OLS as BLP estimator

川田恵介

keisukekawata@iss.u-tokyo.ac.jp

2025-04-09

1 OLS の再解釈

1.1 OLS

- データ分析の”主力”選手: 多様な推定対象を、**悪くない性質**を保証しながら、推定ができる
  - 背後には、推定対象についての**別解釈**の存在がある
    - 別解釈を理解することで、分析の透明性を高めることができる
  - 機械学習の手法で補完することで、より妥当に推定できる

1.2 OLS の代表的解釈

- $$Y \sim \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$$

  を OLS 推定した場合の推定対象は何か?
- 古典的解釈: Y の(条件付き)母平均 $\mu(X) = E[Y \mid X]$ を推定対象とする
  - 例: Introductory Econometrics (Wooldridge), Introduction to Econometrics (Stock and Watson).
- $\mu(X) = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$ を仮定する必要があり、非現実的

1.3 OLS の別解釈

- 二つの別解釈: OLS の推定対象は
  - $\mu(X)$ の母集団上での線形近似モデル (Best Linear Projection)
  - $\mu(D = 1, X) - \mu(D = 0, X)$ 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、明確な推定対象を持ち、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
  2. **母集団上**での推定対象
- 次のスライドで、社会上での研究課題 (予測問題)、との関連性を議論
  - 先取りすると、"最善の予測モデルは母平均 $\mu(X)$" であり、OLS は予測問題においても有益

2 データのでの計算

2.1 例: ある事例

- データから、以下の事例を発見

Price (万円)	Size	T	District
150	80	1	杉並区

- 杉並区の 75 平米の物件は、1 億 5000 千万円で取引される傾向があると主張できる?

2.2 例: 他の事例

- 杉並区、75 平米の物件は以下の通り

Price	Size	T	District
90	80	1	杉並区
150	80	1	杉並区
110	80	1	杉並区
51	80	1	杉並区

- かなりの上振れ事例であることが確認できる
  - 可能な説明: Size 以外の要因(公園の近く/デザイナーズマンション….)

2.3 データ上の平均値

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

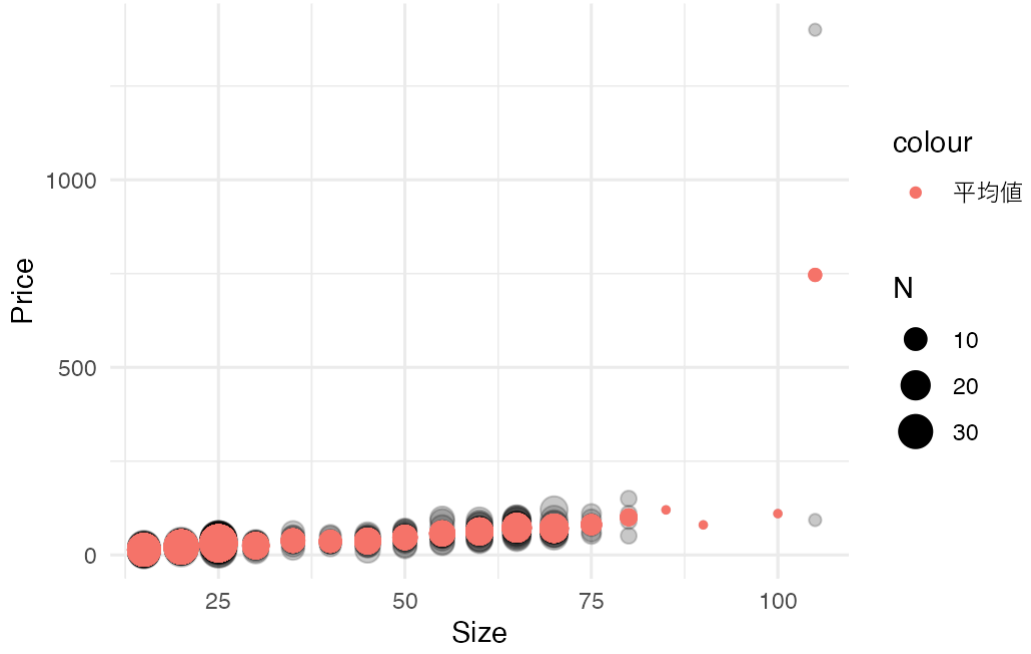
$$\hat{\mu}(X) = \frac{1}{(X_i = x)\text{である事例数}}(Y_1 + Y_2 + ..)$$

