

# 回帰分析

## モデルの評価

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

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

- 誤差項を含む確率モデルで観測データを表現

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

### 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)

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

## 問題の記述

- 確率モデル

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

- 回帰式の推定: **残差平方和** の最小化

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

## 解の表現

- 解の条件: **正規方程式**

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

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

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

## 最小二乗推定量の性質

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

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

- 回帰式は説明変数と目的変数の **標本平均** を通過

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

## 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

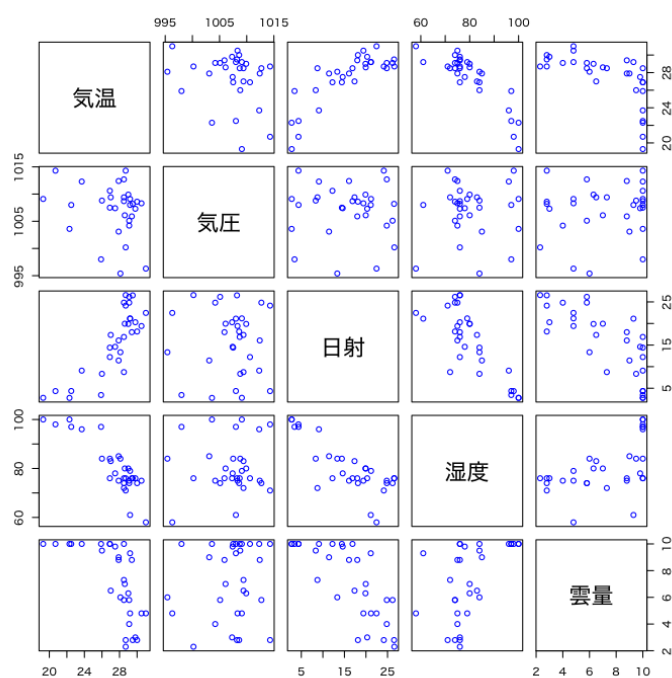


図 1: 散布図

- 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

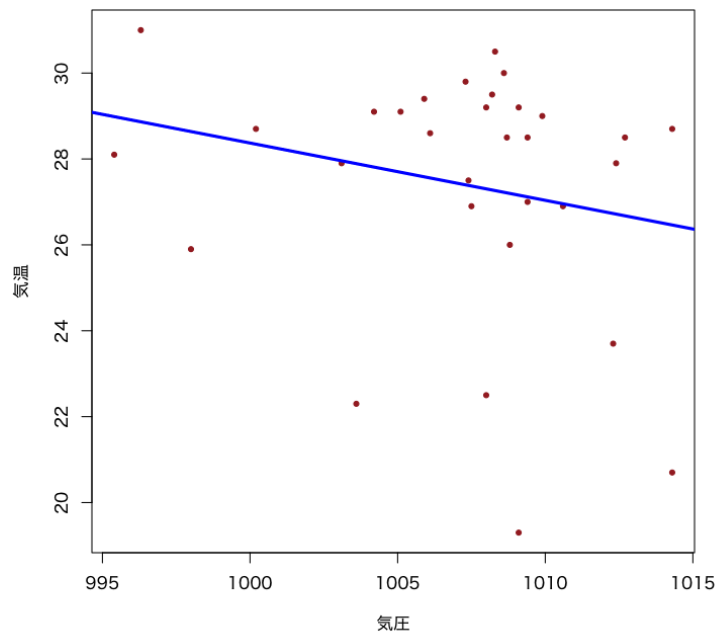


