回帰分析

モデルの推定

村田 昇

2020.10.09

疑似乱数

疑似乱数とは

- コンピューターで生成された数列のこと
- 完全にランダムに数字を発生されることは不可能
- R の既定値は “Mersenne-Twister 法” (`?Random` 参照)
- 数値シミュレーションにおいて再現性が要請される場合には，乱数の “シード値” を指定して再現性を担保 (関数 `set.seed()`)

基本的な乱数

- **ランダムサンプリング**: 与えられた集合の要素を無作為抽出することで発生する乱数
- **二項乱数**:「確率 p で表がでるコインを n 回投げた際の表が出る回数」に対応する乱数
- **一様乱数**: 決まった区間 (a, b) からランダムに発生する乱数
- **正規乱数**: 平均 μ, 分散 σ^2 の正規分布に従う乱数

乱数を生成する関数

- 関数 `sample()`: ランダムサンプリング
- 関数 `rbinom()`: 二項乱数
- 関数 `runif()`: 一様乱数
- 関数 `rnorm()`: 正規乱数

練習問題

- ヘルプを用いて以下の関数を調べよ
  - 関数 `sample()`
  - 関数 `rbinom()`
  - 関数 `runif()`
  - 関数 `rnorm()`
  - 関数 `set.seed()`
- 以下の試行を実装してみよ
  - サイコロを 10 回振る
  - 4 枚のコインを投げたときの表の枚数

モンテカルロ法

モンテカルロ法とは

- 乱数を使った統計実験
- 計算機上でランダムネスを実現 (擬似乱数)
- ランダムネスから導かれる種々の数学的結果を観察

例: 中心極限定理

- 定理

  $X_1, X_2, \ldots$ を独立同分布な確率変数列とし，その平均を μ，標準偏差を σ とする．このとき，すべての実数 $a < b$ に対して

  $$P\Big(a \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le b\Big) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \quad (n \to \infty)$$

  が成り立つ．

- 直感的には “多数の独立な確率変数の和はほぼ正規分布に従う” ことを主張している
- 中心極限定理のシミュレーション例

```
## 確率変数の分布の設定 (例: 区間 [-1,1] の一様乱数)
myrand <- function(n) { # n 個の乱数を生成
    return(runif(n,min=-1,max=1))
}
## 標本平均の計算
mymean <- function(n) { # n 個のデータで計算
    return(mean(myrand(n)))
}
```

```
## Monte-Carlo 実験
set.seed(123) # 実験を再現したい場合はシードを指定する
mu <- 0; sigma <- sqrt(1/3) # 理論平均と標準偏差
mc <- 5000 # 実験の繰り返し回数
for(n in c(1,2,4,8,16)){ # n を変えて実験
    xbars <- replicate(mc, mymean(n)) # mc 回実験し標本平均を記録
    hist(xbars, breaks=25, freq=FALSE, # 分布を表示
         col="orchid", border="slateblue",
         xlab=expression(bar(X)), main=paste0("n=",n))
    thdist <- function(x){dnorm(x,mean=mu,sd=sigma/sqrt(n))}
    curve(thdist, add=TRUE, col="orange", lwd=2) # 理論曲線を重ねる
}
```

例: コイン投げの賭け

- A と B の二人で交互にコインを投げる．最初に表が出た方を勝ちとするとき，A と B それぞれの勝率はいくつとなるか？
- コイン投げは関数 `sample()`, `rbinom()` などを用いて模擬できる

```
sample(0:1,1) # 0 と 1 が入った壺からから 1 つ選ぶ
rbinom(1,size=1,prob=0.5) # 表裏が等確率で出る 1 枚のコインを 1 回投げる
```

- コイン投げの賭けのシミュレーション例

```r
## コイン投げの試行（いろいろな書き方があるので以下は一例）
mytrial <- function(){
    while(TRUE){ # 永久に回るループ
        if(rbinom(1,size=1,prob=0.5)==1){return("A")} # A が表で終了
        if(rbinom(1,size=1,prob=0.5)==1){return("B")} # B が表で終了
        ## どちらも裏ならもう一度ループ
    }
}
```

```r
## Monte-Carlo 実験
set.seed(8888) # 実験を再現したい場合はシードを指定する
mc <- 10000 # 実験回数を設定
mydata <- replicate(mc,mytrial())
## 簡単な集計
table(mydata)    # 頻度
table(mydata)/mc # 確率（推定値）
```

練習問題

- 以下の簡単な双六ゲームの実験を行ってみよう
  - ゴールまでのます目は 100
  - さいころを振り出た目の数だけ進む
  - ゴールに辿り着くまで繰り返す
  - さいころを振る回数の分布は?

