OLS as Best Linear Projection esitmator

労働経済学 1

川田恵介

1 OLS

1.1 動機となる研究

- “教育/経験年数と時給との関係性を把握するために、ミンサー型賃金モデル (川口大司, 2011)

$$\log(wage) \sim \beta_0 + \beta_1 \times EducationYear$$

$$+\beta_2 \times Experience + \beta_3 Experience^2$$

  を OLS で推定した"

- 批判: 推定対象/研究目標が、不透明であり、解釈が難しい
  - 実際には何を推定し、どのように解釈できるのか?

1.2 動機

- OLS = 現代的な比較研究においても、代表的**推定方法**
  - 研究者が**事前**に設定した**線型モデル**を、データから**推定する計算方法**
  - **推定対象** について、**複数の解釈** がある (Angrist & Pischke, 2009; Chattopadhyay & Zubizarreta, 2023)
  - 多くの発展的手法が、OLS の特定の問題点を改善する方法である、と解釈できる

1.3 OLS の入門書的解釈

- 賃金を年齢で OLS で計算した**推定値**は、

```
lm(wage ~ age, CPS1985) # Price ~ beta_0 + beta_1*Size
```

- 以上の**推定対象**は
  - Price の(条件付き)母平均 $\mu(age) = E[wage \mid age]$ (Stock & Watson, 2020; Wooldridge, 2022)
    - $\mu(wage) = \beta_0 + \beta_1 \times age$ を仮定する必要があり、非現実的

1.4 OLS の別解釈

- 二つの別解釈: OLS の推定対象は
  1. 母平均 $\mu(X)$ の**母集団上**での線形近似モデル
  2. $\mu(D=1,X)-\mu(D=0,X)$ の母集団上での近似的な Balancing comparison
- モデルが”正しくない”場合でも、明確な推定対象を定義でき、解釈が容易
- 本ノートでは、線形近似モデルの推定値であることを紹介

1.5 構成

- OLS について、
  1. **データ上**で行なっている計算
  2. **母集団上**での推定対象
  3. 社会の記述

1.6 まとめ

- OLS の推定対象は、複数存在する
  - 母平均の最善の線型モデル (Best Linear Projection)
- 母平均そのものの優れた推定値であるとは**限らない**

2 データ上の計算

2.1 線形近似モデル

- モデル = データや社会、母集団の特徴を要約する”式”
- 例 “単回帰”:

$$g(Age)=\beta_0+\beta_1\times Age$$

- 例 “重回帰”:

$$g(Age,Educ)=\beta_0+\beta_1\times Age+\beta_2\times Educ$$

2.2 線形近似モデル

- β について足し算であれば、X を変形しても線型モデル
- 例 X について非線形モデル:

