補論: 異質性分析

労働経済学 2

川田恵介

Table of contents

1 動機

1.1 動機

- 因果効果や格差、あるいは単なる” 差” が、個人や集団間で異なることは、政策/基礎研究において、極めて重要

1.2 例: 時系列

- コロナ下 (2020 年 4 - 6 月) とその前年 (2019 年 4 - 6 月) で、労働状態にどのような差があるのか？
- Fukai, Ichimura, and Kawata (2021)：全体平均では大きな変化はないが、一部層で大きな影響がある

1.3 例: 因果推論

- 就業支援プログラムの内容は、受講者の状態に影響を与えるのか？
- Behaghel, Crépon, and Gurgand (2014): フランスにおける RCT データを用いて、公的主体のプログラムは、民間主体のプログラムに比べて、平均的な再就職確率が高い
- Kallus (2023): 同じデータを用いた異質性の推定
  - 一部の被験者についてのみ、公的主体のプログラムが有効であり、民間主体のプログラムが有効な被験者も多い

1.4 例: 男女間格差の分解

- どのような層で男女間格差は大きいのか?
  - Bach, Chernozhukov, and Spindler (2024)

1.5 方法論

- 機械学習 × 経済/医療統計の人気があるテーマであり、大量の方法が提案
  - CausalML の 14-15 章 などを参照
- grf package を紹介
  - 残差回帰 (Robinson 1988) の拡張

- 元論文 (Wager and Athey 2018; Athey, Tibshirani, and Wager 2019)
- 拡張 + 入門 (Athey and Wager 2019)

2 残差回帰

2.1 復習

- 以下を OLS 推定して得られる β_D は、X の**平均値**をバランスさせた後の比較結果

$$Y \sim \beta_D D + \beta_0 + \beta_1 X_1 + ..\beta_L X_L$$

2.2 復習

- X の**分布**をバランスさせるためには、以下を推定する必要がある

$$Y \sim \beta_D D + \underbrace{f(X_1, .., X_L)}_{\text{十分に複雑}}$$

- X の数が多いと困難
  - 変数選択がうまく機能しない場合も多い
  - より柔軟な手法が必要

2.3 母集団における残差回帰

- 母集団において、$f(X)$ を十分に複雑した β_D の推定結果は、以下の手順でも得られる

1. $E[Y \mid X], E[D \mid X]$ を計算
2. $Y - E[Y \mid X] \sim D - E[D \mid X]$ を OLS

2.4 データ上での残差回帰

- 以下の手順での推定値は、母集団における残差回帰を推論できる

1. 機械学習などのデータ主導の方法を用いて、$\mu_Y(X) \simeq E[Y \mid X], \mu_D(X) \simeq E[D \mid X]$ を推定
2. $Y - \mu_Y(X) \sim D - \mu_D(X)$ を OLS

- 1 段階目への推定精度について、緩やかな仮定を課すことで、信頼区間を計算できる

2.5 発展学習

- 1 段階目に活用できる機械学習の推定手法は、大量に存在 (Boosting/Nural net/Deep Learning)
  - 一般に教師付き学習の回帰問題を念頭においている手法は活用できる
  - CausalML あるいは Introduction to Statistical Learning 参照

3 R-learner

3.1 Nie and Wager (2021)

- 以下で定義される関数 $\tau(X)$ を推定

$$Y \sim \tau(X) \times D + \underbrace{f(X_1, .., X_L)}_{\text{十分に複雑}}$$

- X に応じて、D と Yの平均差が異なることを許容
- Robinson (1988) にちなみ、R-leanrer と呼んでいる

3.2 母集団における R-learner

1. $E[Y \mid X], E[D \mid X]$ を計算
2. $(Y - E[Y \mid X] - \tau(X)[D - E[D \mid X]])^2$ の母平均を最小化するように $\tau(X)$ を算出