補足

- より詳細な確率シミュレーションについては以下を参照して下さい
  - 講義ノート
    - R 言語の基礎 第 5 章 (pp71-82)
      - Buffon の針
      - Monty Hall 問題
      - 秘書問題 (最適停止問題)

      などの実装例がある
  - 統計データ解析 I スライド
    - 講義 5 確率シミュレーション

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

回帰分析の考え方

回帰分析 (regression analysis)

- ある変量を別の変量で説明する関係式を構成
- 関係式: **回帰式** (regression equation)
  - 説明される側: **目的変数**, 被説明変数, 従属変数, 応答変数

  - 説明する側: **説明変数**, 独立変数, 共変量
- 説明変数の数による分類:
  - 一つの場合: **単回帰** (simple regression)
  - 複数の場合: **重回帰** (multiple regression)

一般の回帰の枠組

- 説明変数: $x_1, \dots, x_p$ (p 次元)
- 目的変数: y (1 次元)
- 観測データ: n 個の $(y, x_1, \dots, x_p)$ の組

$$\{(y_i, x_{i1}, \dots, x_{ip})\}_{i=1}^n$$

- y を $x_1, \dots, x_p$ で説明するための関係式を構成:

$$y = f(x_1, \dots, x_p)$$

線形回帰 (linear regression)

- 任意の f では一般的すぎて分析に不向き
- f として 1 次関数を考える
  ある定数 $\beta_0, \beta_1, \dots, \beta_p$ を用いた以下の式:

$$f(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - 1 次関数の場合: **線形回帰** (linear regression)
  - 一般の場合: 非線形回帰 (nonlinear regression)
- 非線形関係は新たな説明変数の導入で対応可能
  - 適切な多項式 $x_j^2, x_j x_k, x_j x_k x_l, \dots$
  - その他の非線形変換 $\log x_j, x_j^\alpha, \dots$

回帰係数

- 線形回帰式:

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

- $\beta_0, \beta_1, \dots, \beta_p$: **回帰係数** (regression coefficients)
- β_0: **定数項 / 切片** (constant term / intersection)
- 線形回帰分析: 未知の回帰係数をデータから決定

回帰の確率モデル

- 回帰式の不確定性
  - データは一般に観測誤差などランダムな変動を含む
  - 回帰式がそのまま成立することは期待できない
- 確率モデル: データのばらつきを表す項 ϵ_i を追加

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

  - $\epsilon_1, \ldots, \epsilon_n$: **誤差項 / 撹乱項** (error / disturbance term)
    - 誤差項は独立な確率変数と仮定
    - 多くの場合，平均 0，分散 σ^2 の正規分布を仮定
  - **推定** (estimation): 観測データから $(\beta_0, \beta_1, \ldots, \beta_p)$ を決定

回帰係数の推定

残差

- **残差** (residual): 回帰式で説明できない変動
- 回帰係数 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}}$ を持つ回帰式の残差:

$$e_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \quad (i = 1, \ldots, n)$$

- 残差 $e_i(\boldsymbol{\beta})$ の絶対値が小さいほど当てはまりがよい

最小二乗法 (least squares)

- **残差平方和** (residual sum of squares):

$$S(\boldsymbol{\beta}) := \sum_{i=1}^{n} e_i(\boldsymbol{\beta})^2$$

- **最小二乗推定量** (least squares estimator):
  残差平方和 $S(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathsf{T}} := \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

行列の定義

- デザイン行列 (design matrix):

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

ベクトルの定義

- 目的変数，誤差，回帰係数のベクトル:

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

行列・ベクトルによる表現

- 確率モデル:

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

- 残差平方和:

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

解の条件

- 解 $\boldsymbol{\beta}$ では残差平方和の勾配は零ベクトル

$$\nabla S(\boldsymbol{\beta}) := \Big(\frac{\partial S}{\partial \beta_0}(\boldsymbol{\beta}), \frac{\partial S}{\partial \beta_1}(\boldsymbol{\beta}), \dots, \frac{\partial S}{\partial \beta_p}(\boldsymbol{\beta})\Big)^\mathsf{T} = \boldsymbol{0}$$

- 成分 $(j = 0, 1, \dots, p)$ ごとの条件式

$$\frac{\partial S}{\partial \beta_j}(\boldsymbol{\beta}) = -2\sum_{i=1}^{n}\Big(y_i - \sum_{k=0}^{p}\beta_k x_{ik}\Big)x_{ij} = 0$$

  但し $x_{i0} = 1\ (i = 1, \dots, n)$

正規方程式 (normal equation)

