部分線形モデル

Semiparametric 推定への応用

川田恵介

Table of contents

Semiparametric 推定への応用

- Partial out 推定が持つ統計的性質を理解するために以下を” 復習”
  - 何を目標とするのか? (Well-Specified Model について成り立つ漸近性質)
  - 何について議論しているのか？ (収束)

- Partialling out　推定の持つ、教師付き学習の収束の遅さを補う性質を紹介
  - SeminalPaper の一つは、Chernozhukov et al. (2018)

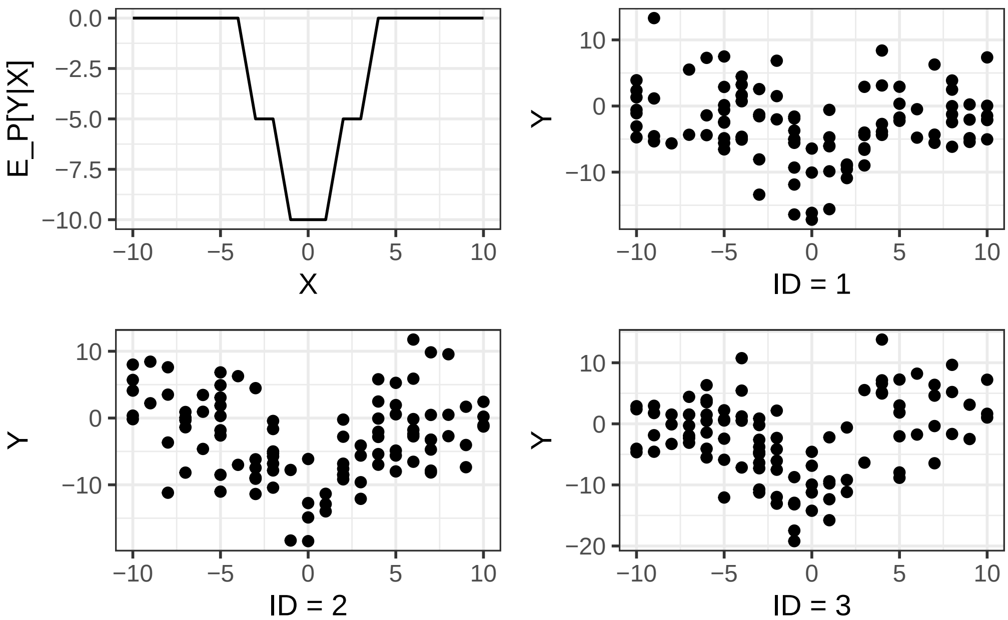
準備: 収束

- 大標本性質: 事例数が大きくなると近似的に成り立つ性質
  - “収束の速度” をざっくり理解
  - 詳しく知りたい人は、大学院レベルの計量分析のテキストなどを参照
- 基本アイディア: “いいレースになる” 単純な数式で記述

