中間まとめ

Machine Learning require sophisticated research roadmap

川田恵介

Table of contents

1 研究/学習のコツ

- RoadMap(工程表) を見失わない
  - 研究: 今何をやっていて、何をやりたかったか
  - 学習: 何を、何のために勉強しているのか
- 機械学習を含む様々な手法は RoadMap の中に埋め込んで整理すべき

1.1 Research Design の RoadMap: Recap

- 利用可能なデータの特徴 (事例数、変数、可能であれば欠損値の数など) を把握
- **Research question**: 研究課題の設定
- **Identification/Summary Strategy**: 研究課題に回答できる Estimand (推定対象) の定義
  - 含む Population (母集団) の定義
  - **人間**によるモデル化 (変数選択も含む) が必要
- **Estimation Strategy**: Estimand を近似する Estimator (推定結果) を得る方法を検討
  - **データ**によるモデル化 (変数選択も含む) が利用可能

1.2 Estimation Strategy

- 予測: Estimand $E[Y|X]$ に対して、データから近似結果 (Estimator) $g_Y(X) \simeq E[Y|X]$ を計算する
  - 例えば LASSO
- 記述: Estimand $\tau = E[Y|d, X] - E[Y|d', X]$ に対して、データから Estimator $\beta_D \simeq \tau$ を計算する
  - 例えば Double Selection
- 「Estimand を決めた**後に**、どのように Estimator を得るのか?」を (今後も) 議論

1.3 Quiz

- 以下の研究計画の問題箇所を選べ
- “本講義への参加が 30 歳時点での所得に与える因果効果を推定したい。データからは、事例数 500、変数としては講義への参加、30 歳時点での所得に加えて、$X = \{$ 講義への参加、30 歳時点での所得、生

まれ年、所属学部、初職の企業規模、初職の職種 } が活用できる。Double Selection を用いて、**X から重要な変数 Z を選択**し、**Y~ D+Z を OLS で推定**し、**D の係数値を通常の信頼区間**ともに報告する。”

- 回答ページ

1.4 Recap: Double LASSO algorithm

- 限れた事例数で、τ をどのように近似するか?
  - 適切な推定モデルを構築する
    - X を十分に複雑化する
    - 近似する上での有用ではない変数を排除する
      - $\tau \sim E[Y|d, Z] - E[Y|d', Z]$ となるような部分集合 $Z \subset X$ を探す
- 注意点: 推定目標 (Estimand) はあくまでも $\tau = E[Y|d, X] - E[Y|d, X]$

1.5 Recap: Why “Double”?

- $Y = \tau D + \beta_X X + u$ を想定
- $$\begin{aligned} &E[Y|D=1] - E[Y|D=0] \\ &= \tau + \underbrace{\beta_X \times \{E[X|D=1] - E[X|D=0]\}}_{Confounding\ term} \end{aligned}$$
- Bias の大きさは、β_X と $E[X|D=1] - E[X|D=0]$ **双方に**依存

1.6 Recap: Why “Double”?

- 例: 以下は同じ推定結果をもたらす
  - $\beta_X = 5$ & $E[X|D=1] - E[X|D=0] = 0.1$
    - D の予測モデルのみでは、X は除外されやすい
  - $\beta_X = 1$ & $E[X|D=1] - E[X|D=0] = 0.5$
    - Y の予測モデルのみでは、X は除外されやすい
- Y/D の予測モデルを用いて、double check する必要がある

1.7 Recap: Approximation

- (くどいが) あくまでも $E[Y|D=1,X]-E[Y|D=0,X]$ の近似が目標
  - $E[Y|D=1,Z]-E[Y|D=0,Z]$ ではない
- Bias-Variance Tradeoff を解く
  - 事例数が少なければ、より多くの変数を落とさざる得ない
  - 無限大の事例数があれば、X が有限である限り、$Z=X$ で OK

2 Estimand の設定

- 現代的研究において、特に強調され、多くの労力を注ぎ込む。
  - 因果推論、潜在結果 (Potential Outcome;PO), Directed Acyclical Graphs (DAG) を活用する場面
  - Estimand を**定義**するために、変数を選択する
- 「Y に関係していそうなものを全てを X として活用する」は、ほとんどの応用で不適切

2.1 例: 格差研究 (記述研究)

- Estimand = 何を社会的に問題のある" 差" と考えるのか、価値判断への (論文内での)"コミット" (Rose 2023)
- 「Racial group W/B ($=D$) の間で、交通違反の検挙率 ($=Y$) にどのような差が存在するのか」
  - X 候補 = 違反速度

2.2 Estimand の設定: 格差研究

- 警察による差別研究: $\tau=E[Y|B,\text{違反速度}]-E[Y|W,\text{違反速度}]$ が妥当な Estimand
  - 同じ" 罪" を犯したとしても、Race 間で司法判断に差が存在するのであれば、差別の証拠
- Race 間の格差研究: $\tau=E[Y|B,\text{違反速度}]-E[Y|W,\text{違反速度}]$?
  - Race によって違法速度での運転に" 追い込まれている" 人の割合は異なるかもしれない
  - $\tau=E[Y|B]-E[Y|W]$ が妥当

