まとめ

川田恵介

Table of contents

1 おすすめ復習方法

- 手法の整理を常に意識
  - 推定対象の定義、仮定、性質を区別
- 手を動かす
  - データ分析法であれば、手持ちのデータに適用してみる
- 周辺の論点も目をとおす
  - 学んだ手法の特徴や欠点が見えやすくなる

1.1 定義: 母平均の近似モデル

- 本講義の推定対象: 平均値の” 良い” 近似モデル
  - 変数 Y の**平均値**を近似するモデル $g_Y(X)$ を推定:

$$E[Y|X] \sim g_Y(X)$$

  - **平均**差を近似するモデル $g_\tau(X)$ を推定:

$$E[Y|1,X] - E[Y|0,X] \sim g_\tau(X)$$

- “良い” の定義が複数あることに注意
  - 漸近正規性を優先的重点項目とするか?

1.2 整理: Y の平均値の近似

- $E[Y|X]$ を近似するモデル $g_Y(X)$ は、以下の**母集団上での**最適化問題を解くことができれば、獲得できる

$$\min_{g_Y(X)} E[(Y - g_Y(X))^2]$$

  - 母集団が観察できないので、直接解くことはできない
    - 代替的な最適化問題を解く: OLS, RandomForest, LASSO で異なる

1.3 整理: モデルと実現値の乖離

- 母平均とモデルの乖離は以下のように分解できる

$$E[Y|X] - g_Y(X) =$$

$$\underbrace{E[Y|X] - g_{Y,\infty}(X)}_{ApproximationError} + \underbrace{g_{Y,\infty}(X) - g_Y(X)}_{EstimationError}$$

where $g_{Y,\infty}(X) =$ 無限大の事例数で推定されたモデル

1.4 整理: 機械学習

- RandomForest, LASSO, .. で $g_Y(X)$ を推定すれば、

$$\underbrace{E[Y|X] - g_{Y,\infty}(X) + g_{Y,\infty}(X) - g_Y(X)}_{\text{頑張って減らす}}$$

1.5 整理: OLS 推定

- Simple な Linear model $g_Y(X) = \beta_0 + .. + \beta_L X_L$ を OLS で推定すれば、

$$\underbrace{E[Y|X] - g_{Y,\infty}(X)}_{\neq 0} + \underbrace{g_{Y,\infty}(X) - g_Y(X)}_{\sim N(0,\sigma^2)}$$

- $g_{Y,\infty}(X) =$ **B**est **L**inear **P**rojection
  - BLP に対する推定誤差の近似的性質が" 判明" している

1.6 整理: 教科書的理想例

- β を適切に選べば $g_Y(X) = E[Y|X]$ を達成できるモデルを OLS で推定すれば、

$$\underbrace{E[Y|X] - g_{Y,\infty}(X)}_{=0} + \underbrace{g_{Y,\infty}(X) - g_Y(X)}_{\sim N(0,\sigma^2)}$$

- 推定値の分布が、近似的に判明する

1.7 まとめ: OLS VS 機械学習

- 十分にシンプルな (研究者が定義する)BLP について、漸近的信頼区間が計算可能
  - モデルの個別パラメタについて、誤差の議論が可能
- 多くの機械学習アルゴリズムは、Estimation/Approximation error を同時に削減する

- 無限大の事例数の元で、$g_Y(X) = E[Y|X]$ (一致性) を達成するアルゴリズムも多い
- 完璧な予測モデルに到達できない限り、漸近性質が BlackBox な場合が多い

1.8 比較研究への拡張

- $E[Y|1, X] - E[Y|0, X]$ の近似モデル $g_\tau(X)$ を推定
- AIPW: 2 段階の推定を行う
  - Psude-outcome $\phi(g_Y(d, X), g_D(X))$ を機械学習を用いて推定
  - psude-outcome の近似モデルとして $g_{\tau(X)}$ を推定
    - 2 段階目は、Y の近似モデル生成と同じ議論が適用できるようにする (Neyman の直行条件、1 段階目の予測モデルの収束速度)
- Moment 推定法を用いて、一般可能 (例: R-learner)

