Introduction: イメージの共有

経済学のための機械学習入門

川田恵介

講義コンセプト

目標

- 統計コンセプト（平均、OLS）について、初歩的な理解を持ち、R (ないし Python) に触れたことがある院生/学部生が執筆する論文の Quality Up
  - 各自の Research Question について、より頑強、現実的な計算時間かつ多様な推定量を紹介
  - キャッチアップする意欲がる完全初学者も歓迎
  - 経済学以外の背景 (他の社会科学や医学、工学など) の受講生も歓迎

実習

- 講義 + Live coding + 実習
- 実習では R をサポート
  - Main Package: mlr3verse, mlr3pipelines, DoubleML, grf
- 課題は python でも OK

関連講義

- 関連性が高くおすすめの科目: 坂口さんの提供する
  - Machine Learning for Economics
  - 経済学とコンピューターサイエンス I /II
- 本講義の比較優位: 言語が日本語!?, セミパラ推定への応用において Treatment Effect Risk (Kallus 2022) や Sensitivity (Chernozhukov et al. 2022) を紹介

スケジュール

- S1/S2

1. 教師付き学習 (Stacking with OLS and RandomForest)
2. セミパラ推定への応用 ((Population) Parameter Estimation with mixed-bias property (Rotnitzky, Smucler, and Robins 2021))
3. 時間があれば他のアルゴリズム紹介 (LASSO/Boosting/BART など)

- 3 回程度課題を設定

具体的イメージ

- “more flexible and more principled” に課題解決
- 定式化問題
  - 予測モデルを作れと言われたが、どのような**式**を推定すれば良いのかわからない
  - コントロール変数を加えた分析をしたいが、どのように定**式**化すれば良いのかわからない
  - 効果の異質性を検証したいが、どのように定**式**化すれば良いのかわからない
- **どのように研究課題をデータ分析に落とし込めばいいのか、わからない**

やらないこと

- Deep Learning, Generative Model (Koenecke and Varian 2020; Kaji, Manresa, and Pouliot 2020), Text Analysis (Gentzkow, Kelly, and Taddy 2019), Causal Discovery (Nogueira et al. 2022), Reinforcement learning (Iskhakov, Rust, and Schjerning 2020)

次回までに

- 講義中に R で作業できる環境整備
- おすすめは
  - Local に R+ Rstudio をインストール
  - Posit cloud (旧 R cloud) に登録
- 講義資料と ExampleData を講義レポジトリからダンロード

教師付き学習

機械学習

- 統計学とは異なるルーツを持つデータ分析方法
  - AI の開発
- 学術・実務研究において、幅広く活用されている手法群を提供
  - Estimand は明確に定義できるが、変数間の具体的関係性は BlackBox な応用 (経済学!!!) において高い比較優位
  - 計量〇〇において、母関数への”Fitting” を行うツールとして広く利用される
  - 計算機〇〇において、さらに高い期待?

分野

- **教師付き学習**
  - **予測研究**のみならず、記述・比較・因果研究においても応用法が” 確立” されている
  - Semiparametric 推定の議論 (Neway, Ichimura, Robinson, Robins…) を活用
- 教師なし学習, 強化学習, 敵対学習等々

教師付き学習

- $\{Y, X\}$ が観察できるデータ（事例集）を用いて、$\{Y, X\}$ の**一般的**な関係性を要約する関数を推定する
  - 予測や社会の推論に有益
- 一般的とは?

根本問題

- 研究のゴール: “合意可能” かつ” 有益な” 結論を得る
- 同じ事例集であれば、同じ結論を得ることは難しくない
- **同じ社会を対象として同じ方法で事例収集しても、研究者によって事例が異なり、厳密な合意はできない**
- 母集団を用いて論点整理

  - 伝統的統計学と同じ!!!

繰り返しサンプリング

- サンプリング & 母集団
- 母集団から、データが**発生** (サンプリング) する
  - ゴール = 母集団の性質理解
- データは研究者によって異なるが、母集団は共通 (“一般的”)
  - 母集団上では、“共通” のゴールを定義できる

母集団

- “無限大のサンプルサイズを持つデータ”
- 同時分布 $f_P(Y, X)$ を用いて、記述
  - $\{Y, X\}$ の母集団における割合 (密度)
- 直接観察 (正確に推定) されることは **“あり得ない”**

典型的ゴール

- $f_P(Y, X)$ を全て推定することは極めて困難
  - “十分に単純” で” 正しい” モデルを推定する必要がある
- 応用上、有益な一側面 (Estimand) を推定
  - 原則、Estimand は母集団上で定義
- OLS と” 同じ”!!!

母平均関数

- 典型的な Estimand: 条件付き母平均関数

$$E_P[Y|X] := \int Y f_P(Y|X) dY$$

- $f_P(Y|X)$ を正確に推定できれば、$E_P[Y|X]$ は正確に計算できるが、逆はそうとは限らない

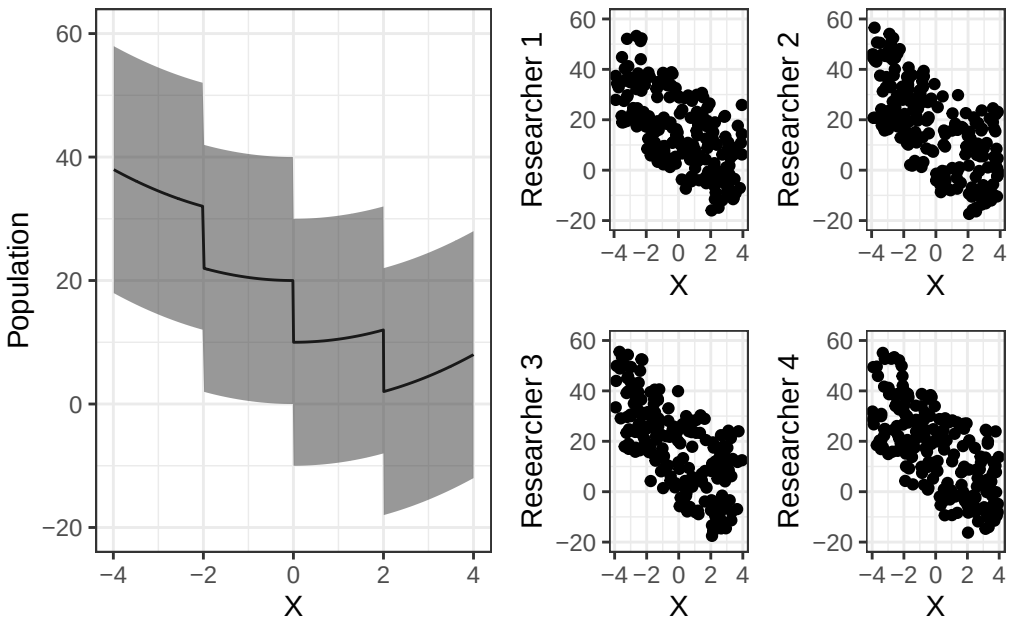
サンプリング

- 母集団から生成された、有限の事例数 (データ) のみ活用可能とする
  - 母集団を完全に観察することはできない
- 生成は確率的に行われる

仮定

- データのみから、母集団について得られる含意はほとんどない (“世の中いろんな人がいる” どまり)
- 仮定を追加し、推論を進める
- ランダムサンプリングの仮定: 事例は母集団から、ランダムサンプリングされる
  - 何を主張しているのか分かりやすく、データ収集のデザインによって保証可能
- “不透明” な仮定は極力減らす

