

OLS as Best Linear Projection estimator

労働経済学 1

川田恵介

1 論点整理法と OLS をしっかり”復習”

1.1 動機

- 実証分析は難しいので、論点整理が非常に重要
- OLS = 「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、別解釈が複数あり、現代的な予測/比較研究においても、現実的な選択肢 (Angrist & Pischke, 2009; Chattopadhyay & Zubizarreta, 2023)
 - ▶ 発展的手法も、OLS の特定の問題点を改善する方法である、と解釈できる場合が多い

1.2 OLS の入門書的解釈

- 賃金を年齢で OLS 回帰

```
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 $\mu(\text{age}) = E[\text{wage} \mid \text{age}]$ (Stock & Watson, 2020; Wooldridge, n.d.)
 - $\mu(\text{wage}) = \beta_0 + \beta_1 \times \text{age}$ を仮定する必要がある、非現実的

1.3 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

1.4 構成

- OLS について、

1. データ上で行なっている計算 —平均の線形近似—
2. 母集団上での推定対象 —母平均の線形近似—
3. 社会上での研究対象 —線型記述モデル—

1.5 まとめ

- OLS の推定対象は、複数存在する
 - ▶ 異なる解釈を有する
- よく紹介されてきた推定対象は、母平均
 - ▶ 比較研究/予測研究においては、他を推定対象とする手法と解釈した方が有益な場合が多い

2 個別事例分析の難しさ

2.1 実例: CPS1985

wage	ed- uca- tion	ex- peri- ence	age	eth- nic- ity	re- gion	gen- der	occu- pa- tion	sec- tor	union	mar- ried
5.10	8	21	35	hispanic	other	fe- male	worker	man- ufac- tur- ing	no	yes
4.95	9	42	57	cauc	other	fe- male	worker	man- ufac- tur- ing	no	yes
6.67	12	1	19	cauc	other	male	worker	man- ufac- tur- ing	no	no
4.00	12	4	22	cauc	other	male	worker	other	no	no
