Introduction

川田恵介

Table of contents

1 概要

- 機械学習初学者向けに、コンセプトの紹介と社会分析への応用を紹介

1. 教師付き学習の紹介と予測への応用
2. モーメント推定への応用: 格差や因果効果推定に有益

- 随時、R による実装と復習

1.1 おすすめ無料 (英語) 教材

- Supervised Learning: Introduction to Statistical Learning
- Supervised Learning + 比較研究 (“因果効果”): Applied Causal Inference Powered by ML and AI
- R
  - Tidyverse Style Guide
  - R for Data Science

1.2 将来の学習

- 坂口さんの講義
- 包括的な入門書
  - ビジネスデータサイエンスの教科書
  - Computer Age Statistical Inference
- より理論的
  - Understanding Machine Learning: From Theory to Algorithms
  - The Elements of Statistical Learning

1.3 機械学習とは?

- 統計学とその派生 (計量経済学など) とは異なるルーツを有する
  - 主として、**予測**のために **Non-(Over-) parametric model** を推定する

        * 計算機科学、データ主導の AI 開発

- Moment 推定との融合が進む
  - 社会の**特徴把握 (記述/要約)** のために、**Semi-parametric model** を推定する

1.4 導入の動機

- データ分析への” 研究者の介入 (Researcher degree of freedom)“を**適切に**減らす
- データ分析 = 前提 (観察したデータの特徴、具体的な分析対象、データの収集法) から結論を得る
  - 前提の中に、研究者によるモデル定式化が含まれる
    * 根拠/透明性が低い仮定への、極力避けたい
      · 機械学習の活用が有力
  - 他の部分については、研究者の介入を前提

1.5 講義の動機

- 機械学習 + “因果効果” に対して、学際的な関心 (Athey and Imbens 2019; Grimmer, Roberts, and Stewart 2021; Brand, Zhou, and Xie 2023)
- 実務家からの関心も高い (例: microsoft, netflix, サイバーエージェント)
  - モーメント法との融合は、有力なアプローチとして、必ず紹介されている
- 研究/実務機関ともに、機械学習を生かした研究経験は、関心を持たれやすい?

1.6 講義の予定

- 概観: OLS により推定した LinearModel の予測モデル/記述モデルとしての解釈とその拡張として、LASSO と Double Selection を紹介
- 予測の発展: Random forest, Boosting, Stacking など
- 記述の発展: 平均効果、効果の異質性等

1.7 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ィアのみ紹介)

- 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

1.8 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 ohthgonality 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

1.9 課題

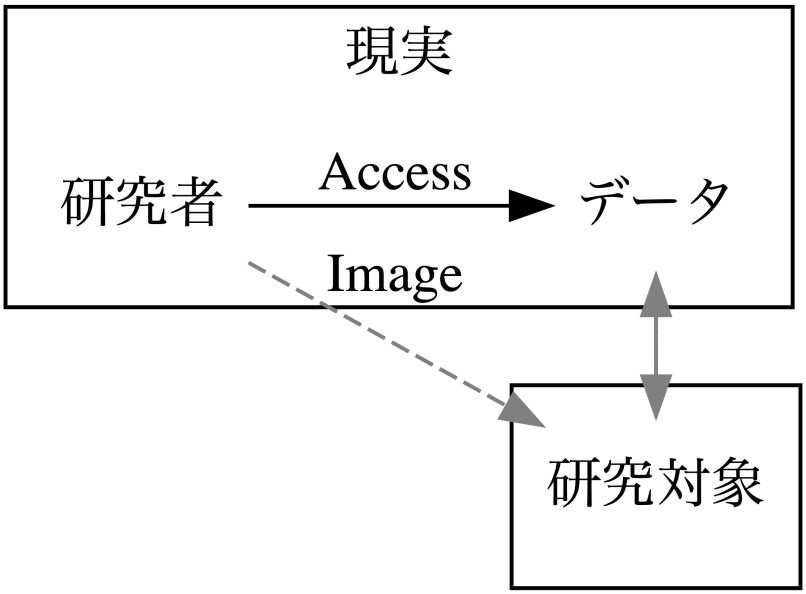
- 3回のレポート
  - 理解を確認するために、講義中にミニクイズも行うが成績には反映しない
- 受講者は次回までに **R/Rstudio の設定/パッケージ (tidyverse/glmnet) のダウンロード/データのダウンロード/プロジェクトフォルダの作成/データの格納**まで行うこと
- 講義資料やアーカイブは、すべてのレポジトリ (https://github.com/tetokawata/TargetML) から入手可能