数値例



典型的教師付き学習

- $E_P[Y|X]$ を近似する関数 $g_Y(X)$ を推定する
  - 以下の削減を頑張る

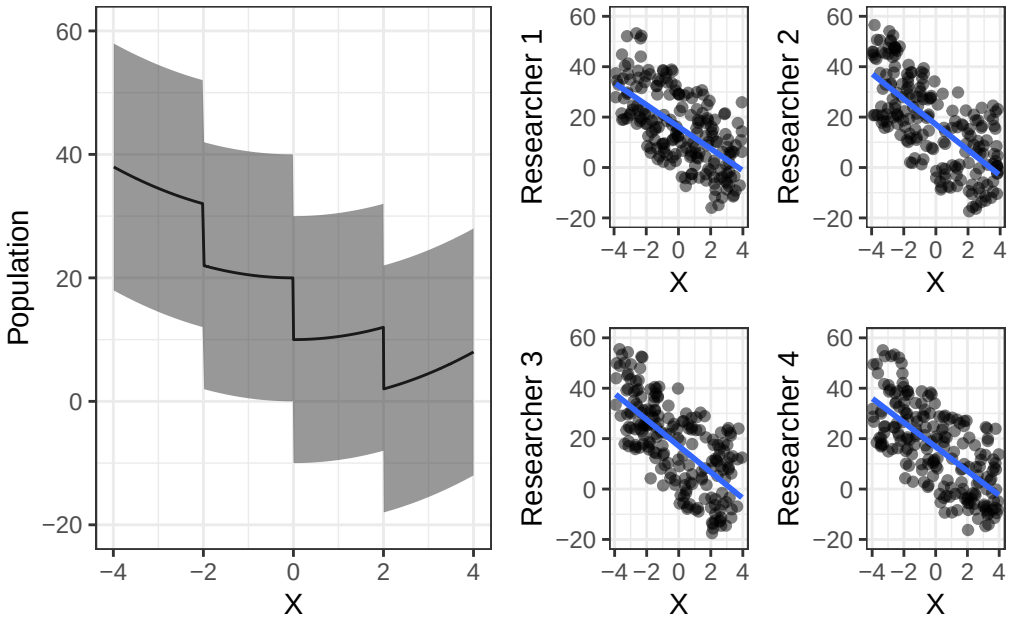
$$E_P[(E_P[Y|X] - g_Y(X))^2]$$

- 伝統的推定と” 同じ”!!!
  - 伝統的推定: 推定するモデルの” 複雑さ” を研究者が事前に指定 → 複雑すぎたり、**単純すぎたり**
  - 機械学習: モデルの複雑さもデータが決定

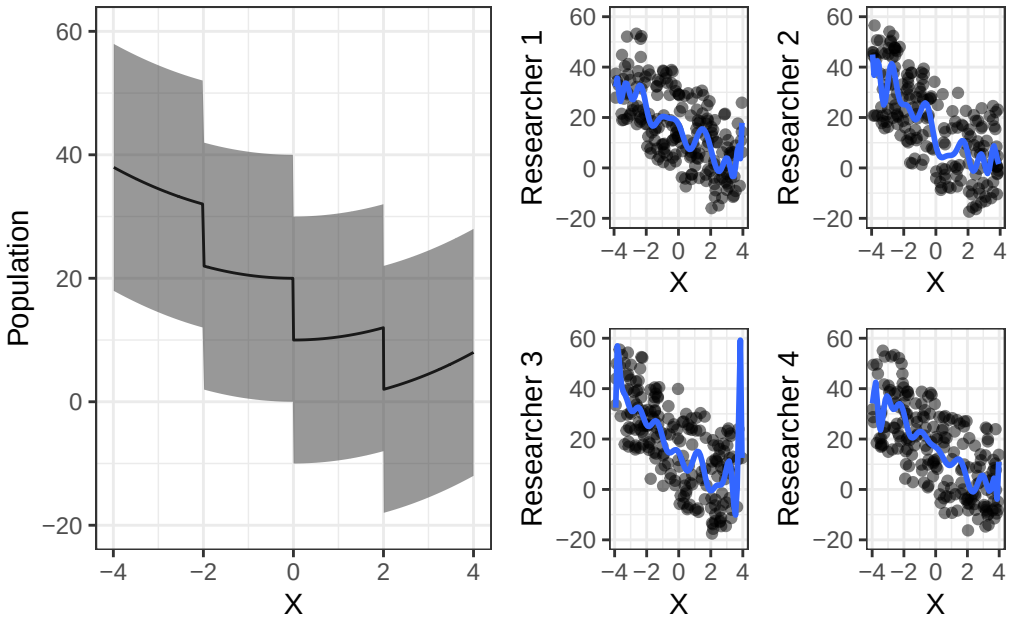
数値例

- ShortOLS: $g_Y(X) = \beta_0 + \beta_1 X$ と” 決め打ち” し、β をデータにもっとも適合するように推定
- LongOLS: $g_Y(X) = \beta_0 + \beta_1 X + .. + \beta_{20} X^{20}$ と” 決め打ち” し推定
- Stacking: OLS と RandomForest の加重平均

