Recap: OLS for Best Linear Projection Model

川田恵介

Table of contents

1 OLS

- 研究者が**事前に設定した** Linear Model を、データに最も当てはまるように推定する **Algorithm**
  - ランダムサンプリングであれば、**母集団上**の解釈 (最善の線形近似, Best Linear Projection) を有する
  - 事例数に比べて、パラメタの数が少ないモデルであれば、**上手く推定**できる

1.1 Linear Model

- Y と X の Linear model: $g_Y(X)$

$$g_Y(X) = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$$

  - $\beta = [\beta_0, .., \beta_L]$: パラメタ (Parameter)
  - $X = [X_1, .., X_L]$: 変数 (Variable)

- 注: X については、NonLinear でも良い

$$g_Y(X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

  - β について Additive (足し算) である必要がある

1.2 Algorithm

- データをモデルに変換する手順
- モデルと Algorithm は分離して理解すべき
  - Linear Model を推定する Algorithm は大量に存在 (OLS, 最尤法, ベイズ法, LASSO, Ridge)
  - OLS は、いくつか望ましい性質を持つ

1.3 OLS アルゴリズム

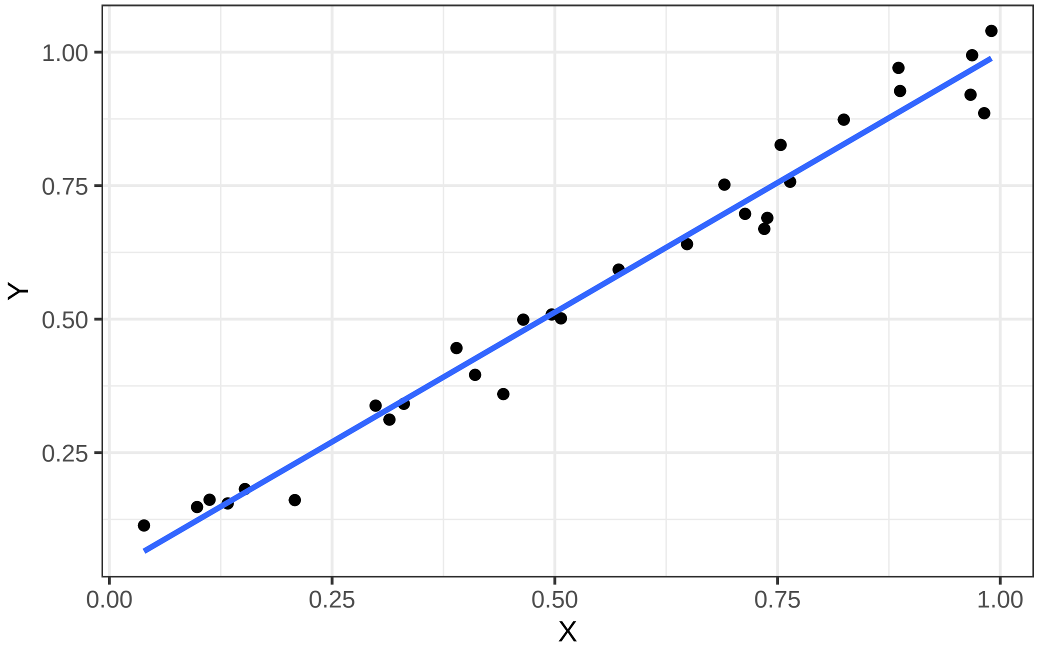
- 仮定: 多重共線性 (wiki) が無い

0. **分析者**が、モデル $g_Y(X) = \beta_0 + .. + \beta_L X_L$ を設定
1. $\beta = [\beta_0, .., \beta_L]$ を二乗誤差の総和を最小にするように決定