2 概念整理: 分析目標

- 分析目標を明確に理解し、イメージし続けることが分析を進める上で有益
  - 本講義では、分析目標を大きく、予測と特徴の記述に分類

2.1 Naive なイメージ

- 手元にデータがあり、そのデータを使って研究対象 (例: 日本社会) への示唆を得る



2.2 分析目標

- もう少し具体的な分析目標を定める必要がある: 本講義では以下に焦点
  - 予測 (Prediction)
  - 特徴の記述 (Summary/Description)
- **明確に区別すること**が今後重要
  - 現状は、混同されがち

2.3 例

- Einav et al. (2018) (予測): 患者の一年後の死亡を予測できるか？
  - 動機: 終末期医療論争への含意
  - 結論: 予測精度は低い
- Athey et al. (2023) (記述): 失職の影響は、どのような層で大きいか?
  - 動機: 所得 - 失職の関係性を、顕著な異質性も含めて理解
  - 結論: 日常業務 - 高齢者において特に顕著な影響

3 概念整理: モデル

- データからモデルを推定し、分析目標に応える
  - 機械学習/統計学: 推定方法 (Algorithm) を提供
- 分析目標に応じて、好ましいモデルは異なる (予測モデル/記述モデル)
  - かつては (今も?)、「真の一つのモデルを推定する」、という発想が強いが、本講義では不採用

3.1 モデル = 模型

- モデルは、現実に比べて、単純すぎる" からこそ" 有益
  - データは複雑すぎ、直接理解/活用できないので、データの持つ情報を、要約 (=モデル化) する必要がある
- 例: 不動産取引価格と築年数と広さのモデル
- $\text{Price} = 13.6 + 0.8 \times \text{Size} - 0.6 \times \text{Tenure}$

3.2 モデルの活用目的

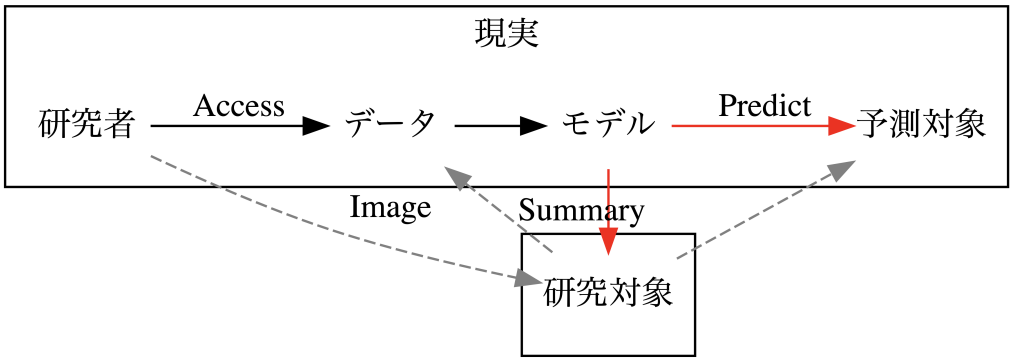
- 伝統的な方法では統計モデルと呼ばれ、予測/記述、双方に使用されてきた
- 予測モデルとしての活用: $\{\ \text{Size} = 10, \text{Tenure} = 5\ \}$ の予測価格は、$18.6 = 13.6 + 0.8 \times 10 - 0.6 \times 5$
- 記述モデルとしての活用
  - Price と Tenure の記述モデル上の関係性 =-0.6
  - "築年数が古くなると、価格が下がる"

3.3 モデルの活用目的

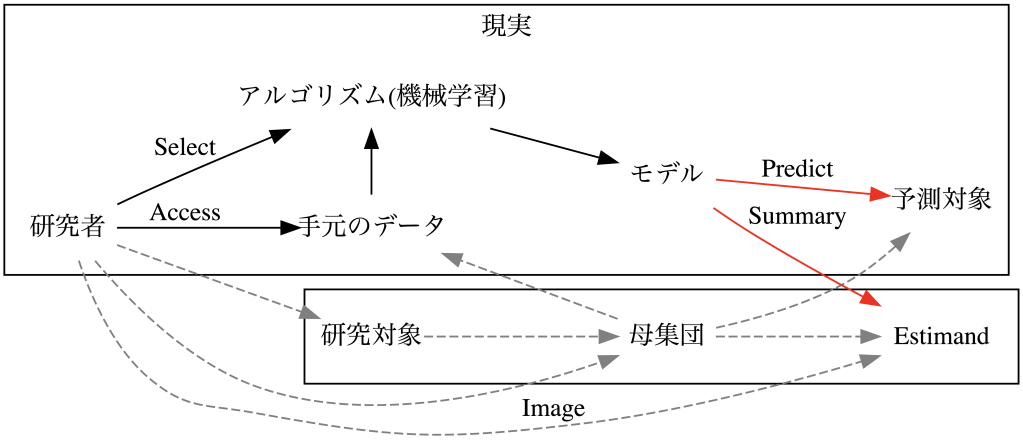
- モデルの目的は、**大きく２種類**あることに注意
  - 人間が理解できるようにする
    - 記述への活用において相対的に重要
      - 研究者によるモデル作りに比較優位
  - 大量の変数に起因する、モデルのデータへの過剰な依存度を減らす
    - 多くの変数を活用する必要がある、予測において相対的に重要