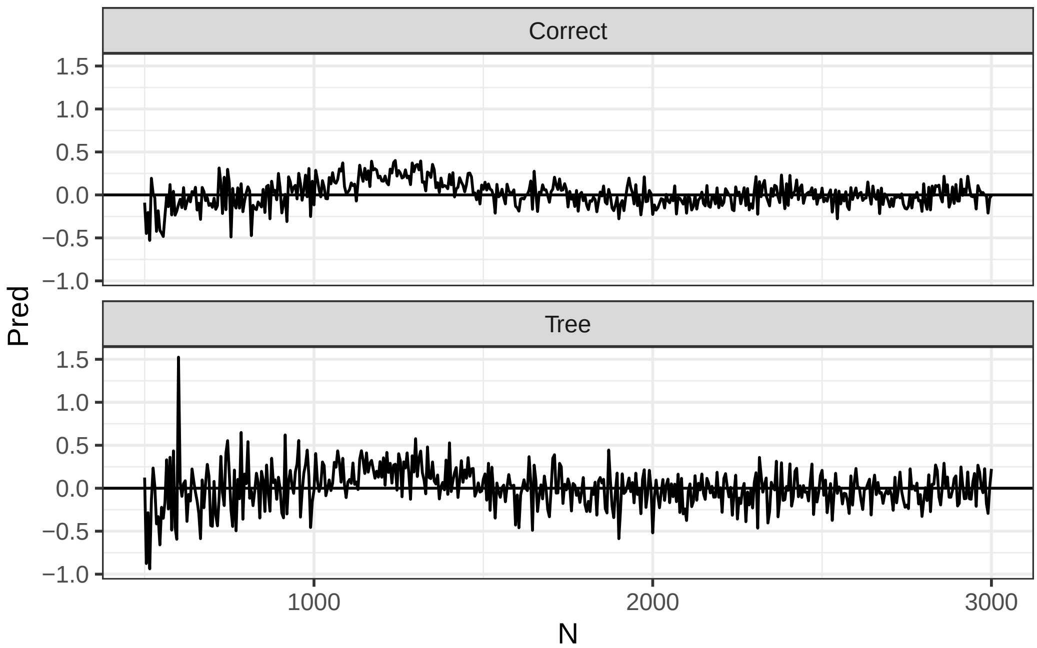
数値例

- $E[Y|X]$
  - $X \in [-10, 10]$ の**整数**
- Well-specified model: $Y \sim factor(X)$
- Prune Tree: 剪定をした決定木
- ともに一致推定量だが、
  - Well-specified model の方が、収束速度は同等以上

数値例



$X = 4$ についての比較



収束速度

- 収束速度の上限をシンプルに記述する
  - シンプルな" 式" とのスピード勝負
- 発展: ランダウ記法 $\mathcal{O}(n^\alpha)$ and $o(n^\alpha)$

Well Specified Model

- $\beta = g(X=4)$ が母集団における値 $\beta^P = E_P[Y|X=4]$ に収束するかどうか
  - OLS で推定すると $\beta = E[Y_i|X_i=4] := \sum_i Y_i/N|X_i=4$
- 一致性を満たすので、$\lim_{N\to\infty}(\beta - \beta^P) \to 0$
- 総和 $N \times \beta := \sum_i Y_i$ は発散する

