Double/Debiased Machine Learning

機械学習入門

川田恵介

Table of contents

Y の平均値	D	Size	N	Target	データ上の割合
62.7	0	75	301	0.82	0.796
71.7	1	75	414	0.82	0.838
84.0	0	90	77	0.18	0.204
101.9	1	90	80	0.18	0.162

1 バランス後の比較: 復習

- 因果推論/格差/比較研究における中心的な手法

1.1 目標: バランス後の比較

- X の分布についての差を補正した後の、平均値の比較
- $\theta(1) - \theta(0)$
- $$\theta(d) = \underbrace{E[Y|d,X] \times Target(X)のXについての和}_{\int E[Y|d,X] \times Target(X) dX}$$
- $Target(X)$: 研究者が指定する集計用荷重

1.2 例: Size のバランス

1.3 OLS

- $Y \sim D + Size + Tenure$ を OLS で推定すると、
- $\beta_D =$
  - Size, Tenure の平均値を D 間でバランスさせた上での Y の平均値の比較結果

1.4 OLS

- $Y \sim D + Size + Tenure + Size^2 + Tenure^2 + Size \times Tenure$ を OLS で推定すると、
- $\beta_D =$

- Size, Tenure の平均値と分散、共分散を D 間でバランスさせた上での Y の平均値の比較結果

1.5 OLS

- $$Y \sim D + \underbrace{f(Size, Tenure)}_{\text{非常に複雑なモデル}}$$

  を OLS できれば、
- Size,Tenure の分布を完璧にバランスできる
- 問題: モデルを” 長くしすぎると”、推定精度が低下する (07MomentBalance 参照)

1.6 Post LASSO

- 重要な X をデータ主導で選択する
- 問題: 線型モデルが、平均値の優れたモデルとは限らない (08Aggregation 参照)

1.7 動機

- $$Y \sim D + \underbrace{f(Size, Tenure)}_{\text{非常に複雑なモデル}}$$

  をなんとか推定できないか?
- Stacking などを駆使することで、より高い精度の予測モデルを推定できる
  - 精度が高い $=$ より 平均値に近い
  - バランス後の比較に、予測モデルを直接活用できないか?
  - OLS の伝統的な別解釈が活用できる

2 OLS の別解釈

- FWL 定理

2.1 OLS/Post selection の別解釈

- $Y \sim D + X_1 + .. + X_L$ を OLS 推定して得られる D の係数値は、以下の手順で得られる推定値と同じ値になる

1. Y と D の予測モデル $\beta_0 + \beta_1 X_1 + .. + \beta_L X_L$ を OLS で推定する

2. Y の予測誤差 $Y - Y$の予測値 、D の予測誤差 $D - D$の予測値 を計算

3. Y の予測誤差と D の予測誤差で OLS 推定する

2.2 理想的な状況の再解釈

- Y または D と関係している変数を全て含めた複雑なモデルを十分な事例数で推定できれば、D の係数値 $= X$ の分布を完璧にバランスさせた後の平均差

2.3 直感

- バランスさせることで、平均差が変わる理由
  - D の間で X の分布がずれている: X と D が関係している
  - X が Y が関係している
- X と D, Y の関係性が問題

2.4 直感

- 十分に複雑なモデルを十分な事例数で推定できれば、X から Y/D への最善の予測モデルを得られる
  -「予測できない部分 $= X$ とは” 無関係” の部分」
- 予測できない部分同士の回帰 $= X$ とは無関係の部分同士の回帰
- 含意: Y と D の良い予測モデルを頑張って作れば良い
  - Random Forest, LASSO, Deep Learning, Stacking などが活用できる!!!