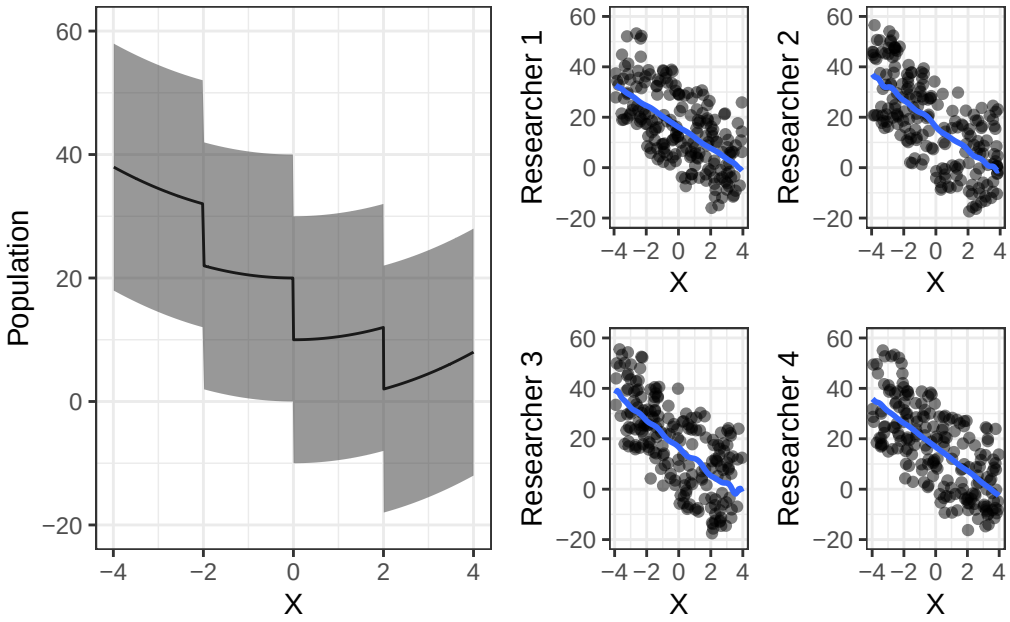
数値例: ShortOLS with N=200



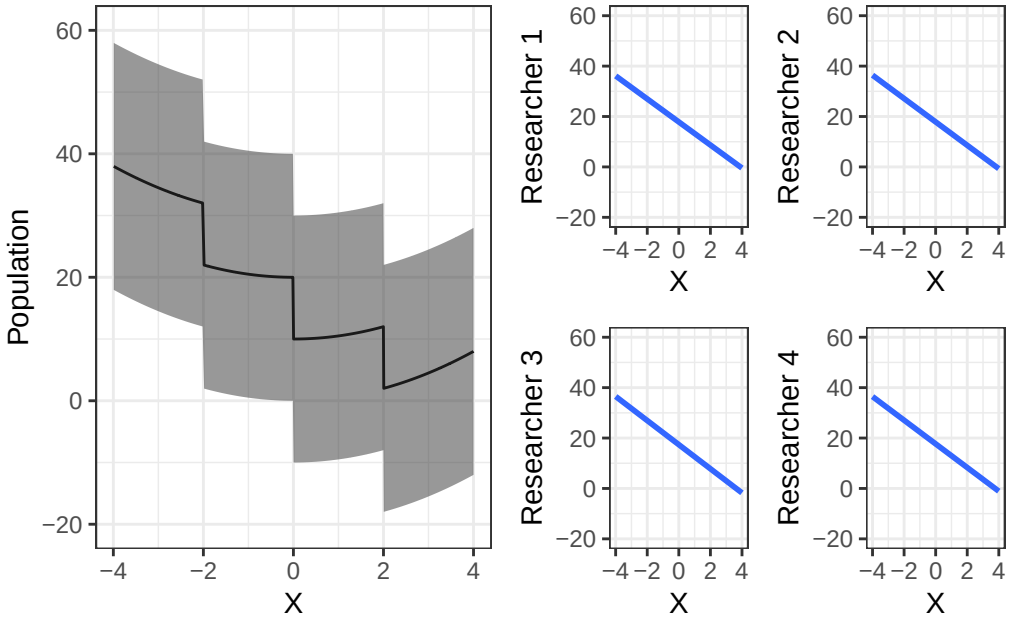
数値例: LongOLS with N=200



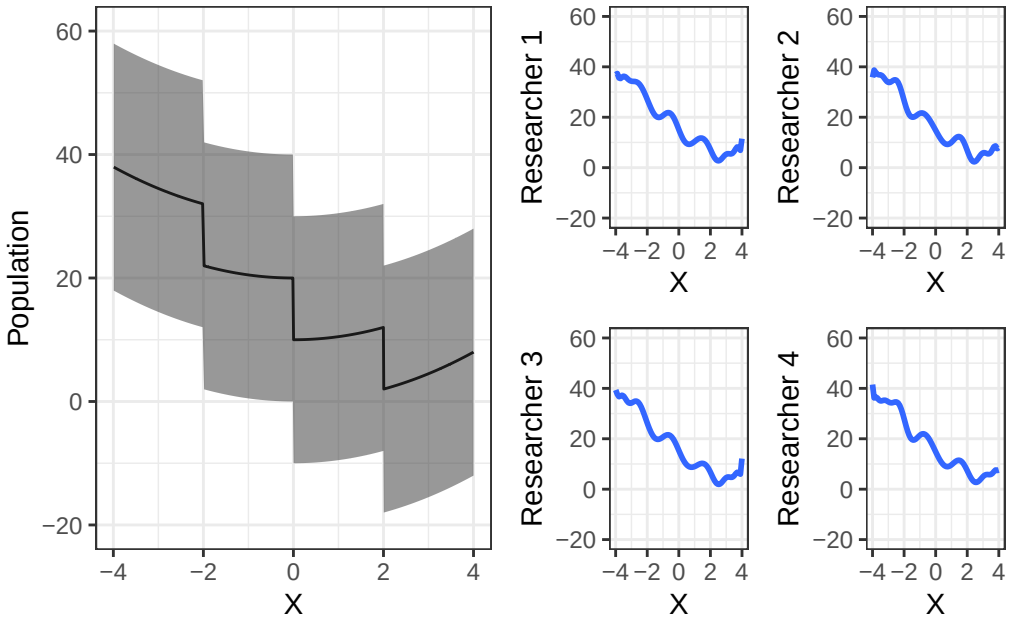
数値例: SL with N=200



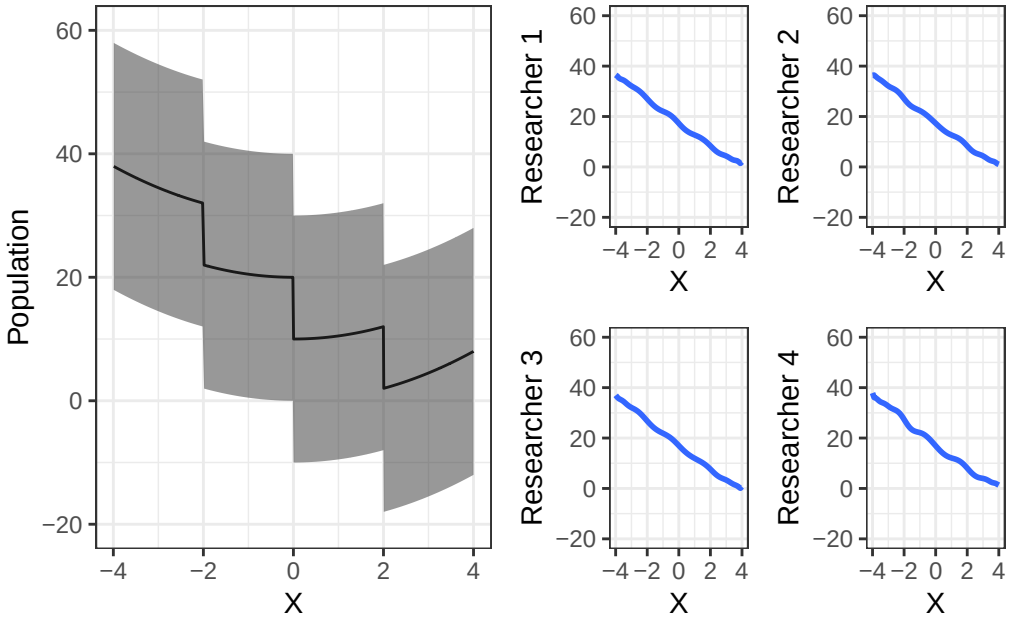
数値例: ShortOLS with N=5000



数値例: LongOLS with N=5000



数値例: SL with N=5000



予測研究

- 新しい事例について、X から Y を予測できるか？
  - 同じ母集団の事例であれば、$E_P[Y|X]$ は理想的な予測モデル
  - 教師付き学習で生成される $g_Y(X)$ は、実用的な予測モデル
- 教師付き学習のそもそもの動機
  - 実務 (政策) 研究において、極めて重要
  - 経済学研究としては、動機付けに工夫が必要!!!

