

# 母平均

## 機械学習

川田恵介 (keisukekawata@iss.u-tokyo.ac.jp)

### Table of contents

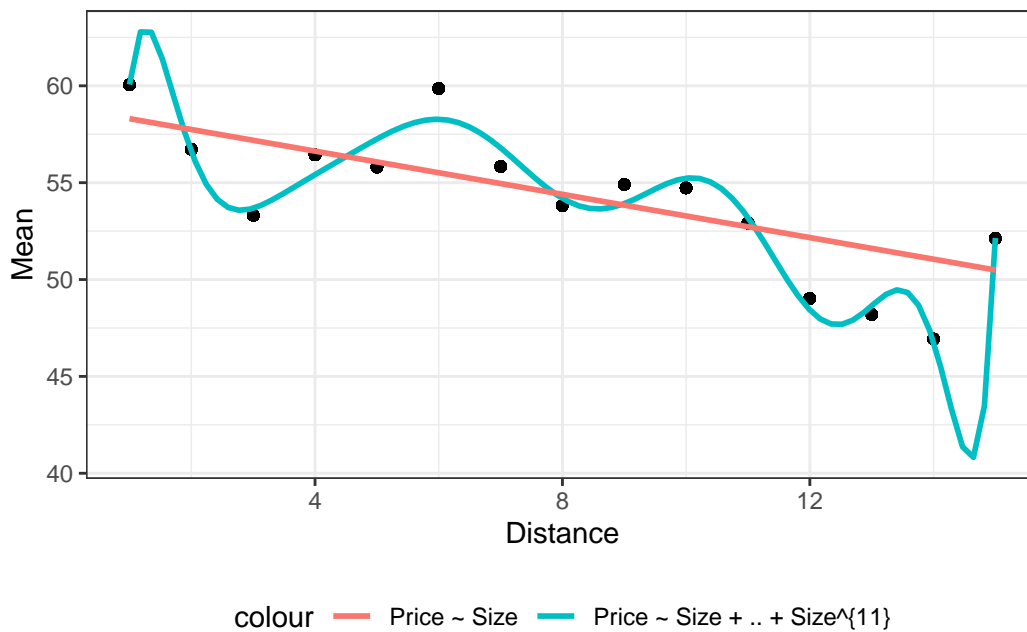
1	キーワード	2
1.1	例	2
1.2	過学習/過剰適合	2
2	母分布	3
2.1	データ分析の根本問題と解決策	3
2.2	解決策: 母集団とサンプリング	3
2.3	解決策: 評価	3
2.4	解決策: 評価	4
2.5	解決策: モデル推定	4
2.6	予測の誤差	4
2.7	最善の予測	4
2.8	数値例	5
2.9	数値例: データ (100 事例)	5
2.10	数値例: 想像上の母平均の追記	6
2.11	数値例: OLS との比較	6
2.12	数値例: OLS との比較	7
2.13	性質	7
2.14	数値例: 事例数の増加	7
2.15	数値例: 30	8
2.16	数値例: 300	8
2.17	数値例: 3000	9
2.18	数値例: 30000	9
2.19	性質	10
2.20	過学習/過剰適合	10
2.21	実戦への示唆	10

# 1 キーワード

- 過学習/過剰適合
- 推定されたモデルが、事例と過剰に適合してしまう (事例から過度に学び過ぎてしまう) 現象
  - 矛盾して聞こえるが、データ分析において最も注意すべき
  - その理由とともに必ず理解を!!!

## 1.1 例

- Price を Size (とその 11 乗まで) で OLS 推定すると？



## 1.2 過学習/過剰適合

- モデルを複雑化すると、データ上の平均に予測値が必ず近づく
  - データとの矛盾は減る
- 予測性能は悪化する!!!

