Inference on heterogeneity aware estimand

川田恵介

Table of contents

1 Conditional Average Difference

- Conditional average difference $\tau(X) = \underbrace{\mu_Y(1,X)}_{=E[Y|1,X]} - \underbrace{\mu_Y(0,X)}_{=E[Y|0,X]}$ の" 特徴" として Estimand を定義
  - 因果推論では Conditional average treatment effect (CATE) と呼ばれる

1.1 Conditional Average Difference

- 一般に $D \in \{0,1\}$ 間での差は、他の変数 X に依存していると考えられる:
  - 自然な Estimand は
$$\tau(x) = \mu_Y(1,x) - \mu_Y(0,x)$$
- $X = x$ を満たすサブサンプルサイズが十分にあれば、ただのサブサンプル平均差を推定値にできる
- ほとんどの応用で、大量の X を用いるので、サブサンプルサイズは不足する

1.2 Estimand: Weighted average difference

- 現実的な Estimand の第一候補は、$\tau(X)$ の平均値:
$$\theta_0 = \int_X \tau(X) \times \omega(X) dX$$
- $\omega(X) =$ “研究者” が暗黙のうちに設定する集計用 Weight
- **重要ポイント:** Estimand は Y, D, X のみならず、$\omega(X)$ にも依存して定義される
  - かつてはそれほど意識されてこなかった

1.3 Estimand: Proportion weight

- $\omega(X) = f(X)$
  - 因果推論では Average Treatment Effect と呼ばれる
-
$$\theta_0 = \int \tau(X) \times f(X) dX$$

1.4 単純平均との比較

-
$$\theta_0 = \int_X \mu_Y(1, X) \times \omega(X) dX - \int_X \mu_Y(0, X) \times \omega(X) dX$$

-
$$E[Y|1] - E[Y|0] = \int_X \mu_Y(1, X) \times f(X|1) dX - \int_X \mu_Y(0, X) \times f(X|0) dX$$

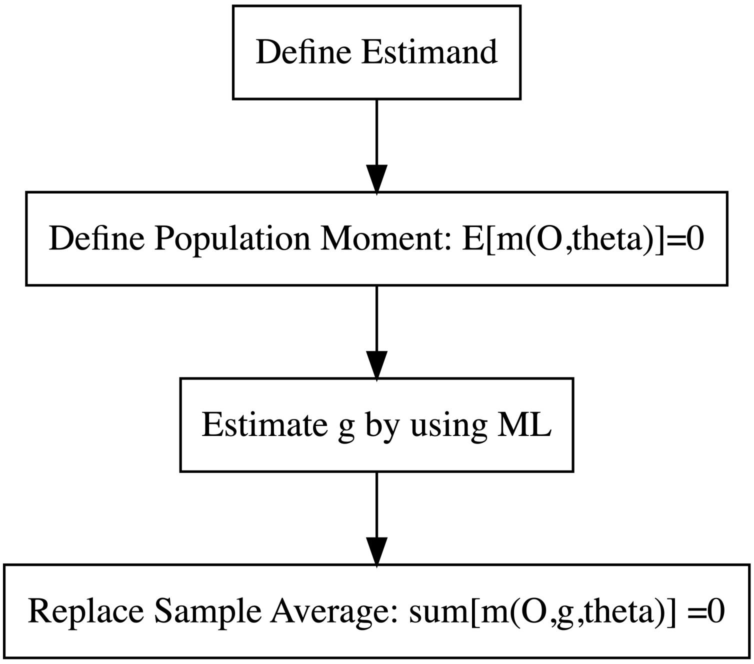
2 Estimation

- $g_Y(d, X), g_D(X)$ が中心的な役割を果たす
  - $g_D(X) =$ Propensity score と呼ばれる

2.1 Propensity score をめぐる混乱

- 「Propensity score では因果効果を識別できない」
  - 正しいが identification の手法ではなく、estimation の手法であり、議論が噛み合っていない場合が多い
- OLS との違いがわからない
  - 本章の主要論点: OLS や R learner では proportion weight を用いた平均効果を推定できない

