OLS/LASSO を活用した近似的バランス

労働経済学 2

川田恵介

Table of contents

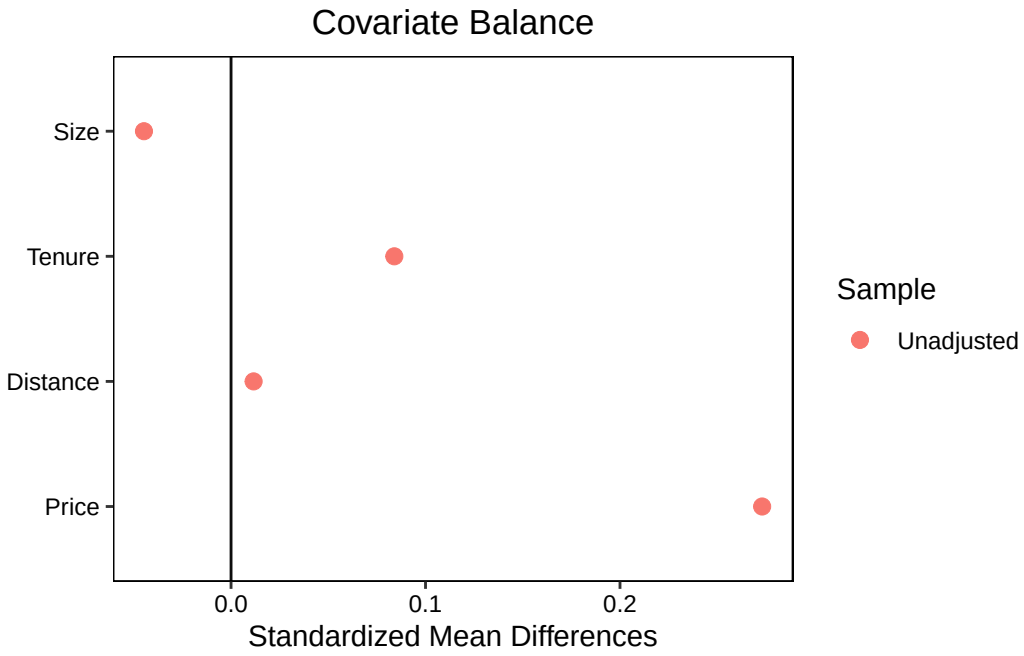
1 近似的バランス

- 事例数が限られる多くの応用において、X の完璧なバランスは不可能
  - 事例数が限られる一方で、バランスさせたい X が多いため
- 近似的なバランスを目指す
  - 多くの発展的手法 (含む機械学習の応用) は、近似的バランスの一つの手法であると解釈できる

1.1 近似的なバランスを達成する手法

- X 分布全体をバランス
  - 代表例: 傾向スコアを活用した Inverse Probability Weight
- X 分布の特徴 (平均や分散、共分散) をバランス
  - 代表例: OLS
    - 発展として Post-Double-LASSO
- 事例数に比べて、X の組み合わせが多い (連続変数が含まれている/X の種類が多い) 場合でも活用できる

1.2 例: バランス



1.3 例: 事例数

- $Size$の種類 $\times$ $Distance$の種類 $\times$ $Tenure$の種類 $= 9660$
- 事例数 5577 を超える

1.4 定義: 平均値のバランス

- X の平均値を $D = 1/0$ で均質化する
  - 一般に平均値を Balance する Weight は無数に存在する

1.5 OLS による平均値のバランス

- $Y \sim D + \underbrace{X_1 + ..X_L}_{\text{ダミーである必要はない}}$ を OLS 推定することで計算される D の係数値は、
  - $X_1, .., X_L$ の平均値をバランスさせた「バランス後の比較」と解釈できる!!!

