回帰分析 - モデルの評価

数理科学続論J

(Press `?` for help, `n` and `p` for next and previous slide)

村田 昇

2018.10.25

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

3.1

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:

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

行列・ベクトルによる簡潔な表現

- デザイン行列:

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

行列・ベクトルによる簡潔な表現

- ベクトル:

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

問題の記述

- 確率モデル:

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

- 回帰式の評価: **残差平方和** の最小化による推定

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

解の表現

- 解の条件: **正規方程式**

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

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

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

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 $\hat{\boldsymbol{y}}$ と直交する

$$\hat{\boldsymbol{\epsilon}} \cdot \hat{\boldsymbol{y}} = 0$$

- 回帰式は説明変数と目的変数の **標本平均** を通る

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

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)

残差の統計的性質

4.1

あてはめ値と誤差の関係

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

- (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

残差と誤差の関係

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

- (A)残差は誤差の重み付けの和で表される

4.3

ハット行列

- 定義:

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

- ハット行列 H の性質:
  - あてはめ値や残差は H を用いて簡潔に表現される

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

  - 観測データの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

4.4

推定量の統計的性質

5.1

推定量の性質

- 推定量と誤差の関係

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

- 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: 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:

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

- 通常 σ^2 は未知，必要な場合には不偏分散で代用

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

分析の評価

6.1

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)

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

6.2

標準誤差

- 推定されたパラメータの精度を評価:
  - 誤差の分布は平均0, 分散 σ^2 の正規分布
  - $\hat{\boldsymbol{\beta}}$ の分布は 平均 $\boldsymbol{\beta}$， 共分散 $\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の $p+1$ 変量正規分布
  - $\hat{\beta}_j$ の分布は， 行列 $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ の対角成分 を $\xi_0, \xi_1, \dots, \xi_p$ とすると， 平均 β_j, 分散 $\sigma^2 \xi_j$ の正規分布
  - 未知パラメータ σ^2 は不偏分散 $\hat{\sigma}^2$ で推定
- **標準誤差 (standard error)**: $\hat{\beta}_j$ の精度の評価指標

$$\hat{\sigma}\sqrt{\xi_j} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2} \cdot \sqrt{\xi_j}$$

演習: 標準誤差

- 05-se.r を確認してみよう

6.4

t -統計量

- **回帰係数の分布に関する定理**:
  t -統計量 は自由度 $n-p-1$ の t 分布に従う:

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

- 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
  - $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{\xi_j})$ 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

6.5

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: $\beta_j = 0$
  - 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

6.6

演習

- 05-tpval.r を確認してみよう

6.7

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** は自由度 $p, n{-}p{-}1$ の F 分布に従う

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

- 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

6.8

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: $\beta_1 = \cdots = \beta_p = 0$
  - 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

6.9

演習

- 05-fstat.r を確認してみよう

演習

- 先週用いたデータの回帰分析の結果を，寄与率・標準誤差・t-統計量・F-統計量を用いて 評価してみよう
  - datasets::airquality
  - datasets::LifeCycleSavings