回帰分析

モデルの評価

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

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

簡潔な表現のための行列

- デザイン行列 (説明変数)

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

簡潔な表現のためのベクトル

- ベクトル (目的変数・誤差・回帰係数)

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

問題の記述

- 確率モデル

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

- 回帰式の推定 : **残差平方和** の最小化

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

解の表現

- 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

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

最小二乗推定量の性質

- **あてはめ値** $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$ は X の列ベクトルの線形結合
- **残差** $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ はあてはめ値 $\hat{\boldsymbol{y}}$ と直交

$$\hat{\boldsymbol{\epsilon}}^\mathsf{T}\hat{\boldsymbol{y}} = 0$$

- 回帰式は説明変数と目的変数の **標本平均** を通過

$$\bar{y} = (1, \bar{\boldsymbol{x}}^\mathsf{T})\hat{\boldsymbol{\beta}}, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i,$$

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

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

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

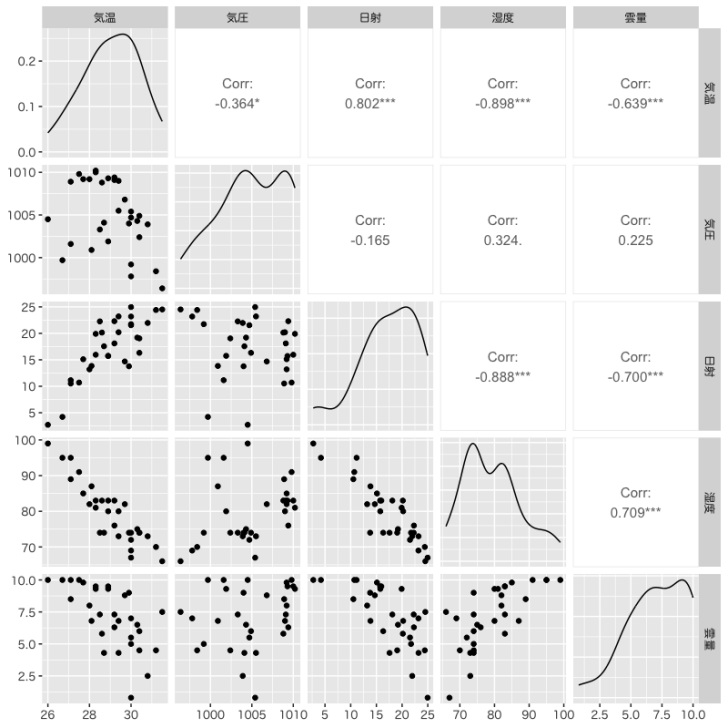

Figure 1: 散布図

	モデル 1		モデル 2		モデル 3		モデル 4		モデル 5	
変数	**係数**	**標準誤差**	**係数**	**標準誤差**	**係数**	**標準誤差**	**係数**	**標準誤差**	**係数**	**標準誤差**
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

- 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)

