バランス後の比較
機械学習

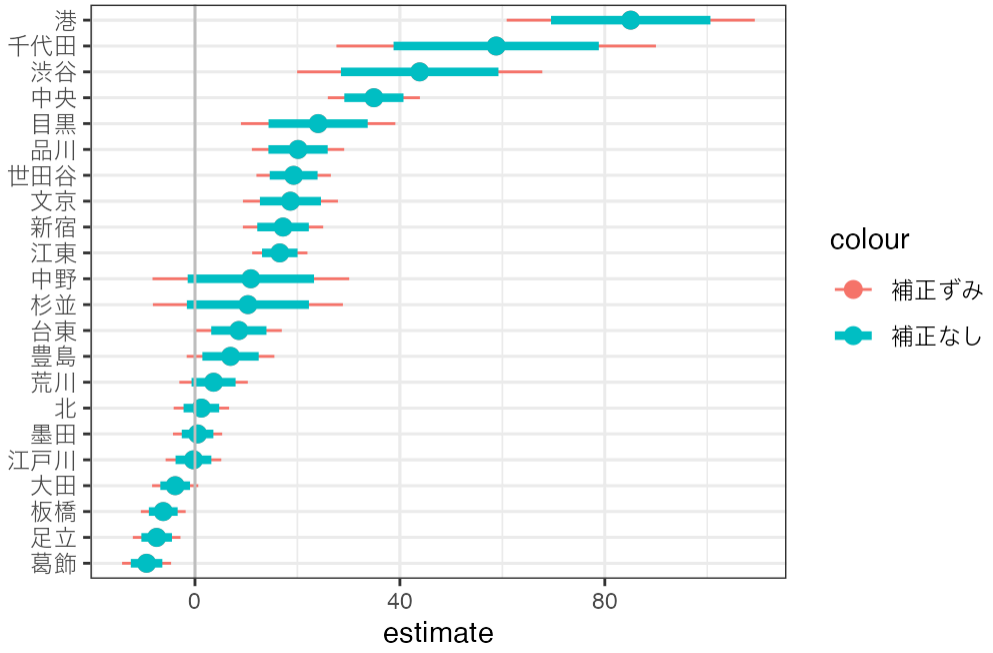
川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-10-14

1 分析例

- To do: PDS の例として、Y,D とそれぞれ強く相関している X_1,X_2 を作る
- サンプル分割は不要であることを強調
- 機械学習と伝統的な推定の共同作業 🤝

1.1 23 区間不動産価格格差



1.2 指導教員からの”コメント”

- 「区の魅力”ではなく、取引される物件の性質の違いを反映しているだけではないか?」

1.3 バランス表

```
gtsummary::tbl_summary(
  data,
  by = District
)
```

Characteristic	港区 N = 283[1]	練馬区 N = 278[1]
Price	86 (40, 150)	34 (24, 50)
Size	50 (25, 75)	50 (25, 65)
Tenure	19 (11, 23)	20 (9, 32)
Distance	6.0 (3.0, 8.0)	7.0 (4.0, 10.0)
RoomNumber		
1	153 (54%)	119 (43%)
2	82 (29%)	59 (21%)
3	48 (17%)	88 (32%)
4	0 (0%)	12 (4.3%)

[1] Median (Q1, Q3); n (%)

- 「中央値(50%) (下位 25%, 上位 25%)」を表示

1.4 可能性

- 練馬区の方が、駅から遠く、築年数が古く、狭めの部屋が取引されている
  - 地区に関わらず、取引価格が低い傾向
- 同じような物件で比べれば、港区との格差は縮まるのではないか

1.5 R の例: 単純比較

```
estimatr::lm_robust(Price ~ District, data)
```

```
                Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper
(Intercept)    123.21555   7.854036  15.68818 3.143109e-46  107.7885 138.64258
District練馬区 -85.05972   7.934338 -10.72046 1.631327e-24 -100.6445 -69.47496
                DF
(Intercept)    559
District練馬区 559
```