- 条件を整理 (x_{ij} は行列 X の (i, j) 成分)

$$\sum_{i=1}^{n} x_{ij}\Big(\sum_{k=0}^{p} x_{ik}\beta_k\Big) = \sum_{i=1}^{n} x_{ij}y_i \quad (j = 0, 1, \dots, p)$$

- **正規方程式** (normal equation):

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

- **Gram 行列** (Gram matrix): $X^\mathsf{T}X$

正規方程式の解

- 正規方程式の基本的な性質
  - 正規方程式は必ず解をもつ (一意に決まらない場合もある)
  - 正規方程式の解は最小二乗推定量であるための必要条件
- Gram 行列 $X^{\mathsf{T}}X$ が正則ならば解が一意に決定
- 正規方程式の解

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

R: 関数 `lm()` による推定

- ベクトルを用いる基本的な使い方:

```
lm(formula= y ~ x1 + ... + xp)
## formula: 目的変数 ~ 説明変数 (複数ある場合は + で並べる)
## y: 目的変数のベクトル
## x1,...,xp: 各説明変数のベクトル
```

- データフレームを用いる方法: **(こちらを推奨)**

```
lm(formula= y の変数名 ~ x1 の変数名 + ... + xp の変数名,
   data = データフレーム)
## formula: 目的変数名 ~ 説明変数名
## data: 目的変数, 説明変数を含むデータフレーム
```

データセットの準備

- 回帰分析では以下のデータセットを使用します
  - `tokyo_weather_reg.csv`

    気象庁より取得した東京の気候データを回帰分析用に整理したもの

    https://www.data.jma.go.jp/gmd/risk/obsdl/index.php
  - `Advertising.csv`

    広告費 (TV,radio,newspapers) と売上の関係を調べたもの

    “Datasets in this presentation are taken from ”An Introduction to Statistical Learning, with applications in R“ (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani ”

    http://faculty.marshall.usc.edu/gareth-james/ISL/Advertising.csv

練習問題

- 前掲のデータセットを用いて回帰式を構成しなさい
  - 東京の 8 月の気候データ

    temp ~ solar + press
  - 広告費と売上データ

    sales ~ TV

    sales ~ radio

    sales ~ TV + radio

最小二乗推定量の性質

解と観測データの関係

- 解析の上での良い条件:
  - 最小二乗推定量がただ一つだけ存在する (以下同値条件)
    - $X^\mathsf{T}X$ が正則
    - $X^\mathsf{T}X$ の階数が $p+1$
    - X の階数が $p+1$
    - X の列ベクトルが 1 次独立
- 解析の上での良くない条件:
  - 説明変数が 1 次従属: **多重共線性** (multicollinearity)
  - 説明変数は多重共線性が強くならないように選択するべき
    - X の列 (説明変数) の独立性を担保する
    - 説明変数が互いに異なる情報をもつように選ぶ
    - 似た性質をもつ説明変数の重複は避ける

推定の幾何学的解釈

- **あてはめ値 / 予測値** (fitted values / predicted values):

$$\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = \hat{\beta}_0 X_{\text{第 0 列}} + \cdots + \hat{\beta}_p X_{\text{第 p 列}}$$

- 最小二乗推定量 $\hat{\boldsymbol{y}}$ の幾何学的性質:
  - $L[X]$: X の列ベクトルが張る $\mathbb{R}^n$ の部分線形空間
  - X の階数が $p+1$ ならば $L[X]$ の次元は $p+1$ (解の一意性)
  - $\hat{\boldsymbol{y}}$ は $\boldsymbol{y}$ の $L[X]$ への直交射影
  - **残差** (residuals) $\hat{\boldsymbol{\epsilon}} := \boldsymbol{y} - \hat{\boldsymbol{y}}$ はあてはめ値 $\hat{\boldsymbol{y}}$ に直交

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

- 幾何学的な考察からも一意に決まる

線形回帰式と標本平均

- $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$: 説明変数の i 番目の観測データ
- 説明変数および目的変数の標本平均:

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

- $\hat{\boldsymbol{\beta}}$ が最小二乗推定量のとき以下が成立:

$$\bar{y} = (1, \bar{\boldsymbol{x}}^\mathsf{T})\hat{\boldsymbol{\beta}}$$

- 以下の関係から簡単に示すことができる

$$\mathbf{1} \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \mathbf{1} \cdot \hat{\boldsymbol{\epsilon}} = 0$$