3.3 データ上での R-learner

1. 機械学習などのデータ主導の方法を用いて、$\mu_Y(X) \simeq E[Y \mid X], \mu_D(X) \simeq E[D \mid X]$ を推定
2. $(Y - \mu_Y(X) - \tau(X)[D - \mu_D(X)])^2$ の母平均を**極力**最小化するように $\tau(X)$ を推定

- 2 段階目にも機械学習を検討

3.4 Causal Forest

- 2 段階目について、Random Forest を活用
- 1 段階目については、なんでも OK だが、grf パッケージの default では、Random Forest を活用
  - 利点: X の数が少なければ、信頼区間も計算可能

3.5 实例

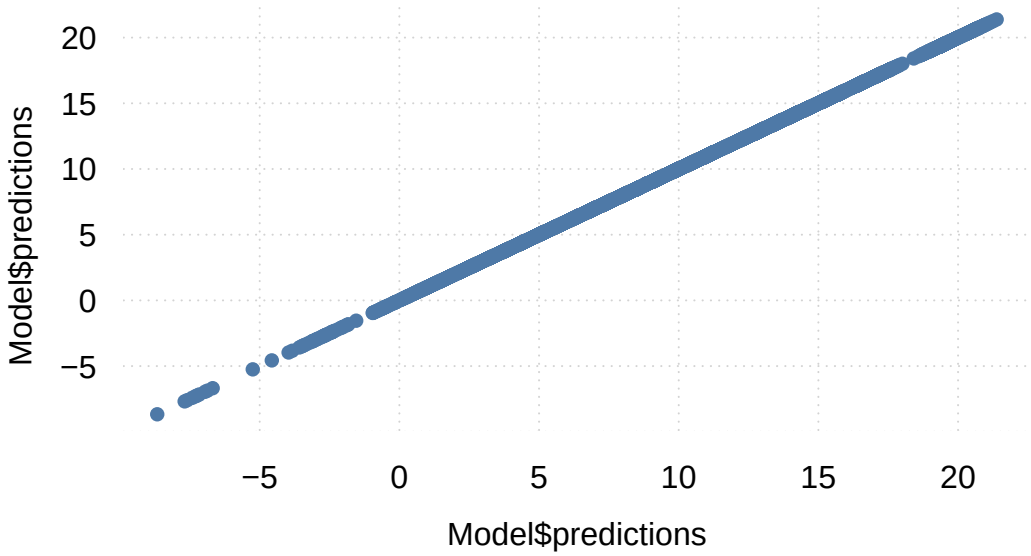
```
library(tidyverse)
library(grf)

data(CPS1988,package = "AER")

Y = CPS1988$wage
W = CPS1988$experience
X = model.matrix(
  ~ education + ethnicity + smsa + region,
  CPS1988)
X = X[,-1]
```

3.6 实例

```
Model = causal_forest(
  X = scale(X),
  W = W,
  Y = Y
)
```

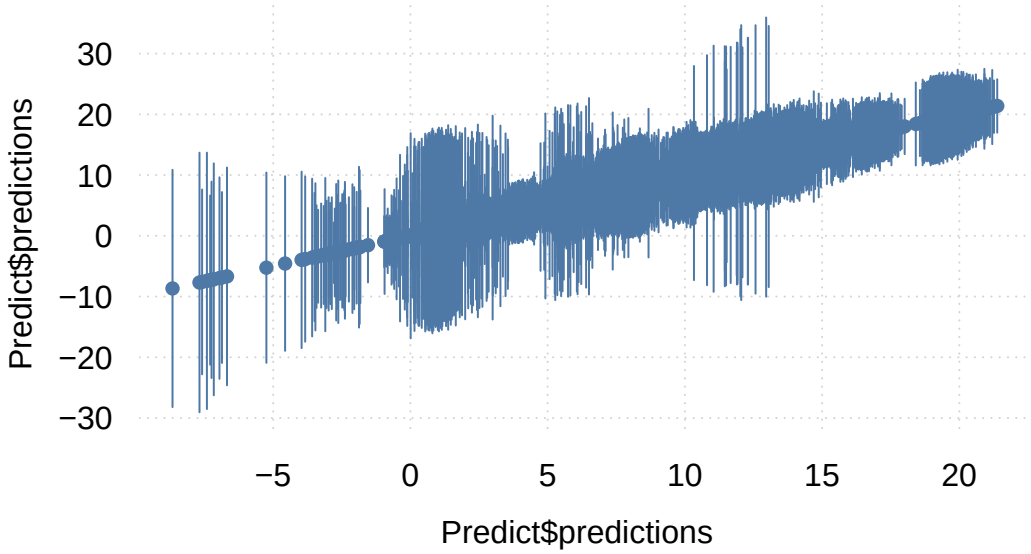
3.7 实例

```
library(tinyplot)

tinytheme("clean2")

plt(x = Model$predictions,
    y = Model$predictions)
```



