部分線形モデル

Semiparametric 推定への応用

川田恵介

Table of contents

Semiparametric 推定への応用 2
動機 2
例 Stacking (OLS with 2 次項 + 剪定ずみ決定木) 3
主要参考文献 3
実装 3

Quick Start 4
部分線形モデル 4
Partialling-out algorithm 4
主要な性質 4
他の手法の問題点 4
数値例 5
数値例 5
数値例 6
まとめ 6

Reserch RoadMap 6
実証分析の RoadMap 7
大雑把な整理 7
例: Research Question 7
例: Identification 7
例: Causal Identification 8
例: Summary 8
例: Estimation 8
まとめ 8

Summary 9
条件付き平均差 9

周辺化 9
部分線形モデルの Estimand 9
詳細: 単回帰の復習 10
詳細: Partialling-out の推定結果 10
詳細: Partialling-out の推定結果 10
まとめ 10
Reference 11

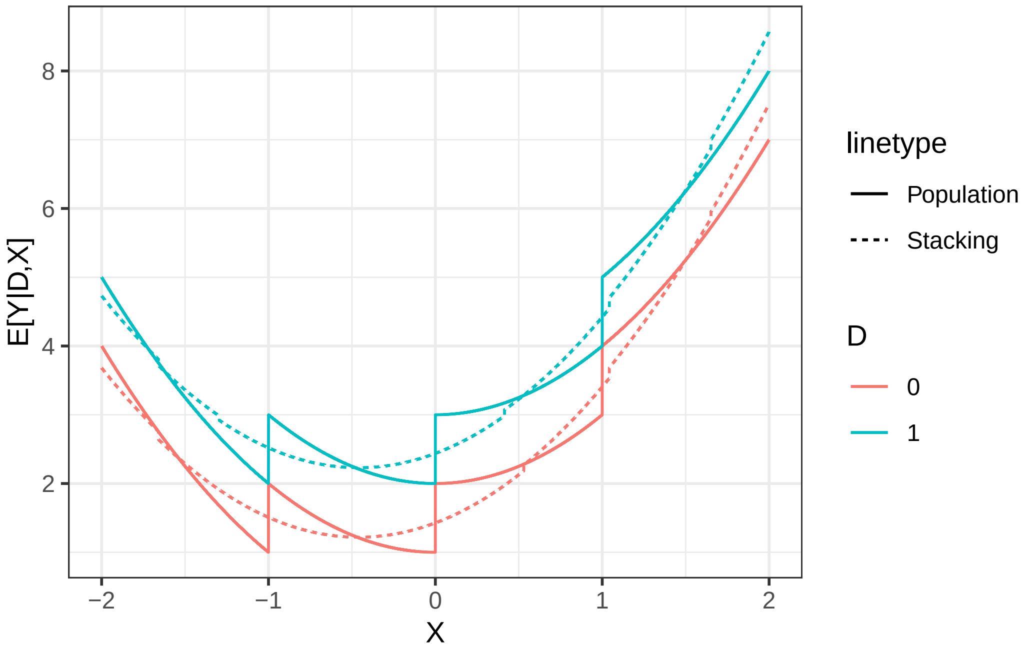
Semiparametric 推定への応用

- 教師付き学習を、母集団の記述統計量推定に応用
- Semiparametric 推定の文脈に落とし込む
  - 教師付き学習 = 多次元でも実用的な (擬似)Nonparametric 推定法

動機

- 経済学におけるデータ分析の主要な関心は、母集団の重要な特徴 (因果効果, 格差など) の理解
- 教師付き学習 := 母平均関数 $E_P[Y|X]$ の近似関数 $g(X)$ の推定
  - $g(X)$ を母平均関数理解に使えるか？
  - 複雑な $g(X)$ の特徴を理解する手法は多く存在 (Molnar 2022)
- 問題点: Well-specified model を"OLS" 推定した場合と比べて、推定誤差の定量化が難しい

例 Stacking (OLS with 2 次項 + 剪定ずみ決定木)