· データによるモデル作りに比較優位

3.4 まとめ



3.5 最終イメージ



4 Gallary

- 2022 年の東京 23 区で取引された中古マンションデータを使用した分析例
  - 本講義の中で、学ぶ分析方法

4.1 研究目的: 取引価格の予測

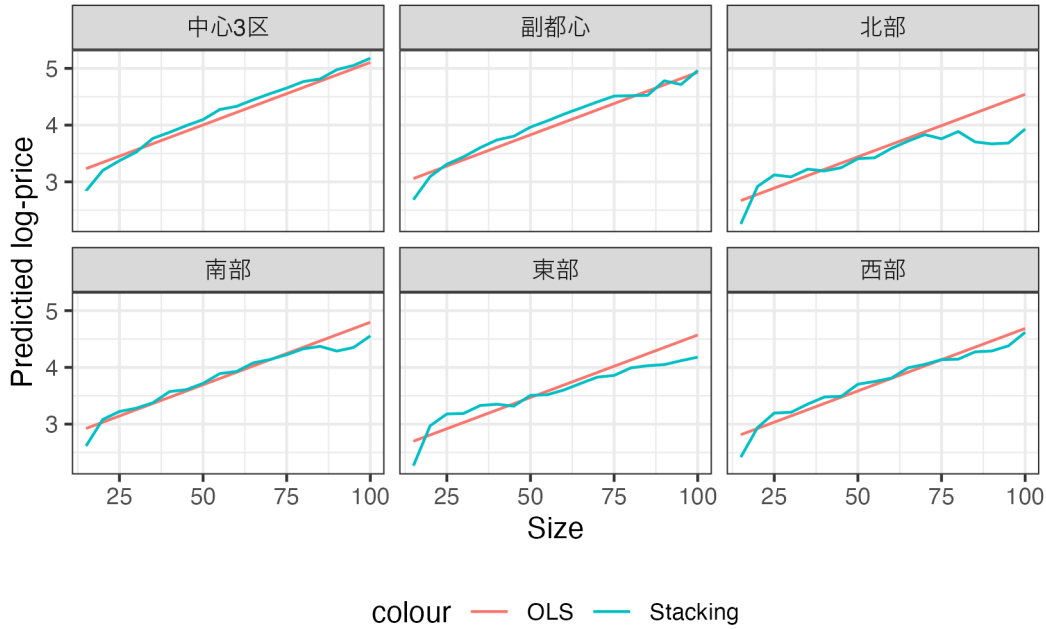
- 取引価格の予測モデル: $X = [$ 部屋の広さ (Size), 立地 (District)$] \rightarrow$ 取引価格 (Price)
  - Linear Model

