母平均

機械学習

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

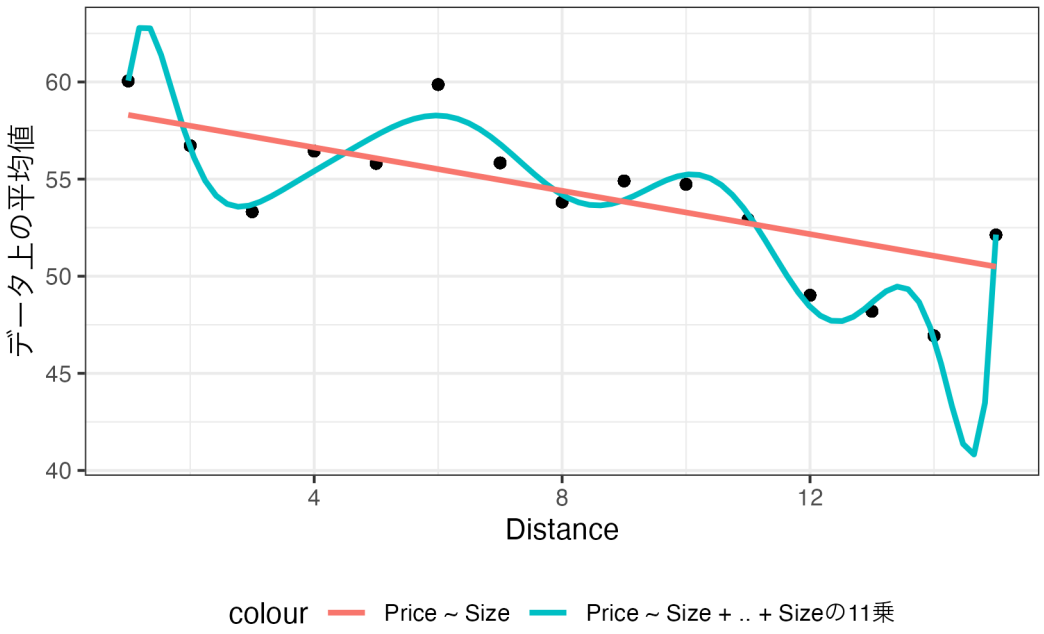
Table of contents

1 キーワード

- 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 ~ Distance (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
- 予測性能は悪化しうる!!!

2 くじ引き

- どのような回答を得られるかは、本質的に確率的に決まる (“くじ引き”)、と考える
  - データを” くじ引き”(データくじ）で入手し、分析しているため
- “データくじ” はどうしようもないものとして、推定方法を工夫する

2.1 くじ引きの評価

- 仮想的な結果を想像する必要がある: くじを引いた人の半数が
  - 100 万円分の商品/残りが 10 万円の商品 (くじ A)
  - 60 万円分の商品/残りが 40 万円の商品 (くじ B)
- 「自分がどちら側になるのかわからない」ことを前提に、どちらのくじを引くのか、くじの特徴を踏まえて考える必要がある

- くじ A の方が当たれば大きい、くじ B のばらつきが小さい、平均値（期待値）は A が $0.5 \times 100 + 0.5 \times 10 = 55$、B が $0.5 \times 60 + 0.5 \times 40$=50、なので A が大きい

2.2 データ分析のイメージ

- 不正確なイメージ:「手元にあるデータを分析する」
- 正確なイメージ:「**関心のある社会や市場から抽出された**データを、**背後にある社会や市場を理解するために**、分析する」
  - データ取集（抽出）と データ分析 (推定) に分解される
  - 多くの実践は、「データ抽出は変更できないので、推定方法の工夫」を議論する必要がある

2.3 データ収集

- 本質的には運ゲー
  - 多くの公的調査は、ランダムに回答者を選んでいる
  - 昨日のコンビニに売り上げデータなども、昨日誰がコンビニを利用したかはある程度ランダムに決まる

