

Introduction

川田恵介

keisukekawata@iss.u-tokyo.ac.jp

2025-10-05

1 概要

1.1 講義の動機

- 「社会科学/医学 + 統計学 + 機械学習(AI)」に対して
 - ▶ 学術研究からも高い関心 (Athey & Imbens, 2019; Brand et al., 2023; Dell, 2025; Grimmer et al., 2021; Imbens, 2024)
 - 労働経済学においても同様 (Angrist & Frandsen, 2022)
 - ▶ 実務研究からも高い関心 (Amazon, Cyber Agent, Microsoft, Mizuho, Netflix, Uber)
- 特に「Semiparametric Theory + 教師付き学習」を用いた「(因果/格差を含む)比較研究」が大きく進展

1.2 労働経済学における例

- Vafa et al. (2025) : 詳細な就業履歴が同質でも、男女間賃金格差はどのくらい存在するのか?
 - ▶ “詳細な就業履歴が同質”の部分に機械学習を活用
- Kallus (2023) : 職業訓練の運営主体 (公営 VS 民営) の因果効果は、どのくらい異質なのか?
 - ▶ 異質性の探索に機械学習を活用

1.3 おすすめ予習/復習教材

- 教師付き学習 + 比較研究: Applied Causal Inference Powered by ML and AI
 - ▶ 特に[PREDICTION],[INFERENCE]のタグがついている章
- 補助として
 - ▶ 教師付き学習: Introduction to Statistical Learning

- 他は講義レポジトリに記載

1.4 講義全体の takeaway

- 機械学習を応用することで、データの持つ情報をさらに引き出す
- 労働経済学研究で重要視される、推定誤差の評価もしっかり行う
 - ▶ 伝統的な機械学習の弱点
 - ▶ Semiparametric Theory + 機械学習 (Debiased/Target Machine Learning) により、手法が確立

2 作業工程

2.1 議論の混乱

- 因果推論や機械学習など、データ分析法の急速な発展/拡大の一方で、混乱した議論も見られる
- 例: 以下の議論は”意味不明”
 - ▶ “機械学習/OLS を用いると、因果効果を推定できる/できない”
 - ▶ “機械学習を用いると、OLS よりも良いモデルが推定できる/できない”
- 整理された議論が理解/応用に必須

2.2 論点整理の方法

- 明確な分析工程の中に、緒概念/手法を位置付ける
- 大雑把な工程「研究目標について、データから回答する」を精緻化する
 - ▶ 経済理論、因果推論や機械学習などの導入に対応し、見直しが進んできた

2.3 本講義の作業工程

1. 研究目標 (Research question) = **社会**の特徴
2. 推定目標(Estimand) = 研究目標に対応する**母集団**の特徴
 - 以上から「データが回答すべきこと」が明確化
3. 推定値(Estimates) = データから計算された推定目標の推論値
4. コーディング → 発信…

2.4 識別/推定戦略

- 「研究目標、推定対象、推定値」を接続する”戦略”
- 識別 (Identification)

研究課題



推定目標

経済理論/因果推論などを用いた接続

- 推定 (Estimation/Inference/Learning)

推定目標



推定値

統計学/機械学習などを用いた接続

2.5 ゲームの因果効果: Egami et al. (2024)

- 研究課題: ゲームは、健康状態を（因果的に）悪化させるのか？
 - ▶ 識別: コロナ下で生じた「ゲーム機の過小供給」と、それに対応するための「購入券くじ」を社会実験として活用
- 推定対象: くじ引き参加者について、居住地を含む背景属性を均一化した後に、精神的健康状態の比較
 - ▶ 推定: Debiased Machine Learning を活用
- 推定値: 点推定量 + 信頼区間 (精神的健康状態は平均的に改善する)

2.6 まとめ

- 分析工程の全体像をまず把握することが重要
 - ▶ 伝統的な教科書は、しばしば、特定のパートに焦点を当ててきた
 - “因果推論の教科書” → 識別に焦点
 - “統計学/機械学習の教科書” → 推定に焦点
 - “経済学の教科書” → 研究目標/識別に焦点
- 本講義の主目的は、「機械学習を活用した推定」だが、識別の議論にもしっかり言及