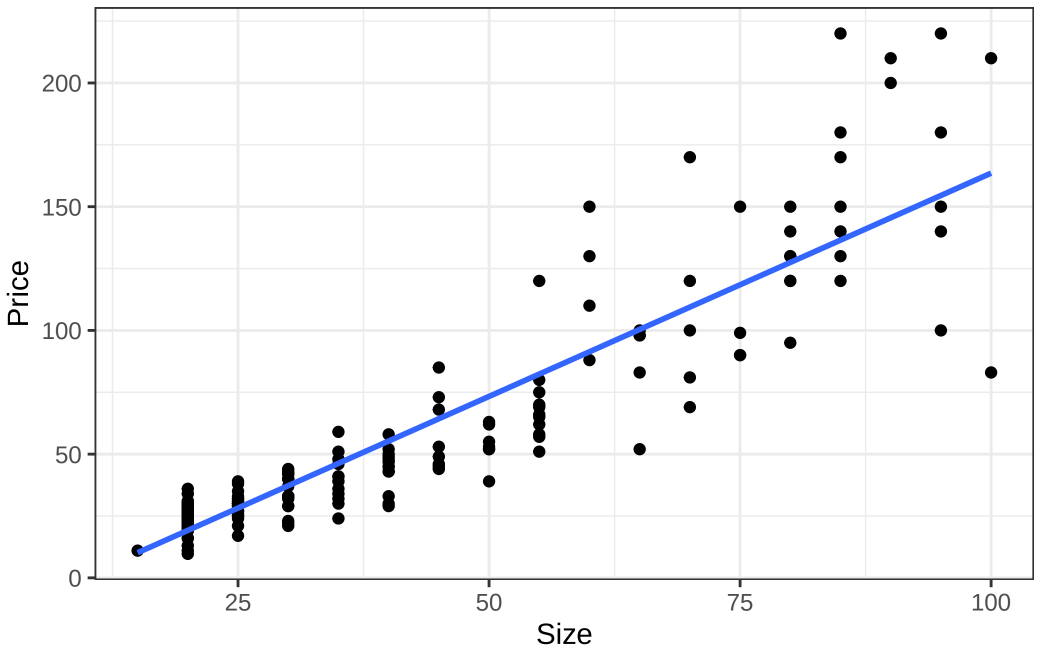
$$\min \sum_i^N (y_i - g_Y(x_i))^2$$

  - $X := [X_1, .., X_L]$

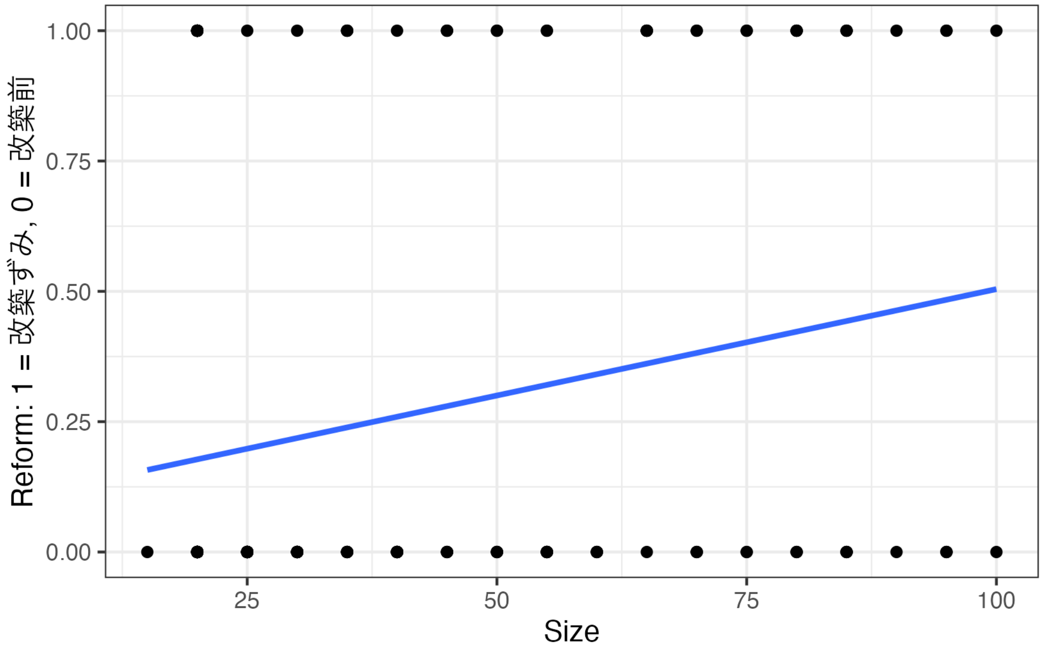
1.4　理想的な例



1.5　実際



1.6 実際



1.7 データの要約

- Y に極力合うように推定 = "Y の要約モデル" として紹介されることも多いが、
  - Y のモデルに" 見えない" 応用も多い
    - 多くの応用で、X 以外の Y の決定要因が大量に存在し、観察できない個体差が顕著
- 有力な別解釈が存在

1.8 OLS アルゴリズム (その 2)

- 仮定: 多重共線性 (wiki) が無い

0. **分析者**が、モデル $g_Y(X) = \beta_0 + .. + \beta_L X_L$ を設定
1. $\beta = [\beta_0, .., \beta_L]$ を二乗誤差の総和を最小にするように決定

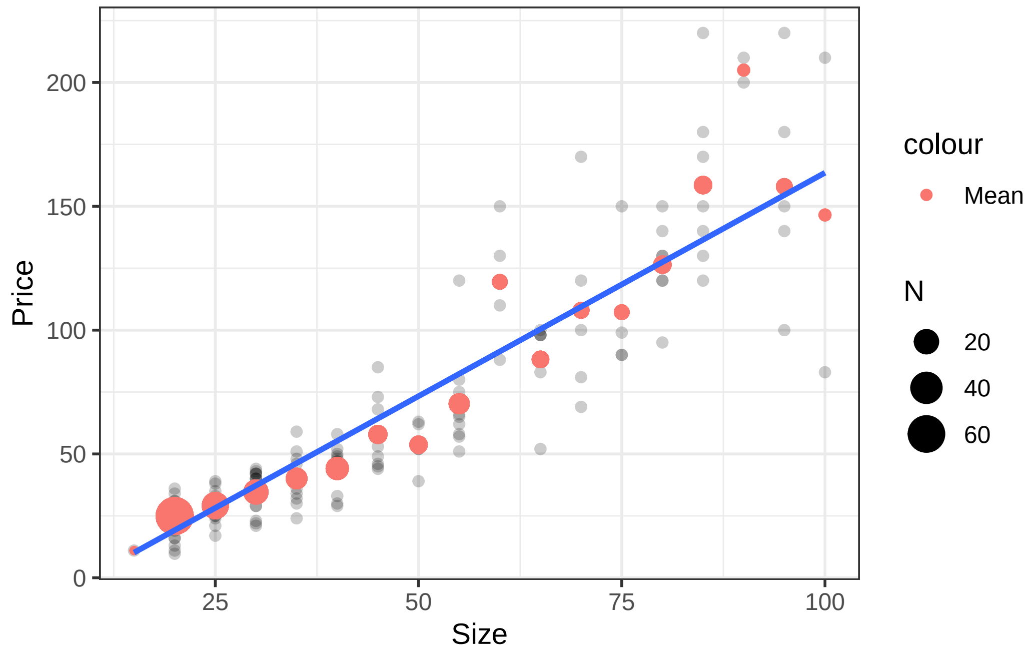
$$\min \sum_X (\mu_Y(X) - g_Y(X))^2 \times N_x$$

