統計的性質
機械学習

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-09-30

1 サンプルリングと母集団

1.1 目標

- ここまでの議論を整理
  - シンプルなモデルや LASSO を活用する動機は、
    - 複雑なモデルを OLS で推定した予測は、“下振れ/上振れ”しやすい
      - 何と比べて”下振れ/上振れ”するのか?
  - 「母集団とサンプリング」を用いた枠組みを用いる

1.2 目標

- 母集団という概念をしっかり理解
  - 統計学、計量経済学、機械学習等、“全ての”データ分析法を学ぶ際の必須概念
    - 同時に初学者最大の壁
- 個人的には、特殊な”世界観”を例え話として用いた論点整理法、ぐらいの感覚が実践的だと思います
  - 類似例: 国家を”人間”に例えて、国際政治を論じる

2 母分布

2.1 サンプリング

- データ収集 = サンプリング
  - 業務記録、聞き取り/電話/ネット調査等
- 本講義では、**母集団**からランダムに選ばれた事例が、データに収録されていると考える

- 同じ研究対象、課題、分析手法であったとしても、サンプリングされるデータが異なるため、結果が変化してしまう。

2.2 母集団

- 本講義の範囲内では、無限大の事例数を持つ巨大データをイメージして OK
  - 推定 = 手元のデータから母集団の特徴を推測するプロセス

2.3 例

- 武蔵大学生の実態調査を行うために、ランダムに選ばれた 100 名にインタビューし、データ化
  - 母集団 = 武蔵大学生、データ = 選ばれた 100 名
- 9 月の生協利用者の購入履歴をデータ化し、売れ筋商品を把握したい
  - 母集団 = “潜在的な”生協利用者、 データ = “実際”の利用者
  - 「潜在的な利用顧客(母数団)の一部が偶然利用した」とも解釈できる

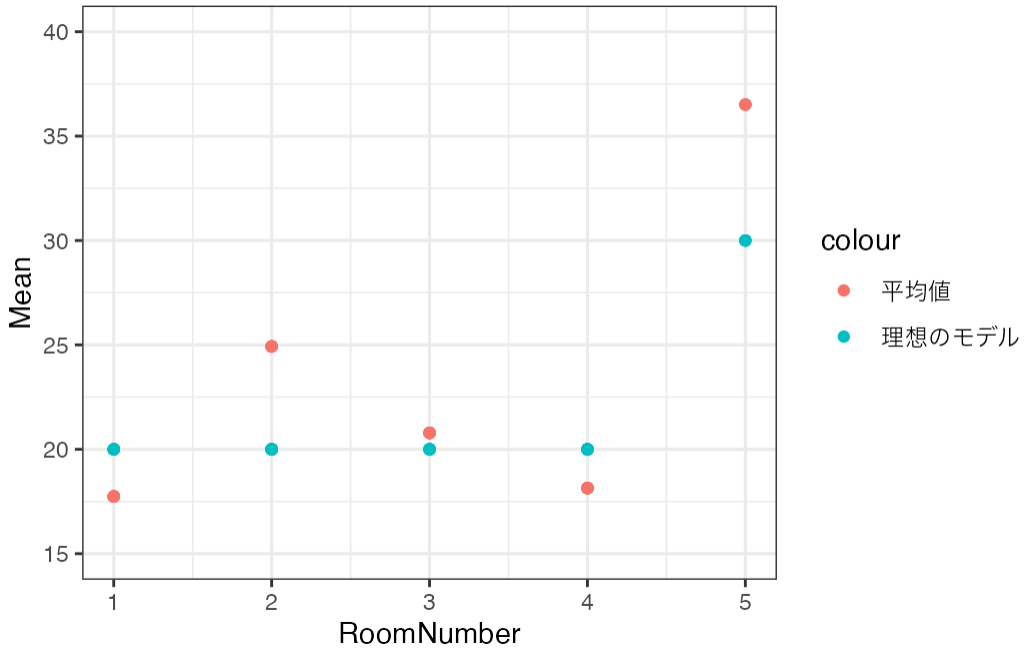
2.4 母平均

- データである以上、母集団に対してもここまでの議論が適用できる
- 母平均 = 母集団において”仮想的に”計算された平均値
- 実際の計算は、不可能であることに**厳重注意**