$$g(Age)=\beta_0+\beta_1\times Age+\beta_2\times Age^2$$

2.3 OLS

- データに極力適合するように、**推定モデル**を計算する方法

  - 以下を最小化するように β を推定する

$$(Y-g(X))^2 \text{のデータ上の平均値}$$

- Y を近似するモデルと解釈できる
  - **多重共線性**がなければ計算できる

2.4 データ上の平均値

- (条件つき)平均値 $(\hat{\mu}(X))$：$X_i=x$ である事例内でのYの平均値

$$\hat{\mu}(X)=\frac{1}{(X_i=x)\text{である事例数}}(Y_1+Y_2+..)$$

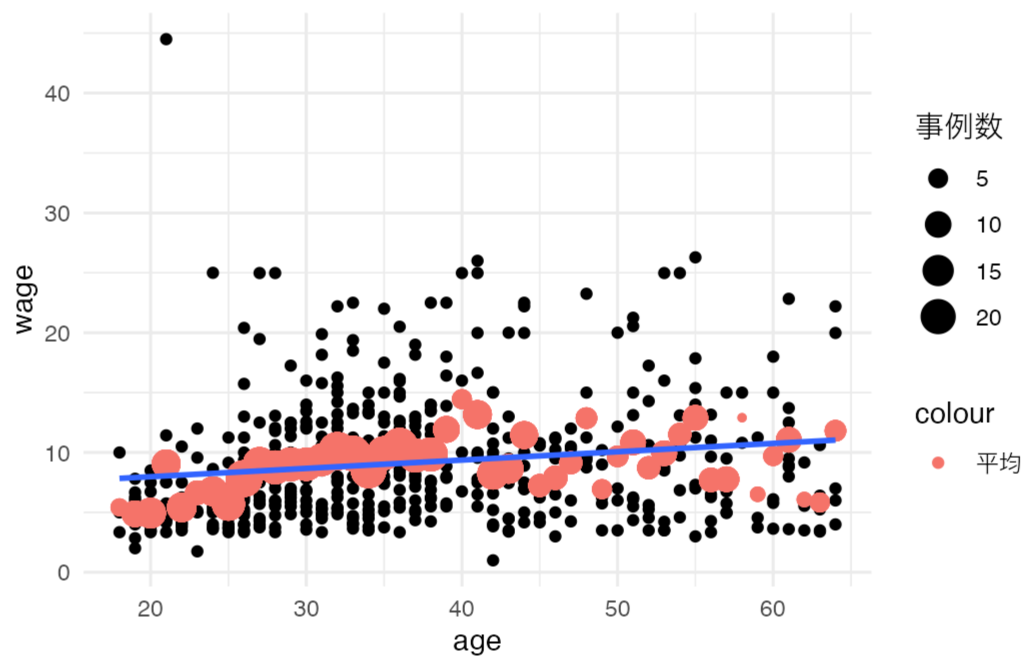
2.5 別解釈

- 以下を最小化しても、同じモデル $\hat{g}(X)$ が計算される
- 全てのXの組み合わせ $[x_1,..]$について、

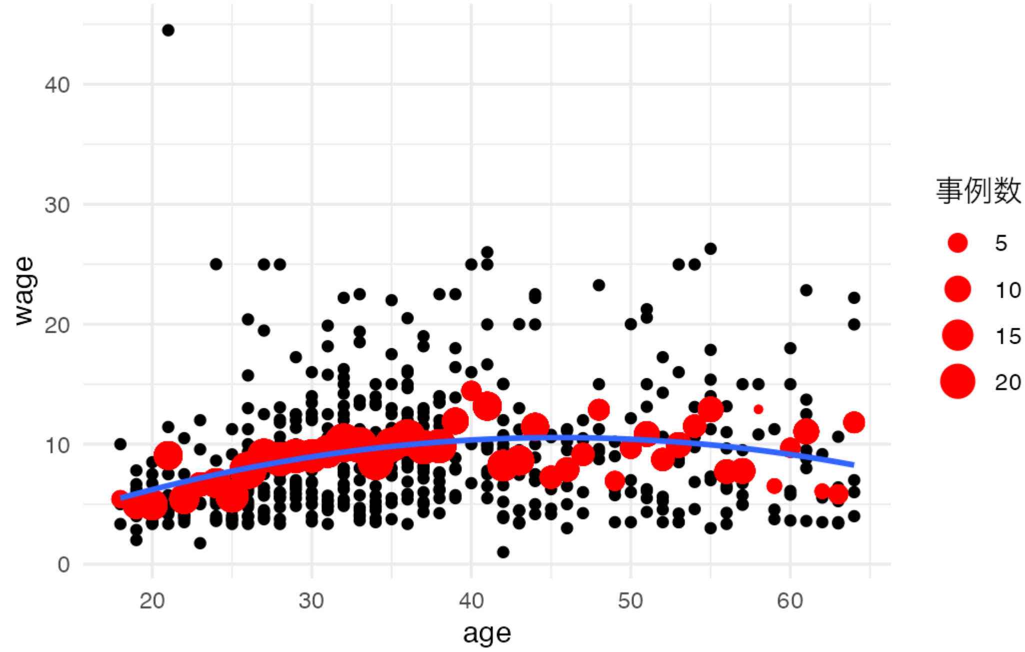
$$\left[\underbrace{(\hat{\mu}(x)-g(x))^2}_{\text{平均からの乖離}} \times [X=x\text{となる事例割合}]\right]\text{の平均値}$$

- ” Y の**平均値** “を近似するモデルと解釈できる

2.6 例 $g(Age)=\beta_0+\beta_1 Age$



2.7 例 $g(Age) = \beta_0 + .. + \beta_2 Age^2$



2.8 Y のモデル VS 平均値のモデル

- 実際に計算される推定値は同じ
  - あくまで”解釈”の問題
- 研究対象次第で、有益な解釈は変化する
  - 経済学研究においては、平均値のモデルと解釈した方が有益な場面が多い
    - 個人差が大きく、Yのモデルに見えない
    - 平均値は、予測/比較研究における中核的関心

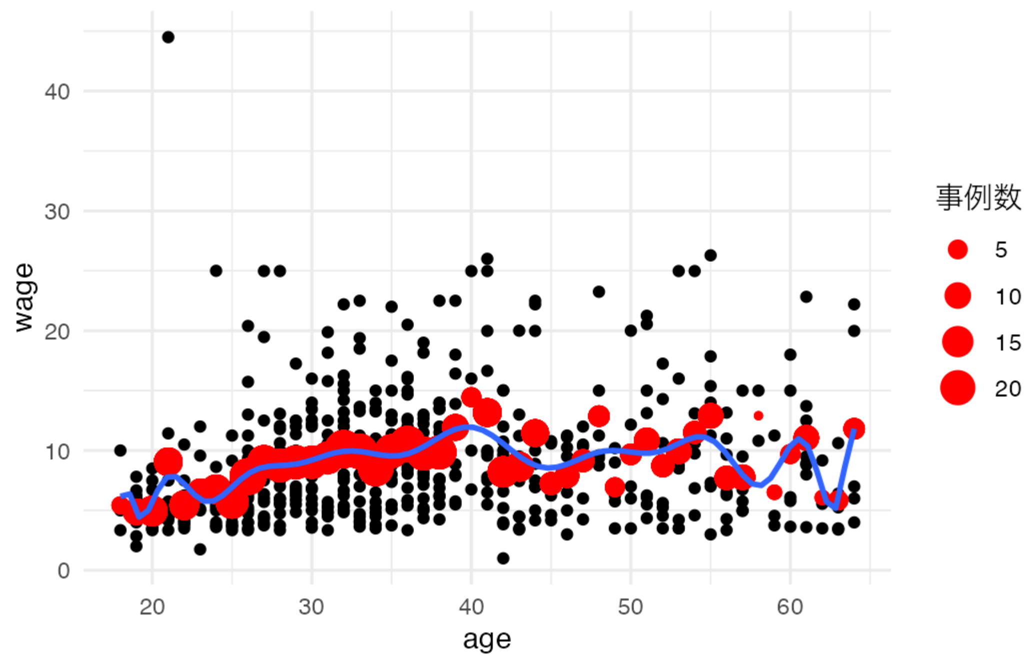
2.9 OLS の特性

- 「どのようなモデルを推定するのか」によって、推定されうるパターンがある程度決まってしまう
  - $g(Age) = \beta_0 + \beta_1 Age$ をデータに当てはめると、年齢と平均賃金の間に”一直線”の関係性しか推定されない
    - $g(Age) = \beta_0 + \beta_1 Age + \beta_2 Age^2$ について、 $\beta_2 = 0$ と事前に研究者が決めてしまっている
    - “研究者主導"の方法