1.6 R の例: Size, Tenure, Distance をバランス

- OLS を用いたバランスが可能

```
estimatr::lm_robust(Price ~ District + Size + Tenure + Distance, data)
```

```
                 Estimate Std. Error    t value     Pr(>|t|)   CI Lower
(Intercept)     37.921304  6.1745184   6.141581 1.556690e-09  25.793070
District練馬区 -63.875987  4.4707171 -14.287638 1.136352e-39 -72.657548
Size              2.698450  0.2130456  12.666069 1.754278e-32   2.279977
Tenure           -1.841204  0.2164363  -8.506909 1.660098e-16  -2.266336
Distance         -3.180468  0.7897796  -4.027032 6.433225e-05  -4.731784
                 CI Upper  DF
(Intercept)     50.049539 556
District練馬区 -55.094427 556
Size              3.116922 556
Tenure           -1.416071 556
Distance         -1.629151 556
```

- Size, Tenure, Distance を”バランス”すると、練馬/港区間格差が 2118 万円縮まる

1.7 What if 分析

- 広義には What if 分析 (もし〜ならば、どうなるか?) の一部
  - **もし**部屋の広さや築年数、駅からの距離の分布に差がなければ、練馬/港区の価格格差はどうなるか？

1.8 実際の例

- 合計特殊出生率
  - 出生率の国比較や時系列比較に理由される
    - 年齢構造も異なる
- 合計特殊出生率 = 年齢のバランス

$$= \frac{\text{15歳の女性が一年間で生んだ子供の数}}{\text{15歳の女性人口}} +$$

$$.. + \frac{\text{49歳の女性が一年間で生んだ子供の数}}{\text{49歳の女性人口}}$$

1.9 Takeaway

- 実務で伝統的に用いられてきた方法は、X の数が多い場合には適用不可能
- 重回帰を用いた方法は、ある程度対応可能だが、X の数が非常に多くなると対応不可能

- LASSO は候補になるが、信頼区間の計算困難であり、非実用的

- 次のスライドで、Post-double LASSO を紹介

2 サブグループ法

2.1 推定対象

- $Y =$ Price の平均値の $D =$ 地区間での差
  - ただし、$X =$ 物件の属性の差を無視できるように調整(バランス)
- 実務で最も用いられてきた方法は、サブグループ分析

2.2 特定グループの比較

- 最も単純な方法は、X 同じサブグループ内での比較
- 例: $X =$ RoomNumber のみがバランスの対象

```
estimatr::lm_robust(
  Price ~ District,
  data,
  subset = RoomNumber == 3) # 部屋数3
```

```
                Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper
(Intercept)     266.8333   27.36748  9.750014 2.649876e-17  212.7052  320.9614
District練馬区 -216.1174   27.42044 -7.881619 9.946072e-13 -270.3503 -161.8846
                DF
(Intercept)    134
District練馬区 134
```

2.3 実務での例

- 既存店における前年同月
  - イオン・グループ

2.4 サブグループ分析

- X の組み合わせごとに、事例をサブグループに分割し、平均取引価格を比較
- 例: $X =$ 部屋数

```
# A tibble: 7 × 4
  平均値 District `X=RoomNumber` 事例数
   <dbl> <chr>              <dbl>  <int>
1   54.3 港区                   1    153
2   25.6 練馬区                 1    119
3  168.  港区                   2     82
```

```
4   38.4 練馬区                2     59
5  267.  港区                  3     48
6   50.7 練馬区                3     88
7   69.4 練馬区                4     12
```

2.5 サブグループ分析

```
# A tibble: 4 × 4
  `X=RoomNumber` 事例数_港区 事例数_練馬区 平均差
           <dbl>       <int>         <int>  <dbl>
1              1         153           119   28.7
2              2          82            59  129.
3              3          48            88  216.
4              4          NA            12   NA
```

