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 : $x_1, \dots, x_p$ (p 次元)
  - 目的変数 : y (1 次元)
- **回帰係数** $\beta_0, \beta_1, \dots, \beta_p$ 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

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

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

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

- 回帰式の推定 : **残差平方和** の最小化

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

解の表現

- 解の条件 : **正規方程式**

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

- 解の一意性 : **Gram 行列** $X^{\top}X$ が正則

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

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

$$\hat{\boldsymbol{\epsilon}}^{\top}\hat{\boldsymbol{y}} = 0$$

- 回帰式は説明変数と目的変数の **標本平均** を通過

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

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

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

	month	day	day_of_week	temp	rain	solar	snow	wdir	wind	press	humid	cloud
213	8	1	Sun	28.7	0.0	26.58	0	SSE	3.2	1000.2	76	2.3
214	8	2	Mon	28.6	0.5	19.95	0	SE	3.4	1006.1	80	7.0
215	8	3	Tue	29.0	3.0	19.89	0	S	4.0	1009.9	80	6.3
216	8	4	Wed	29.5	0.0	26.52	0	S	3.0	1008.2	76	2.8
217	8	5	Thu	29.1	0.0	26.17	0	SSE	2.8	1005.1	74	5.8
218	8	6	Fri	29.1	0.0	24.82	0	SSE	2.9	1004.2	75	4.0
219	8	7	Sat	27.9	2.0	11.43	0	NE	2.5	1003.1	85	9.0
220	8	8	Sun	25.9	90.5	3.43	0	N	3.0	998.0	97	10.0
221	8	9	Mon	28.1	2.0	13.34	0	S	6.1	995.4	84	6.0
222	8	10	Tue	31.0	0.0	22.45	0	SSW	4.7	996.3	58	4.8
223	8	11	Wed	29.2	0.0	21.12	0	SE	2.9	1008.0	61	9.3
224	8	12	Thu	26.0	0.5	8.34	0	SSE	2.4	1008.8	84	9.5
225	8	13	Fri	22.5	20.5	4.36	0	NE	2.7	1008.0	97	10.0
226	8	14	Sat	22.3	77.0	2.76	0	N	2.7	1003.6	100	10.0

- 関連するデータの散布図

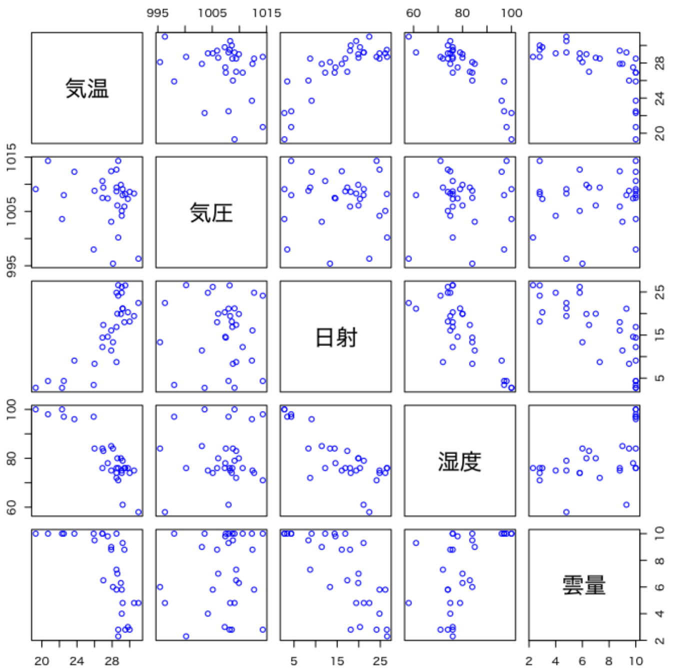

図 1: 散布図

- 気温を説明する 5 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(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)
- モデル 1 の推定結果

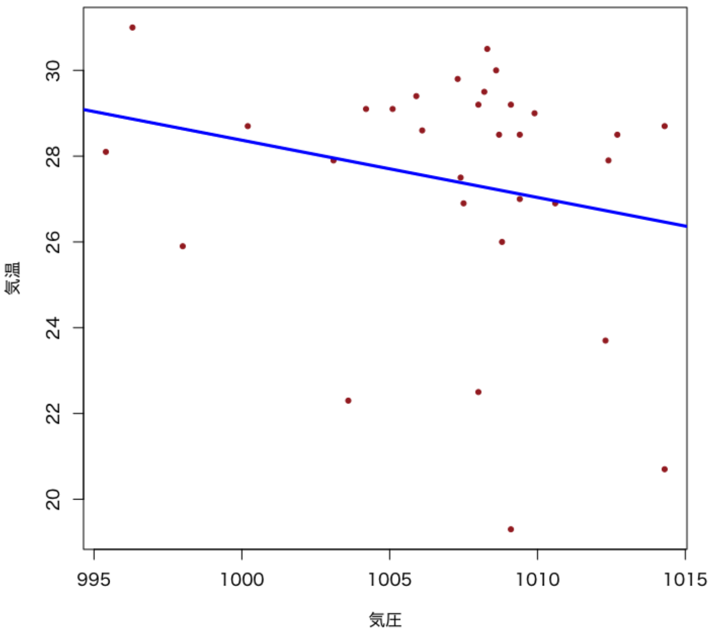

図 2: モデル 1

- 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

- 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: 0.0483 ; adj. R2: 0.0155"
    ```
  - モデル 2 : 気温 = F(日射)

