R

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(recipes)                                                  ②
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
② データ整備用の関数を提供

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") |>
  summry()
```

- 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(District == "文京区")

Y = Data$Price
D = case_when(
  Data$TradeYear == 2022 ~ 1,
  Data$TradeYear == 2017 ~ 0)
```

1.7 PreProcess: recipes

- 伝統的な方法に比べて、X に対して、より多くの処理が必要
  - recipe を活用すると、コード量を減らし、ミスを減らせる

```
X = recipe(
  ~  Size + Distance + Tenure + Youseki + Reform + TradeQ + Area,
  Data
  ) |> # Data から X を指定
  step_interact(
    ~ all_predictors():all_predictors()
    ) |> # すべての X について交差項を作成
  step_poly(
    Size,
    Distance,
    Tenure,
    Youseki,
```

```
    TradeQ,
    degree = 2
  ) |> # ２乗項まで作成
  step_dummy(
    all_nominal_predictors()
  ) |> # 文字/ファクター変数について、ダミーを作成
  step_zv(
    all_predictors()
  ) |> # 全ての変数について、定数であれば排除
  step_normalize(
    all_predictors()
  ) |> # 全ての変数について、標準化
  prep() |>
  bake(
    new_data = NULL,
    composition = "matrix"
  ) # matrix、として出力 (default は data.frame)
```

1.8 LASSO

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

Model = rlasso(
  x = X[Group == 1,],
  y = Y[Group == 1])
```

1.9 LASSO: Selected variable

```
Model$index
```

```
           Reform    Size_x_Distance      Size_x_Tenure     Size_x_Youseki
            FALSE             FALSE              FALSE               TRUE
     Size_x_Reform      Size_x_TradeQ       Size_x_AreaH       Size_x_AreaI
            FALSE             FALSE              FALSE              FALSE
```

Distance_x_Tenure	Distance_x_Youseki	Distance_x_Reform	Distance_x_TradeQ
FALSE	FALSE	FALSE	FALSE
Distance_x_AreaH	Distance_x_AreaI	Tenure_x_Youseki	Tenure_x_Reform
FALSE	FALSE	FALSE	FALSE
Tenure_x_TradeQ	Tenure_x_AreaH	Tenure_x_AreaI	Youseki_x_Reform
FALSE	FALSE	FALSE	FALSE
Youseki_x_TradeQ	Youseki_x_AreaH	Youseki_x_AreaI	Reform_x_TradeQ
FALSE	FALSE	FALSE	FALSE
Reform_x_AreaH	Reform_x_AreaI	TradeQ_x_AreaH	TradeQ_x_AreaI
FALSE	FALSE	FALSE	FALSE
Size_poly_1	Size_poly_2	Distance_poly_1	Distance_poly_2
TRUE	FALSE	FALSE	FALSE
Tenure_poly_1	Tenure_poly_2	Youseki_poly_1	Youseki_poly_2
TRUE	FALSE	FALSE	FALSE
TradeQ_poly_1	TradeQ_poly_2	Area_H	Area_I
FALSE	FALSE	FALSE	FALSE

1.10 Evaluation

```
PredLASSO = Model |>
  predict(
    X,
    post = FALSE)

PredOLS = lm(Y ~ X, subset = Group == 1) |>
  predict(
    X |> as_tibble())

mean((Y - PredLASSO)[Group == 2]^2)/var(Y)
```

```
[1] 0.1560659
```

```
mean((Y - PredOLS)[Group == 2]^2)/var(Y)
```

```
[1] 0.1590627
```

1.11 Double Selection

- サンプル分割は不要

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

Model |> summary()
```

```
[1] "Estimates and significance testing of the effect of target variables"
   Estimate. Std. Error t value Pr(>|t|)
d1    11.812      0.617   19.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.12 Double Selection: Selected Variable

```
Model$selection.index
```

Reform	Size_x_Distance	Size_x_Tenure	Size_x_Youseki
FALSE	FALSE	TRUE	TRUE
Size_x_Reform	Size_x_TradeQ	Size_x_AreaH	Size_x_AreaI
FALSE	FALSE	FALSE	FALSE
Distance_x_Tenure	Distance_x_Youseki	Distance_x_Reform	Distance_x_TradeQ
FALSE	FALSE	FALSE	FALSE
Distance_x_AreaH	Distance_x_AreaI	Tenure_x_Youseki	Tenure_x_Reform
FALSE	FALSE	FALSE	FALSE
Tenure_x_TradeQ	Tenure_x_AreaH	Tenure_x_AreaI	Youseki_x_Reform
FALSE	FALSE	FALSE	FALSE
Youseki_x_TradeQ	Youseki_x_AreaH	Youseki_x_AreaI	Reform_x_TradeQ
FALSE	FALSE	FALSE	FALSE
Reform_x_AreaH	Reform_x_AreaI	TradeQ_x_AreaH	TradeQ_x_AreaI
FALSE	FALSE	FALSE	FALSE
Size_poly_1	Size_poly_2	Distance_poly_1	Distance_poly_2
TRUE	FALSE	FALSE	FALSE
Tenure_poly_1	Tenure_poly_2	Youseki_poly_1	Youseki_poly_2
TRUE	FALSE	FALSE	FALSE
TradeQ_poly_1	TradeQ_poly_2	Area_H	Area_I
FALSE	FALSE	FALSE	FALSE