2.3 例: 因果効果 (記述研究)

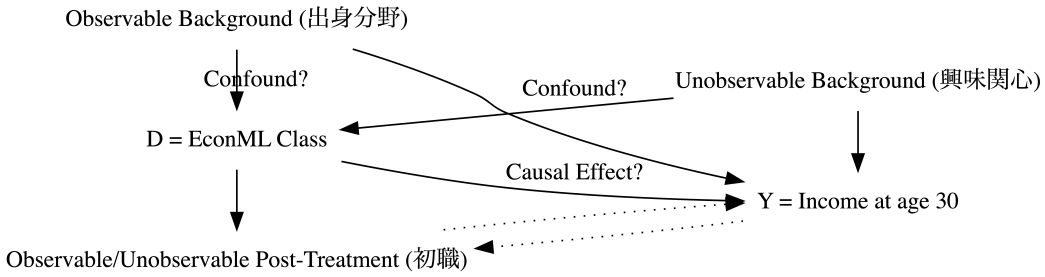
- Y について観察される差は、何が原因で生じたのか?
- ある変数 D を変化させた場合、Y にどのような差をもたらすのか?
  - 重要だが不毛な議論になりやすい
    - 有益な概念装置 (PO,DAG) が、複数提案され、補完的に活用できる状況になっている (Chap 2, 4-8 in CausalML 参照、他には Heckman and Pinto (2024))
  - 共通見解: X の選択 = 仮想的な実験結果 (Target Trial) へのコミットとして (も) 解釈できる

2.4 例: 本講義の因果効果

- 本講義を受講することが、30 歳時点での所得にどの程度影響を与えるのか?
  - $E[Income|Attend] - E[Income|NotAttend]$ = 因果効果?
  - 本講義は何の付加価値をもたらさなかったとしても、他の変数についての差があり、参加者と非参加者の間で賃金格差は観察されうる
    - 所得につながりやすい学部・大学院出身/ない
    - データ分析について関心がある/ない

2.5 例: 本講義の因果効果

- DAG による表現 (Chap 7 in CausalML 参照)
  - 矢印 = “因果効果”



2.6 Estimand as Target Trial

- 矢印はどのように決まるのか?

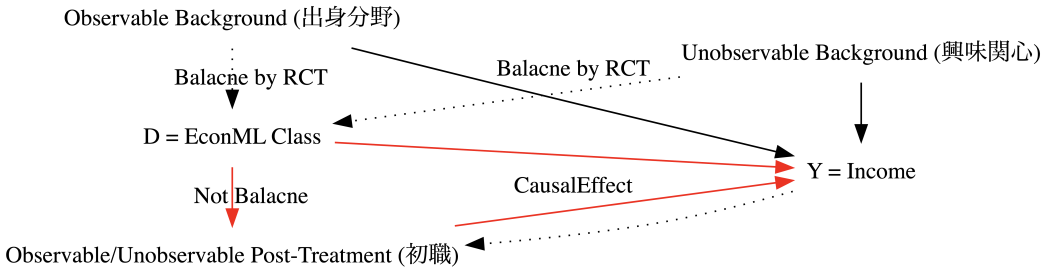
- Estimand の決定には、実用的な定義が必要
    - 母集団上での (仮想的な) 実験結果 (Target Trial)

2.7 Randomized Controlled Trial (RCT)

- 最も有名な Target Trial (Chap 2 in CausalML 参照)
- 無限大の被験者について、相互作用がない状態で、D を被験者にランダムに割り振る
    - Y の差を D 因果効果と” 見做す”
    - Observable/Unobservable 問わず、Background の分布は D 間でバランスする
    - Post-Treatment はバランスしないが、因果効果の一部 (Mediation Effect) として解釈する

2.8 例: RCT の実行した場合

- 赤線を因果効果と定義する



2.9 推定への含意

- 現実に実験できたとしても、被験者数は有限
    - **偶然**、背景属性はズレる
    - 推定手法で対処可能 (信頼区間等)

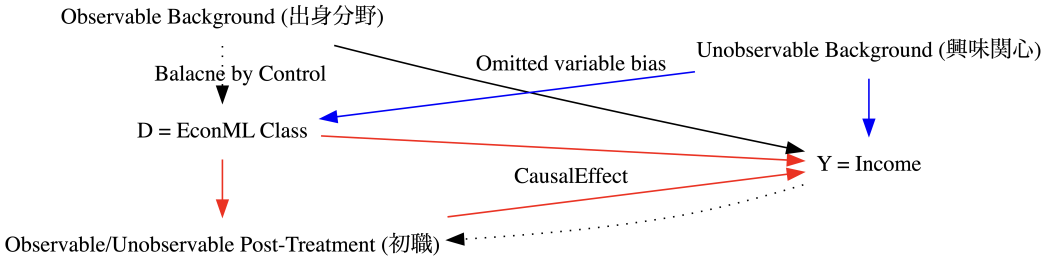
2.10 X の選択への含意

- Observable Background を選び、X に加える
    - Post-Treatment は除外
- Unobservable background は加えたいができない (Omitted variable bias を引き起こす)