- 注: 単純な平均差は、 85.1
- 部屋数が増えるにつれて、平均差は増加傾向にある
  - 部屋数 4 については、このデータでは、比較不可能

2.6 サブグループ分析の限界

- X の組み合わせが増えると、
  - サブグループの事例数が減る
    - 推定の精度が悪化
    - 練馬/港区のどちらかしかないグループでは、比較不可能
  - 大量の平均差が計算され、人間が認識できなくなる
- 追加的な仮定のもとで、より単純なモデルの推定を行う方が現実的

2.7 例

- $X =$ RoomNumber, Tenure, Distance
- 例

```
# A tibble: 387 × 6
   RoomNumber Tenure Distance 事例数_港区 事例数_練馬区 平均差
        <dbl>  <dbl>    <dbl>       <int>         <int>  <dbl>
 1          1      1        3           3            NA     NA
 2          1      1        5          NA             5     NA
 3          1      1        6           1            NA     NA
 4          1      1        7           1             1     15
 5          1      1        8          NA             2     NA
 6          1      1       12           1            NA     NA
```

```
 7          1      2        9          NA             1     NA
 8          1      2       11           1            NA     NA
 9          1      2       14          NA             1     NA
10          1      3        8           1             1     64
# ℹ 377 more rows
```

3 同質性の仮定のもとでの推定

3.1 OLS による調整

```
estimatr::lm_robust(Price ~ District + RoomNumber, data)
```

```
                Estimate Std. Error    t value     Pr(>|t|)   CI Lower
(Intercept)      41.96570   6.006662   6.986526 8.056762e-12   30.16727
District練馬区 -102.30973   8.343091 -12.262809 8.899901e-31 -118.69744
RoomNumber       49.87789   5.174809   9.638596 1.934738e-20   39.71340
                CI Upper  DF
(Intercept)     53.76413 558
District練馬区 -85.92203 558
RoomNumber      60.04237 558
```

- $\beta_D = -102.3$ は、RoomNumber をバランスさせた後の比較結果と見做せるか？

3.2 妥当な定式化

- 十分な事例数 + ランダムサンプルング + “妥当な定式化” であれば、OK
- 妥当な定式化
  - 適当なβ_0, β_1 を選べば、

    $$E[Y \mid D, X] = \beta_0 + \beta_D \underbrace{District}_{D} + \beta_1 \underbrace{RoomNumber}_{X}$$

  - 妥当ではない = 誤定式化

3.3 妥当な定式化の前提

- $\beta_0 + \beta_D D + \beta_1 X + ..$

  $$= \underbrace{\beta_D}_{\text{推定対象}} \times D + \underbrace{\beta_0 + \beta_1 X}_{\text{局外 } (Nuisance)}$$

- 推定対象が一定: どんな d, X についても

  $$E[Y \mid D = d + 1, X] - E[Y \mid D = d, X] = \beta_D$$

- Nuisance が十分に複雑に定式化されている

3.4 例 $X =$ RoomNumber

- 二乗項まで導入

```
estimatr::lm_robust(
  Price ~ District + RoomNumber + I(RoomNumber^2),
  data)
```

```
                Estimate Std. Error    t value     Pr(>|t|)   CI Lower
(Intercept)     -30.21843  13.827749  -2.185347 2.927920e-02  -57.37934
District練馬区  -98.68988   8.217596 -12.009580 1.041826e-29 -114.83115
RoomNumber      135.62844  18.090127   7.497373 2.580539e-13  100.09523
I(RoomNumber^2) -20.92321   4.273036  -4.896568 1.279563e-06  -29.31645
                 CI Upper  DF
(Intercept)      -3.057526 557
District練馬区  -82.548612 557
RoomNumber      171.161646 557
I(RoomNumber^2) -12.529978 557
```