- 10 万サンプルで推定しても、ズレている

主要参考文献

- (Double/)Debiased Machine Learning (Chernozhukov et al. 2018)
  - 関連ワード: Neyman's ohtogonality/Locally robust score/Efficient influence function (Chernozhukov et al. 2022), Mixed bias property (Rotnitzky, Smucler, and Robins 2021)
- 大量の解説論文 (Ichimura and Newey 2022; Fisher and Kennedy 2021; Hines et al. 2021)
- 教師付き学習の有力な応用 (Leist et al. 2022)

実装

- DoubleML (R/Python)
- grf (R)
- tlverse (R)
- econml (Python)

- STATA

- 日本語での紹介

Quick Start

- X を一定とした下 (Control) で、D と Y の関係性を推定する

部分線形モデル

- Partial Linear Model

$$Y_i = \underbrace{\tau_P}_{Estimand(\text{関心となる特徴})} \times D_i + \underbrace{b(X_i)}_{\text{未知}} + \underbrace{u_i}_{E[u|D,X]=0}$$

- 主要な仮定: τ_P は、X, D に" 依存していない"
  - Misspecification が生じていたとしても、母分布上で解釈可能 (Vansteelandt and Dukes (2022))

Partialling-out algorithm

1. Y, D の予測モデル $g_Y(X), g_D(X)$ を、何らかの方法で交差推定する
2. 予測誤差 $Y - g_Y(X), D - g_d(X)$ を単回帰する $(Y - g_Y(X) \sim D - g_D(X))$
3. τ_P の推定値 = 単回帰の係数

主要な性質

- "緩やかな" 仮定のもとで、一致/漸近正規性を満たす
  - **C**onsistent and **A**symptotically **N**ormal (CAN) estimator
  - 信頼区間の近似計算が可能

他の手法の問題点

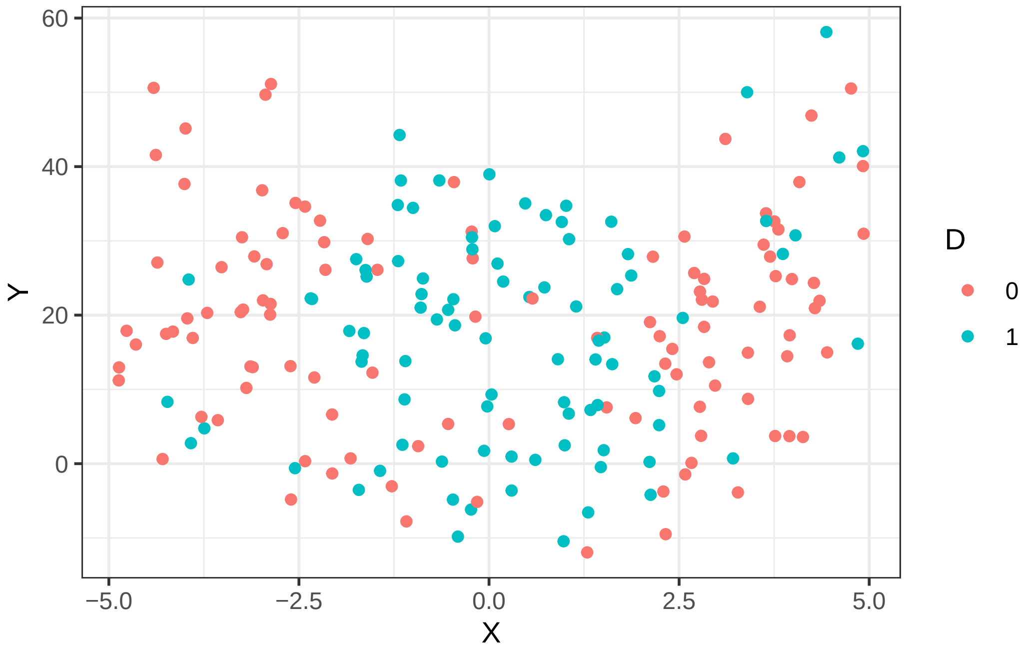
- 近似モデル $g(D, X) \simeq E[Y|D, X]$ を推定し、$E[g(1, X) - g(0, X)]$ を計算する
  - Plugin-in estimator
  - 一般に CAN estimator にならない
- OLS で推定: 深刻な定式化依存 → Not consistent