2.10 モデルの複雑化

- よりβが多い、複雑なモデルを OLS で推定することもできる
  - 例: $\beta_0 + \beta_1 \times X + .. + \beta_{10} \times X^{10}$
- より多くのβをデータで決める
  - β の数を増やすと、平均 $\hat{\mu}(X)$ により近づく

2.11 例 $g(Age) = \beta_0 + .. + \beta_{20} Age^{20}$



2.12 まとめ

- OLS $= Y$ の要約値である平均値 $\hat{\mu}(X)$ を、さらに要約したモデル $\hat{g}(X)$ (“線”)を算出
- 要約しているので、一般に、$\hat{g}(X) \neq \hat{\mu}(X)$
  - モデルを複雑化すると、$\hat{g}(X) \simeq \hat{\mu}(X)$
  - 平均値に近づけることの弊害はあるのか?
    - 母集団を導入し、推定精度を定義する必要がある

3 母集団上での推定対象

3.1 議論の枠組み: 頻度論

- 研究課題/推定対象/推定方法は同じだが、**データは独立して収集する**、仮想的研究者群をイメージ
  - 自分はその中の一人

- 同じ手法を用いても、データが偶然異なるので、推定値は異なる
  - 自分の結果は、「偶然生じた」信用できないものと考える方が合理的
- 詳細: Section 6 , StatLec などを参照

3.2 推定対象と推定値

- データ分析法を、建設的に議論するために
  - 全ての研究者が原理的に合意できる正答 (推定対象) と 自身のデータから得られる回答 (推定値)を個別に定義する
    - 推定対象を定義するために、母集団を導入する

3.3 母集団

- 手元にあるデータに含まれる事例を、ランダムに選んできた仮想的な集団
  - 本講義の範囲内では、手元にあるデータと同じ変数が観察できる”超巨大データ”をイメージしても OK
- 注: 時系列などの独立ではないデータは、本講義の対象外

3.4 推定対象

- 推定対象 = 母集団を用いて**仮想的に**計算される値
- 例: 母集団上で計算される OLS の**仮想的**な結果 (Population OLS)
  - 同じ方法でデータ収集するのであれば、母集団は全ての研究者で共通

3.5 まとめ

- 分析計画が確定したとしても、実際に取集される事例が異なるため、異なる推定値が算出される
  - データ”くじ”に伴う不確実性
    - Sampling Uncertainly
  - 信頼区間やp値、機械学習におけるさまざまな工夫などは、この不確実性への対処がメイン
    - よい統計的手法 $\simeq$ データくじの影響を受けにくい/影響を適切に評価できる

3.6 注意点

- データ分析は入門段階から、「**厳密に定義されるが、根本的に測定不可能な推定対象を、頑張って推定したい**」という複雑な問題を論じる必要がある
  - 初学者が混乱するのは当たり前
  - 随時質問しながら、ゆっくり消化してください

4 Population OLS

4.1 Population OLS

- OLS の推定対象 = 母集団上で仮想的に行われる OLS (Population OLS)の結果

$$g^{Pop}(Y) = \beta_0^{Pop} + .. + \beta_L^{Pop} X_L$$

- 以下、Population OLS は定義できる、と仮定する

4.2 Population OLS の推定

- OLS の推定値 $\hat{g}(X)$ = Population OLS $g^{Pop}(X)$ の推定値
  - β の数に比べて、事例数が大きければ、$g^{Pop}(X)$ とよく似た推定結果 $\hat{g}(X)$ を得る可能性が高い (Section 7)
    - Threorem 1.2.1 (Chapter 1, CausalML)

4.3 複雑なモデルの推定対象

- モデルの複雑化 → 推定対象が変化する

```
lm(Price ~ poly(Size, 2), Data) # Price ~ beta_0 + beta_1*Size  + beta_2*Size^2
```

- 推定対象は、$\beta_0 + \beta_1 \times Size$ ではなく、$\beta_0 + \beta_1 \times Size + \beta_2 \times Size^2$ の Population OLS

4.4 十分に複雑なモデル: 推定対象

- モデルを複雑にすれば、**Population OLS は**、母平均に近づく
  - Section 2.10 と同じ理屈
- OLS の推定対象 $\underbrace{=}_{\text{常に}}$ Population OLS $g^{Pop}(X)$
  - $\underbrace{\simeq}_{\text{十分に複雑であれば}}$ 母平均 $\mu(X)$

