hdm パッケージを用いた LASSO/Double Selection の実装

川田恵介

Table of contents

1 流れ

1. Set up
2. Import Data (arrow)
3. PreProcess (recipes)
4. LASSO/Double selection (hdm)

1.1 Set up

```
set.seed(111)

library(tidyverse)
```

```
library(arrow)                                                    ①
library(hdm)
```

① parquet 形式 (大規模データに比較優位) によるデータの保存/読み込み

1.2 Install package

- 基本、通常のやり方で OK だが、
  - 現時点では、arrow を Mac にインストール際には以下を実行

```
install.packages("arrow", repos = c("https://apache.r-universe.dev"))
```

1.3 Import Data

- csv 形式の導入

```
Raw = read_csv("Public/Data.csv")
```

- parquet 形式の導入

```
Raw = read_parquet("Public/Data.parquet")
```

1.4 データの形式

- 多くの選択肢が存在 (R for Data Science 20-23 参照)
- 個人的には、parquet 形式が最もバランスが良い印象
  - Size が小さくなる/読み込みが早い/メモリを使わない操作が可能/DataBase よりも初学者にとって簡単?
- csv 形式を parquet 形式に変換して保存

```
open_csv_dataset("Public/Data.csv") |>
  write_parquet("Data.parquet")
```

1.5 Pipe

- input を、名前をつけずに、output として利用可能
- 例

```
Raw = read_parquet("Public/Data.parquet")
summary(Raw)
```

- 以下は同じ結果を出す

```
read_parquet("Public/Data.parquet") |>
  summary()
```

- Shortcut: command(control) + m
  - Tools -> Global option -> Code -> Use native pipe oeprator をチェック

1.6 PreProcess: dplyr

- 基本は dplyr(tidyverse に含まれる) の使用を推奨 (R for Data Science 4 章参照)

```
Data = Raw |>
  filter(TradeYear == 2022,
         TradeQ == 4) |>
  select(
    Size:Reform,
    Price,
    District
  )
```

1.7 LASSO

```
Group = sample(
  1:2,
  nrow(Data),
  replace = TRUE,
  prob = c(0.8,0.2))

Model = rlasso(
  Price ~ .^2 + I(Size^2) + I(Distance^2) + I(Youseki^2),
  Data[Group == 1,],
  post = FALSE
)
```

1.8 LASSO: Selected variable

```
Model$index
```

Size	Distance	Tenure
TRUE	TRUE	TRUE
Youseki	Reform	District 中央区
FALSE	FALSE	FALSE
District 中野区	District 北区	District 千代田区
FALSE	FALSE	FALSE
District 台東区	District 品川区	District 大田区
FALSE	FALSE	FALSE
District 文京区	District 新宿区	District 杉並区
FALSE	FALSE	FALSE
District 板橋区	District 江戸川区	District 江東区
FALSE	FALSE	FALSE
District 渋谷区	District 港区	District 目黒区
FALSE	FALSE	TRUE
District 練馬区	District 荒川区	District 葛飾区
FALSE	FALSE	FALSE
District 豊島区	District 足立区	District 墨田区
FALSE	FALSE	FALSE
I(Size^2)	I(Distance^2)	I(Youseki^2)
TRUE	TRUE	FALSE
Size:Distance	Size:Tenure	Size:Youseki
FALSE	TRUE	TRUE
Size:Reform	Size:District 中央区	Size:District 中野区
TRUE	FALSE	TRUE
Size:District 北区	Size:District 千代田区	Size:District 台東区
TRUE	TRUE	TRUE
Size:District 品川区	Size:District 大田区	Size:District 文京区
FALSE	TRUE	FALSE
Size:District 新宿区	Size:District 杉並区	Size:District 板橋区
FALSE	TRUE	TRUE
Size:District 江戸川区	Size:District 江東区	Size:District 渋谷区
TRUE	TRUE	TRUE
Size:District 港区	Size:District 目黒区	Size:District 練馬区
TRUE	TRUE	TRUE