7.50	12	17	35	cauc	other	male	worker	other	no	yes
13.07	13	9	28	cauc	other	male	worker	other	yes	no
4.45	10	27	43	cauc	south	male	worker	other	no	no
19.47	12	9	27	cauc	other	male	worker	other	no	no

wage	ed- uca- tion	ex- peri- ence	age	eth- nic- ity	re- gion	gen- der	occu- pa- tion	sec- tor	union	mar- ried
13.28	16	11	33	cauc	other	male	worker	man- ufac- tur- ing	no	yes
8.75	12	9	27	cauc	other	male	worker	other	no	no

2.2 実例: ある事例

- データから、以下の事例を発見

wage	age	education	gender
1	42	12	male

- 42 歳/高校卒/男性は、時給 1 ドルで働いていた?

2.3 実例: 他の事例

wage	age	education	gender
10.75	42	12	male
1.00	42	12	male

- かなりの下振れ事例であることが確認できる
 - 有力な説明: 年齢/性別/学歴以外の賃金の決定要因

2.4 X を増やす

- X の増やせば、賃金が低い理由がわかる?
 - Y の決定要因が全て観察できれば、可能かもしれない
 - 少なくとも X 内で Y の個人差がなくなるほど X を増やせば、可能性がある
- 「ほとんどの応用では、複雑な個人差が存在し、絶望的」が一つの相場観

3 集団の記述

3.1 集団の特徴把握

- 労働経済学の重要な研究課題は、個別事例ではなく、集団の特徴把握
 - どのような家計/企業が、働いているか?/高い賃金を得ているか?/結婚しているか?/子供を持つか?/雇用を増やしているか?/賃金を増額しているか? 等

3.2 特徴把握の課題

- 事例 (=データ)や社会の特徴を直接的に把握することは困難
 - ▶ 大量の変数 (wage, education,..) について、大量の事例が存在しており、人間の認知能力をそもそも超えている
 - ▶ “誰が見ても明らかな”特徴”は、存在しない場合が多い

3.3 関心とする特徴の明示

- どのような特徴を分析対象とするのか、分析前に決定する
 - ▶ 例: 平均、分散、中央値
 - データから社会の特徴を推論する上でも必須
- 労働経済学では、変数間の関係性把握に焦点が当たりがち
