Post Double Selection

川田恵介

1 機械学習の活用

1.1 OLS の問題点

- 本来の推定目標 = X の分布を完璧にバランスさせた後に Y を比較する
- OLS のアプローチ
  - X の高次項や交差項などを導入した Population OLS が推定対象 ≃ 本来の推定目標
  - X が(事例数に比べて)少ない場合、データ上での OLS によって、高い精度で推定できる

1.2 OLS の問題点

- X の数が多い場合、Population OLS とデータ上の OLS の乖離が激しく、実用的ではない
- 本来の推定目標 ≃ 複雑な Population OLS
  - ≠ データ上の OLS

1.3 LASSO の性質

- LASSO の推定対象も、Population OLS
  - 無限大の事例数のもとでは、LASSO もXの平均値を完全にバランスする
- 限られた事例数でも、
- 本来の推定目標 ≃ 複雑な Population OLS
  - ≠ データ上の OLSX の数が多い場合の代替案?

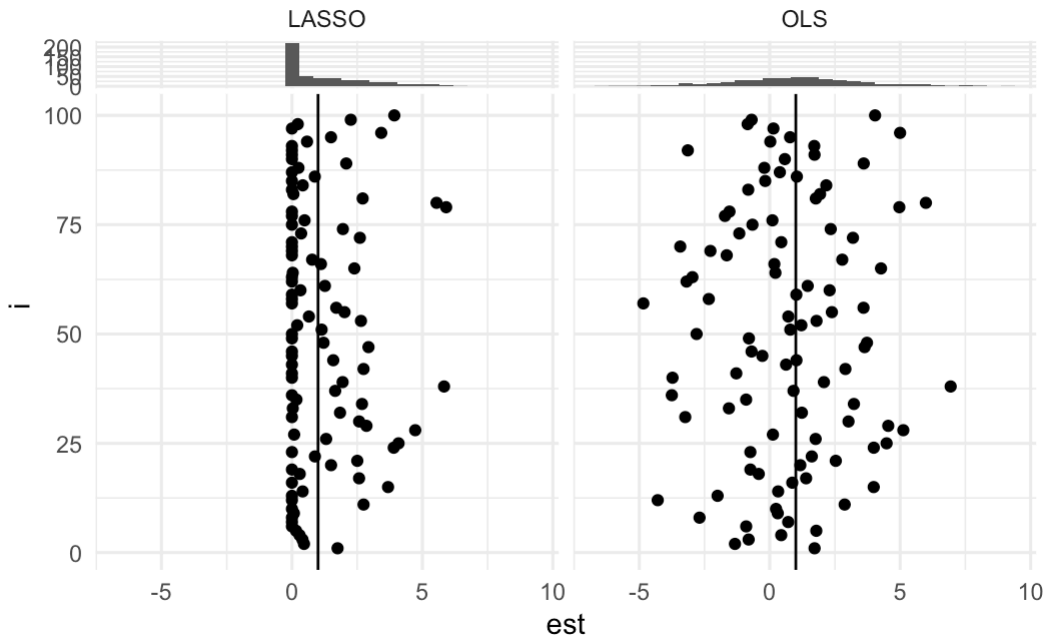
1.4 LASSO の問題点

- 推定対象 (Population OLS)について、Valid Inference (Section 3) が提供できない
  - (ブートストラップ法も含め)近似的な信頼区間導出ができない
    - 推定値の分布に対して、中心極限定理が適用できず、推定値の分布がバイアスの無い正規分布として近似できない
- “データが異なるため、結論が異なる"問題への対処が難しい

1.5 数値例

- $\{Y, D, X, Z_1, .., Z_{100}\}$
- $Y = 5 \times D + X + Z_1 + Z_2 + Z_3 + Z_4 + \underbrace{u}_{N(0,5)}$
- $X = 5 \times D + \underbrace{v}_{N(0,0.5)}$
  - $Z_l \sim [-1, 1]$ までの一様分布
  - $D \sim \{0, 1\}$ の一様分布

1.6 数値例: 500 事例



1.7 まとめ

- OLS も LASSO も、母集団に適用すると、Population OLS を算出する
  - 同じ推定目標
- LASSO は、推定目標について、Valid inference を提供するのが難しい
  - 機械学習一般についても同様
- この問題解決が、今後の議論の大きな目標