- Sensitivty 分析などを検討
  - Ding et al. (2023), Section 16-19 in Ding (2023) などを参照

2.11 例: コントロール変数による識別

- 赤線を因果効果と定義する
  - τ = 赤 (Causal Effect) + 青 (Confound)



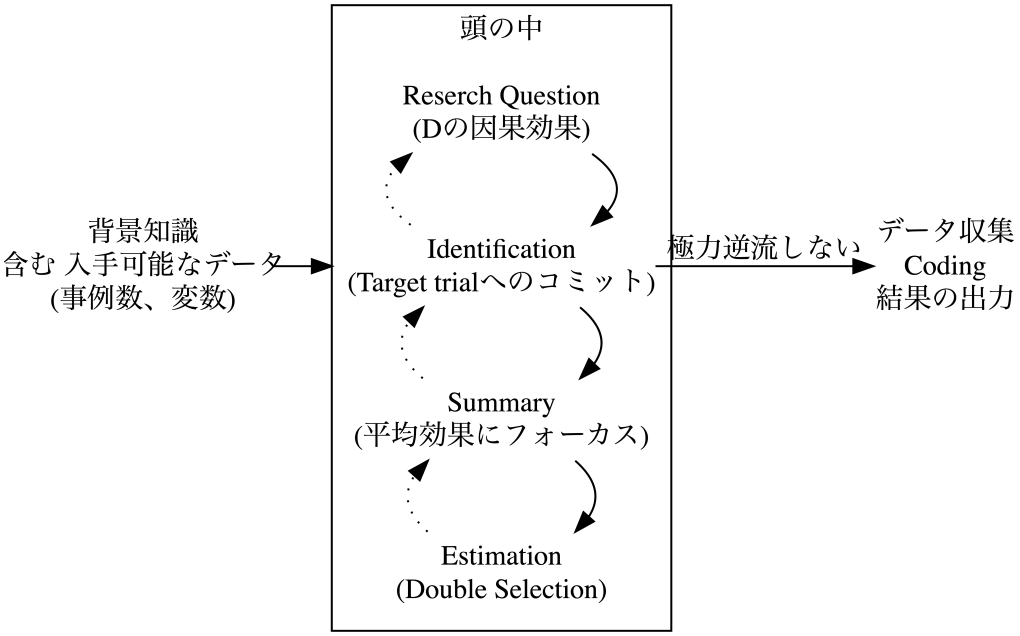
2.12 データによる Estimand の定義

- Double Selection に選択された $Z \subset X$ を用いて、Estimand を定義するのであれば、事例数に応じて推定対象が変化することを受け入れる必要がある
- “Post-treatment/Background”/“問題のある差か否か”、が分析に用いる事例数に依存する
  - ???

2.13 データによる Estimand の定義: 悪例

- 例: 本講義の因果効果研究: データに含まれる (Y,D 以外の) すべての変数 X から、Double selection で Z を選択
  - 事例数が増えれば、$X = Z$ に含まれる
    - 事例数が増えれば、**母集団 (あるいは社会) において**、初職によって、本講義への参加**が**変化するようになる
      - ???

2.14 まとめ: Research Design



2.15 まとめ

- Estimand を定義する議論と推定する議論 (Estimation) を分離する
  - “CausalInference/因果推論入門” 系は、定義を議論
  - “統計学/機械学習入門” 系は、推定を議論
    - 本講義の力点もこちら
  - “経済学” 系は、研究課題を議論

2.16 まとめ

- 私見：推定方法の進歩（含む機械学習の導入）により、信頼できる Estimation を得やすくなっている
  - 方法論の進歩を取り込むことが重要
    - 研究者は、Estimand の定義により注力できる

2.17 発展

- 因果推論において、Unobservable background (Unobservable confounders) への対処は難しい課題
- 大量のアプローチが提案
  - Instrumental Variable, Pallarel Trend (Panel Data), Regression-Discontinuity, Senstivity
    - Chap 12,13,16,17 in CausalML, Ding (2023) 参照
- 大量の変数の中から、因果関係を発見する方法 (Causal Discovery) も研究されているが、現状、(私見では) あまりにも強すぎる仮定を要求する (Daoud and Dubhashi 2023)

Reference

Daoud, Adel, and Devdatt Dubhashi. 2023. "Statistical Modeling: The Three Cultures." *Harvard Data Science Review* 5 (1).

Ding, Peng. 2023. "A First Course in Causal Inference." *arXiv Preprint arXiv:2305.18793*.

Ding, Peng, Yixin Fang, Doug Faries, Susan Gruber, Hana Lee, Joo-Yeon Lee, Pallavi Mishra-Kalyani, et al. 2023. "Sensitivity Analysis for Unmeasured Confounding in Medical Product Development and Evaluation Using Real World Evidence." https://arxiv.org/abs/2307.07442.

Heckman, James, and Rodrigo Pinto. 2024. "Econometric Causality: The Central Role of Thought Experiments." *Journal of Econometrics*, 105719.

Rose, Evan K. 2023. "A Constructivist Perspective on Empirical Discrimination Research." *Journal of Economic Literature* 61 (3): 906–23.