

Linear Model for Description

川田恵介

Table of contents

1	線形モデルによる記述	2
1.1	Linear Model による記述	2
1.2	OLS	2
1.3	数値例: OLS	3
1.4	例: OLS (独立の場合)	3
1.5	例: OLS (相関の場合)	4
1.6	LASSO	4
1.7	例: LASSO	5
1.8	まとめ	5
2	比較	5
2.1	シンプルな比較研究	5
2.2	シンプルな比較研究	6
2.3	踏み込んだ研究課題: 差の理由	6
2.4	例: Dube et al. (2020)	6
2.5	伝統的な方法	6
3	OLS の別解釈	7
3.1	OLS Algorithm: 単回帰	7
3.2	OLS Algorithm: General case	7
3.3	OLS の解釈	7
4	Constant Difference モデルによる解釈	8
4.1	Constant Difference	8
4.2	単回帰の解釈	8
4.3	単回帰の解釈	8
4.4	母集団への含意 (事例数無限大)	9
4.5	重回帰の解釈	9
4.6	重回帰の解釈	9

4.7	母集団への含意	9
4.8	Mis-specification	10
4.9	Overfit	10
4.10	まとめ	10
5	Double Selection	10
5.1	Naive なアイデア	10
5.2	問題点	10
5.3	Double Selection Algorithm	11
5.4	重要な仮定: Sparsity	11
5.5	実装	11
5.6	実践	11
5.7	Next Step	12
	Reference	12

1 線形モデルによる記述

- Y と X の関係性を”簡潔な”関数で記述する
 - 非常に難しいチャレンジであり、(私見では) 経済学においては主流ではない

1.1 Linear Model による記述

- $$g_Y(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_L X_L$$
 - 記述モデルの研究者が設定
- “モデル上の”関係性は容易に解釈可能
 - $\beta_1 = X_1$ が一単位大きい時に、 Y の平均値はどのくらい大きいか?

1.2 OLS

- X の数が少ない (記述モデルがシンプル) であれば、OLS は有効
 - 信頼区間も導出可能
 - * 多重検定への対応は必要 (ISL Chap 13 等参照)
- Estimand (Best Linear Projection) $\neq E[Y|X]$ であり、必ずしも直感的ではないことに注意

1.3 数値例: OLS

$$\bullet Y = \underbrace{D}_{\in \{0,1\}} + \underbrace{X^2}_{\sim Uniform(-2,2)} + \underbrace{u}_{\sim N(0,10)}$$

– 独立している場合:

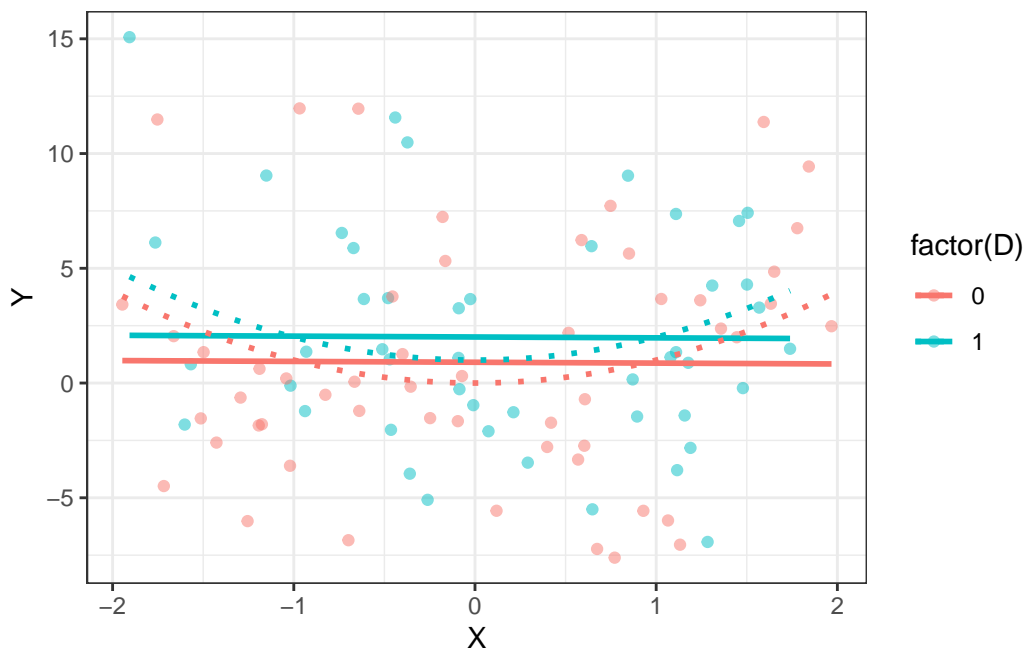
$$\ast \Pr(D = 1) = 0.5$$

– 相関している場合:

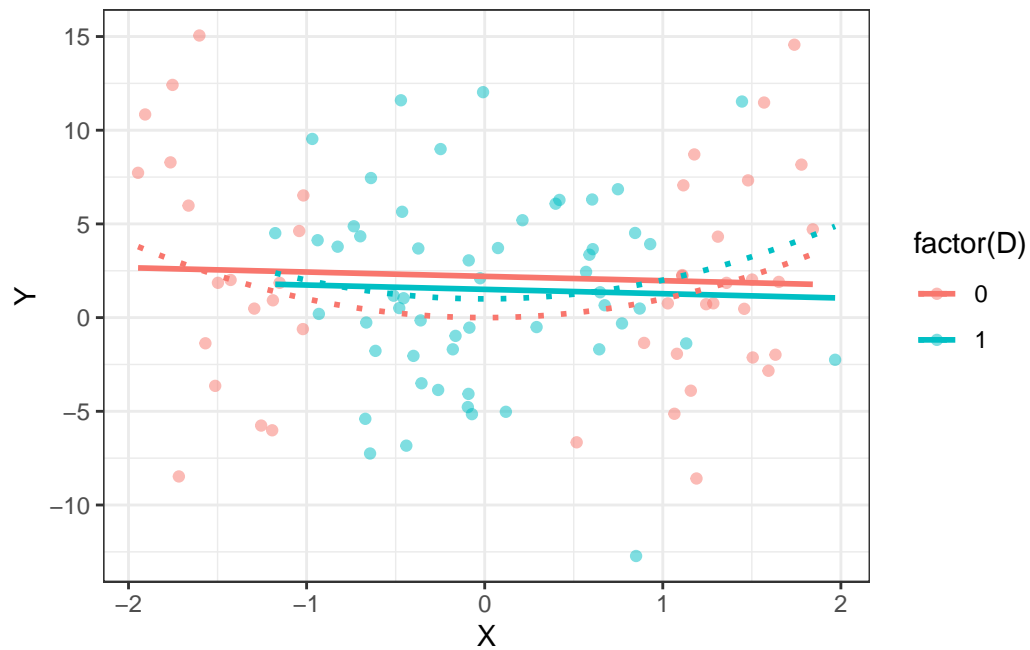
$$\ast \Pr(D = 1 | 1 \geq X \geq -1) = 0.9$$

$$\ast \Pr(D = 1 | X \geq 1 | X \leq -1) = 0.1$$

1.4 例: OLS (独立の場合)