  - $\mu_Y(x) = \sum_{i; X_i = x} y_i / N_x$, $N_x : X_i = x$ を満たす事例数

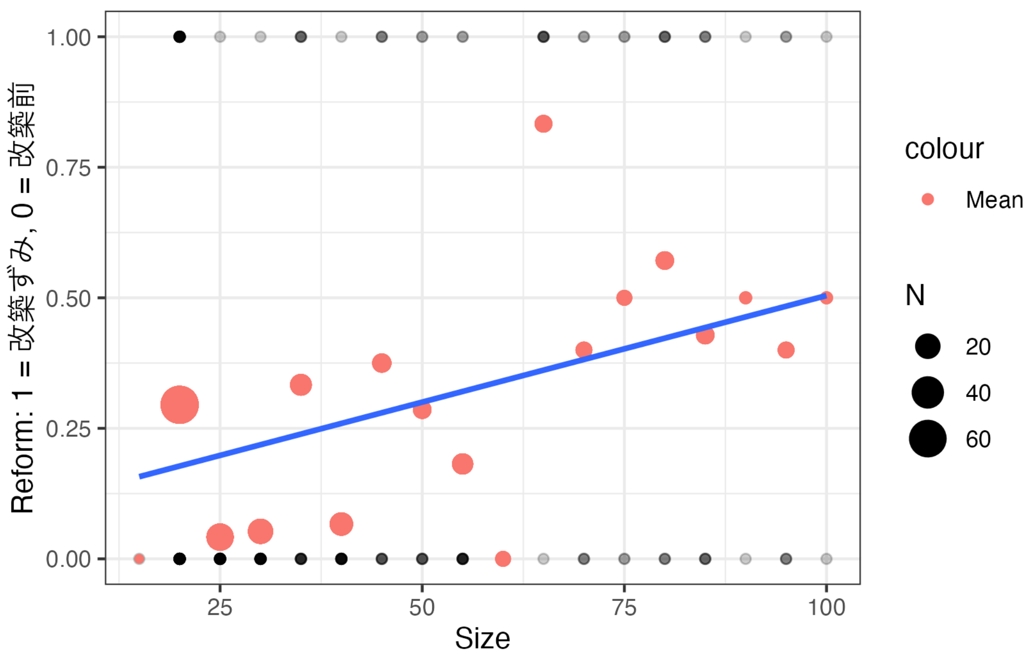
- Y の**平均値**のモデル

– OLS Algorithm と同じ推定結果を導く

1.9 実際



1.10 実際



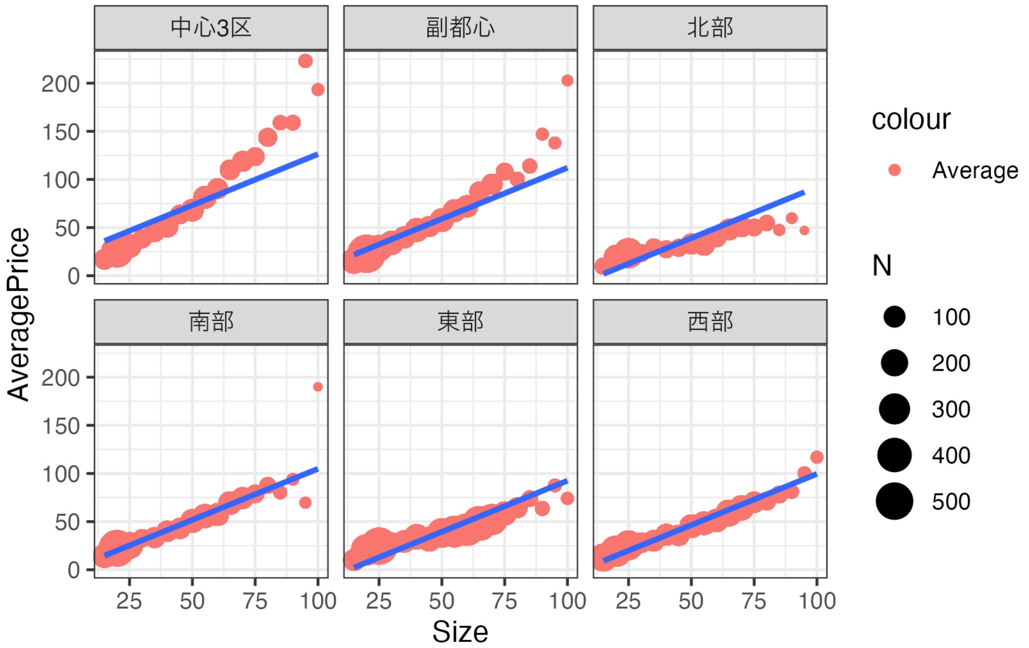
1.11 実例

- 2022 年の不動産取引データを用いて、以下の Linear model を OLS で推定

$$g(X) = \beta_0 + \beta_1 \times Size$$
$$+\beta \times Dummies(District)$$

- Dummies(x): x のダミー変数
- 必ず共通の傾きを持った直線モデルが推定される

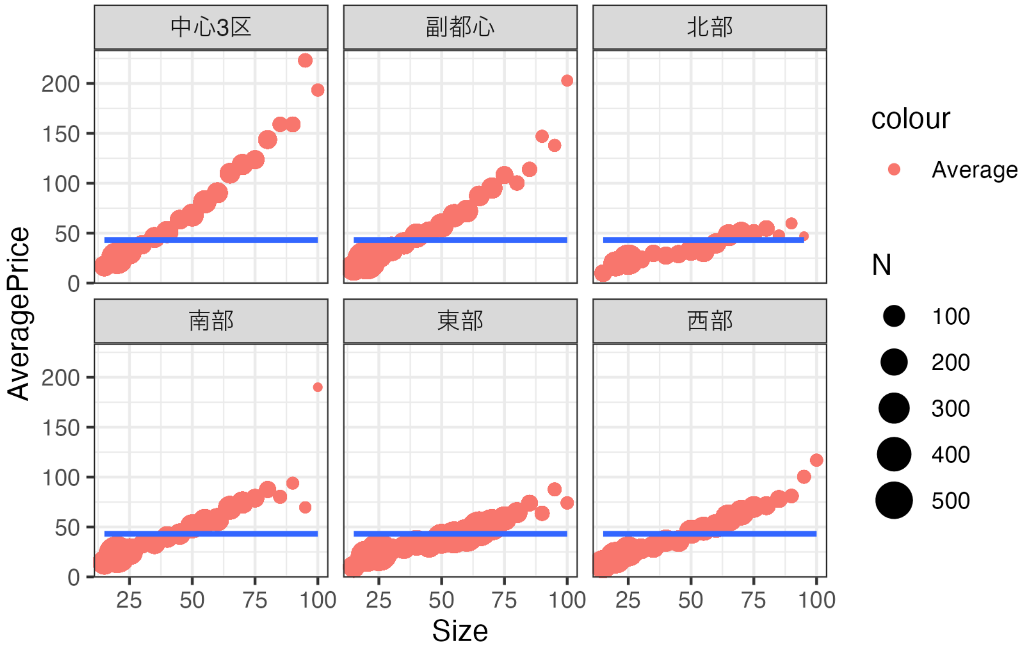
1.12 実例: シンプルモデル



1.13 実例: 最も単純なモデル

- $$g_Y(X) = \beta_0$$
- OLS で推定すると $\beta_0 = Y$ の平均値

1.14 実例: シンプルモデル



1.15 実例

- より複雑なモデル

$$log(Price) =$$

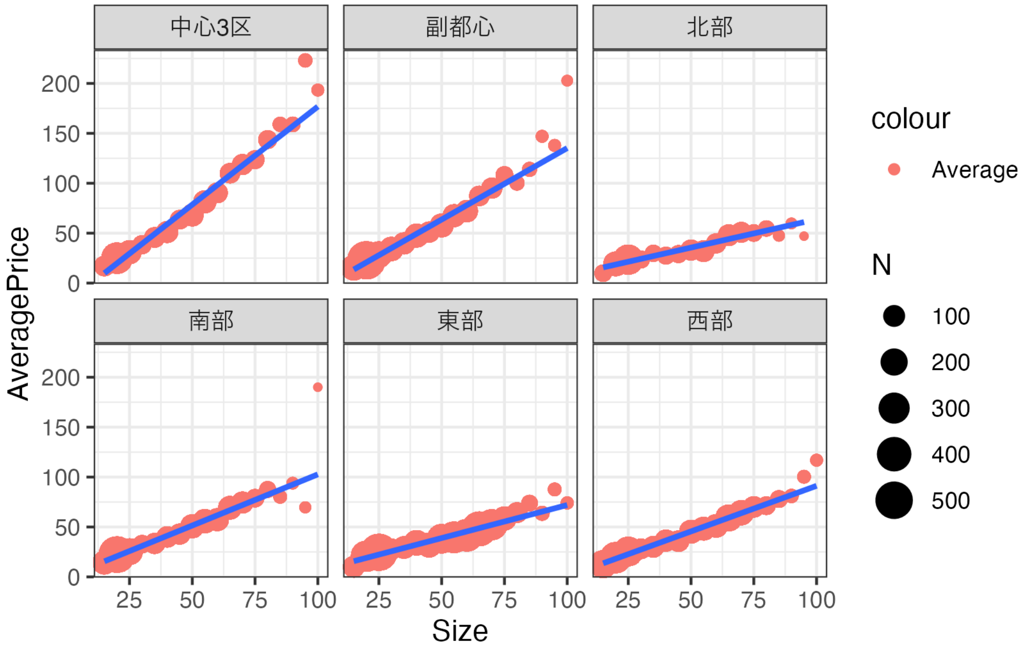
$$\beta_1 \times Size \times Dummies(District) \tag{1}$$

