回帰分析

予測と発展的なモデル

村田 昇

講義概要

- 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**

回帰分析の復習

線形回帰モデル

- **目的変数** を **説明変数** で説明する関係式を構成
  - 説明変数: $x_1, \ldots, x_p$ (p 次元)
  - 目的変数: y (1 次元)
- **回帰係数** $\beta_0, \beta_1, \ldots, \beta_p$ を用いた一次式

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

- **誤差項** を含む確率モデルで観測データを表現

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

問題設定

- 確率モデル

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{確率分布}$$

- 式の評価 : **残差平方和** の最小化による推定

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

解とその一意性

- 解の条件 : **正規方程式**

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

- 解の一意性 : **Gram 行列** $X^\mathsf{T}X$ が正則

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

日付	気温	降雨	日射	降雪	風向	風速	気圧	湿度	雲量
2024-10-01	23.3	0.5	11.45	0	NNW	2.6	1006.0	81	5.8
2024-10-02	26.5	0.0	18.32	0	S	2.9	1007.9	77	6.0
2024-10-03	23.1	11.0	5.88	0	E	2.7	1015.9	87	10.0
2024-10-04	25.9	2.0	12.60	0	S	3.5	1015.4	87	10.0
2024-10-05	21.3	9.5	1.88	0	NNE	2.5	1018.4	94	10.0
2024-10-06	21.3	0.0	5.01	0	NNW	1.7	1017.1	93	10.0
2024-10-07	25.0	0.0	14.99	0	S	2.9	1008.9	83	8.0
2024-10-08	18.8	33.5	1.98	0	NE	3.0	1008.9	97	10.0
2024-10-09	16.0	53.5	3.58	0	NNW	2.9	1009.3	93	10.0
2024-10-10	17.8	0.0	7.52	0	NNW	2.6	1009.8	75	6.0
2024-10-11	19.0	0.0	16.14	0	SSE	1.9	1013.1	69	7.5
2024-10-12	20.6	0.0	16.44	0	N	1.9	1019.0	73	2.5
2024-10-13	20.9	0.0	16.27	0	NNW	2.2	1021.1	70	0.8
2024-10-14	20.8	0.0	16.02	0	NNW	2.3	1022.6	71	4.0
2024-10-15	22.1	0.0	16.53	0	SSW	2.2	1020.3	72	3.8
2024-10-16	22.6	0.0	8.50	0	NNE	1.5	1017.3	76	7.5
2024-10-17	22.8	0.0	8.10	0	ENE	2.3	1020.0	79	9.3
2024-10-18	21.6	2.0	3.27	0	N	1.8	1019.5	92	10.0
2024-10-19	24.2	1.5	11.29	0	S	2.7	1009.2	84	10.0
2024-10-20	17.4	0.0	13.59	0	ENE	3.6	1023.6	55	5.8
2024-10-21	16.2	0.0	12.31	0	NW	2.7	1029.2	61	7.0
2024-10-22	19.7	0.0	12.02	0	NNW	1.9	1022.1	69	8.3
2024-10-23	21.9	6.5	4.24	0	NW	2.3	1012.3	90	10.0
2024-10-24	22.6	0.0	9.18	0	NE	2.2	1013.5	79	8.0
2024-10-25	20.2	0.5	3.61	0	NNE	2.4	1021.1	77	9.8
2024-10-26	19.0	0.0	3.90	0	NNW	1.7	1019.1	80	10.0
2024-10-27	19.7	4.0	8.46	0	NW	1.5	1011.4	87	9.5
2024-10-28	18.8	8.0	3.54	0	NE	1.9	1006.4	87	9.8
2024-10-29	15.4	24.5	2.79	0	NE	2.9	1017.4	79	10.0
2024-10-30	16.8	17.5	9.07	0	NW	3.2	1012.4	79	7.5
2024-10-31	16.2	0.0	11.86	0	NNE	2.0	1021.9	65	7.5