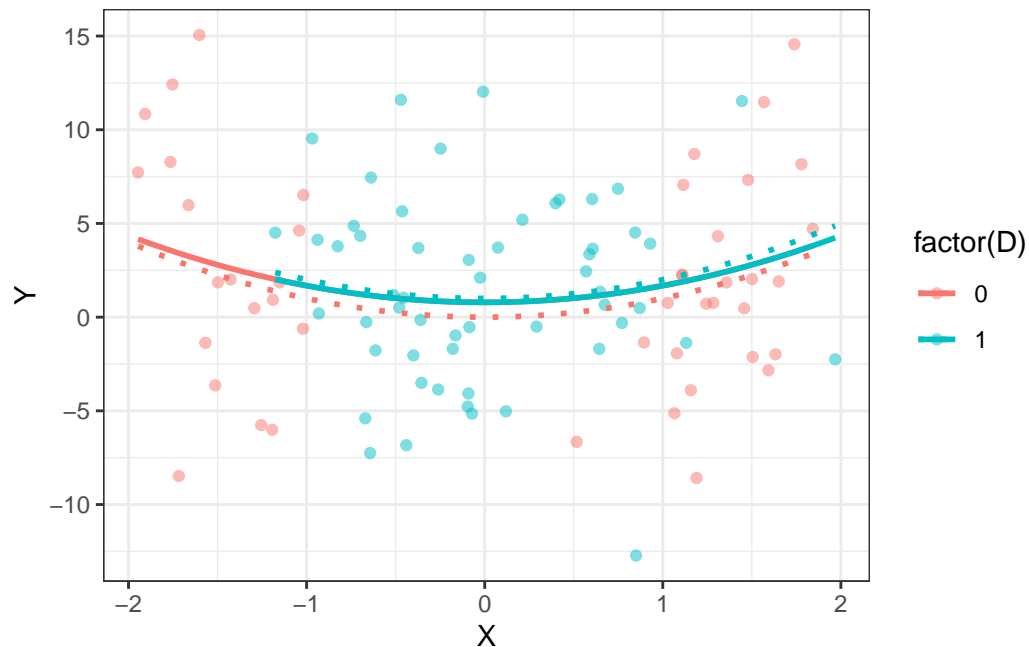
1.5 例: OLS (相関の場合)



1.6 LASSO

- X の数が多くても、変数の数を減らしてくれるので、一見良さそうだが、
 - 信頼区間の導出が難しい
 - 変数選択の精度はそこまで高くない

1.7 例: LASSO



1.8 まとめ

- Y と大量の X の関係性を記述する、は非常に難しい課題
 - X の中から特に興味がある変数 D を選んで、 Y との関係性を記述することが現実的

2 比較

- 本講義では、研究課題の段階で、関心とする変数を絞り込むことを推奨
 - 比較研究に持ち込む

2.1 シンプルな比較研究

- 特定の Y と D の関係性について関心があるケースが多い
 - 例: 男女間賃金格差
 - * $Y = \text{Wage}, D = \text{Gender}$
 - 年功型賃金体系の程度

* $Y = Wage, D = Tenure$

2.2 シンプルな比較研究

- 有力な Estimand は、母平均の差 $E[Y|D=1] - E[Y|D=0]$ ないし、Population OLS $Y \sim D$
 - データ上で Y を D で回帰すれば OK
- Y と関係していそうな X がデータに含まれていたとしても、“無視” して良い

2.3 踏み込んだ研究課題: 差の理由

- なぜ差が生まれるのか?
 - データから観察可能な他の変数 X に注目
 - * X についての格差が、 Y の差をもたらしている
 - * X 以外についての格差が、 Y の差をもたらしている
- 注目されてきた Estimand であり、多様な方法論開発が進む
 - 機械学習の導入も有力視されている

2.4 例: Dube et al. (2020)

- Online 労働市場において、求人が提示する賃金水準 ($= D$) と応募者数 ($= Y$) はどのような関係にあるのか?
 - 賃金が高い仕事は、高い技能が要求される/きつい... ($= X$) 可能性がある
 - 求人内容 ($= X$) 以外の要因で、どの程度の差が生じているのか?
 - * 労働市場の不完全性 (独占力) の指標 (Langella and Manning 2021)

2.5 伝統的な方法

- X をコントロールする: 以下を推定

$$g_Y(D, X) = \underbrace{\beta_D}_{=X以外による格差} D + \underbrace{\beta_0 + \beta_1 X_1 + \dots}_{Xの影響を除去}$$

- 課題
 - 何が Estimand なのか?
 - * コントロールとは?