実例: Einav et al. (2018)

- 1 年後生存 $(= Y)$ を、個人属性 (病歴含む) $(= X)$ から予測
- 研究動機: 終末期医療問題への基礎研究
  - 倫理的議論が目立つが、技術的に予測できるのか？
- 結論: できない

  - “死ぬ” とわかっている人に多くの医療資源を注ぎ込んでいるわけではない

Model Interpretation

- 一般に機械学習は、“複雑な” 予測モデルを生み出し、そのモデル自体の理解も難しい
  - 単純な OLS のように式で示されても????
- モデルの可視化 (Interpretation): Molnar (2022)
- 注意: “モデル” の” 可視化” であって、(経済学的な意味での) 解釈でも、母集団の可視化でもない

Individual Conditional Expectation

- 注目する X を一つ Pick Up し、予測値との関係性を図示
- 他の属性の値はどのように設定？
  - 他の属性は、データの値をそのまま使用
- 各事例について、部屋の広さ (Size) のみを仮想的に変化させた場合の予測値の推移

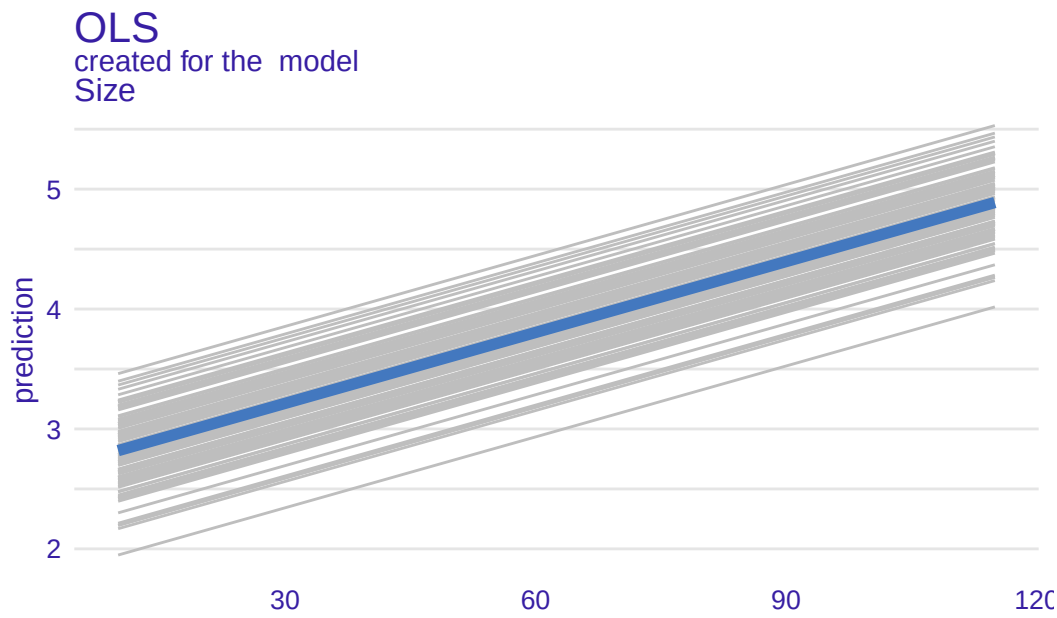
練習例

- 国交省が提供する不動産取引価格情報から東京 2 3 区の 2017/22 年に取引された中古マンション取引事例を取得
- 中古マンションの取引価格、取引時期を予測
  - $Y =$ 取引価格 (100 万円), 取引年 ($= 1$ 2021, $= 0$ 2019)
  - $X =$ 立地, 駅からの距離（分)、部屋の広さ、構造など

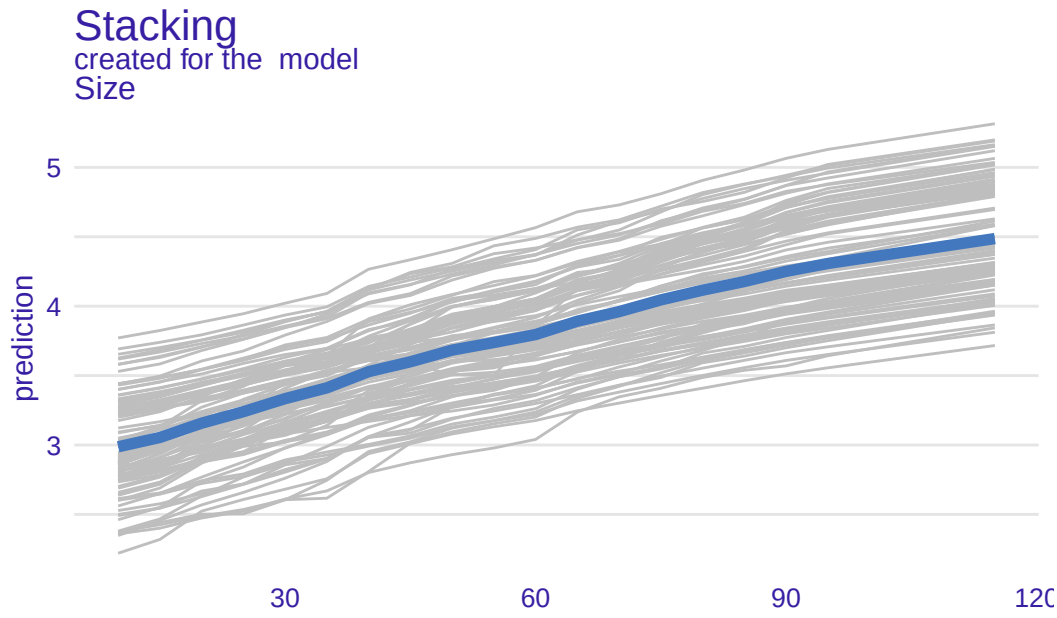
練習例: 推定方法

- OLS: $g_Y(X) = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$
- Stacking: OLS と RandomForest の加重平均

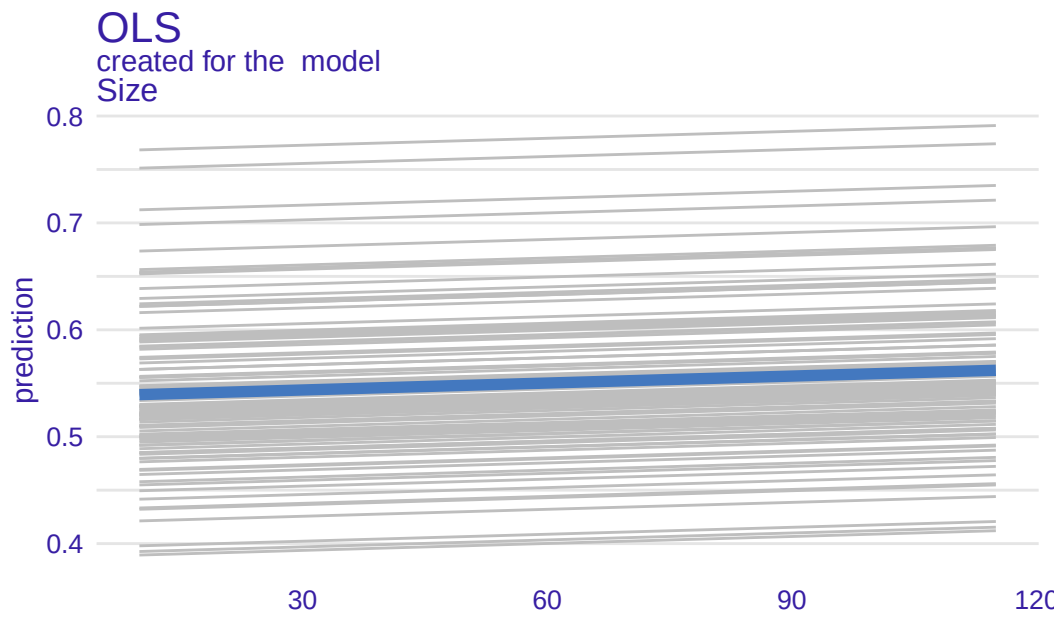
練習例: 価格予測 (OLS)