2.2 推定手順



2.3 Moment condition

- 複数存在する: $0 = E[m(\theta_0, O)]$ where
  - $$m(\theta_0, O) = \theta_0 - \mu_Y(1, X) + \mu_Y(0, X)$$
    - g method, plugin method などと呼ばれる
  - $$m(\theta_0, O) = \theta_0 - \frac{DY}{\mu_D(X)} + \frac{(1-D)Y}{1-\mu_D(X)}$$
    - inverse propensity score などと呼ばれる

2.4 Argumented inverse propensity score (Robins and Rotnitzky 1995)

- おすすめの Moment condition
- $$m(\theta_0, O) = \theta_0 - \mu_Y(1, X) + \mu_Y(0, X)$$ $$-\underbrace{\frac{D(Y - \mu_Y(1, X))}{\mu_D(X)} + \frac{(1-D)(Y - \mu_Y(0, X))}{1 - \mu_D(X)}}_{Adjustment}$$

- Neyman's orthogonal condition を満たす

2.5 Psude-outcome

- 以下のように書き換えられる: $\theta_0 = E[\phi(O)]$ where

$$\phi(O) = \mu_Y(1, X) - \mu_Y(0, X)$$
$$+\frac{D(Y - \mu_Y(1, X))}{\mu_D(X)} - \frac{(1-D)(Y - \mu_Y(0, X))}{1 - \mu_D(X)}$$

- $\phi(O) =$ Psude-outcome と呼ばれる
  - ϕ の平均 $=$ τ の平均

2.6 Estimator

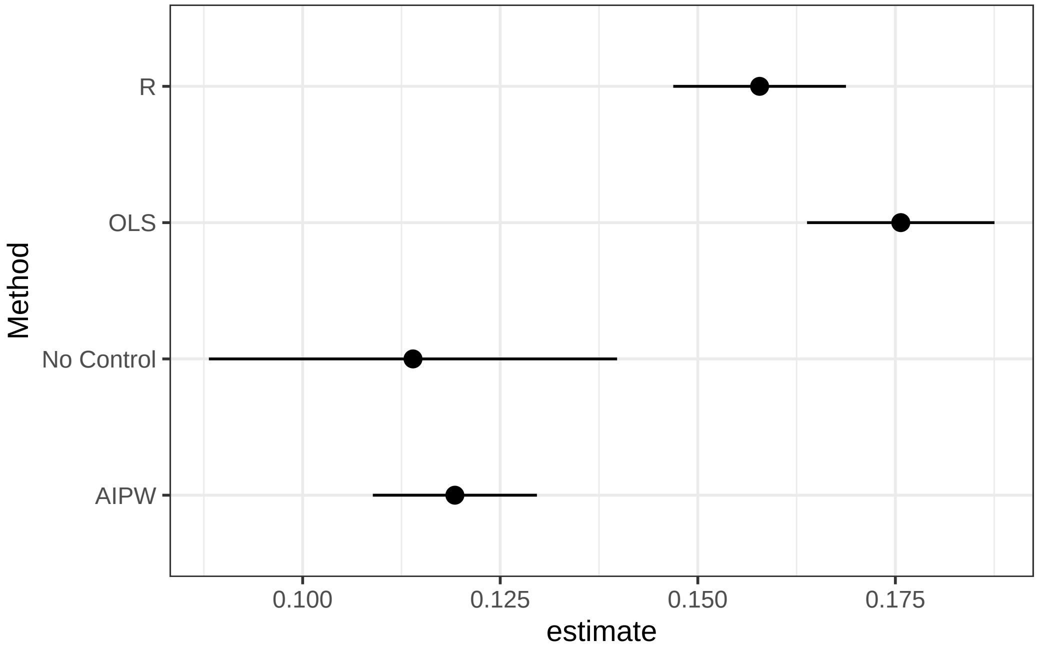
- $0 = \sum \phi(O, g)/N$ where

$$\phi(O, g) = g_Y(1, X) - g_Y(0, X)$$
$$+\frac{D(Y - g_Y(1, X))}{g_D(X)} - \frac{(1-D)(Y - g_Y(0, X))}{1 - g_D(X)}$$