 - ▶ 良い出発点: 平均値を要約する線型モデル
 - Y と X の関係性を簡潔に要約できる

4 平均値

4.1 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

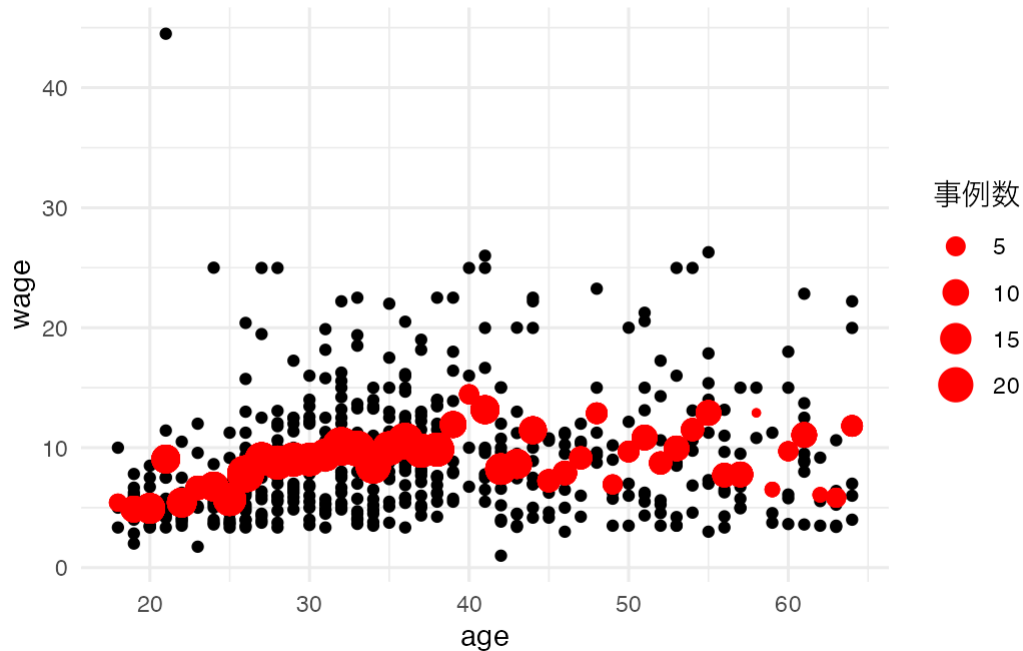
$$\hat{\mu}(X) = \frac{1}{(X_i = x) \text{ である事例数}} (Y_1 + Y_2 + \dots)$$

- 一般に、母平均 $\mu(X) \neq$ データ上の平均 $\hat{\mu}(X)$ であることに注意

4.2 平均値の利点

- 社会データは、 X 内での Y のばらつきが大きい傾向
 - ▶ Y の”重要な決定要因”が無数にあり、多くのデータで観察できない
- 平均値は Y と X の関係性を捉える、現実的な”要約方法”
 - ▶ 事例数が多ければ、 X 以外の要因による上振れ/下振れを抑制できる

4.3 例: 賃金と年齢



4.4 平均値の(労働研究における)問題点

- 非常に少ない事例のみから計算される平均値が発生しうる
 - ▶ X 以外の要因による上振れ/下振れの影響が強く、多くの問題が発生
 - 詳細は後述
- 多くの社会分析で、 X の組み合わせが多くなる
- 例: 年齢 \times 性別 \times 教育年数 = 1598
 - ▶ 0~1 事例しかないサブグループが頻出する

5 線形近似モデル

5.1 線形近似モデル

- 平均値を、さらに要約し、少数事例の影響を緩和するモデル
- β の足し算となるモデル
 - ▶ 単回帰:

$$\mu(\text{Age}) = \beta_0 + \beta_1 \times \text{Age}$$

- ▶ 重回帰:

$$\mu(\text{Age}, \text{Education}) = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Education}$$

5.2 線形近似モデル

- β の足し算であれば良いので、 X については変形できる
 - ▶ “ X ”について線形モデル:

$$\mu(\text{Age}) = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Age}^2$$

5.3 OLS

- データに極力適合するように β を選ぶ方法
 - ▶ β_0, β_1 は、以下を最小化するように推定する
- $$(\beta_0 + \beta_1 \times \text{Age} - \text{Wage})^2 \text{のデータ上の平均値}$$
- Y を近似するモデルと解釈できる