4.5 モデルの複雑化: 推定

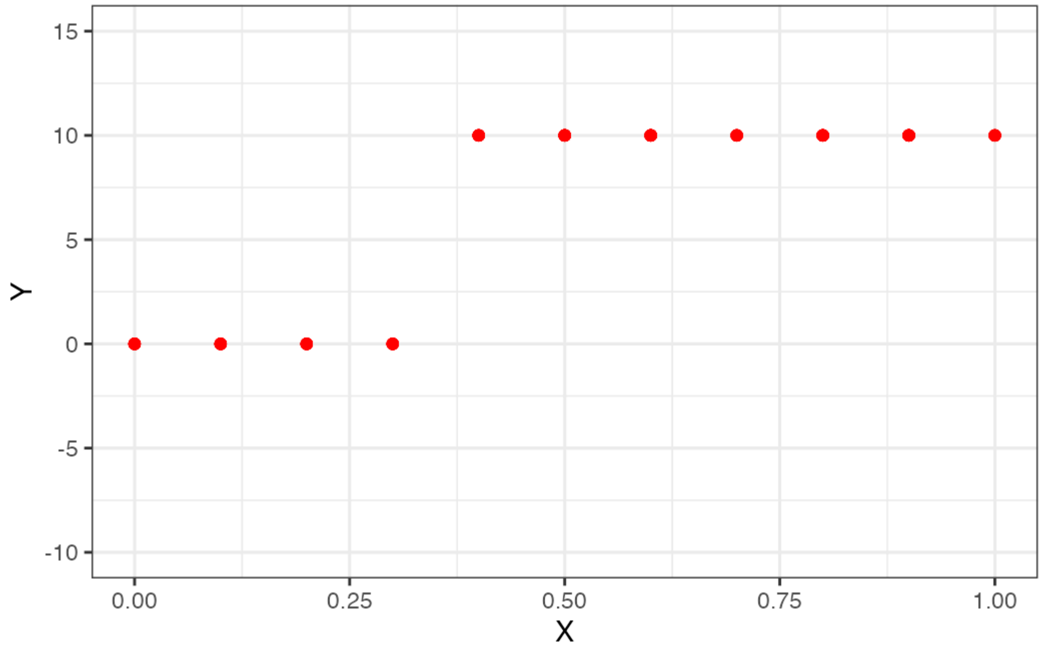
- モデルの複雑化 → 推定値の性質が変化し、**推定誤差**が拡大する
  - Population OLS とデータ上での OLS との乖離が広がる傾向が大きくなる
- OLS の推定値 $\hat{g}(X)$ $\underbrace{\simeq}_{\text{十分に単純であれば}}$ Population OLS

4.6 データ上の平均値 $\hat{\mu}(X)$

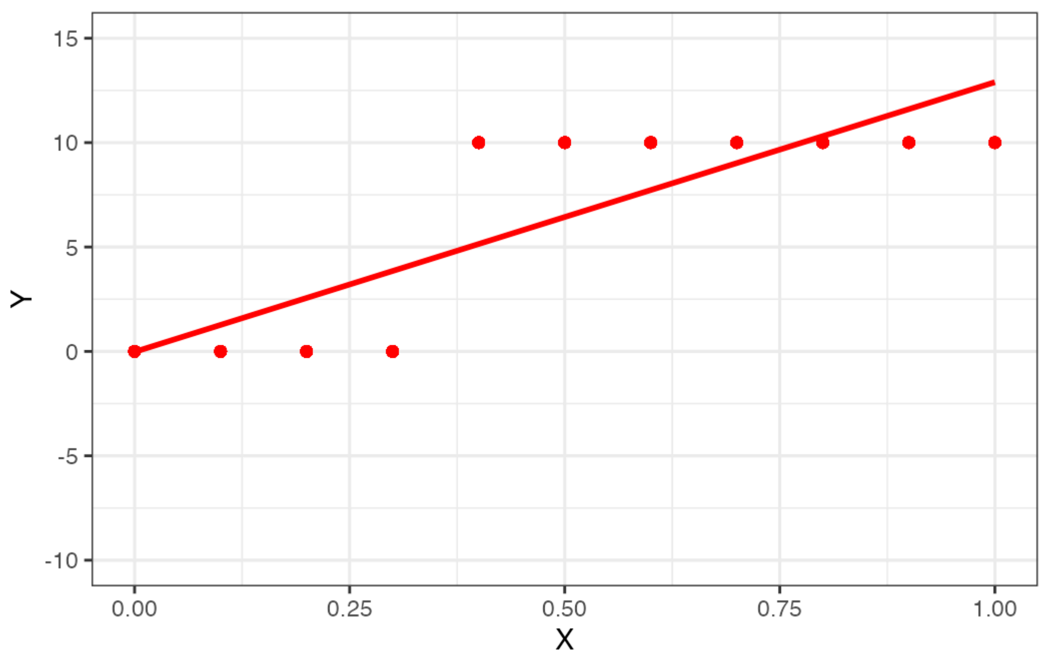
- X の組み合わせが多いと、$\hat{\mu}(X)$ は$\mu(X)$ の複雑すぎる推定値
- 例: 年齢 × 性別 × 教育年数 = 1598
- 1 事例で平均値を計算する組み合わせが頻出する