## 2 母分布

- データ分析における論点整理のために、母分布を導入
  - 機械学習/統計学/計量経済学、全ての分野で用いられており、今後の講義や自学で学ぶ際に必須
  - なぜ”単純すぎる”経済モデルが実用的な場面があるのか、を理解する上でも重要
    - \* 直接観察できない概念であり、人間の想像力に依拠

### 2.1 データ分析の根本問題と解決策

- 同じ社会や市場を対象にしたとしても、研究者によって
  - 異なる予測モデルを推定する
  - モデルについて異なる評価を下す
    - \* どんなモデルもまぐれあたりする可能性がある
- 解決策: 全研究者共通の正答と各人の回答を分離

### 2.2 解決策: 母集団とサンプリング

- 評価に用いる事例 (テストデータ) は、仮想的な集団 (母集団) からランダムに選ばれたデータから、さらにランダムに選ばれた考える
  - 母集団全体を用いた評価やモデルが正答
  - 自身のデータから計算された評価やモデルが、各人の回答
- 現実においては、自身の回答しかわからないので、正答は誰も知らない
  - 高校までの勉強とは決定的に異なる

### 2.3 解決策: 評価

- 正答: あるモデルの予測値  $f(X)$  と  $Y$  との乖離を、母集団において計算
$$(Y - f(X))^2 \text{ の母集団における平均値}$$
- 回答: 自身のテストデータ上で、上記を計算

## 2.4 解決策: 評価

- 理論的性質を用いて、評価の信頼性を議論
- **大数の法則**: テストデータの事例数が十分あれば、回答 (データ上での評価) と正答 (母集団上での評価) は十分に近い値となる
  - データ全体の 2 割程度をテストに割くのが一般的
    - \* 誤差の範囲も計算できる (後述)

## 2.5 解決策: モデル推定

- 想定: モデル推定に用いるデータ (訓練データ) も、評価に用いるデータと同じ母集団からランダムに選ばれているとする
- 最善の予測 (正答) と完璧な予測を区別できる
  - 「最善の予測に近い予測を生み出す」をガイドラインとして、推定方法を評価できる

## 2.6 予測の誤差

- 予測誤差:  $Y - \underbrace{f(X)}_{\text{予測値}}$
- 完璧な予測は以下を要求: 全ての事例について

$$f(X) = Y$$

- $X$  内で個人差があれば、不可能
  - $X = \{\text{年齢、学歴、性別}\}$  から  $Y = \text{賃金}$  を完璧に予測するためには、「同じ年齢、学歴、性別であれば、賃金が全く同じ社会」が前提だが、ありえない

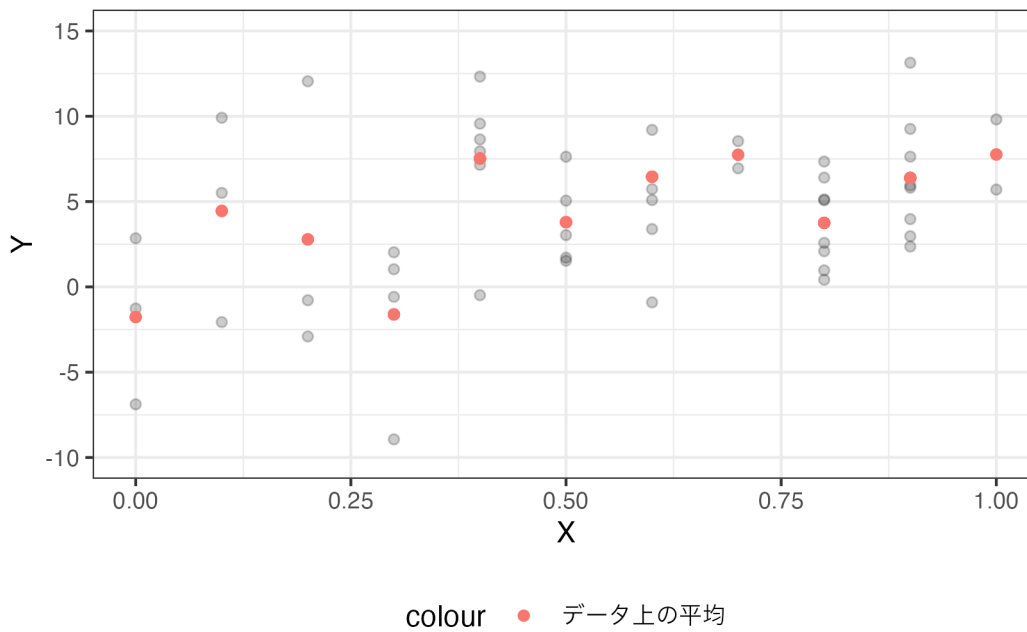
## 2.7 最善の予測

- 最善のモデル =  $(Y - \underbrace{f(X)}_{\text{予測値}})^2$  の母集団における平均値を最小化するモデル
  - 動機: まぐれあたりではなく、平均的に上手くいく予測モデルを採用したい
- 母集団を直接活動できれば、その平均値 (母平均) が最善の予測モデル