$$\beta_1 \times poly(Size, 2) \times Dummies(District) \tag{2}$$

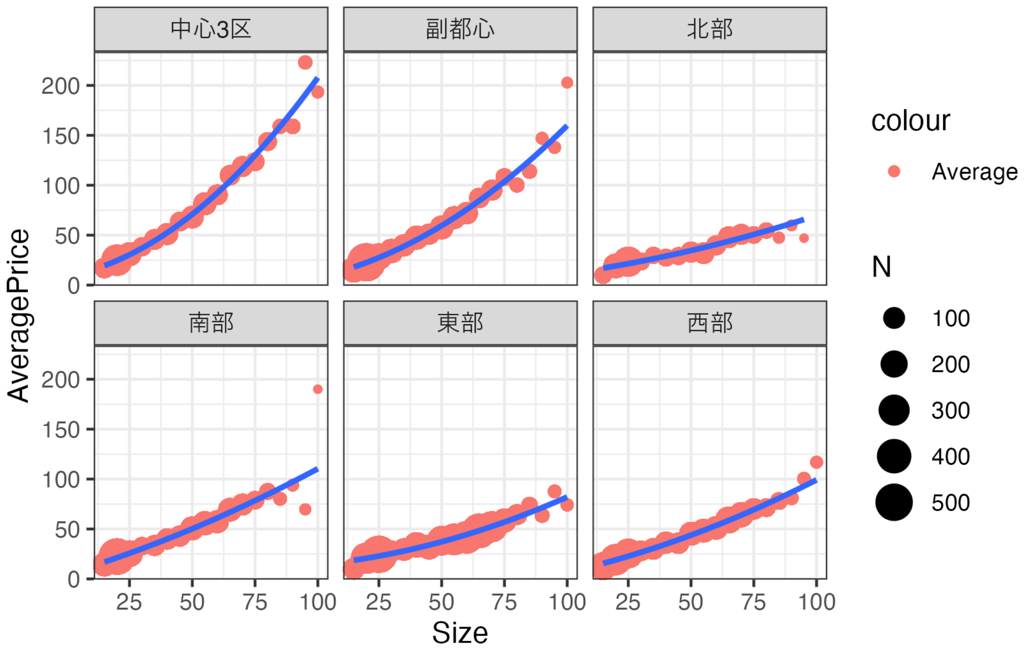
$$\beta_1 \times poly(Size, 15) \times Dummies(District) \tag{3}$$

- poly(x,l): x の l 乗まで作る

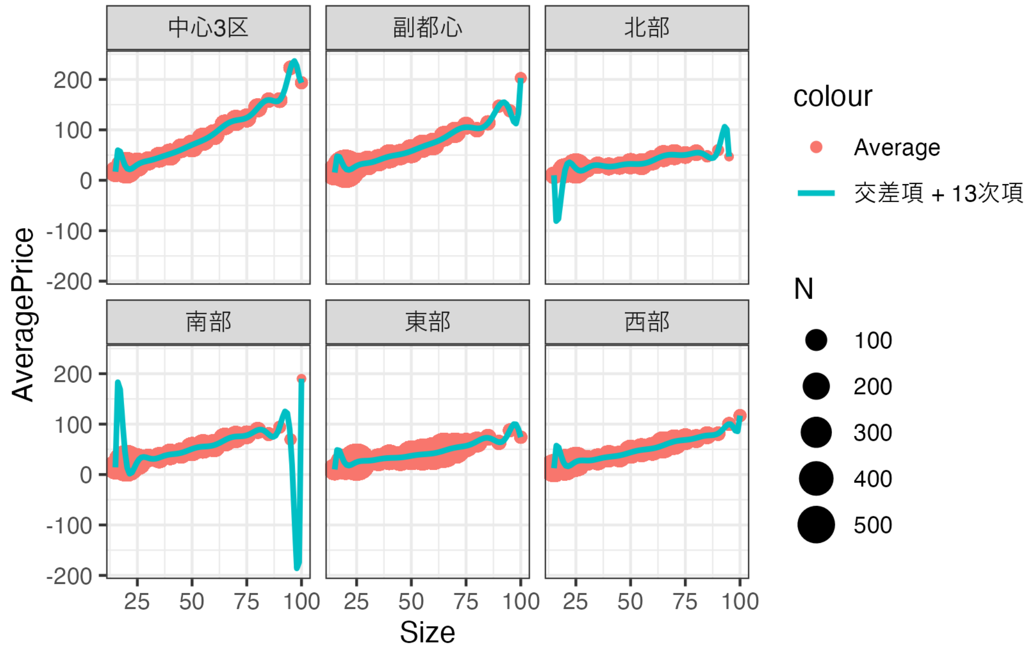
1.16 実例: 交差項 (Equation 1)



1.17 実例: 交差項 + 二乗 (Equation 2)



1.18 実例: 飽和モデル (Equation 3)



1.19 まとめ

- Y の (X 以外の要因による) 個人差が大きい応用においては、平均値のモデルであると解釈することが有効
- モデルの設定は研究者が行っていることに注意
  - 複雑なモデルを設定すれば、データと無矛盾なモデルが設定可能
    - 飽和モデル/丸暗記モデル (Learning by memoraization) とも呼ばれる

2 母集団への含意

- データ分析の目的は、「データの理解」ではなく、その背後にある社会の理解/予測
  - 統計学/機械学習の中核的だが達成困難な目標であり、“丁寧な議論” が必要

2.1 社会理解への含意

- 独立してデータ収集した研究者は、同じ社会を対象にしていたとしても異なる結論を得る
  - 例: 報道機関による世論調査

- 「自身で再現できるので合意する」ができない
  - 母集団を導入することで乗り越える

2.2 コンセプト: Estimand/Estimator

- 正答 (Estimand) と回答 (Estimator) を分離する
- Estimand(推定対象):
  - 実際には研究課題 (含む関心地域) に応じて、研究者が適切に設定
    - 理論上、すべての研究者が合意可能な答え
  - ただし誰も辿り着けない仮想的な答え
- Estimator(推定結果): データから算出される値

2.3 コンセプト: Population と Sampling

- Population(母集団): **全ての**事例が属する集団
  - “無限大” の事例数をもつ
    - 正確には、同時分布として定義される
  - Estimand が仮想的に定義される
- Sampling(サンプリング): 研究者は母集団の一部を収集する
  - ランダムサンプリング: 収集する事例は、ランダムに選ぶ
    - 事例間の独立・無相間 (IID) の仮定を正当化

2.4 合意

- 母集団を直接観察できれば、合意可能だが、
  - **母集団は直接観察できない**
    - “全数調査” であれば、Hyper-population を想定
- 推定値 (Estimator): 母集団からサンプリングされたデータから算出
  - 母集団について、**部分的な**情報をもつ
    - 一般に Estimand $\neq$ Estimator

2.5 例: 母平均

- 研究課題: 日本社会における賃金の特徴
- データ: 賃金構造基本統計調査
  - 一定数の従業員が所属する事業所をランダム抽出
- Estimand: 平均賃金 (と設定)
  - 観察できない
- Estimator: データから計算した平均賃金
  - $\neq$ 母集団における平均賃金