3 賃金予測研究

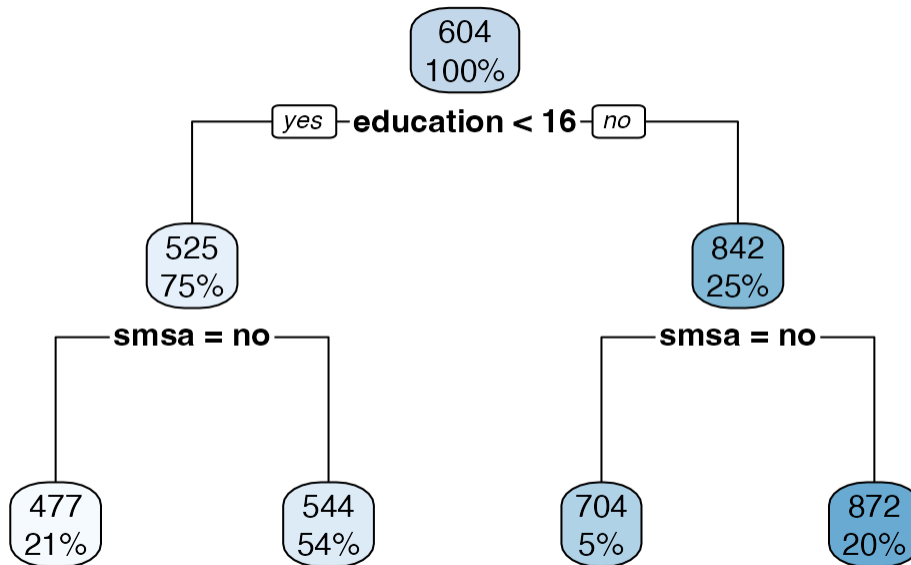
3.1 実習例: 賃金予測モデル

- 研究対象: 労働者の属性 X から賃金 Y を予測する
 - ▶ X = education, experience, age, ethnicity, region, gender, occupation, sector, union, married
- 推定対象: 母平均 $EY | X$
 - ▶ OLS/LASSO/Boosting/DeepLearning…を用いて推定の柔軟性と精度を両立を目指す
- 推定値: CPS1988 から計算

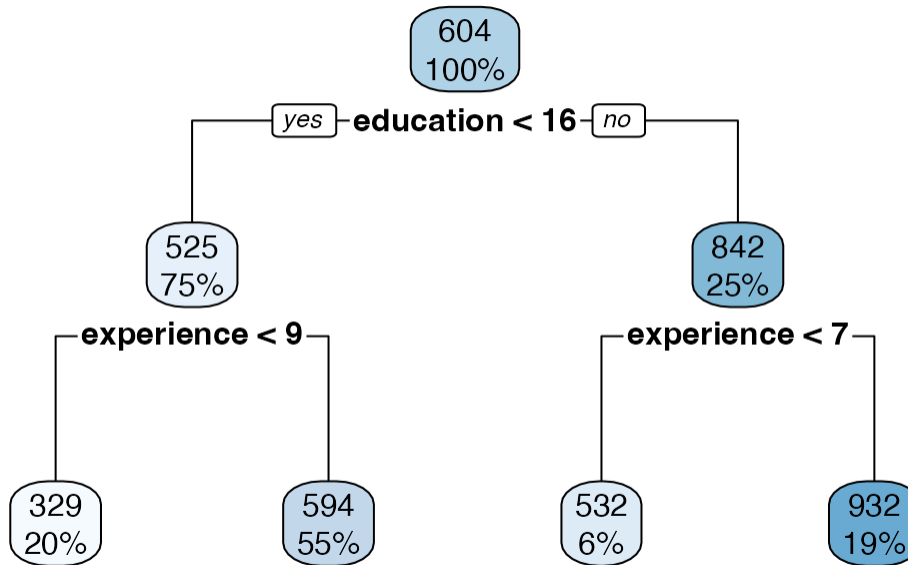
3.2 推定戦略: 回帰木モデルの推定

- X に応じて、サブグループに分け、サブグループごとの平均価格を予測値とする
- 問題はサブグループをどのように定義するか
 - ▶ 研究者主導: 事前知識や”長年のかん”などにに基づき決定
 - ▶ データ主導: データに最も適合するように決定