- 母平均と大きく乖離する可能性が高い

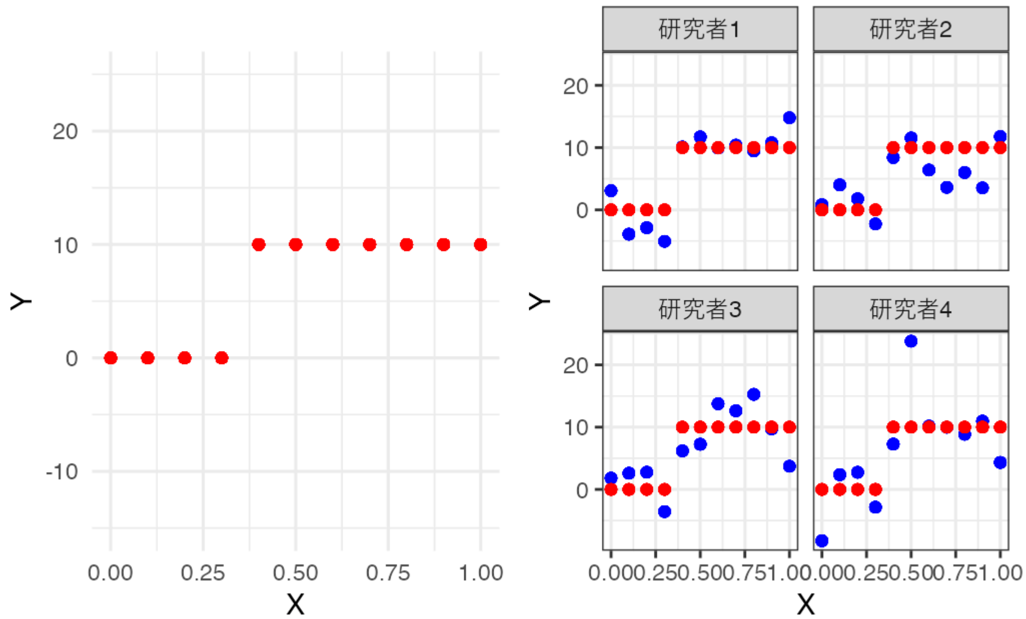
4.7 数値例: 母平均



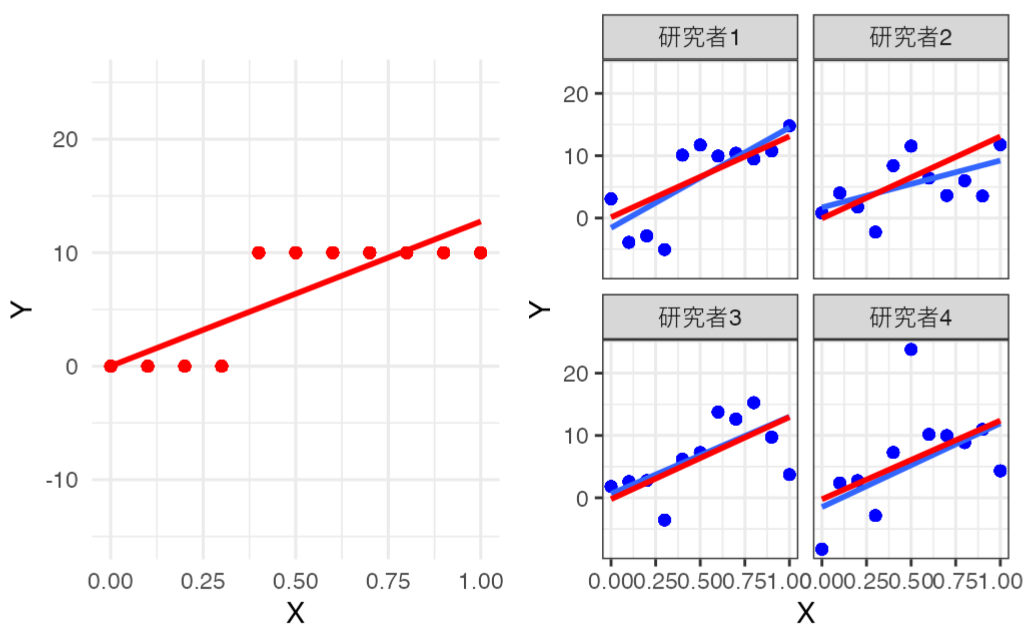
4.8 数値例: Population OLS



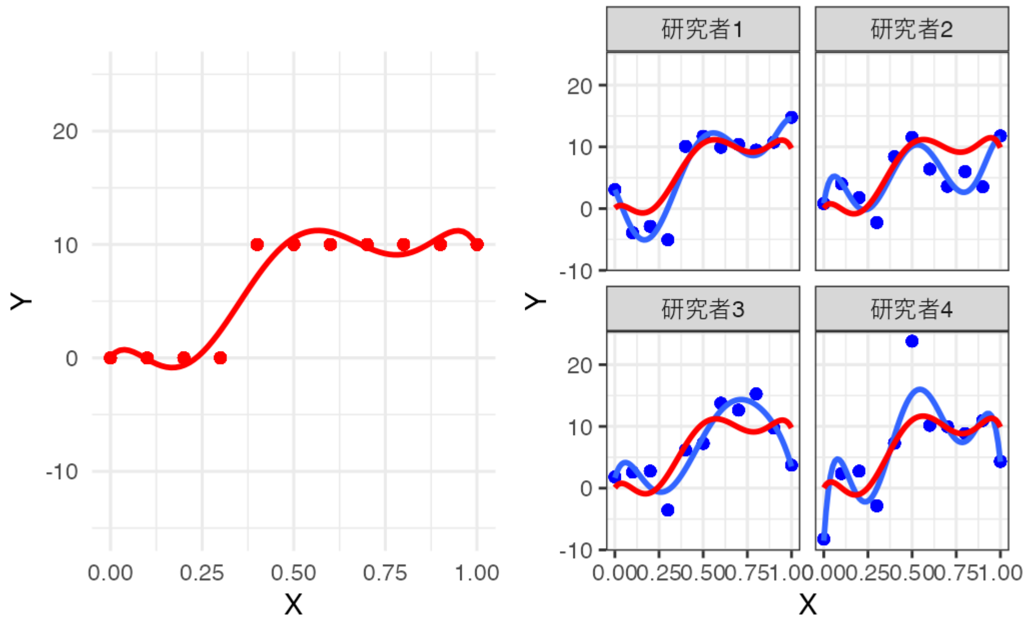
4.9 数値例: データ上の平均値



4.10 数値例: データ上の単純な OLS



4.11 数値例: データ上の複雑な OLS



4.12 まとめ

- Population OLS は**常に**、データ上での OLS の推定対象
    - 十分に複雑な Population OLS は、母平均を近似するので、母平均も推定対象
- 複雑な Population OLS を、データから推定しようとすると、推定精度が悪化する

