実例: バランス後の比較と推定対象の設定

機械学習入門

川田恵介

Table of contents

1 バランス後の比較

- 因果効果の推定や格差分析など、重要な社会/市場分析で活用されている
  - 現代的分析では、「研究課題」を「推定対象」に落とし込むプロセスも意識する必要がある

1.1 研究工程

1. 研究課題の設定: 前年比較、地域間格差の解明、学部の因果的効果
    - 社会/政策/業務課題から設定
2. 推定対象の設定: 研究課題に答えることができる母集団の特徴を指定
    - バランス後の比較においては、分析する事例と $X/D/Y$
        - 因果推論や格差の規範理論を活用できる
3. 推定値の計算: 推定対象の推定値をデータから計算

2 格差分析

- 「倫理的/規範的に望ましくない差」を推定
    - 「望ましくない差」を定義するためにバランス後の差を活用

2.1 研究計画

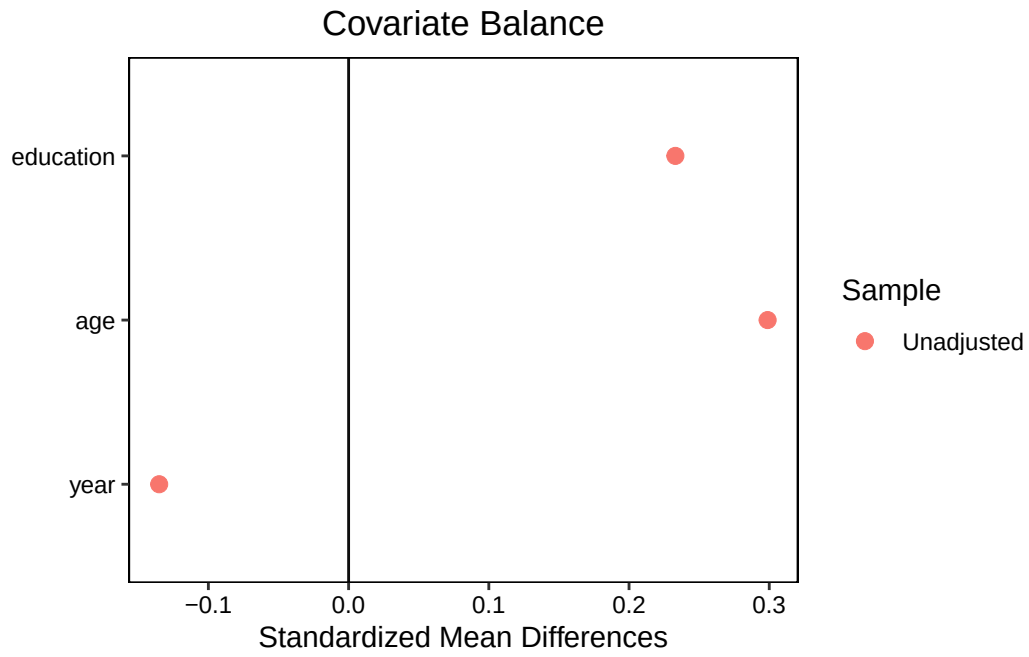
- GSS7402 (AER パッケージに収録) を用いて、米国における「人種間」教育格差を推定
- 同データは、1974 年から 2002 年まで回答を結合しており、多様な年齢層の回答者が含まれている
    - 「人種」は、「白人 (cauc)」と「白人以外 (other)」として記録される
    - 教育年数もわかる

2.2 SetUp

```
library(tidyverse)
library(estimatr)
library(hdm)
library(dotwhisker)
library(cobalt)
library(AER)

data("GSS7402")
```

2.3 単純比較

```
Balance = bal.tab(
  ethnicity ~ education + age + year,
  GSS7402
)

love.plot(Balance)
```



- 平均教育年数は、白人の方が顕著に長いが、他の変数についても差がある

2.4 批判的思考

- (自身の主張への) 批判的思考が有効:「単純差は、望ましくない差なのか?」
  - –「白人」には、昔に生まれた世代が多い
    - ∗ 背景知識より、昔に生まれたせいで、教育年数が短い可能性がある
      - · 格差が過小に推定されている可能性がある