3 R learner

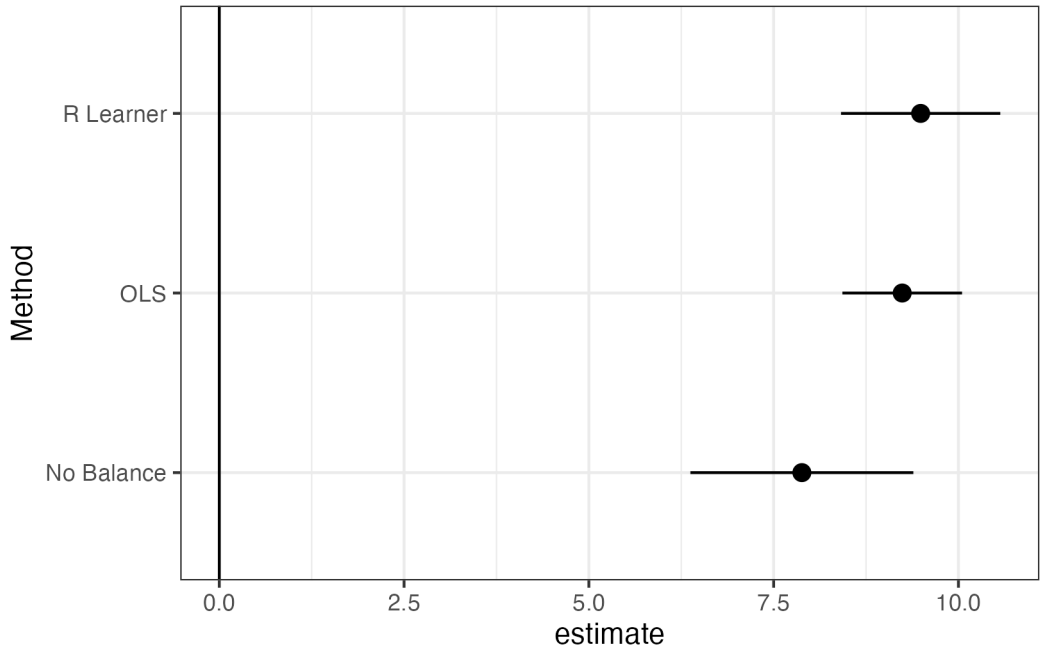
- 代表的な方法の一つ: Robinson/Residual learner
- OLS/Post Double Seletion の一般化

3.1 アルゴリズム

1. データを訓練とテストに分割
2. Y と D の予測モデルを、訓練データとなんらかの推定手法で推定する
3. Y の予測誤差 $Y - Y$の予測値 、D の予測誤差 $D - D$の予測値 をテストデータで計算

4. Y の予測誤差と D の予測誤差を、テストデータで OLS 推定する

3.2 例



4 R-learner の性質

- OLS や naive な機械学習の応用と比べて、優れた性質を持つ
  - 特に「信頼できる」信頼区間を計算できる

4.1 OLS の問題点

- X の” 平均値” のみをバランスするので、研究者の定式化に結果が強く依存
- 例: X の分散も Y に影響を与えるのに、モデルに入れてない場合、X の完全なバランスの結果と乖離してしまう

4.2 Naive な機械学習の応用

- 古典的なアイディア (Varian 2014):
  - $D = 1$ の事例のみを用いた Y の予測モデルと $D = 0$ のみを用いたモデルの予測値の差を集計する
  - 2 つのモデルを用いるため、T - Learner と呼ばれる

Price	D	Size	Tenure	Distance	PredY_D1	PredY_D0	Difference
53	0	40	6	4	50.79	49.79	1.00
27	0	25	15	2	29.54	24.93	4.61
49	0	50	8	5	68.76	50.90	17.86
71	0	75	16	3	68.20	56.29	11.90
22	0	30	41	1	25.70	23.00	2.70
47	0	60	40	4	43.47	34.39	9.07
170	0	65	7	3	74.42	87.94	-13.52
44	0	35	9	2	54.63	36.69	17.94
56	0	60	18	6	67.69	61.89	5.80
18	0	25	39	2	18.90	16.80	2.11

4.3 例: RandomForest の活用

```
library(ranger)

Data_D1 = Data[Data$D == 1,]
Data_D0 = Data[Data$D == 0,]

Model_D1 = ranger(Price ~ Size + Distance + Tenure, Data_D1)
Model_D0 = ranger(Price ~ Size + Distance + Tenure, Data_D0)

Data$PredY_D1 = predict(Model_D1,Data)$predictions
Data$PredY_D0 = predict(Model_D0,Data)$predictions
```

4.4 例

4.5 問題点

- Y の平均値のモデルとしての性能に強く依存
- OLS などの伝統的手法も機械学習的手法も、一般的には要求される水準を満たさない
  - OLS などの伝統的な手法は、研究者による定式化（の誤り）が推定結果に大きな悪影響を与える
  - Stacking などを用いたデータ主導のモデル化は、研究者による定式化への依存は減らせるが、データへの依存度が高くなる

