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

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

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

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

- ベクトルによる微分を行うと以下のようになる

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

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

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

補足

- 成分ごとの計算は以下のようになる

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

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

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

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)

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

- $X^\top X$: **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

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

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
  - $X^{\top}X$ が正則
  - $X^{\top}X$ 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)

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

- 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

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

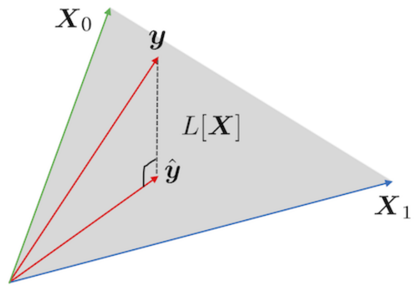

図 1: $n = 3$, $p + 1 = 2$ の場合の最小二乗法による推定

線形回帰式と標本平均

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

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

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

    目的変数や残差のベクトルについて以下を示せばよい

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

    ただし $\mathbf{1} = (1, \dots, 1)^\mathsf{T}$ とする
  - 回帰式が標本平均を通る

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

解答例

- 残差の表現を整理する

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

- 左から X^T を乗じる

$$X^\top \boldsymbol{y} - X^\top X(X^\top X)^{-1}X^\top \boldsymbol{y} = X^\top \boldsymbol{y} - X^\top \boldsymbol{y} = 0$$

- 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

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

- したがって以下が成立する

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

残差の分解

最小二乗推定量の残差

- 観測値と推定値 $\hat{\boldsymbol{\beta}}$ による予測値の差

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

  - 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

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

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

$$S_y = S + S_r$$

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

$$S_y = S + S_r$$

解答例

- 残差の表現を整理する

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

- 左から $\hat{\boldsymbol{y}}$ を乗じる

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

- 以下の関係を用いて展開すればよい

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

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

- このとき以下の項は 0 になる

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

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

東京の 8 月の気候の分析

- データの一部

- 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)

表 1: 東京の 8 月の気候

日付	気温	降雨	日射	降雪	風向	風速	気圧	湿度	雲量
2022-08-01	30.6	0	24.53	0	SSE	2.8	1010.1	72	8.8
2022-08-02	31.6	0	24.78	0	SSE	2.5	1008.8	71	9.8
2022-08-03	31.5	0	21.24	0	SSE	2.3	1005.1	75	7.3
2022-08-04	24.6	18	3.46	0	NE	2.7	1006	89	10
2022-08-05	23.8	0	7.65	0	NE	2.9	1006.1	83	9.8
2022-08-06	25.2	0	17.06	0	SSE	2.4	1008.1	73	10
2022-08-07	27.6	0	14.45	0	SSE	2.2	1009.3	80	8.3
2022-08-08	29.8	0	22.52	0	S	4.5	1008.5	75	4.8
2022-08-09	30.9	0	25.5	0	S	5.5	1006.9	69	6.8
2022-08-10	30.5	0	25.99	0	S	5.3	1007.2	70	6
2022-08-11	29.5	0	22.9	0	S	5.4	1007.5	75	6
2022-08-12	28.3	2	15.36	0	S	5.8	1007.5	81	9.8
2022-08-13	25.5	47.5	4.53	0	S	4.8	1005.6	94	10
2022-08-14	28.2	0	16.28	0	SSE	2.6	1003	84	8.8
2022-08-15	29.4	0	18.65	0	S	2.5	1003.4	78	8.8
2022-08-16	31	0	20.5	0	SSW	4.8	1000.6	70	8.3
2022-08-17	27.3	5	8.87	0	NE	2.5	1005.8	77	10
2022-08-18	26.8	13	8.74	0	S	2.8	1001.7	81	6
2022-08-19	27.5	0	23.52	0	SSE	3.4	1001.7	62	3
2022-08-20	26.4	1.5	13.5	0	NW	1.8	1000.6	82	9.8
2022-08-21	26	1	8.96	0	NE	2.1	1002.3	87	10
2022-08-22	26.2	0	9.05	0	NNE	2.5	1005.5	82	10
2022-08-23	28.7	0	17.94	0	S	3.2	1003.2	83	8.3
2022-08-24	27.8	2	12.86	0	NE	2.9	1003.2	79	10
2022-08-25	25.7	0	9.83	0	SE	2	1004.1	77	10
2022-08-26	27	3.5	10.05	0	SSE	2.1	1002.5	89	10
2022-08-27	29	0	19.87	0	SSE	3.3	1002.7	80	5.5
2022-08-28	23.7	5	4.58	0	NE	3	1009.2	87	9.8
2022-08-29	23.3	0.5	15.45	0	NE	2.8	1016.1	69	8
2022-08-30	22.8	5	10.12	0	NNE	1.9	1012.5	88	10
2022-08-31	27.1	1	17.46	0	S	3.2	1007.6	85	8.8