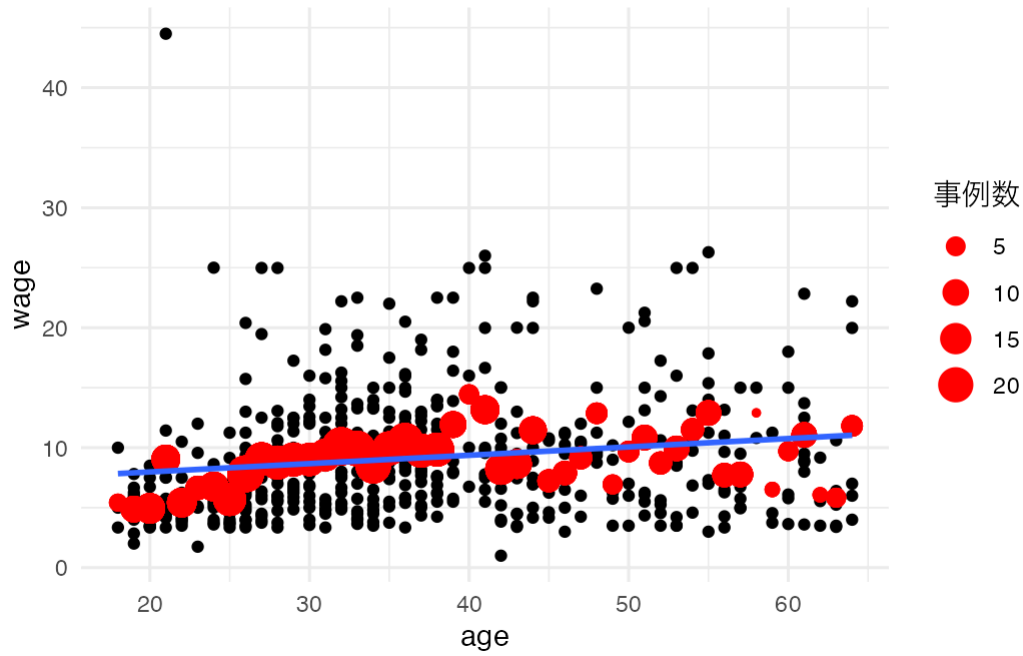
5.4 OLS の別解釈

- 以下を最小化しても、同じ β_0, β_1 が計算される
- 全ての $\text{age} = 15, 16, \dots$ について、

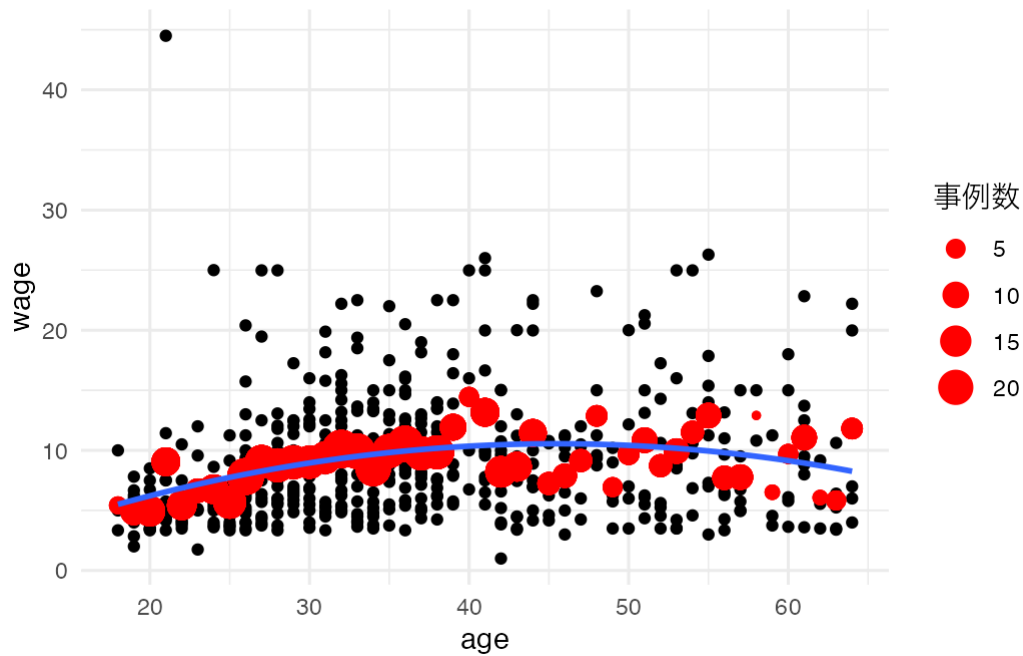
$$\left[\underbrace{(\beta_0 + \beta_1 \times \text{age} - \hat{\mu}(\text{age}))^2}_{\text{平均からの乖離}} \times [\text{Age} = \text{age} \text{となる事例割合}] \right] \text{の平均値}$$

- Y の平均値を近似するモデルと解釈できる

5.5 例 $\beta_0 + \beta_1 \text{Age}$



5.6 例 $\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Age}^2$



5.7 解釈の比較

- 労働経済学における多くの応用では、「 X がよく似ていても Y の値が大きく異なる」

- ▶ Y のモデルには見えない
- Y の平均値のモデルとみなしたほうが実践的

6 母集団上での推定対象

6.1 素朴な疑問

- ここまでは議論は、「同じデータ → OLS の推定値」
 - ▶ 同じデータなので、全員が必ず同じ計算結果を得る
 - 再現可能性がある
- 実際にはデータの収集も行う必要がある
 - ▶ 誰がやっても同じ結果がでる？
 - 水の沸騰温度を測定する実験であれば OK

6.2 想定する問題

- 研究計画は確定しており、分析コードも書いており、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
 - ▶ 分析計画: データの収集方法(対象地域や時期)や Coding すべき分析の内容
 - 「研究目標、推定目標 (Estimand)、推定値 (Estimator)の算出方法」 は、データを収集/入手する前に議論し、決定済み
 - 例: Pre-Analysis Plan by Kasy and Lehner

6.3 事例分析の問題点

- 同じ分析計画を実行する、複数の”独立した”研究者をイメージ: 事例を独立して収集し、データ化する
- 同じデータ収集方法(同じ地域/時点/サンプリング方法)を採用したとしても、**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

6.4 推定対象と推定値

- 建設的な議論のために、以下を分離する
 - ▶ 全ての研究者が原理的に合意できる正答 (推定対象)
