回帰分析

予測と発展的なモデル

村田 昇

2020.10.23

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:

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

- **誤差項** を含む確率モデルで観測データを表現:

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

問題設定

- 確率モデル:

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

- 式の評価: **残差平方和** の最小による推定

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

解

- 解の条件: **正規方程式**

$$X^{\mathsf{T}}X\boldsymbol{\beta} = X^{\mathsf{T}}\boldsymbol{y}$$

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

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

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):

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

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する
  - 帰無仮説: $\beta_j = 0$
  - 対立仮説: $\beta_j \neq 0$ (β_j は役に立つ)
- t-統計量: 各係数ごと，ξ は $(X^\top X)^{-1}$ の対角成分

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\xi_j}}$$

- p -値: 自由度 $n-p-1$ の t 分布を用いて計算

F-統計量による検定

- 説明変数のうち 1 つでも役に立つか否かを検定する
  - 帰無仮説: $\beta_1 = \cdots = \beta_p = 0$
  - 対立仮説: $\exists j\ \beta_j \neq 0$ (少なくとも 1 つは役に立つ)
- F-統計量: 決定係数 (または残差) を用いて計算

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

- p -値: 自由度 $p, n-p-1$ の F 分布で計算

回帰モデルによる予測

予測

- 新しいデータ (説明変数) $\boldsymbol{x}$ に対する予測値

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

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = \boldsymbol{w}(\boldsymbol{x})^\top \boldsymbol{y}$$

- 重みは元データと新規データの説明変数で決定

$$\boldsymbol{w}(\boldsymbol{x})^\top = (1, \boldsymbol{x}^\top)(X^\top X)^{-1}X^\top$$

予測値の分布

- 推定量は以下の性質をもつ多変量正規分布

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

- この性質を利用して以下の 3 つの値の違いを評価

$$\begin{aligned}\hat{y} &= (1, \boldsymbol{x}^\top)\hat{\boldsymbol{\beta}} && \text{(回帰式による予測値)} \\ \tilde{y} &= (1, \boldsymbol{x}^\top)\boldsymbol{\beta} && \text{(最適な予測値)} \\ y &= (1, \boldsymbol{x}^\top)\boldsymbol{\beta} + \epsilon && \text{(観測値)}\end{aligned}$$

($\hat{y}$ と y は独立な正規分布に従うことに注意)

最適な予測値との差

- 差の分布は以下の平均・分散の正規分布

$$\begin{aligned}\mathbb{E}[\tilde{y} - \hat{y}] &= \boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\ \mathrm{Var}(\tilde{y} - \hat{y}) &= \underbrace{\sigma^2 \boldsymbol{x}^\top (X^\top X)^{-1} \boldsymbol{x}}_{\hat{\boldsymbol{\beta}}\text{ の推定誤差による分散}} = \sigma^2 \gamma_c(\boldsymbol{x})^2\end{aligned}$$

- 正規化による表現

$$\frac{\tilde{y} - \hat{y}}{\sigma\gamma_c(\boldsymbol{x})} \sim \mathcal{N}(0, 1)$$

信頼区間

- 未知の分散を不偏分散で推定

$$Z = \frac{\tilde{y} - \hat{y}}{\hat{\sigma}\gamma_c(\boldsymbol{x})} \sim \mathcal{T}(n-p-1) \quad (t\text{-分布})$$

- 確率 α の信頼区間 (最適な予測値 $\tilde{y}$ が入ることが期待される区間)

$$(\hat{y} - C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}))$$

ただし C_α は以下を満たす定数

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

観測値との差

- 差の分布は以下の平均・分散の正規分布

$$\begin{aligned}\mathbb{E}[y-\hat{y}] &= \boldsymbol{x}^\top\boldsymbol{\beta} - \boldsymbol{x}^\top\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0\\ \mathrm{Var}(y-\hat{y}) &= \underbrace{\sigma^2\boldsymbol{x}^\top(X^\top X)^{-1}\boldsymbol{x}}_{\hat{\boldsymbol{\beta}}\text{ の推定誤差による分散}} + \underbrace{\sigma^2}_{\text{誤差の分散}} = \sigma^2\gamma_p(\boldsymbol{x})^2\end{aligned}$$

- 正規化による表現

$$\frac{y-\hat{y}}{\sigma\gamma_p(\boldsymbol{x})} \sim \mathcal{N}(0,1)$$

予測区間

- 未知の分散を不偏分散で推定

$$Z = \frac{y-\hat{y}}{\hat{\sigma}\gamma_p(\boldsymbol{x})} \sim \mathcal{T}(n{-}p{-}1) \quad (t\text{-分布})$$

- 確率 α の予測区間 (観測値 y が入ることが期待される区間)

$$(\hat{y} - C_\alpha\hat{\sigma}\gamma_p(\boldsymbol{x}),\ \hat{y} + C_\alpha\hat{\sigma}\gamma_p(\boldsymbol{x}))$$

  ただし C_α は以下を満たす定数

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n{-}p{-}1)) = \alpha$$

- $\gamma_p > \gamma_c$ なので信頼区間より広くなる

発展的なモデル

非線形な関係のモデル化

- 目的変数 Y
- 説明変数 $X_1, \dots, X_p$
- 説明変数の追加で対応可能
  - 交互作用 (交差項): X_iX_j のような説明変数の積
  - 非線形変換: $\log(X_k)$ のような関数による変換

R: 線形でないモデル式の書き方

- 交互作用を記述するためには特殊な記法がある
- 非線形変換はそのまま関数を記述すればよい
- 1 つの変数の多項式は関数 `I()` を用いる

```r
## 目的変数 Y, 説明変数 X1,X2,X3
## 交互作用を含む式 (formula) の書き方
Y ~ X1 + X1:X2        # X1 + X1*X2
Y ~ X1 * X2           # X1 + X2 + X1*X2
Y ~ (X1 + X2 + X3)^2 # X1 + X2 + X3 + X1*X2 + X2*X3 + X3*X1
## 非線形変換を含む式 (formula) の書き方
Y ~ f(X1)             # f(X1) (f は任意の関数)
Y ~ X1 + I(X1^2)      # X1 + X1^2
```

カテゴリデータ

- 悪性良性や血液型などの数値ではないデータ
- 適切な方法で数値に変換して対応:
  - 2 値の場合は 0,1 を割り当てる
    - 悪性:1
    - 良性:0
  - 3 値以上の場合は **ダミー変数** を利用する (カテゴリ数-1 個)
    - A 型: (1,0,0)
    - B 型: (0,1,0)
    - O 型: (0,0,1)
    - AB 型: (0,0,0)

R: カテゴリデータの取り扱い

- 通常は適切に対応してくれる
- カテゴリデータとして扱いたい場合は `factor()` を利用する

```r
## データフレーム mydat1, mydat2, mydat3
## 変数名 X, Y, Z
mydat2 <- transform(mydat1, Y=as.factor(X))
mydat3 <- transform(mydat1, Z=as.factor(X > 0))
```