

OLS の比較分析への応用

労働経済学 2

川田恵介

Table of contents

1	過剰適合	2
1.1	OLS の問題点	2
1.2	数値例: 母平均	2
1.3	数値例: データ	3
1.4	数値例: 平均値による推定	3
1.5	含意	4
1.6	数値例: 平均値 VS シンプルなモデル	4
1.7	数値例: 平均値 VS 複雑なモデル	5
1.8	含意	5
1.9	Takeaway	5
1.10	Takeaway: 過剰適合	6
1.11	課題	6
1.12	数値例: 50000 事例	6
1.13	数値例: 平均値 VS シンプルなモデル	7
1.14	数値例: 平均値 VS 複雑なモデル	7
2	LASSO	8
2.1	実装	8
2.2	実装: OLS	8
2.3	実装: LASSO	9
2.4	罰則付き回帰	10
2.5	罰則付き回帰の基本手順	10
2.6	入門経済学による例え話	10
2.7	罰則の定式化	11
2.8	λ の設定	11
2.9	伝統的な推定方法との関係性	11
2.10	罰則付き回帰の問題点	11

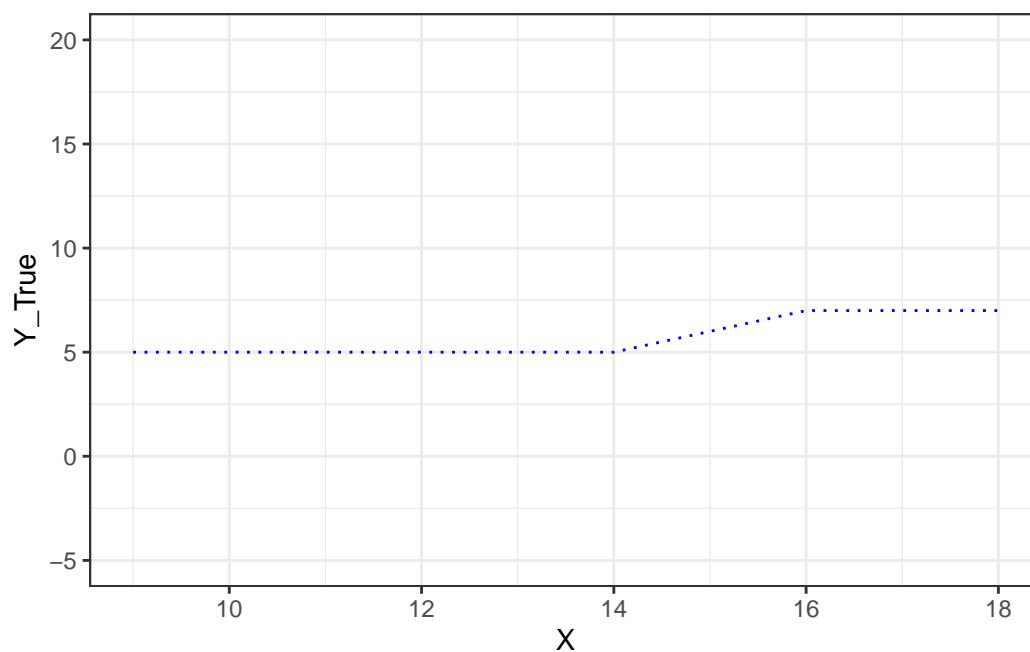
2.11	Takeaway	11
2.12	よくある誤解: 完璧なモデル	12
2.13	よくある誤解: 予測と比較	12
	Reference	12