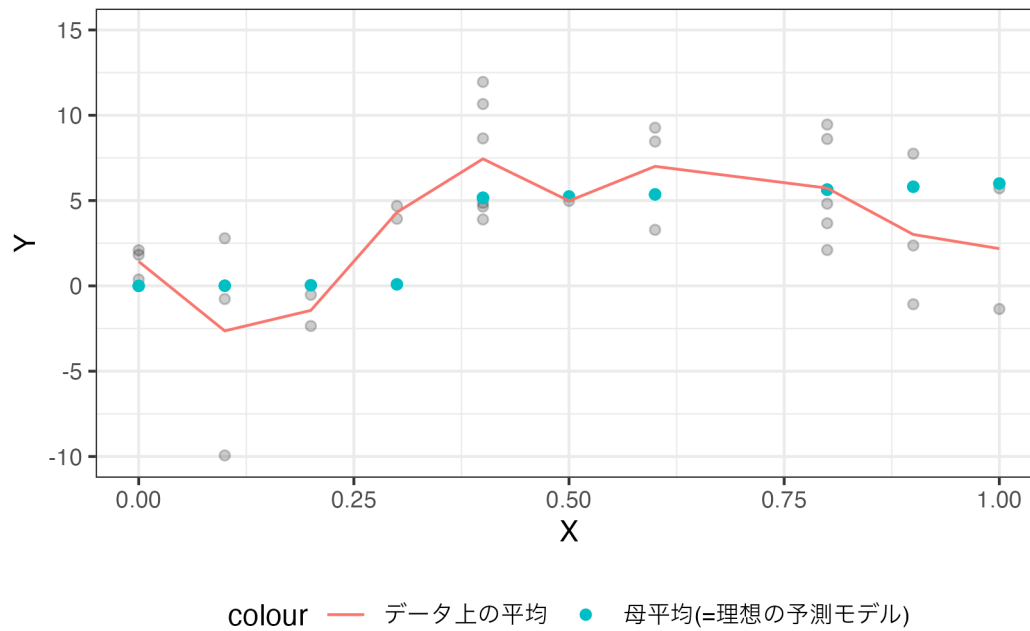
## 2.8 数値例

- $Y$  の平均値  $= X^2$  もし  $X < 0.4$  ならば
- $Y$  の平均値  $= X^2 + 2$  もし  $X \geq 0.4$  ならば