    ```
    [1] "R2: 0.663 ; adj. R2: 0.651"
    ```
  - モデル 3 : 気温 = F(気圧, 日射)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.681"
    ```
  - モデル 4 : 気温 = F(気圧, 日射, 湿度)

    ```
    [1] "R2: 0.83 ; adj. R2: 0.811"
    ```
  - モデル 5 : 気温 = F(気圧, 日射, 雲量)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.67"
    ```

残差の性質

あてはめ値

- さまざまな表現

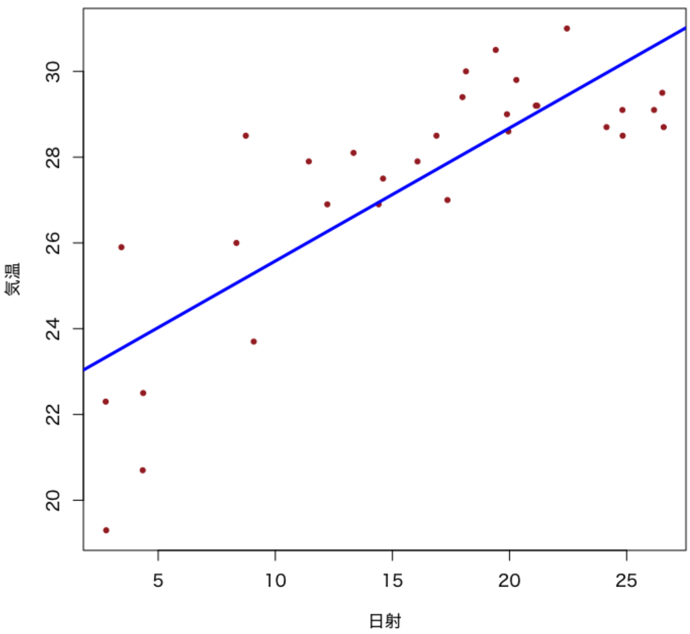

図 3: モデル 2

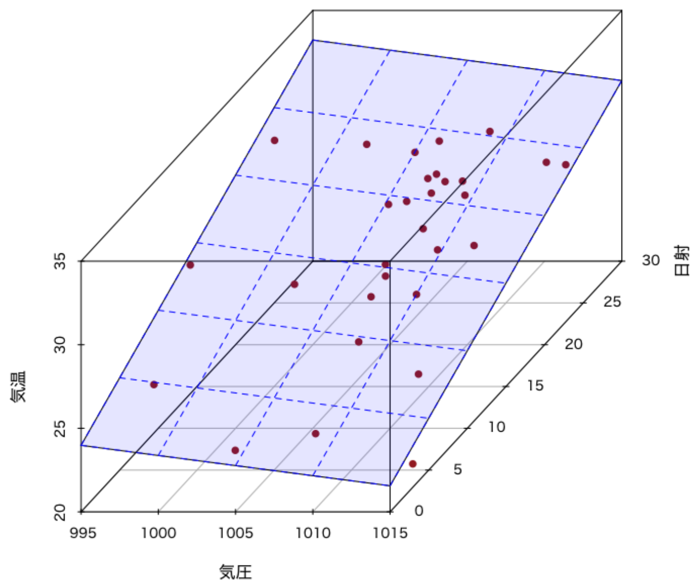

図 4: モデル 3

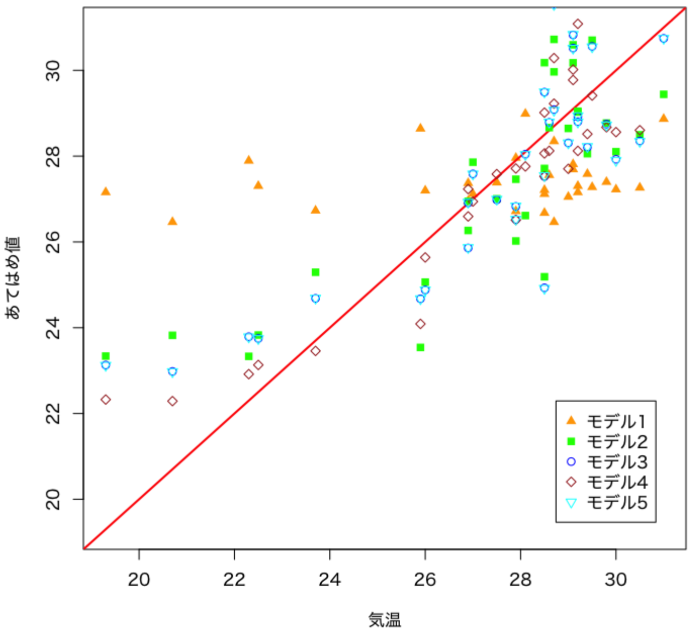

図 5: モデルの比較

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

- (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

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

- (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

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

- 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

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

- 正規分布の重要な性質

  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

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}\\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \sigma^2(X^\top X)^{-1}\end{aligned}$$

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

実習

R：乱数を用いた人工データの生成

- 正規乱数を用いた線形単回帰モデルの例

```
### 人工データによる推定量の性質の確認
set.seed(987) # 乱数のシード
x_obs <- c(1, 3, 5, 7) # 説明変数の観測値
epsilon <- rnorm(length(x_obs),sd=0.5) # 誤差項の生成
y_obs <- 2 - 3*x_obs + epsilon # 目的変数の観測値
my_data <- data.frame(x=x_obs,y=y_obs) # データフレームの作成
beta_est <- lm(y ~ x, data=my_data) # 回帰係数の推定
coef(beta_est) # 回帰係数の取得
summary(beta_est) # 分析結果の概要の表示
```

R：数値実験 (Monte-Carlo 法)

- 実験のためのコードは以下のようになる

```
mc <- 5000 # 実験回数を指定
my_trial <- function(){ # 1 回の試行を行うプログラム
                              # 乱数生成と推定の処理
    return(返り値)}
my_data <- as.data.frame(t( # 実験結果を転置してデータフレームに変換
    replicate(mc, my_trial()))) # Monte-Carlo 実験
## 適切な統計・視覚化処理 (下記は例)
apply(my_data,2,var) # 各列の分散の計算
plot(my_data) # 散布図行列の描画
hist(my_data[[k]]) # k 列目のデータのヒストグラム
```

練習問題

- 最小二乗推定量の性質を数値実験 (Monte-Carlo 法) により確認しなさい
  - 以下のモデルに従う人工データを生成する

    説明変数の観測データ：

    $$\{1, 20, 13, 9, 5, 15, 19, 8, 3, 4\}$$

    確率モデル：

    $$y = -1 + 2 \times x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 2)$$