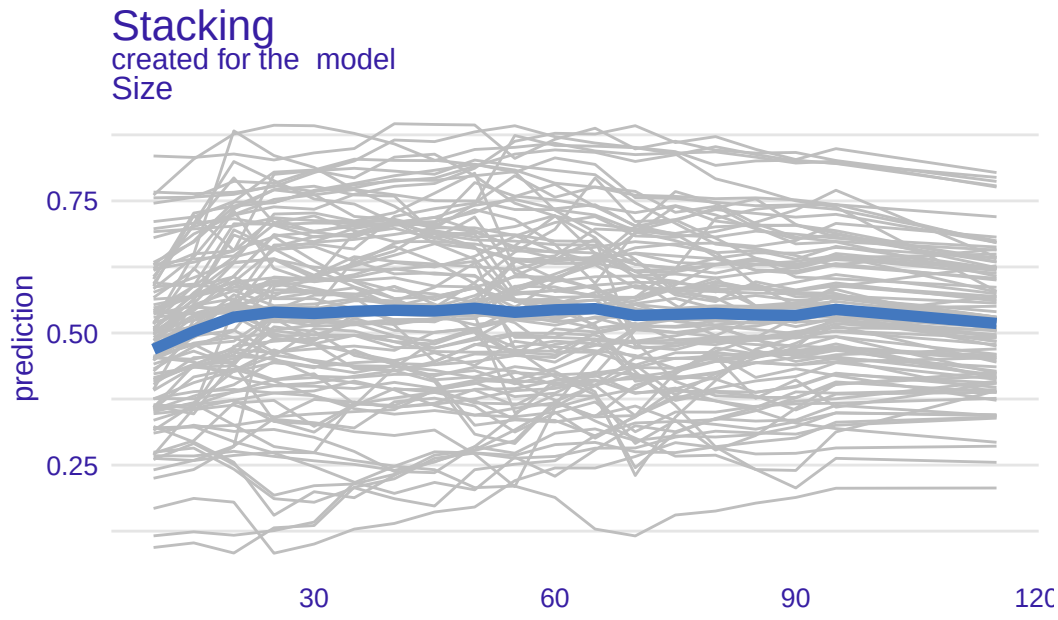
練習例: 価格予測 (Stakcing)



練習例: **取引年予測** (OLS)



練習例: **取引年予測** (Stakcing)



まとめ

- 母集団の複雑さを捉える VS 有限のデータから推定する
  - 伝統的アプローチ: 研究者が経験 (ヤマカン) で設定
  - 教師付き学習: よりデータ主導
- 注意: 良い予測モデル $\neq$ 母集団の特徴理解に有益なモデル

母集団推定の応用

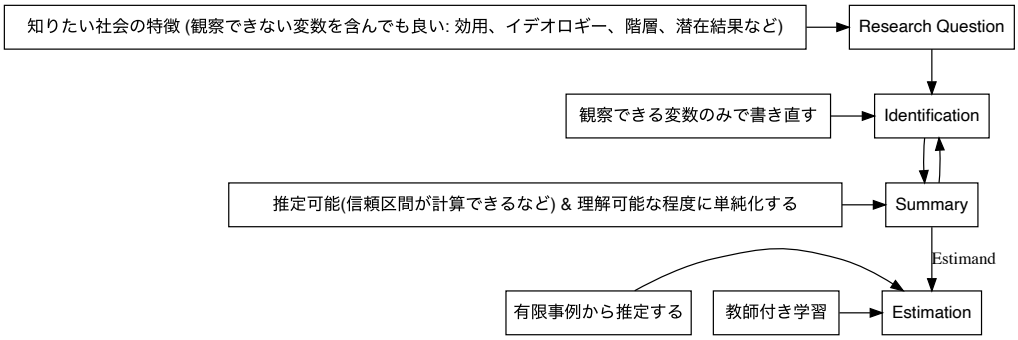
混乱

- 機械学習を何に応用できるか?
- 統計学の伝統的な用語 (最尤法、ベイズ推定など)、“因果推論の用語” (RCT, マッチングなど) などともに、応用上の混乱がみられる
  - できること/できないことが、不正確に喧伝される
    - 予測しかできない VS　なんでもできる
  - 過剰なナワバリ、縦割り的理解が散見される

記述/比較/因果研究

- 経済学における典型的な研究課題
  - 一般に社会の” 重要” な特徴の” 推論” を目指す
- Descriptive Comparison: (例) 同一学歴内男女間平均賃金格差
- Causal Inference: (例) 最低賃金の増加が雇用に与える平均効果
  - 研究プロジェクトの RoadMap の一部に貢献

実証分析の RoadMap



練習例: Research Question

- 2019-2021 年にかけて、中古マンションの市場価格はどのように変化したのか？
  - 市場とは? (一物一価を用いて**定義**)
  - マンションの属性ごとに細分化されている
- 全く同じ属性の物件間で 2019/2021 年比較したい
  - $Y =$ 取引価格, $D =$ 取引時点, $X =$ 物件属性
- 教師付き学習の出番なし?
  - 出番があったら怖い? (Ludwig and Mullainathan 2023)

練習例: Naive なアプローチ

- “以下のモデルを推定します”

$$Y_i = \beta_0 + \tau \times D_i + \beta_1 X_{1i} + .. + \underbrace{u_i}_{Normal}$$

- 何が仮定されているのか、不明確

練習例: Identification

- データから観察できる属性 $\neq$ 物件の全属性
- 例えば以下が仮定できれば OK

$$f_P(Y_i|D, X, \underbrace{U}_{\text{観察不可能}}) = f_P(Y_i|D, X)$$

- 背景（実験デザインなど）知識を用いて、より説得的な議論が可能な場合も

機械学習の応用?

- データは、データにない変数について何か語れるか？
- 例えば X の選択 (Gupta, Childers, and Lipton 2023)

練習例: Summary

- $f_P(Y|D,X)$ の推定は難しい
- 重要な特徴を捉え、人間が認知でき、推定できる程度に単純化
- 例
  - $\tau_P(X) = E_P[Y|2021,X] - E_P[Y|2019,X]$: 条件付き平均差
  - $\tau_{P,Average} = \int_X \omega(X) \times \tau_P(X) dX$: 周辺化条件付き平均差 ($\omega(X)$: 加重)

機械学習の応用?