1 過剰適合

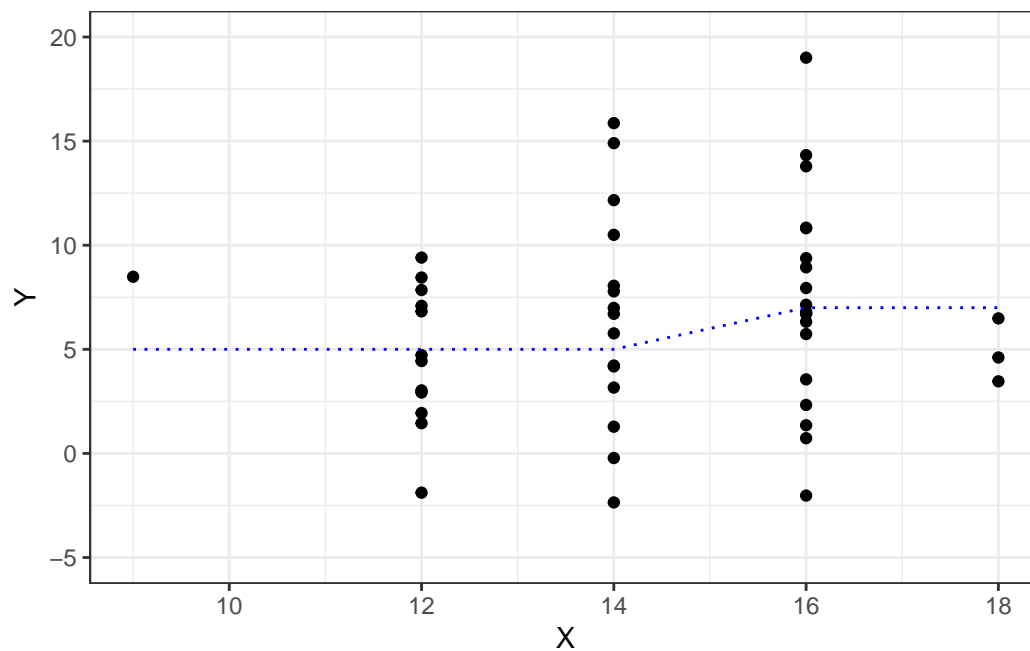
1.1 OLS の問題点

- 事例数に比べて、単純なモデルを推定するのであれば、OLS は有効な選択肢
 - 実用的な推定精度
 - 信頼区間の計算
- 複雑なモデルを推定しようとする、データ上の平均値に近づくが、母平均との乖離が大きくなる
 - 過剰適合/過学習
 - * データではなく、母集団を関心とするのであれば、大問題

