

回帰分析

モデルの評価

村田 昇

2020.10.20

講義の予定

- 第1日: 回帰モデルの考え方と推定
- 第2日: モデルの評価
- 第3日: 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:

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

行列・ベクトルによる簡潔な表現

- デザイン行列:

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

行列・ベクトルによる簡潔な表現

- ベクトル:

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

問題の記述

- 確率モデル:

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

- 回帰式の評価: **残差平方和** の最小化による推定

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

解の表現

- 解の条件: **正規方程式**

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

- 解の一意性: **Gram 行列** $X^\top X$ が正則

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

最小二乗推定量の性質

- あてはめ値** $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ は 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)

残差の性質

あてはめ値

- あてはめ値のさまざまな表現:

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

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

- (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

あてはめ値と誤差の関係

- 残差と誤差の関係:

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

- (A) 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:

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

- 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

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

但し $(A)_{ij}$ 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 について以下の問に答えなさい.
 - 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.

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

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:

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

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

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

演習

問題

- 最小二乗推定量 $\hat{\beta}$ について以下の問に答えなさい.
 - 平均が β となることを示せ.
 - 分散が $\sigma^2 (X^T X)^{-1}$ となることを示せ.

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
 - $\hat{\beta}$ の分布:

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

$p+1$ 変量正規分布

- $\hat{\beta}_j$ の分布:

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

$(A)_{jj}$ 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

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

- 未知パラメタ 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

- 以下の問に答えなさい.
 - 不偏分散 $\hat{\sigma}^2$ が母数 σ^2 の不偏な推定量となることを示しなさい.
- 以下を示せばよい

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

係数の評価

t -統計量

- 回帰係数の分布に関する定理:

$$(t\text{-統計量}) \quad t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{\xi_j}}$$

t -統計量は自由度 $n-p-1$ 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
 - $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{\xi_j})$ は標準正規分布に従う
 - $(n-p-1)\hat{\sigma}^2/\sigma^2 = S/\sigma^2$ 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: $\beta_j = 0$ (t -統計量が計算できる)
 - 対立仮説: $\beta_j \neq 0$
- 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 $\beta_j = 0$ 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:

$$(F\text{-統計量}) \quad F = \frac{\frac{1}{p} S_r}{\frac{1}{n-p-1} S} = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

$\beta_1 = \dots = \beta_p = 0$ ならば, **F -統計量** 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: $\beta_1 = \dots = \beta_p = 0$ (S_r が χ^2 分布になる)
 - 対立仮説: $\exists j \beta_j \neq 0$
- 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 $\forall j \beta_j = 0$ 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_reg.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)

	date	temp	rain	solar	snow	wind	press	humid	cloud
213	2019/8/1	30.5	0.0	20.55	0	2.5	1008.5	80	1.8
214	2019/8/2	30.2	0.0	20.24	0	2.7	1008.4	80	2.8
215	2019/8/3	29.4	0.0	25.03	0	2.9	1008.7	78	1.0
216	2019/8/4	29.4	0.0	24.62	0	2.8	1009.5	76	3.0
217	2019/8/5	29.8	0.0	26.72	0	3.0	1009.5	75	2.8
218	2019/8/6	30.3	0.0	24.18	0	3.8	1008.4	76	7.5
219	2019/8/7	30.4	0.0	24.10	0	3.1	1007.4	74	6.5
220	2019/8/8	29.9	0.0	22.46	0	2.8	1006.6	78	4.3
221	2019/8/9	30.1	0.0	25.10	0	3.3	1005.5	74	6.5
222	2019/8/10	29.6	0.0	22.69	0	3.2	1005.4	76	4.3
223	2019/8/11	29.4	0.0	23.77	0	2.8	1005.9	76	6.0
224	2019/8/12	28.8	0.5	17.16	0	2.6	1005.7	81	10.0
225	2019/8/13	29.3	0.0	15.57	0	2.6	1003.8	83	6.8
226	2019/8/14	29.2	8.5	15.38	0	3.8	1003.4	85	9.0

- 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
- 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
-2.9372	-1.5395	0.3867	1.4446	2.3344