1.6 例: Size のバランス

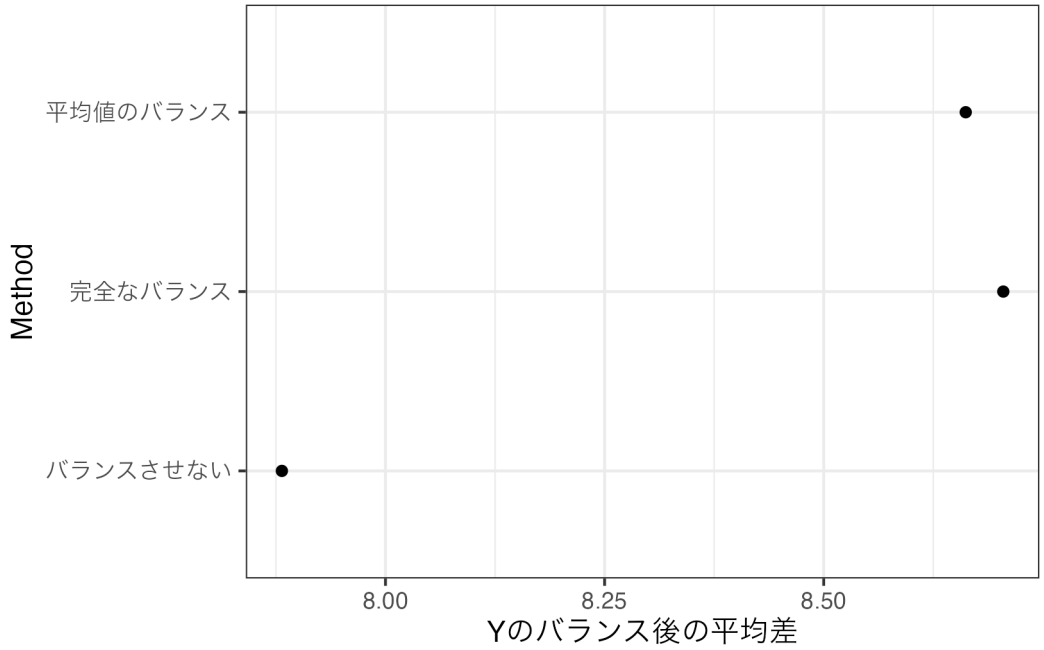
Y の平均値	D	Size	N	Size の割合
45.9	0	55	216	0.364
54.1	1	55	252	0.338
62.7	0	75	301	0.507
71.7	1	75	414	0.555
84.0	0	90	77	0.130
101.9	1	90	80	0.107

1.7 例: OLS の結果

Y の平均値	D	Size	N	Size の割合	OLS	完璧なバランス
45.884	0	55	216	0.364	0.361	0.353
54.075	1	55	252	0.338	0.341	0.353
62.688	0	75	301	0.507	0.508	0.528
71.725	1	75	414	0.555	0.553	0.528
84	0	90	77	0.13	0.131	0.119
101.887	1	90	80	0.107	0.106	0.119

- $$\beta_D = 0.341 \times 54.075 + 0.553 \times 71.725 + 0.106 \times 101.887$$
$$-0.361 \times 45.884 + 0.508 \times 62.688 + 0.13 \times 84$$

1.8 例: OLS VS 完璧なバランス



1.9 まとめ: 完璧 VS OLS による近似的バランス

- OLS を用いても、X の分布の完全なバランスは達成できない
- 平均値はバランス: $D = 0/1$ の平均 Size $\simeq 69.76$
- 元々は $D = 1$ については 69.85 、$D = 0$ については 69.67

2 OLS の性質

- 複数の変数についても、平均値をバランスできる
- 「母集団で仮想的に実行した」バランス後の比較の優れた推定値となる

2.1 OLS の利点: 複数の X のバランス

- Size, Distance, Tenure, District, 全ての平均値をバランスさせる
- ダミー化した変数については、分布がバランスする
  - District については、分布がバランスする