 - ▶ 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

6.5 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
 - ▶ パネルデータについては、議論します

6.6 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平均値 (母平均)/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
 - 仮想的で誰も知ることができないが全員共通の値

6.7 まとめ

- 分析計画が確定したとしても、実際に収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 Uncertainty)
 - ▶ 信頼区間や p 値、機械学習におけるさまざまな工夫は、この不確実性への対処がメイン
 - よい手法 \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

6.8 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が、推定値はブレる」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

7 Population OLS

7.1 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
 - ▶ 以下、Population OLS は定義できる、と仮定する

7.2 Population OLS の推定

- OLS の推定値 = Population OLS の推定値

- ▶ β の数に比べて、事例数が非常に大きければ、全ての研究者が Population OLS とよく似た推定結果を得ることができる (一致性; Consistency)
 - Theorem 1.2.1 (Chapter 1, CausalML)
 - Sampling uncertainty は無視できる
- 詳細は、Section 9

7.3 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、 $\beta_0 + \beta_1 \text{Size}$ ではなく、 $\beta_0 + \beta_1 \text{Size} + \beta_2 \text{Size}^2$ の Population OLS

7.4 十分に複雑なモデル: 推定対象

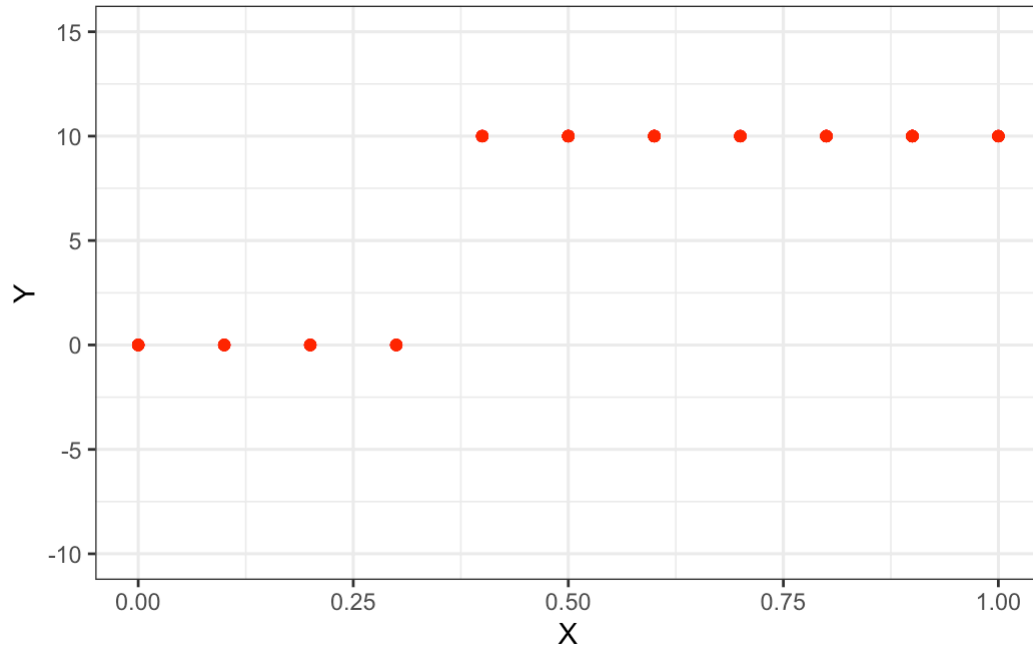
- モデルを複雑にすれば、Population OLS は、母平均に近づく