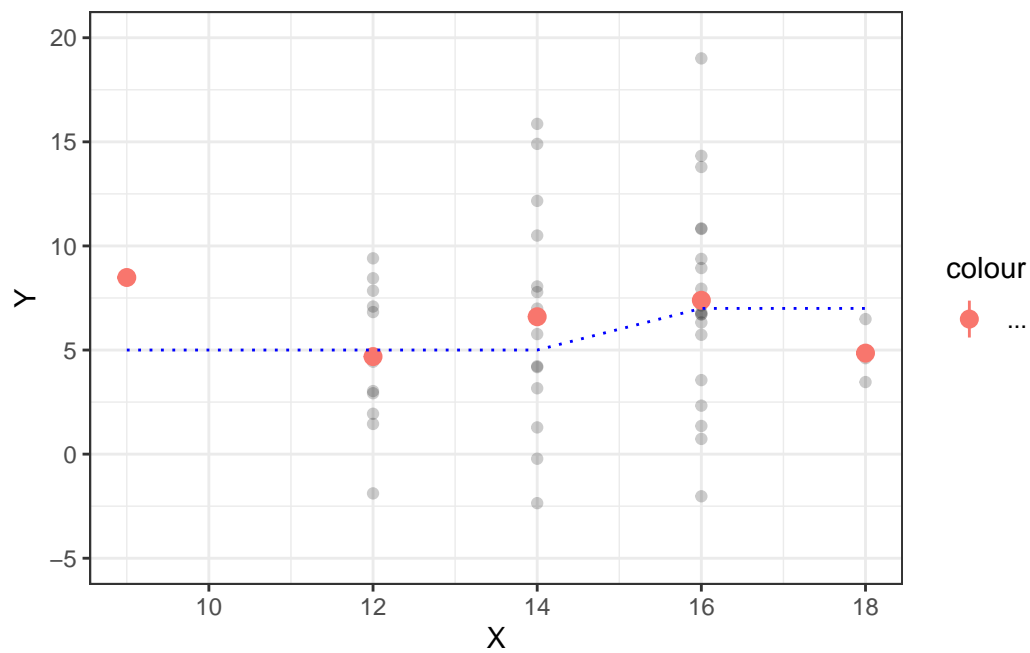
1.2 数値例: 母平均



1.3 数値例: データ



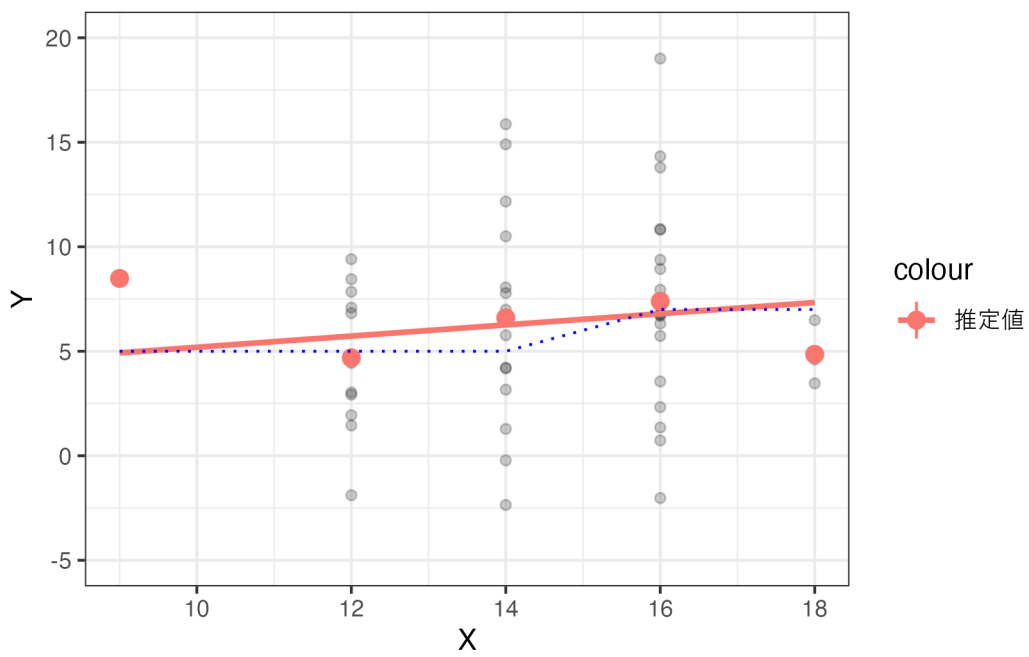
1.4 数値例: 平均値による推定



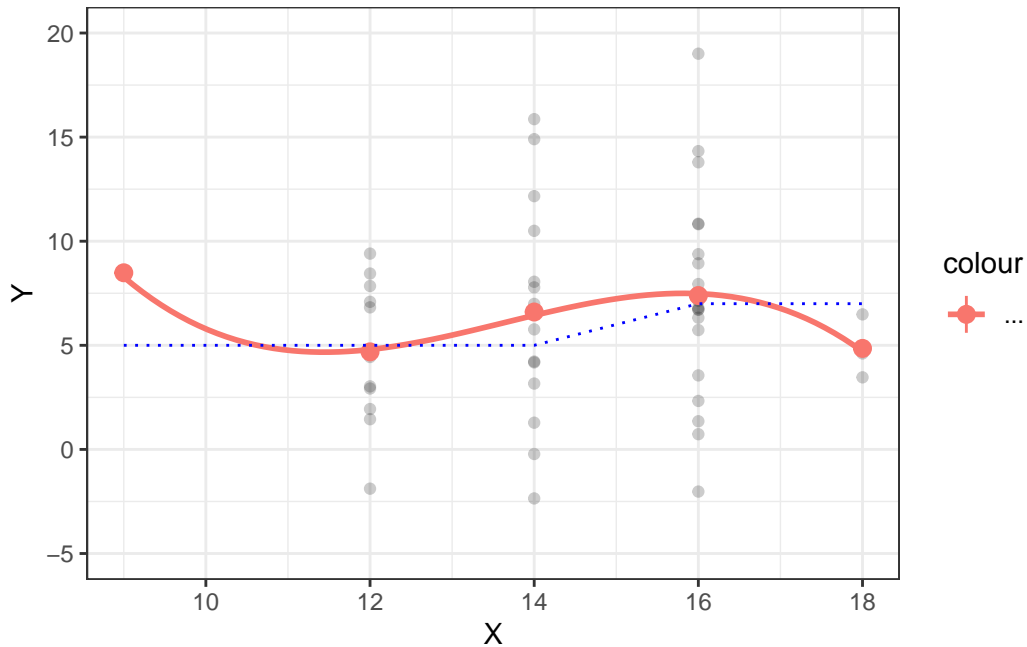
1.5 含意

- データ上の平均値は、母平均から基本的に乖離する
 - 事例数が少ないグループ ($X = 9, 16$) において特に乖離する

1.6 数値例: 平均値 VS シンプルなモデル



1.7 数値例: 平均値 VS 複雑なモデル



1.8 含意

- シンプルなモデルは、データ上の平均値よりも、母平均に近い
 - 線型モデルへの仮定 (Y は X に応じて、“緩やかに変化する” (smoothness)) が事例の少なさを補う
- 複雑なモデルは、データ上の平均値と一致する