あてはめ値の性質

あてはめ値

- さまざまな表現

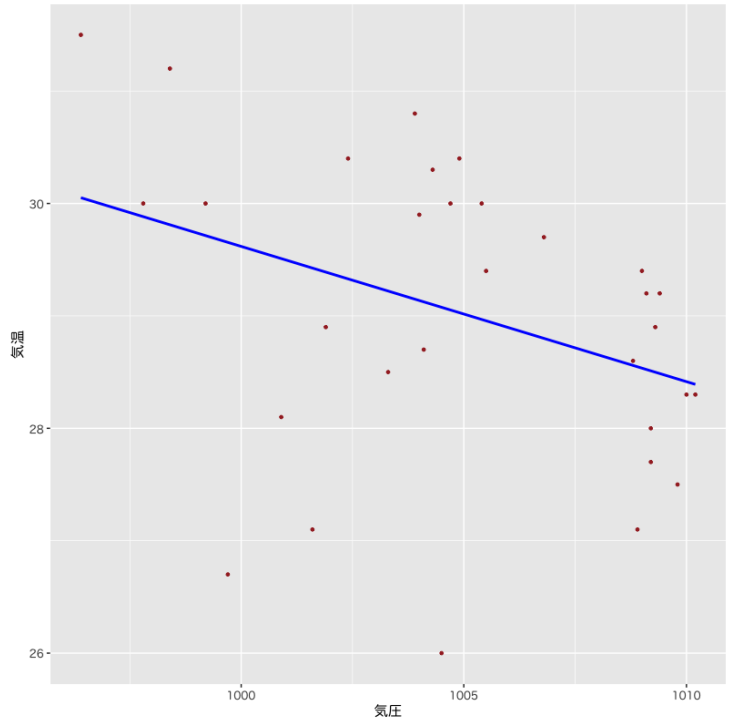

Figure 2: モデル 1

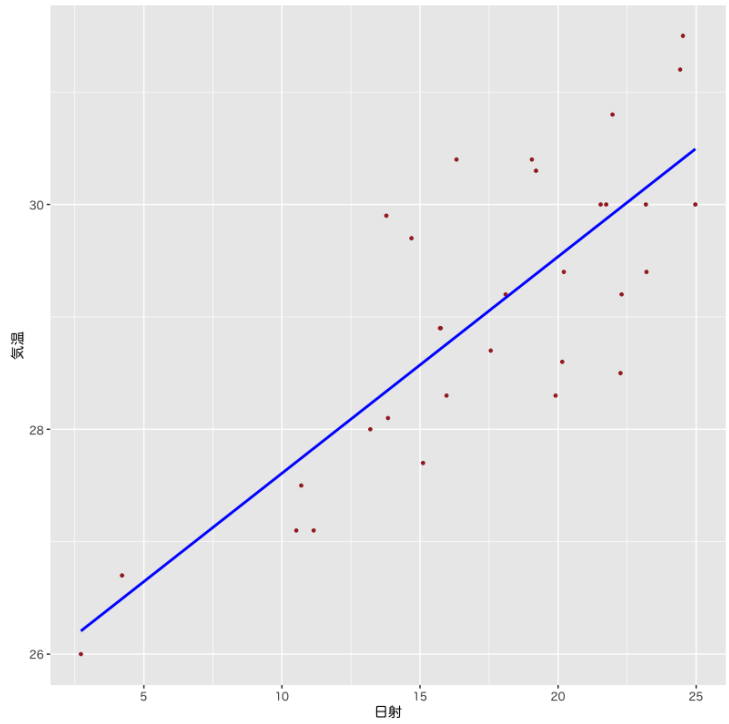

Figure 3: モデル 2

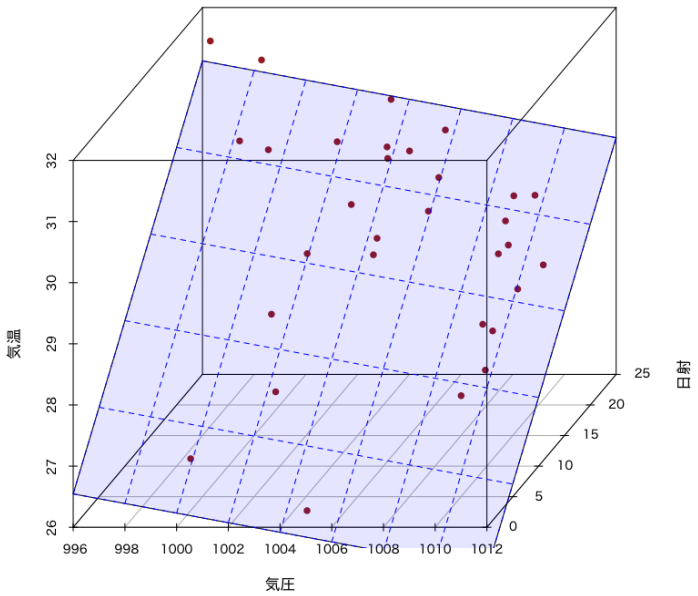

Figure 4: モデル 3

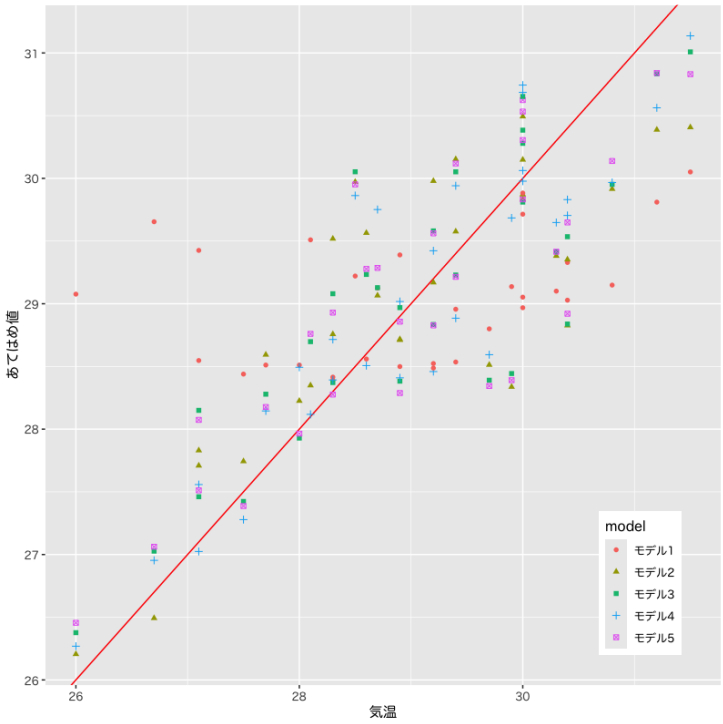

Figure 5: モデルの比較

$$
\begin{aligned}
\hat{\boldsymbol{y}} &= X\hat{\boldsymbol{\beta}} \\
&\quad (\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}\text{を代入}) \\
&= X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y} && (A) \\
&\quad (\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}\text{を代入}) \\
&= X(X^\mathsf{T}X)^{-1}X^\mathsf{T}X\boldsymbol{\beta} + X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{\epsilon} \\
&= X\boldsymbol{\beta} + X(X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{\epsilon} && (B)
\end{aligned}
$$

- (A) あてはめ値は **観測値の重み付けの和** で表される
- (B) あてはめ値と観測値は **誤差項** の寄与のみ異なる

あてはめ値と誤差

- 残差と誤差の関係

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

- (C) 残差は **誤差の重み付けの和** で表される

ハット行列

- 定義

$$
H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}
$$

- ハット行列 H による表現

$$
\begin{aligned}
\hat{\boldsymbol{y}} &= H\boldsymbol{y} \\