2.5 母平均の推定

- 母平均の推定: もし事例数が十分にあれば、データ上の平均値は優れた推定値
  - データ上の平均≈ 母平均、
- 事例数が不十分な場合は、
  - データ上の平均値 ≈ 母集団上の平均値

2.6 例



3 予測モデルの推定

3.1 予測対象のサンプリング

- 予測対象も、データと同じ母集団からランダムに選ばれる状況を想定
  - **実践への含意:** 予測対象と時間、地理、文化的に近い集団から得られたデータを活用できる
  - **違反している例:** 今の武蔵大生を予測するために、1970 年のアメリカの高校生のデータから予測モデルを推定する

3.2 完璧な予測

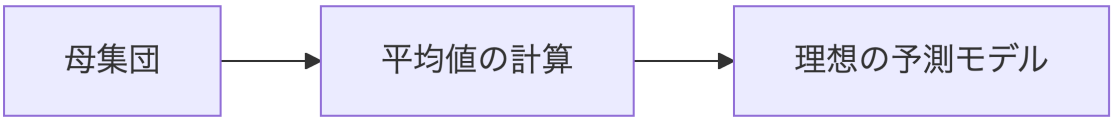
- 完璧な予測: 全ての事例について

$$Y = \textbf{予測値}$$

  - X 内で個人差があれば、不可能
  - $X = \{\textbf{年齢}, \textbf{学歴}, \textbf{性別}\}$ から $Y = \textbf{賃金}$ 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”非現実的”社会」を前提とする必要がある