2.4 データ分析

- 「運よく綺麗なデータを引いたら上手くいくが、それ以外では上手くいかない手法」が存在することに注意
  - 典型例は、「事例数が少ないのに、複雑なモデルを OLS 推定する」

2.5 まとめ

- 推定方法 = イメージとしては、麻雀やトランプ (7 並べなど) の戦略
- データ = 初期の手札/配牌
- 有効な戦略 = 手札がいい時だけではなく、悪かったとしてもそこそこ機能する必要がある

3 母分布

- データ分析における「くじ引き問題」の論点整理のために、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

3.1 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

3.2 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 (くじ引きの結果)
        - 正答に近い回答が望ましい（賞品が良い）
- 自身の回答しかわからないので、正答を誰も知らない（賞品の価値がわからない）
    - くじ引きとは高校までの勉強とは決定的に異なる

4 評価

- ある予測モデルの性能をどのように評価するか?

4.1 正答と回答

- 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

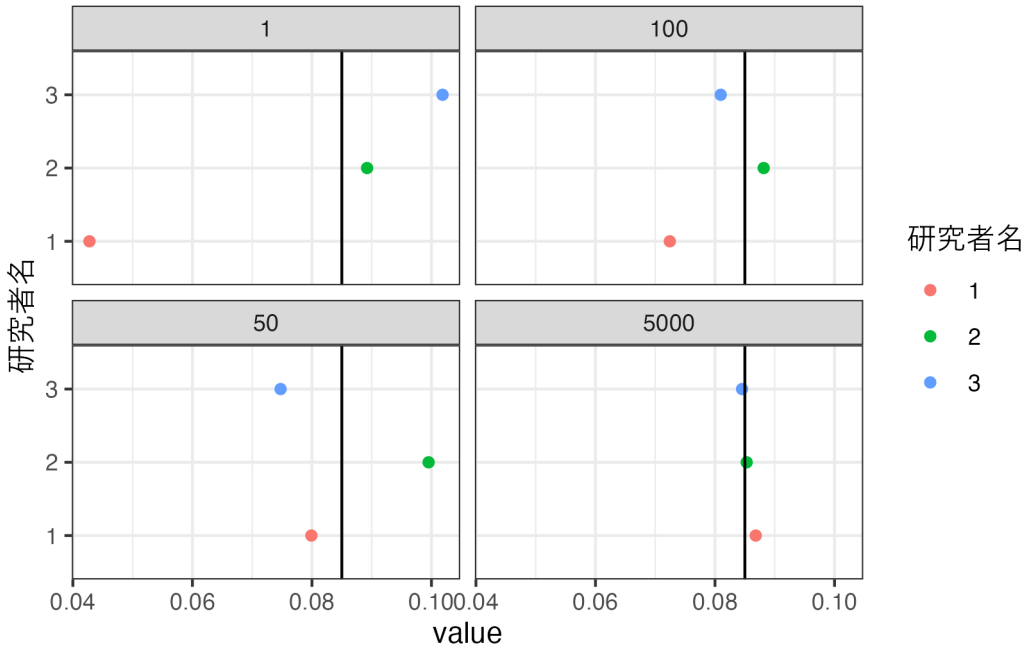
$$(Y - f(X))^2\text{の母集団における平均値}$$

- 回答: 自身のテストデータ上で、上記を計算

4.2 数値例

- 共通の予測モデルの性能を 3 名の研究者が調べる
  - モデルは共通だが、テストデータは独立して収集
  - テストデータの事例数は 1, 100, 5000, または 50000
- Y と X は、[0,1] の値をランダムに選ばれる
- モデルは、100 事例の訓練データを用いて、OLS (Y ~ X) で推定される

4.3 数値例



4.4 解決策: 評価

- 理論的性質を用いて、評価の信頼性を議論できる
- **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
  - 全員、大当たりのくじを引ける
  - 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)

