回帰分析

モデルの評価

村田 昇

2020.10.16

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:

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

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:

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

問題の記述

- 確率モデル:

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

- 回帰式の評価: **残差平方和** の最小化による推定

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

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)

練習問題 (前回の宿題)

- 決定係数を用いてモデルの比較を行ってみなさい
  - 東京の 8 月の気候データ

    temp ~ solar

    temp ~ solar + press

    temp ~ solar + press + cloud

残差の性質

あてはめ値

- あてはめ値のさまざまな表現:

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

  - (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

あてはめ値と誤差の関係

- 残差と誤差

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

  - (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:

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

- ハット行列 H の性質:

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

  - あてはめ値や残差は H を用いて簡潔に表現される
  - 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

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

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

- 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: 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

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

- 通常 σ^2 は未知，必要な場合には不偏分散で代用

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

- この性質を利用してモデルの評価を行う

R: 乱数を用いた人工データの生成

- 正規乱数を用いた線形回帰モデルの例

```
### 人工データによる推定量の性質の確認
set.seed(987) # 乱数のシード
xobs <- c(1, 3, 5, 7) # 説明変数の観測値
epsilon <- rnorm(length(xobs),sd=0.5) # 誤差項の生成
yobs <- 2 - 3*xobs + epsilon # 目的変数の観測値
myData <- data.frame(x=xobs,y=yobs) # データフレームの作成
est <- lm(y ~ x, data=myData) # 回帰係数の推定
coef(est) # 回帰係数の取得
summary(est) # 分析結果の概要の表示
```

R: 数値実験

- 実験のためのコードは以下のようになる

```
mc <- 5000 # 実験回数を指定
myTrial <- function(){ # 1 回の試行を行うプログラム
    # 乱数生成と推定の処理
    return(返り値)}
myData <- as.data.frame(t( # 実験結果を転置してデータフレームに変換
    replicate(mc, myTrial()))) # Monte-Carlo 実験
```

```r
## 適切な統計・視覚化処理 (下記は例)
apply(myData,2,var) # 各列の分散の計算
plot(myData) # 散布図行列の描画
hist(myData[[k]]) # k 列目のデータのヒストグラム
```

練習問題

- 数値実験により最小二乗推定量の性質を確認しなさい
  - 以下のモデルに従う人工データを生成する

    説明変数の観測データ:

    $$\{1, 20, 13, 9, 5, 15, 19, 8, 3, 4\}$$

    確率モデル:

    $$y = -1 + 2 \times x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 2)$$
  - 観測データから回帰係数を推定する
  - 実験を複数回繰り返し推定値 $(\hat{\beta}_0, \hat{\beta}_1)$ の分布を調べる

誤差の評価

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)

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

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:

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

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

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

    $(A)_{jj}$ は行列 A の (j, j) (対角) 成分

- **標準誤差** (standard error): $\hat{\beta}_j$ の標準偏差の推定量

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

  - 未知パラメータ σ^2 は不偏分散 $\hat{\sigma}^2$ で推定
  - $\hat{\beta}_j$ の精度の評価指標

練習問題

- 数値実験により標準誤差の性質を確認しなさい
  - 人工データを用いて標準誤差と真の誤差を比較しなさい

```
### 標準誤差は以下のようにして取り出せる
est <- lm(formula, data)
summary(est)$coef[,"Std. Error"] # 列名での指定
summary(est)$coefficients[,2] # 列番号での指定. coef と省略してもよい
```

  - 広告費と売上データを用いて係数の精度を議論しなさい
  - 東京の気候データを用いて係数の精度を議論しなさい

係数の評価

t-統計量

- **回帰係数の分布に関する定理**:

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

  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 $\hat{\boldsymbol{\beta}}$ 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

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

練習問題

- 数値実験により t-統計量の性質を確認しなさい
  - 人工データを用いて t-統計量の分布を確認しなさい

```
### t-統計量とその p-値は以下のようにして取り出せる
est <- lm(formula, data)
summary(est)$coef[,c("t value","Pr(>|t|)")] # 列名での指定
summary(est)$coef[,3:4] # 列番号での指定
```

  - 広告費と売上データを用いて係数の有意性を議論しなさい
  - 東京の気候データを用いて係数の有意性を議論しなさい

モデルの評価

F-統計量

- **ばらつきの比に関する定理**:

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

$\beta_1 = \cdots = \beta_p = 0$ ならば，F**-統計量** は自由度 $p, n-p-1$ の F 分布に従う

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

練習問題

- 数値実験により F-統計量の性質を確認しなさい
  - 人工データを用いて F-統計量の分布を確認しなさい

```
### f-統計量とその自由度は以下のようにして取り出せる
est <- lm(formula, data)
summary(est)$fstat
summary(est)$fstatistic # 省略しない場合
```

  - 広告費と売上データのモデルの有効性を議論しなさい
  - 東京の気候データのモデルの有効性を議論しなさい