解析の事例

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

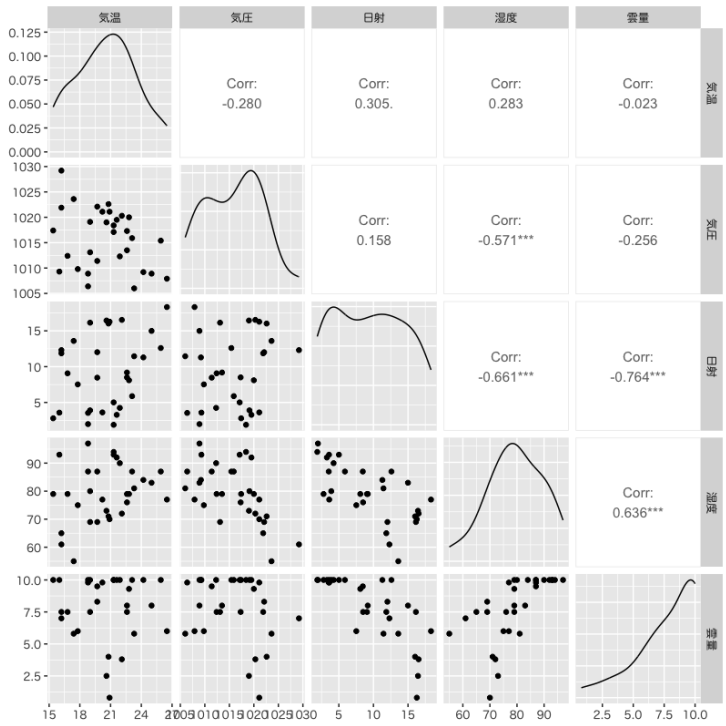

Figure 1: 散布図

- 観測値とあてはめ値の比較

寄与率

- **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

モデルの評価

- 決定係数 (R^2・Adjusted R^2) によるモデルの比較

F 統計量による検定

- 説明変数のうち 1 つでも役に立つか否かを検定する
  - 帰無仮説 H_0: $\beta_1 = \cdots = \beta_p = 0$
  - 対立仮説 H_1: $\exists j\ \beta_j \neq 0$ (少なくとも 1 つは役に立つ)
- F 統計量: 決定係数 (または残差) を用いて計算

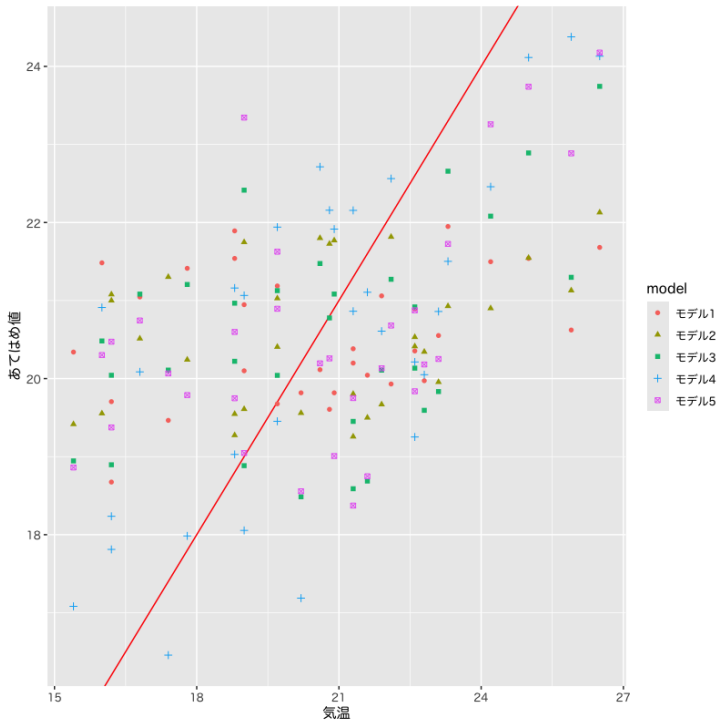

Figure 2: モデルの比較

	モデル 1		モデル 2		モデル 3		モデル 4		モデル 5	
Characteristic	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**
気圧	-0.14	-0.32, 0.04			-0.17	-0.35, 0.01	0.06	-0.12, 0.24	-0.14	-0.32, 0.04
日射			0.17	-0.03, 0.38	0.21	0.00, 0.41	0.53	0.30, 0.75	0.38	0.08, 0.68
湿度							0.28	0.14, 0.42		
雲量									0.49	-0.14, 1.1
R^2	0.078		0.093		0.204		0.519		0.272	
Adjusted R^2	0.047		0.062		0.147		0.466		0.191	

