

変数選択

労働経済学 (補論)

川田恵介

Table of contents

1	変数選択	2
1.1	問題設定	2
1.2	アイデア	2
1.3	仮定: Sparsity	2
2	Penalized Regression	3
2.1	LASSO Algorithm	3
2.2	Constrained optimization としての書き換え	3
2.3	λ の役割: OLS	3
2.4	λ の役割: 平均	3
2.5	λ の役割	4
2.6	変数選択	4
2.7	注意点	4
3	Double Selection	4
3.1	Naive なアイデア	5
3.2	問題点	5
3.3	Double Selection Algorithm	5
3.4	実装	5
3.5	実践	5
3.6	Example: Bonaccolto-Töpfer and Briel (2022)	6
4	Example: CPS1988	6
4.1	Data	6
4.2	推定方法	6
4.3	Comparison	6
4.4	Selected Variable	7
	Reference	7

1 変数選択

- $\tau = E[Y|D = d', X] - E[Y|D = d, X]$ を近似する $Y = \beta_D D + \beta_0 + \beta_1 Z_1 + \dots + \beta_L Z_L$ を明確な統計的な基準に基づいて選ぶ
 - OLS では、事前に研究者が選ぶ必要がある!!!

1.1 問題設定

- 労働研究では、バランスさせたい X が大量に存在するケースも多い
- 関心のある比較は、 $E[Y|D = d', X] - E[Y|D = d, X]$
 - 例: $Y =$ 年収、 $D =$ 性別、 $X =$ ついている仕事の特徴
 - * 同じ働き方をしている男女内賃金格差
- X に大量のデータが含まれているケースがある
 - 例: 労働時間、勤続年数、業務内容、それらの交差項...
 - * 全てを Balance させることができない/推定精度が大幅に悪化する

1.2 アイディア

- X 全てが”重要” なのわけではないかもしれない
 - X の一部 Z のみをバランスさせれば十分
 - $E[Y|D = d', X] - E[Y|D = d, X] \simeq E[Y|D = d', Z] - E[Y|D = d, Z]$

1.3 仮定: Sparsity

- $$E[Y|D, X] = \tau D + \beta_0 + \underbrace{\beta_1 X_1 + \dots + \beta_L X_L}_{L > \text{事例数でも OK}}$$
- (Approximately) sparsity: 事例数に比べて、十分に少ない変数数 $S \ll$ 事例数で、母平均をうまく近似できる
- 実戦: 十分に複雑なモデルについて LASSO を推定し、変数選択
 - もともとのモデルには、“trivial” な変数も含まれていると仮定
 - 詳細は [Capter 4 in Causal ML 参照](#)

2 Penalized Regression

- データ主導の変数選択については、多くの提案がなされている
 - 代表例は LASSO (機械学習の代表的な手法)

2.1 LASSO Algorithm

- $E[Y|X]$ を近似するモデル $g(X)$ を推定
0. 十分に複雑なモデルからスタート: 例: X について二乗項と交差項を作成
 1. 何らかの基準 (後述) に基づいて Hyper (Tuning) parameter λ を設定
 2. 以下の最適化問題を解いて、Linear model $g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$ を推定

$$\min \sum (y_i - g(x_i))^2 + \lambda(|\beta_1| + |\beta_2| + \dots)$$

2.2 Constrained optimization としての書き換え

1. 何らかの基準 (後述) に基づいて Hyper parameter λ を設定
2. 以下の最適化問題を解いて、Linear model $g(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$ を推定

$$\min \sum (y_i - g(x_i))^2$$

where

$$|\beta_1| + |\beta_2| + \dots \leq A$$

2.3 λ の役割: OLS

- $\lambda = 0$ と設定すれば、(複雑なモデルを) OLS で推定した推定結果と一致
 - すべての変数が活用される

2.4 λ の役割: 平均

- $\lambda = \infty$ と設定すれば、必ず $\beta_1 = \beta_2 = \dots = 0$ となる
 - β_0 のみ、最小二乗法で推定: $g(X) = \text{サンプル平均}$
- すべての変数が排除される

2.5 λ の役割

- やりたい事: $E[Y|X]$ を上手く近似するように λ を設定し、単純すぎるモデル (Approximation error が大きすぎる) と複雑すぎるモデル (Estimation error が大きすぎる) の間の”ちょうどいい”モデルを構築する
- 設定方法: 理論値 ([hdm](#) で採用)
 - サンプル分割 (交差推定, [glmnet](#) で実装)/情報基準 ([gamlr](#) で採用) などもあり

2.6 変数選択

- いくつかの変数について、係数値 β が厳密にゼロになるうる
 - モデルから除外される
 - * OLS では”生じない”

2.7 注意点

- LASSO および他の Penlized Regression (Ridge など) によって推定された係数値について、解釈を与えることは難しい
 - $E[Y|X]$ の近似が目的であり、係数値について明確な母集団上の解釈がない
 - * Y の予測が目的であれば、優れた手法
 - Y とそこそこの関係性がある変数であったとしても、除外される可能性がある
 - 信頼区間の計算も難しい
- さらに学びたい人は、Chap 1 in [CausalML](#), in Chap 6 in [ISL](#) などを参照

3 Double Selection

- 変数選択には一般にミスが生じる
 - 重要な変数を除外してしまうリスクがある
- リスクを緩和するために、“ダブルチェック”を行う必要がある
- Belloni, Chernozhukov, and Hansen (2014)
 - Gentle introduction: Angrist and Frandsen (2022)

3.1 Naive なアイデア

- X を全てバランスさせるのではなく、 Y との相関が強いものだけをバランスさせる
 - $g_Y(X)$ を LASSO で推定し、選択された変数だけを OLS に加える