表 2: 寄与率によるモデルの比較

	目的変数				
			気温		
	モデル 1	モデル 2	モデル 3	モデル 4	モデル 5
気圧	−0.178 (0.127)		−0.223 (0.068)	−0.214 (0.067)	−0.242 (0.068)
日射		0.297 (0.041)	0.306 (0.036)	0.366 (0.056)	0.348 (0.045)
湿度				0.071 (0.051)	
雲量					0.238 (0.161)
Constant	206.535 (127.430)	22.969 (0.690)	247.477 (68.433)	231.843 (68.254)	263.717 (67.941)
R^2	0.064	0.641	0.741	0.758	0.760
Adjusted R^2	0.031	0.628	0.722	0.731	0.733

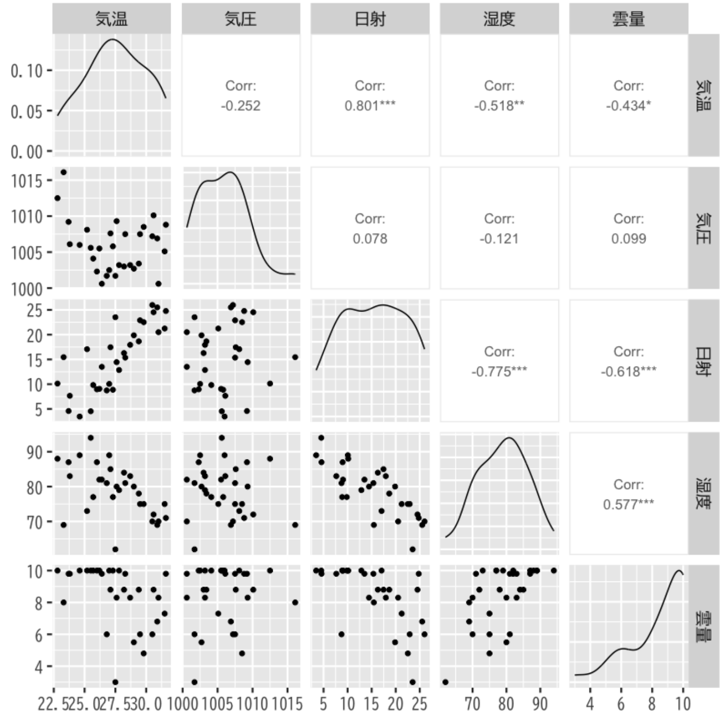

図 2: 散布図

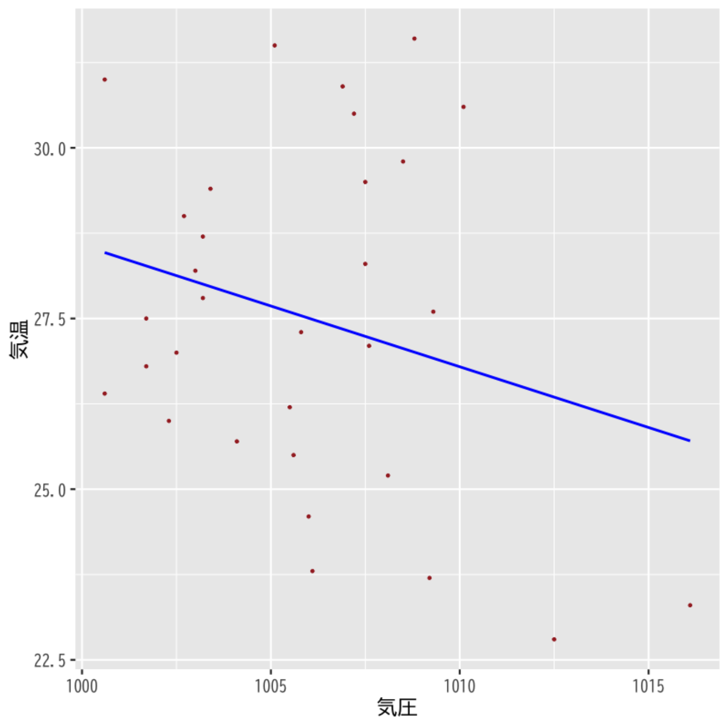

図 3: モデル 1

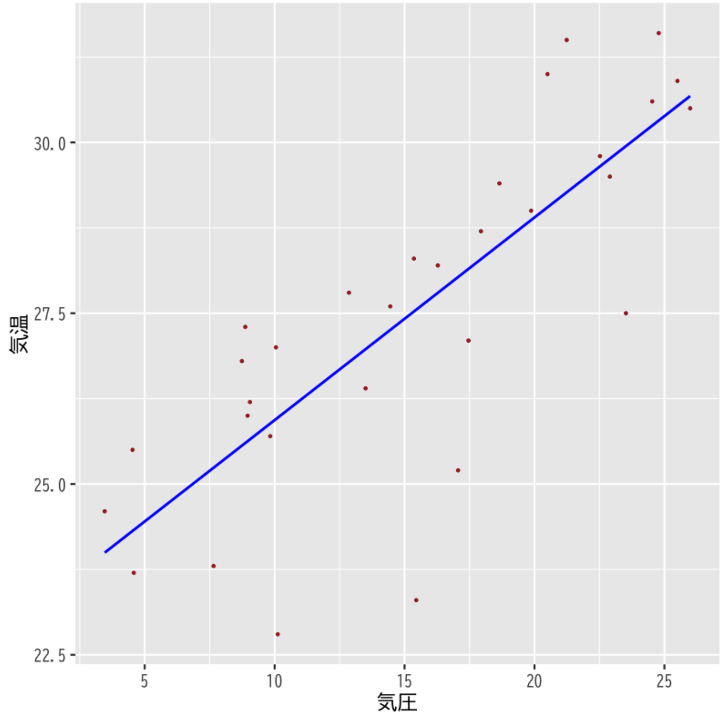

図 4: モデル 2

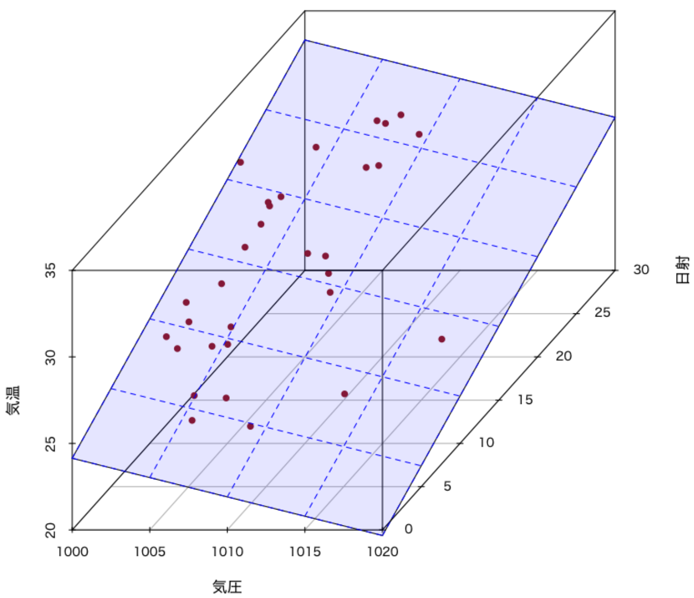

図 5: モデル 3

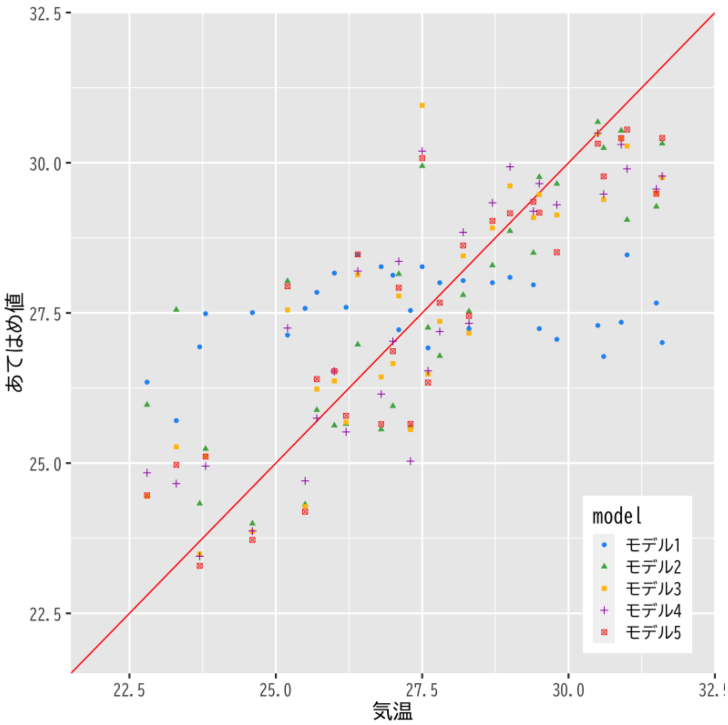

図 6: モデルの比較

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