- データ主導で認知可能なモデリングは可能だが (LASSO など)、推定誤差の評価が難しい
  - Kuchibhotla, Kolassa, and Kuffner (2021)

練習例: Estimation

- 教師付き学習の有力な応用先
  - 伝統的な推定方法が持つ、モデルに強く推定結果が依存してしまう性質を緩和
- Naive な応用は、教師付き学習の推定結果がもつ悪い性質 (収束が遅い) の影響をまともに受ける
- 教師付き学習を Nonparametric 推定に応用:
  - Data adaptive な推定法一般が持つ、収束速度が遅い性質を緩和

練習例: Partialling-out 推定

1. $E_P[Y|X], E_P[D|X]$ を機械学習などで推定 $\rightarrow g_Y(X), g_D(X)$ を得る
2. $Y - g_Y(X)$ を $D - g_D(X)$ で OLS 回帰
3. 回帰係数を、$\tau_{P,Average}$ の推定値として使用

- Chernozhukov et al. (2018)

Partialling-out 推定の利点

- $E_P[Y|X], E_P[D|X]$ の推定誤差の影響を、緩和できる
- OLS 推定の一般化

- $E_P[Y|X], E_P[D|X]$ を同じ線形モデルで回帰すれば OLS
- 一般に一致推定量にならない

練習例

```
Estimates and significance testing of the effect of target variables
  Estimate. Std. Error t value Pr(>|t|)
d  0.109884   0.003444   31.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

練習例: Naive plugin-in 推定

- $E_P[Y|D, X]$ を最尤法, ベイズ, 機械学習などで推定
- 推定値を $E_P[Y|D, X]$ に代入して、Estimand を計算
- $E_P[Y|D, X]$ の推定精度に決定的に依存する
  - 教師付き学習は収束が遅く、信頼区間も計算できない
  - Parametric Model は、一般に収束しない (一致性がない)

応用例

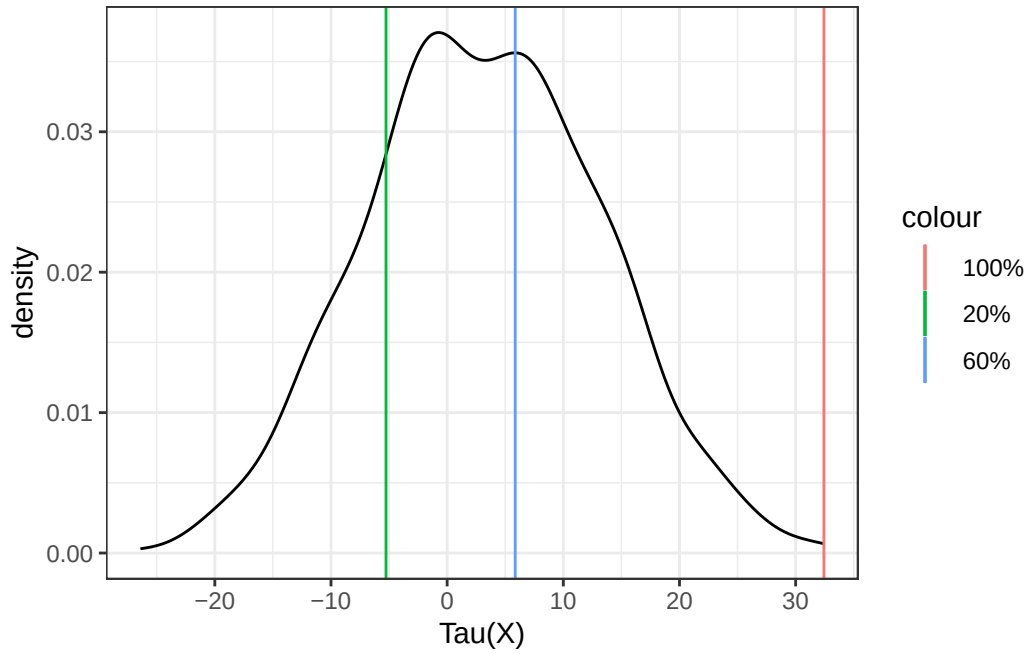
- Partialling-out は一般化できる
  - Efficient influence function を用いた収束速度の改善 (Hines et al. 2022; Ichimura and Newey 2022)
- ATE 推定, Conditional Average Treatment Effect 推定 (Semenova and Chernozhukov 2021; Kallus 2022; Wager and Athey 2018) , Mediation Analysis (Farbmacher et al. 2022; Díaz et al. 2021) , Sensitivity Analysis (Chernozhukov et al. 2022) などなど
- Estimation が改善したことで、活用できる Identification, Summary が増える!!!

因果研究 VS 比較研究

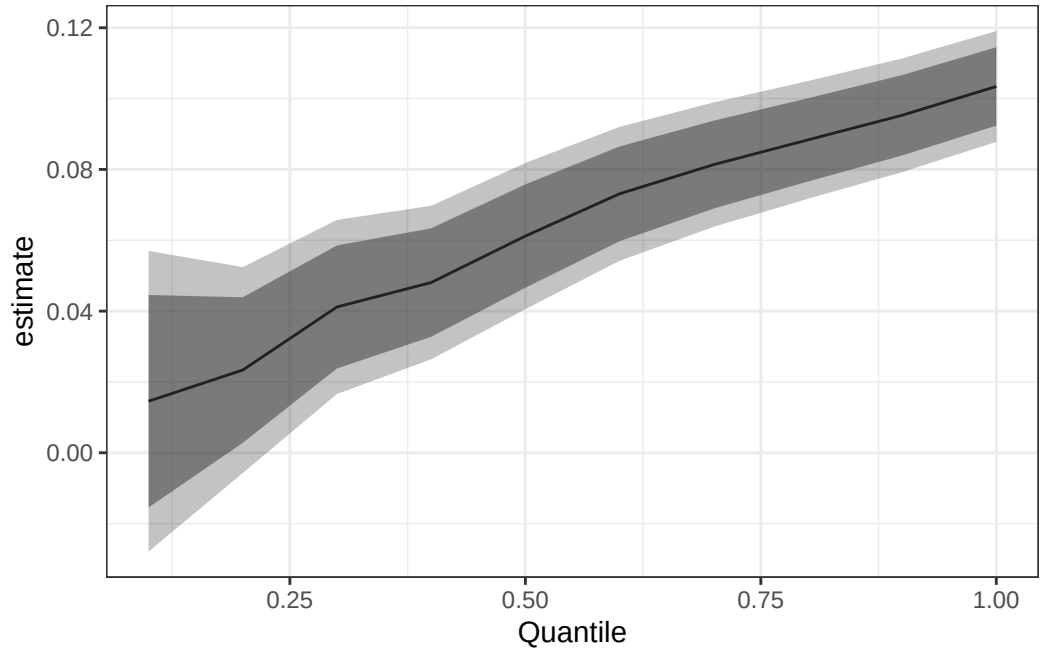
- 識別の議論は決定的に異なる: 観察できない変数への仮定、Interference への仮定
- Summary, Estimation は多くの場合よく似ている

– サンプリングに伴う不確実性を考慮する場合は、“同じ手法” を用いて、$\tau_{P,Average}, \tau_P(X)$ の推定を目指す。

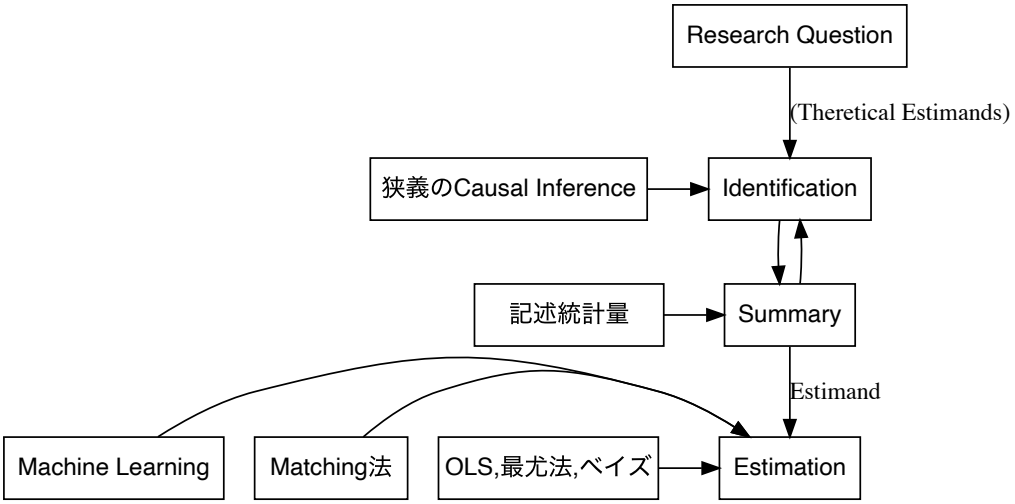
Treatment Effect Risk (Kallus 2022)



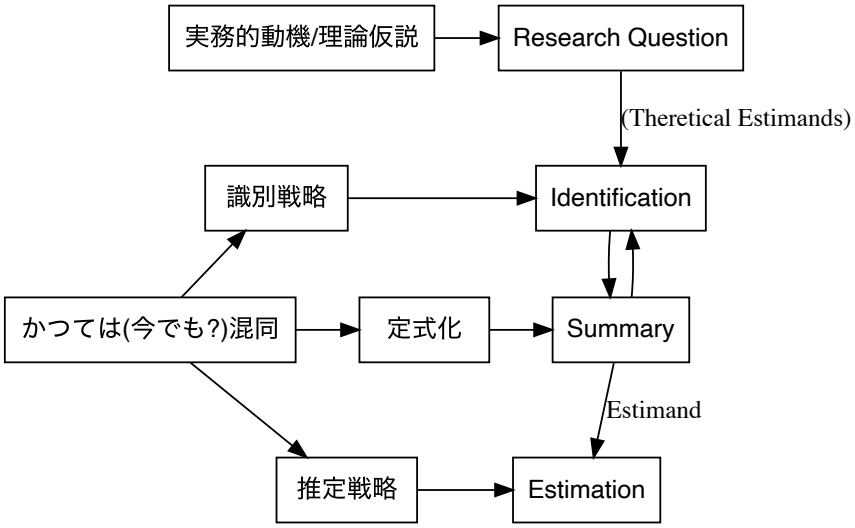
練習例



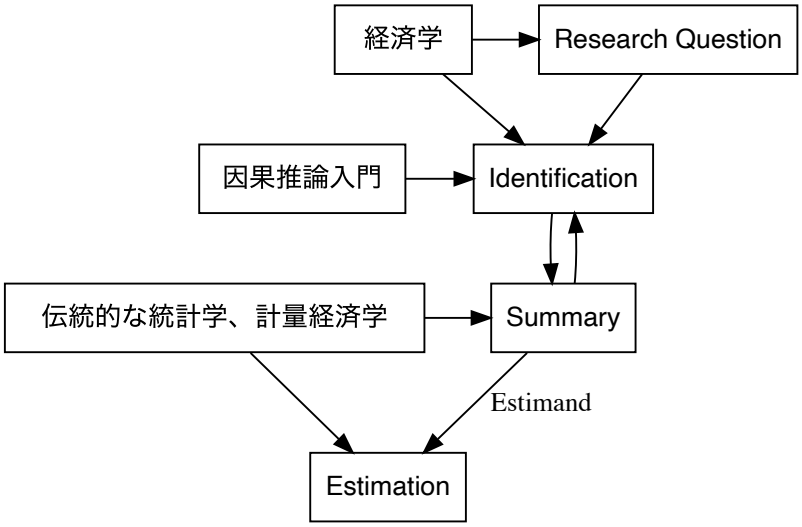
RoadMap: **手法**


Research Question
(Theretical Estimands)
狭義のCausal Inference
Identification
記述統計量
Summary
Estimand
Machine Learning
Matching法
OLS,最尤法,ベイズ
Estimation

RoadMap: **論文の章立て**



余談: 教科書


経済学
Research Question
因果推論入門
Identification
伝統的な統計学、計量経済学
Summary
Estimand
Estimation

まとめ

- 応用上は、以下の二つをしっかり区別することが重要

1. $\mathbb{E}[(\mathbb{E}[Y|X] - f(X))^2]$ を可能な限り削減する関数 $f(X)$ の推定
   - 予測問題と親和的