$$Price \sim \beta_0 + \beta_1 Size + \beta_D dummy(District)$$

を OLS で推定

– OLS/RandomForest/LASSO/Boosting を Stacking

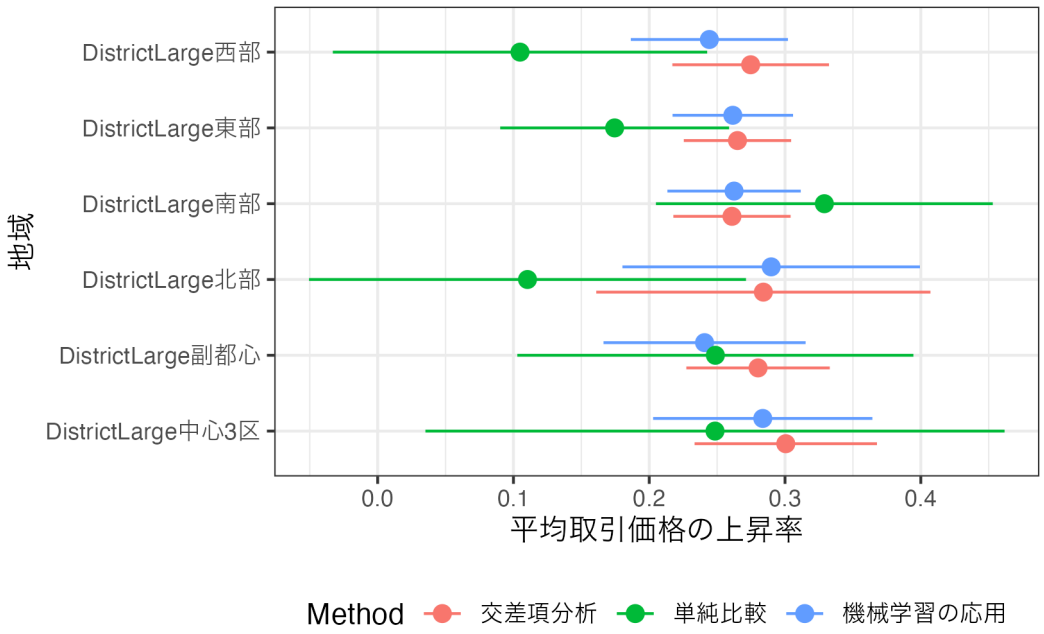
4.2 予測研究



4.3 研究目的: 特徴の理解

- 各地域内で、不動産価格は 2017-2022 年にかけて、どの程度変化したのか?
  - 単純な地域内の平均変化: 機械学習は" 不要"
  - 同一物件 (同じ取引物件の部屋の広さ/駅からの距離/最寄駅/築年数/容積率) についてどの程度変化したか?
    - 2017-2022 年にかけて、価格以外についても取引物件の性質は変化しており、この影響を除外する
- 予測モデル" のみ" を用いると、推定の誤差 (信頼区間) を評価できない (Chernozhukov et al. 2018)

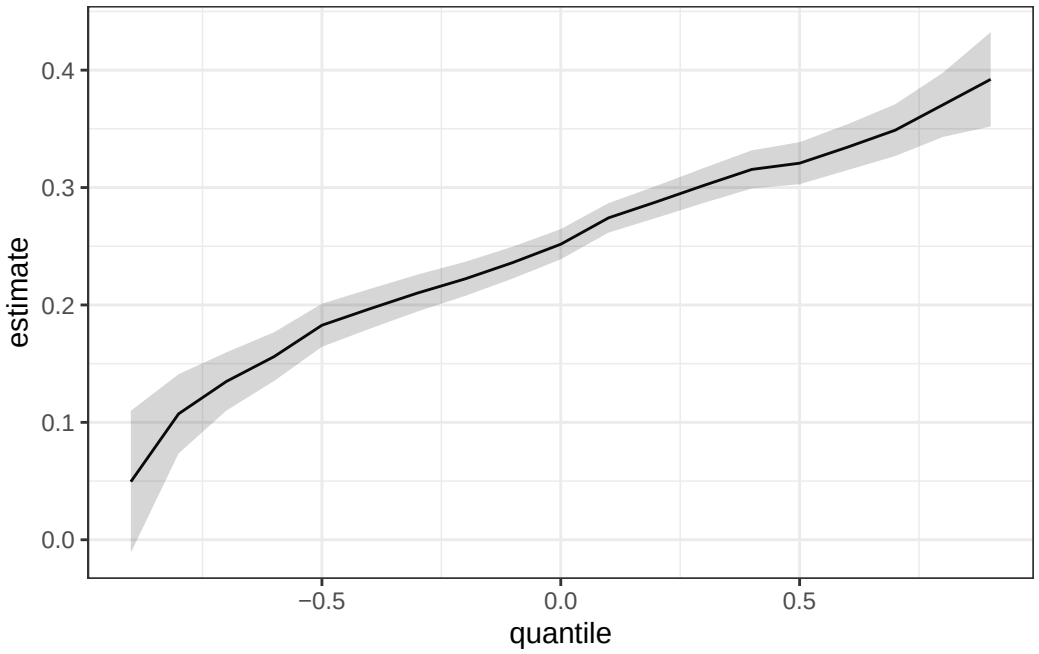
4.4 BLP



4.5 研究目的: 特徴の理解

- 背景変数 (立地区/同じ取引物件の部屋の広さ/駅からの距離/最寄駅/築年数/容積率) のすべての組み合わせについて、異質性を評価
- 変化が上位 $q\%(quantile)$ あるいは下位 $q\%(-quantile)$ 内での平均的な変化を推定
  - Kallus (2023)

4.6 Risk



Reference

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

Athey, Susan, Lisa K Simon, Oskar N Skans, Johan Vikstrom, and Yaroslav Yakymovych. 2023. "The Heterogeneous Earnings Impact of Job Loss Across Workers, Establishments, and Markets." *arXiv Preprint arXiv:2307.06684*.

Brand, Jennie E, Xiang Zhou, and Yu Xie. 2023. "Recent Developments in Causal Inference and Machine Learning." *Annual Review of Sociology* 49: 81–110.

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." Oxford University Press Oxford, UK.

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.

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.

Kallus, Nathan. 2023. "Treatment Effect Risk: Bounds and Inference." *Management Science* 69 (8): 4579–90.