4.13 まとめ

- 推定対象:

$$\text{母平均}\underbrace{\simeq}_{\text{モデルが十分に複雑}} PopulaitonOLS$$

- 推定値:

$$\underbrace{\simeq}_{\text{モデルが十分に単純}} \text{データ上の}OLS$$

4.14 関連文献

- BLP としての解釈
    - Applied Causal Inference Powered by ML and AI : 第 1 章
    - Angrist & Pischke (2009)
    - Aronow & Miller (2019)

- 川田作成のノート

5 線型記述モデル

5.1 研究目標

- 推定対象とすべき Population OLS の定式化は、研究目標に決定的に依存する
- 典型的な予測研究: 母平均 $\mu(X)$
  - Population OLS を母平均に近づけたい
- 社会のシンプルな記述: 社会の重要な特徴を把握しつつ、人間による認知が簡単な Population OLS を推定したい

5.2 研究目標

- 社会における教育/経験年数と平均時給との重要な関係性を、線形近似モデルとして把握したい
- ミンサー型賃金モデル (川口大司, 2011)

$$\log(wage) \sim \beta_0 + \beta_1 \times EducationYear$$

$$+\beta_2 \times Experience + \beta_3 Experience^2$$

- β_1 = "Return to education", β_2/β_3 = "Return to experience"
  - "Retrun to human capital"

5.3 識別

- 社会上でのミンサー型賃金モデルの計算値と、母集団(Target Population)上での計算値が一致する必要がある
  - データは、社会からランダムサンプリングされていると仮定する必要がある
  - 母集団上での分布 = 社会上での分布
- 標準的であり本講義でも想定する仮定だが、常に疑わしい仮定
  - 近年のチャレンジ

5.4 仮定のまとめ

- 社会における平均賃金の重要な特徴
  - $\underbrace{\simeq}_{\text{人的資本理論?}}$ Study population 上での OLS
  - $\underbrace{\simeq}_{\text{Study = Source}}$ 母集団上での OLS
  - $\underbrace{\simeq}_{\text{データ数に比べて、モデルが十分単純}}$ データ上での OLS

5.5 記述モデル

- Populaiton OLS を複雑にしすぎなければ、パラメータの解釈が明確
  - $\beta_1 =$ “モデル上"でX_1が”1 単位”大きかった時に、 Y の平均値がどの程度大きいか?
- 複数のXがあるケースで、特に重要
  - 人間には 4 次元以上が認識できず、可視化ができない

5.6 例: 賃金モデル (関数)

```
Fit <- CPS1985 |>
  lm(log(wage) ~ education + poly(experience, 2),
    data = _
  )

Fit$coefficients |>
  round(3)
```

```
          (Intercept)            education poly(experience, 2)1
                0.891                0.090                3.223
poly(experience, 2)2
               -2.025
```

5.7 記述モデルの問題点

- あくまでも母平均をさらに単純化したモデルであり、モデルの定式化次第では、母集団の特徴を見逃す恐れが高い
- 「重要な特徴をしっかり捉えるモデル」を事前に設定することは、難しい場合が多い
- 労働経済学においては、OLS を社会の線型記述モデルとして利用する研究は、少なくなっている
  - Balancing comparison の手法としての解釈が有力

6 補論: 頻度論

6.1 Replicability

- “科学的事実"を検証する黄金戦略: 独立した研究者が、同じ研究計画を実行すると、"同じ"結果を得る
  - 例: 水の沸騰温度を測定する実験室実験
- より複雑な社会/人間を対象とした研究でも、同じ戦略を適用したい

6.2 実証研究の研究計画

- 研究計画: 研究目標 (含む対象地域/時点)、推定目標、推定値の算出方法、データの収集方法や Cording すべき分析の内容
- 研究計画が確定しているのであれば、あとはデータを実際に入手し、パソコンにデータを流し込むだけ
  - 同じ結果を得ることができるか?

6.3 実証研究の Replicability

- ここまでは議論は、「同じ分析計画 → 同じデータ → OLS の推定値」
  - 同じデータなので、全員が必ず同じ推定値
- これからの議論は、「同じ分析計画 → データの入手 → OLS の計算」
  - どのようなデータが入手できるのかは、偶然(Sampling, データくじ)に決まる
    - 人によってデータが異なるので、異なる結果となる

6.4 手法検証/応用上の含意

- 多くの応用で、同じ研究計画を実施する研究者は自分達しか存在しない
  - 仮想的に、無数の”独立した”研究者をイメージする必要がある
  - あるいは、異なるデータを入手した場合の計算結果をイメージする

6.5 例

- 労働力調査を利用した推定: 日本全体から選ばれた 4 万家計を調査
  - 現実は一つの調査しか存在しないが、もし独立した研究者が同じ調査をやっていた場合、OLS の結果はどのように異なるのか?
- 国勢調査を利用した推定: 日本の全家計を調査
  - 労働力調査と同じイメージ
    - 対象家計は同じだが、回答が変化するかもしれない (測定誤差)

6.6 まとめ

- 同じ分析計画を実行する、"独立した"研究者をイメージ
- 同じ研究計画を採用したとしても、データを独立して取集すると、**推定値は異なる**
  - データに含まれる事例が、"偶然"異なるため
- 自身の推定結果は、「"偶然"計算された信用できない値」、と考える方が合理的
- 頻度論の枠組み

7 補論: 推定値の分布

7.1 サンプリングに伴う分布

- 分析計画 = データを推定値に変換
  - データくじの結果によって、推定値も異なる
    - 推定値の分布
- 現実に実現し、自身が観察する値はその中の一つだが、どれになるかは操作できない