図 2: モデル 1

- 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

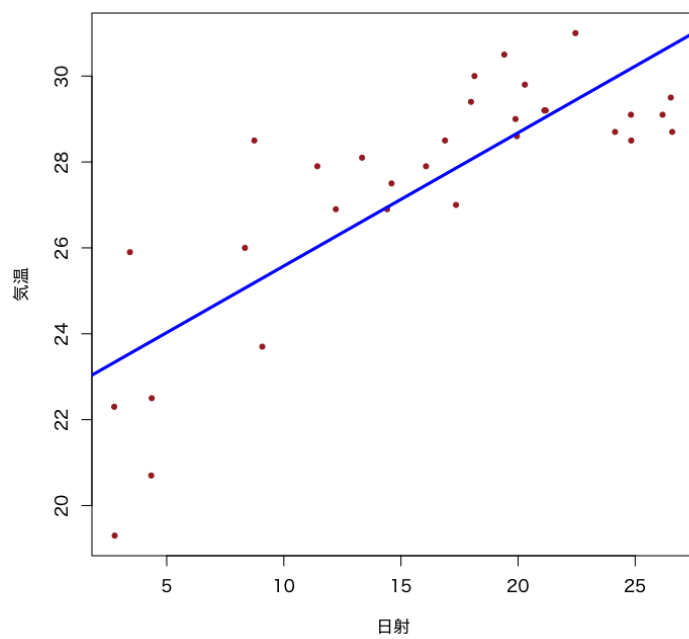


図 3: モデル 2

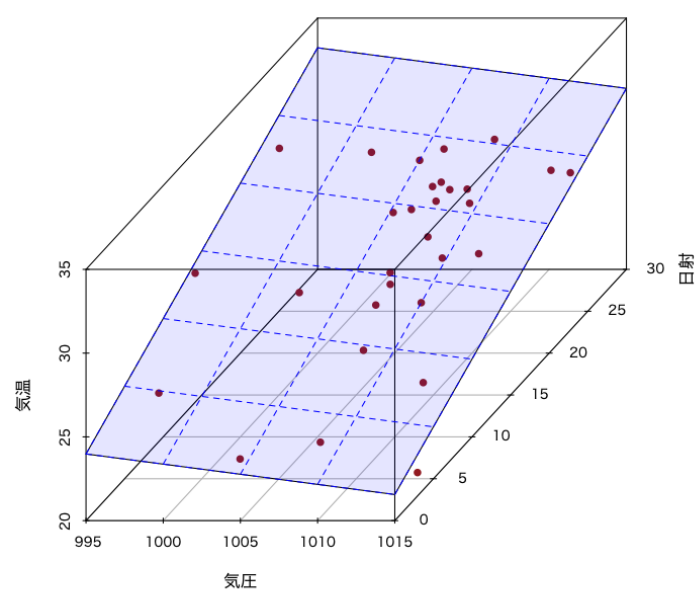


図 4: モデル 3

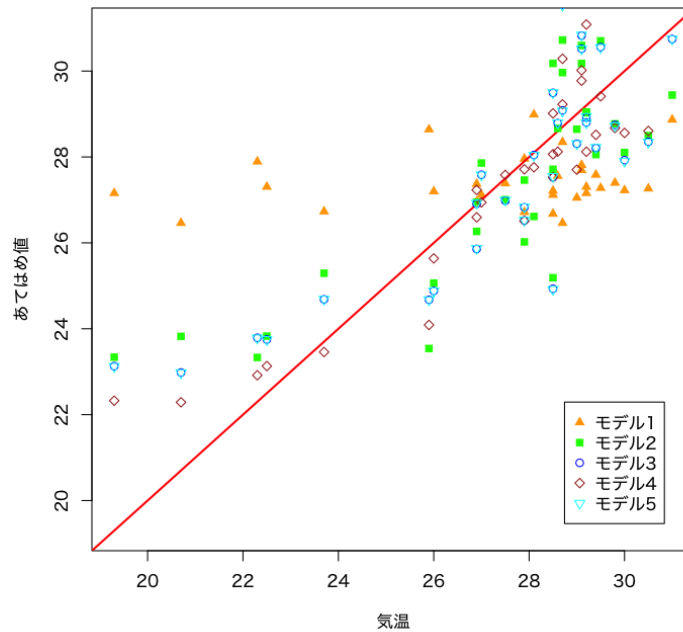


図 5: モデルの比較

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

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

- (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

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

- (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

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

- ハット行列  $H$  による表現

$$\hat{y} = Hy$$

$$\hat{\epsilon} = (I - H)\epsilon$$

- 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

## 演習

### 問題

- ハット行列  $H$  について以下を示しなさい
  - $H$  は対称行列であること
  - $H$  は冪等であること

$$H^2 = H, \quad (I - H)^2 = I - H$$

- 以下の等式が成り立つこと

$$HX = X, \quad X^T H = X^T$$

## 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

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

- 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  $\sigma^2$  の正規分布
- 推定量は以下の多変量正規分布に従う