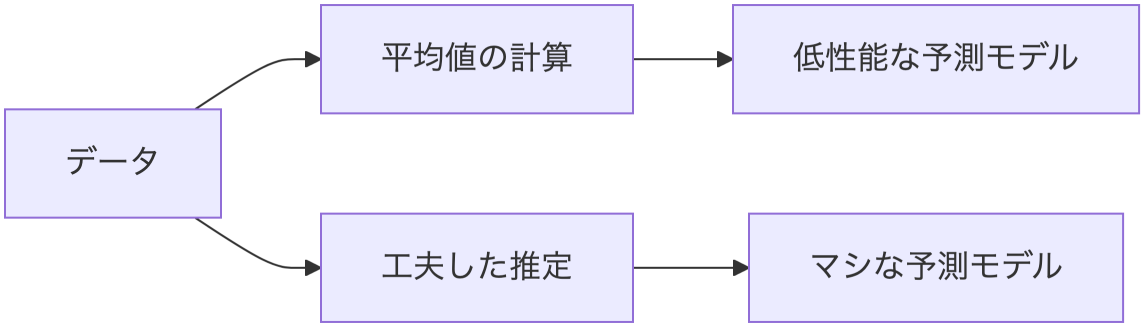
3.3 最善の予測

- もし母集団を予測モデル推定に活用できる場合、どのような”推定”方法がベストか?
  - 平均二乗誤差を最小化したいのであれば、**母平均** が最善

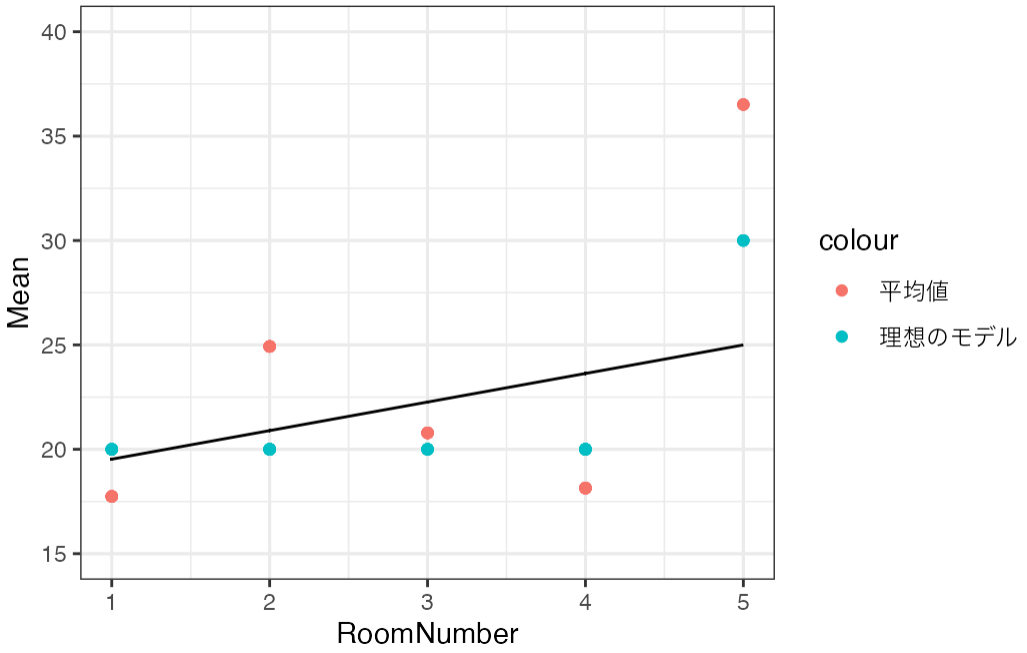


3.4 現実の予測

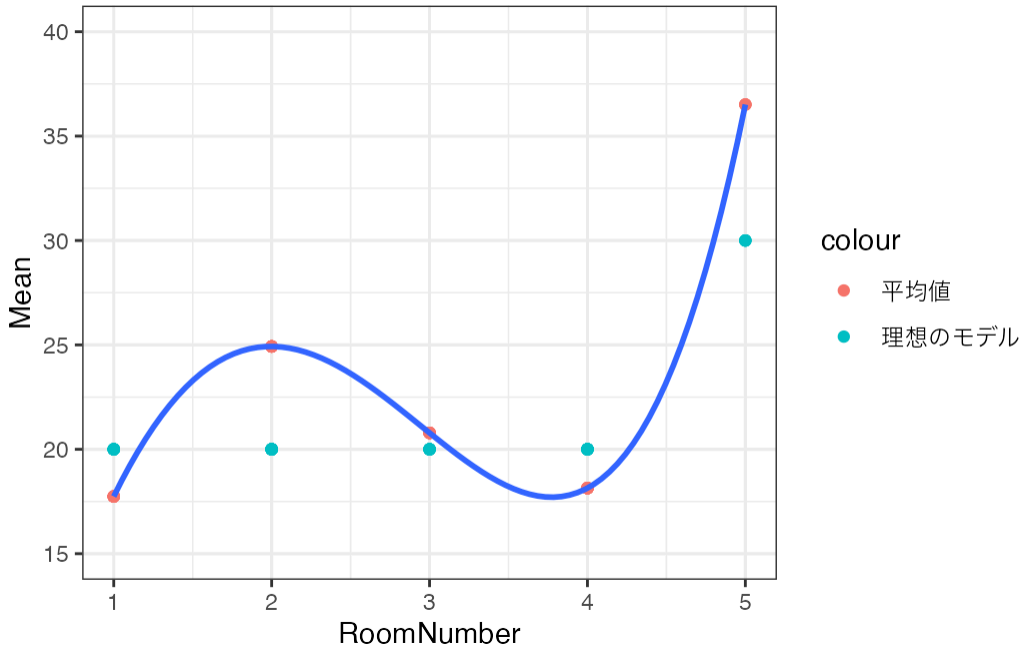
- 限られたデータから、母平均に近い予測モデルを生み出せる推定方法を採用すべき
  - 多くの応用で、事例数の問題から、平均値は母平均との乖離が大きい