Abbreviation: CI = Confidence Interval

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

- p 値: 自由度 $p, n-p-1$ の F 分布で計算

モデルの評価

- F 統計量によるモデルの比較

t 統計量による検定

- 回帰係数 β_j が回帰式に寄与するか否かを検定する
  - 帰無仮説 H_0: $\beta_j = 0$
  - 対立仮説 H_1: $\beta_j \neq 0$ (β_j は役に立つ)
- t 統計量: 各係数ごと，ζ^2 は $(X^\mathsf{T} X)^{-1}$ の対角成分

	モデル 1		モデル 2		モデル 3		モデル 4		モデル 5	
Characteristic	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**	**Beta**	**95% CI**
気圧	-0.14	-0.32, 0.04			-0.17	-0.35, 0.01	0.06	-0.12, 0.24	-0.14	-0.32, 0.04
日射			0.17	-0.03, 0.38	0.21	0.00, 0.41	0.53	0.30, 0.75	0.38	0.08, 0.68
湿度							0.28	0.14, 0.42		
雲量									0.49	-0.14, 1.1
R^2	0.078		0.093		0.204		0.519		0.272	
Statistic	2.47		2.98		3.58		9.72		3.36	
p-value	0.13		0.10		0.041		<0.001		0.033	

Abbreviation: CI = Confidence Interval

	モデル 1				モデル 2				モデル 3				
Characteristic	**Beta**	**SE**	**Statistic**	**p-value**	**Beta**	**SE**	**Statistic**	**p-value**	**Beta**	**SE**	**Statistic**	**p-value**	**Beta**
(Intercept)	164	91.3	1.80	0.083	19	1.08	17.6	<0.001	191	87.3	2.19	0.037	-70
気圧	-0.14	0.090	-1.57	0.13					-0.17	0.086	-1.97	0.059	0.06
日射					0.17	0.101	1.73	0.10	0.21	0.098	2.10	0.045	0.53
湿度													0.28
雲量													

Abbreviations: CI = Confidence Interval, SE = Standard Error

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\zeta_j}$$

- p 値: 自由度 $n-p-1$ の t 分布を用いて計算

モデルの評価

- t 統計量によるモデルの比較

診断プロットによる評価

- 回帰モデルのあてはまりを視覚的に評価
  - Residuals vs Fitted: あてはめ値 (予測値) と残差の関係 (誤差の独立性)
  - Normal Q-Q: 残差の正規性の確認
  - Scale-Location: あてはめ値と正規化した残差の関係 (分散の一様性)
  - Residuals vs Leverage: 正規化した残差とテコ比の関係 (外れ値)
- モデル 2
- モデル 3
- モデル 4