- 定式化の影響は?

3 OLS の別解釈

- $g_Y(D) = \beta_0 + \beta_D D + \beta_1 X_1 + \dots$ を OLS で推定する
 - 議論の簡略のために、 X は標準化されているとする
- (BLP ではなく) Weight 推定としても再解釈できる

3.1 OLS Algorithm: 単回帰

- $g_Y(D) = \beta_D D + \beta_0$ を OLS 推定すると

$$\beta_D = \sum_{i:D_i=1} \underbrace{\frac{1}{N_1}}_{=W_i} Y_i - \sum_{i:D_i=0} \underbrace{\frac{1}{N_0}}_{=W_i} Y_i$$

- 事例数の逆数を Weight とした比較と解釈できる

$$\text{– 注: } \sum_{i:D_i=1} W_i = \sum_{i:D_i=0} W_i = 1$$

3.2 OLS Algorithm: General case

1. 全ての $X = [X_1, \dots, X_L]$ について、

$$\begin{aligned} \sum_{i:D_i=1} W_i X_{i,l} &= \sum_{i:D_i=0} W_i X_{i,l}, \\ \sum_{i:D_i=1} W_i &= \sum_{i:D_i=0} W_i = 1 \end{aligned}$$

を満たす W から、分散が最も小さいものを選ぶ

- 2.

$$\beta_D = \sum_{i:D_i=1} W_i Y_i - \sum_{i:D_i=0} W_i Y_i$$

3.3 OLS の解釈

- Y の Weight 付き平均差として解釈できる
 - データ上で、 D 間で X の平均値が”Balance”するように Weight は選ぶ
- X^2 もモデルに加えれば、 X^2 の平均値 (分散) も等しくなるように選ばれる
- $X_1 * X_2$ もモデルに加えれば、 X_1, X_2 の共分散も等しくなるように選ばれる

4 Constant Difference モデルによる解釈

- 母集団に対する、かなり強い仮定を用いて、OLS の推定結果を解釈
- 注記: 不必要に強い仮定であり、将来緩める

4.1 Constant Difference

- $E[Y|1, X] - E[Y|0, X] = \tau$ を母集団上で仮定

– $\tau = X$ が全く同じ集団間での平均格差

* ” X をコントロール/Ceteris paribus”

- 以下のモデルで表現できる

$$Y = \tau \times D + \underbrace{h(X)}_{\text{なんらかの関数}} + \underbrace{u}_{=E[u|X]}$$

– Semiparametric estimation では、 $h(X)$ は Nuisance function と呼ばれる

4.2 単回帰の解釈

- Y を代入すると

$$\begin{aligned}\beta_D &= \frac{\sum_{i:D_i=1} Y_i}{N_1} - \frac{\sum_{i:D_i=0} Y_i}{N_0} \\ &= \frac{\sum_{i:D_i=1} (\tau_D + h(X_i) + u_i)}{N_1} \\ &\quad - \frac{\sum_{i:D_i=0} (h(X_i) + u_i)}{N_0}\end{aligned}$$

4.3 単回帰の解釈

$$\begin{aligned}\beta_D &= \tau_D + \underbrace{\left[\frac{\sum_{i:D_i=1} h(X_i)}{N_1} - \frac{\sum_{i:D_i=0} h(X_i)}{N_0} \right]}_{\text{属性のずれ}} \\ &\quad + \underbrace{\left[\frac{\sum_{i:D_i=1} u_i}{N_1} - \frac{\sum_{i:D_i=0} u_i}{N_0} \right]}_{\text{観察できない属性のずれ}}\end{aligned}$$

4.4 母集団への含意 (事例数無限大)

$$\begin{aligned}\beta_D = \tau_D + \beta_X & \left[\frac{\sum_{i:D_i=1} h(X_i)}{N_1} - \frac{\sum_{i:D_i=0} h(X_i)}{N_0} \right] \\ & \xrightarrow[N_1, N_0 \rightarrow \infty]{} E_X[h(X)|D=1] - E_X[h(X)|D=0] \\ & + \underbrace{\frac{\sum_{i:D_i=1} u_i}{N_1} - \frac{\sum_{i:D_i=0} u_i}{N_0}}_{\rightarrow 0}\end{aligned}$$