2.6 OLS の Estimand

- 代表的なものだけでも複数存在する
- **Y の母平均関数 $E[Y|X]$ の線形近似モデル (Best Linear Projection)**
  - 「誤定式化していない」などの強い仮定のもとで、さらに明確な解釈も有する

2.7 $E[Y|X]$ の線形近似モデル

- Estimand: **母集団上**で研究者が設定したモデル $g_Y(X)$ に OLS を適用した結果得られるモデル

$$g_Y^*(X) = \beta_0^* + \beta_1^* X_1 + ..$$

- 以下の Algorithm により、**仮想的に** 算出される

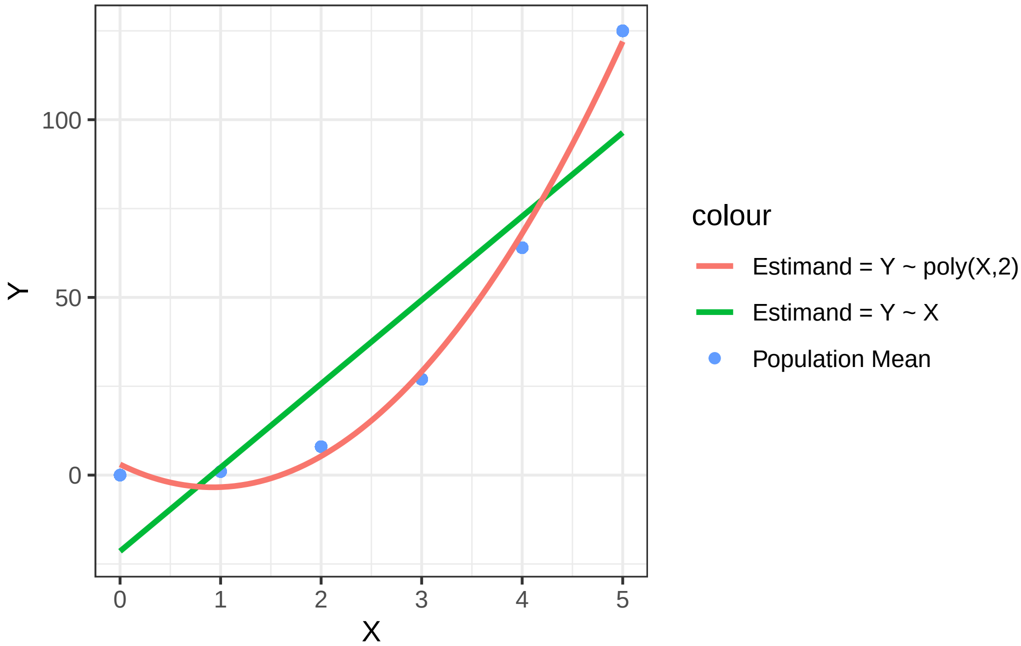
$$\min \int \Big(E[Y|X] - g_Y^*(X)\Big)^2 \times f(X)dX$$

  - $f(X)$: 母集団における属性 X を持つ事例の割合

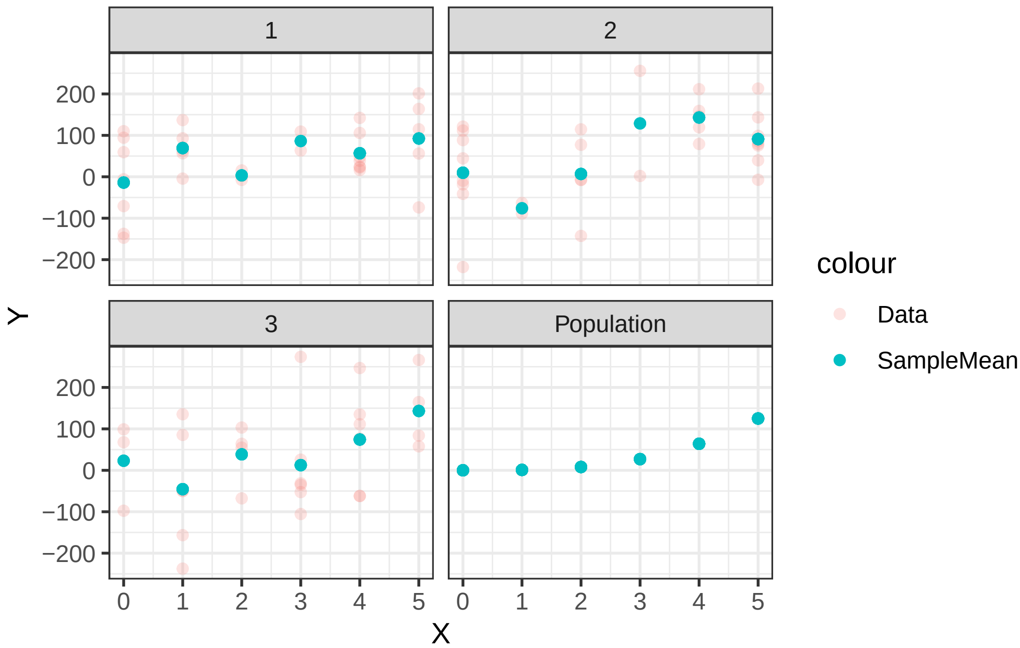
2.8 Estimator の性質

- **IID(ランダムサンプリング) であれば、**データ上で同じモデルに OLS を適用し得られるモデル $g_Y(X)$ は、$g_Y^*(X)$ の優れた Estimator