2 Post Selection

2.1 Post-selection OLS

1. LASSO などを活用し、X から”重要ではない変数”を除外する
2. 除外されなかったX $(= Z)$とDのみを用いて、重回帰 $Y \sim D + Z$を行う

- 機械学習を活用した”下準備”

2.2 例: Selection by $Y - X$

- OLS は、全てのXの平均値をバランスする
  - X の中には、Yと関係ない変数も含まれているかもしれない
    - 例: $Y =$ 成績、$X =$ 昔の出席番号、$D =$ 性別
- $Y \sim X$ を LASSO で推定し、重要ではない変数を除外

2.3 問題点

- 復習: データが異なるために推定結果が異なる
- Post-selection においては、問題が複雑化する
  - Step 2: 同じモデルを推定したとしても、データが異なるため、推定結果が異なる
    - 中心極限定理から、推定結果の分布を正規分布で近似することで、定量化できる
  - + Step 1: **変数選択: 除外される変数が異なる**

2.4 推定値の分布

- 事例数が十分に大きくなると、β_Dの推定値の分布 $=$

$$\underbrace{\underbrace{\textbf{変数選択の不確実性}}_{?} + \underbrace{OLS\textbf{の不確実性}}_{\Rightarrow \textbf{正規分布}}}_{?}$$

- 変数選択がβ_Dの分布に影響を与えてしまい、正規分布に収束しない

2.5 Double Selection (Belloni, Chernozhukov and Hansen, 2014)

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) 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”AI”とDの予測”AI”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

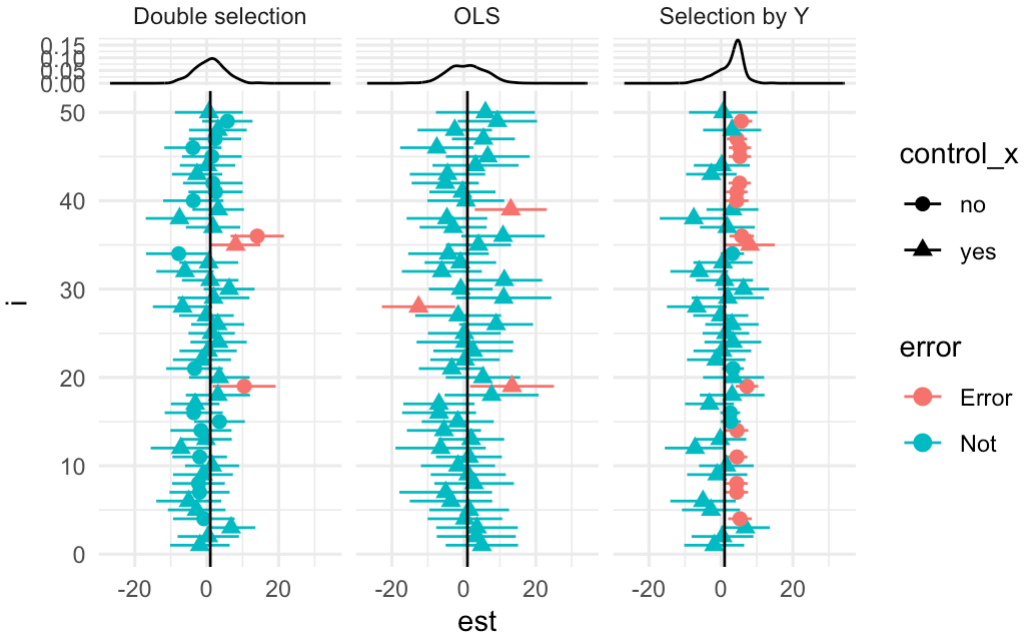
2.6 推定値の分布

- 仮定: (Approximately) sparsity: 事例数に比べて、十分に少ない変数数で、母平均をうまく近似できる
  - Xの中には、"trivial"な変数も含まれている

- 事例数が十分に大きくなると、β_Dの推定値の分布 $=$

$$\underbrace{\underbrace{\text{変数選択の不確実性}}_{\to 0} + \underbrace{OLSの不確実性}_{\Rightarrow 正規分布}}_{\Rightarrow 正規分布}$$