## 2.9 数値例: データ (100 事例)

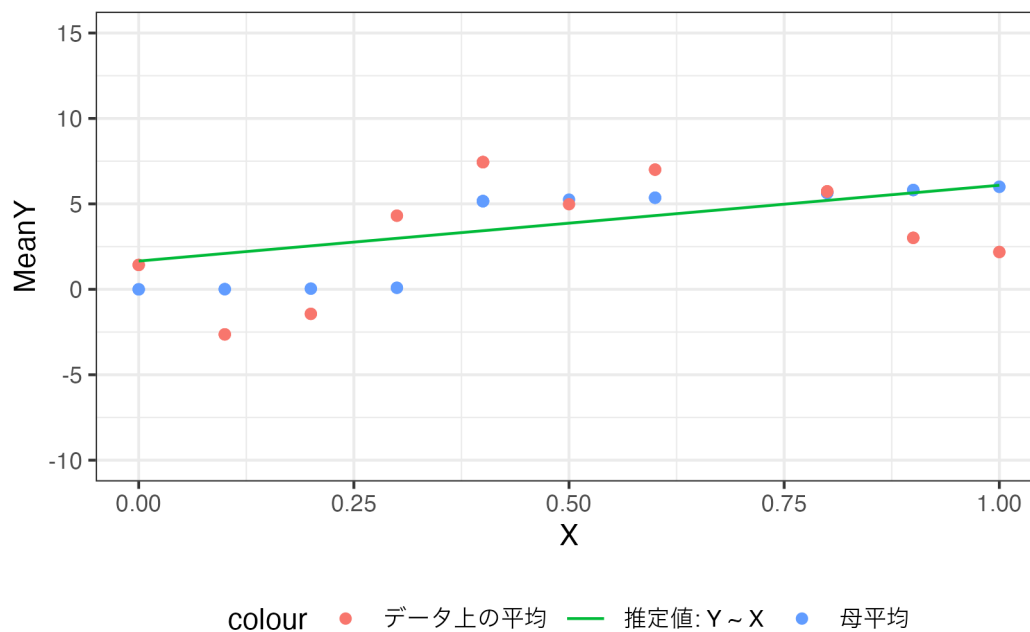


## 2.10 数値例: 想像上の母平均の追記

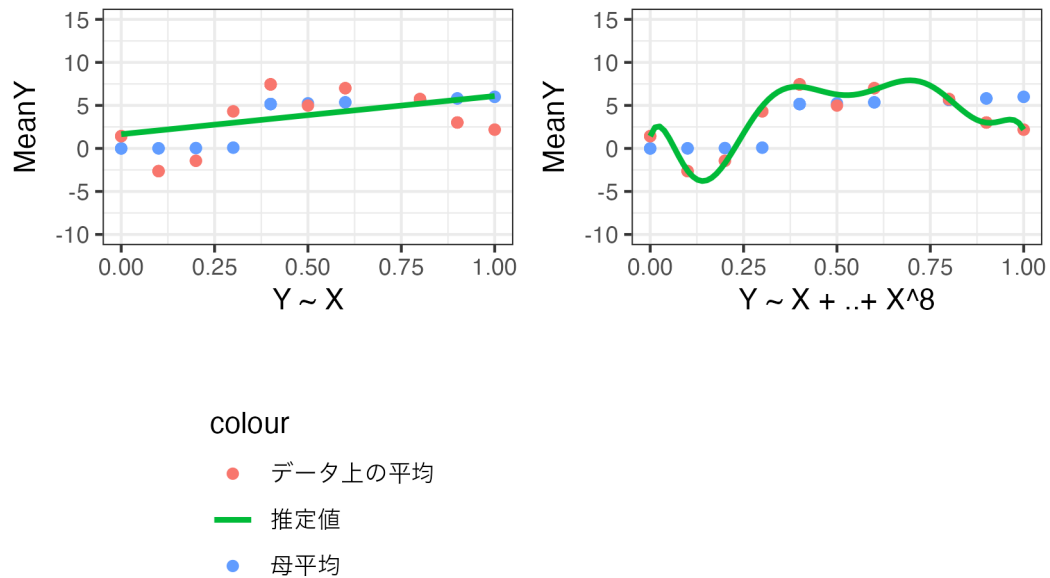


- 嚴重注意: 青点は、想像上の存在

## 2.11 数値例: OLS との比較



## 2.12 数値例: OLS との比較



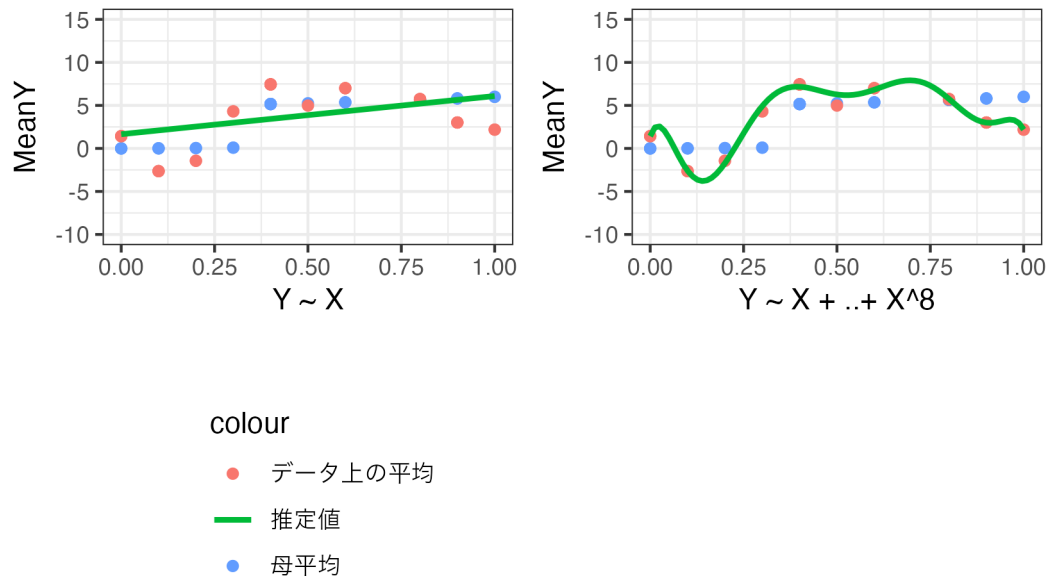
## 2.13 性質

- 限られた事例数 (N=100) では、