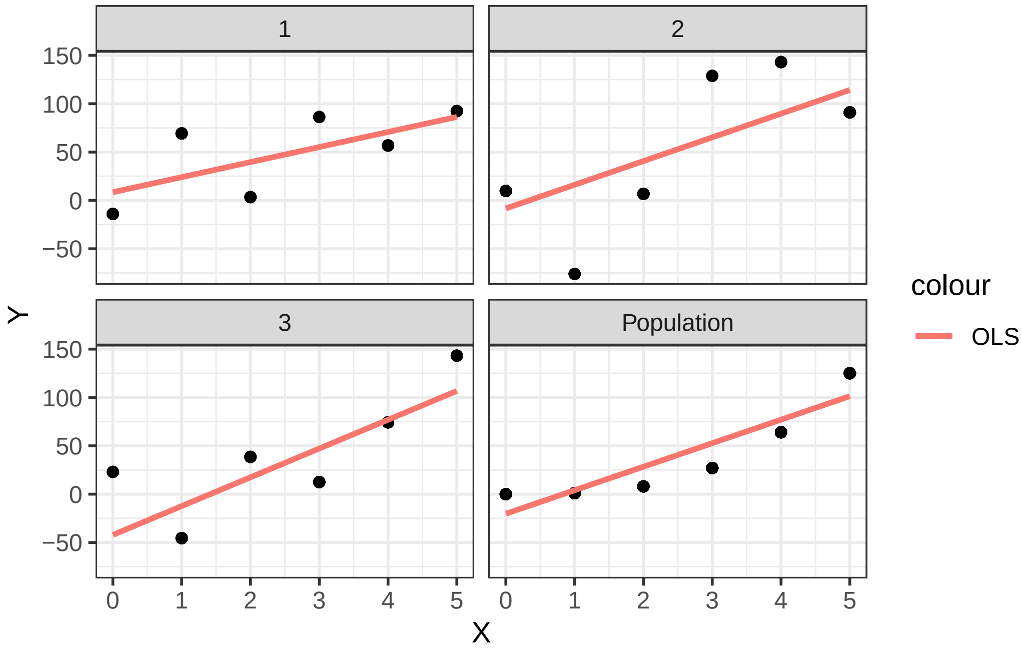
2.9 例: Population/Estimand



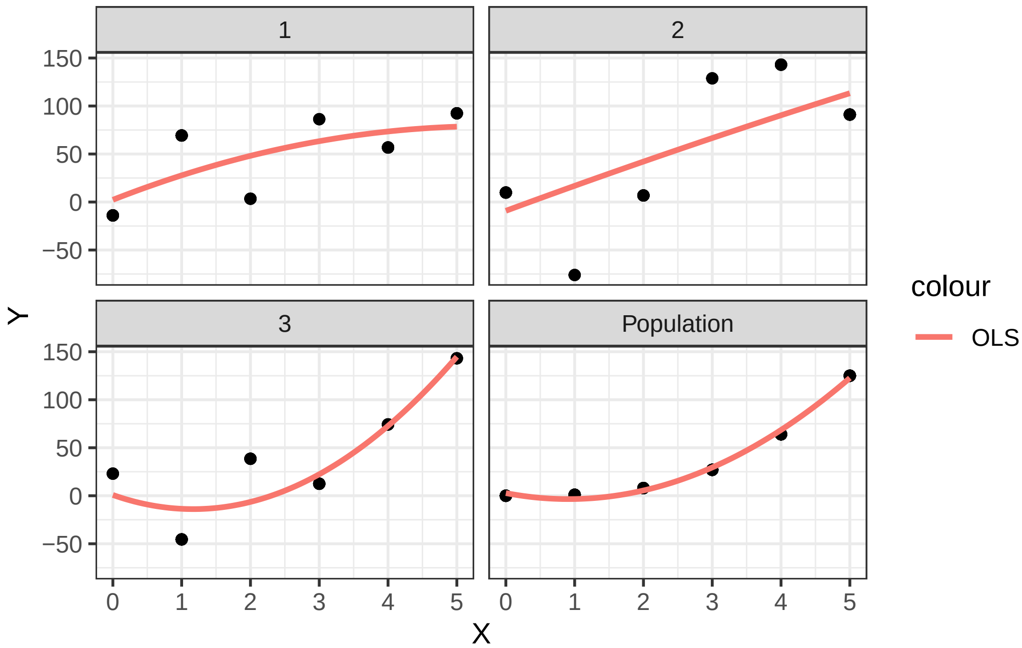
2.10 例: Data/SampleMean



2.11 例: 1 次



2.12 例: 2 次



2.13　OLS による推定結果の特徴

- IID であれば、**推定結果のバラツキ方**について、一定の規則性が生じる
- OLS が推定するモデルは
  - $\frac{\text{事例数}}{\text{パラメタの数}}$ が大きくなれば、Population OLS が推定するモデルに近い値を得られる
    - CausalML (Chap-1 pp-19) とその参考文献を参照
  - $\frac{\text{事例数}}{\text{パラメタの数}}$ が無限大になれば、Population OLS と一致する (一致推定量; Consistency)
- 詳細は Section 4

3　母平均の正確なモデル

- 伝統的な教科書では、しばしば、OLS は $Y-X$ の母集団における" 真の" 確率的関係性を理解する手法として" 説明される"
  - 例えば、$g_Y(X)$ を 母平均関数 $E[Y|X]$ の良い推定結果であるためには、IID 以上の仮定が必要

3.1　Mis-specification

- データ上での OLS により得られる $g(X)$ を、母平均 $E[Y|X]$ の良い推定結果であるためには、母平均について Mis-specification がないことを仮定する必要がある
- パラメタ β_0.. を**どのように**選んでも、

  $$E[Y|X] = g(X)(= \beta_0 + \beta_1 X_1 + ..)$$

  は達成できない

3.2　母平均の推定

- Mis-specification がないと仮定できれば、$g^*(X) = E[Y|X]$
  - データ上での OLS により得られる $g(X)$ は、$g^*(X)$ の一致推定量なので、$E[Y|X]$ の一致推定量でもある

3.3　母平均の推定の難しさ

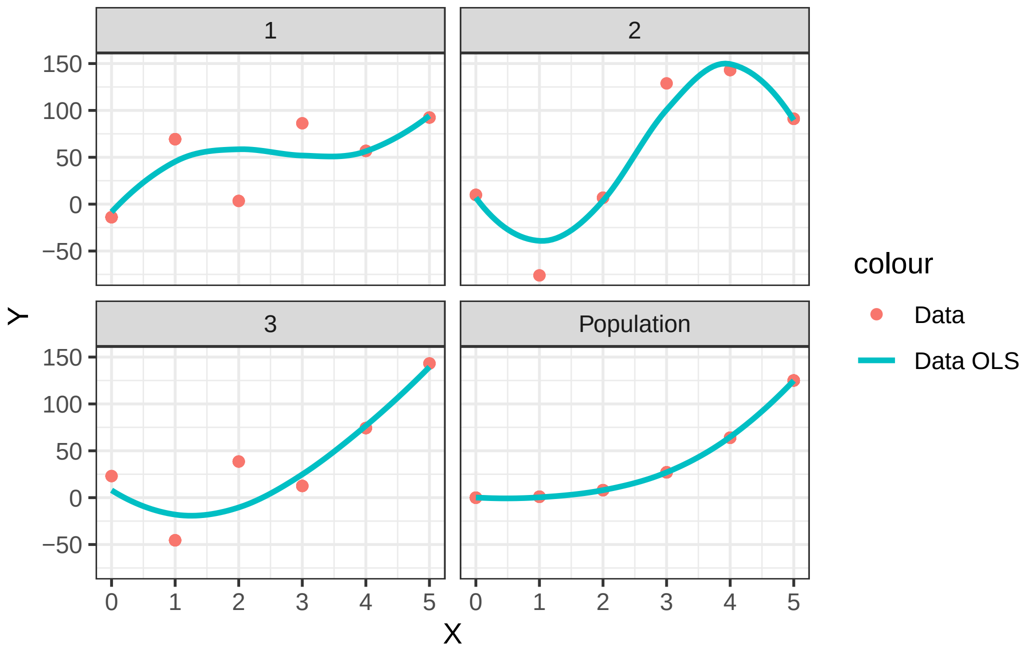
- 実践においては $E[Y|X]$ は非常に複雑な形状をしていることが予想される
  - 大量のパラメタを導入しないと、顕著な Mis-specification が発生する可能性がある

