# ポイント

教師付き学習の経済学への応用

川田恵介

## Table of contents

# 推定問題

- $\theta := T(f(X,Y))$ の推定
  - $T :=$ 既知の関数
  - $f(X,Y) :=$ 母分布

## 伝統的な推定

- “伝統的な推定”: $\theta$ について、“信頼できる” 信頼区間を形成する (Unbiasedness を重視)
  - “2 次的” な目標として効率性 (点推定量の精度アップ: 信頼区間の縮小)
- しばしば非常に厳しい関数系への仮定を要求
  - 改善策: マッチング法 + 回帰

## 機械学習

- 本講義では、
$$\theta = E_P[Y|X]$$
を推定する便利なツール
  - “パラメタ” ではなく、関数を推定
- “機械学習”: 極力近い関数を推定する
  - Unbiasedness は Must requirement ではない
  - 一般に信頼区間の計算も難しい

## 応用: Bruns-Smith, Feller, and Nakamura (2023)

- 母平均関数そのものを Estimand とする研究

- ランダム抽出された家計について” 推移”, $\tau_i = \{Z_i, Y_i\}$ , が観察可能
  - $Z_{it}$ := 家計 $i$ の時点 $t$ についての、年齢、学歴、時点、資産
  - $Y_{it}$ := 所得
- 予測される所得 $E_P[Y_t|X_t]$
  - $X_{it} = [Y_{i,t-1}, X_{i,t-1,..}]$

### Estimand: Income Shock

- $\Delta_t := Y_t - E_P[Y_t|X_t]$
  - 予測できない所得変化

### Estimand: Persistency

- $\phi_{t,h} := E_P[Y_{t+h}|X_t] - E_P[Y_{t+h-1}|X_{t-1}]$
  - $t$ 期目の所得変化を織り込んだ後の所得変化

### Estimand

- $t$ 期の所得変化が、どの程度 Persistecy をもたらすのか?
  - 何期後まで続くのか

## 生涯所得の変化との接続

- 期待生涯所得 $Y_t^{perm}$ := 将来所得の予測値の割引現在価値
- $$Y_t^{perm} := E_P[\sum_{k=t} \gamma^{k-t} Y_k | X_t]$$
- 予期せぬ生涯所得の変化
- $$Y_t^{perm} - E[Y_t^{perm}|X_{t-1}] = \Delta_t + \sum_{h=1} \gamma^h \psi_{t,h}$$

## 結果

- Boosting を使用

  - 所得の予測力: 線形モデルよりは高い
  - $E_P[Y|X]$ の近似力が改善

## 予測への応用

- $E_P[Y|X]$ は、$X$ から $Y$ を予測する” 最善” のモデル
- ただし完璧に予測できるモデルではない
  - テストが重要

### 削減不可能な誤差

- $E_P[Y|X]$ の推定値 $:= g(X)$
- $$Y - g(X) = \underbrace{Y - E_P[Y|X]}_{\text{削減不可能}} + \underbrace{E_P[Y|X] - g(X)}_{\text{削減可能}}$$

### 削減不可能な誤差

- Social Outcome についての予測は、根本的に疑わしいとの主張も
- かなり $X$ を増やしても、顕著な個人差が存在
  - 例: 双子
  - 削減不可能な誤差が大きい
- 必ずテストを!!!
  - 単純な予測アルゴリズム (OLS + 研究者による変数選択など) との比較も

## 信頼区間の計算

- $E[Y|X]$ の推定はかなり野心的なゴール
  - どれだけ巨大なデータを使ったとしても、誤差が” 常に” 生じる

### 機械学習の問題点

- $E[Y|X]$ の近似モデルを柔軟に推定
  - Nonparametric 推定と同様に、一般に推定誤差の性質を調べることが難しい

- 本講義の中心アイディア: 推定プロセスの一部を機械学習で行う
  - 機械学習の推定誤差を” 無視” できるような工夫を行う
    - 誤差の性質について、緩やかな仮定のみを要求

## モーメント法への応用

- $$E_P[m(\theta, Data, g)] = 0$$

  として定義される $\theta$ については、一般的な推定方法が提案されている

1. $\partial m/\partial g = 0$ を満たすような $m$ を使用
2. $g$ を機械学習を用いて、推定
3. 推定された $g$ を代入し、$\theta$ を推定

## 利点

- **近似計算**の精度を高めることができる
- 機械学習の品質への要求 (収束速度) を緩めることができる
- $\theta$ の**漸近分布**において、バイアスの正規分布を保証

## 有限標本性質

- 現実のサンプルサイズは有限
  - 漸近性質がどこまで適用できるのか、シュミレーションを使って議論される (Knaus, Lechner, and Strittmatter 2021)
  - 有限標本での保証 (Chernozhukov, Newey, and Singh 2023)

## 有限標本性質: サンプル分割

- サンプリングに伴う不確実性 + **サンプル分割に伴う不確実性**
- 漸近分布において、後者は無視できる
  - 有限標本では微妙
- マニュアル (Bach et al. 2021) で、推定プロセス自体を繰り返す方法 (chernonozhukov et al. 2018) が推奨されている