  - 平均値は実用的ではない: データ上の平均値と母平均は、大きく乖離している
  - 複雑なモデルも実用的ではない: データ上の平均値に近づいており、母平均から乖離している
- 単純なモデルは実用的:  $Y \sim X$  の OLS 推定結果は、母平均に近い
  - 問題点もある: “0.4 でジャンプする” という性質を捉えられない

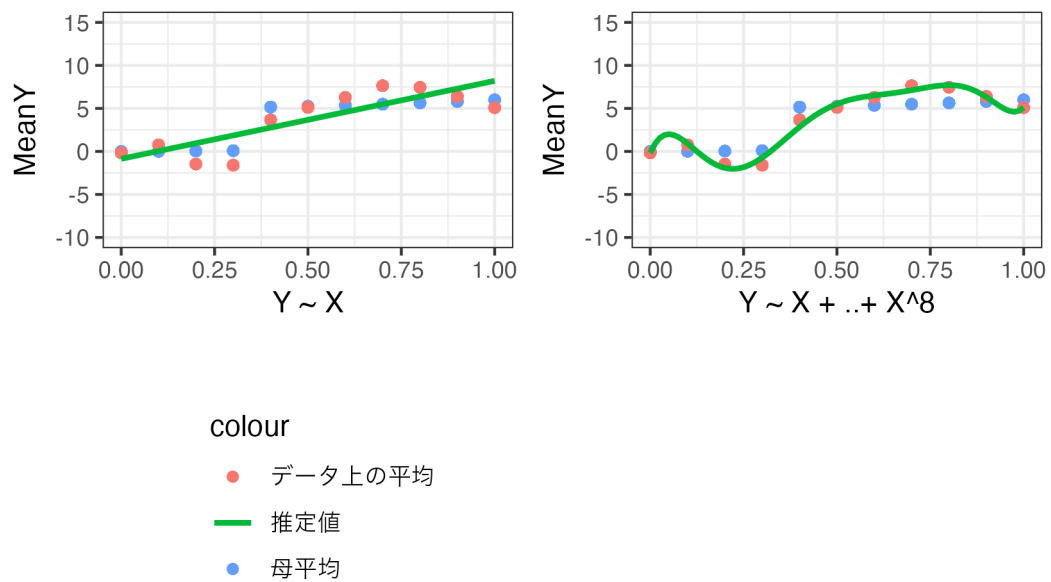
## 2.14 数値例: 事例数の増加

- 以上は事例数が少ないことにも起因
  - 300/3000/30000 事例まで増やすと?