 - シンプルなモデルよりも、母平均から乖離している

1.9 Takeaway

- OLS は、研究者が設定したモデルを、データに極力適合するように推定する
 - 複雑なモデルを用いると、データ上の平均値を”なぞった”モデルが推定される
 - * データにより適合する
- データ上の平均値と母平均は、基本的に乖離する
 - 事例数が限られており、上振れ/下振れする

* 特に少数しかないカテゴリ ($X = 9, 18$)

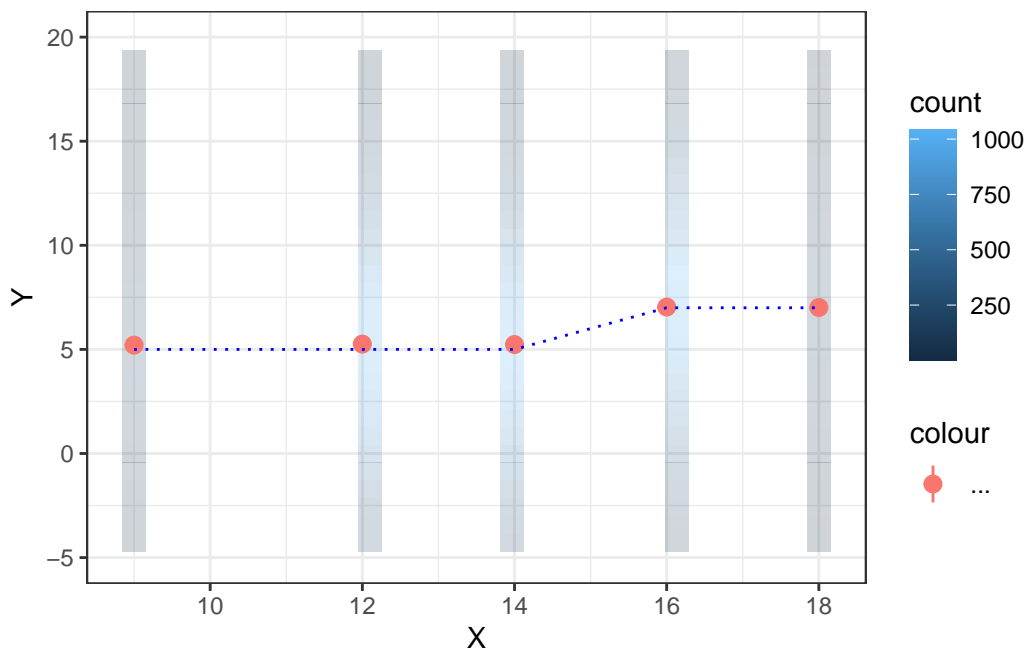
1.10 Takeaway: 過剰適合

- 適度に複雑なモデルを OLS で推定: データ上の平均と母平均を、ほどほど近似する
- 複雑すぎるモデル: データ上の平均をよく近似するが、母平均から乖離する
 - 過剰適合
- 単純すぎるモデル: データ上の平均と母平均から、乖離する
 - 過小適合

