

比較研究: 発展 機械学習

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-11-18

1 比較問題

1.1 バランス後の比較

- 比較分析: $E[Y \mid D = 1, X] - E[Y \mid D = 0, X]$ の推定
- 本スライド: 以下を想定


$$E[Y \mid D, X] = \underbrace{\beta_D \times District}_{D \text{に関する部分}} + \underbrace{\beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \dots}_{X \text{に関する部分}}$$

- X の偏り/影響を排除し、 Y と D の関係性を推定

1.2 予測の直接活用

- 二重選択法 (LASSO によって、 X のなから重要な変数を選ぶ) を紹介
- LASSO よりも性能の良い予測モデルの推定方法 (Stacking) や 汎用性のある AI が活用できないか?
 - ▶ 古典的なアイデア (Varian, 2014) を改良した手法を紹介

1.3 汎用 AI(大規模言語モデル)の活用

- Chat-GPT  などに、直接質問する
- 改装済み/未改装間での平均取引格差を推定した
 - ▶ X = 部屋の広さ、築年数
- 🤖: 55 平米、築 30 年、改装済み、物件価格は?
 - ▶ 🤖: 6800 万円

- 🤖: 55 平米、築 30 年、未改築、物件価格は？
 - 🏠: 5800 万円
- 🤖: 6800 万円 - 5800 万円 = 1000 万円

1.4 自作 AI の活用

```
library(tidyverse)

data <- read_csv("Public/example.csv")

Y <- data$Price

X <- select(
  data,
  Reform,
  Tenure,
  Size)
```

1.5 価格予測 AI の推定

```
library(SuperLearner)

model = SuperLearner(
  Y = Y,
  X = X,
  SL.library = c("SL.lm", "SL.ranger")
)
```

1.6 自作 AI による推定

```
predict(model, tibble(Reform = 1, Tenure = 30, Size = 55))$pred
```

```
      [,1]
[1,] 45.57694
```

```
predict(model, tibble(Reform = 0, Tenure = 30, Size = 55))$pred
```

```
      [,1]
[1,] 37.09001
```

- 差は 8.49

1.7 問題点

- 価格予測 AI の性能に、推定結果が強く依存
 - ▶ 汎用 AI や Stacking などを持ちいた自作 AI も、母平均を十分に近似できていない可能性が高い

2 二つの AI を活用

2.1 ダブルチェック

- 「(自作した) Y を予測する AI 🤖 と D を予測する AI 🤖 にダブルチェックさせる」が有効
- 🏠 : 55 平米、築 30 年、の物件価格は？
 - ▶ 🤖 : 6000 万円
- 🏠 : 55 平米、築 30 年、の改装率は？
 - ▶ 🤖 : 35%

2.2 予測誤差同士の回帰

- データ上のすべての事例について、予測値を計算した後に、
- $(Y - \hat{Y})$ の予測値) を $(D - \hat{D})$ の予測値) で OLS 推定する
 - ▶ どちらのモデルも、母平均をある程度 捉えているならば、高い精度で「改築/比改築物件間での平均価格差」を推定可能
 - ▶ 信頼区間も計算可能
 - 二つの AI が同時に母平均から大きく乖離しない限り、推定結果が大きな悪影響を受けない

2.3 手順

1. データの事例について、 Y/D の予測値を算出する
 - 本講義では、Stacking 法を用いて自作 AI を作るアプローチを推奨
 - ▶ 汎用 AI の活用も議論されている (Engh and Aronow, 2025)
2. $(Y - \hat{Y})$ の予測値) を $(D - \hat{D})$ の予測値) で OLS 推定する

2.4 実装

- 専用のパッケージを利用した方が、安全に実装できる
- ここでは ddml package を利用
 - ▶ doubleml package も有力な選択肢

2.5 実装

```
library(tidyverse)
library(ddml)

data <- read_csv("Public/example.csv")

Y <- data$Price

D <- data$Reform

X <- model.matrix(
  ~ 0 + Tenure + Size,
  data)
```

2.6 実装

```
model <- ddml::ddml_plm(
  y = Y,
  D = D,
  X = X,
  learners = list(
    list(fun = ols),
    list(fun = mdl_ranger)
  ),
  shortstack = TRUE,
  silent = TRUE
)
```

2.7 実装

```
summary(model)
```

PLM estimation results:

, , nnls

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.459	0.294	-1.56	1.18e-01
D_r	6.141	0.747	8.22	2.10e-16

2.8 注意点

- Stacking を構成するアルゴリズムの数が極端に多い場合、追加的なサンプル分割が必要
 - 経済学の多くの応用においては、問題なく活用可能

2.9 Reference

Bibliography

Engh, C. and Aronow, P. (2025) “Using LLMs to Directly Guess Conditional Expectations Can Improve Efficiency in Causal Estimation,” arXiv preprint arXiv:2510.09684 [Preprint].

Varian, H.R. (2014) “Big data: New tricks for econometrics,” *Journal of economic perspectives*, 28(2), pp. 3–28.