- 教師付き学習: 収束が遅い → (May be) consistent but not Asymptotically normal

数値例

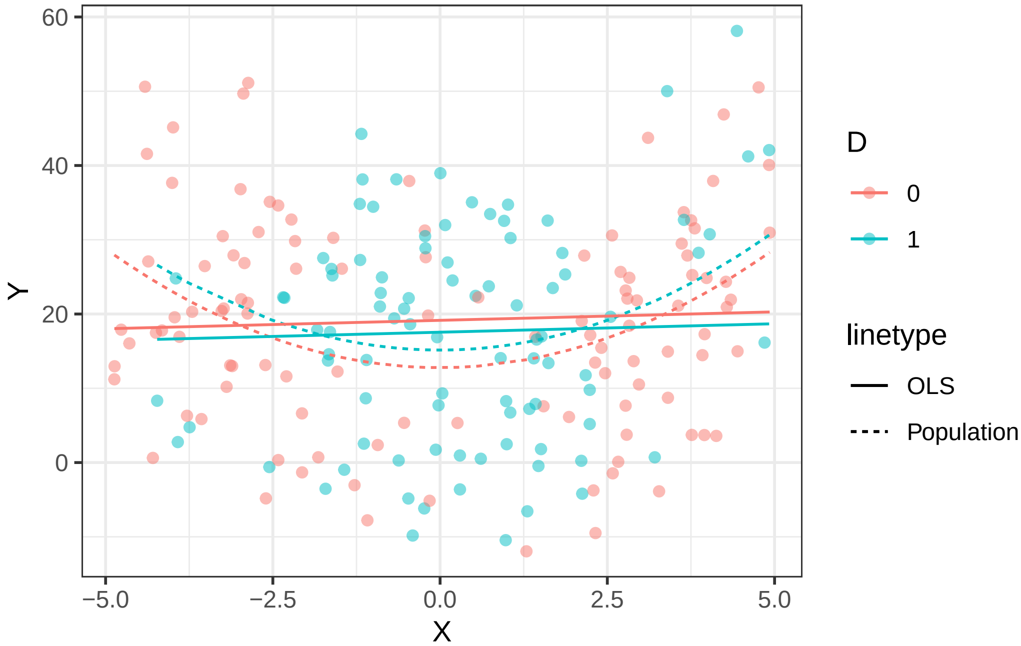
- 「格闘ゲームをプレイした経験間で、主観的幸福度はどの程度異なるのか？」
  - 年齢と主観的幸福度、格闘ゲームのプレイ経験には強い相関がされるので、“コントロール”
- 母集団
  - 格闘ゲームのプレイ経験があるグループの方が、主観的幸福度は高い
  - 40 歳前後が最も格闘ゲームのプレイ経験は高い
  - 年齢と主観的幸福度の間には、U 字の関係がある

数値例



- $Y \sim \beta_0 + \beta_1 D + \beta_2 X$ を行うと?

数値例



まとめ

- Partialling-out 自体は古典的なアイディア (少なくとも Robinson (1988))
  - 理論性質についても、精緻な議論がされていた
- 教師付き学習 + 交差推定を組み込むことで、より幅広い状況に応用が可能に
- 他の応用 (例: AIPW, Sensitivity, Panel Data, Mediation Analysis) と、同様の原理を共有する

Reserch RoadMap

- データ分析をめぐる手法や概念について、依然として混乱が見られる
  - 過度に" 万能" 視
  - 過剰な縦割り/縄張り意識
- "研究 RoadMap" の中にしっかり埋め込んで理解/整理される必要がある