7.2 推定値の分布についての性質

- 推定手法に応じて、推定値の**分布**の性質は操作できる
  - 研究者は、良い性質の**分布**を持つ手法を採用したい
- 現実生活の例: 旅行保険に入るかどうか
  - 現実に事故に遭うかどうかはわからないので、結果の分布を”良く”するように決定(保険に入った場合の被害、事故確率など)から判断

7.3 OLS の分布

- Population OLS の計算式

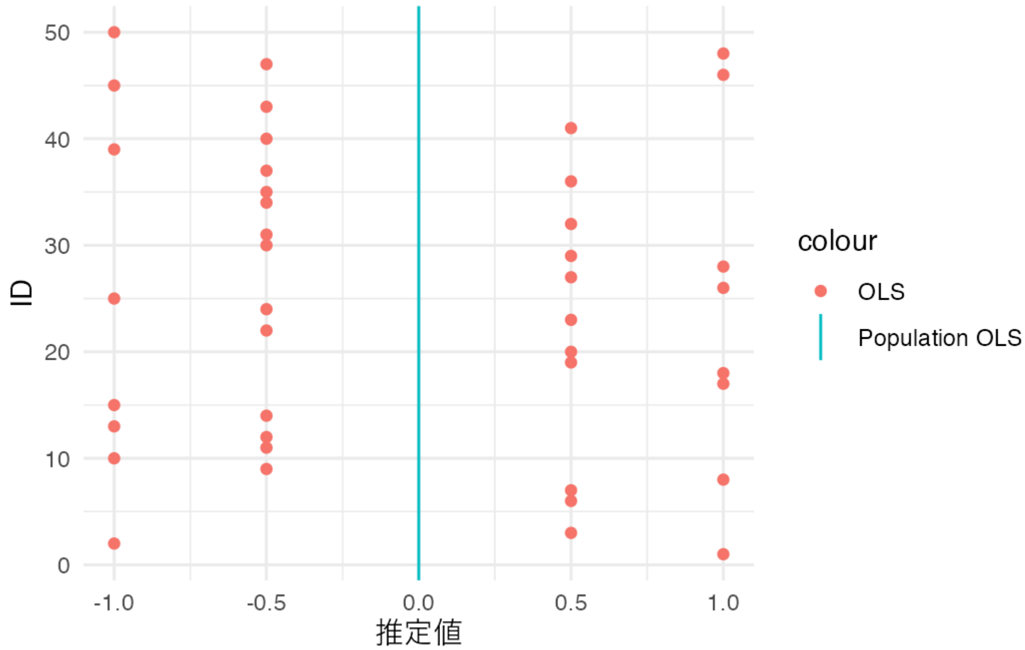
$$\hat{\mu}(X)^{Pop} = \hat{\beta}_0^{Pop} + ... + \hat{\beta}_L^{Pop} X_L$$

  - $\hat{\beta}^{Pop}$ は全員共通
- データ上の OLS

$$\hat{\mu}(X) = \hat{\beta}_0 + ... + \hat{\beta}_L X_L$$

  - データが異なるので、$\hat{\beta}$ の値も異なる
    - **推定値** の平均などを定義できる

7.4 イメージ: 3 事例



7.5 OLS の分布: 収束

- 事例数が大きくなれば、Population OLS に近い推定値を、ほとんどの研究者が得ることができる (収束する)
  - 「自分もそのような値を得ている可能性が高い」と考えられる

7.6 OLS の分布: 二つの収束性質

- 事例数が β の数に比べて、非常に大きければ、

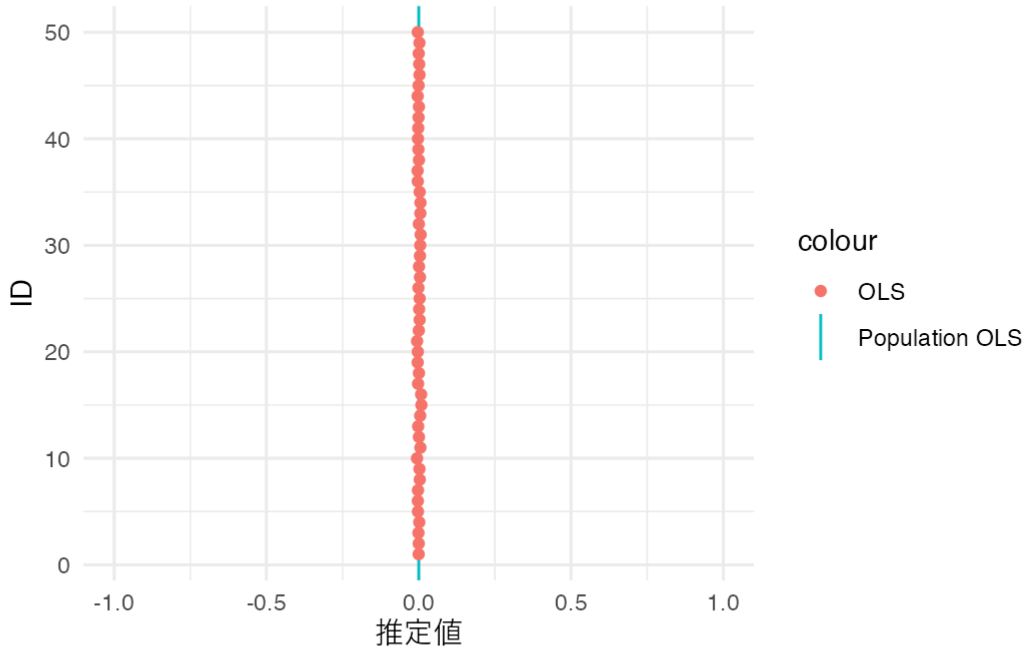
$$\left(\hat{\beta}_l^{Pop} - \hat{\beta}_l\right)^2 \textbf{の平均値} \rightarrow 0$$