4.6 問題点: 詳細

- 推定値が満たして欲しい性質の代表例
  - 無限大の事例数を持つデータを用いることができれば、母集団におけるバランス後の比較結果と一致する
    - 研究者による定式化に依存する OLS などでは、保証することは一般に難しい
  - ある程度の事例数があれば、信頼区間を近似計算できる
    - データへの依存が大きい機械学習を naive に活用すると、保証することは一般に難しい

4.7 R learner の特徴

- Post-selection を一般化した性質を持つ
  - Y または D の予測モデルのどちらかでも機能すれば、もう一方はそこそこでも良い
- 「AI によるダブルチェックが機能」

4.8 まとめ

- 多くのアプローチは、(1) 研究者によるモデル設定、(2) データの偏り、のどちらかの悪影響を強く受けてしまう
- 「予測モデルによるダブルチェック」を活用し、悪影響を低下させることができる
  - セミパラメトリック理論との融合
- 近年、盛んに研究されている
  - 継続学習したい人には、以下を推奨
  - CausalML

4.9 補論: カリブーレーション

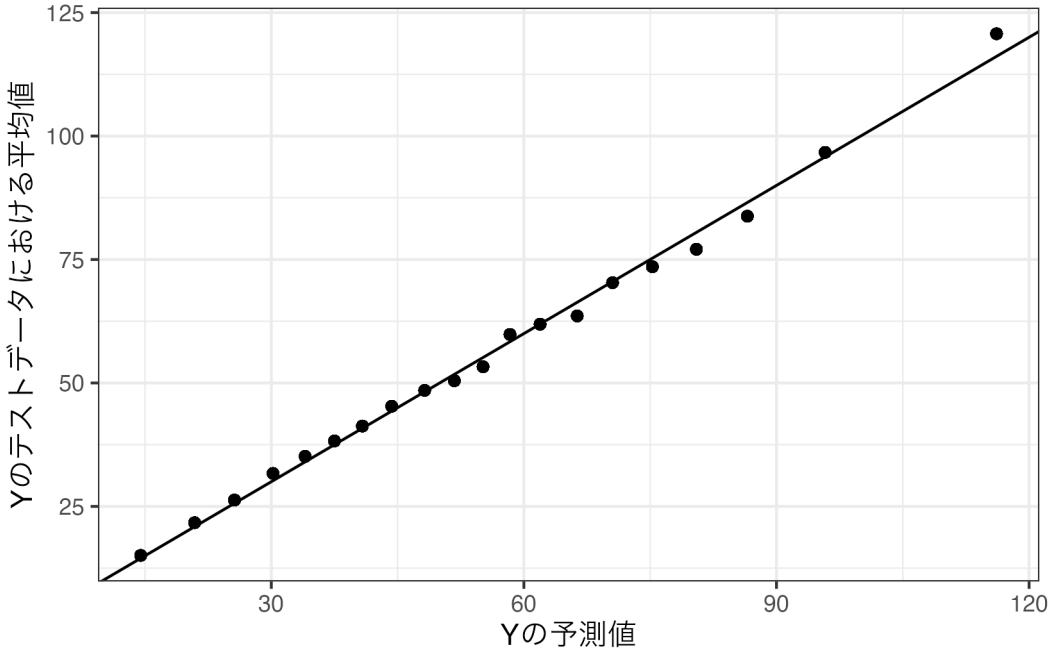
- テストデータにおいて、予測値と「平均値」の乖離を示す
  - 予測値と Y の乖離ではない (03Population)
- 予測値に応じて、テストデータを等分割する
  - 各サブグループごとに、Y と予測値の平均値を算出

Bin	平均価格	平均予想
[8.88,23.7]	19.19	17.74
(23.7,32]	30.91	27.75
(32,38.5]	38.40	35.27
(38.5,44.8]	45.10	41.65
(44.8,51.5]	50.42	48.12
(51.5,58.8]	58.35	55.34
(58.8,67]	67.05	62.79
(67,77.3]	77.01	71.85
(77.3,90.5]	86.25	83.23
(90.5,153]	117.81	105.48

- 理想的な予測モデルであれば、一致するはず

4.10 例: カリブレーション

4.11 補論: 予測誤差



Reference

Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28.