2. $\mathbb{E}[Y|D, Z]$ を特徴づける**研究者により事前に定義された**有限個のパラメータの推論
   - 信頼区間も得たい

予習

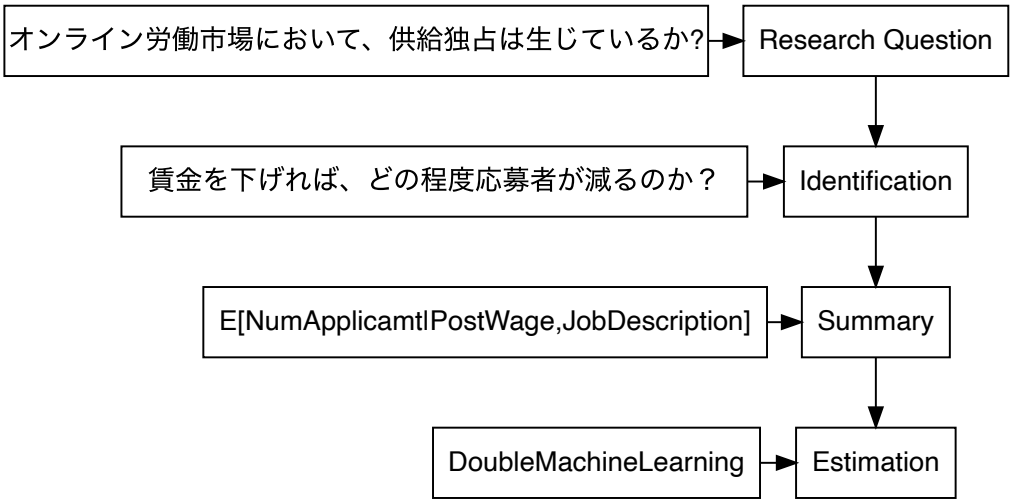
- Short Introduction: Daoud and Dubhashi (2020)
- TextBook:
  - Modern Business Analytics
  - Introduction to Statistical Learning
- Article
  - Brand, Zhou, and Xie (2022), Athey and Imbens (2019), Grimmer, Roberts, and Stewart (2021), Harding and Lamarche (2021)
  - Leist et al. (2022)

実例

Dube et al. (2020)

- Estimand: $E[NumApplicamt|PostWage, JobDescription]$
  - Identification: 価格モデルを前提とした場合、オンライン労働市場における供給独占 (Monopsony) を測定可能
  - Summary: 平均差に注目
- Estimation 上の問題: Web Scraping で収集した実際の求人データ
  - Job Description が非常に多次元 (テキスト情報も含む)

RoadMap: Dube et al. (2020)



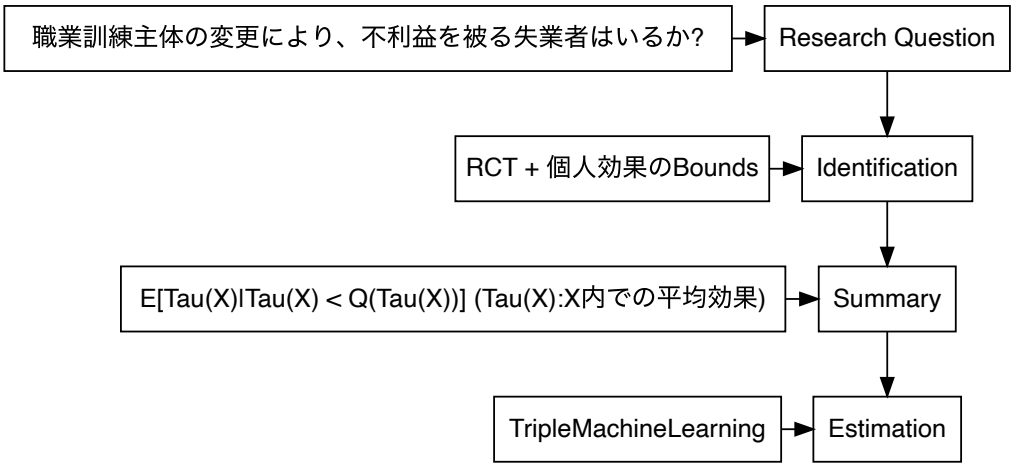
Behaghel, Crépon, and Gurgand (2014)

- 研究課題: 民間が行う新職業訓練 VS 政府が行う新職業訓練の因果効果
- Estimand: E[6ヶ月以内再就職|職業訓練の種類]
  - Identification: RCT
  - Summary: 平均差に注目
- Estimation 上の問題: ほぼない (平均差の推定で OK)!!!

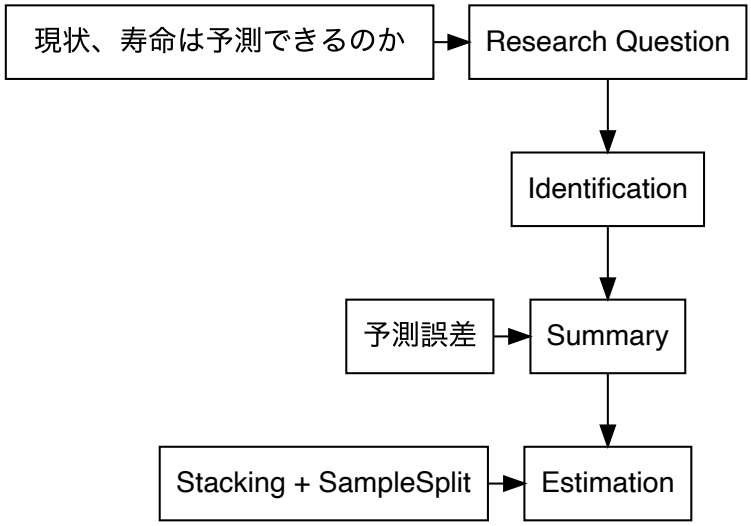
Kallus (2022)

- Estimand: $E_P[\tau_P(X)|\tau_P(X) \leq Q(\tau_P(X), q)]$
  - $\tau_P(X) = E_P[Y|D = 2021, X] - E_P[Y|D = 2019, X]$
  - $Q(\tau_P(X), q) =$ qth quantile
- 因果効果が低い（マイナス）のサブグループにおける平均効果
  - ” 全体” では正だとしても、負の影響を受けるグループがいるかもしれない

RoadMap: Kallus (2022)



RoadMap: Einav et al. (2018)



Reference

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Behaghel, Luc, Bruno Crépon, and Marc Gurgand. 2014. "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics* 6 (4): 142–74.

Brand, Jennie E, Xiang Zhou, and Yu Xie. 2022. "Recent Developments in Causal Inference and Machine Learning." *SocArXiv (Forthcoming in Annual Review of Sociology)*.

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, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. 2022. "Long Story Short: Omitted Variable Bias in Causal Machine Learning." National Bureau of Economic Research.

Daoud, Adel, and Devdatt Dubhashi. 2020. "Statistical Modeling: The Three Cultures." *arXiv Preprint arXiv:2012.04570*.

Díaz, Iván, Nima S Hejazi, Kara E Rudolph, and Mark J van Der Laan. 2021. "Nonparametric Efficient Causal Mediation with Intermediate Confounders." *Biometrika* 108 (3): 627–41.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in Online Labor Markets." *American Economic Review: Insights* 2 (1): 33–46.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. 2018. "Predictive Modeling of US Health Care Spending in Late Life." *Science* 360 (6396): 1462–65.