5 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

5.1 予測の誤差

- 予測誤差: $Y - \underbrace{f(X)}_{\text{予測値}}$

5.2 完璧な予測

- 重要: 完璧な予測はほぼ不可能
  - 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
    * $X = \{$年齢、学歴、性別$\}$ から $Y =$ 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

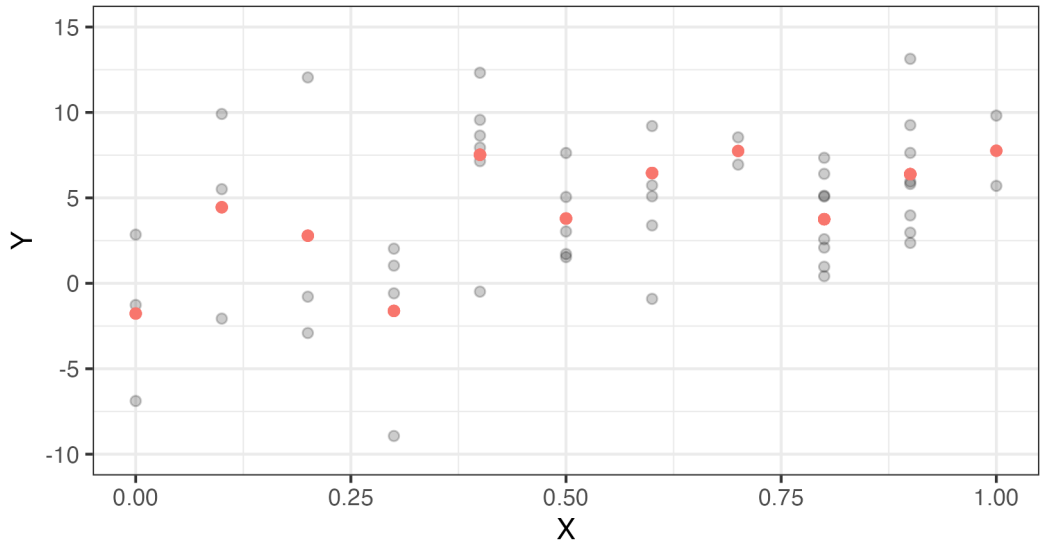
5.3 最善の予測

- 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
- 母集団における平均値 (母平均) が最善の予測モデル

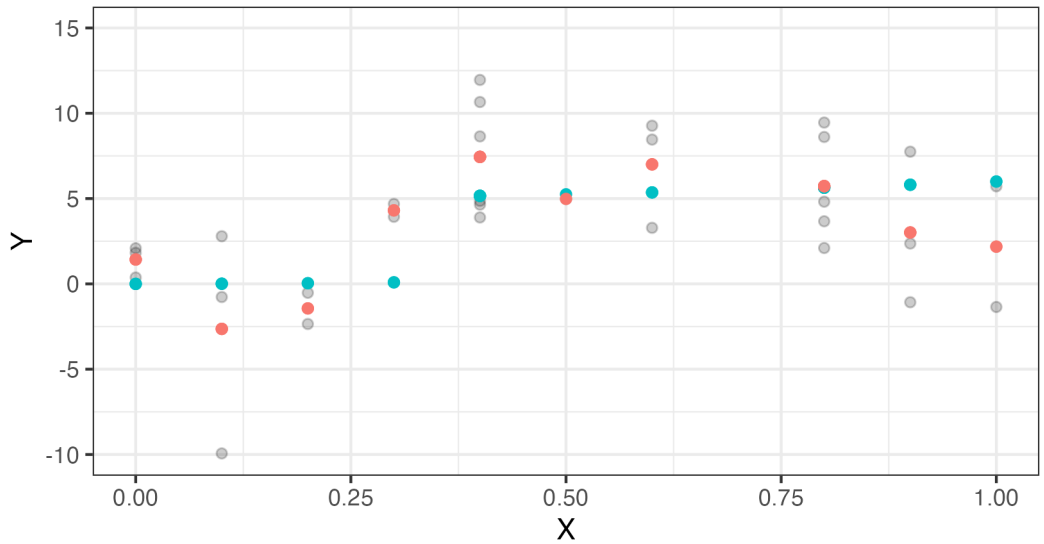
5.4 数値例

- 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 $X >= 0.4$ ならば

5.5 数値例: データ (100 事例)