3.3 実例: 研究者主導のモデル



3.4 実例: データ主導のモデル



3.5 データへの適合度

- 決定係数は、
 - ▶ 研究者主導のモデル = 0.0997606
 - ▶ データ主導のモデル = 0.1749204
- 基本的に、人間行動や社会的な相互作用の結果として決まる変数 (Social outcome) についての予測精度は低い (Narayanan & Kapoor, 2024)
 - ▶ 予測以外への活用が現実的

3.6 寿命予測: Einav et al. (2018)

- 研究対象: 1年後の生存率を予測できるか?
 - ▶ 予測対象とデータは同じ母集団からサンプリングされている
- 推定対象: 平均生存率
 - ▶ 複数の予測モデルを組み合わせる Stacking
- 推定値: 医療機関領収書データを用いた予測モデルは、生存率を上手く予測できない

3.7 Takeaway: 教師付き学習

- 本講義で紹介する機械学習: 「教師付き学習における回帰/分類問題」に関する手法群
- 経済学の視点から、ざっくり言うと

- ▶ 「flexible nonparametric estimation (Vasilis Syrgkanis)」
- ▶ 「new generation of nonparametric estimators that seem to do well in some empirical applications (Jiafeng Chen)」

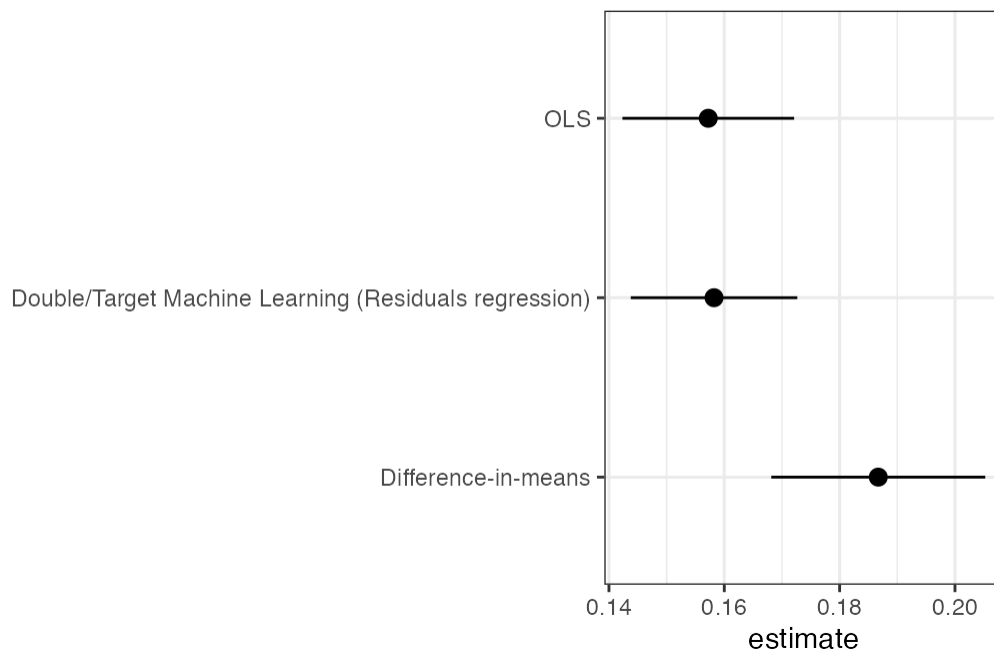
4 賃金格差研究

4.1 実習例: 居住地間賃金格差改装の因果効果予測

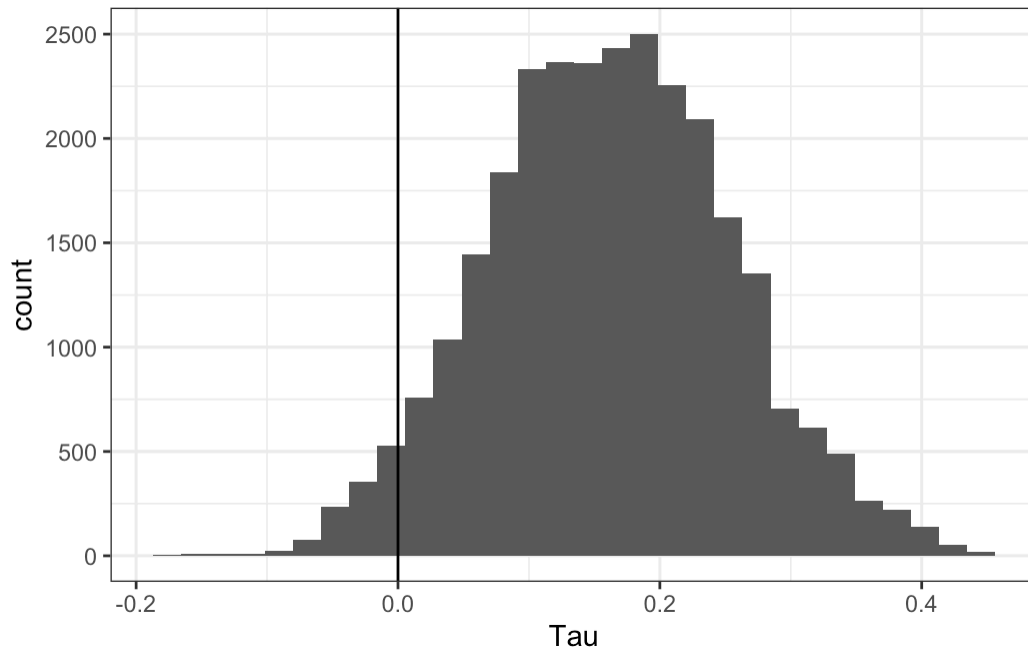
- 研究課題: 都会と田舎で賃金格差はどのくらいあるか?
- 推定対象: 対数賃金の平均差 $\tau(X) = E[Y \mid \text{Urban}, X] - E[Y \mid \text{Rural}, X]$
 - ▶ Double/Target Machine Learning + Causal Forest (Athey et al., 2019)
- 推定値: CPS1988 から計算