回帰モデルによる予測

予測

- 新しいデータ (説明変数) $\boldsymbol{x}$ に対する **予測値**

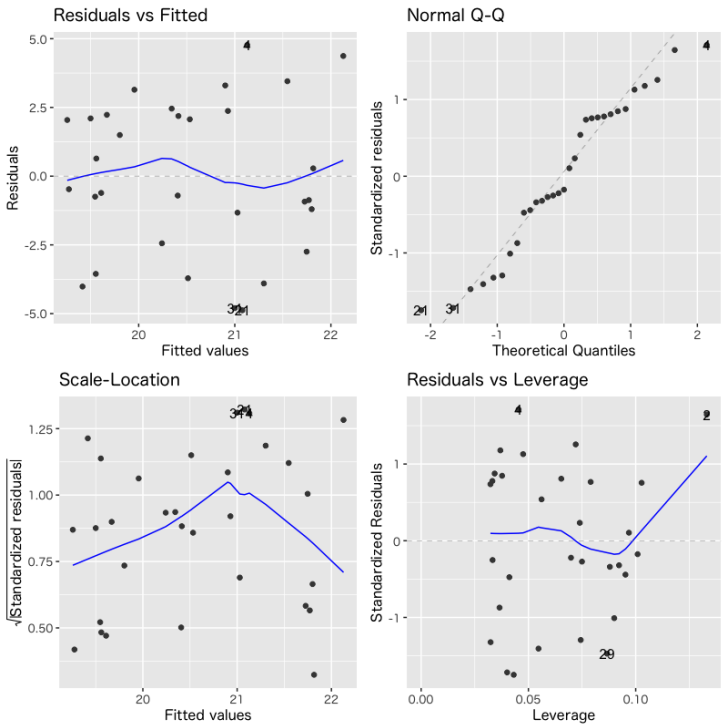

Figure 3: モデル 2 の診断

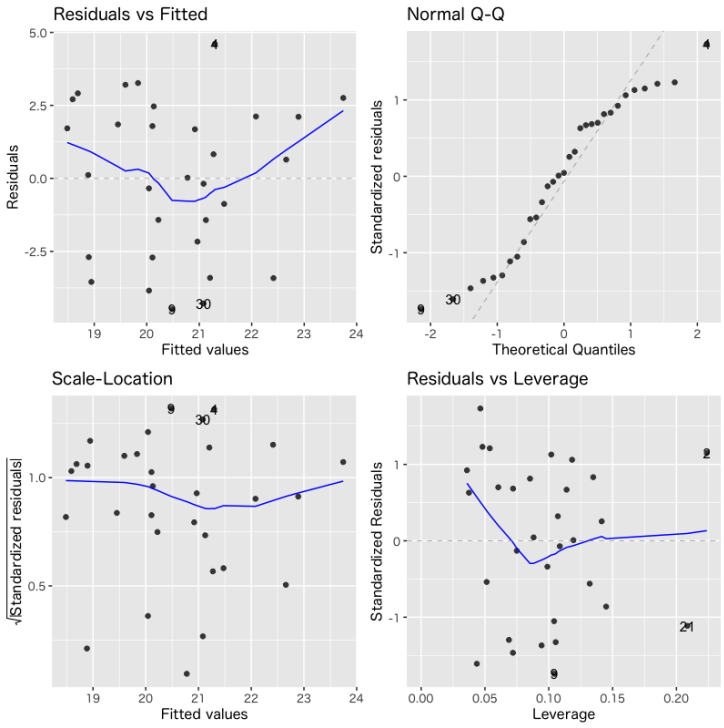

Figure 4: モデル 3 の診断

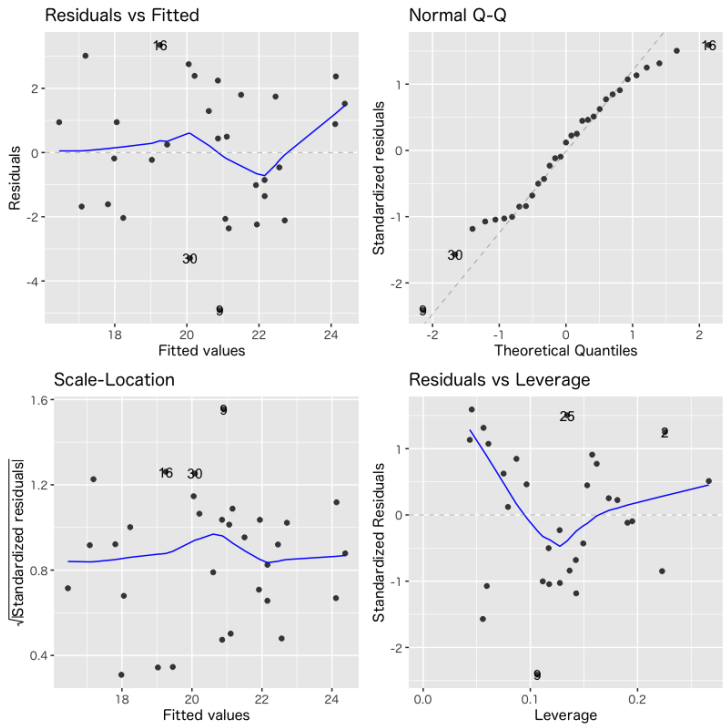

Figure 5: モデル 4 の診断

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

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = \boldsymbol{w}(\boldsymbol{x})^\mathsf{T}\boldsymbol{y}, \qquad \boldsymbol{w}(\boldsymbol{x})^\mathsf{T} = (1, \boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}X^\mathsf{T}$$

  - 重みは元データと新規データの説明変数で決定

予測値の性質

- 推定量は以下の性質をもつ多変量正規分布

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$
$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(X^\mathsf{T}X)^{-1}$$

