

まとめと因果推論への接続

機械学習

川田恵介

2025-12-10

ここまで

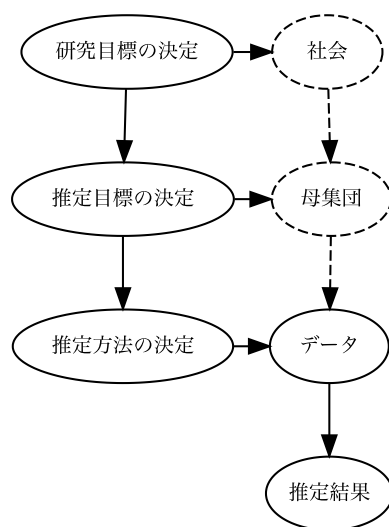
論点整理の重要性

- データ分析は、学際的・産学連携に大きく発展
 - 大量の有益な概念、手法が提案される (機械学習、因果推論など)
 - 多くの選択肢から適したものを選ばなければならない
- 論点整理が重要

例: 旅行計画

- 宿泊先と移動方法を決めなければならない
 - 整理された議論: 「ホテル VS コテージ」と「レンタカー VS 電車」をまず議論し、全体として不都合があれば見直す
 - 混乱を産みがちな議論: 「ホテル VS 電車」をいきなり議論する
- 論点整理について、絶対的なルールがあるわけではないが、“おすすめ”の方法は存在する

データ分析の論点



推定目標

- データから、その背後にある**母集団** (≒ 現実の市場や社会) の「有益な」特徴を把握する
 - 取引価格を予測したい → 取引価格の**母平均**

$$E[Y \mid X]$$

- 改装/未改装で取引価格を比較したい → 取引価格の**母平均の差**

$$E[Y \mid D = 1, X] - E[Y \mid D = 0, X]$$

伝統的なアプローチの課題

- X の数が多いと、**過剰適合問題**が深刻になる
- 伝統的には、以下を人間が行った後に、OLS などで推定する
 - 重要そうではない変数を除外する
 - モデルをなんとなく単純化する
- 機械学習は、上記を部分的に補助する

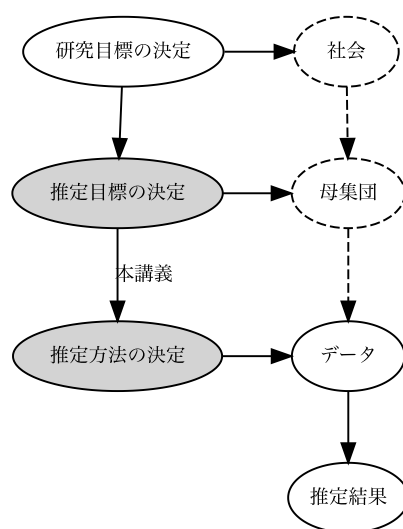
議論した手法の整理

推定目標	推定方法	推定誤差
平均	OLS	信頼区間
平均	LASSO/Random Forest	?
平均の差	OLS	信頼区間
平均の差	二重選択法/予測誤差の回帰	信頼区間
平均の差	Causal Forest	?

予測誤差と推定誤差の区別

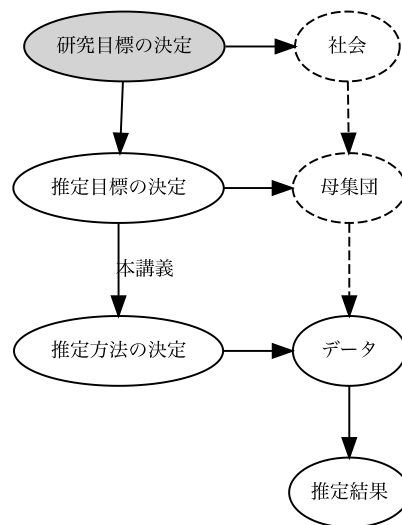
- サンプル分割**を用いれば、Random Forest 等についても、予測誤差は測定できる
 - 母集団の特定の特徴についての推定誤差は**難しい**
- 予測誤差の例: 価格予測モデル (立地、広さ) を、あるデータに適用した際に、どのくらい予測を外すのか?
- 推定誤差の例: (練馬区, 50平米) の物件の、母集団における平均取引価格

データ分析の論点



研究目標: 状況 VS 仕組みの把握

研究の流れ



研究目標: 状況の把握

- ここまでの議論は、「現状や過去の社会、かつ、データから把握できる変数」についてのみ言及する研究目標に活用できる
 - **予測研究:** ある物件を、今日売りに出したら、いくらで売れるのか?
 - **比較研究:** 過去五年間で、不動産価格は、どのように変化したのか?
- どちらも変数の「表面的な」状況の把握がキモ