4.2 平均差

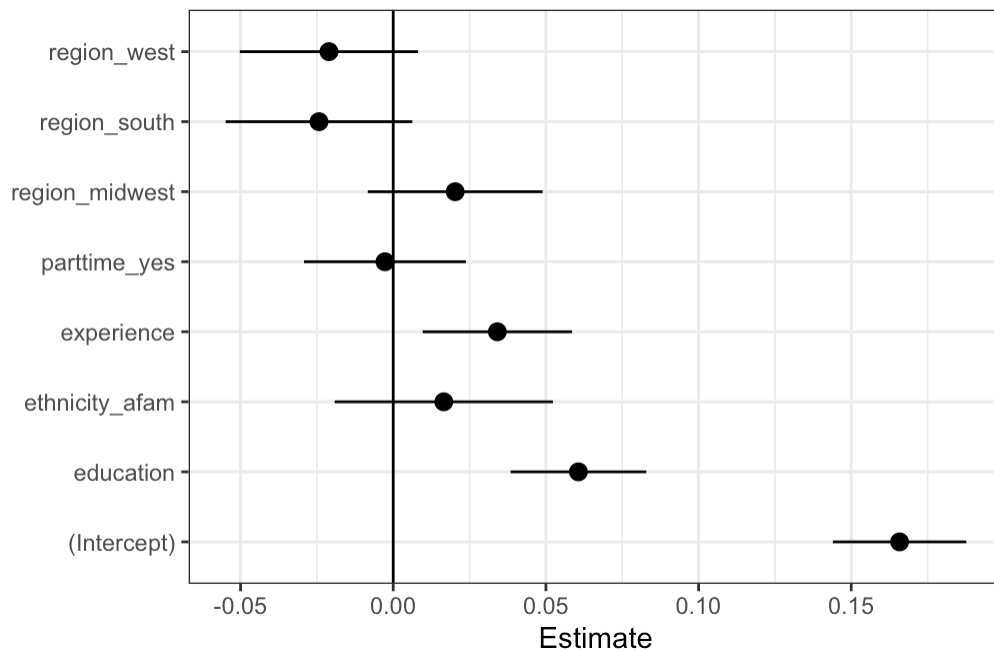
sample fold 1/2
sample fold 2/2



4.3 異質性: $\tau(X)$ の分布



4.4 異質性: $\tau(X)$ の線型近似



4.5 まとめ

- 機械学習等を活用した予測分析は、注目され、多くの応用が存在する

- ▶ ただし、Social outcome の予測は一般に困難
- ▶ Narayanan & Kapoor (2024) : 画像やテキストへの応用の”成功”イメージを、安易に適用すべきではない
 - 社会分析への便益は限定的???
- 応用には、直面している意思決定問題から、研究課題と推定対象を再考する必要がある

5 次回までに

5.1 講義の方針

- 応用に向けた手法のコンセプトと実装方法の紹介に注力
 - ▶ 証明は省略 (必要に応じてアイデアのみ紹介)
- 応用時に質問が集中しがちな既存の手法との接続/比較に注力
 - ▶ 線形モデルの OLS/最尤法/ベイズによる推定、傾向スコア/Balancing Weight
- 「手法の動機と研究課題との relevance を説明しながら、実際の研究に応用する」ことを目指す

5.2 講義の方針: 既習者向け

- 教師付き学習とモーメント推定を、大表本性質を確保しながら融合し、柔軟に活用する方法を提示
 - ▶ 鍵となる性質: Neyman’s orthogonality condition
 - Nuisance 関数の推定値の収束速度が多少遅くても、Target parameter の推定値は漸近的に正規分布に従う

5.3 課題

- 3回のレポート
 - ▶ 理解を確認するために、講義中にミニクイズも行うが成績には反映しない
- 受講者は次回までに R/Rstudio の設定/パッケージ(tidyverse/glmnet)のダウンロード/データのダウンロード/プロジェクトフォルダの作成/データの格納まで行うこと
- 講義資料やアーカイブは、すべてのレポジトリ (<https://github.com/tetokawata/TargetML>) から入手可能

5.4 Rstudio

- 講義では R (4.5.1 以降) + Rstudio (2025.09.1 以降) をサポート
 - ▶ tidyverse のインストール
- 他の IDE や Python を活用しても OK

- ▶ Julia などの他言語は、(教員が読めないので)、課題では使用しないこと

5.5 Reference

Bibliography

- Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1), S97–S140.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, 47(2), 1148–1178.
- Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49, 81–110.
- Dell, M. (2025). Deep learning for economists. *Journal of Economic Literature*, 63(1), 5–58.
- Egami, H., Rahman, M. S., Yamamoto, T., Egami, C., & Wakabayashi, T. (2024). Causal effect of video gaming on mental well-being in Japan 2020–2022. *Nature Human Behaviour*, 1–14.
- Einav, L., Finkelstein, A., Mullainathan, S., & Obermeyer, Z. (2018). Predictive modeling of US health care spending in late life. *Science*, 360(6396), 1462–1465.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1), 395–419.
- Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.
- Kallus, N. (2023). Treatment effect risk: Bounds and inference. *Management Science*, 69(8), 4579–4590.
- Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.
- Vafa, K., Athey, S., & Blei, D. M. (2025). Estimating wage disparities using foundation models. *Proceedings of the National Academy of Sciences*, 122(22), e2427298122.