2.7 Algorithm

1. データ分割 (auxiliary/main data)
2. μ_Y, μ_D を auxiliary data (+ 機械学習) で推定する
3. main data を用いて、$\phi(O, g)$ の平均値を計算し、信頼区間とともに報告する

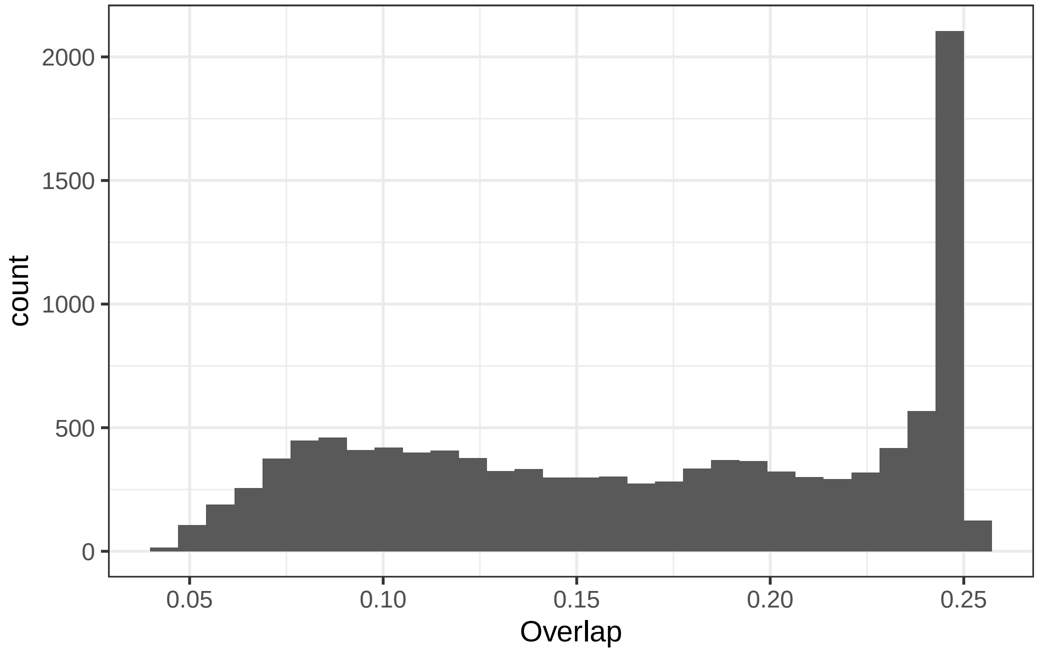
2.8 Example: R VS AIPW



2.9 R VS AIPW

- R learner も heterogeneity aware estimand として解釈できる
- 異なる集計用 Weight を用いた Estimand
  - AIPW: $\omega(X) = f(X)$
  - R: $\omega(X) = \frac{\mu_D(X)(1-\mu_D(X))f(X)}{\int \mu_D(X)(1-\mu_D(X))f(X)dX}$
    - Overlap Weight
    - $\mu_D(X) = 0.5$ (バランスよく $D = 1, 0$ が混在しているサブグループ) の平均差をより強く反映している

2.10 Exmaple. Overlap weight



- 最大で 6 倍以上の格差が存在

2.11 Example. Overlap weight

- Lowest/Largest Overlap weight

```
# A tibble: 6 x 5
   Size Distance Tenure Youseki Overlap
  <dbl>    <dbl>  <dbl>   <dbl>   <dbl>
1    15        3      1     300  0.0399
2    15        3      1     300  0.0410
3    15        3      1     300  0.0418
4    20        6      1     500  0.0429
5    20        6      1     500  0.0432
6    20        7      1     300  0.0442
```

```
# A tibble: 6 x 5
   Size Distance Tenure Youseki Overlap
  <dbl>    <dbl>  <dbl>   <dbl>   <dbl>
1    55       15     29     300   0.250
2    55        2     44     400   0.250
```