研究目標: 仕組みの把握

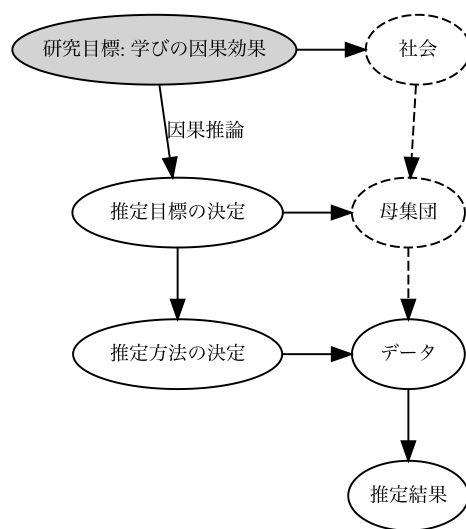
- 経済学やビジネス、政策研究では、「社会の仕組み (原因、構造、因果関係)」が研究目標になることも多い
- データから観察できない概念や関係性、仮想的な社会に言及する
 - 例: **なぜ東京 23 区の不動産価格は上昇したのか?**
 - * 経済学だけでなく一般的な解説記事などでも、しばしば**需要と供給**という”概念”によって説明される
 - * **需要と供給**などは、データから観察できない

研究目標: 因果効果

因果効果研究

- 「機械学習を学ぶことによって、30 歳時点での所得はどの程度変化するのか?」
- しばしば「(機械学習を) 学びの因果効果は?」と言い換えられる
 - 仕組みについての研究課題
 - * 「因果効果」をデータから観察することができない
- 因果推論: 因果効果と推定目標を接続する議論

研究の流れ



因果効果の重要性

- 非常に学際的な分野であり、実務家からの関心も高い
 - actionable (実用的な) 含意を得られる
 - * 各「action」がもたらす帰結は、意思決定の基盤
 - 社会の仕組みを理解する基盤
 - * 大学卒/高校卒の賃金格差の原因は?
 - ・ 有益な研究目標は、「もし入学試験を廃止し、完全ランダム選別にしたい場合に、賃金格差はどのように変化するか」

注意: 予測モデルの誤用

- 因果効果推定に、予測モデル (AI) を安直に応用すべきではない
- 例: $Y = \text{築年数}$ 、 $D = \text{改装の有無}$
 - $\beta_0 + \beta_1 \times D$ は容易に推定できる、**現実の築年数**を予測するモデル (AI)

注意: 予測モデルの誤用

```
lm(Tenure ~ Reform, data)
```

Call:

```
lm(formula = Tenure ~ Reform, data = data)
```

Coefficients:

(Intercept)	Reform
17.92	9.54

- 改築によって、築年数が平均的に 9.54 年のびる?
 - 正しい主張は、改築した物件の方が、築年数が平均的に 9.54 年ながい

仮想的な社会による定義

- まずは因果効果とは何か、正確に定義する必要がある
 - 非常に難しい問題

- 多くの議論があるが、共通するのは、「仮想的な社会」をイメージすることで、因果効果を定義する
- 例: “全員が機械学習を学んだ社会” と “誰も機械学習を学ばなかった社会”
 - ここまでのモデルでは、仮想的な社会で何が起きるのかわからないことに注意

平均因果効果

- Y_1 : 全員が学んだ社会における所得
- Y_0 : 全員が学ばなかった社会における所得
 - 個別効果 = $Y_1 - Y_0$
 - 平均効果 = $E[Y_1] - E[Y_0]$
- どのように Y_1, Y_0 を推測するか?
 - 予測モデルが予測するのは、現実の所得 Y

現実世界と仮想世界の関係性

$$\begin{aligned}
 & E[Y_1] - E[Y | D = 1] \\
 & = \underbrace{E[Y_1] - E[Y_1 | D = 1]}_{\text{背景の違い}} \\
 & \quad + \underbrace{E[Y_1 | D = 1] - E[Y | D = 1]}_{\text{波及}}
 \end{aligned}$$

波及

- $E[Y_1 | D = 1] - E[Y | D = 1]$ = 現実に機械学習を学んだ人について、「他の人も学んだ仮想世界と現実世界の差」
- どちらの世界でも本人の状況は変わらない
 - 周りの状況が違うことによる変化
- 例: 機械学習の知識を持った労働者の供給が増えるので、“skill premium” が低下する

$$E[Y_1 | D = 1] - E[Y | D = 1] < 0$$

背景の違い

- $E[Y_1] - E[Y_1 | D = 1]$ = 仮想世界における、全体平均と現実世界で学んでいる人の平均所得の差
 - 現実世界の”背景の違い”に起因する
- 例: 機械学習を学ぼうと思う人は、「学習意欲が高く」、他の skill も身につけやすい

$$E[Y_1] - E[Y_1 | D = 1] < 0$$

現実世界と仮想世界のギャップの例

-

$$E[Y_1] - E[Y \mid D = 1]$$

-

$$\begin{aligned} &= \underbrace{E[Y_1] - E[Y_1 \mid D = 1]}_{\text{背景の違い}(<0)} \\ &\quad + \underbrace{E[Y_1 \mid D = 1] - E[Y \mid D = 1]}_{\text{波及}(<0)} < 0 \end{aligned}$$