3.5 例: ベンチマーク



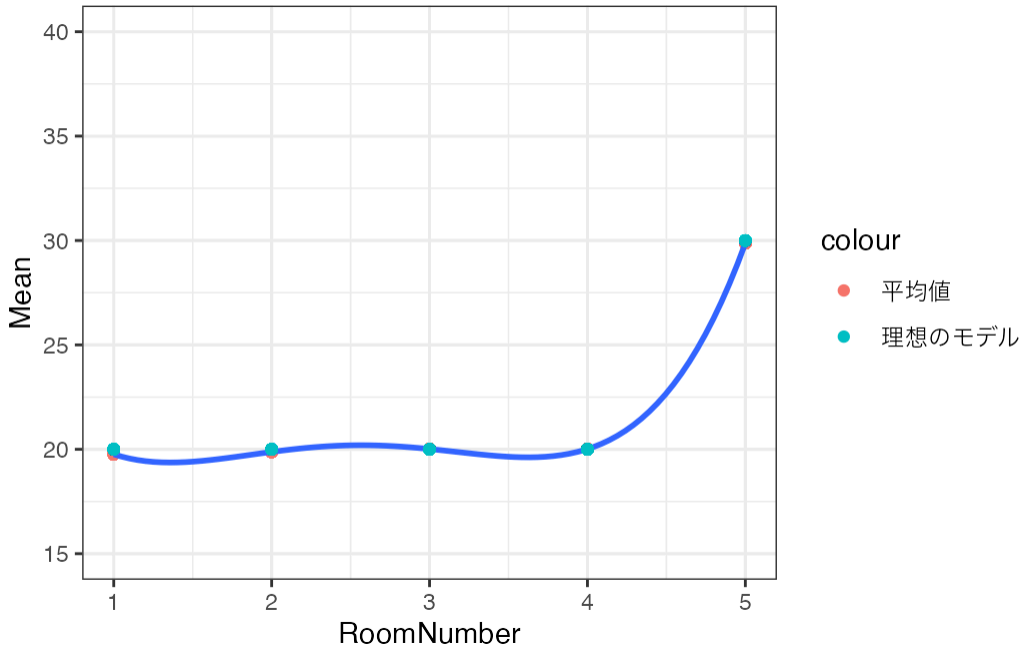
3.6 例: 複雑なモデル



3.7 サンプルサイズの問題

- 先のシュミレーションは、小規模な事例 ($N = 20$) からなるデータ
  - データ上の平均値に近づく複雑なモデルは、予測性能が低い
- 大規模事例を用いれば、複雑なモデルの方が、理想の予測モデル (母平均) に近づく
  - データ上の平均値が母平均に近づくため

3.8 例: 複雑なモデル with 50000 事例



3.9 Takeaway

- 基本アイディアは、母集団 ≈ データを期待して、データ上の平均値の性質を予測に活用
- 事例数が少ない場合は、母集団の詳細な特徴 ≈ データの詳細な特徴であり、データの大雑把な特徴のみを予測に活用することが現実的
  - 一つの方法は、単純なモデルを OLS で推定

3.10 Takeaway

- **無限大**の事例数を持つデータが利用できるのであれば、平均値が理想の予測モデル
- **非常に大きい**事例数であれば、データ上の平均値が優れた予測モデル
- **十分に大きい**事例数であれば、複雑な予測モデルを OLS で推定する方法は実用的
- **限られた**事例数であれば、過剰適合を避けるために、単純な予測モデルを OLS で推定する方法が実用的

3.11 Takeaway

- OLS は、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、人間が単純なモデルを設定する必要がある
    - 実践は難しい
- LASSO は、過剰適合を減らす仕組みを導入

- 例えば、事例数が少なければ、複雑性への”課税額”が自動的に増加し、単純なモデルが推定されやすくなる