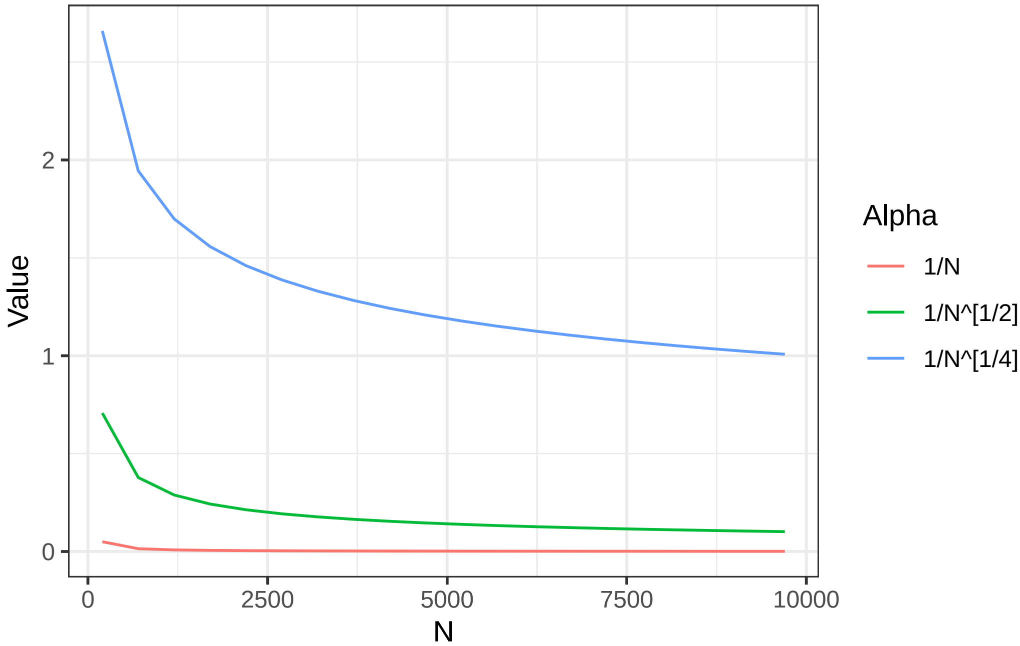
$\sqrt{N}$ CAN estimator

- $\lim_{N\to\infty} N^\alpha \times (\beta - \beta^P) = ?$
- 両極端な性質
  - $\lim_{N\to\infty} N^{\alpha=0} \times (\beta - \beta^P) \to 0$
  - $\lim_{N\to\infty} N^{\alpha=1} \times (\beta - \beta^P) \to \infty$
- 収束も発散もしない、"ちょうどいい" α があるのでは?
  - Yes !!!
  - $\alpha = 1/2$!!!

イメージ

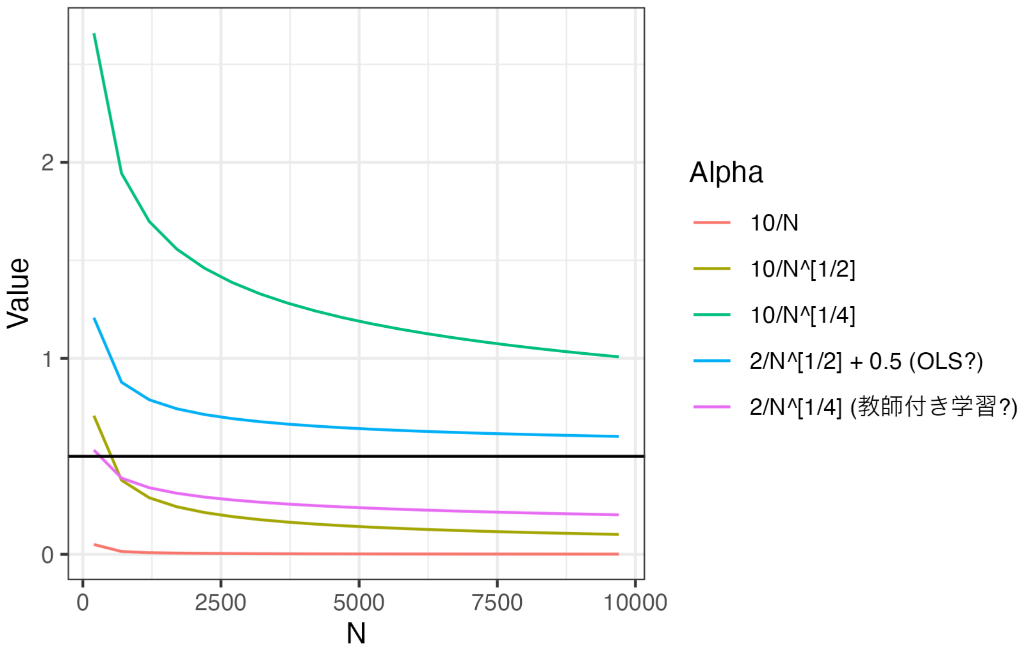
- $N^\alpha(\beta - \beta^P) = \frac{\beta-\beta^P}{1/N^\alpha}$
  - 分子分母ともに 0 に向かって収束する
  - $\alpha = 1/2$ について、分子の収束スピードは分母と同等
- 多くの教師付き学習 (Nonparametric 推定) で得られる予測値 $g(X)$ は、$E_P[Y|X]$ に収束するが、
  - $\alpha = 1/2$ に収束速度で負ける!!!

例



Value
N
Alpha
1/N
1/N^[1/2]
1/N^[1/4]

例



Value
N
Alpha
10/N
10/N^[1/2]
10/N^[1/4]
2/N^[1/2] + 0.5 (OLS?)
2/N^[1/4] (教師付き学習?)