- ただし 事例 i について、$X_i = x$
- 一般に、母平均 $\mu(X) \neq \hat{\mu}(X)$ であることに注意

2.4 平均値の利点

- 社会データは、一般にX内でのYのばらつきが大きい傾向
  - Yの”決定要因”が、X 以外にも多い傾向
- 平均値はYとXの関係性を捉える、有力な”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

2.5 例



2.6 平均値の問題点

- 多くの応用で、非常に少ない事例のみから計算される平均値が発生
  - 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

2.7 社会分析との相性

- 多くの社会分析で、Xの組み合わせが多くなる
  - $[\textbf{広さ}, \textbf{立地}] = [\{15, \textbf{文京区}\}, \{20, \textbf{新宿区}\}, ..]$：437 個の組み合わせが存在
  - $[\textbf{広さ}, \textbf{立地}, \textbf{築年数}, \textbf{駅からの距離}, \textbf{区域}]$：4931982 個の組み合わせが存在

2.8 OLS

- 平均値を、"さらに要約する"モデルを計算する
- 例 $Price = \beta_0 + \beta_1 \times Size$ で $\hat{\mu}(Size)$ の特徴を捉える
  - β_0, β_1 は、以下を最小化するように推定する

$(\beta_0 + \beta_1 \times Size - Price)^2$のデータ上の平均値

2.9 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 size $= 15, 20, 25, ...$ について、

$$\underbrace{(\beta_0 + \beta_1 \times size - \hat{\mu}(size))^2}_{\text{平均からの乖離}}$$

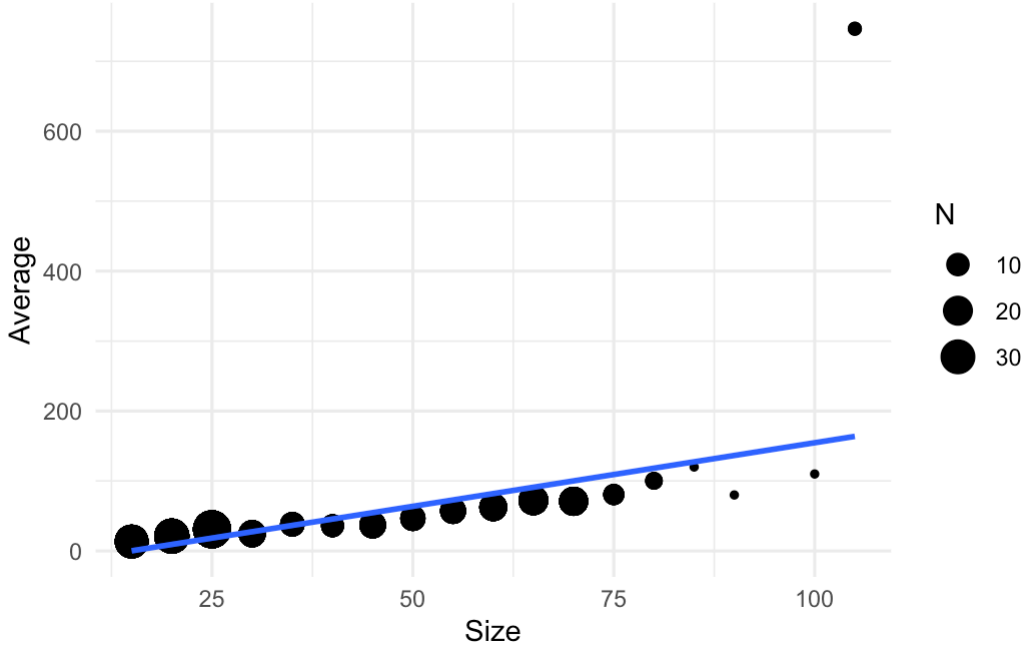
$$\times [Size = size\text{となる事例割合}]$$

の平均値

2.10 例

Average	OLS	乖離	Size	N
746	164	338724	105	2
110	155	2025	100	1
80	136	3136	90	1
120	127	49	85	1
100	118	324	80	4
81	109	784	75	7