2 発展: Predictive interval

- Recap: 本講義の中心的議題は、**母平均**
  - **個人ではなく、集団の特徴**
  - 個々の Y についての予測誤差と混同しないことが重要

2.1 予測問題

- Y_i ないし 個人因果効果 τ_i の**最善の**予測モデル (最小の平均二乗誤差) は、$E[Y_i|X]$ ないし $E[\tau_i|X]$
- 以下の分解が可能

$$Y_i - g_Y(X_i) =$$

$$\underbrace{Y_i - E[Y|X] + \underbrace{E[Y|X] - g_Y(X)}_{\text{平均値についての誤差}}}_{\text{予測誤差}}$$

2.2 Average Prediction Error

- 予測誤差 $(Y - g_Y(X))$ の測定方法として、テストデータにおける平均二乗誤差 (または R^2) を紹介

$$E[(Y - g_Y(X))^2]$$

  - あくまでも「平均的な予測性能」の評価

2.3 Average Prediction Error の限界

- 仮に予測性能が非常に高かったとしても、事例によっては予測を大はずししているかもしれない
- 予測性能が低かった場合の活用法が難しい
  - Double Debiased Learning への活用では、平均値を捉えれば良いので問題ない
  - 予測には活用しにくい、かといって判断の参考になる数字は欲しい

2.4 Predictive interval

- 最善 (ただし必ず外れる) の点推定ではなく、高い確率で当たる予測幅を知りたい
  - 取引価格は、90 % の確率で 2000 - 3000 万円の範囲内になる

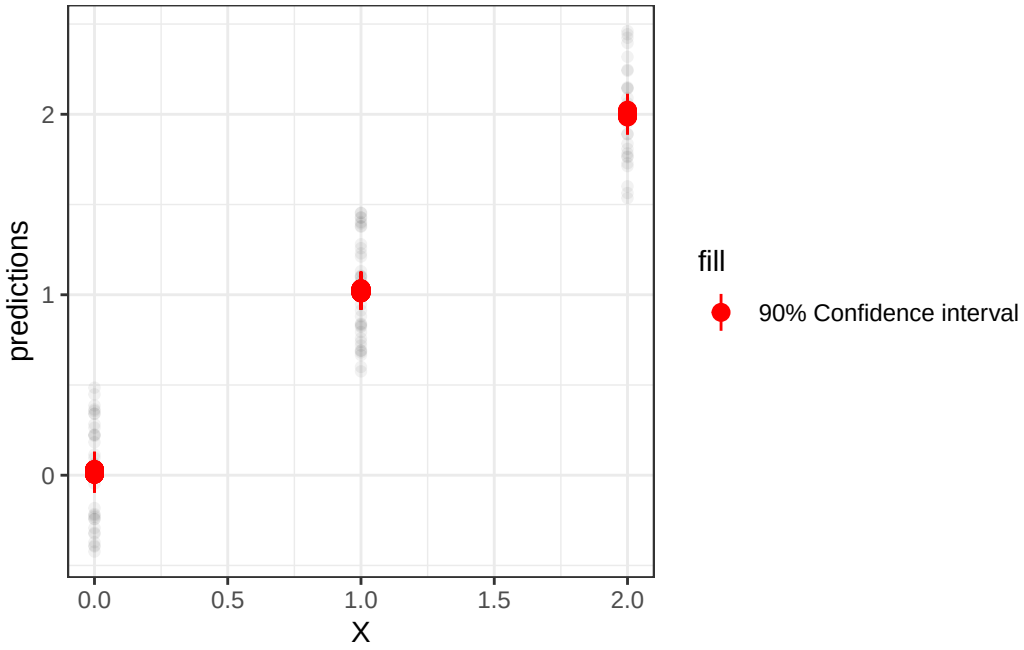
2.5 典型的間違い

- 信頼区間を活用すれば良い???
- 母平均に対する信頼区間であり、個別の Y についての議論ではない
  -「Sampling の結果に応じて、無数の信頼区間が実現する」イメージを持つことが重要
- 95 % 信頼区間 $= 95\%$ の信頼区間が母平均を含む
  - 区間内に、95% の事例があるわけではない

