二重選択法
機械学習

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

2025-10-28

1 復習

1.1 研究計画

1. 研究目標: データ分析から、**社会**のどのような特徴を知りたいか
2. 推定目標: **データ**を使って、**母集団**のどのような特徴を知りたいか
3. 推定方法: **R/Python/Execel 等**を使って、**データ**のどのような特徴を計算したいのか

1.2 研究目標

- **予測研究**: Y (取引価格)の予測に役に立つような特徴を知りたい
- **比較研究**: Y の D (改築済み VS 未改築)間での違いを知りたい
  - **バランス後の比較**: X (物件の属性)が同じマンションで比較したい

1.3 推定目標

- 予測/比較ともに、**母平均**は有力な推定目標
  - 予測: 母平均 $E[Y \mid X]$ は、最善の予測モデル
  - 比較: X をバランスさせた母平均の差 $E[Y \mid \textbf{改築済み}, X] - E[Y \mid \textbf{未改築}, X]$

1.4 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 $E[Y \mid X]$ は複雑な関数であることが予想されるが、事例数は限られている
- 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

1.5 予測 VS 比較

- $$E[Y \mid D, X] = \underbrace{\beta_D \times District}_{D\text{に関する部分}}$$
$$+\underbrace{\beta_0 + \beta_1 \times Tenure + \beta_2 \times Distance}_{X\text{に関する部分}}$$
- 比較: D に関する部分のみを正確に推論したい
  - 信頼区間も計算したい
- 予測: D も X も同じくらい重要

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: $Distance$は”重要ではないので”モデルから除外する

$$E[Y \mid D, X] = \beta_D \times District$$
$$+\beta_0 + \beta_1 \times Tenure + \underbrace{\beta_2 \times Distance}_{=0}$$

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た結果、得られる β_D と同じような値を計算したい
- 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

学籍番号	学年	テスト	欠席の有無
1	2	100	0
2	3	60	0
3	4	70	0
4	3	60	1
5	3	60	1

学籍番号	学年	テスト	欠席の有無
6	1	80	1

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(二重選択法)を活用

3 LASSO を用いた二重選択法

3.1 アイディア

- X の中から **重要な変数** を選ぶ
  - 予測のための変数選択が行われる LASSO を利用
- 機械学習/AI も、“ミスを犯す可能性”を考慮する

3.2 コード例

```
Y <- data$Price

D <- data$District港区

X <- model.matrix(
  ~ 0 + Size + Tenure + Distance +
    I(Size^2) + I(Tenure^2) + I(Distance^2),
  data = data
)
```

3.3 コード例

```
PDS = hdm::rlassoEffect(
  y = Y,
  d = D,
  x = X)

summary(PDS)
```

```
[1] "Estimates and significance testing of the effect of target variables"
   Estimate. Std. Error t value Pr(>|t|)
d1    58.628      4.041   14.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 コード例

- X の選択結果

```
PDS$selection.index
```

```
         Size        Tenure      Distance     I(Size^2)   I(Tenure^2)
        FALSE          TRUE          TRUE          TRUE         FALSE
I(Distance^2)
         TRUE
```

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 二重選択

- 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(
  Price ~ Size + Tenure + Distance +
```

```
    I(Size^2) + I(Tenure^2) + I(Distance^2),
  data = data)

model_Y$index
```

```
         Size        Tenure      Distance     I(Size^2)   I(Tenure^2)
        FALSE          TRUE          TRUE          TRUE         FALSE
I(Distance^2)
         TRUE
```

3.8 イメージ

```
model_D = hdm::rlasso(
  District港区 ~  Size + Tenure + Distance +
    I(Size^2) + I(Tenure^2) + I(Distance^2),
  data = data)

model_D$index
```

```
         Size        Tenure      Distance     I(Size^2)   I(Tenure^2)
        FALSE         FALSE          TRUE          TRUE         FALSE
I(Distance^2)
        FALSE
```

3.9 イメージ

```
lm(
  Price ~ District港区 + Tenure + Distance + I(Size^2) + I(Distance^2),
  data = data)
```

```


Call:
lm(formula = Price ~ District港区 + Tenure + Distance + I(Size^2) +
    I(Distance^2), data = data)

Coefficients:
  (Intercept)   District港区         Tenure       Distance      I(Size^2)
     28.33512       58.62804       -1.73680       -3.48083        0.02595
I(Distance^2)
      0.03898
```

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目

3.13 Reference

Bibliography