3	50	5	43	500	0.250
4	50	5	44	400	0.250
5	60	5	37	300	0.250
6	50	8	43	500	0.250

2.12 まとめ

- R learner とは異なる Estimand (異なる集計用 Weight) を推定している
- 一般に、R learner よりも解釈しやすい
  - Overlap weight とは?
  - $\tau(X), \mu_D(X)$ の異質性が大きい場合、乖離幅が大きくなる
- D がカテゴリカルであり、positivity の問題 (後述) がなければ、AIPW を用いることを推奨

2.13 補論: Marginal means

- X の分布をバランスさせた Y の平均値

$$\theta_0(d) = \int \mu_Y(d, X) \times f(X) dX$$

  を $D = 1, 0$ ごとに提示することもできる
  - 差 + 絶対水準を示すことができ、誤解が減らせる
- Average difference $= \theta_0(1) - \theta_0(0)$

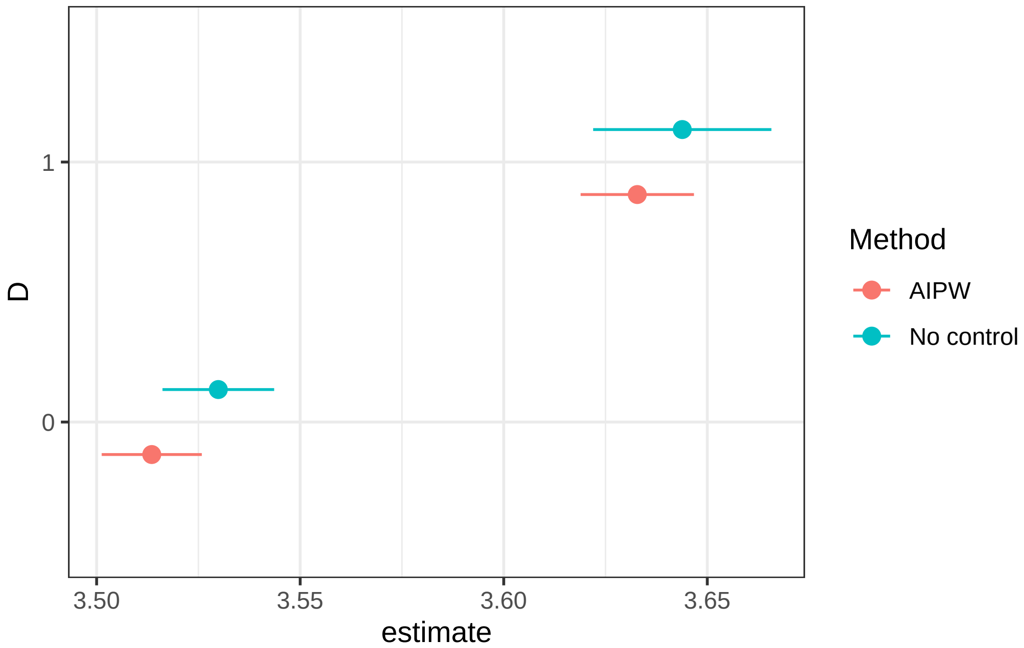
2.14 補論: Marginal means

- Estimator は、

$$\theta(1) = E\left[g_Y(1, X) + D\frac{Y - g_Y(1, X)}{g_D(X)}\right]$$

$$\theta(0) = E\left[g_Y(0, X) + (1 - D)\frac{Y - g_Y(0, X)}{1 - g_D(X)}\right]$$

2.15 Example



3 Best Lienar Projection for CATE

- 平均差では、X との関係性について、含意を持たない
  - 信頼区間もしっかり推定しつつ、CATE の持つ特徴を、もう少し理解することを目指す
- $\tau(X)$ を近似するシンプルな線形近似モデル $g_\tau(Z) = \beta_0 + \beta_1 Z_1 + ..$ を推定する