3.2 問題点

- 問題点: LASSO による変数選択は、 Y とそこそこ相関がある変数も除外されてしまう可能性がある
 - $E[Y|X]$ の近似が目的であれば、(Tuning parameter が正しく選ばれている限り)、許容できる
- D との相関が強い (分布が Unbalance) 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
 - τ の推定という目標に対して、モデルが過度に単純化される (Regularization bias)

3.3 Double Selection Algorithm

1. $g_Y(X)$ および $g_D(X)$ を LASSO で推定し、選択された変数を記録
 2. どちらかの予測モデルで選択された変数 (Z) のみを用いて、 $Y \sim D + Z$ を回帰
- Y の予測モデルと D の予測モデルによる”ダブルチェック”

3.4 実装

- hdm package が有益

```
rlassoEffect(  
  x = X, # Must be matrix  
  d = D, # Must be vector  
  y = Y # Must be vector  
)
```

- 注: Tuning parameter は、交差推定ではなく、理論値を使用

3.5 実践

- かなり制約的なアプローチ (Variable selection を行う Algorithm しか使えない)
 - より柔軟なアプローチが多く提案
- 今でも多くの応用研究が、Robustness check として活用

- 最終的には OLS なので、Editor/Referee に理解させやすい!?
- すぐに活用できるという意味で、十分に実践的
- * OLS でコントロールしている自身の研究があれば、使ってみてください!!!

3.6 Example: Bonaccolto-Töpfer and Briel (2022)

- 就業形態や教育歴、家族背景等をバランスさせたもとの、男女間賃金格差を推定
 - 二乗項と交差項を加えて、9045 変数が元々のコントロール変数
 - Double Selection により、5,821 変数を選択

4 Example: CPS1988

4.1 Data

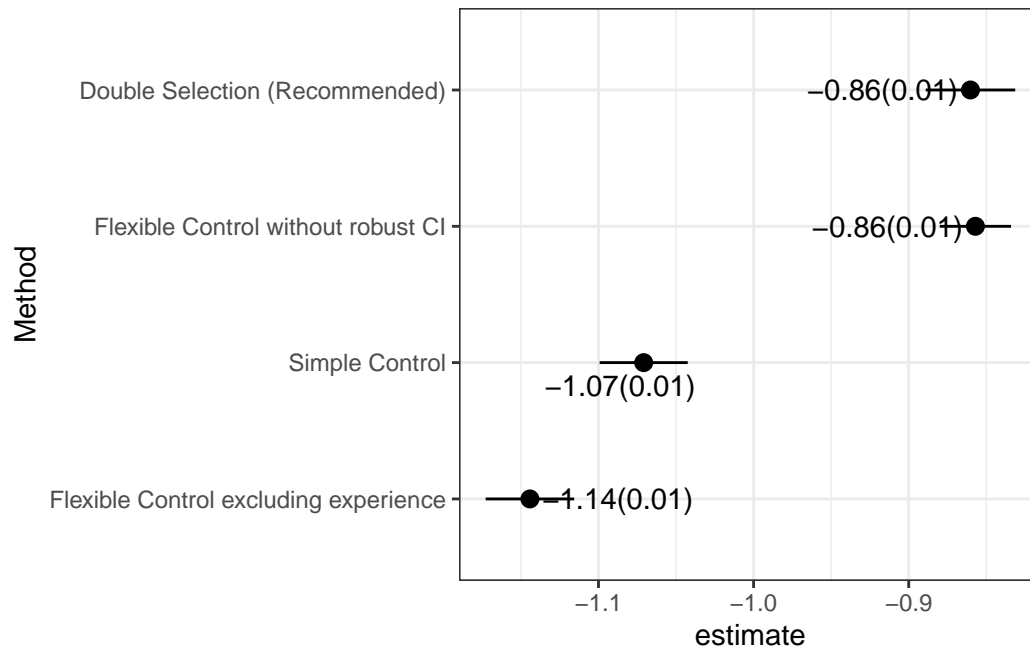
- Use CPS1988 from AER package
 - Sample size 28155
 - $Y = \log \text{ wage}$, $D = \text{partime}$ (Parttime wage penalty)
 - $X = \text{BaseLine (education, ethnicity, smsa, region) + experience}$
 - * 7 variables

4.2 推定方法

- Flexible control excluding experience: BaseLine と、その二乗項と交差項をコントロールし、OLS 推定
- Simple control: BaseLine + experience。ただし二乗項や交差項は除外
- Flexible control without robust CI: BaseLine + experience と、その二乗項と交差項をコントロールし、OLS 推定
 - Robust stadard error は計算できなかったなので、classical standard error を報告
- Double selection: Double selection で変数選択

4.3 Comparison

- 二乗項や交差項を加えることで、経験年数を追加的にバランスされた推定結果が顕著に異なる



4.4 Selected Variable

- 教育年数や学歴の二乗項も残る

education	experience	ethnicityafam
TRUE	TRUE	TRUE
regionsouth	I(education^2)	I(experience^2)
TRUE	TRUE	TRUE
education:experience	education:smsayes	education:regionwest
TRUE	TRUE	TRUE
experience:smsayes	smsayes:regionsouth	
TRUE	TRUE	

Reference

- Angrist, Joshua D, and Brigham Frandsen. 2022. "Machine Labor." *Journal of Labor Economics* 40 (S1): S97–140.
- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *Review of Economic Studies* 81 (2): 608–50.
- Bonaccolto-Töpfer, Marina, and Stephanie Briel. 2022. "The Gender Pay Gap Revisited: Does Machine Learning Offer New Insights?" *Labour Economics* 78: 102223. <https://doi.org/10.1016/j.labeco.2022.102223>.