- OLS の推定対象 = Population OLS

- ▶ $\underbrace{\hspace{1cm}}_{\text{十分に複雑であれば}} \cong \text{母平均}$

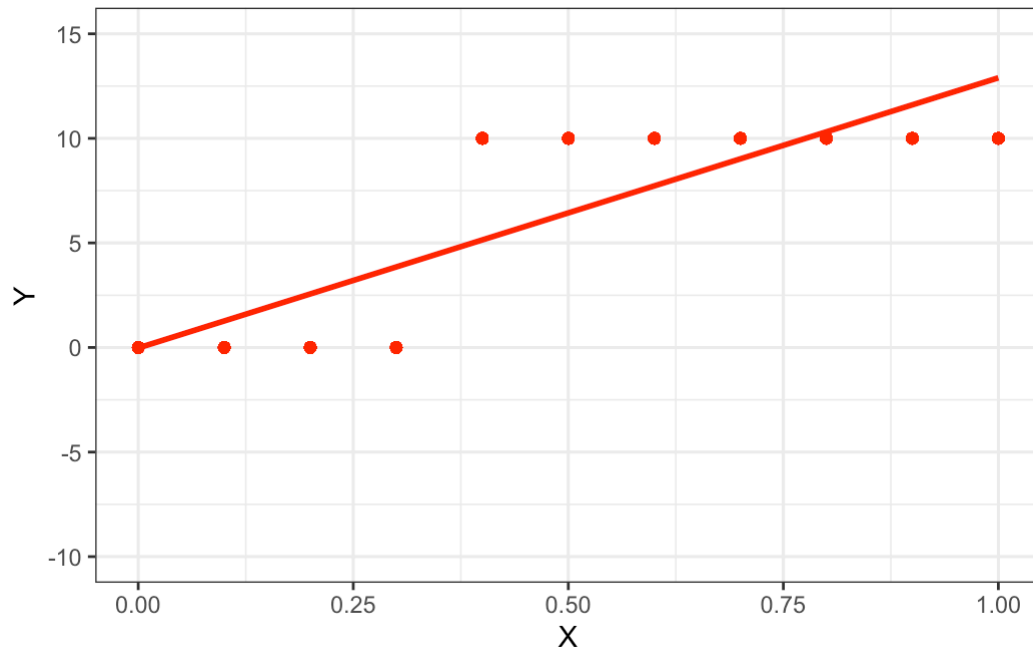
7.5 モデルの複雑化: 推定

- モデルの複雑化 → 推定値の性質が変化する、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
- 詳細は Section 9 参照

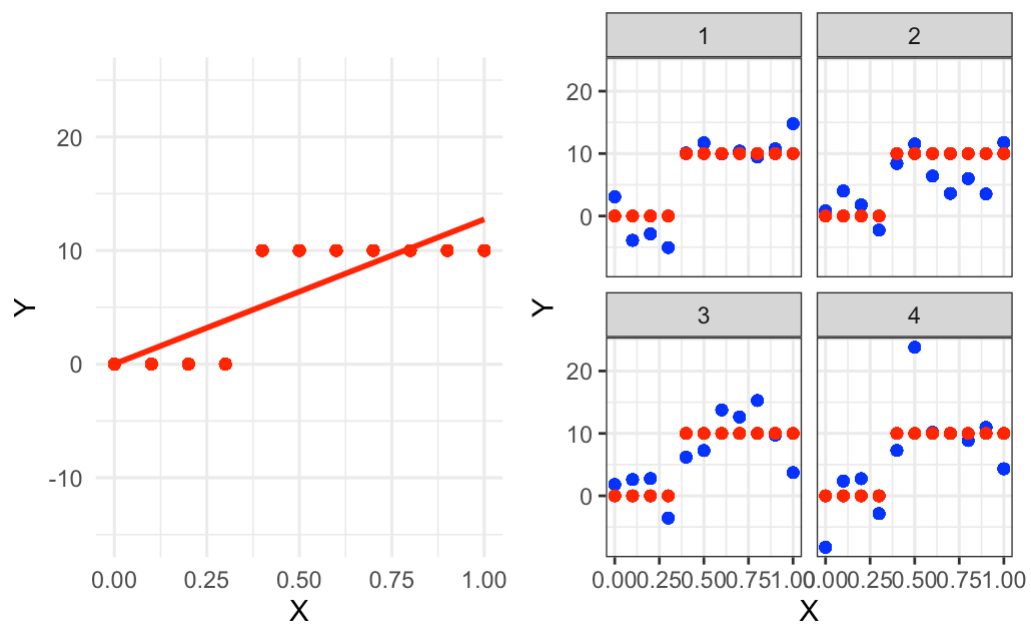
7.6 数值例: 母平均



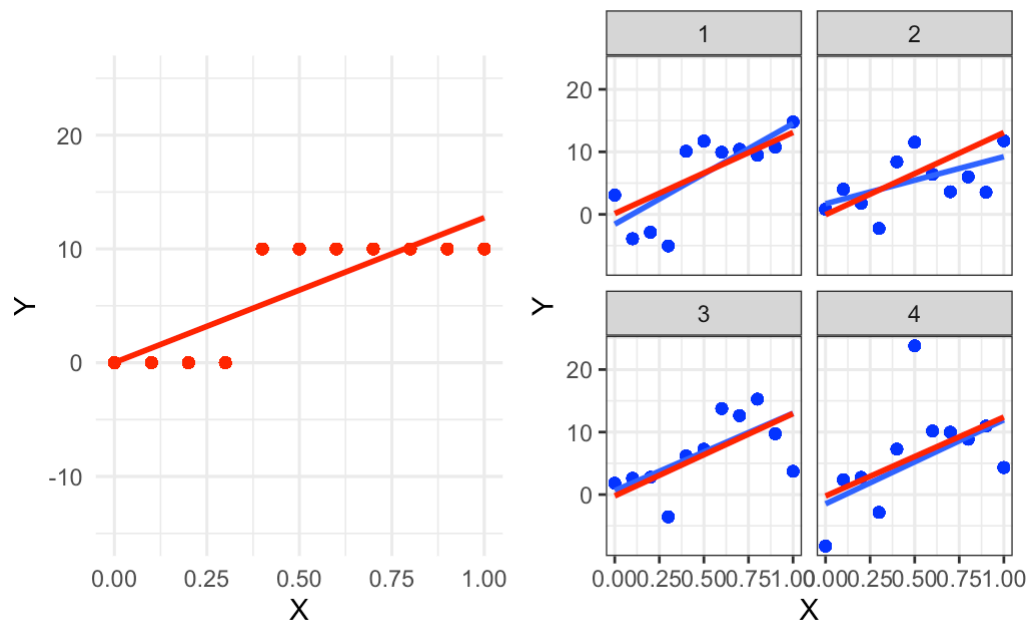
7.7 数值例: Population OLS



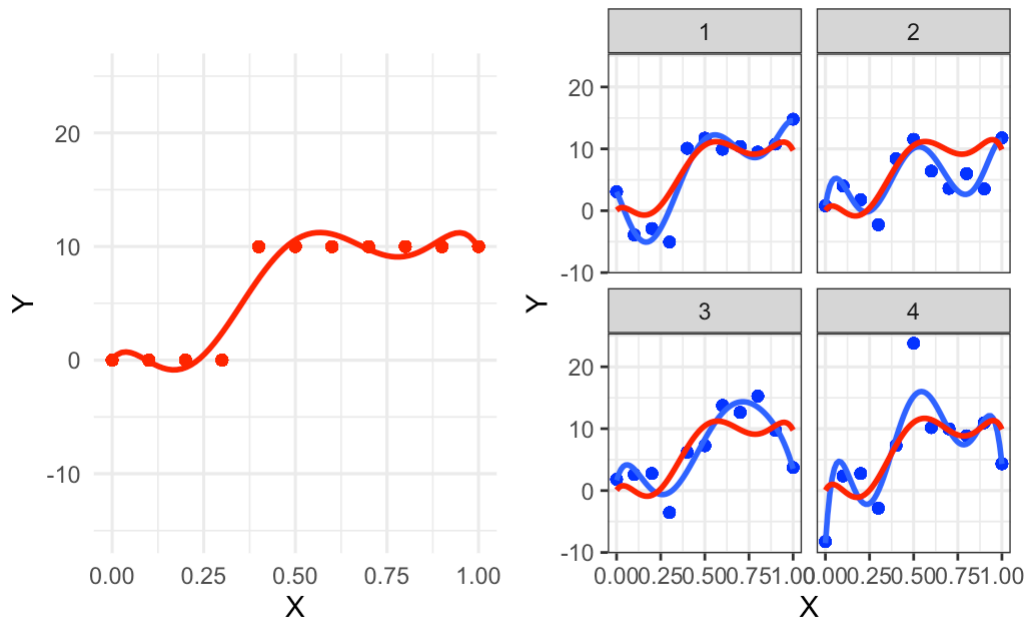
7.8 数値例: データ上の平均値



7.9 数値例: データ上の OLS



7.10 数値例: データ上の OLS



7.11 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、推定精度が悪化する
 - ▶ 一般に母平均を推定対象とするためには、推定精度悪化を受け入れる必要がある

8 線型記述モデル

8.1 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: 極力母平均 $\mu(X)$ に近い Population OLS を推定したい
- 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

8.2 研究目標

- 社会における教育/経験年数と平均時給との関係性を、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”
 - ▶ “Return to human capital”

8.3 識別

- 社会(Study population)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、Study population 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
 - ▶ 研究対象となる社会に対して、直接サンプリング調査ができている
- 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

8.4 仮定のまとめ

- 社会における平均賃金の特徴
 - ▶ \cong Study population 上での OLS
人的資本理論?