## 2.15 数値例: 30

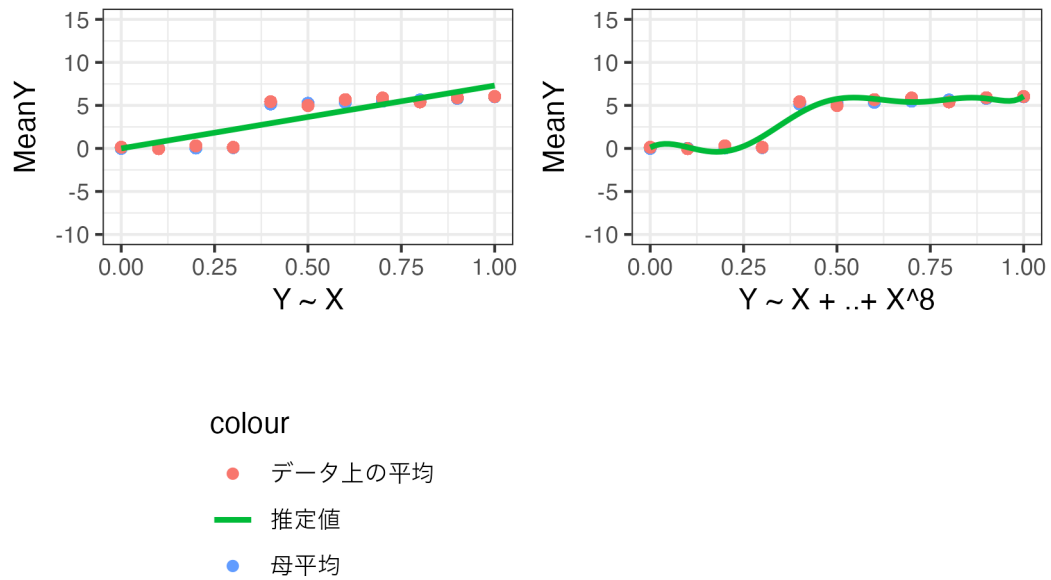


## 2.16 数値例: 300

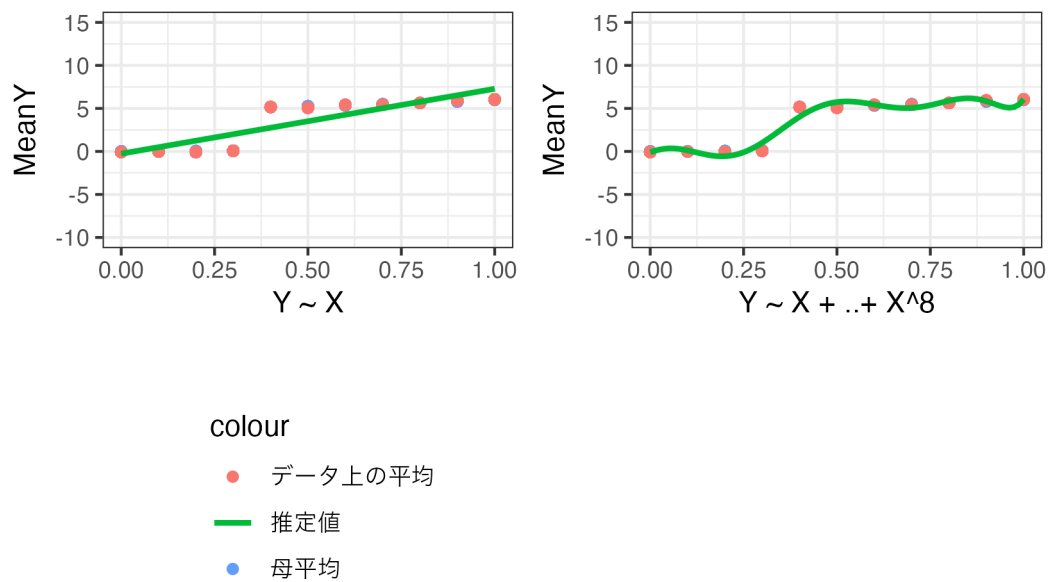




## 2.17 数値例: 3000



## 2.18 数値例: 30000



## 2.19 性質

- 事例数が増えると、複雑なモデルが実用的になる
  - 母平均とデータ上の平均が近づくので
  - “複雑な OLS” も、母平均に近づく
- 単純な OLS の予測精度は頭打ち
  - 母平均の乖離が、ほとんど変化しない
- 複雑なモデルの予測力が上がる!!!

## 2.20 過学習/過剰適合

- 事例数が少ないと、データ上の平均と母平均は大きく乖離する
  - 複雑なモデルはデータ上の平均に近いが、母平均から乖離する
    - \* 最善の予測モデル (母平均) を推定するという目標に対して、(データへの) 過剰適合/(データから) 過学習
- 事例数が増えると、過学習/過剰適合は緩和

## 2.21 実戦への示唆

- 事例数が多ければ複雑なモデルを推定できるが、少なければ単純なモデルで妥協する必要がある
- “推定するモデルの複雑さを適切に変えるべき”
  - 具体的には?
    - \* かつては人力で頑張っていたが、難しい
    - \* 次のスライドでは、データ主導のアプローチ (LASSO) を紹介