- – 大量のパラメタを導入すると、推定精度が悪化する
- 事例数とパラメタの数の競争となる
  - – 大量の事例があれば、過剰にパラメタを投入しても OK

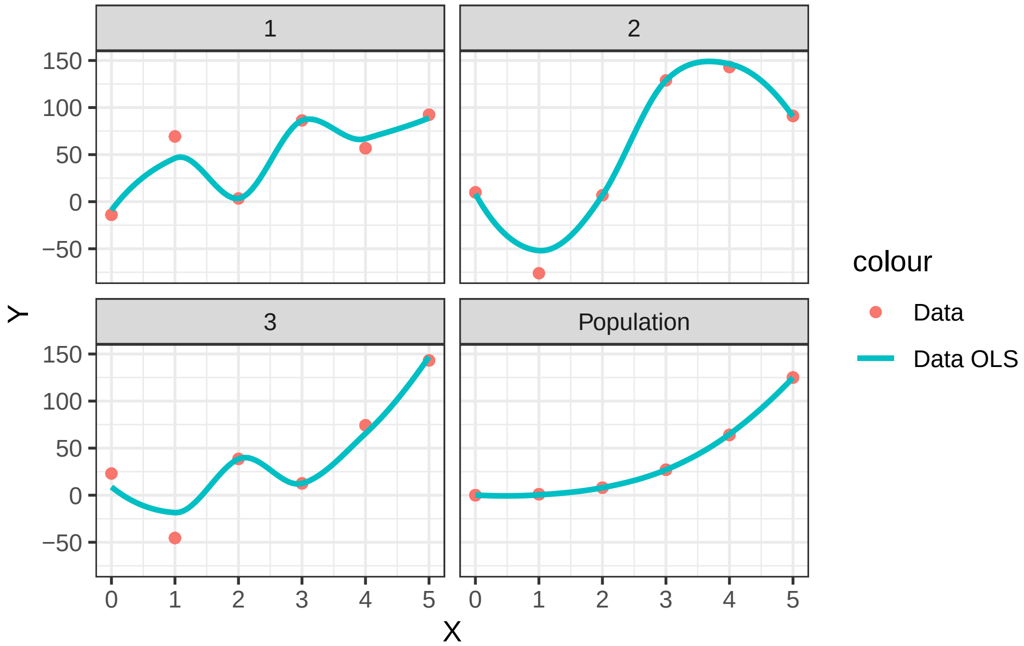
3.4 例:

- 真の母平均関数 $E[Y|X] = X^3$
- 真の母平均関数を推定できない
  - – $Y \sim X$ または $Y \sim X + X^2$
- 真の母平均関数を推定できる
  - – $Y \sim X + X^2 + X^3 + ..$
  - – ただし事例数が十分にないと、推定誤差が大きい

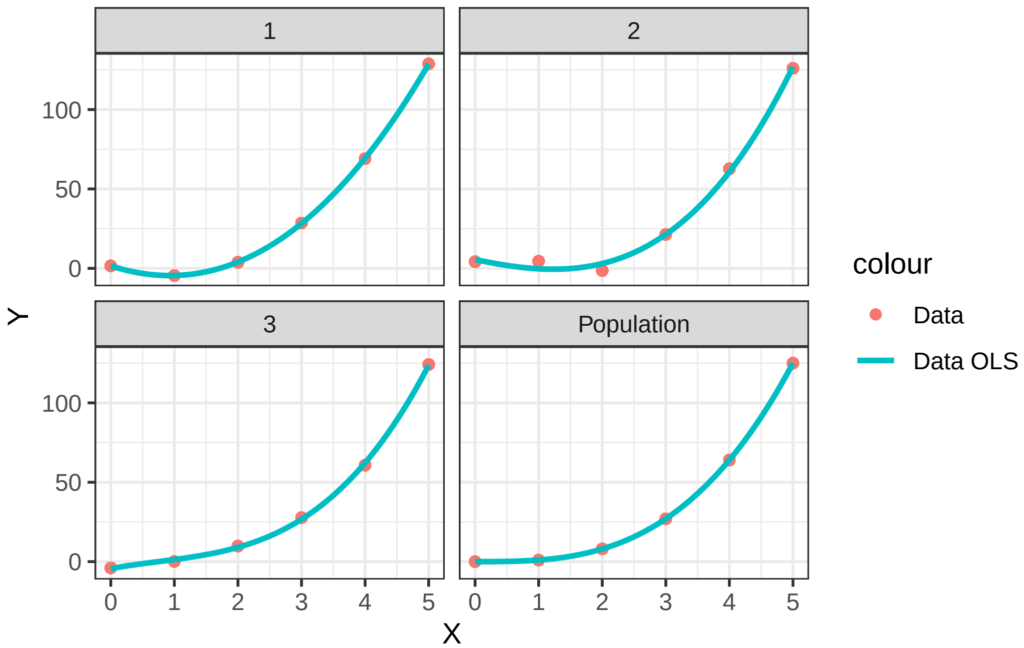
3.5 例: 3 次 $\&N = 30$



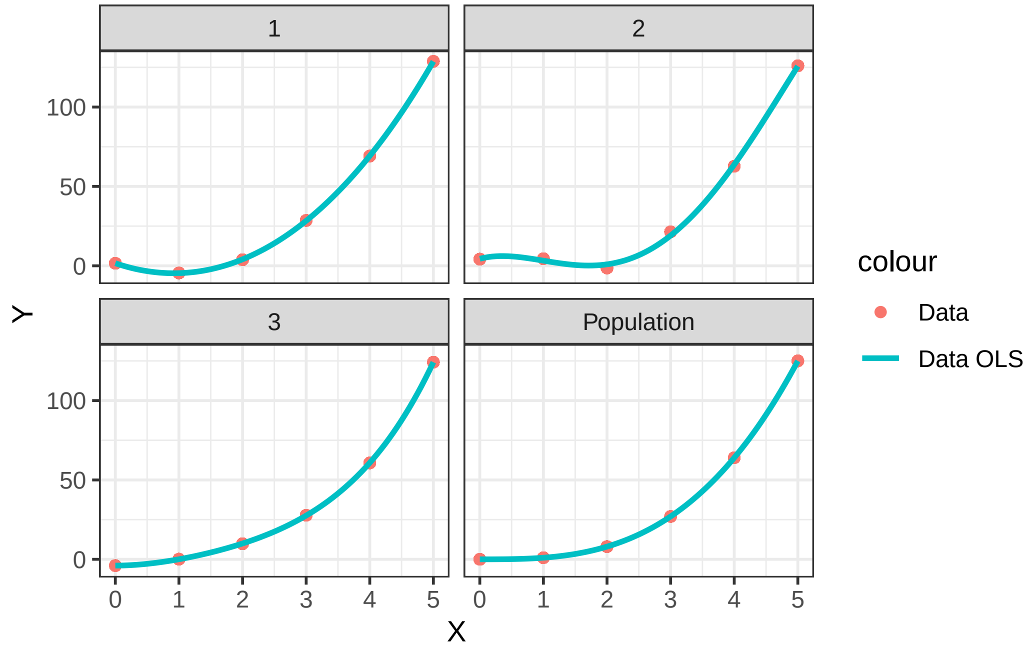
3.6 例: 5 次 $\& N = 30$



3.7 例: 3 次 $\& N = 5000$



3.8 例: 5 次 $\&N = 5000$



3.9 まとめ

- 最低限の仮定から始める議論を紹介
  - IID(ランダムサンプリング) であれば、**少なくとも**、Population OLS の結果についての推定であると解釈できる
  - モデルが正しければ、母平均関数の推定結果であると解釈できる
- Angrist and Pischke (2009), Aronow and Miller (2019) などで採用
- 多くの入門書では、より強い仮定から議論をスタート Section 5