 - ▶ \cong 母集団上での OLS
Study=SourcePopulation
 - ▶ \cong データ上での OLS
データ数に比べて、モデルが十分単純

8.5 記述モデル

- Population OLS を複雑にしすぎなければ、パラメータの解釈が明確
 - ▶ β_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

8.6 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		

8.7 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と捉えるモデル」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

9 推定値の分布

9.1 サンプルングに伴う分布

- 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

9.2 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一般に異なる
 - ▶ 研究者は、実現する値を操作することはできないので、良い分布の性質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

9.3 OLS の分布

- Population OLS の計算式

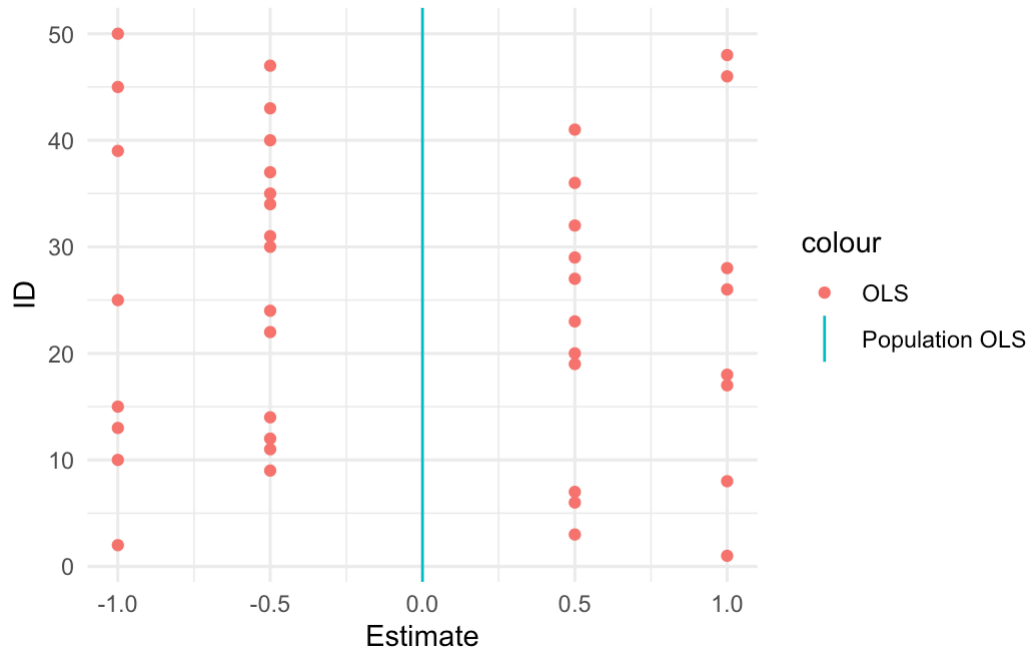
$$\hat{\mu}(X)^{Pop} = \hat{\beta}_0^{Pop} + \dots + \hat{\beta}_L^{Pop} X_L$$

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

