Weighted average of conditional difference

機械学習の経済学への応用

川田恵介

違和感

- 線形回帰モデル (Partial Linear model も含む) は、データ全体で D のばらつきがある限り、" どんな" データを使ったとしても何らかの係数値を推定する
  - 原理的に差が推定できないグループがデータに含まれていたとしても
- 例: 中古マンション取引データから、改築済み・前グループの比較を行う
  - 築 1 年の改築済みマンションが存在しない
  - $E[Y|D=1, BuildYear=1] - E[Y|D=0, BuildYear=1]$ とは？

Conditional Difference

- $\tau(X) = E[Y|D=1, X] - E[Y|D=0, X]$
  - サブグループ内での平均差
  - サブグループ内での平均効果（後述）
- 経済学において一般に重要だが、X が多次元の場合推定が難しい
  - 要約統計量を推定する

Weighted average

- 周辺化 (Marginalization): 代表的戦略
- 母集団上で定義される関心 (Estimand) を、（非）明示的に以下のように定義

$$\int \tau(X) \times \underbrace{\omega(X)}_{Weight} \, dX$$

Partial Linear Model

- Variance Weighted Average Difference (Angrist and Pischke 2009; Vansteelandt and Dukes 2022)

$$\omega(X) = \frac{E[D|X]\{1 - E[D|X]\}}{\int E[D|X]\{1 - E[D|X]\}g(X)dX}$$

- $g(X)$: 母分布
- D のばらつきが大きいサブグループをより代表する
- D のばらつきが 0 のサブグループは、一切反映されない

Equal weight

- 最もシンプルな weight は、$\omega(X) = 1$
  - Variance-weight に比べて解釈が容易な局面も多い
- 例: 学歴間隔差が男女間でどのように異なっているか？
  - 男性・女性サンプルを使って、平均差をそれぞれ推定
  - 同じ X を推定に用いたとしても、variance weight の値は異なってくる可能性 (例: 女性の方が生まれ年と学歴の相関が強い)

Plug-in and IPW Estimaters

- Plugin-in

$$\sum_i f_Y(D = 1, X_i) - f_Y(D = 0, X_i)$$

- Inverse propensity score weight (IPW)

$$\sum_i \frac{D_i Y_i}{f_D(X_i)} - \frac{(1 - D_i)Y_i}{1 - f_D(X_i)}$$

- $f_Y = Y$の予測モデル, $f_D = D$ の予測モデル

AIPW (Doble Robust)

- Plugin-in, IPW は、Nuisance functions f_Y, f_D の推計誤差の影響をモロに受ける
  - Neyman's ohthogornality 条件を満たさない

- Robins and Rotnitzky (1995) : Argumented IPW (AIPW)

$$\sum_i \underbrace{f_Y(D=1,X_i) - f_Y(D=0,X_i)}_{Plug-in}$$

$$+ \underbrace{\frac{D_i(Y_i - f_Y(D=1,X_i))}{f_D(X_i)} - \frac{(1-D_i)(Y_i - f_Y(D=0,X_i))}{1-f_D(X_i)}}_{AdjustTerm}$$

アルゴリズム

- DoubleML で容易に実装可能

1. f_Y, f_D を交差推定
2. 代入し、AIPW スコアを計算
3. AIPW スコアの母平均を推定・信頼区間計算を行い、平均差の推定値とする

まとめ

- 統計モデル上は一つのパラメータ = 何らかの集計量として解釈
  - 「格差や因果効果が均一である」と仮定しているわけではない
- しばしば暗黙の内に設定される Weight に注意
- AIPW はより直感的な Estimand を推定するかも
  - Variance weight を厳密な研究関心とするケースは？

注意点

- AIPW が推定できないケースも多いので、Partial Linear Model(Variance weights) も依然として現実的かつ妥協可能な Summary として活用される
- より一般的な枠組み (Influence function による修正) として導出できる
- 次回紹介

Reference

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Robins, James M, and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429): 122–29.

Vansteelandt, Stijn, and Oliver Dukes. 2022. "Assumption-Lean Inference for Generalised Linear Model Parameters." *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. https://doi.org/10.48550/arXiv.2006.08402.