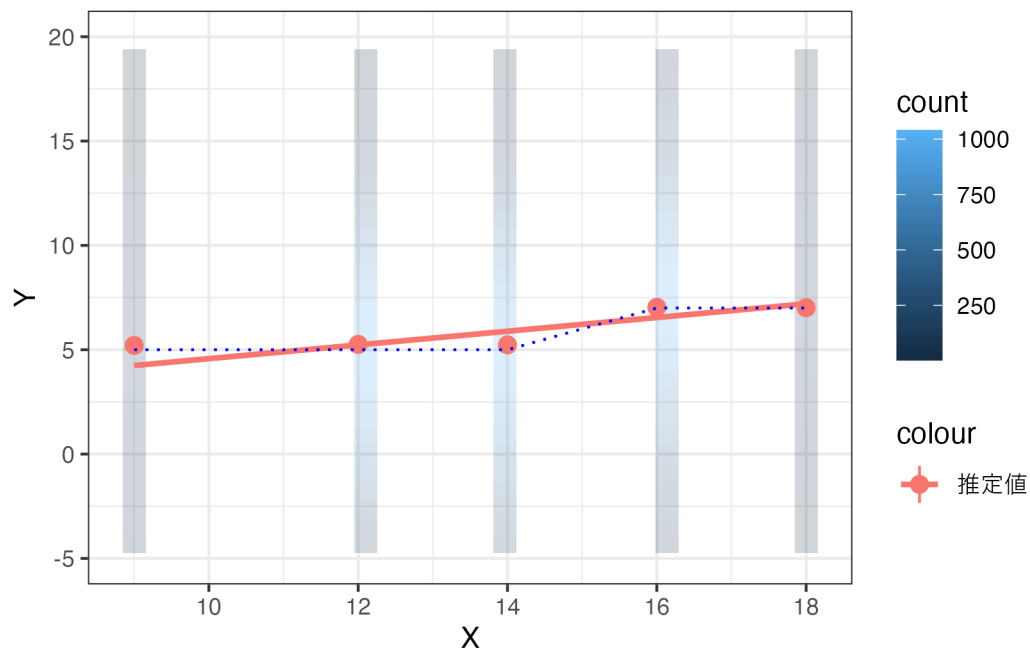
1.11 課題

- 実践において、適度に複雑なモデルを定式化することが非常に難しい
 - 特に X の数が多い場合は、ほぼ不可能
 - “(T)here are parts of the (statistical) model where economic theory is silent” (Imbens and Athey 2021) の代表例
 - * () は川田が補った部分