3.8 实例

```
Predict = predict(Model, estimate.variance = TRUE)

plt(x = Predict$predictions,
    y = Predict$predictions,
    ymin = Predict$predictions -
      1.96*sqrt(Predict$variance.estimates),
    ymax = Predict$predictions +
      1.96*sqrt(Predict$variance.estimates),
    type = "pointrange")
```



3.9 Nonparametric model

- $\tau(X)$ の性質を直接調べるのは、一般に困難
  - 信頼区間計算が難しい/計算できても広い
  - 大量の推定値を出しても、人間の理解が追いつかない
    - X との関係性が理解できない
- 因果効果の予測には有効だが、把握には Too much な可能性

3.10 記述モデル

- $\tau(X)$ の” 特徴” を、シンプルなモデルを用いて推定する
  - $\tau(X) \sim \beta_0$
    - 平均差
  - $\tau(X) \sim \beta_0 + \beta_1 X_1 + ..$
    - 線型モデル

3.11 例

```
average_treatment_effect(
  Model
  )
```

```
  estimate    std.err
10.5127023  0.2489033
```

3.12 例

```
best_linear_projection(
  Model,
  X |>
    scale()
  )
```

```
Best linear projection of the conditional average treatment effect.
Confidence intervals are cluster- and heteroskedasticity-robust (HC3):

              Estimate Std. Error t value  Pr(>|t|)
(Intercept)   10.51270    0.24805 42.3807 < 2.2e-16 ***
education      2.80779    0.32056  8.7590 < 2.2e-16 ***
ethnicityafam -0.74686    0.20162 -3.7043 0.0002124 ***
smsayes        1.55467    0.23618  6.5824 4.709e-11 ***
regionmidwest  1.39115    0.29274  4.7521 2.023e-06 ***
regionsouth    0.75889    0.33268  2.2811 0.0225491 *
regionwest     0.85342    0.30788  2.7719 0.0055764 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.13 Scaling

- $\mathrm{scale(x)} = \frac{x - x\text{の平均}}{x\text{の標準偏差}}$
- 特に記述モデルを推定する際に有効
- $Y \sim \beta_0 + \beta_1 X$ の $\beta_0 \simeq X$ が 0 の時の Y の平均値

  - 多くの経済分析で、$X = 0$ はデータの範囲外

- scale すると、X が平均値の時の Y の平均値

Reference

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–78.

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *Observational Studies* 5 (2): 37–51.

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2024. "Heterogeneity in the US Gender Wage Gap." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (1): 209–30.

Behaghel, Luc, Bruno Crépon, and Marc Gurgand. 2014. "Private and Public Provision of Counseling to Job Seekers: Evidence from a Large Controlled Experiment." *American Economic Journal: Applied Economics* 6 (4): 142–74.

Fukai, Taiyo, Hidehiko Ichimura, and Keisuke Kawata. 2021. "Describing the Impacts of COVID-19 on the Labor Market in Japan Until June 2020." *The Japanese Economic Review* 72 (3): 439–70.

Kallus, Nathan. 2023. "Treatment Effect Risk: Bounds and Inference." *Management Science* 69 (8): 4579–90.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika* 108 (2): 299–319.

Robinson, Peter M. 1988. "Root-n-Consistent Semiparametric Regression." *Econometrica*, 931–54.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.