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:

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

- **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
214	8	1	Sat	26.1	0.5	19.79	0	NE	2.6	1009.3	77	7.8
215	8	2	Sun	26.3	0.0	19.53	0	SSE	2.4	1011.0	75	5.5
216	8	3	Mon	27.2	0.0	24.73	0	SSE	2.4	1011.0	74	3.8
217	8	4	Tue	28.3	0.0	24.49	0	SSE	2.9	1012.2	77	4.3
218	8	5	Wed	29.1	0.0	24.93	0	S	2.9	1013.4	76	3.3
219	8	6	Thu	28.5	0.0	24.02	0	SSE	3.9	1010.5	79	7.8
220	8	7	Fri	29.5	0.0	22.58	0	S	3.4	1005.0	71	7.5
221	8	8	Sat	28.1	0.0	15.49	0	SE	2.7	1006.1	79	8.3
222	8	9	Sun	28.7	0.0	19.96	0	SSE	2.4	1006.9	77	9.5
223	8	10	Mon	30.5	0.0	20.26	0	SE	2.4	1010.3	73	10.0
224	8	11	Tue	31.7	0.0	25.50	0	S	4.0	1009.7	67	2.8
225	8	12	Wed	30.0	0.5	18.24	0	SSE	2.5	1009.0	79	6.8
226	8	13	Thu	29.4	21.5	19.01	0	N	2.2	1006.4	82	5.0
227	8	14	Fri	29.4	0.0	19.85	0	SE	2.8	1005.5	78	2.0

- 気温を説明する 4 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(気圧, 日射)
  - モデル 3: 気温 = F(気圧, 日射, 湿度)
  - モデル 4: 気温 = F(気圧, 日射, 雲量)
- 関連するデータの散布図