2.6 数値例

- $E[Y|X] = X$
  - $Y - E[Y|X] \sim Uniform(-0.5, 0.5)$
  - $X = \{0, 1, 2\}$

2.7 数値例

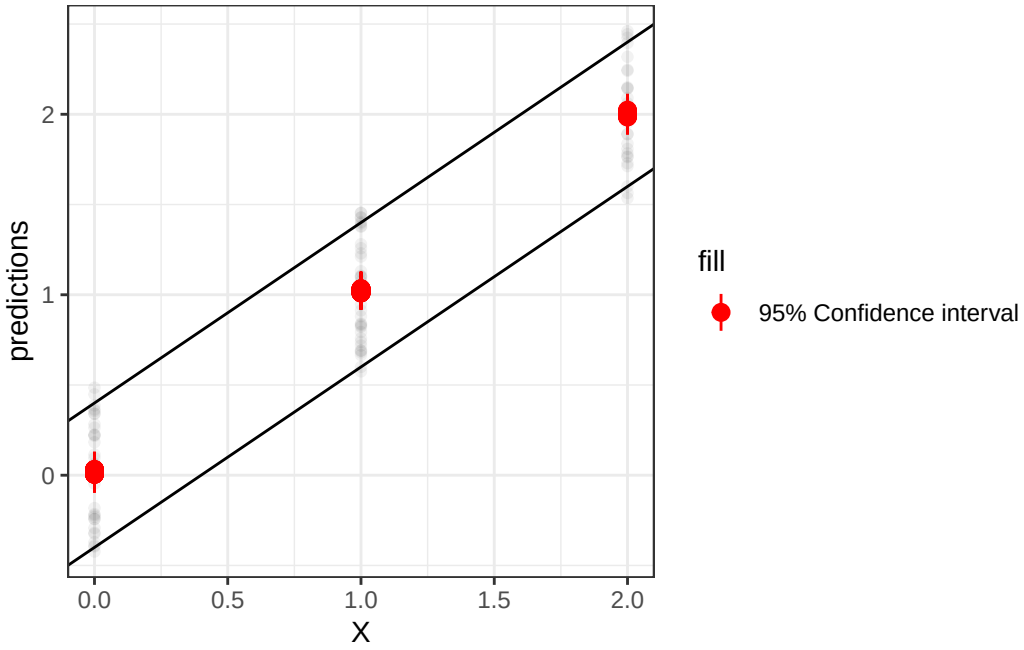


- Irreducible error が大きく、予測誤差は大きい
  - 90 % 以上の事例が、信頼区間の外部に存在

2.8 Conformal inference

- Predictive inference を推定する代表的手法
  - わかりやすい入門論文 (Angelopoulos, Bates, et al. 2023)
- 個人因果効果への応用: Lei and Candès (2021)
  - cfcausal

2.9 Estimand: 数値例



2.10 Conformalized quantile regression

- 90% の確率で当たる予測はどのように推定できるか?
  - 母集団が観察できのでれば、区間 $C(X; 0.1) = [Q_{.05}(X), Q_{.95}(X)|X]$ に予測対象の 90% が含まれる
  - $Q_q(X) = X$ 内で下から数えて q% 目の Y の値
- 理論的性質改善のための補正: Romano, Patterson, and Candes (2019)

2.11 実例

- 築 20 年、60 平米、駅から 5 分の物件について、各立地ごとに平均取引価格の推定値 (+ 信頼区間)、および 90% の conformal interval を計算する
  - 平均取引価格: grf package の regression_forest を使用 (信頼区間の近似計算可能)
  - Conformal inference: cfcausal package の conformal を使用

