Applications: Sensitvitiy Analysis

川田恵介

Table of contents

1 Sensitivity Analysis

- 識別に必要な変数が観察できない問題 (Omitted variable problem) に対して、推定結果の頑健性を評価する手法
  - Debiased Machine Learning を使用することで、Model specification に依存しない推定が可能になる

- サーベイとしては、Ding (2023) の 17-19 章など

1.1 Omitted variable problem

- 研究課題に応える理想の Esitmand は、$E[m(\theta_0, O, U)] = 0$ だが、U はデータから観察できない
- 本スライドでは、平均差における Omitted variable problem を考える: 理想の Estimand は

$$\theta_+ = E[E[Y|1, X, U] - E[Y|0, X, U]]$$

- U もバランスさせたいが、観察できない....

1.2 例: OS の因果効果

- $$E[Y|1, X, \text{プログラミングへの関心}]$$
$$-E[Y|0, X, \text{プログラミングへの関心}]$$
- 多くの応用で、プログラミングへの関心は omitted variable

1.3 例: 同一学歴内賃金格差

- $$E[Y|1, X, \text{出身学部/大学}]$$
$$-E[Y|0, X, \text{出身学部/大学}]$$
- 多くの応用で、出身学部/大学は omitted variable

1.4 論点整理

- $E[E[Y|1, X, U] - E[Y|0, X, U]]$ を推定する際の問題として
  - 定式化問題: $Y \sim D, X, U$ の関係性がよくわからない
    - ここまでの議論を活用することで、緩和できる
  - 観測問題: U が観察できない
    - 新しいアプローチが必要

1.5 Sensitivity model

- Omitted variable の推定結果への影響について、上/下限を" 想定する"
- 本スライドでは、Effect difference model (McClean, Branson, and Kennedy 2024) を紹介
- 因果効果の上限/下限は

$$\{\theta + c_U, \theta - c_L\}$$

where $\theta = E[Y|1, X] - E[Y|0, X]$

  - 問題は Sensitivity parameter $\{c_U, c_L\}$ の設定

1.6 Calibrated sensitivity model

- **観察可能** な X の影響を、Sensitivity parameter の設定に用いる
  - X の一部 X_- のみを残して平均差を定義する

$$\theta_- = E[E[Y|1, X_-] - E[Y|0, X_-]]$$

  - U の影響は、落とした変数の影響 $\theta - \theta_-$ の一定割合 γ 以内であると仮定する

1.7 Example

- 出身大学/学部はわからないが、教育年数はわかる
- 知りたいのは、$X = \{$教育年数､年齢$\}$ と $U =$ 出身大学/学部 をバランスさせた平均差
- "最善" の推定値でも、X のみをバランスさせた平均差
  - Calibration のために、$X_- =$ 年齢 のみをバランスさせた推定を行う
- "出身大学の" 追加的な影響" は、教育年数の影響よりも小さい" と想定できるのであれば、$|\theta_+ - \theta| < |\theta - \theta_-|$

1.8 Identification

- 任意の γ のもとで、τ_+ の上限/下限は、以下のように識別される

$$\tau_+ \in \{\tau + \gamma \times |\tau - \tau_-|, \tau - \gamma \times |\tau - \tau_-|\}$$

1.9 Estimation

- McClean, Branson, and Kennedy (2024) にて、Neyman's ohthogonality を満たすモーメント条件が提案されている
- Psude-outcome を用いる

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

1.10 Estimation

- $\sum \phi(X)/N > \sum \phi(X_-)/N$ ならば、上限は、

$$\frac{\sum \phi(X) + \gamma(\phi(X) - \phi(X_-))}{N}$$

- 下限は

$$\frac{\sum \phi(X) - \gamma(\phi(X) - \phi(X_-))}{N}$$

1.11 他のやり方との関係性

- Selection-on-Observable を仮定 $\iff \gamma = 0$
- OLS のみを用いた分析でも、同様の senstivity 分析は行われてきた (Oster 2019; Cinelli and Hazlett 2020)
  - 機械学習も活用することで、定式化問題も削減できる
- 注意: Sensitivity 分析として、Main estimation に変数を加える分析が見られるが、望ましくない
  - Main estimation とは?

1.12 発展

- McClean, Branson, and Kennedy (2024) では全 X について、逐次除外し、その中で最も影響が大きいものを使用する方法を紹介
  - Maximum leave-one-out
- Effect difference model 以外の定式化も提案され、今でも議論が続いている
  - Ding (2023) などを参照

2 Practical Example

2.1 Simple example

- 改築前/後の中古マンション取引価格を、広さ、築年数、容積率、と**立地**をバランスさせた上で比較したい
- 立地については、区、駅からの距離、ゾーニングしかわからない。
  - 立地の詳細は Omitted variable

2.2 Simple example

```
set.seed(111)
library(tidyverse)
library(DoubleML)
library(mlr3verse)

lgr::get_logger("mlr3")$set_threshold("warn")

Data = read_csv("Public/Data.csv")

Y = Data$Price |> log()

D = Data$Reform

X = Data |>
  select(
    Size,
    Tenure,
    Youseki,
    Distance,
    District,
    Area
  ) |>
  mutate(
    District = District |>
      factor(
        labels = "Dist"
```

```
      ),
    Area = Area |>
      factor(
        label = "Area"
      )
  )

X_ = Data |>
  select(
    Size,
    Tenure,
    Youseki
  ) # Drop geographical parameters

Task = double_ml_data_from_matrix(
  y = Y,
  d = D,
  X = X
)

Task_ = double_ml_data_from_matrix(
  y = Y,
  d = D,
  X = X_
)

PLR = DoubleMLIRM$new(
  Task,
  lrn("regr.ranger"),
  lrn("classif.ranger"),
  n_folds = 2
  )$fit()

PLR_ = DoubleMLIRM$new(
  Task_,
  lrn("regr.ranger"),
  lrn("classif.ranger"),
  n_folds = 2
  )$fit()
```

2.3 Simple example

```
PLR$summary()
```

```
Estimates and significance testing of the effect of target variables
  Estimate. Std. Error t value Pr(>|t|)
d  0.105892   0.006828   15.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PLR_$summary()
```

```
Estimates and significance testing of the effect of target variables
  Estimate. Std. Error t value Pr(>|t|)
d  0.098265   0.005786   16.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.4 Simple example

```
Phi = PLR$psi_b[,1,1] # Get Psude-outcome

Phi_ = PLR_$psi_b[,1,1]

Gamma = 1 # Set parameters

Phi_U = Phi + Gamma*(Phi - Phi_)

Phi_L = Phi - Gamma*(Phi - Phi_)

estimatr::lm_robust(Phi_U ~ 1)
```

```
             Estimate Std. Error t value     Pr(>|t|)   CI Lower  CI Upper
(Intercept) 0.1135203  0.0104682 10.8443 2.486527e-27 0.09300188 0.1340387
               DF
(Intercept) 22138
```

```
estimatr::lm_robust(Phi_L ~ 1)
```

```
             Estimate  Std. Error  t value     Pr(>|t|)   CI Lower  CI Upper
```

```
(Intercept) 0.09826463 0.005785966 16.98327 2.790116e-64 0.08692373 0.1096055
               DF
(Intercept) 22138
```

Reference

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (1): 39–67.

Ding, Peng. 2023. "A First Course in Causal Inference." *arXiv Preprint arXiv:2305.18793*.

McClean, Alec, Zach Branson, and Edward H Kennedy. 2024. "Calibrated Sensitivity Models." *arXiv Preprint arXiv:2405.08738*.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.