2.11 例



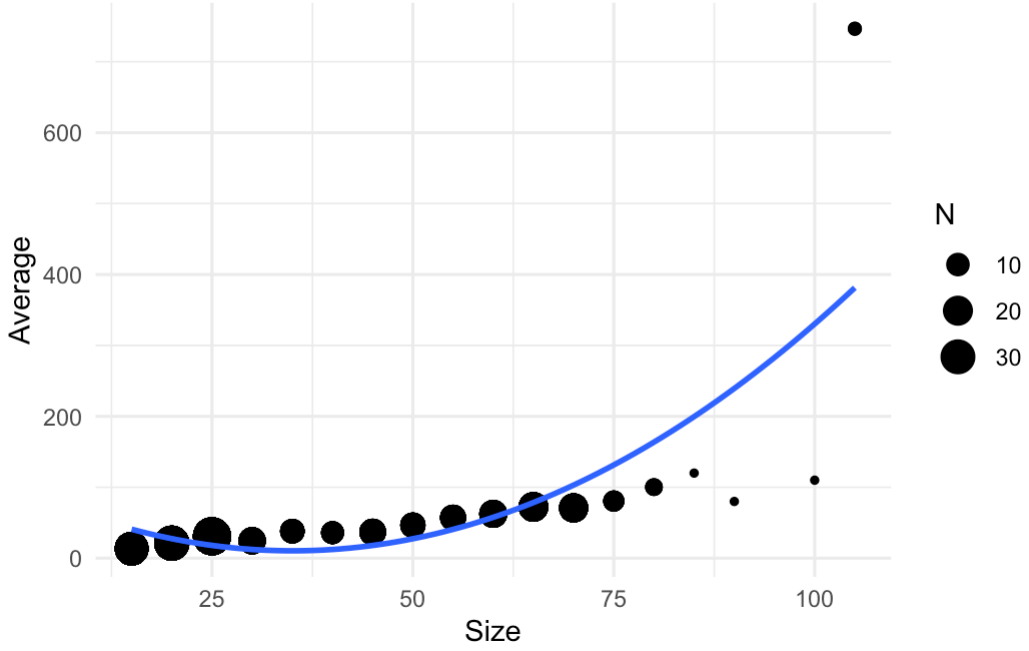
2.12 モデルの複雑化

- よりβの数が多い(複雑な)モデルも当てはめられる
  - 単純なモデル例: $\beta_0 + \beta_1 \times X$
  - 複雑なモデル例: $\beta_0 + \beta_1 \times X + \beta_2 \times X^2 + .. + \beta_{10} \times X^{10}$
- 複雑なモデルを推定すると、$\hat{\mu}(X)$ により近づく
  - 注意: 元々のXを増やしている (新しい属性を追加している) わけではない