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

- ▶ 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
 - 母平均やデータ上の平均値としっかり区別

9.4 イメージ: 3 事例



9.5 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

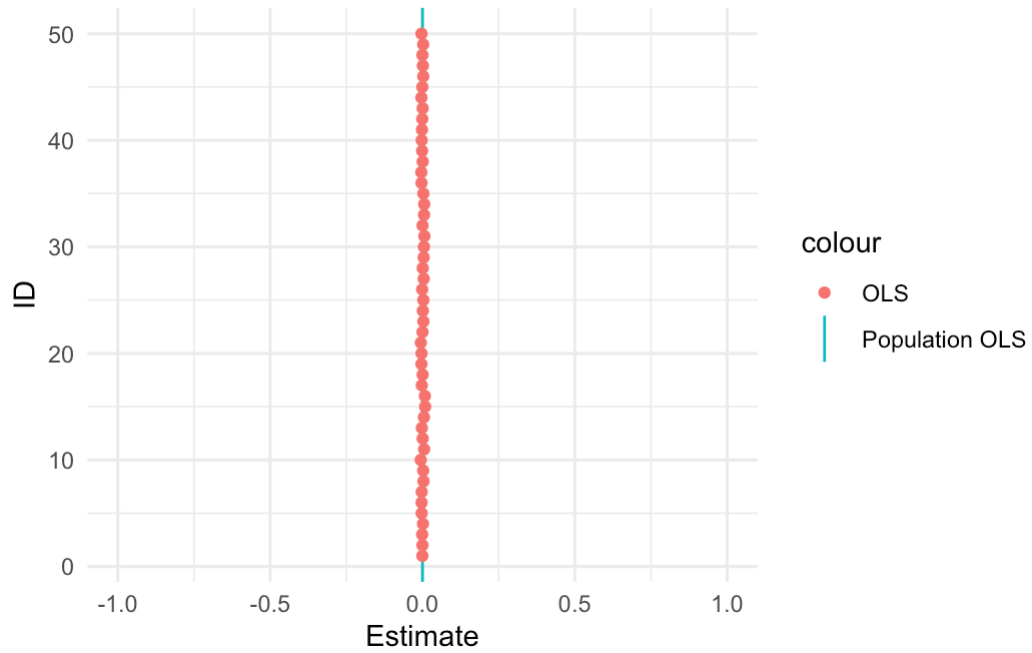
9.6 OLS の分布: 二つの収束性質

- 事例数が β の数に比べて、非常に大きければ、

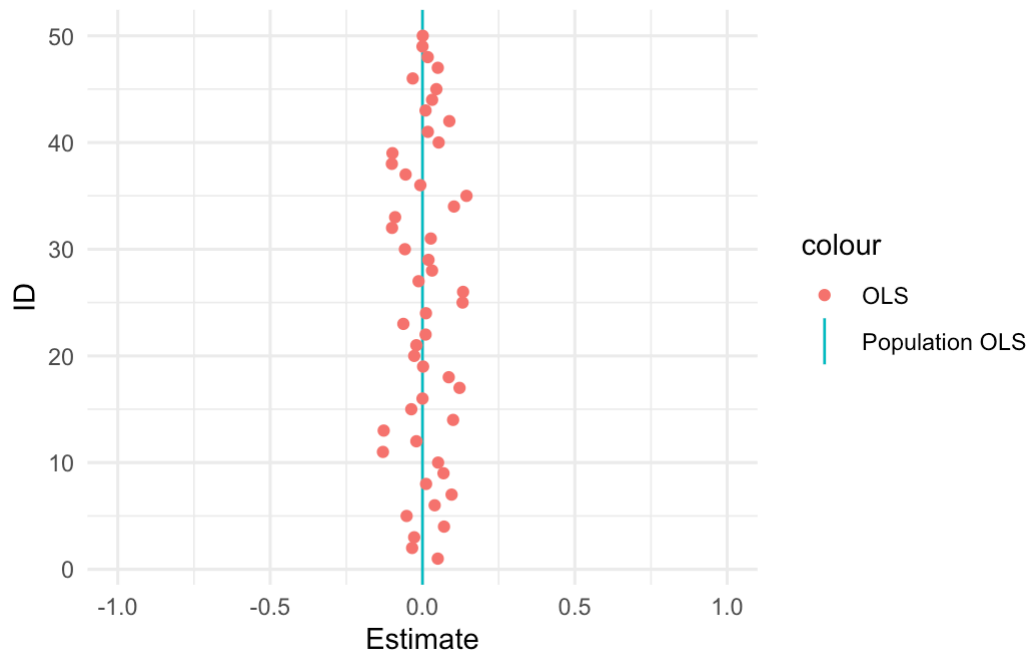
$$(\hat{\beta}_l^{Pop} - \hat{\beta}_l^{Pop}) \text{ の平均値} \rightarrow 0$$
- 事例数が β の数に比べて、ある程度大きければ、

$$(\hat{\beta}_l^{Pop} - \hat{\beta}_l^{Pop}) \rightarrow \text{正規分布}$$
 - ▶ 統計的推論の基礎となる

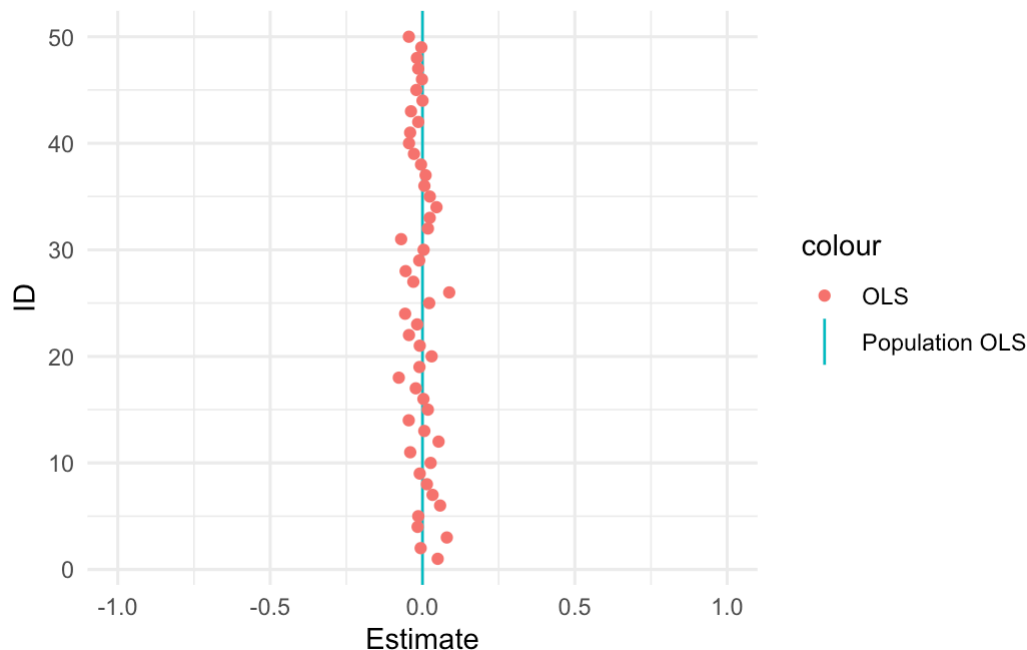
9.7 イメージ: 5 万事例



9.8 イメージ: 200 事例



9.9 イメージ: 1000 事例



9.10 まとめ

- 復習したい人は、以下も参照ください
 - ▶ 線形近似モデル
- よりしっかり復習したい人には、以下がおすすめです。
 - ▶ StatLect
 - 特に Plug-in Principle

9.11 Reference

Bibliography

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