補論: 交差推定

機械学習入門

川田恵介

Table of contents

0.1 ２度づけ回避

- ある事例の予測値は、その事例を含まないデータから推定する必要がある
- ここまでの方法: データを２分割し、片方で予測モデルを推定し、残りで OLS 推定を行う
  - 問題点: 半分のデータしか、OLS に使えず、推定精度が悪化する
  - 改善策: 交差推定

0.2 交差推定

0. データを細かく分割 (第 1,..,10 サブグループなど)
1. 第 1 サブグループ**以外**で推定して、第 1 サブグループの予測値を算出

2. 第 2...サブグループについて、繰り返し、全事例に対して予測値を算出

0.3 数値例: 3 分割

```
# A tibble: 9 x 3
  StationDistance    Price Group
            <int>    <dbl> <fct>
1               9  6.05    3
2               4  3.94    2
3               7 31.0     3
4               1  8.64    1
5               2 -5.99    3
6               7 -4.48    1
7               2 -0.895   1
8               3  0.00785 2
9               1 -3.12    2
```

0.4 数値例: Step 1

```
# A tibble: 9 x 5
  StationDistance    Price Group   OLS RandomForest
            <int>    <dbl> <fct> <dbl>        <dbl>
1               9  6.05    3     NA           NA
2               4  3.94    2     NA           NA
3               7 31.0     3     NA           NA
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3     NA           NA
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2     NA           NA
9               1 -3.12    2     NA           NA
```

0.5 数値例: Step 2

```
# A tibble: 9 x 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    3     NA           NA
2               4  3.94    2      4.86        -0.189
```

```
3                7 31.0     3     NA           NA
4                1  8.64    1     -4.12        -1.89
5                2 -5.99    3     NA           NA
6                7 -4.48    1     12.9         16.7
7                2 -0.895   1     -1.29        -1.91
8                3  0.00785 2      3.55        -0.189
9                1 -3.12    2      0.938        1.91
```

0.6 数値例: Step 3

```
# A tibble: 9 x 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    3     -4.88        -1.84
2               4  3.94    2      4.86        -0.189
3               7 31.0     3     -3.03        -1.84
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3      1.61         0.945
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2      3.55        -0.189
9               1 -3.12    2      0.938        1.91
```

0.7 Stacking

- 交差推定から Stacking も可能

```
lm(Price ~ OLS + RandomForest, PopData)
```

```

Call:
lm(formula = Price ~ OLS + RandomForest, data = PopData)

Coefficients:
 (Intercept)           OLS  RandomForest
       5.056        -1.248         0.243
```

0.8 実装

- 以上の手続きは ddml package で実装可能

0.9 例

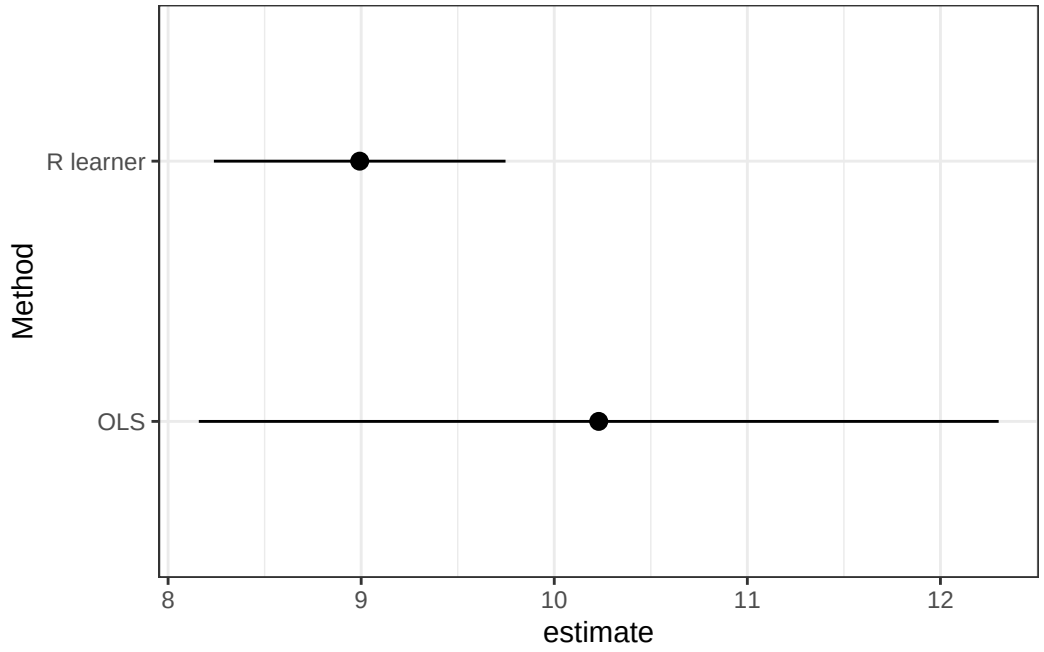
```
library(tidyverse)
library(ddml)

Y = Data$Price
D = Data$D
X = select(
  Data,
  Size,
  District,
  Distance,
  Tenure)
```

0.10 例

```
Model = ddml_plm(
  y = Y,
  D = D,
  X = data.matrix(X),
  learners = list(
    list(fun = ols),
    list(fun = mdl_glmnet),
    list(fun = mdl_ranger)
  ),
  shortstack = TRUE,
  silent = TRUE)

summary(Model)
```

```
PLM estimation results:

, , nnls

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -0.156      0.188  -0.829  4.07e-01
D_r            8.992      0.385  23.337 1.85e-120
```

0.11 比較



0.12 例

```
library(estimatr)

HatY = Model$ols_fit$model$y_r
HatD = Model$ols_fit$model$D_r

lm_robust(
  HatY ~ 0 + HatD +
    HatD:scale(Size) + HatD:scale(Tenure) + HatD:scale(Distance),
  Data)
```

```
                      Estimate Std. Error   t value      Pr(>|t|)  CI Lower
HatD                 9.0738250  0.3877271 23.402609 1.293111e-115  8.313729
HatD:scale(Size)     3.4971204  0.4196271  8.333876  9.756498e-17  2.674488
HatD:scale(Tenure)  -0.6944621  0.3404948 -2.039567  4.144052e-02 -1.361965
HatD:scale(Distance) -1.4511045  0.3913597 -3.707854  2.110542e-04 -2.218322
                      CI Upper   DF
HatD                 9.8339212 5573
```

```
HatD:scale(Size)      4.3197531 5573
HatD:scale(Tenure)   -0.0269595 5573
HatD:scale(Distance) -0.6838869 5573
```