  - 観測データから回帰係数を推定する
  - 実験を複数回繰り返し推定値 $(\hat{\beta}_0, \hat{\beta}_1)$ の分布を調べる

誤差の評価

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
  - $\hat{\beta}$ の分布

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

    * p+1 変量正規分布

- $\hat{\beta}_j$ の分布

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

  * 1 変量正規分布
  * $(A)_{jj}$ 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

$$\hat{\sigma}\zeta_j = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2} \cdot \sqrt{((X^\top X)^{-1})_{jj}}$$

  - 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

練習問題

- 数値実験により標準誤差の性質を確認しなさい
  - 人工データを用いて標準誤差と真の誤差を比較する

```
### 標準誤差は以下のようにして取り出せる
est <- lm(formula, data)
summary(est)$coef[,"Std. Error"] # 列名での指定
summary(est)$coefficients[,2] # 列番号での指定．coef と省略してもよい
```

  - 広告費と売上データを用いて係数の精度を議論する
  - 東京の気候データを用いて係数の精度を議論する

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

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

  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

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

  - $f(x)$ は自由度 $n{-}p{-}1$ の t 分布の確率密度関数
- 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

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**

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

  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

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

  - $f(x)$ は自由度 $p, n{-}p{-}1$ の F 分布の確率密度関数
- 帰無仮説 H_0 が正しければ p 値は小さくならない

実習

練習問題

- 数値実験により F 統計量の性質を確認しなさい
  - 人工データを用いて F 統計量の分布を確認しなさい

```r
### F統計量とその自由度は以下のようにして取り出せる
est <- lm(formula, data)
summary(est)$fstat
summary(est)$fstatistic # 省略しない場合
```

- 広告費と売上データのモデルの有効性を議論しなさい
- 東京の気候データのモデルの有効性を議論しなさい

補足

R：診断プロット

- 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
  - **Normal Q-Q**: 残差の正規性の確認
  - **Scale-Location**: あてはめ値と標準誤差で正規化した残差の関係
  - **Residuals vs Leverage**: 正規化した残差とテコ比の関係

  などが用意されている

```r
### 関数 lm() による推定結果の診断プロットの使い方
est <- lm(temp ~ press + solar + rain, data=tw_subset)
plot(est) # 指示に従って <Return> キーを押すと順次表示される
## help(plot.lm) を参照
```

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**