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

## 演習

### 問題

- 誤差が独立で，平均 0 分散  $\sigma^2$  の正規分布に従うとき，最小二乗推定量  $\hat{\beta}$  について以下を示しなさい
  - 平均は  $\beta$  (真の母数) となること
  - 共分散行列は  $\sigma^2(X^T X)^{-1}$  となること

## 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  $\epsilon$  の分布は平均 0 分散  $\sigma^2$  の正規分布
  - $\hat{\beta}$  の分布

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

\*  $p+1$  変量正規分布

–  $\hat{\beta}_j$  の分布

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

\* 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

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

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



## 演習

### 問題

- 以下を示しなさい
  - 不偏分散  $\hat{\sigma}^2$  が母数  $\sigma^2$  の不偏な推定量となる  
以下が成り立つことを示せばよい

$$\mathbb{E} \left[ \sum_{i=1}^n \hat{\epsilon}_i^2 \right] = (n-p-1)\sigma^2$$

## 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  $\hat{\beta}$  は独立となる
  - $(\hat{\beta}_j - \beta_j)/(\sigma \zeta_j)$  は標準正規分布に従う
  - $(n-p-1)\hat{\sigma}^2/\sigma^2 = S(\hat{\beta})/\sigma^2$  は自由度  $n-p-1$  の  $\chi^2$  分布に従う

### $t$ 統計量による検定

- 回帰係数  $\beta_j$  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

## モデルの評価

### $F$ 統計量

- ばらつきの比 に関する定理:

$\beta_1 = \dots = \beta_p = 0$  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
  - $S_r/\sigma^2$  は自由度  $p$  の  $\chi^2$  分布に従う
  - $S/\sigma^2$  は自由度  $n-p-1$  の  $\chi^2$  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  $H_0: \beta_1 = \dots = \beta_p = 0$  ( $S_r$  が  $\chi^2$  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

## 解析の事例

### データについて

- 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](https://noboru-murata.github.io/multivariate-analysis/data/tokyo_weather.csv)

### 東京の 8 月の気候の分析

- 気候 (気温, 降雨, 日射, 降雪, 風速, 気圧, 湿度, 雲量) に関するデータ (の一部)

	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

- 作成した線形回帰モデルを検討する
  - モデル 1: 気温 =  $F(\text{気圧})$
  - モデル 2: 気温 =  $F(\text{日射})$

- 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)