        * 23 区の立地割合が、2021 年と 2023 年で同じになるように調整される

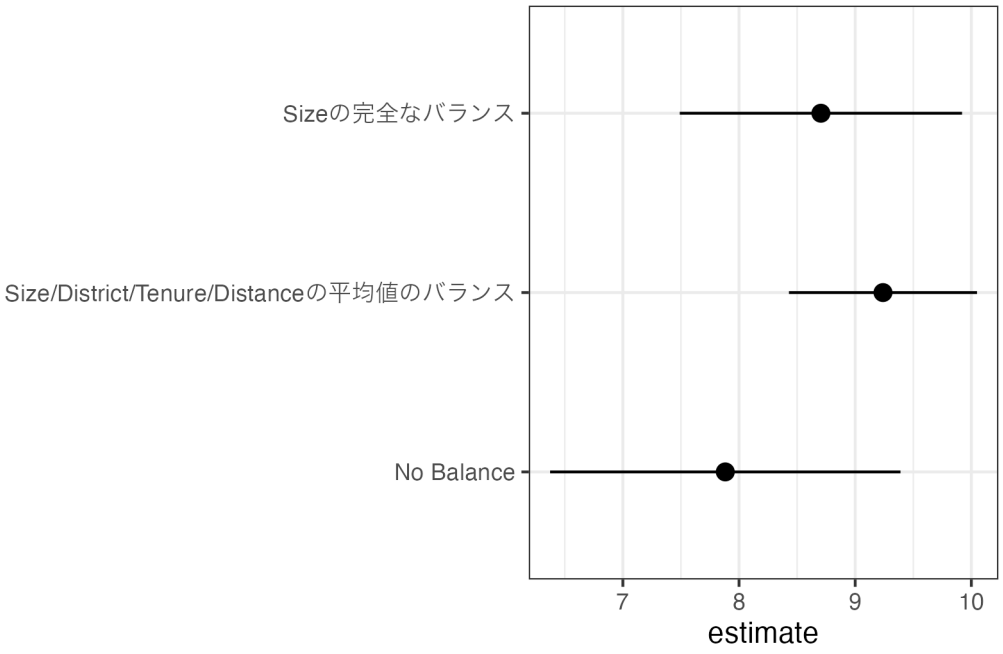
2.2 OLS の利点

- 本来やりたい比較 (Size,Distance,Tenure,District のバランス) に到達するためには、以下の二つの手法が考えられる
  - 一部の変数 (Size) に絞って、完璧にバランスさせる
  - すべての変数と OLS を用いて、平均値のみバランスさせる
- 一般に後者の方が、本来やりたい比較に近づける場合が多いと考えられる

2.3 母集団への含意

- データ上での OLS の推定結果は、「Population における OLS による平均値のバランス後の比較」の優れた推定値とみなせる
  - OLS は母集団における OLS の優れた推定値であり、信頼区間も計算できる
    * 事例数が、(組み合わせではなく) X の数に比べて、十分に多いことが前提

