Double selection

1 復習

1.1 研究計画

1. 研究目標: 職種間賃金格差
2. 推定目標: $X =$ [年齢、学歴、人種、性別] をバランスさせた後に、$D =$ [専門職、非専門職] を比較

1.2 推定方法

- 線型モデルの推定

$$\beta_0 + \beta_D D + \beta_1 X_1 + ..$$

- バランス後の比較を行うためには、十分に複雑なモデルを推定する必要がある
  - 大量の β を含むモデルを推定する必要がある

1.3 推定方法

- OLS: 事例数に比べて、単純なモデル (β の数が少ない)の推定に向く
  - 弱点: ほとんどの応用で、事例数は限られており、複雑なモデルが推定できない
- LASSO: 複雑なモデルの推定に活用できる
  - 弱点: 信頼区間が計算できない
    - 予測分析では大きな問題ではないが、比較分析では大問題

1.4 推定方法: 本スライドの提案

- Semi-parameric estimation のアプローチを導入
- 線型モデルを Interest と Nuisance に分割する
- $$E[Y \mid D, X] = \underbrace{\beta_D \times D}_{Interest}$$
$$+ \underbrace{\beta_0 + \beta_1 \times X_1 + ...}_{X\text{に関する部分}(Nuisance)}$$

1.5 推定方法: 本スライドの提案

- 本スライドでは、Interest は単純に定式化 (D と Y の関係は、X にかかわらず一定)
- Nuisance を十分に複雑にする

1.6 推定方法: 本スライドの提案

- 基本アイディア: β_D **のみ** をしっかり推論できれば、推定目標に回答できる
  - Nuisance (Y と X の関係性)は、推定できていなくても OK
- β_D の推論を目指して、変数選択を行う

2 人間による変数選択

2.1 古典的なアプローチ

- X に関する部分から、“重要ではない要素”を、(経験や”かん”によって)取り除く
- 例: X_1は”重要ではないので”モデルから除外する

$$E[Y \mid D, X] = \beta_D \times D$$

$$+\beta_0 + \beta_1 \times X_1 + \underbrace{\beta_2 \times X_2}_{=0}$$

2.2 問題点

- 問題点: 取り除く基準が曖昧であり、分析結果を恣意的に操作できる余地も大きい
- そもそも”重要ではない”は、正確に何を意味しているのか?

2.3 重要性

- 目標: 無限大のデータで複雑なモデルを推定し得られる β_D と同じような値を推定したい
- Y や D と”関係ない”要素を、モデルから除外すべき

2.4 例: 成績データ

2.5 例: 推定

- 推定目標: 学生の背景をバランスさせた上で、欠席の有無 (D) 間で、テストの点 (Y) を比較したい
- 理想的な推定方法: 無限大の事例数を用いて、“複雑な”モデルを OLS で推定

$$\beta_D \times \textbf{欠席} + \beta_0 + \beta_1 \times \textbf{学年} + \beta_2 \times \textbf{学籍番号}$$

2.6 例: 変数選択

- 背景知識から、学籍番号はランダムに振られていることを**知っている**
  - テストとも欠席とも関係がないので、モデルから除外した方が、β_D に近い推定値を得やすい
  - 学年は、テストや欠席と関係している可能性が高いので、除外しない方が良い

2.7 問題点

- 信頼できる変数選択を行うだけの背景知識がないケースが多い
- 本講義の提案: データ主導のアプローチ(doble selection)を活用

3 LASSO を用いた Double selection

3.1 アイディア

- X の中から **重要な変数** を、データ主導で選ぶ
  - 予測のための変数選択が行われる LASSO を利用 (Belloni, Chernozhukov and Hansen, 2014)
- 機械学習/AI も、“ミスを犯す可能性”を考慮する
  - 重要な変数が誤って除外されるリスクを考慮

3.2 コード例

```
library(tidyverse)

data("CPS1985", package = "AER")

Y <- CPS1985$wage |> log()

D <- if_else(CPS1985$occupation == "technical",1,0)

X <- model.matrix(
  ~ 0 + education + age + gender + experience,
  CPS1985)
```

3.3 コード例

```
model <- hdm::rlassoEffect(
  x = X,
  y = Y,
  d = D
)

summary(model)
```

3.4 コード例

- X の選択結果

```
model$selection.index
```

3.5 基本手順

1. LASSO を使って、X (含む二乗、交差項)の変数選択を行い、その一部 Z を抽出
2. Z と D のみを用いて、Y について OLS 推定する $Y \sim D + Z$

- 機械学習による”下準備”をしたのちに、OLS で推定する

3.6 Double selection

- Step 1.を以下の手順で行う
  - X から Y を予測するモデルを LASSO で推定し、選択された変数を記録
  - X から D を予測するモデルを LASSO で推定し、選択された変数を記録
  - $Z = D$ **または** Y の予測に用いられた変数を として用いる

3.7 イメージ

```
model_Y = hdm::rlasso(
  x = X,
  y = Y)

model_Y$index
```

3.8 イメージ

```
model_D = hdm::rlasso(
  x = X,
  y = D,
  data = data)

model_D$index
```

3.9 イメージ

```
lm(
  Y ~ D + education + age + gender,
  data = CPS1985)
```

3.10 性質

- 以下の仮定が成り立てば、「複雑なモデルを無限大の事例数で推定した結果」を近似でき、信頼区間も計算できる
- 仮定: 事例数に比べて、十分に少ない変数数で、母平均を近似できる
  - 「もともとのモデルには、“重要ではない”変数も含まれている」を仮定
- D **または** Y の予測に役立つ変数を残していることが重要

3.11 非推奨の方法

- Y の予測の役に立たない変数は、D の予測に役立つとしても除外
- 問題点: 限られた事例数のもとで、LASSO による変数選択は、Y とそこそこ関係ある変数も、誤って除外されてしまう可能性がある
  - D との関係が強い (分布の分断が激しい)な変数が除外されると β_D の推定結果が大きな影響を受ける

3.12 Takeaway

- 二重選択法は、重要な変数を誤って除外しないように、Yの予測モデルとDの予測モデルに”ダブルチェック”を行わせている
  - 二つのモデルが同時に重要な変数を見落とさない限り、推定結果の大幅な悪化は主じない
- 研究者の主観的な変数選択を補完できる
- 推定対象は、引き続き研究者が決めていることにも注意
- 読みやすいサーベイ (Angrist and Frandsen, 2022)

3.13 Reference

Bibliography

Angrist, J.D. and Frandsen, B. (2022) “Machine labor,” Journal of Labor Economics, 40(S1), pp. S97–S140.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) “Inference on treatment effects after selection among high-dimensional controls,” Review of Economic Studies, 81(2), pp. 608–650.