3.1 Algorithm

1. データ分割 (auxiliary/main data)
   - 交差推定も活用可能
2. μ_Y, μ_D を auxiliary data (+ 機械学習) で推定する
3. main data を持ちいて AIPW と同じ psude-outcome を回帰する: $\phi(O, g) \sim Z$ を OLS で推定、信頼区間とともに推定する

3.2 Best Linear projection

- Estimand:

$$\min_{\beta} E[(\tau(X) - g_\tau(Z))^2]$$

  where $Z \subset X$ and

$$g_\tau = \beta_0 + \beta_1 Z_1 + .. + \beta_L Z_L$$

- 注: Average Difference は特殊ケース: $g_\tau(Z) = \beta_0$

3.3 補論: Normalization

- g_τ は" 記述" のために推定する
  - 全ての Z を事前に標準化 $(Z - mean(Z))/sd(Z)$ することがおすすめ
- β_0 は通常、解釈を持たない
  - 全ての Z が 0 における τ の近似値
    - 通常、そのような事例がないので、近似モデルを作る際に無視されるサブグループ
- 標準化すれば、$\beta_0 =$ 全ての Z が平均値であった時の τ の近似値と解釈できる

4 Proportional weight の問題点と対策

- 比較研究 (含む格差、因果) における根本的な仮定は、Positivity:

$$1 > E[D = d, X] > 0$$

- 実践においては、しばしば満たされない
  - 推定、識別に対して、解決困難な問題をもたらす

4.1 Assumption: Positivity for identification

- 任意の d について

$$1 > \Pr[D = d|X] > 0$$

  - 直感: X の中での比較研究をしているので、$D = 1$ または 0 のサブグループが存在する場合、比較不可能
- 因果推論であれば、Positivity + 因果効果の識別用の仮定 (Conditonal independence, No interference など)

4.2 Assumption: Positivity for estimation

- 任意の d について 0 ないし 1 に非常に近い

$$\Pr[D = d|X]$$

  が存在する場合、推定が難しくなる

  – "推定が上手くいっているのであれば"、$g_D(X)$ が 1 ないし 0 に近いサブグループが出てくる

4.3 Trouble with AIPW

- $\theta = \sum m(O, g)/N$ where

$$m(O, g) = g_Y(1, X) - g_Y(0, X)$$
$$+ \frac{D(Y - g_Y(1, X))}{g_D(X)} - \frac{(1 - D)(Y - g_Y(0, X))}{1 - g_D(X)}$$

- Adjust term の分母が"0" に近づく事例が出てくる

  – 推定誤差が非常に大きくなる

4.4 Overlap weight

- R learner であれば、estimand は $\int \tau(X) \times \omega(X)dX$ where

$$\omega(X) = \frac{\mu_D(X)(1 - \mu_D(X))f(X)}{\int \mu_D(X)(1 - \mu_D(X))f(X)dX}$$

- 定義上、$\mu_D(X)$ が 1 または 0 に近いグループは、"無視" する

  – より安定的な推定が可能

  – Average Difference との乖離が大きくなり、より解釈が難しくなる...