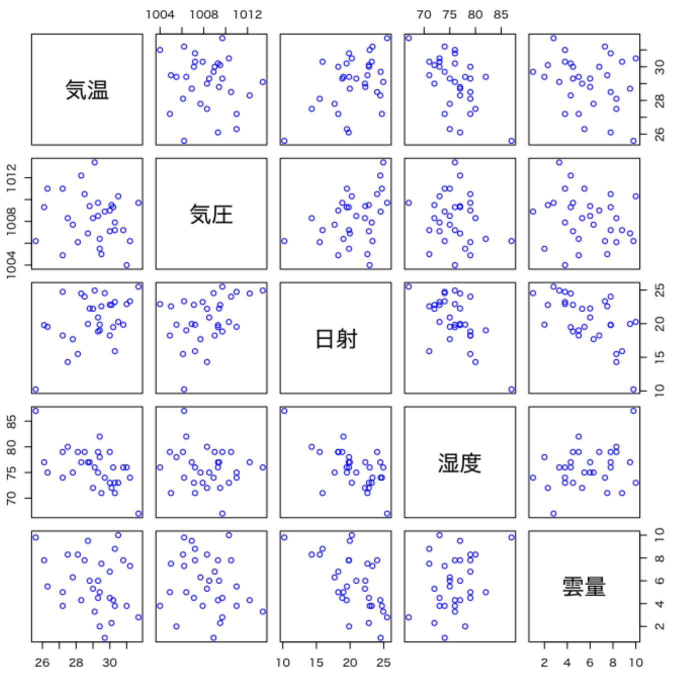

図 1: 散布図

- 観測値とあてはめ値の比較

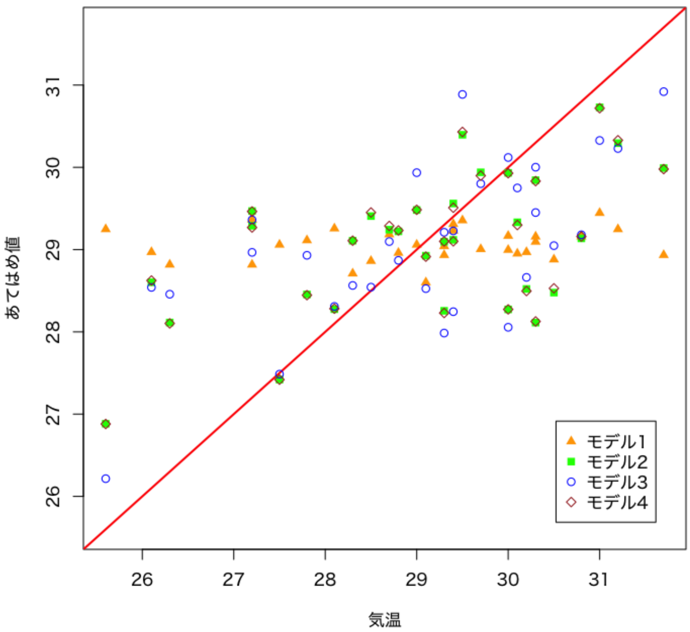

図 2: モデルの比較

- 決定係数・自由度調整済み決定係数の比較
  - モデル 1: 気温 = F(気圧)

    `[1] "R2: 0.0169 ; adj. R2: -0.017"`
  - モデル 2: 気温 = F(気圧, 日射)

    `[1] "R2: 0.32 ; adj. R2: 0.271"`
  - モデル 3: 気温 = F(気圧, 日射, 湿度)

    `[1] "R2: 0.422 ; adj. R2: 0.358"`
  - モデル 4: 気温 = F(気圧, 日射, 雲量)

    `[1] "R2: 0.32 ; adj. R2: 0.245"`

残差の性質

あてはめ値

- さまざまな表現:

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

  - (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:

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

  - (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:

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

- 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

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
  - H は冪等である

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

  - 以下の等式が成り立つ

$$HX = X, \quad X^\mathsf{T}H = X^\mathsf{T}$$

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:

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

- 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

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

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

演習

問題

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

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:
  - 誤差の分布は平均 0 分散 σ^2 の正規分布
  - $\hat{\boldsymbol{\beta}}$ の分布:

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

    * p+1 変量正規分布
  - $\hat{\beta}_j$ の分布:

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

    * $(A)_{jj}$ 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

演習

問題

- 以下を示しなさい
  - 不偏分散 $\hat{\sigma}^2$ が母数 σ^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 $\hat{\boldsymbol{\beta}}$ 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

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

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

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
214	8	1	Sat	26.1	0.5	19.79	0	NE	2.6	1009.3	77	7.8
215	8	2	Sun	26.3	0.0	19.53	0	SSE	2.4	1011.0	75	5.5
216	8	3	Mon	27.2	0.0	24.73	0	SSE	2.4	1011.0	74	3.8
217	8	4	Tue	28.3	0.0	24.49	0	SSE	2.9	1012.2	77	4.3
218	8	5	Wed	29.1	0.0	24.93	0	S	2.9	1013.4	76	3.3
219	8	6	Thu	28.5	0.0	24.02	0	SSE	3.9	1010.5	79	7.8

```
220       8    7          Fri  29.5   0.0  22.58     0     S   3.4  1005.0     71    7.5
221       8    8          Sat  28.1   0.0  15.49     0    SE   2.7  1006.1     79    8.3
222       8    9          Sun  28.7   0.0  19.96     0   SSE   2.4  1006.9     77    9.5
223       8   10          Mon  30.5   0.0  20.26     0    SE   2.4  1010.3     73   10.0
224       8   11          Tue  31.7   0.0  25.50     0     S   4.0  1009.7     67    2.8
225       8   12          Wed  30.0   0.5  18.24     0   SSE   2.5  1009.0     79    6.8
226       8   13          Thu  29.4  21.5  19.01     0     N   2.2  1006.4     82    5.0
227       8   14          Fri  29.4   0.0  19.85     0    SE   2.8  1005.5     78    2.0
```

- 作成した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(気圧, 日射)
  - モデル 3: 気温 = F(気圧, 日射, 湿度)
  - モデル 4: 気温 = F(気圧, 日射, 雲量)
- 観測値とあてはめ値の比較