3.5 例 $X =$ たくさん

- $X =$ RoomNumber,Tenure,Distance,Size
  - 二乗と交差項を導入

```
estimatr::lm_robust(
  Price ~ District + (RoomNumber + Tenure + Distance + Size)^2 +
    I(RoomNumber^2) + I(Tenure^2) + I(Distance^2) + I(Size^2),
  data)
```

```
                       Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)          36.97051755 14.218380043  2.6001920 9.569917e-03
District練馬区      -31.94108957  4.306790608 -7.4164482 4.634298e-13
RoomNumber           48.45086983 15.577445412  3.1103219 1.966587e-03
Tenure               -1.96845086  0.521329698 -3.7758272 1.770032e-04
Distance             -2.20845269  1.190711841 -1.8547331 6.417423e-02
Size                 -0.50733896  0.808251130 -0.6276997 5.304636e-01
I(RoomNumber^2)      -9.60540818  5.114857469 -1.8779425 6.092275e-02
I(Tenure^2)           0.05277823  0.009188361  5.7440304 1.538156e-08
I(Distance^2)         0.18871909  0.103970844  1.8151155 7.005568e-02
I(Size^2)             0.07892157  0.009063450  8.7076734 3.705147e-17
RoomNumber:Tenure     0.89762107  0.298417856  3.0079335 2.751756e-03
RoomNumber:Distance   2.28827531  0.955769680  2.3941702 1.699497e-02
RoomNumber:Size      -1.10937993  0.369990714 -2.9983994 2.837865e-03
Tenure:Distance       0.07866583  0.040301612  1.9519276 5.145854e-02
Tenure:Size          -0.09108435  0.016490309 -5.5235079 5.152146e-08
Distance:Size        -0.16207740  0.061744873 -2.6249532 8.909327e-03
```

```
                          CI Lower     CI Upper  DF
(Intercept)           9.040980e+00  64.90005533 545
District練馬区       -4.040103e+01 -23.48114754 545
RoomNumber            1.785168e+01  79.05005545 545
Tenure               -2.992512e+00  -0.94438923 545
Distance             -4.547399e+00   0.13049388 545
Size                 -2.095008e+00   1.08032998 545
I(RoomNumber^2)      -1.965266e+01   0.44184084 545
I(Tenure^2)           3.472929e-02   0.07082717 545
I(Distance^2)        -1.551357e-02   0.39295175 545
I(Size^2)             6.111799e-02   0.09672514 545
RoomNumber:Tenure     3.114310e-01   1.48381111 545
RoomNumber:Distance   4.108318e-01   4.16571882 545
RoomNumber:Size      -1.836162e+00  -0.38259744 545
Tenure:Distance      -4.996867e-04   0.15783135 545
Tenure:Size          -1.234767e-01  -0.05869200 545
Distance:Size        -2.833645e-01  -0.04079032 545
```

3.6 Takeaway

- OLS を用いても、バランス後の比較は可能
- 本スライドでは、平均差の同質性を仮定し、X に関する部分 (nuisance)を複雑に定式化するアプローチを紹介
- 結果
  - 単純比較: 港区の方が 8506 万円 (信頼区間 = [6947 万円, 10065 万円]) 程度平均的に高い
  - RoomNumber,Tenure,Distance,Size をバランス: 港区の方が 3194 万円 (信頼区間 = [2348 万円, 4040 万円]) 程度平均的に高い

3.7 本スライドのアプローチの問題

- OLS は、複雑なモデルの推定に向かいない
  - β の数が、事例数の 1/3 を超えるほど大きくなると、推定精度の大幅な悪化、信頼区間が信頼できなくなる
- 同質性の仮定が怪しい
  - データ上、部屋数に応じて、平均差が大きく異なっている
  - 正当化するには、「データ上では”偶然の上振れ/下振れによって、異質に見えるだけ」と強弁するしかないが 🤔

3.8 Reference

Bibliography