4 補論: Sampling Distribution

4.1 コンセプト: 信頼区間

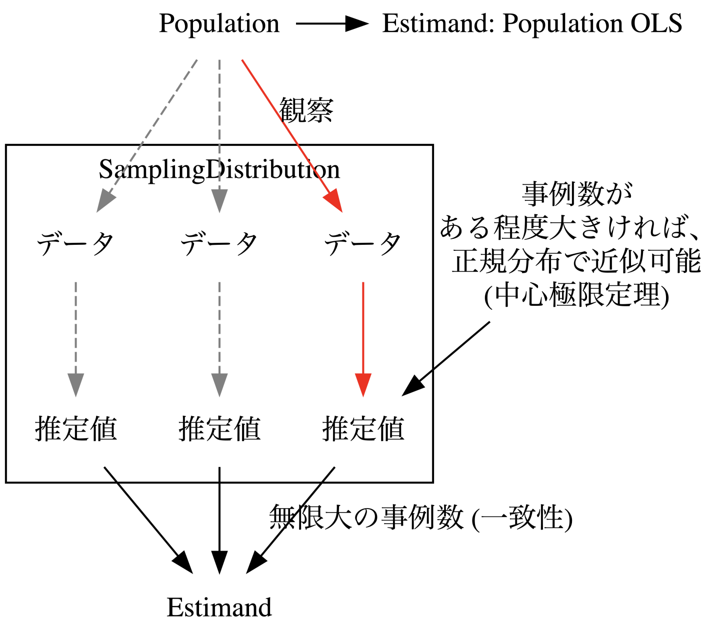
- 一般に Estimator $\neq$ Estimand
  - 少なくとも無限大の事例数が必要
    - 全ての研究者が" 間違った" 結果を得ている

- 代替的に信頼区間を計算
  - 95 % の研究者は、Estimand を含んだ区間を得られる
- Sampling Distribution が漸近的に正規分布で近似可能であることを活用

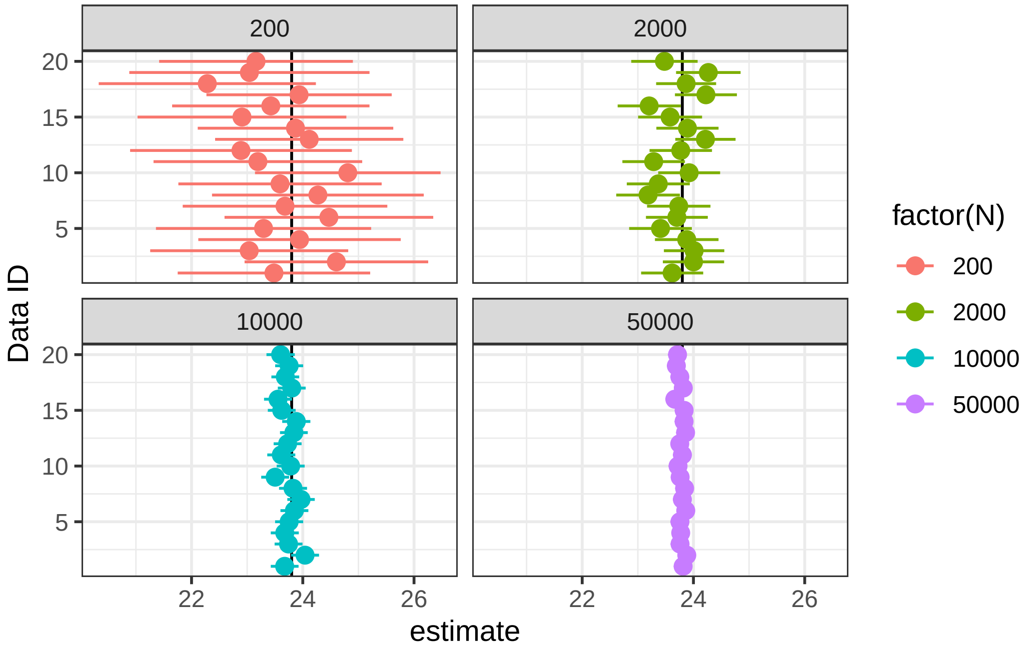
4.2 コンセプト: Sampling Distribution

- データから計算される推定結果の**仮想的な**分布
  - "独立した大量の研究者をイメージし"、それぞれが得られる推定結果の分布を想像する
- ポイント: 統計的性質の多くは、個別の推定結果ではなく、推定結果の Sampling Distribution の性質であることに注意

4.3 例: IID(+ 技術的な仮定) が導く OLS Estimator の性質



4.4 例: 信頼区間



5 補論: 伝統的な議論

- 伝統的な入門書 (の最初の方の章) における議論と比較する
  - Wooldridge (2016), Stock and Watson (2020) など
    - 「正しい確率モデルのパラメタを推定する問題」に落とし込まれることが多い

5.1 確率モデル

- Y の正しい確率モデルを想定する:

$$Y = \underbrace{g(X)}_{\beta_0+\beta_1 X_1+..} + \underbrace{u}_{\text{誤差項}=Y-g(X)}$$

- 恒等式として成り立つ
- 誤差項の分布に "仮定" を追加することで β を推定する

5.2 $E[u \times X] = 0$

- $g(X)^*$ は Population OLS の結果であれば、$E[u \times X] = 0$ は成り立つ

- 多くの実践で、β を推定するために用いられているモーメント条件

5.3 $E[u|X] = 0$

- 母平均について Mis-specification がなければ、$E[u|X]$ は成り立つ
- 分散均一: 誤差項 u の分散が、X に対して一定
  - $E[u^2|X] = E[u^2]$
  - OLS は最善の不偏推定量を提供
    - ガウス・マルコフの定理 (wiki)
  - 古典的な方法で標準誤差 (信頼区間) を計算可能
- 多くの応用で非現実的と判断し、本講義では不採用

5.4 $u \sim N(0, \sigma)$

- 仮定: 誤差項 u が正規分布に従う (古典的回帰モデル)
  - Estimand $f(Y|X)$ (Y の条件付き分布) の優れた Estimator (最尤法の推定値と一致)
  - 推定結果は、有限の事例数の元で、正規分布に従う
- より非現実的な仮定

Reference

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Aronow, Peter M, and Benjamin T Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.

Stock, James H, and Mark W Watson. 2020. *Introduction to Econometrics*. Pearson.

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics*. Cengage AU.