | 0 | N | 2.7 | 997.8 | 69 | 7.0 |
| 2024-08-11 | 31.5 | 0.0 | 24.52 | 0 | WNW | 2.8 | 996.4 | 66 | 7.5 |
| 2024-08-12 | 31.2 | 0.0 | 24.42 | 0 | SSE | 3.9 | 998.4 | 70 | 4.5 |
| 2024-08-13 | 30.8 | 0.0 | 21.97 | 0 | SSE | 3.6 | 1003.9 | 73 | 2.5 |
| 2024-08-14 | 30.4 | 0.0 | 16.32 | 0 | S | 2.7 | 1004.9 | 74 | 6.0 |
| 2024-08-15 | 30.3 | 0.0 | 19.20 | 0 | ESE | 2.7 | 1004.3 | 75 | 6.5 |
| 2024-08-16 | 26.7 | 90.0 | 4.21 | 0 | NNE | 4.2 | 999.7 | 95 | 10.0 |
| 2024-08-17 | 30.4 | 0.0 | 19.05 | 0 | S | 2.7 | 1002.4 | 74 | 4.5 |
| 2024-08-18 | 29.7 | 0.0 | 14.69 | 0 | SSE | 2.5 | 1006.8 | 82 | 8.8 |
| 2024-08-19 | 29.2 | 22.0 | 18.10 | 0 | NW | 2.4 | 1009.1 | 83 | 7.3 |
| 2024-08-20 | 28.3 | 0.5 | 19.91 | 0 | S | 2.4 | 1010.2 | 81 | 9.3 |
| 2024-08-21 | 28.3 | 21.0 | 15.96 | 0 | SSE | 2.8 | 1010.0 | 83 | 9.5 |
| 2024-08-22 | 27.5 | 16.0 | 10.70 | 0 | SSE | 3.1 | 1009.8 | 91 | 10.0 |
| 2024-08-23 | 28.9 | 0.0 | 15.72 | 0 | S | 4.0 | 1009.3 | 83 | 9.5 |
| 2024-08-24 | 29.4 | 1.0 | 20.21 | 0 | S | 3.2 | 1009.0 | 80 | 6.8 |
| 2024-08-25 | 28.6 | 1.5 | 20.15 | 0 | SSE | 3.5 | 1008.8 | 83 | 5.8 |
| 2024-08-26 | 29.2 | 0.0 | 22.30 | 0 | S | 4.4 | 1009.4 | 76 | 6.3 |
| 2024-08-27 | 27.7 | 30.0 | 15.11 | 0 | S | 4.7 | 1009.2 | 85 | 9.8 |
| 2024-08-28 | 28.0 | 0.0 | 13.20 | 0 | SSE | 3.5 | 1009.2 | 82 | 8.0 |
| 2024-08-29 | 27.1 | 23.0 | 10.52 | 0 | SSE | 2.7 | 1008.9 | 89 | 8.5 |
| 2024-08-30 | 26.0 | 84.0 | 2.72 | 0 | SSE | 3.2 | 1004.5 | 99 | 10.0 |
| 2024-08-31 | 27.1 | 15.5 | 11.15 | 0 | SSE | 3.6 | 1001.6 | 95 | 10.0 |

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

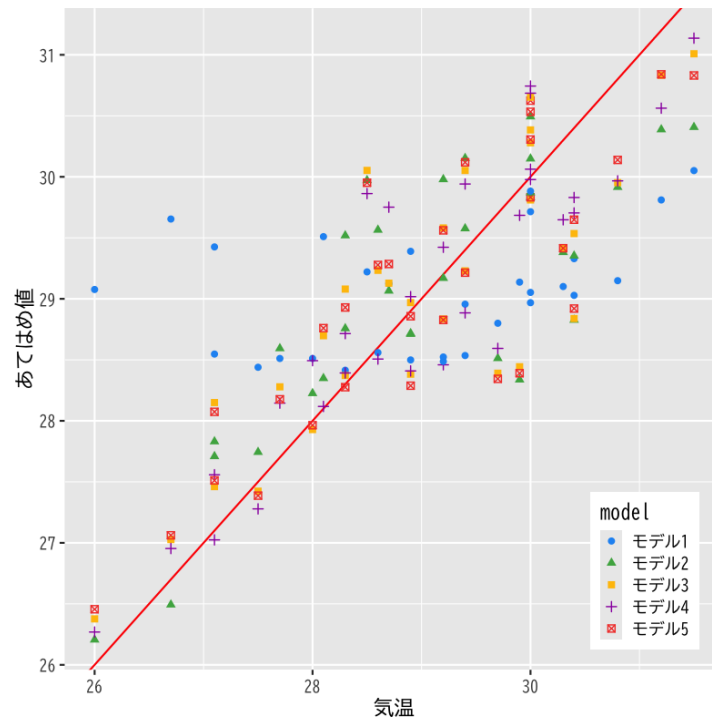


Figure 6: モデルの比較

| Characteristic | モデル 1 | | モデル 2 | | モデル 3 | | モデル 4 | | モデル 5 | |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | Beta | SE | Beta | SE | Beta | SE | Beta | SE | Beta | SE |
| 気圧 | -0.12 | 0.057 | | | -0.08 | 0.035 | -0.03 | 0.030 | -0.07 | 0.036 |
| 日射 | | | 0.19 | 0.027 | 0.18 | 0.025 | 0.02 | 0.045 | 0.17 | 0.035 |
| 湿度 | | | | | | | -0.13 | 0.031 | | |
| 雲量 | | | | | | | | | -0.06 | 0.083 |
| R ² | 0.132 | | 0.644 | | 0.699 | | 0.813 | | 0.704 | |
| Adjusted R ² | 0.102 | | 0.631 | | 0.677 | | 0.792 | | 0.671 | |

Abbreviations: CI = Confidence Interval, SE = Standard Error