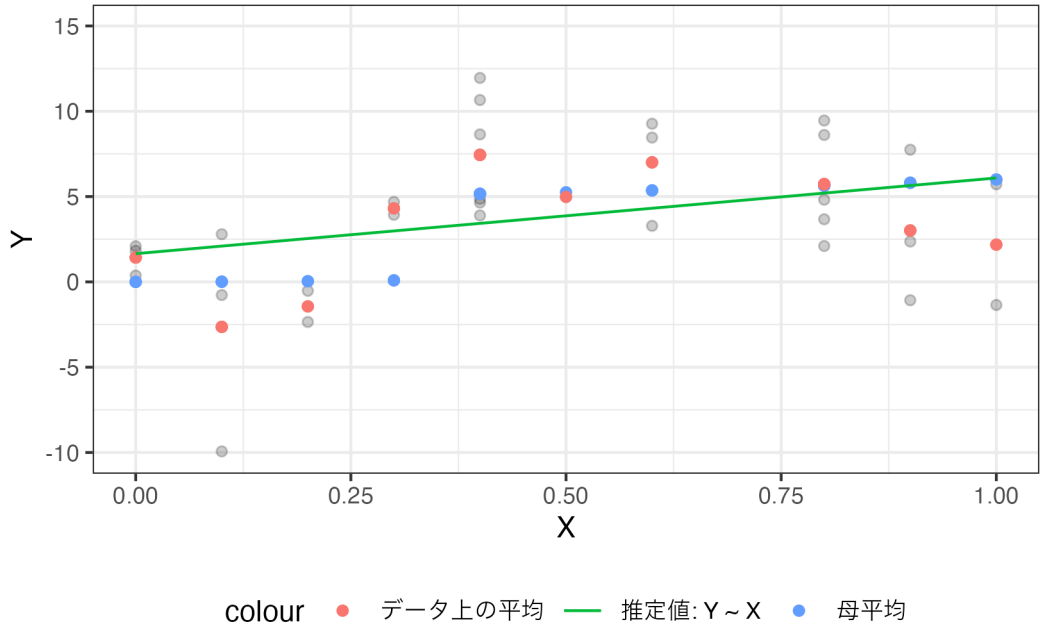


5.6 数値例: 想像上の母平均の追記

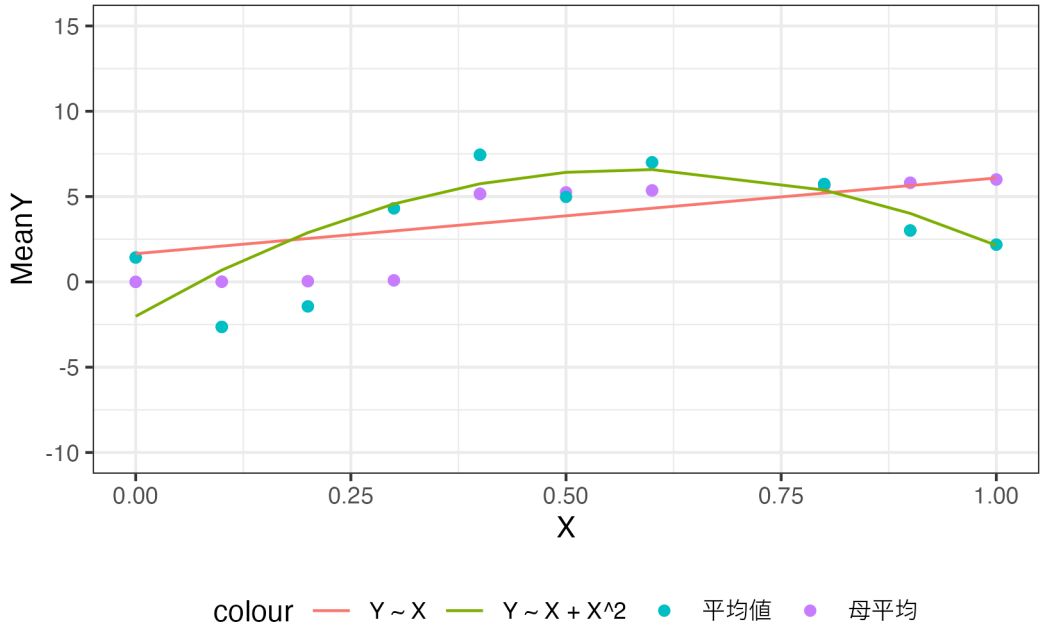


- 厳重注意: 青点は、想像上の存在

5.7　数値例: OLS との比較



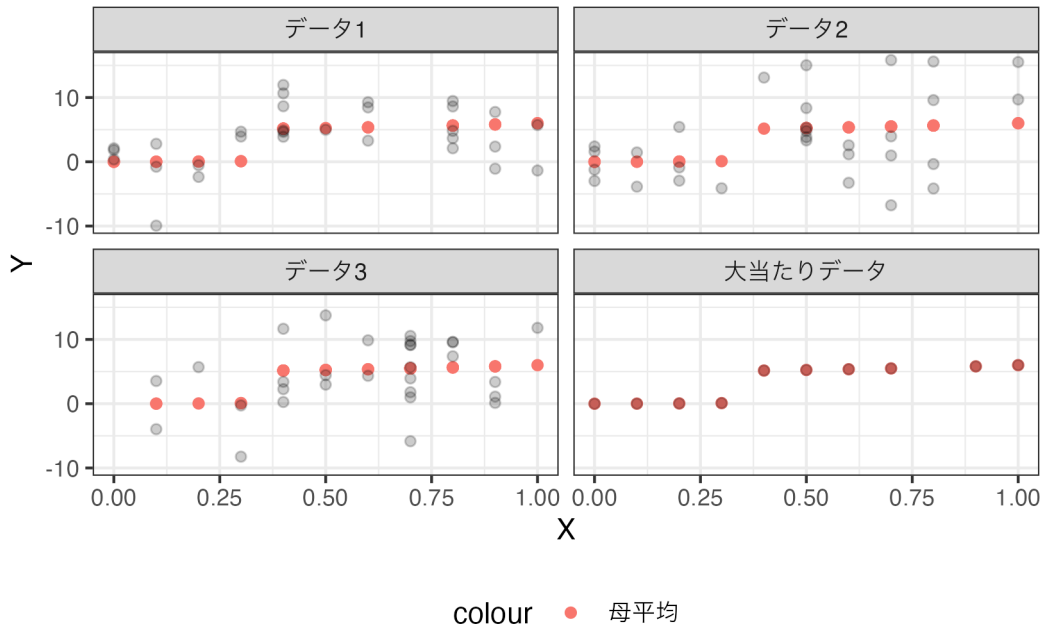
5.8　数値例: OLS との比較