2.4 例: バランス後の平均差



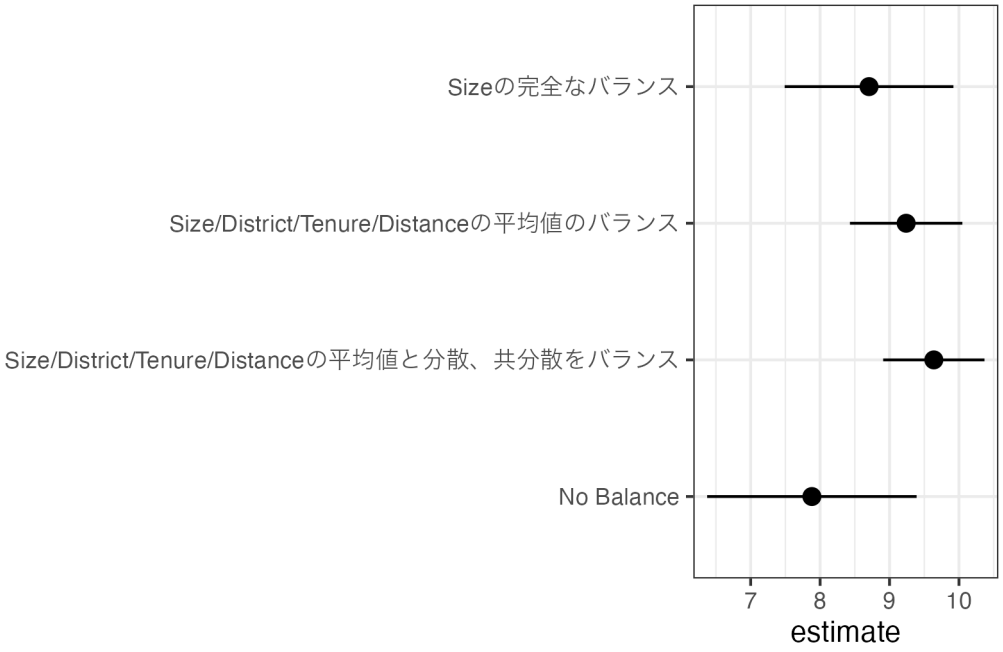
2.5 さらなるバランス

- “平均値” のみならず分散などもバランスできる
- $Y \sim D + X + X^2$ を推定すれば、X の平均値と分散もバランス
- $Y \sim D + X_1 + X_2 + X_1^2 + X_2 + X1 * X2$ を推定すれば、X_1, X_2 の平均値と分散、共分散もバランス
- 増やしすぎると、推定誤差が大きくなる

2.6 補論: 分散/共分散

- 分散: X_1 のばらつきを捉える指標
  - $(X_1 - X_1$の平均値$)^2$ の平均値
- 共分散: X_1 と X_2 の相関関係を捉える指標
  - $(X_1 - X_1$の平均値$) \times (X_2 - X_2$の平均値$)$ の平均値

2.7 例: バランス後の平均差



2.8 OLS の解釈

- 平均値をバランスさせる方法は、無数に存在する
  - OLS はその中の一つ
- $Y = \beta_0 + \beta_D D + \beta_1 X_1 + .. + \beta_L X_L + u$ を OLS で推定した β_D は、以下の性質を満たしたバランス後の比較と一致
  - $X_1..X_L$ の平均値をバランス
  - 推定の精度が高く、信頼区間が狭くなる傾向
  - 問題のある性質も持つ (後述)

2.9 母集団におけるバランス後の比較

- 平均値のバランスではなく、「母集団における分布のバランス後の比較」の推定値とみなすには、追加の仮定が必要
  - 母集団において overlap が成り立っている
    - すべての X の組み合わせについて、$D = 0/1$ どちらの事例も、**母集団においては**、存在する
  - $Y \sim D + X_1 + .. + X_L$ の母集団における OLS が「分布のバランス後の比較」と一致するためには、母平均が X の平均値のみに依存している

2.10 実践

- 一般に分布をどこまでバランスさせれば十分なのか、よくわからない
- 変数選択を活用しつつ、X の二乗項と交差項までをバランスさせるのが、現状の私的おすすめ
- 元々の X が多い場合、機械学習による変数選択の併用が有効 (次章)

3 OLS の問題点

- モデルに投入した全ての X を、“強引” バランスさせてしまう
  - Y との関係性/Overlap しているかしていないかも無視指定しまう

3.1 例: “昔の出席簿” データ

ID	Gender	TestScore
1	男性	60
2	男性	60
3	男性	60
4	女性	60
5	女性	60
6	女性	100

- 女性の平均点の方が高い
- 出欠番号は、“男性” からあいうえお順
  - “成績と関係ない”
  - 男女間で” バランスさせよう” がない

3.2 例: “昔の出席簿” データ

```
lm(TestScore ~ Gender + ID,
   data = Temp)
```

```
Call:
lm(formula = TestScore ~ Gender + ID, data = Temp)

Coefficients:
(Intercept)    Gender男性           ID
      23.33        16.67        10.00
```

- ID をバランスさせると「男性の方が平均的が高くなる」
  - どうやってバランスさせているのか？

3.3 例: 問題点

ID	Gender	TestScore	Target
1	男性	60	-0.4166667
2	男性	60	0.3333333
3	男性	60	1.0833333
4	女性	60	1.0833333
5	女性	60	0.3333333