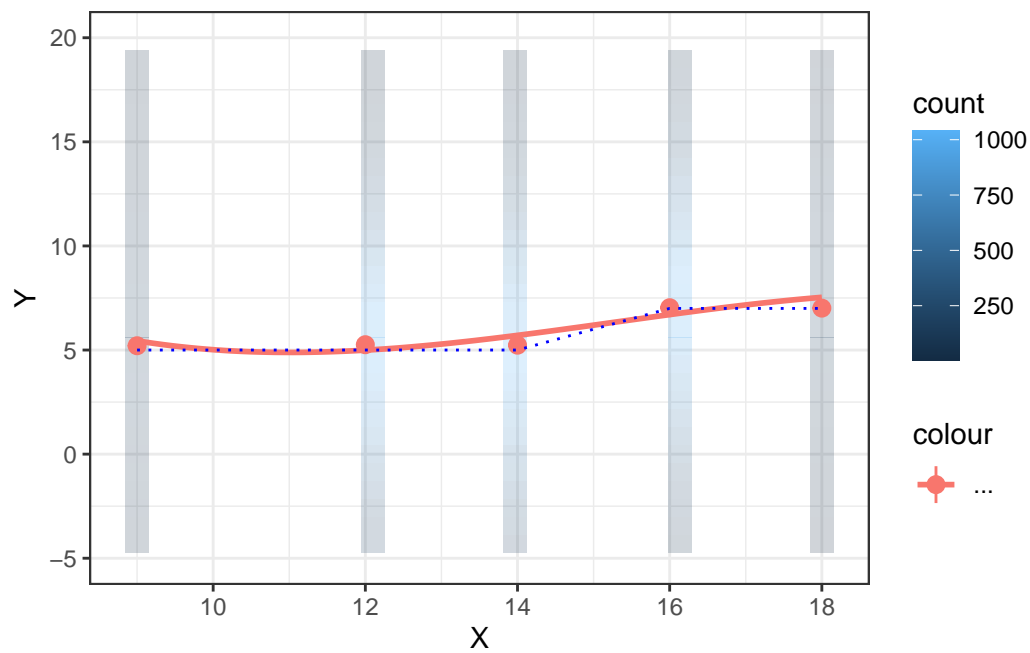
1.12 数値例: 50000 事例



1.13 数値例: 平均値 VS シンプルなモデル



1.14 数値例: 平均値 VS 複雑なモデル



2 LASSO

2.1 実装

```
library(tidyverse)

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

data <- mutate(
  CPS1985,
  Y = log(wage), # log of wage
  D = if_else(
    occupation == "technical",
    1,
    0
  ) # occupation dummy
)
```

2.2 実装: OLS

```
lm(Y ~ D +
    (education + ethnicity + region + gender)^2 +
    I(education^2) + I(age^2),
    data
)
```

Call:

```
lm(formula = Y ~ D + (education + ethnicity + region + gender)^2 +
    I(education^2) + I(age^2), data = data)
```

Coefficients:

(Intercept)	D
1.2342971	0.1432014
education	ethnicityhispanic
0.0746214	-0.1399608
ethnicityother	regionother
0.1422315	-0.4574510

genderfemale	I(education ²)
-0.8052570	-0.0015238
I(age ²)	education:ethnicityhispanic
0.0001293	-0.0117403
education:ethnicityother	education:regionother
-0.0259754	0.0388527
education:genderfemale	ethnicityhispanic:regionother
0.0336461	0.3266717
ethnicityother:regionother	ethnicityhispanic:genderfemale
0.0948690	-0.0006260
ethnicityother:genderfemale	regionother:genderfemale
0.1618201	0.1242733

2.3 実装: LASSO

```
hdm::rlasso(Y ~ D +
  (education + ethnicity + region + gender)2 +
  I(education2) + I(age2),
  data,
  post = FALSE
)
```

Call:

```
rlasso.formula(formula = Y ~ D + (education + ethnicity + region +
  gender)2 + I(education2) + I(age2), data = data, post = FALSE)
```

Coefficients:

(Intercept)	D
1.318e+00	1.187e-01
education	ethnicityhispanic
3.779e-02	0.000e+00
ethnicityother	regionother
0.000e+00	0.000e+00
genderfemale	I(education ²)
-1.823e-01	6.254e-04
I(age ²)	education:ethnicityhispanic
8.555e-05	0.000e+00
education:ethnicityother	education:regionother

	0.000e+00	7.712e-03
education:genderfemale		ethnicityhispanic:regionother
	0.000e+00	0.000e+00
ethnicityother:regionother		ethnicityhispanic:genderfemale
	0.000e+00	-4.452e-02
ethnicityother:genderfemale		regionother:genderfemale
	0.000e+00	0.000e+00