実証分析の RoadMap

1. Research Question: 知りたい (母) 分布の特徴は?
2. Identification Step: 観察できる変数のみで書き下せるか?
3. Summary Step: 推定 & 理解可能な程度に単純化
4. Estimation Step: 推定 & 理解可能な程度に単純化
5. Coding & 分析 Step

大雑把な整理

1. Research Question: 実務/研究 (理論的) 動機: 因果推論、比較研究、予測研究
2. Identification Step: 潜在結果/RegressionDiscontinuity/IV/DAG
3. Summary Step: 線形モデル, 周辺化条件付き平均差, 分位点差
4. Estimation Step: 教師学習/セミパラ推定/最尤法・ベイズによるパラメトリック推定
5. Coding & 分析 Step

- 異なる Question について、同じ推定手法は使える場合も多い

例: Research Question

- 賃金 ($=Y$)、大卒/高卒 ($=X$)、Windows/LinuxOS User($=OS$) が使えるとして、

1. 同一教育年数内 OS 間賃金格差 (比較/格差研究)
2. 同一教育経験内 OS 間賃金格差 (比較/格差研究)
3. OS が賃金に与える因果効果 (因果効果)

例: Identification

1. 同一教育年数内賃金格差 (比較/格差研究): 以下を比較すればいいので" 不要"

$$f_P(Y|Linux,X)\ VS\ f_P(Y|Windows,X)$$

2. 同一教育経験内賃金格差 (比較/格差研究): 教育経験が観察できないので、たとえば、以下の仮定が必要

$$f_P(Y|OS, \text{教育経験}) = f_P(Y|OS, \text{教育年数})$$

例: Causal Identification

- 因果効果の定義と Identification については、膨大な議論が存在 (Imbens 2022 など)
- 典型的な仮定は、Conditional unconfounderness/independence

$$f_P(Y|OS, X) = f_P(Y|OS, X, U)$$

- U : OS 選択の" 前に" 決まる全ての観察できない変数
  - 例えば、使用する OS がランダムに強制決定されていれば OK

例: Summary

- 一般に条件付き分布 $f_P(Y|OS, X)$ の OS についての比較は難しい
- 解釈が容易で推定可能な母集団上での記述統計量 (Estimand) を定義する
- 典型的な Estimand は、
  - $\tau_P(X) = E_P[Y|Linux, X] - E_P[Y|Windows, X]$
  - $\tau_P = E_P[\tau_P(X)]$
- Research Question 1-3 まで全てに" 有効"

例: Estimation

- Estimand を有限サンプルから推定する
- OLS, 最尤法, ベイズ, 傾向スコア, 教師付き学習などなど
  - ここではセミパラ推定 + 教師付き学習

まとめ

- Identification が大きく異なったとしても、同じ Summary や Estimation が活用可能なケースは多い
- 現状、教師付き学習の最も確立された応用先は、Estimation
  - 最後に他の Step への応用可能性も紹介

Summary

- Misspecification が生じた部分線形モデルは何を推定している?
    - Estimand は?

条件付き平均差

```
# A tibble: 2 x 4
  X       `Tau_P(X)` `E_P[Linux|X]` `f_P(X)`
  <chr>        <dbl>          <dbl>    <dbl>
1 高校卒          20          0.6        0.6
2 大学卒          10          0.001      0.4
```

- サブサンプル平均差で推定できる場合もある
    - X の値が増えると、サブサンプルサイズが非常に小さくなり、不可能になる

周辺化

- より推定が容易な目標

$$\tau_P = \underbrace{W(\text{高校})}_{Weight} \times \underbrace{\tau_P(\text{高校})}_{=20} + \underbrace{W(\text{大学})}_{Weight} \times \underbrace{\tau_P(\text{大学})}_{=10}$$

- Weight は、"本質的には"、任意
    - $W(\text{大学}) = W(\text{高校}) = 0.5$ ならば、$\tau = 15$
    - $W(\text{大学}) = 0.4(\text{大卒比率}), W(\text{高校}) = 0.6(\text{高卒比率})$ 　ならば、$\tau = 16$