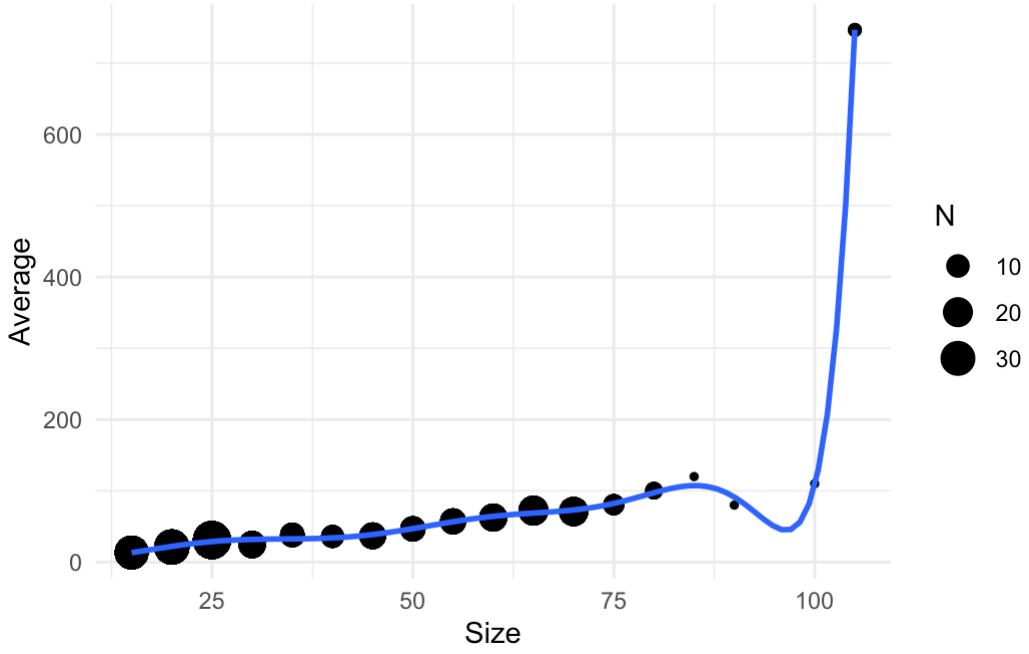
2.13 例

Average	単純	単純なモデルの誤差	複雑	複雑なモデルの誤差	Size	N
746	164	338724	747	1	105	2
110	155	2025	108	4	100	1
80	136	3136	91	121	90	1
120	127	49	107	169	85	1
100	118	324	97	9	80	4
81	109	784	82	1	75	7

2.14 例: X の二乗



2.15 例: X の 10 乗



2.16 複雑化の問題点

• モデルを複雑化すると、平均値にいくらでも近づけることができる

- データとの矛盾が減るので、一見よさそうだが、
  - そもそもの動機は「極めて少数の事例の集計を避けるために」と矛盾
  - 推定精度が悪化する
- より正確に議論するには、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 データ分析の問題点

- 複数の”独立した”研究者をイメージ: データを独立して収集する
- 同じ推定手法/データ収集計画(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
  - 例: 報道機関による世論調査
- 自身の推定結果は、「"偶然"計算された値」、と考える方が合理的
  - データ分析から、建設的なメッセージを引き出せるか?

3.2 推定対象と推定値

- 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
- 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
  - 母集団を用いて**仮想的に**計算される値は、全員共通
    - 推定対象 = **仮想的**で**誰も知ることができない**値

3.4 注意点

- 推定対象は、**仮想的な値**であり、その正確な値は**"誰も知ることができない"**
  - データから正確に知るためには、無限大の事例数が必要なため
- 「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」という複雑な問題設定であり、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

3.5 OLS の整理

- 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
  - 全員共通
- OLS の推定値 = Population OLS の推定値
  - 人によって異なるが、Population OLS の優れた推定値となりうる
    - β の数に比べて、事例数が十分に大きければ、全ての研究者が Populaiton OLS とよく似た結果を得ることができる (一致性; Consistency)

3.6 複雑なモデルのコスト

- β の数が増えると**推定精度**が悪化する
  - 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
- Threorem 1.2.1 (Chapter 1, CausalML): βの数/事例数が大きくなると、Population OLS と データ上での OLS の乖離も大きくなる傾向

3.7 複雑なモデルの利点

- 伝統的な教科書の序盤の章では、OLS は母平均の推定値であると紹介されることが多い
- もし”Population OLS = 母平均”であれば、正しい
- モデルを複雑にすれば、**Population OLS は**、母平均に近づく
  - Section 2.12 と同じ理屈

3.8 数値例

3.9 まとめ

- Population OLS は**常に**、データ上での OLS の推定対象
  - 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
- 母平均を推定対象とするためには、複雑な Population OLS を推定する必要がある
  - 推定精度悪化とのトレードオフが生じる

3.10 関連文献

- Applied Causal Inference Powered by ML and AI : 第 1 章
- Angrist & Pischke (2009)
- Aronow & Miller (2019)

3.11 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.