Size:District 荒川区	Size:District 葛飾区	Size:District 豊島区
TRUE	TRUE	FALSE
Size:District 足立区	Size:District 墨田区	Distance:Tenure
TRUE	TRUE	FALSE
Distance:Youseki	Distance:Reform	Distance:District 中央区
FALSE	FALSE	FALSE
Distance:District 中野区	Distance:District 北区	Distance:District 千代田区
TRUE	FALSE	FALSE
Distance:District 台東区	Distance:District 品川区	Distance:District 大田区
TRUE	FALSE	FALSE
Distance:District 文京区	Distance:District 新宿区	Distance:District 杉並区
FALSE	FALSE	FALSE
Distance:District 板橋区	Distance:District 江戸川区	Distance:District 江東区
FALSE	FALSE	FALSE
Distance:District 渋谷区	Distance:District 港区	Distance:District 目黒区
FALSE	FALSE	FALSE
Distance:District 練馬区	Distance:District 荒川区	Distance:District 葛飾区
FALSE	TRUE	FALSE
Distance:District 豊島区	Distance:District 足立区	Distance:District 墨田区
FALSE	FALSE	FALSE
Tenure:Youseki	Tenure:Reform	Tenure:District 中央区
TRUE	TRUE	FALSE
Tenure:District 中野区	Tenure:District 北区	Tenure:District 千代田区
FALSE	FALSE	FALSE
Tenure:District 台東区	Tenure:District 品川区	Tenure:District 大田区
FALSE	FALSE	FALSE
Tenure:District 文京区	Tenure:District 新宿区	Tenure:District 杉並区
FALSE	FALSE	FALSE
Tenure:District 板橋区	Tenure:District 江戸川区	Tenure:District 江東区
FALSE	FALSE	FALSE
Tenure:District 渋谷区	Tenure:District 港区	Tenure:District 目黒区
FALSE	FALSE	FALSE
Tenure:District 練馬区	Tenure:District 荒川区	Tenure:District 葛飾区
FALSE	TRUE	FALSE
Tenure:District 豊島区	Tenure:District 足立区	Tenure:District 墨田区
FALSE	FALSE	FALSE
Youseki:Reform	Youseki:District 中央区	Youseki:District 中野区
FALSE	FALSE	FALSE
Youseki:District 北区	Youseki:District 千代田区	Youseki:District 台東区
FALSE	FALSE	FALSE

Youseki:District 品川区	Youseki:District 大田区	Youseki:District 文京区
FALSE	FALSE	FALSE
Youseki:District 新宿区	Youseki:District 杉並区	Youseki:District 板橋区
FALSE	TRUE	TRUE
Youseki:District 江戸川区	Youseki:District 江東区	Youseki:District 渋谷区
FALSE	TRUE	FALSE
Youseki:District 港区	Youseki:District 目黒区	Youseki:District 練馬区
FALSE	FALSE	FALSE
Youseki:District 荒川区	Youseki:District 葛飾区	Youseki:District 豊島区
FALSE	FALSE	FALSE
Youseki:District 足立区	Youseki:District 墨田区	Reform:District 中央区
FALSE	FALSE	FALSE
Reform:District 中野区	Reform:District 北区	Reform:District 千代田区
FALSE	FALSE	FALSE
Reform:District 台東区	Reform:District 品川区	Reform:District 大田区
FALSE	FALSE	FALSE
Reform:District 文京区	Reform:District 新宿区	Reform:District 杉並区
FALSE	FALSE	FALSE
Reform:District 板橋区	Reform:District 江戸川区	Reform:District 江東区
FALSE	FALSE	FALSE
Reform:District 渋谷区	Reform:District 港区	Reform:District 目黒区
FALSE	FALSE	FALSE
Reform:District 練馬区	Reform:District 荒川区	Reform:District 葛飾区
FALSE	FALSE	FALSE
Reform:District 豊島区	Reform:District 足立区	Reform:District 墨田区
FALSE	FALSE	FALSE

1.9 LASSO: Evaluation

```
PredLASSO = Model |>
  predict(
    Data)

PredOLS = lm(
  Price ~ .,
  Data[Group == 1,]) |> # 比較のために OLS を推定
  predict(Data)

1 - mean((Data$Price - PredLASSO)[Group == 2]^2)/var(Data$Price)
```

```
[1] 0.7926756
```

```
1 - mean((Data$Price - PredOLS)[Group == 2]^2)/var(Data$Price)
```

```
[1] 0.7075551
```

1.10 Double Selection

- サンプル分割は不要

```
Y = Data$Price
D = Data$Reform

X = model.matrix(
  ~ 0 + .^2 + I(Size^2) + I(Distance^2) + I(Tenure^2) + I(Youseki^2),
  Data |>
    select(
      -Price,
      -Reform
    )
)

Model = rlassoEffect(
  x = X,
  y = Y,
  d = D)

Model |> summary() # 推定結果
```

```
[1] "Estimates and significance testing of the effect of target variables"
   Estimate. Std. Error t value Pr(>|t|)
d1     8.404      1.511   5.563 2.65e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.11 OLS

```
estimatr::lm_robust(Price ~ .,
                    Data) |>
  estimatr::tidy() |>
  filter(
```

```
    term == "Reform"
  )
```

```
    term estimate std.error statistic      p.value conf.low conf.high   df
1 Reform 7.907015  1.645964  4.803882 1.639577e-06 4.679567  11.13446 2753
  outcome
1   Price
```

1.12 Double Selection: Selected Variable