- 事例数が β の数に比べて、**ある程度**大きければ、

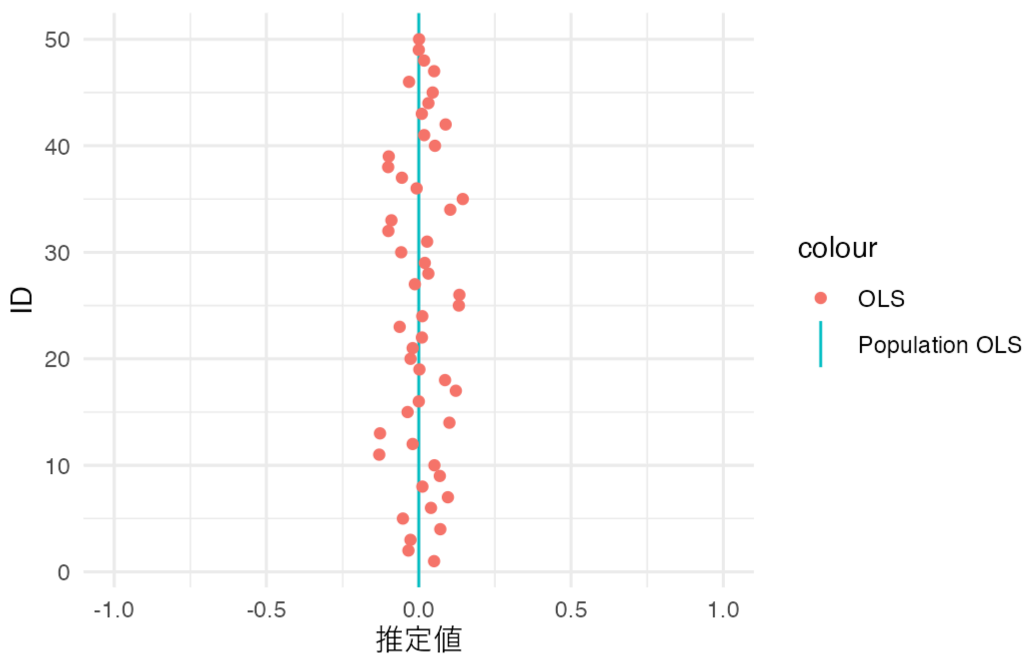
$$\left(\hat{\beta}_l^{Pop} - \hat{\beta}_l\right) \rightarrow \textbf{正規分布} N(0, \sigma^2)$$

  - 統計的推論の基礎となる

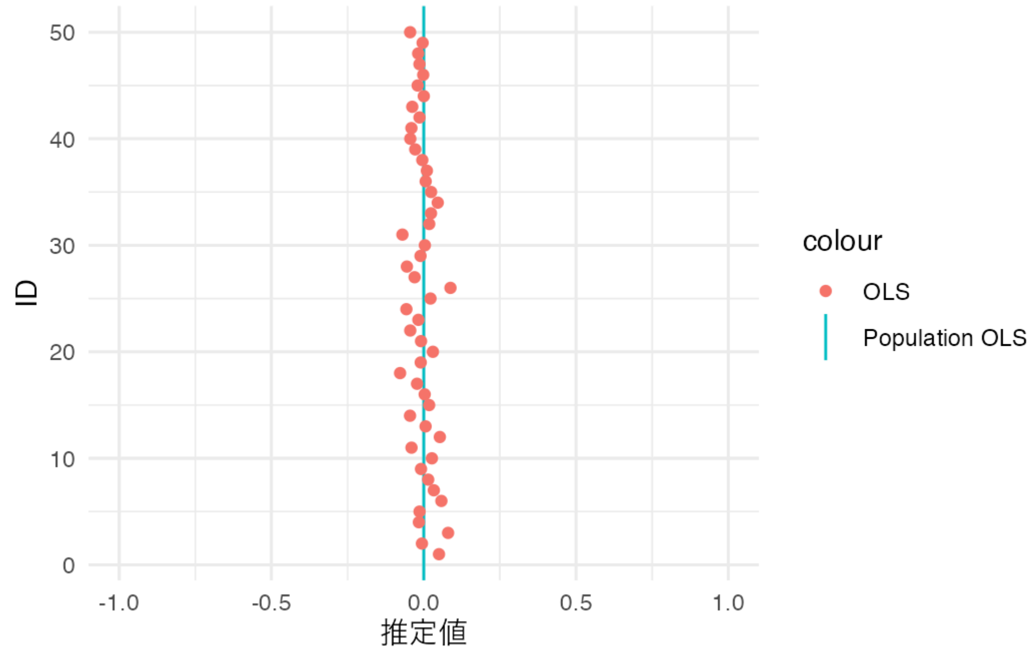
7.7 イメージ: 5 万事例



7.8 イメージ: 200 事例



7.9 イメージ: 1000 事例



7.10 Reference

Bibliography

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Aronow, P. M., & Miller, B. T. (2019). Foundations of agnostic statistics. Cambridge University Press.

Chattopadhyay, A., & Zubizarreta, J. R. (2023). On the implied weights of linear regression for causal inference. Biometrika, 110(3), 615–629.

Stock, J. H., & Watson, M. W. (2020). Introduction to econometrics. Pearson.

Wooldridge, J. M. (2022). Introductory Econometrics: A Modern Approach. Cengage learning.

川口大司. (2011). ミンサー型賃金関数の日本の労働市場への適用. 現代経済学の潮流, 67–98.