- 観測値とあてはめ値の比較

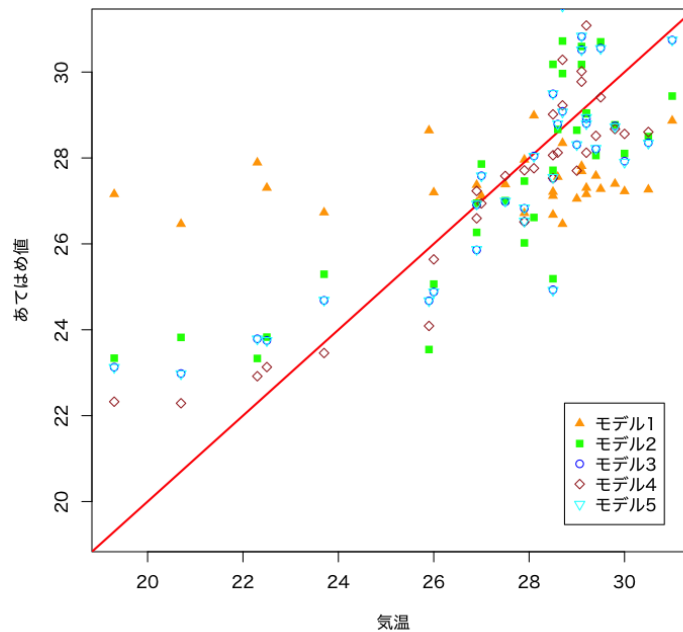


図 6: モデルの比較

- モデル 1: 係数とモデルの評価

Call:

```
lm(formula = tw_model1, data = tw_subset, y = TRUE)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.858	-0.680	1.183	1.922	3.235

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	161.9708	110.9113	1.460	0.155
press	-0.1336	0.1101	-1.213	0.235

Residual standard error: 2.846 on 29 degrees of freedom

Multiple R-squared: 0.0483, Adjusted R-squared: 0.01548

F-statistic: 1.472 on 1 and 29 DF, p-value: 0.2348

- モデル 2: 係数とモデルの評価

Call:

```
lm(formula = tw_model2, data = tw_subset, y = TRUE)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-4.0385	-1.2347	0.1714	1.1857	3.3124

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	22.47596	0.72144	31.154	< 2e-16 ***
solar	0.31026	0.04108	7.552	2.52e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.694 on 29 degrees of freedom  
Multiple R-squared: 0.6629, Adjusted R-squared: 0.6513  
F-statistic: 57.03 on 1 and 29 DF, p-value: 2.521e-08

- モデル 3: 係数とモデルの評価

Call:

```
lm(formula = tw_model3, data = tw_subset, y = TRUE)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.8296	-1.0254	0.2546	1.0629	3.5691

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	144.53120	63.13599	2.289	0.0298 *
press	-0.12116	0.06267	-1.933	0.0634 .
solar	0.30833	0.03928	7.849	1.5e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.619 on 28 degrees of freedom  
Multiple R-squared: 0.7026, Adjusted R-squared: 0.6814  
F-statistic: 33.08 on 2 and 28 DF, p-value: 4.232e-08

- モデル 4: 係数とモデルの評価

Call:

```
lm(formula = tw_model4, data = tw_subset, y = TRUE)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.0251	-0.6246	0.1873	0.9281	1.8950

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	147.00793	48.67327	3.020	0.005466 **
press	-0.10823	0.04839	-2.236	0.033783 *
solar	0.13429	0.04922	2.728	0.011058 *
humid	-0.15840	0.03532	-4.485	0.000121 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.248 on 27 degrees of freedom  
Multiple R-squared: 0.8296, Adjusted R-squared: 0.8107  
F-statistic: 43.81 on 3 and 27 DF, p-value: 1.649e-10

- モデル 5: 係数とモデルの評価

```

Call:
lm(formula = tw_model5, data = tw_subset, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8230 -1.0230  0.2534  1.0684  3.5829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.90781    64.97793   2.230  0.0342 *
press       -0.12161     0.06481  -1.876  0.0715 .
solar        0.31025     0.06243   4.969 3.31e-05 ***
cloud        0.00686     0.17144   0.040  0.9684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.649 on 27 degrees of freedom
Multiple R-squared:  0.7026, Adjusted R-squared:  0.6696
F-statistic: 21.27 on 3 and 27 DF,  p-value: 2.809e-07

```

- 決定係数と  $F$  統計量
  - モデル 1
 

```
[1] "R2: 0.0483 ; adj. R2: 0.0155 ; F-statistic: 1.47"
```
  - モデル 2
 

```
[1] "R2: 0.663 ; adj. R2: 0.651 ; F-statistic: 57"
```
  - モデル 3
 

```
[1] "R2: 0.703 ; adj. R2: 0.681 ; F-statistic: 33.1"
```
  - モデル 4
 

```
[1] "R2: 0.83 ; adj. R2: 0.811 ; F-statistic: 43.8"
```
  - モデル 5
 

```
[1] "R2: 0.703 ; adj. R2: 0.67 ; F-statistic: 21.3"
```

## 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