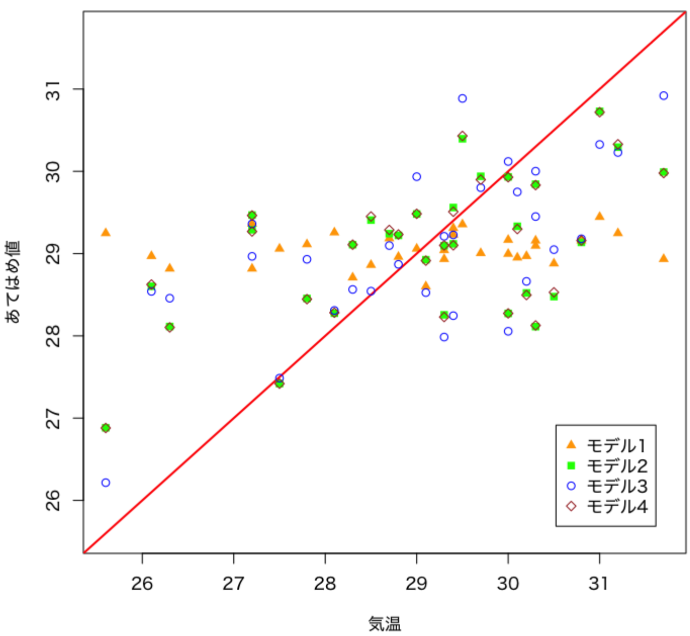

図 3: モデルの比較

- 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
-3.6478 -0.8208  0.1702  1.1452  2.7664

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.56133  128.23971   0.932    0.359
```

```
press        -0.08976    0.12719  -0.706    0.486

Residual standard error: 1.539 on 29 degrees of freedom
Multiple R-squared:  0.01688,Adjusted R-squared:  -0.01702
F-statistic: 0.498 on 1 and 29 DF,  p-value: 0.486
```

- 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
-2.50259 -0.73147  0.06766  0.83716  2.18776

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 273.68079  117.00384   2.339  0.02670 *
press        -0.24793    0.11661  -2.126  0.04245 *
solar         0.26057    0.07379   3.531  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.303 on 28 degrees of freedom
Multiple R-squared:  0.3198,Adjusted R-squared:  0.2712
F-statistic: 6.582 on 2 and 28 DF,  p-value: 0.00454
```

- 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
-2.44058 -0.50661  0.01425  0.81490  1.94439

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 262.65623  109.96118   2.389   0.0242 *
press        -0.22210    0.11012  -2.017   0.0537 .
solar         0.14203    0.08801   1.614   0.1182
humid        -0.16572    0.07589  -2.184   0.0379 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.223 on 27 degrees of freedom
Multiple R-squared:  0.4219,Adjusted R-squared:  0.3577
F-statistic: 6.568 on 3 and 27 DF,  p-value: 0.001772
```

- 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
-2.52396 -0.72721  0.07162  0.83623  2.17339
```

```
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 274.00410  119.16813   2.299  0.02945 *
press        -0.24843    0.11883  -2.091  0.04610 *
solar         0.26598    0.09155   2.905  0.00723 **
cloud         0.01295    0.12509   0.104  0.91829
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.327 on 27 degrees of freedom
Multiple R-squared:  0.3201,Adjusted R-squared:  0.2445
F-statistic: 4.236 on 3 and 27 DF,  p-value: 0.0141
```

- 決定係数と F-統計量
  - モデル 1: 気温 = F(気圧)

    ```
    [1] "R2: 0.0169 ; adj. R2: -0.017 ; F-statistic: 0.498"
    ```
  - モデル 2: 気温 = F(気圧, 日射)

    ```
    [1] "R2: 0.32 ; adj. R2: 0.271 ; F-statistic: 6.58"
    ```
  - モデル 3: 気温 = F(気圧, 日射, 湿度)

    ```
    [1] "R2: 0.422 ; adj. R2: 0.358 ; F-statistic: 6.57"
    ```
  - モデル 4: 気温 = F(気圧, 日射, 雲量)

    ```
    [1] "R2: 0.32 ; adj. R2: 0.245 ; F-statistic: 4.24"
    ```

次週の予定

- 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**