\hat{\boldsymbol{\epsilon}} &= (I - H)\boldsymbol{\epsilon}
\end{aligned}
$$

- あてはめ値や残差は H を用いて簡潔に表現される

ハット行列の性質

- 観測データ (デザイン行列) のみで計算される
- 観測データと説明変数の関係を表す
- 対角成分 (**テコ比**; leverage) は観測データが自身の予測に及ぼす影響の度合を表す

$$
\hat{y}_j = (H)_{jj}y_j + (\text{それ以外のデータの寄与})
$$

- $(A)_{ij}$ は行列 A の (i, j) 成分
- テコ比が小さい : 他のデータでも予測が可能
- テコ比が大きい : 他のデータでは予測が困難

推定量の統計的性質

最小二乗推定量の性質

- 推定量と誤差の関係

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

- 正規分布の重要な性質 (**再生性**)
  正規分布に従う独立な確率変数の和は正規分布に従う

推定量の分布

- **誤差の仮定** : 独立，平均 0 分散 σ^2 の正規分布
- 推定量は以下の多変量正規分布に従う

$$
\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[\boldsymbol{\beta} + (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{\epsilon}] = \boldsymbol{\beta} \\
\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\mathsf{T}] = \sigma^2(X^\mathsf{T}X)^{-1} \\
\hat{\boldsymbol{\beta}} &\sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(X^\mathsf{T}X)^{-1})
\end{aligned}
$$

- 通常 σ^2 は未知，必要な場合には不偏分散で代用

$$
\hat{\sigma^2} = \frac{S}{n-p-1} = \frac{1}{n-p-1}\hat{\boldsymbol{\epsilon}}^\mathsf{T}\hat{\boldsymbol{\epsilon}} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2
$$

- これらの性質を利用してモデルの評価を行う

実習

誤差の評価

寄与率 (再掲)