まとめ

- OLS は、$1/N^{1/2}$ と同等の収束速度
  - ただし、Misspecified であれば、BLP に収束
- 代表的な教師付き学習の推定値は、$1/N^{1/2}$ に収束速度で負ける
  - “収束が遅い”
- Partialling Out: $1/N^{1/4}$ には勝つ教師付き学習を前提

準備: 大表本理論に基づく推論

- サンラムサンプリング (IID) の仮定が持つ、母集団への含意は？
  - 母集団への詳細な仮定なしで何が言える?

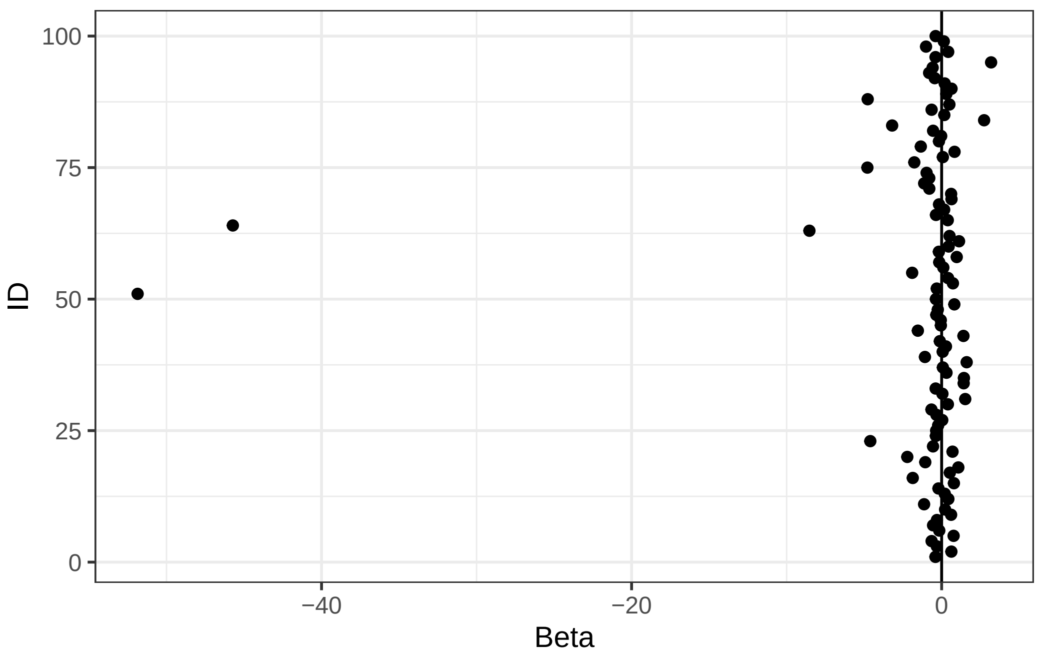
Well-specified model

- “入門教科書” 的な問題設定
- $g(D,X)=\beta_0+\beta_D D+\beta_1 X_1+..+\beta_L X_L$
- $E_P[Y|D,X]=g(D,X)$ を達成する β^P が存在