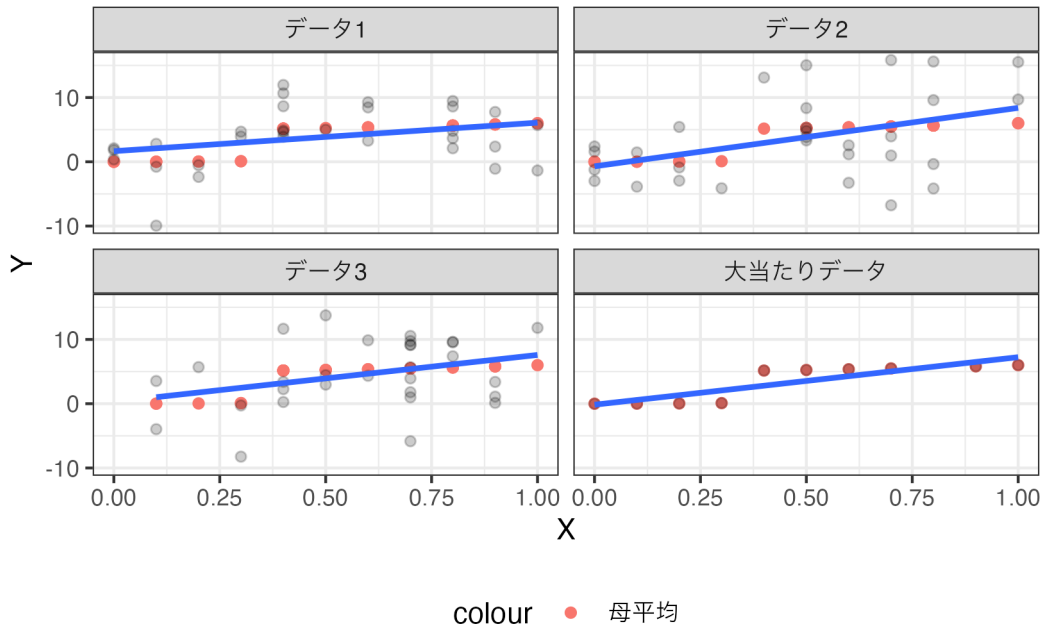
6　くじ引きとしての復習

- データくじ → 推定 → 結果、を再確認
- データくじを前提に平均 (期待値) 的にいい結果を生み出せる方法が望ましい

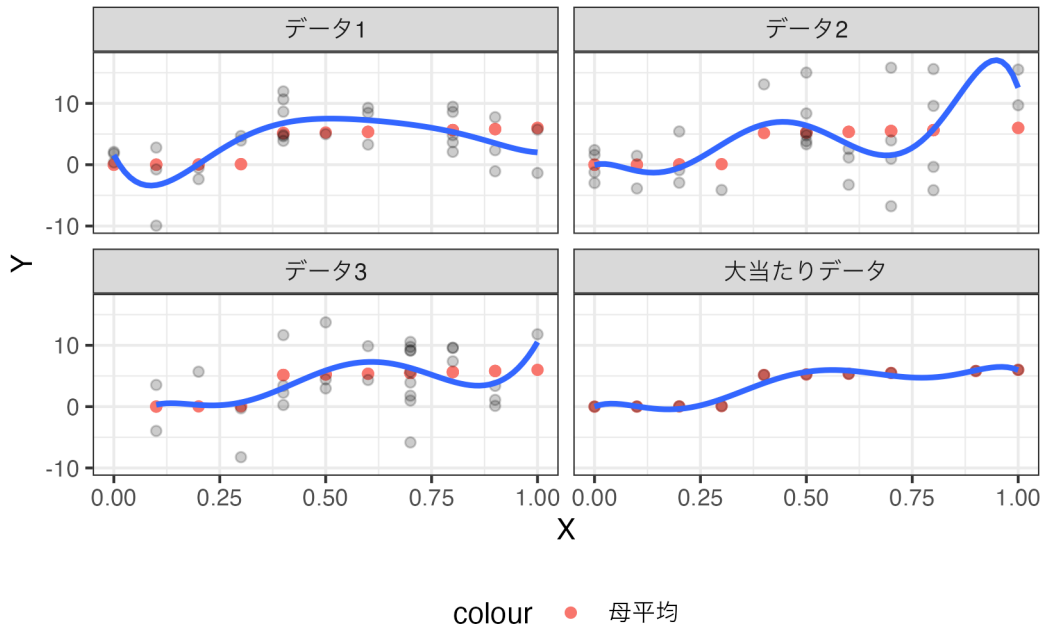
6.1　数値例: データくじ結果



6.2 数値例: シンプルなモデル