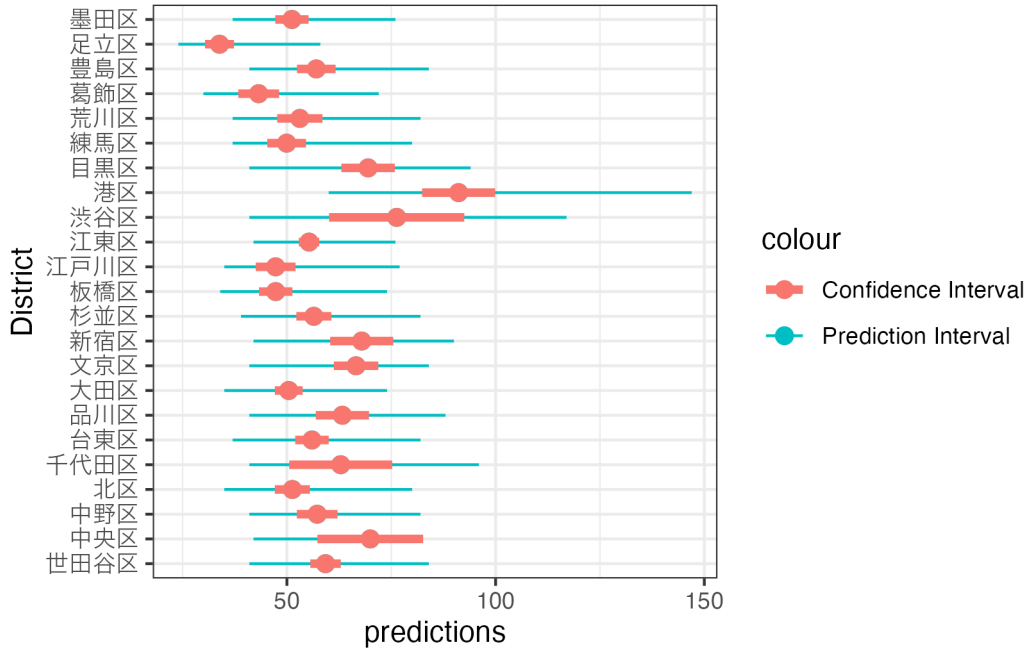
2.12 実例: 築 20 年、60 平米、駅から 5 分

```
PredPoint = grf::regression_forest( # Random Forest
  X = X, # Size, Tenure, District, Distance
  Y = Y, # Price
  num.trees = 5000
  ) |>
  predict(
    TestX,
    estimate.variance = TRUE # Show estimator's variance
    )

Model = cfcausal::conformal( # Conformal inference
  X = X,
  Y = Y,
  type = "CQR",
  quantiles = c(0.05, 0.95),
  outfun = "quantRF",
  wtfun = NULL,
  useCV = FALSE)

Pred = Model |>
  predict(
    TestX,
    alpha = 0.1
  )
```

2.13 実例: 築 20 年、60 平米、駅から 5 分



2.14 実例: 築 20 年、60 平米、駅から 5 分

- 築 20 年、60 平米、駅から 5 分の物件について、各立地ごとに Reform の平均効果の推定値 (+ 信頼区間)、および個別因果効果についての、90% の conformal interval を計算する
  - 平均取引価格: grf package の causal_forest を使用 (信頼区間の近似計算可能)
  - Conformal inference: cfcausal package の conformalIte を使用

