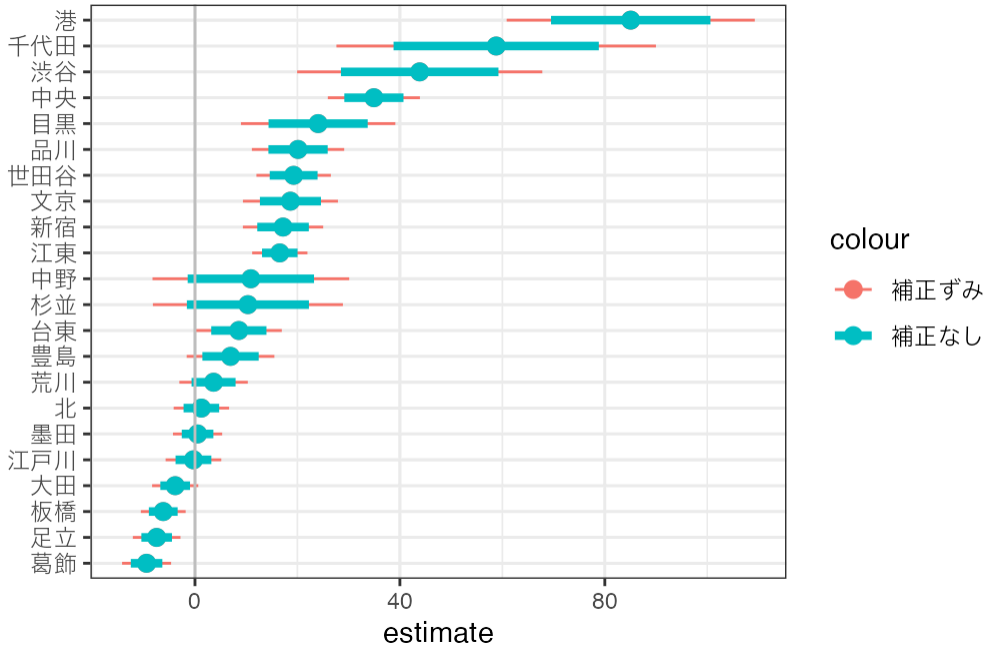
バランス後の比較
機械学習

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

2025-10-21

1 分析例

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る傾向
  - 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

- 企業によるデータ分析で、頻繁に用いられる
- 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: $X =$ [Tenure, Distance]

```
# A tibble: 295 × 5
   Tenure Distance 事例数_港区 事例数_練馬区 平均差
    <dbl>    <dbl>       <int>         <int>  <dbl>
 1      1        2           1            NA     NA
 2      1        3           7            NA     NA
 3      1        5          NA             5     NA
 4      1        6           2            NA     NA
```

```
 5      1        7          1            1     15
 6      1        8         NA            2     NA
 7      1       10          1           NA     NA
 8      1       12          2           NA     NA
 9      2        1          1           NA     NA
10      2        2          2           NA     NA
# ℹ 285 more rows
```

3 OLS を用いた調整

3.1 よくある例

Table 1

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

```
                 Estimate Std. Error     t value     Pr(>|t|)   CI Lower
(Intercept)    157.806101 14.0949558  11.1959273 2.204229e-26 130.120336
District練馬区 -81.987135  7.4424340 -11.0161722 1.144778e-25 -96.605803
Tenure          -1.688495  0.2974546  -5.6764801 2.214360e-08  -2.272765
Distance        -0.158779  0.9753437  -0.1627928 8.707406e-01  -2.074580
                 CI Upper  DF
(Intercept)    185.491866 557
District練馬区 -67.368467 557
Tenure          -1.104225 557
Distance         1.757023 557
```

- $\beta_D = -81.99$ は、Tenure/Distance 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、β_D はバランス後の比較結果と一致する
- 妥当な定式化
  - 適当なβ_0, β_1 を選べば、

$$E[Y \mid D, X] = \beta_0 + \beta_D \times \underbrace{District}_{D} + \beta_1 \times \underbrace{Tenure}_{X_1}$$

$$+\beta_2 \times \underbrace{Distance}_{X_2}$$

3.3 実践上の問題

- シンプルな定式化は、“妥当な定式化”ではない可能性を疑うべき

- より複雑な定式化を用いれば、妥当な定式化である可能性が高まる
  - 過剰適合が生じ、推定精度が悪化する
    - 信頼区間自体も”信頼できなくなる”