2.5 望ましくない差の定義

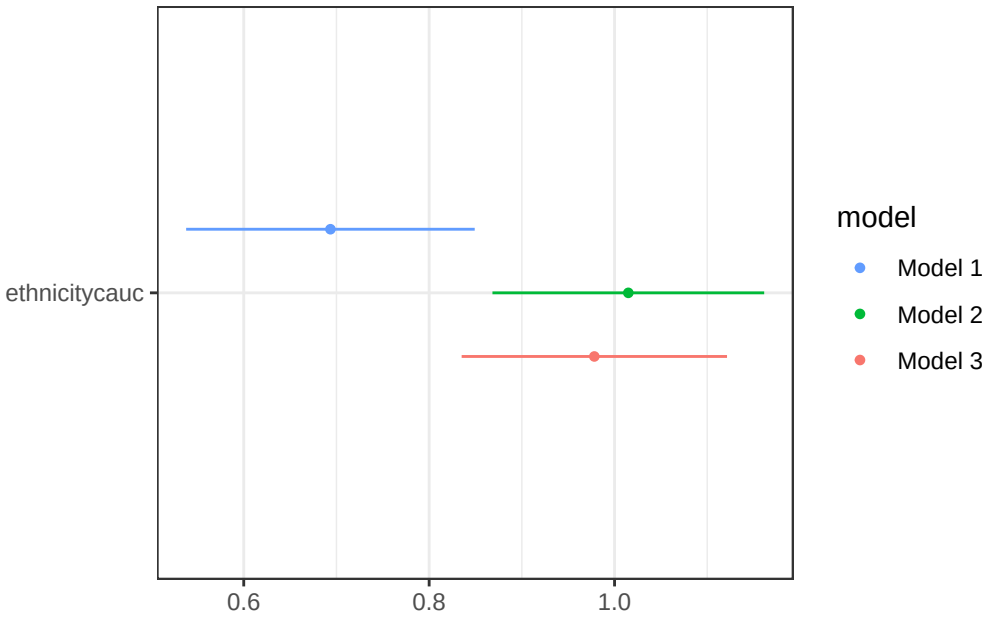
- 「同じ時期に生まれたにも関わらず差がある」ことが問題である、と研究課題を設定
  - 「機会の不平等」
  - 「昔に生まれたから、教育年数が短い」は、“この研究では” 問題にしない
- 推定対象は、$X = \{$年齢、調査年$\}$ をバランスさせた後の人種間平均教育年数格差

2.6 OLS によるバランス

```
Model1 = lm_robust(
  education ~ ethnicity,
  GSS7402
) # Balance なし

Model2 = lm_robust(
  education ~ ethnicity + age + year,
  GSS7402
) # 年齢、調査年の平均をバランス

Model3 = lm_robust(
  education ~ ethnicity + (age + year)**2 + I(age^2) + I(year^2),
  GSS7402
) # 年齢、調査年の平均、分散、共分散をバランス

Fig = dwplot(
  list(Model1,
       Model2,
       Model3
       ),
  vars_order = c("ethnicitycauc")
) +
  theme_bw()
```

2.7 OLS によるバランス

```
Fig
```



2.8 Double selection

```
Y = GSS7402$education

D = if_else(
  GSS7402$ethnicity == "cauc",
  1,
  0)

X = model.matrix(
  ~ 0 + (age + year)**2 + I(age^2) + I(year^2),
  GSS7402
)

X = scale(X)

Model4 = rlassoEffect(
```

```
  x = X,
  d = D,
  y = Y
)

summary(Model4)
```

```
[1] "Estimates and significance testing of the effect of target variables"
   Estimate. Std. Error t value Pr(>|t|)
d1   0.99884    0.07343    13.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.9 Double selection

```
Model4$selection.index
```

```
      age      year  I(age^2) I(year^2)  age:year
     TRUE      TRUE      TRUE     FALSE     FALSE
```

- 調査年の二乗とと年齢との交差項は除外

3 因果推論

- 因果効果の推定においても、中心的な役割を果たす
  - 詳細は計量経済学を受講してください

3.1 例: ゲーム規制論争

- 「主として若年層のゲームを規制すべきかどうか」を一生議論している
  - ゲームの（精神的）健康被害が懸念
- ゲームの利用は、健康被害を**引き起こすのか?**
  - ゲーム利用者と非利用者の差ではない

3.2 例: 因果 VS 単純比較

- 因果効果: “仮想的に” ある集団全員がゲームを利用しない状況から、利用する状況に変化させると何が起きるのか?

- 比較分析: “現実に” ゲームを利用している層と利用していない層を比較した際に、どのような差があるのか?
  - ゲームを利用する層としない層で、そもそもの健康状態の差 (年齢や趣味、その他の様々な要因に起因する)
  - ゲームの影響とそもそもの差が、「混在しており」ゲームの因果効果がわからない

3.3 例: ランダム化対照実験

- 因果効果の理想的な推定には、ランダム化対照実験 (RCT) が要求される
  - ゲームをしない被験者に対して、「ゲーム機を配布する」など
    - ランダムに配布しているので、被験者数が十分いれば、ゲーム機を持っている集団と持っていない集団の間で、そもそもの差が非常に少なくなる
- 金銭的、倫理的、政治的制約があり、現実には難しい

3.4 例: 自然実験

- 社会において” 自然発生した” 実験を活用
- Egami et al. (2024) : コロナ下の日本で生じたゲーム機への超過需要
  - 小売店において生じた、「ゲーム機のランダム割り当て」を活用
    - くじ引きに当選した人だけが買える

3.5 例: 自然実験の補正

- 完璧なランダム化対照実験と自然実験の間には乖離が発生
  - ゲーム機くじ: 当たりやすい地域と当たりにくい地域が発生
    - ゲーム機所有者と非所有者の間で、居住地に違いが発生
      - 完璧なランダム化対照実験では発生しないはず
- 居住地をバランスさせた比較が必要

Reference

Egami, Hiroyuki, Md Shafiur Rahman, Tsuyoshi Yamamoto, Chihiro Egami, and Takahisa Wakabayashi. 2024. “Causal Effect of Video Gaming on Mental Well-Being in Japan 2020–2022.” *Nature Human*

Behaviour, 1–14.