2.7 数値例: $N = 200$



2.8 直感

- $Y \sim X$ で変数選択すると、Y とそこそこ相関がある変数も除外される可能性がある
  - バランス後の比較のおいては、D との相関も重要
    - D 間で分布が大きく異なる変数ならば、バランス後の比較結果に大きな影響を与える
- $D \sim X$ での変数選択結果も活用することで、推定値への影響を減らす

2.9 まとめ

- 予測問題とバランス後の比較は、推定対象が本質的に異なるため、推定方法の活用法が異なる
- 予測問題の推定対象 $= E[Y \mid D, X]$
  - D と X の間の相関は、主要な関心ではない
- バランス後の比較の推定対象 $= X$ の分布をバランスさせた後の平均差

- Y との相関だけではなく、 D 間で大きく分布が異なる変数を重要視すべき

2.10 まとめ

- 伝統的な手法と同様に、AI/機械学習も「ミスを犯す」
  - データが偏ると推定結果も偏り、推定誤差が生じる
- AI/機械学習は、伝統的な手法に比べて、推定誤差の性質が不透明
  - 最終的な推定値の性質にも転嫁(Pass-through)される

2.11 まとめ

- 改善策
  - AI/機械学習の推定精度改善、推定誤差の性質の明確化
    - 現状、限界がある
  - 推定値への**転嫁されにくい**方法で活用
    - 一例が Double selection
      - 紹介論文 (Angrist and Frandsen, 2022)/直近の応用論文(Novella, Rosas-Shady and Freund, 2024; Hill and Stein, 2025)

3 Statistical inference

3.1 Statistical valid inference

- Statistical Inference: 前提(データの特徴 + 仮定(ランダムサンプリングなど)) $\Rightarrow$ 母集団の特徴
- 要求: 推論の Validity を保証
  - Deductive (演繹的) validity = 前提が真ならば、結論は**必ず**真
  - Inductive (帰納的) validity = 前提が真ならば、結論は**概ね**真
  - (Logical methods for AI 参照)

3.2 Invalid (無効な) inference

- 「ランダムサンプリング + OLS における係数値が 50」 $\not\Rightarrow$ 母平均は必ず/概ね 50
  - データが偶然上振れ/下振れした可能性が大きい

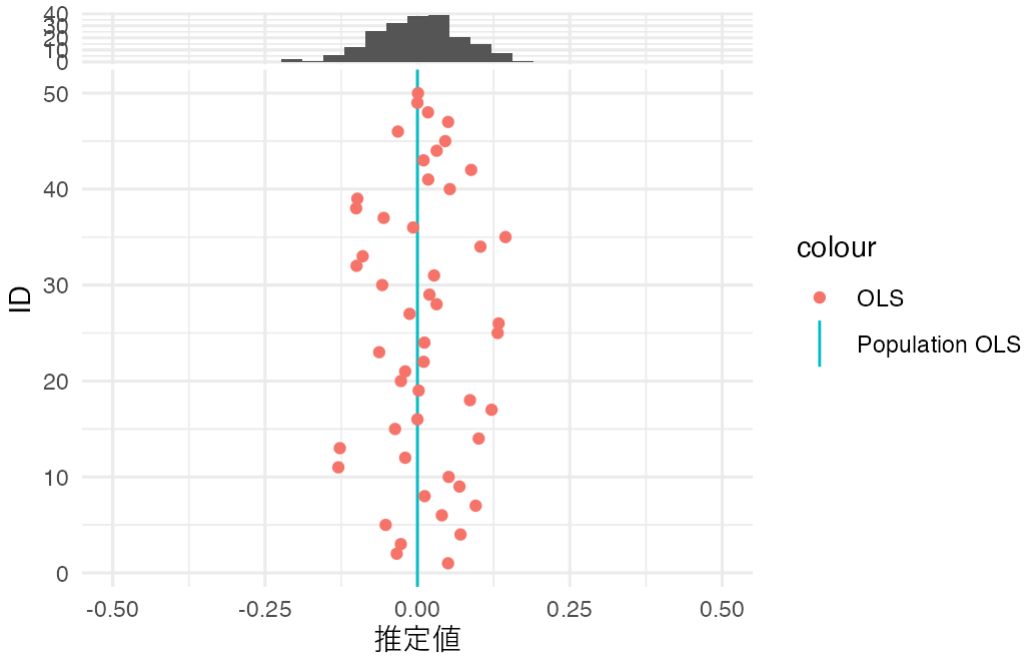
3.3 Deductively valid

- 「ランダムサンプリング + 無限大の事例数 + OLS における係数値が 50」
  - $\Rightarrow$ Population OLS でも 50 (一致性)
  - 応用上、前提が非現実的すぎる