3.4 定式化の分解

- $E[Y \mid D, X] = \beta_0 + \beta_D \times District + \beta_1 \times Tenure$

$$
\begin{aligned}
&+\beta_2 \times Distance \\
= \quad &\underbrace{\beta_D \times District}_{D\text{に関する部分}(Interest)} \\
&\underbrace{+\beta_0 + \beta_1 \times Tenure + \beta_2 \times Distance}_{X\text{に関する部分}(Nuisance)}
\end{aligned}
$$

3.5 妥当な定式化の前提

- “Interest”が十分に複雑に定式化されている
- Table 1 では、「$E[Y \mid D = 1, X] - E[Y \mid D = 0, X] = \beta_D$ は一定」を仮定
  - 現実にはおそらく異なる
    - 広めの部屋の方が、港区-練馬区間の格差が大きい
  - 次回以降に対応策を議論

3.6 妥当な定式化の前提

- $E[Y \mid D = 1, X] - E[Y \mid D = 0, X] \underbrace{\approx}_{\text{ほぼ一定}} \beta_D$ であれば、
- “nuisance”が十分に複雑に定式化されていれば OK

3.7 例: 二乗と交差項の導入

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

```
                    Estimate  Std. Error     t value     Pr(>|t|)      CI
Lower
(Intercept)     204.65833135 36.06561682   5.6746106 2.242903e-08
133.816253292
District練馬区  -83.44966546  7.69192402 -10.8489976 5.359009e-25 -98.558567789
Tenure           -5.05071521  1.72158597  -2.9337572 3.487524e-03
```

```
-8.432349524
Distance         -5.05103264  4.84203891  -1.0431623 2.973281e-01
-14.562033111
I(Tenure^2)       0.04586426  0.02240533   2.0470245 4.112691e-02
0.001854471
I(Distance^2)     0.09347738  0.15968225   0.5853962 5.585197e-01
-0.220179312
Tenure:Distance   0.16577952  0.10777625   1.5381823 1.245750e-01
-0.045920551
                     CI Upper  DF
(Intercept)     275.50040940 554
District練馬区  -68.34076314 554
Tenure           -1.66908090 554
Distance          4.45996783 554
I(Tenure^2)       0.08987405 554
I(Distance^2)     0.40713408 554
Tenure:Distance   0.37747959 554
```

3.8 実践上の問題

- X の数が多い場合、二乗項や交差項を導入すると、変数の数が爆発的に増える
  - $X =$ [Size,Tenure,Distance,RoomNumber] について、二乗項や交差項を作成すると 15 個の β を推定する必要がある
- 過剰適合が生じ、推定誤差の拡大や信頼区間が信頼できなくなる等の問題が生じる

3.9 LASSO?

- OLS に比べて、LASSO は複雑なモデルの推定に”適している”
  - 推定精度を改善できる
- 基本的に、信頼区間がうまく計算できない
  - データと推定結果の関係性が、OLS に比べて”複雑であり”、どう計算すれば良いか、現状でも議論が続いている

3.10 Takeaway

- 目標:「母集団におけるバランス後の比較結果」について、**概ね正しい結論**を得たい
- OLS を用いた暫定的結果は、
  - 港区と練馬区の取引価格の平均格差は、6947 ～ 1 億 65 万円程度
  - [Tenure, Distance]をバランスした場合、6834 ～ 9856 万円程度
  - [Tenure, Distance,Size,RoomNumber]をバランスした場合、404 ～ 2348 万円程度

3.11 Takeaway: Positive

- バランスさせたい属性 X が少数であれば、OLS は有力な選択肢
  - 仮定: $E[Y \mid D = 1, X] - E[Y \mid D = 0, X] = \beta_D$ も緩めることができる
  - ただし伝統に用いられてきた定式化は単純すぎることが多いので、二乗項や交差項を導入すべき
- LASSO は信頼区間の計算が難しいという問題を持つ

3.12 Takeaway: Negative

- バランスさせたい属性 X が多い場合、OLS も LASSO も問題がある
- OLS は、複雑なモデルの推定に向かいない
- LASSO は、引き続き信頼区間の計算が難しい

3.13 Reference

Bibliography