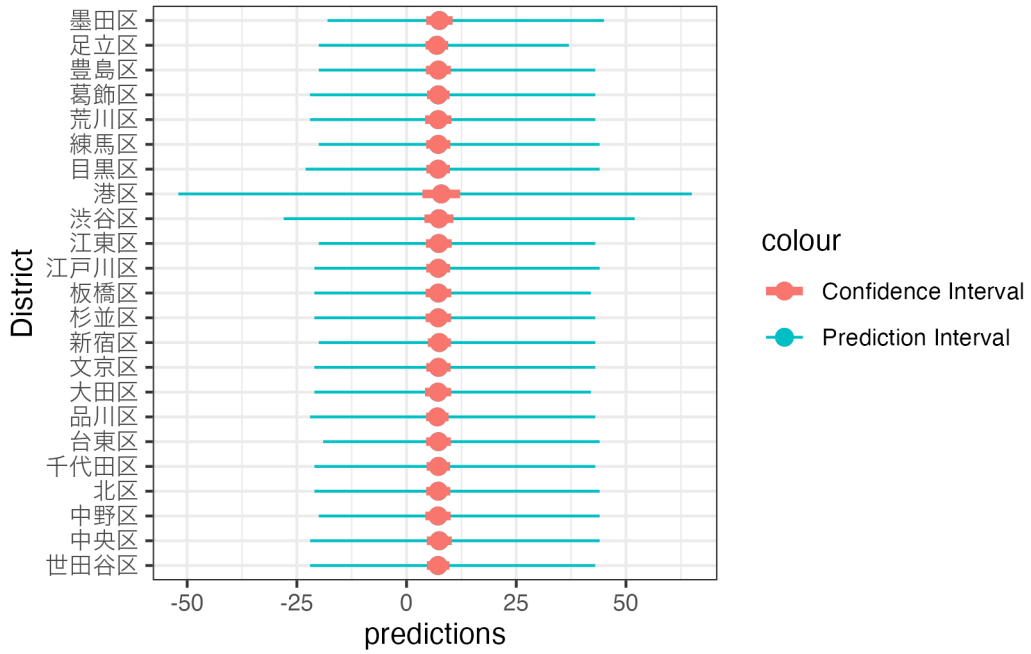
2.15 実例: 築 20 年、60 平米、駅から 5 分

```
Model = cfcausal::conformalIte(
  X = X,
  Y = Y,
  T = D,
  alpha = 0.1,
  algo = "nest",
  exact = FALSE,
  type = "CQR",
  quantiles = c(0.05, 0.95),
```

```
  outfun = "quantRF",
  useCV = FALSE
)

PredictConformal = Model(TestX)
```

2.16 実例: 築 20 年、60 平米、駅から 5 分



3 他の論点

- 因果効果についての論点を紹介

3.1 構造推定

- Parametric model の推定: $E[Y|X]$ ではなく、母分布 $f(Y, X; \theta)$ を規定する有限個のパラメタ θ を推定する
  - 経済理論から導出された 母分布 $f(Y, X; \theta)$ を推定し、政策シミュレーション分析に活用する
  - (自然) 実験による推定では、定義/推定しえない因果効果の議論が可能
    - 典型例: (実験室内も含めて) 過去に実行されたことがない介入

3.2 構造推定: 敵対学習

- 伝統的には最尤法やベイズ法などで推定されてきたが、初期値依存など、多くの推定上の問題が存在
  - Kaji, Manresa, and Pouliot (2023) : 実際のデータか、モデル $f(Y, X; \theta)$ からシュミレートされたデータが、“AI でも区別できない” ように θ を推定する
  - 紹介記事

3.3 因果効果の探索

- ここまでの議論は、事前に定義された少数の変数間の因果効果の大きさを推定する方法を紹介
  - 例: 改築が取引価格に与える影響
- 複数の変数間での因果的関係性 (含む方向性) を、データから発見できるか?
  - Causal Discovery
  - 現状、独立性の検定など、社会データへの応用については課題が多いが、将来的に実用的になるかも
    - 紹介論文 (Huber 2024)

3.4 因果効果の探索

Reference

Angelopoulos, Anastasios N, Stephen Bates, et al. 2023. “Conformal Prediction: A Gentle Introduction.” *Foundations and Trends® in Machine Learning* 16 (4): 494–591.

Huber, Martin. 2024. “An Introduction to Causal Discovery.” https://arxiv.org/abs/2407.08602.

Kaji, Tetsuya, Elena Manresa, and Guillaume Pouliot. 2023. “An Adversarial Approach to Structural Estimation.” *Econometrica* 91 (6): 2041–63.

Lei, Lihua, and Emmanuel J Candès. 2021. “Conformal Inference of Counterfactuals and Individual Treatment Effects.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (5): 911–38.

Romano, Yaniv, Evan Patterson, and Emmanuel Candes. 2019. “Conformalized Quantile Regression.” *Advances in Neural Information Processing Systems* 32.