部分線形モデルの Estimand

- 周辺化された条件付き平均差: ただし

$$W(x) = \frac{V_P(OS|x) \times f_P(x)}{V_P(OS)}$$

- $V_P(OS|X) = E_P[(OS - E_P[OS|E])^2|X]$ (OS の分散)
    - $W(\text{大学}) \simeq 0, W(\text{高校}) \simeq 1$
    - $\tau_P \simeq 20$

- あまり直感的ではないかも (代替案: AIPW)

詳細: 単回帰の復習

- OLS の係数値 = 共分散/分散
- 同じ X を共有するサブグループ (母集団) 内で、$Y = \beta_0 + \beta_1 D + u$ を回帰すると

$$\tau_P(X) := \beta_1 = \frac{cov_P(Y, D|X)}{var_P(D|X)}$$

$$= \frac{E_P[(Y - E_P[Y|X])(D - E_P[D|X])|X]}{E_P[(D - E_P[D|X])^2|X]}$$

詳細: Partialling-out の推定結果

- $Y - E_P[Y|X] \sim D - E_P[D|X]$ を母集団で回帰すると、

$$\tau_P = \frac{E_P[(Y - E_P[Y|X])(D - E_P[D|X])]}{E_P[(D - E_P[D|X])^2]}$$

- 繰り返し期待値 $E[Y] = \int E[Y|X] \times f(X)dX$ を使うと?

詳細: Partialling-out の推定結果

$$\tau = \int \frac{E_P[(Y - E_P[Y|X])(D - E_P[D|X])|X]}{E_P[(D - E[D|X])^2]} f_P(X)dX$$

$$= \int \underbrace{\frac{E_P[(D - E_P[D|X])^2|X]}{E_P[(D - E_P[D|X])^2]}}_{=W(X)}$$

$$\times \underbrace{\frac{E[(Y - E[Y|X])(D - E[D|X])|X]}{E[(D - E[D|X])^2|X]}}_{=\tau(X)} \times f_P(X)dX$$

まとめ

- Partialling-out 推定は、母集団における記述統計" 単回帰の加重平均" (Y と D の BLP) について、信頼区間を提供
  - D が二値であれば、周辺化された条件付き平均差

- OLS は母集団における Best Linear Projection について、信頼区間を提供
- 一般に一致しないが、重要な例外
  - D と X が独立 (D がランダムに決定されているなど) であれば、BLP における D の係数値 = 周辺化された条件付き平均差

Reference

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4): 1501–35.

Fisher, Aaron, and Edward H Kennedy. 2021. "Visually Communicating and Teaching Intuition for Influence Functions." *The American Statistician* 75 (2): 162–72.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2021. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician* 76: 292–304.

Ichimura, Hidehiko, and Whitney K Newey. 2022. "The Influence Function of Semiparametric Estimators." *Quantitative Economics* 13 (1): 29–61.

Imbens, Guido W. 2022. "Causality in Econometrics: Choice Vs Chance." *Econometrica* 90 (6): 2541–66.

Leist, Anja K, Matthias Klee, Jung Hyun Kim, David H Rehkopf, Stéphane PA Bordas, Graciela Muniz-Terrera, and Sara Wade. 2022. "Mapping of Machine Learning Approaches for Description, Prediction, and Causal Inference in the Social and Health Sciences." *Science Advances* 8 (42): eabk1942.

Molnar, Christoph. 2022. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable . Christophm. Github. Io/Interpretable-Ml-Book."

Robinson, Peter M. 1988. "Root-n-Consistent Semiparametric Regression." *Econometrica* 56: 931–54.

Rotnitzky, Andrea, Ezequiel Smucler, and James M Robins. 2021. "Characterization of Parameters with a Mixed Bias Property." *Biometrika* 108 (1): 231–38.

Vansteelandt, Stijn, and Oliver Dukes. 2022. "Assumption-Lean Inference for Generalised Linear Model Parameters." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84 (3): 657–85.