- この性質を利用して以下の 3 つの値の違いを評価

$$y = (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} + \epsilon \qquad \text{(観測値)}$$
$$\tilde{y} = (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} \qquad \text{(最適な予測値)}$$
$$\hat{y} = (1, \boldsymbol{x}^\mathsf{T})\hat{\boldsymbol{\beta}} \qquad \text{(回帰式による予測値)}$$

  - $\hat{y}$ と y は独立な正規分布に従うことに注意

信頼区間

最適な予測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}\mathbb{E}[\tilde{y}-\hat{y}] &= (1,\boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} - (1,\boldsymbol{x}^\mathsf{T})\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0\\ \mathrm{Var}(\tilde{y}-\hat{y}) &= \underbrace{\sigma^2(1,\boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}(1,\boldsymbol{x}^\mathsf{T})^\mathsf{T}}_{\hat{\boldsymbol{\beta}}\text{ の推定誤差による分散}} = \sigma^2\gamma_c(\boldsymbol{x})^2\end{aligned}$$

- 標準化による表現

$$\frac{\tilde{y}-\hat{y}}{\sigma\gamma_c(\boldsymbol{x})} \sim \mathcal{N}(0,1)$$

信頼区間

- 未知の分散を不偏分散で推定

$$Z = \frac{\tilde{y}-\hat{y}}{\hat{\sigma}\gamma_c(\boldsymbol{x})} \sim \mathcal{T}(n-p-1) \quad (t\text{ 分布})$$

- 確率 α の信頼区間

$$\mathcal{I}^c_\alpha = (\hat{y} - C_\alpha\hat{\sigma}\gamma_c(\boldsymbol{x}),\ \hat{y} + C_\alpha\hat{\sigma}\gamma_c(\boldsymbol{x}))$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

  - 最適な予測値 $\tilde{y}$ が入ることが期待される区間

予測区間

観測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}\mathbb{E}[y-\hat{y}] &= (1,\boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] - (1,\boldsymbol{x}^\mathsf{T})\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0\\ \mathrm{Var}(y-\hat{y}) &= \underbrace{\sigma^2(1,\boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}(1,\boldsymbol{x}^\mathsf{T})^\mathsf{T}}_{\hat{\boldsymbol{\beta}}\text{ の推定誤差による分散}} + \underbrace{\sigma^2}_{\text{誤差の分散}} = \sigma^2\gamma_p(\boldsymbol{x})^2\end{aligned}$$

- 標準化による表現

$$\frac{y-\hat{y}}{\sigma\gamma_p(\boldsymbol{x})} \sim \mathcal{N}(0,1)$$

予測区間

- 未知の分散を不偏分散で推定

$$Z = \frac{y - \hat{y}}{\hat{\sigma}\gamma_p(\boldsymbol{x})} \sim \mathcal{T}(n-p-1) \quad (t \text{ 分布})$$

- 確率 α の予測区間

$$\mathcal{I}_\alpha^p = \left(\hat{y} - C_\alpha \hat{\sigma}\gamma_p(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}\gamma_p(\boldsymbol{x})\right)$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

  - 観測値 y が入ることが期待される区間
  - $\gamma_p > \gamma_c$ なので信頼区間より広くなる

実習

発展的なモデル

非線形性を含むモデル

- 目的変数 y
- 説明変数 $x_1, \ldots, x_p$
- 説明変数の追加で対応可能
  - 交互作用 (交差項) : $x_i x_j$ のような説明変数の積
  - 非線形変換 : $\log(x_k)$ のような関数による変換

カテゴリカル変数を含むモデル

- 数値ではないデータ
  - 悪性・良性
  - 血液型 (A 型,B 型,AB 型,O 型)
- 適切な方法で数値に変換して対応:
  - 2 値の場合は 1,0 (真，偽) を割り当てる
    - 悪性 : 1
    - 良性 : 0
  - 3 値以上の場合は **ダミー変数** を利用する (カテゴリ数-1 個)
    - A 型 : (1,0,0)
    - B 型 : (0,1,0)
    - O 型 : (0,0,1)
    - AB 型 : (0,0,0)

実習

次回の予定

- **第 1 回: 主成分分析の考え方**
- 第 2 回: 分析の評価と視覚化