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	117.37523	95.88549	1.224	0.231
press	-0.08846	0.09532	-0.928	0.361

Residual standard error: 1.774 on 29 degrees of freedom

Multiple R-squared: 0.02884, Adjusted R-squared: -0.004651

F-statistic: 0.8611 on 1 and 29 DF, p-value: 0.3611

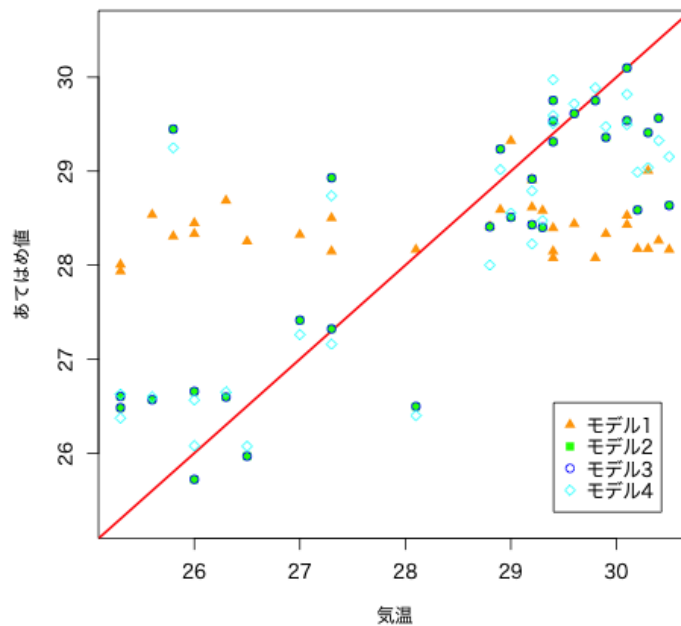


図 1: モデルの比較

- 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
-3.6477	-0.3836	0.0493	0.5511	1.8650

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	197.29993	61.11379	3.228	0.00317 **
press	-0.17149	0.06086	-2.818	0.00877 **
solar	0.20863	0.03072	6.792	2.23e-07 ***

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

Residual standard error: 1.11 on 28 degrees of freedom

Multiple R-squared: 0.6332, Adjusted R-squared: 0.607

F-statistic: 24.17 on 2 and 28 DF, p-value: 7.977e-07

- 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
-3.6475	-0.3836	0.0494	0.5510	1.8652

Coefficients:


```

      Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.973e+02  6.259e+01   3.152  0.00394 **
press        -1.715e-01  6.330e-02  -2.709  0.01158 *
solar         2.085e-01  6.012e-02   3.469  0.00177 **
humid        -1.097e-04  6.796e-02  -0.002  0.99872
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.13 on 27 degrees of freedom
Multiple R-squared:  0.6332, Adjusted R-squared:  0.5925
F-statistic: 15.54 on 3 and 27 DF,  p-value: 4.553e-06

```

- 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
-3.4490 -0.4580 -0.0780  0.7019  1.7003

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 198.22420    60.53758   3.274 0.002902 **
press        -0.17082     0.06028  -2.834 0.008602 **
solar         0.16740     0.04505   3.716 0.000934 ***
cloud        -0.12979     0.10459  -1.241 0.225311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.099 on 27 degrees of freedom
Multiple R-squared:  0.653, Adjusted R-squared:  0.6144
F-statistic: 16.94 on 3 and 27 DF,  p-value: 2.183e-06

```

- 決定係数と F -統計量

– モデル 1

```
[1] "R2: 0.0288 ; adj. R2: -0.00465 ; F-statistic: 0.861"
```

– モデル 2

```
[1] "R2: 0.633 ; adj. R2: 0.607 ; F-statistic: 24.2"
```

– モデル 3

```
[1] "R2: 0.633 ; adj. R2: 0.592 ; F-statistic: 15.5"
```

– モデル 4

```
[1] "R2: 0.653 ; adj. R2: 0.614 ; F-statistic: 16.9"
```

次週の予定

- 第 1 日: 回帰モデルの考え方と推定
- 第 2 日: モデルの評価
- 第 3 日: モデルによる予測と発展的なモデル