- **決定係数 (R-squared)**
  - 回帰式で説明できるばらつきの比率

$$
R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}
$$

- **自由度調整済み決定係数 (adjusted R-squared)**
  - 決定係数を不偏分散で補正

$$
\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}
$$

各係数の推定量の分布

- 推定された回帰係数の精度を評価
  - 誤差 ϵ の分布は平均 0 分散 σ^2 の正規分布
  - $\hat{\boldsymbol{\beta}}$ の分布 : p+1 変量正規分布

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (X^\mathsf{T} X)^{-1})$$

  - $\hat{\beta}_j$ の分布 : 1 変量正規分布

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 ((X^\mathsf{T} X)^{-1})_{jj}) = \mathcal{N}(\beta_j, \sigma^2 \zeta_j^2)$$

    * $(A)_{jj}$ は行列 A の (j, j) (対角) 成分

標準誤差

- **標準誤差** (standard error)
  - $\hat{\beta}_j$ の標準偏差の推定量

$$\text{s.e.}(\hat{\beta}_j) = \hat{\sigma}\zeta_j = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} \hat{\epsilon}_i^2} \cdot \sqrt{((X^\mathsf{T} X)^{-1})_{jj}}$$

  - 未知母数 σ^2 は不偏分散 $\hat{\sigma}^2$ で推定
  - $\hat{\beta}_j$ の精度の評価指標

実習

係数の評価

t 統計量

- **回帰係数の分布** に関する定理

  t **統計量** (t-statistic)

$$t = \frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\zeta_j}$$

  は自由度 $n-p-1$ の t 分布に従う

  - 証明には以下の性質を用いる
    * $\hat{\sigma}^2$ と $\hat{\boldsymbol{\beta}}$ は独立となる
    * $(\hat{\beta}_j - \beta_j)/(\sigma\zeta_j)$ は標準正規分布に従う
    * $(n-p-1)\hat{\sigma}^2/\sigma^2 = S(\hat{\boldsymbol{\beta}})/\sigma^2$ は自由度 $n-p-1$ の χ^2 分布に従う

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
  - 帰無仮説 $H_0 : \beta_j = 0$ (t 統計量が計算できる)
  - 対立仮説 $H_1 : \beta_j \neq 0$
- p 値 : 確率変数の絶対値が $|t|$ を超える確率

- $f(x)$ は自由度 $n-p-1$ の t 分布の確率密度関数

$$(p \text{ 値}) = 2\int_{|t|}^{\infty} f(x)dx \quad \text{(両側検定)}$$

帰無仮説 H_0 が正しければ p 値は小さくならない

実習

モデルの評価

平方和の分解 (再掲)

- いろいろなばらつき
  - $S_y = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \bar{\boldsymbol{y}})$: 目的変数のばらつき
  - $S \ = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{y}})$: 残差のばらつき ($\hat{\boldsymbol{\epsilon}}^\mathsf{T}\hat{\boldsymbol{\epsilon}}$)
  - $S_r = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\mathsf{T}(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$: 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

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

$$S_y = S + S_r$$

F 統計量

- **ばらつきの比** に関する定理

  $\beta_1 = \cdots = \beta_p = 0$ ならば F **統計量** (F-statistic)

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

  は自由度 $p, n-p-1$ の F 分布に従う
  - 証明には以下の性質を用いる
    * S_r と S は独立となる
    * S_r/σ^2 は自由度 p の χ^2 分布に従う
    * S/σ^2 は自由度 $n-p-1$ の χ^2 分布に従う

F 統計量を用いた検定

- 説明変数のうち 1 つでも役に立つか否かを検定
  - 帰無仮説 $H_0 : \beta_1 = \cdots = \beta_p = 0$ (S_r が χ^2 分布になる)
  - 対立仮説 $H_1 : \exists j \ \beta_j \neq 0$
- p 値 : 確率変数の値が F を超える確率
  - $f(x)$ は自由度 $p, n-p-1$ の F 分布の確率密度関数

$$(p \text{ 値}) = \int_{F}^{\infty} f(x)dx \quad \text{(片側検定)}$$

帰無仮説 H_0 が正しければ p 値は小さくならない

実習

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**