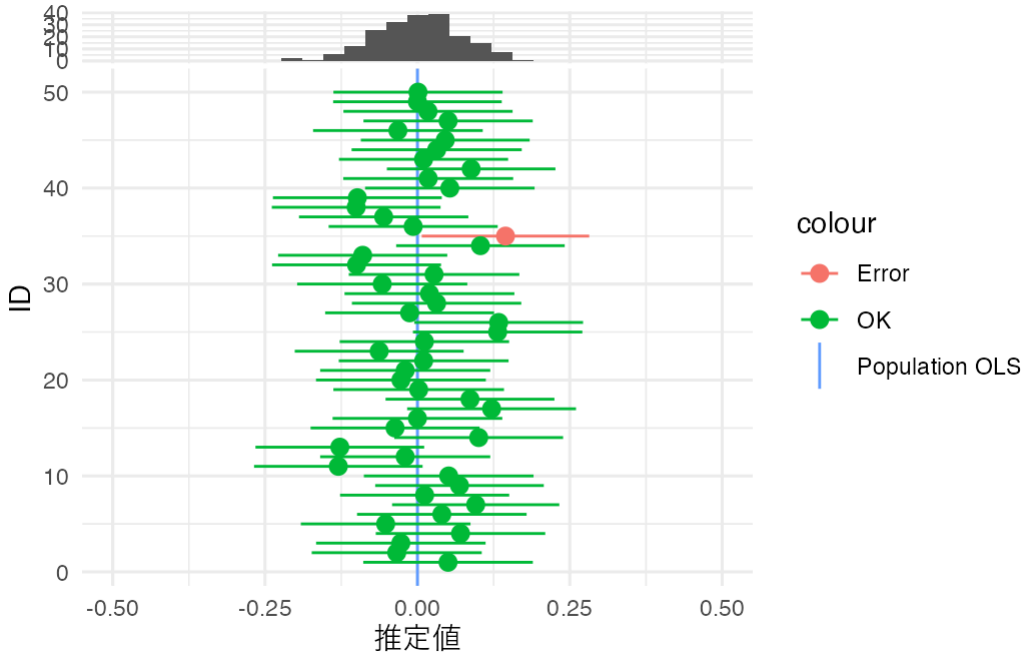
3.4 Inductive validity

- 「ランダムサンプリング + 事例数が十分に大きい + 計算した 95 % 信頼区間は [18 万円,22 万円]」
  - ⇒ 母平均は、概ね[18 万円,22 万円]の間
  - 平均値の推定値の分布が、正規分布で近似できることを利用

3.5 イメージ: 200 事例



3.6 イメージ: 200 事例



3.7 機械学習を活用した推論

- 「完璧な予測ができるモデルの予測値が 50」
  - ⇒ 母平均は必ず 50
- Valid だが、少なくとも社会データにおいては、前提が非現実的
  - 仮にモデルが母平均を完璧に捉えたとしても、削減不可能な誤差のせいで、完璧な予測は不可能
- 補論: 予測値についての Valid inference も議論される (Conformal inference; Angelopoulos, Bates and others (2023) など)

4 Reference

Bibliography

Angelopoulos, A. N., Bates, S. and others (2023) “Conformal prediction: A gentle introduction,” Foundations and Trends® in Machine Learning, 16(4), pp. 494–591

Angrist, J. D. and Frandsen, B. (2022) “Machine labor,” Journal of Labor Economics, 40(S1), p. S97–S140

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) “Inference on treatment effects after selection among high-dimensional controls,” Review of Economic Studies, 81(2), pp. 608–650

Hill, R. and Stein, C. (2025) “Scooped! Estimating rewards for priority in science,” Journal of Political Economy, 133(3), p. 0–1

Novella, R., Rosas-Shady, D. and Freund, R. (2024) “Is online job training for all? Experimental evidence on the effects of a Coursera program in Costa Rica,” Journal of Development Economics, 169, p. 103285–103286