Farbmacher, Helmut, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. 2022. "Causal Mediation Analysis with Double Machine Learning." *The Econometrics Journal* 25 (2): 277–300.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24: 395–419.

Gupta, Shantanu, David Childers, and Zachary C Lipton. 2023. "Local Causal Discovery for Estimating Causal Effects." *arXiv Preprint arXiv:2302.08070*.

Harding, Matthew, and Carlos Lamarche. 2021. "Small Steps with Big Data: Using Machine Learning in Energy and Environmental Economics." *Annual Review of Resource Economics* 13.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician* 76 (3): 292–304.

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

Iskhakov, Fedor, John Rust, and Bertel Schjerning. 2020. "Machine Learning and Structural Econometrics: Contrasts and Synergies." *The Econometrics Journal* 23 (3): S81–124.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot. 2020. "An Adversarial Approach to Structural Estimation." *arXiv Preprint arXiv:2007.06169*.

Kallus, Nathan. 2022. "Treatment Effect Risk: Bounds and Inference." *2022 ACM Conference on Fairness, Accountability, and Transparency (Minor Revision in Management Science)*.

Koenecke, Allison, and Hal Varian. 2020. "Synthetic Data Generation for Economists." *arXiv Preprint arXiv:2011.01374*.

Kuchibhotla, Arun K., John E. Kolassa, and Todd A. Kuffner. 2021. "Post-Selection Inference." *Annual Review of Statistics and Its Application*.

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.

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Hypothesis Generation." National Bureau of Economic Research.

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

Nogueira, Ana Rita, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and Joao Gama. 2022. "Methods and Tools for Causal Discovery and Causal Inference." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12 (2): e1449.

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

Semenova, Vira, and Victor Chernozhukov. 2021. "Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions." *The Econometrics Journal* 24 (2): 264–89.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.