- 観察できる属性のずれの影響が残る

4.5 重回帰の解釈

- Y を代入すると

$$\begin{aligned}\beta_D &= \sum_{i:D_i=1} W_i Y_i - \sum_{i:D_i=0} W_i Y_i \\ &= \sum_{i:D_i=1} W_i (\tau_D + h(X_i) + u_i) \\ &\quad - \sum_{i:D_i=0} W_i (h(X_i) + u_i)\end{aligned}$$

4.6 重回帰の解釈

- Y を代入すると

$$\begin{aligned}\beta_D &= \tau_D \\ &+ \underbrace{\left[\sum_{i:D_i=1} W_i h(X_i) - \sum_{i:D_i=0} W_i h(X_i) \right]}_{h(X)=\beta_0+\beta_1 X \text{であれば}=0} \\ &\quad + \sum_{i:D_i=1} W_i u_i - \sum_{i:D_i=0} W_i u_i\end{aligned}$$

4.7 母集団への含意

$$\begin{aligned}\beta_D = \tau_D + \beta_X & \left[\underbrace{\sum_{i:D_i=1} W_i h(X) - \sum_{i:D_i=0} W_i h(X)}_{h(X)=\beta_0+\beta_1 X \text{であれば}=0} \right] \\ & + \underbrace{\sum_{i:D_i=1} W_i u_i - \sum_{i:D_i=0} W_i u_i}_{\rightarrow 0}\end{aligned}$$

4.8 Mis-specification

- $g_Y(D, X) = \beta_0 + \beta_D D + \beta_1 X$ を OLS 推定するが、 $h(X) = \beta_0 + \beta_1 X + \beta_2 X^2$
 - X の分散 (X^2) は Balance しないので、 β_D は τ_D に (事例数が無限大でも) 収束しない

4.9 Overfit

- Mis-specification を避けるためには、 X を十分に複雑にしてモデルに導入する必要がある
- より多くの変数の平均値を揃える必要があるので、Weight W_i の分散が大きくなる
- 特定の個人 (u_i) の影響が非常に強くなり、推定精度が悪化

4.10 まとめ

- OLS = X の平均値を Balance させる Algorithm
 - 高次項 ($X_1^2, X_1^3, X_1 \times X_2, \dots$) を導入すると、 X の分布を Balance させられる
 - 弊害: Weight の分散が大きくなり、推定精度が悪化する
- 課題: “重要な” X のみ Balance させたい

5 Double Selection

- LASSO の”副産物”である変数選択を利用
 - “AI”によるダブルチェックを行い、変数選択のミスを減らす
- Belloni, Chernozhukov, and Hansen (2014)
 - Gentle introduction: Angrist and Frandsen (2022)

5.1 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

5.2 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

- Y の予測のためであれば、(Tuning parameter が正しく選ばれている限り)、許容できる (Bias-variance Tradeoff)
- 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ついて、モデルが過度に単純化される (Regularization bias)

5.3 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”

5.4 重要な仮定: 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 $S < L$ 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

5.5 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

5.6 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)
 - 後日、より柔軟なアプローチを紹介

- 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!!!

5.7 Next Step

- ここまでの議論は以下に限定
 - Algorithm: Liner Model
 - Estimand: 平均値関数/平均差の推定
- 課題: より幅広い Algorithm (Tree/Stacking model)/Estimand (“Exact” Average Difference/Heterogeneity)

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.
- Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. “Monopsony in Online Labor Markets.” *American Economic Review: Insights* 2 (1): 33–46.
- Langella, Monica, and Alan Manning. 2021. “Marshall Lecture 2020: The Measure of Monopsony.” *Journal of the European Economic Association* 19 (6): 2929–57.