4.5 Propensity score weight

- $\mu_D(X)$ が 0 に近いグループが存在すること**のみ**が問題であれば、解釈と推定を両立する Estimand が存在

- Propensity score weight $\int \tau(X) \times \omega(X)dX$ where

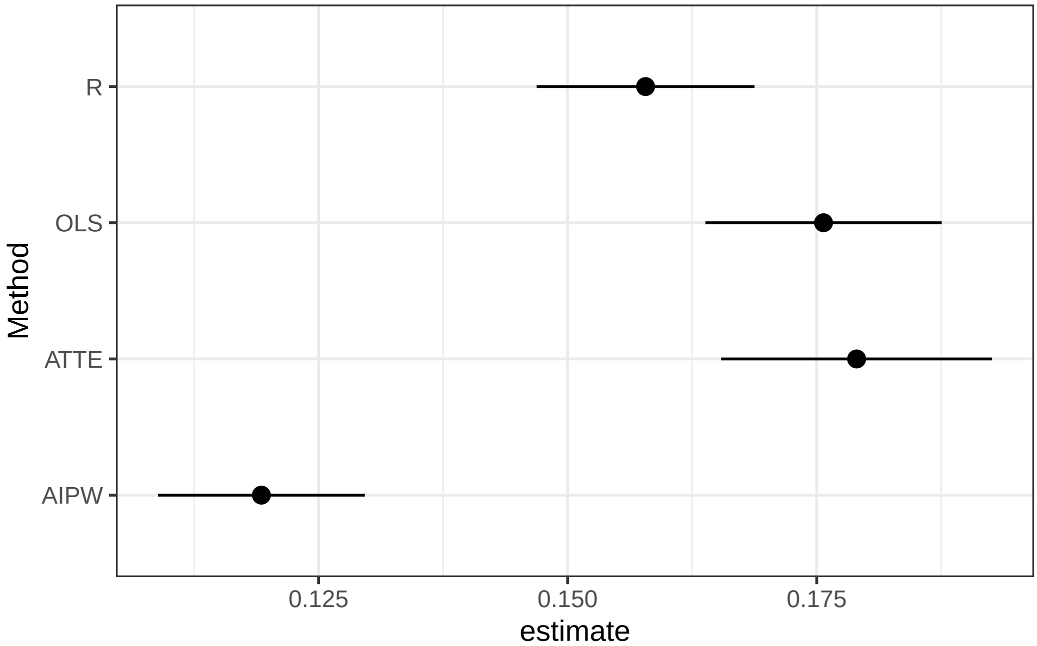
$$\omega(X) = \frac{\mu_D(X)f(X)}{\int \mu_D(X)f(X)dX}$$

- 推定が難しい $g_D(X) \simeq 0$ となるサブグループは、定義上、無視する

4.6 Average treatment effect on treated

- 以下のように書き換えられる $\int \tau(X) \times f(X|1)dX$
- 解釈が容易
  - 男女間格差研究: $D = 1$ が女性であれば、X の分布を男女ともに女性と揃えた時の、男女間格差
  - 因果推論: $D = 1$ における平均因果効果
    - Average Treatment Effect on Treated と呼ばれる

4.7 Example: ATE VS Overlap VS ATTE



4.8 まとめ

- 平均差を estimand とする場合は、どのような集計用 weight を用いるかも定義する必要がある
- proportion weight が最も直感的なので、(D がカテゴリカルであれば)、最有力候補
  - $\mu_D(X)$ が 0 や 1 に近いグループが存在する (positivity に問題がある) 場合、推定困難

4.9 まとめ

- $D = 1$ が少ないことが問題であれば、propensity score weight が解釈も容易な解決策
  - $D = 0$ の場合は、" ひっくり返せば良い" だけ
- $\mu_D(X) \simeq 0$ と $\simeq 1$ が両方存在する場合は?
  - overlap weight は有力だが、$\tau(X)$ が同質でない限り、解釈が難しい
  - 他には Trimming (Yang and Ding 2018) や Balancing weight (Ben-Michael et al. 2021), Targetted learning (Van der Laan and Rose 2011) などを活用する提案もあるが、万能薬は現状ない

Reference

Ben-Michael, Eli, Avi Feller, David A Hirshberg, and José R Zubizarreta. 2021. "The Balancing Act in Causal Inference." *arXiv Preprint arXiv:2110.14831*.

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.

Van der Laan, Mark J, and Sherri Rose. 2011. *Targeted Learning*. Vol. 1. 3. Springer.

Yang, Shu, and Peng Ding. 2018. "Asymptotic Inference of Causal Effects with Observational Studies Trimmed by the Estimated Propensity Scores." *Biometrika* 105 (2): 487–93.