2.4 罰則付き回帰

- OLS の問題点: 定式化のもとで、データへの適合のみを目指して推定する
 - 過剰適合を避けるには、定式化を単純するしかない
- 罰則付き回帰のアイディア: 複雑性への罰則を与えることで、複雑に定式化されたモデルを、過剰適合を減らしながら推定できる
 - LASSO,Ridge,elastic net など

2.5 罰則付き回帰の基本手順

- Step 1. 研究者 が線型モデル $\beta_0 + \beta_1 X_1 + ..$ を設定
- Step 2. 以下を最小化するように、線型モデルの β を算出

データへの不適合度 + 複雑性への罰則
- 「モデルを複雑にしない」ことも”目的関数”に加える

2.6 入門経済学による例え話

- 生産方法を企業の自主的な意思決定に任せると
 - 同じ生産量を達成する方法の中で、最も費用が少ない方法が選ばれる
 - * 希少な資源の利用を減らせ、それなりに望ましい
 - 一般に”負の外部性”が生じる
 - * 温室効果ガスの過剰排出等
 - 社会的に望ましい水準に誘導するための政策が必要
 - * 総量規制、環境税、補助金等

2.7 罰則の定式化

- $$\text{複雑性への罰則} = \underbrace{\lambda}_{\text{税率}} \times \text{複雑性の測定値}$$
 - LASSO においては、
$$\text{複雑性の測定値} = |\beta_1| + |\beta_2|..$$
 - Ridge においては、
$$\text{複雑性の測定値} = \beta_1^2 + \beta_2^2..$$

2.8 λ の設定

- λ は、推定されたモデルが母平均に近づく (予測性能が高まる) ように設定する
 - いろいろな方法が提案されている
 - 本講義では、hdm package で実装されている理論的指標を用いる

2.9 伝統的な推定方法との関係性

- $\lambda = 0$: OLS
- $\lambda = \infty$: $\beta_1 = \beta_2 = .. = 0$
 - 推定されたモデル = $\beta_0 = Y$ の平均値
- OLS と単純平均の中間的なモデルが推定される

2.10 罰則付き回帰の問題点

- LASSO を含む罰則付き回帰 (および Random Forest, Boosting, Deep Learning 等の機械学習の手法) において、推定結果とデータの間には複雑な関係性が生じる
 - 基本的に、中心極限定理が適用できず、信頼区間の計算が難しい
 - 分析結果として、職種間賃金格差は概ね “この範囲” という主張ができない

2.11 Takeaway

- OLS は、「限られたデータで、複雑なモデルを推定する」に適した推定手法ではない
 - 事例数に比べて、単純なモデルを推定するのであれば、優れた手法

- LASSO は、「限られたデータで、複雑なモデルを推定する」ための手法の一つ
 - データと推定結果の関係性が複雑であり、妥当な推論が難しい
- 次回 LASSO の性質である変数選択を適切に利用することで、上記の問題を克服できることを紹介

2.12 よくある誤解: 完璧なモデル

- 「LASSO で推定したモデルが、母平均と一致している」ことが確認できるのであれば、信頼区間の計算は不要
 - “Y を完璧に予測できるモデル” であれば、上記の条件を満たす
 - * ただし人間行動や社会的な現象について、“完璧に予測できるモデル” は非現実的 (Narayanan and Kapoor 2024)

2.13 よくある誤解: 予測と比較

- LASSO は典型的な予測で有効
 - モデル全体の適合度で良し悪しを測定
 - * 例えば、 $(Y - \text{モデルの予測値})^2$ の母平均
- 比較では、モデルの**特定の部分 (特定の変数間の関係性)** の信頼できる推定が重要
 - セミパラ推定 (Double-Debiased/Targeted learning)
 - [Applied Causal Inference Powered by ML and AI](#)
 - [Introduction to Modern Causal Inference](#)

Reference

- Imbens, Guido, and Susan Athey. 2021. “Breiman’s Two Cultures: A Perspective from Econometrics.” *Observational Studies* 7 (1): 127–33.
- Narayanan, Arvind, and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can’t, and How to Tell the Difference*. Princeton University Press.