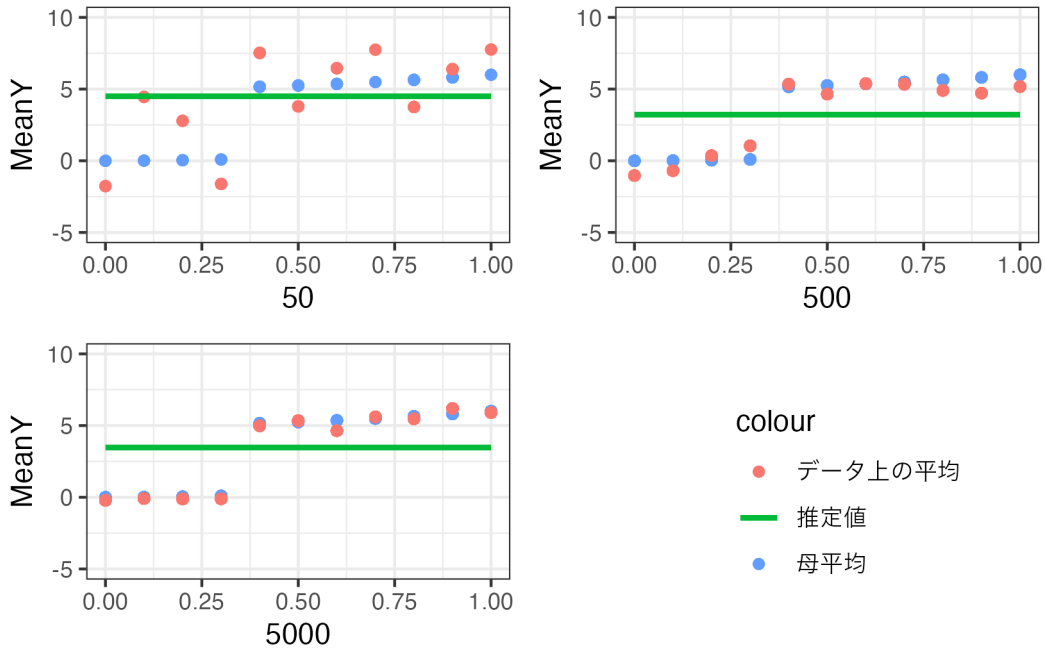
6.3 数値例: シンプルなモデル



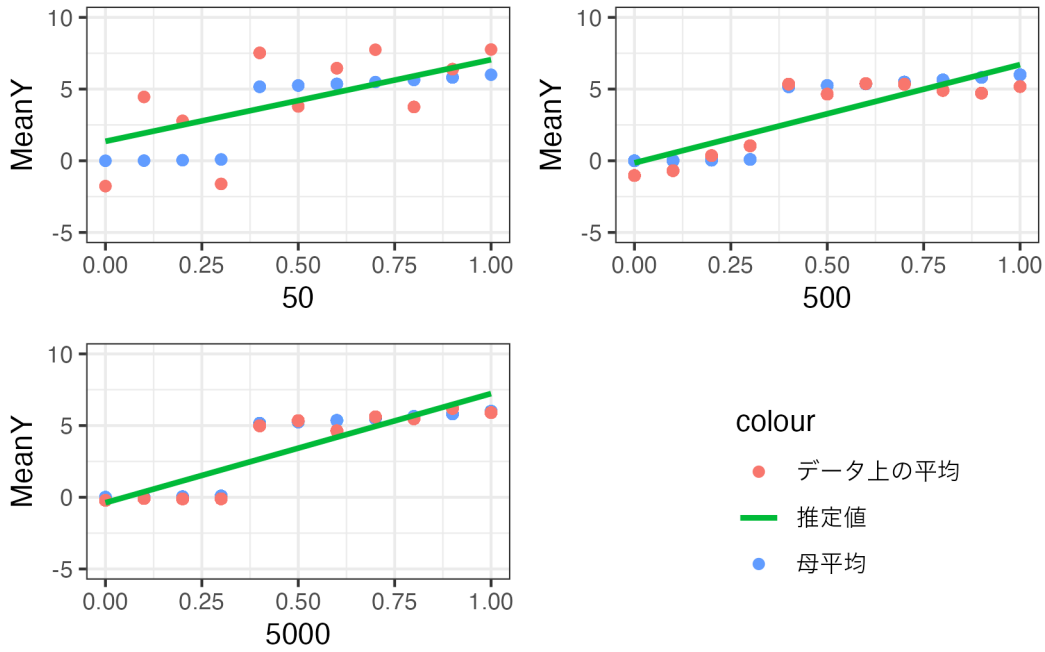
7　事例数とモデルの複雑さ

- 望ましい推定方法は、対象とする社会/市場とデータの収集方法に強く依存する
- ここでは事例数との関係性を確認

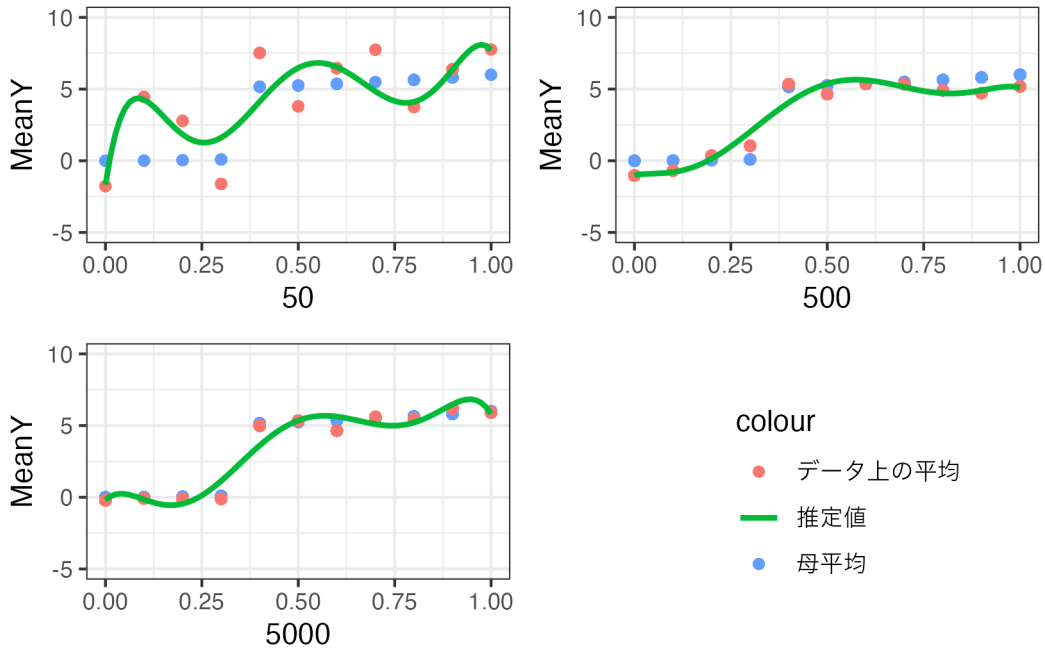
7.1　数値例: OLS との比較



7.2　数値例: OLS との比較



7.3　数値例: OLS との比較



7.4 性質: 少数事例

- 限られた事例数 (N=50) 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

7.5 性質: 大規模事例

- 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

7.6 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

7.7 まとめ

事例数/モデルの複雑さ	複雑	単純
少ない	(過剰適合)	(現実的な妥協)
多い	(理想的)	(不十分なデータ活用)

- 笑顔 = より良い予測モデル

7.8 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