```
Model$selection.index
```

Size	Distance	Tenure
TRUE	TRUE	TRUE
Youseki	District 世田谷区	District 中央区
FALSE	FALSE	FALSE
District 中野区	District 北区	District 千代田区
FALSE	FALSE	FALSE
District 台東区	District 品川区	District 大田区
TRUE	FALSE	TRUE
District 文京区	District 新宿区	District 杉並区
FALSE	FALSE	FALSE
District 板橋区	District 江戸川区	District 江東区
TRUE	FALSE	FALSE
District 渋谷区	District 港区	District 目黒区
FALSE	FALSE	TRUE
District 練馬区	District 荒川区	District 葛飾区
FALSE	TRUE	FALSE
District 豊島区	District 足立区	District 墨田区
FALSE	FALSE	TRUE
I(Size^2)	I(Distance^2)	I(Tenure^2)
TRUE	TRUE	FALSE
I(Youseki^2)	Size:Distance	Size:Tenure
FALSE	FALSE	TRUE
Size:Youseki	Size:District 中央区	Size:District 中野区
TRUE	FALSE	FALSE
Size:District 北区	Size:District 千代田区	Size:District 台東区
TRUE	TRUE	FALSE
Size:District 品川区	Size:District 大田区	Size:District 文京区
FALSE	TRUE	FALSE

Size:District 新宿区	Size:District 杉並区	Size:District 板橋区
FALSE	FALSE	TRUE
Size:District 江戸川区	Size:District 江東区	Size:District 渋谷区
TRUE	TRUE	TRUE
Size:District 港区	Size:District 目黒区	Size:District 練馬区
TRUE	TRUE	TRUE
Size:District 荒川区	Size:District 葛飾区	Size:District 豊島区
FALSE	TRUE	FALSE
Size:District 足立区	Size:District 墨田区	Distance:Tenure
TRUE	TRUE	FALSE
Distance:Youseki	Distance:District 中央区	Distance:District 中野区
FALSE	FALSE	FALSE
Distance:District 北区	Distance:District 千代田区	Distance:District 台東区
FALSE	FALSE	TRUE
Distance:District 品川区	Distance:District 大田区	Distance:District 文京区
FALSE	FALSE	FALSE
Distance:District 新宿区	Distance:District 杉並区	Distance:District 板橋区
FALSE	FALSE	FALSE
Distance:District 江戸川区	Distance:District 江東区	Distance:District 渋谷区
FALSE	FALSE	FALSE
Distance:District 港区	Distance:District 目黒区	Distance:District 練馬区
FALSE	FALSE	FALSE
Distance:District 荒川区	Distance:District 葛飾区	Distance:District 豊島区
TRUE	FALSE	FALSE
Distance:District 足立区	Distance:District 墨田区	Tenure:Youseki
FALSE	FALSE	FALSE
Tenure:District 中央区	Tenure:District 中野区	Tenure:District 北区
FALSE	FALSE	FALSE
Tenure:District 千代田区	Tenure:District 台東区	Tenure:District 品川区
FALSE	FALSE	FALSE
Tenure:District 大田区	Tenure:District 文京区	Tenure:District 新宿区
FALSE	FALSE	FALSE
Tenure:District 杉並区	Tenure:District 板橋区	Tenure:District 江戸川区
FALSE	FALSE	FALSE
Tenure:District 江東区	Tenure:District 渋谷区	Tenure:District 港区
FALSE	FALSE	FALSE
Tenure:District 目黒区	Tenure:District 練馬区	Tenure:District 荒川区
FALSE	FALSE	TRUE
Tenure:District 葛飾区	Tenure:District 豊島区	Tenure:District 足立区
FALSE	FALSE	FALSE

Tenure:District 墨田区	Youseki:District 中央区	Youseki:District 中野区
FALSE	FALSE	FALSE
Youseki:District 北区	Youseki:District 千代田区	Youseki:District 台東区
TRUE	FALSE	FALSE
Youseki:District 品川区	Youseki:District 大田区	Youseki:District 文京区
FALSE	FALSE	FALSE
Youseki:District 新宿区	Youseki:District 杉並区	Youseki:District 板橋区
FALSE	FALSE	TRUE
Youseki:District 江戸川区	Youseki:District 江東区	Youseki:District 渋谷区
FALSE	TRUE	FALSE
Youseki:District 港区	Youseki:District 目黒区	Youseki:District 練馬区
FALSE	FALSE	FALSE
Youseki:District 荒川区	Youseki:District 葛飾区	Youseki:District 豊島区
FALSE	FALSE	FALSE
Youseki:District 足立区	Youseki:District 墨田区	
FALSE	FALSE	