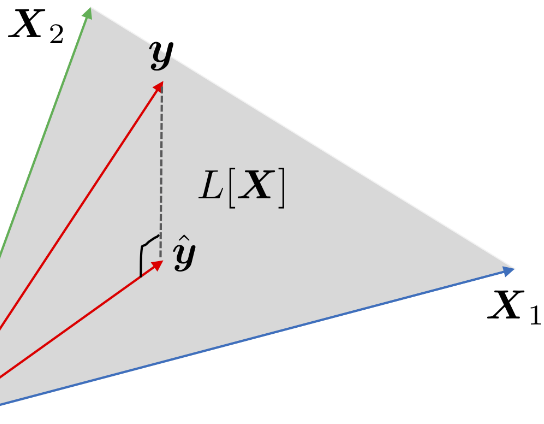

図 1: $n = 3$, $p + 1 = 2$ の場合の最小二乗法による推定

R: 推定結果からの情報の取得

- 関数 `lm()` の出力には様々な情報が含まれる

```
## lm の出力を引数とする関数の例
coef(lm の出力)          # 推定された回帰係数
fitted(lm の出力)        # あてはめ値
resid(lm の出力)         # 残差
model.frame(lm の出力)   # model に必要な変数の抽出 (データフレーム)
model.matrix(lm の出力)  # デザイン行列
```

R: 行列とベクトルの計算

- $X^\top Y$ および $X^\top X$ の計算

```
crossprod(X, Y) # cross product の略
## X: 行列 (またはベクトル)
## Y: 行列 (またはベクトル)
crossprod(X) # 同じものを掛ける場合は引数は 1 つで良い
```

- 行列 A, B の積 AB

```
A %*% B # 行列の大きさは適切である必要がある
```

- 正方行列 A の逆行列 A^{-1}

```
solve(A) # 他にもいくつか関数はある
```

練習問題

- 前問の推定結果を用いて最小二乗推定量の性質を確認しなさい
  - 推定された係数が正規方程式の解

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

となること
  - あてはめ値と残差が直交すること
  - 回帰式が標本平均を通ること

残差の分解

最小二乗推定量の残差

- 観測値と推定値 $\hat{\boldsymbol{\beta}}$ による予測値の差:

$$\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}) \quad (i = 1, \ldots, n)$$

  - 誤差項 $\epsilon_1, \ldots, \epsilon_n$ の推定値
  - 全てができるだけ小さいほど良い
  - 予測値とは独立に偏りがないほど良い

- 残差ベクトル:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n)^{\mathsf{T}}$$

平方和の分解

- 標本平均のベクトル: $\bar{\boldsymbol{y}} = \bar{y}\mathbf{1} = (\bar{y}, \bar{y}, \ldots, \bar{y})^{\mathsf{T}}$
- いろいろなばらつき
  - $S_y = (\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}})$: 目的変数のばらつき
  - $S \ = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}})$: 残差のばらつき ($\hat{\boldsymbol{\epsilon}}^{\mathsf{T}}\hat{\boldsymbol{\epsilon}}$)
  - $S_r = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})$: あてはめ値 (回帰) のばらつき
- 3 つのばらつき (平方和) の関係

$$\begin{aligned}
&(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) \\
&S_y = S + S_r
\end{aligned}$$

- 証明には以下の関係を使う

$$\begin{aligned}
&\boldsymbol{y} - \bar{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{\boldsymbol{y}} \\
&\hat{\boldsymbol{y}} \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}} \cdot \hat{\boldsymbol{\epsilon}} = 0 \\
&\mathbf{1} \cdot (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \mathbf{1} \cdot \hat{\boldsymbol{\epsilon}} = 0
\end{aligned}$$

練習問題

- 前問の推定結果を用いて残差の性質を確認しなさい
  - 以下の分解

$$\begin{aligned}
&(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \bar{\boldsymbol{y}}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathsf{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^{\mathsf{T}}(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) \\
&S_y = S + S_r
\end{aligned}$$

    が成り立つこと

決定係数

回帰式の寄与

- ばらつきの分解:

$$S_y\ (\text{目的変数}) = S\ (\text{残差}) + S_r\ (\text{あてはめ値})$$

- 回帰式で説明できるばらつきの比率:

$$(\text{回帰式の寄与率}) = \frac{S_r}{S_y} = 1 - \frac{S}{S_y}$$

- 回帰式のあてはまり具合を評価する代表的な指標

決定係数 (R^2 値)

- **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):

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

  不偏分散で補正している

練習問題

- 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
  - 広告費と売上データ

    sales ~ TV

    sales ~ radio

    sales ~ TV + radio

補足

- 人工データを作成して回帰分析の性質を調べるコード (Rscript) ついては以下を参照して下さい
  - 講義ノート
    - 多変量解析 第 1 章 単回帰分析
    - 多変量解析 第 2 章 回帰分析