6	女性	100	-0.4166667

- マイナスの割合を目標にする (???) ことで、平均値を 1.61 に「バランス」させている

3.4 例: 問題点

ID	Gender	TestScore
1	男性	10
2	男性	60
3	男性	60
4	女性	60
5	女性	60
6	女性	100

- 一番 (男性) の成績が悪かったとする

3.5 例: 問題点

- 出席番号をバランスさせると、男性の成績が悪くなっているのに、男性の平均点が女性よりもさらに高くなる (!!?)

```
lm(TestScore ~ Gender + ID,
   data = Temp)
```

```
Call:
lm(formula = TestScore ~ Gender + ID, data = Temp)

Coefficients:
(Intercept)   Gender男性           ID
     -39.17        37.50        22.50
```

3.6 実践

- 研究者が背景知識を用いて、推定対象や X を注意深く選ぶ
  - 母集団において、Overlap が成り立つように、分析事例を限定
  - Y と関係ない X の排除
- 後者については、機械学習を活用したデータ主導の変数選択も補完的に用いることができる

4 Double Selection

- X の中から **重要な変数** を選ぶ
  - LASSO を使用
  - 予測とは、機械学習を用いたとしても一定確率でミスを犯すこと、**重要性の基準**が異なることに注意
    - “AI によるダブルチェック” を実施させる

4.1 問題設定

- 労働研究では、バランスさせたい X が大量に存在するケースも多い
- 例: $Y =$ 年収、$D =$ 性別、$X =$ ついている仕事の特徴
  - 同じ働き方をしている男女内賃金格差
  - $X =$ 労働時間、勤続年数、業務内容、それらの交差項...
    - 全てを Balance させることができない/推定精度が大幅に悪化する

4.2 アイディア

- X 全てが” 重要” なわけではないかもしれない
  - X の一部 Z のみをバランスさせれば十分
- 仮定: (Approximately) sparsity: 事例数に比べて、十分に少ない変数数 $<<$ 事例数で、母平均をうまく近似できる
  - もともとのモデルには、“trivial” な変数も含まれていると仮定

4.3 LASSO の活用

- Y **を予測するために**「重要ではない」X を削除する
- Naive なアイディア: X を全てバランスさせるのではなく、LASSO が Y を予測するために選んだ変数のみをバランスさせる

4.4 問題点

- 問題点: 限られた事例数のもとで、LASSO による変数選択は、Y とそこそこ相関がある変数も除外されてしまう可能性がある
  - Y の近似が目的であれば、(Hyper parmeter が正しく選ばれている限り)、許容できる
- D との相関が強い (分布の分断が激しい) な変数が除外されると β_D の推定結果が大きな影響を受ける
  - バランス後の比較という目標に対して、モデルが過度に単純化される (Regulization bias)

4.5 例: Naive なアイディアが機能しやすいケース

- Y, D, X が以下のような関係性があるとすると
  - Yの平均値 $= D + \times X$
  - Dの平均値 $= 0.1 \times X$
- Y と X の関係性が強いので、X は除外されにくい

4.6 例: 機能しにくいケース

- Y, D, X が以下のような関係性があるとすると
  - Yの平均値 $= D + 0.1 \times X$
  - Dの平均値 $= X$
- Y と X の関係性が弱いので、X は除外されやすい
  - D との関係性は強いので、除外すると母集団におけるバランス後の比較から大きく乖離してしまう

4.7 Double Selection

1. Y および D を予測するモデルを、LASSO で推定し、選択された変数を記録
2. **どちらかの**予測モデルで選択された変数 (Z) のみを用いて、$Y \sim D + Z$ を回帰

- 重要な変数を誤って除外しないように、Yの予測モデルと D の予測モデルに” ダブルチェック” を行わせている
  - 今後の機械学習の活用における基本アイディア

4.8 性質

- Approximate sparcity が成り立っていると、緩やかな理論的家庭のもとで、信頼区間を近似計算できる
  - ある程度の事例数が前提

4.9 まとめ

- 「限られた事例数のもとで、多くの X をバランスさせる」を達成するために、適切な近似を重ねていく
  - OLS: 平均値や分散などのみをバランスさせることで、近似的なバランスを達成する
    - 変数が多いと、これでも多すぎるので、さらに近似する
      - Double Selection: 重要な条件のみをバランスする