- $E[Y_1] - E[Y \mid D = 1] < 0$ であり、現実の平均所得は、仮想世界よりも高い

識別: 推定目標と研究目標の接続

統計的因果推論

- 識別: 研究目標と母集団上の推定対象を接続する
- 推定方法: 推定対象とデータから計算した推定値を接続する
- 統計的因果推論: 研究目標となる因果効果と推定対象を接続する
 - 識別の一種

よくある識別

- 波及効果は、“0”であると仮定する (SUTVA, No interference)
 - 経済学の伝統的な関心であり、今でも因果推論の hot issues
- 背景の違いは、本講義で学んだバランス後の比較 + ランダム化対照実験ないし自然実験のデータを使うことで対処

比較分析との違い

- 本講義では、データから観察できる属性 X をバランスする方法を紹介
- 因果推論では、観察で観察できない属性もバランスしたい
 - 例: 「学習意欲」をバランスさせたいが、データから観察することが難しい
- ランダム化対照実験ないし自然実験のデータで対処する

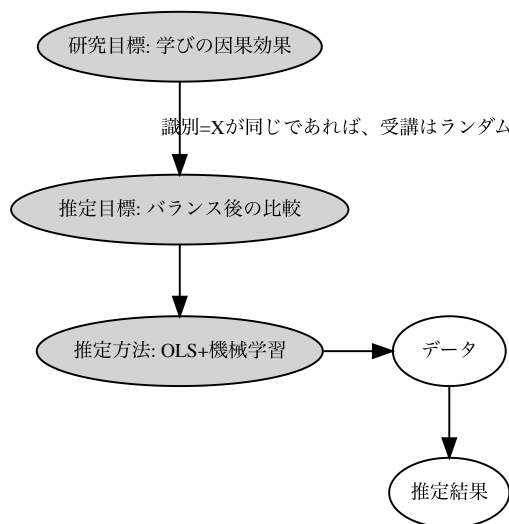
ランダム化対照実験の例

- $D = \text{SNS}$ は、人々の知識や健康状態にどのような影響を与えるのか? (Allcott, Gentzkow, Wittenbrink, et al. 2025; Allcott, Gentzkow, Levy, et al. 2025; Allcott et al. 2024)
 - 大統領選直前の米国において、実験への参加に同意したユーザーの中から、ランダムに一部を選び、SNS (Instagram と Facebook) の利用を一定期間停止する
 - ランダムに選ばれているので、背景は観察できない属性も含めて、偏っていない
 - 偶然の偏りは存在しうるが、信頼区間などで考慮することができる

自然実験の例

- $D = \text{ゲーム機}$ は、人々の健康状態にどのような影響を与えるのか? (Egami et al. 2024)
 - コロナ下で生じたゲーム機への超過需要と小売店によるくじ引きの活用を、「自然に生じた」実験として活用
 - くじ引きで決まるので、**一部を除き**、背景は観察できない属性も含めて偏っていない
 - 立地などはずれる可能性があるが、データから観察できるので、バランス後の比較で対応

本講義内容を活用した因果推論



継続学習

- 佐藤先生の計量経済学
- 因果効果を定義/識別する枠組みの比較 (Wang, Richardson, and Robins 2025)
- 最新の論点: (Cinelli et al. 2025)
- 新書: [データ分析の力因果関係に迫る思考法](#)
- 大学院レベルの教科書: [CausalMLBook](#)

Reference

- Allcott, Hunt, Matthew Gentzkow, Ro'ee Levy, Adriana Crespo-Tenorio, Natasha Dumas, Winter Mason, Devra Moehler, et al. 2025. "The Effects of Political Advertising on Facebook and Instagram Before the 2020 US Election." National Bureau of Economic Research.
- Allcott, Hunt, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, et al. 2024. "The Effects of Facebook and Instagram on the 2020 Election: A Deactivation Experiment." *Proceedings of the National Academy of Sciences* 121 (21): e2321584121.
- Allcott, Hunt, Matthew Gentzkow, Benjamin Wittenbrink, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, et al. 2025. "The Effect of Deactivating Facebook and Instagram on Users' Emotional State." National Bureau of Economic Research.
- Cinelli, Carlos, Avi Feller, Guido Imbens, Edward Kennedy, Sara Magliacane, and Jose Zubizarreta. 2025. "Challenges in Statistics: A Dozen Challenges in Causality and Causal Inference." *arXiv Preprint arXiv:2508.17099*.
- Egami, Hiroyuki, Md Shafur Rahman, Tsuyoshi Yamamoto, Chihiro Egami, and Takahisa Wakabayashi. 2024. "Causal Effect of Video Gaming on Mental Well-Being in Japan 2020–2022." *Nature Human Behaviour*, 1–14.
- Wang, Linbo, Thomas Richardson, and James Robins. 2025. "Causal Inference: A Tale of Three Frameworks." *arXiv Preprint arXiv:2511.21516*.