## 繰り返し DML

1. $\partial m/\partial g = 0$ を満たすような $m$ を使用
2. $g$ を機械学習を用いて、推定
3. 推定された $g$ を代入し、$\theta$ を推定
4. 1-3 を何度か繰り返す

## 実装例

```
FitPLR_R <- DoubleMLPLR$new(
  Task,
  ml_l=RegNuisanceLearner$clone(),
  ml_m=RegNuisanceLearner$clone(),
  n_folds = 5,                                    ①
  n_rep = 3                                       ②
  )
```

① サンプル分割数
② 繰り返し回数

## 実例

- ３回繰り返す
- 点推定量

```
          [,1]      [,2]      [,3]
[1,] 0.1770029 0.1728856 0.1750678
```

- 標準誤差

```
             [,1]         [,2]         [,3]
[1,] 0.005181284 0.005149713 0.005145048
```

- 集計 (中央値)

```
    After
0.1750678
```

### Britto, Pinotti, and Sampaio (2022)

- 裁判所データと従業員-事業所データをマッチング (名寄せ!)
- $D :=$ 大規模解雇に伴う雇用喪失
- $Y :=$（認知された）犯罪への関与
- $X :=$ 年齢, 勤続年数, 学歴,　賃金, 地域レベル変数 (殺人率、非正規率、格差率など)
- Causal Forest を採用

#### 結果

- 98% の事例について、条件付き平均効果は正に” 有意”
  - 背景属性に関わらず、雇用喪失は、犯罪率を上げる
- 低年齢層、勤続年数が短い層で特に顕著

## サンプリング問題

- Semiparametric 推定 (含む機械学習)「関心のある母集団からランダムサンプリングされている事例」の含意を導く
  - そうではなければ何が言える???
  - データ分析の「見果てぬ夢」

#### 問題の類型

- 事例間に相関があるサンプリング
  - Clustered, Time-series, spacial sampling
- (サンプルの) 母集団外への結果の適用

### Clustered sampling

- 実証研究において、事例を集団 (Cluster) で扱う方が妥当なケースは多い
  - 集団 (企業、家計) をランダムサンプリングして、その構成員を調査する
  - 集団レベルで原因変数が固定 (学校への介入、村への介入)

- 集団内の事例は、集団間に比べて、より” 似ている” 可能性がある
- 対応: サンプル分割/Bootstrap を Cluster level で行う

### MLR3 による実装

- Task を定義する際に、Cluster variable (ここでは ClusterVariable) を group として定義

```
Task <- as_task_regr(
  Data
  )

Task$set_col_roles("ClusterVaraible", "group")
```

### DoubleML による実装

- Task を定義する際に、Cluster variable (ここでは ClusterVariable) を cluster_vars として定義

```
double_ml_data_from_matrix(
  X = X,
  y = Y,
  d = D,
  cluster_vars = ClusterVariable
)
```

### 系列相関

- 時系列, 空間相関: 事例間に近い/遠いがあり、近い事例間の方がよく似ている
- 私見ではまだまだ解決が難しい
  - Survey (Masini, Medeiros, and Mendes 2023)

### 異なる母集団への適用

- Hot Issue !!!
  - $X$ の分布についての調整 (Dahabreh et al. 2020)
  - $X$ 以外についても感度分析 (Nie, Imbens, and Wager 2021)

## まとめ: 人間 VS AI

- AI(機械学習) が BlackBox であることは、しばしば批判される
  - その通りだが、人間の脳みそはより BlackBox (!?)
- コードとシード値で再現できるので、推定を極力データ主導で行うことは有力
  - Principle のあるアプローチ (Urminsky, Hansen, and Chernozhukov 2019)
- 研究課題の設定、識別、要約に、人間は集中できる

## Reference


Bach, Philipp, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. 2021. "DoubleML–an Object-Oriented Implementation of Double Machine Learning in r." *arXiv Preprint arXiv:2103.09603*.

Britto, Diogo GC, Paolo Pinotti, and Breno Sampaio. 2022. "The Effect of Job Loss and Unemployment Insurance on Crime in Brazil." *Econometrica* 90 (4): 1393–423.

Bruns-Smith, David, Avi Feller, and Emi Nakamura. 2023. "Using Supervised Learning to Estimate Inequality in the Size and Persistence of Income Shocks." In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1747–56.

chernonozhukov, 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, Whitney K Newey, and Rahul Singh. 2023. "A Simple and General Debiased Machine Learning Theorem with Finite-Sample Guarantees." *Biometrika* 110 (1): 257–64.

Dahabreh, Issa J, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. 2020. "Extending Inferences from a Randomized Trial to a New Target Population." *Statistics in Medicine* 39 (14): 1999–2014.

Knaus, Michael C, Michael Lechner, and Anthony Strittmatter. 2021. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." *The Econometrics Journal* 24 (1): 134–61.

Masini, Ricardo P, Marcelo C Medeiros, and Eduardo F Mendes. 2023. "Machine Learning Advances for Time Series Forecasting." *Journal of Economic Surveys* 37 (1): 76–111.

Nie, Xinkun, Guido Imbens, and Stefan Wager. 2021. "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials Across Locations." *arXiv Preprint arXiv:2112.04723*.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2019. "The Double-Lasso Method for Principled Variable Selection."