OLS 推定

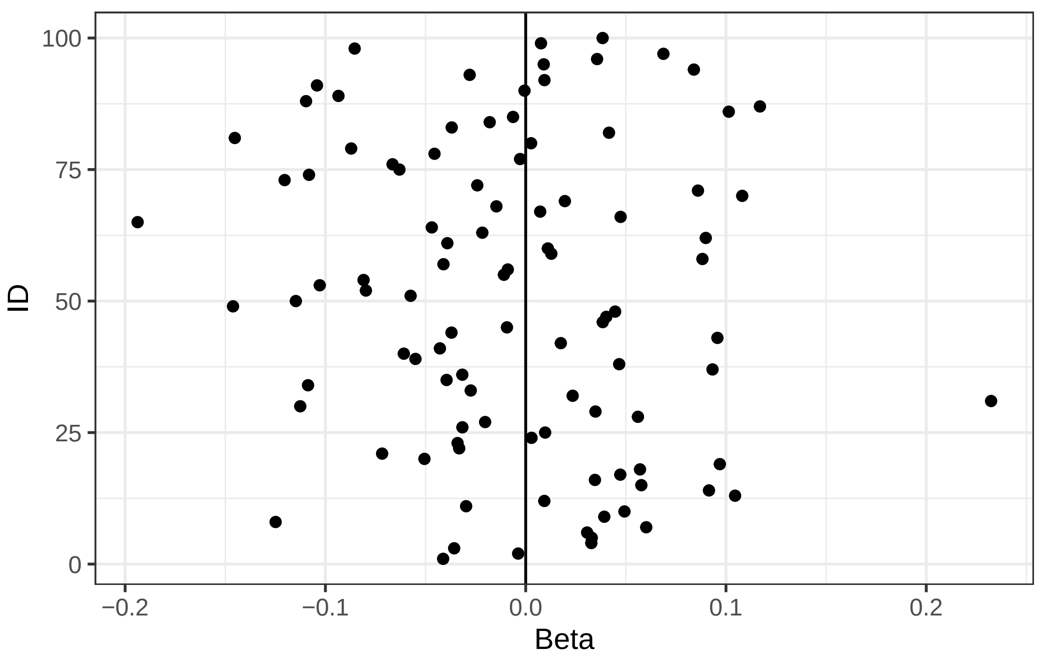
- $Y\sim X$ を Empirical Risk を最小にするように推定
  - 推定値のばらつき (Sampling Uncertainly) の源泉は？
  - Y,X のデータ上の分布の (研究者間での) 違い

N = 10

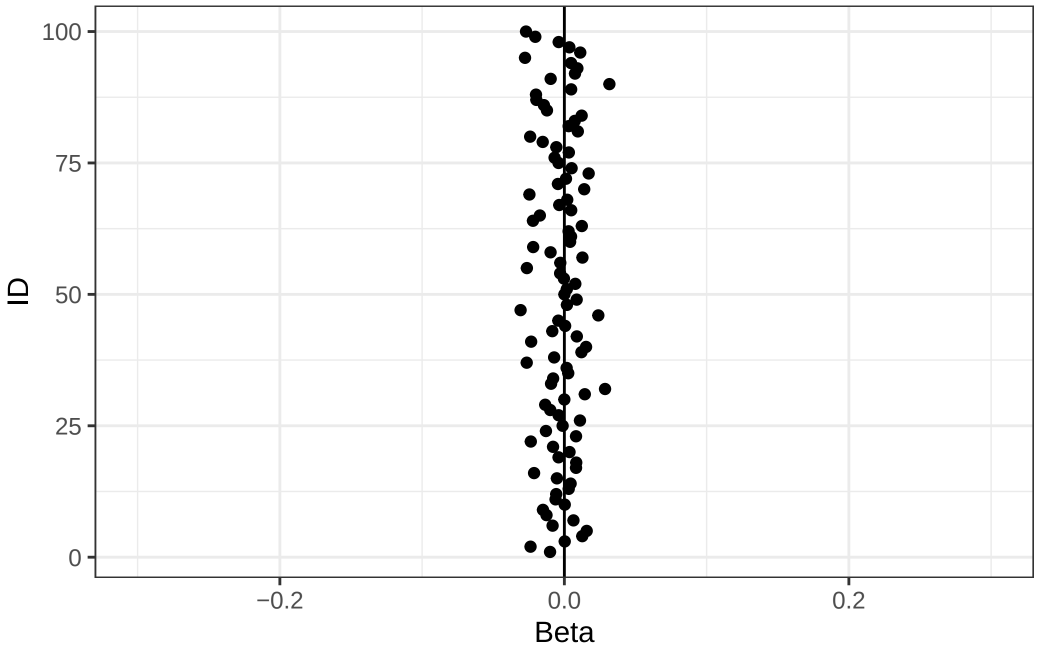


- 100 名研究者が独立して研究 (事例数 = 10)

N = 200



N = 5000



分解

- β_∞：サンプルサイズ無限大の元での推定値, β：推定値

$$\underbrace{\beta^P}_{真の値} - \beta = \underbrace{\beta^P - \beta_\infty}_{=0\ \ Consistency}$$

$$+ \underbrace{\beta_\infty - E_P[\beta]}_{=0\ :\ Bias}$$

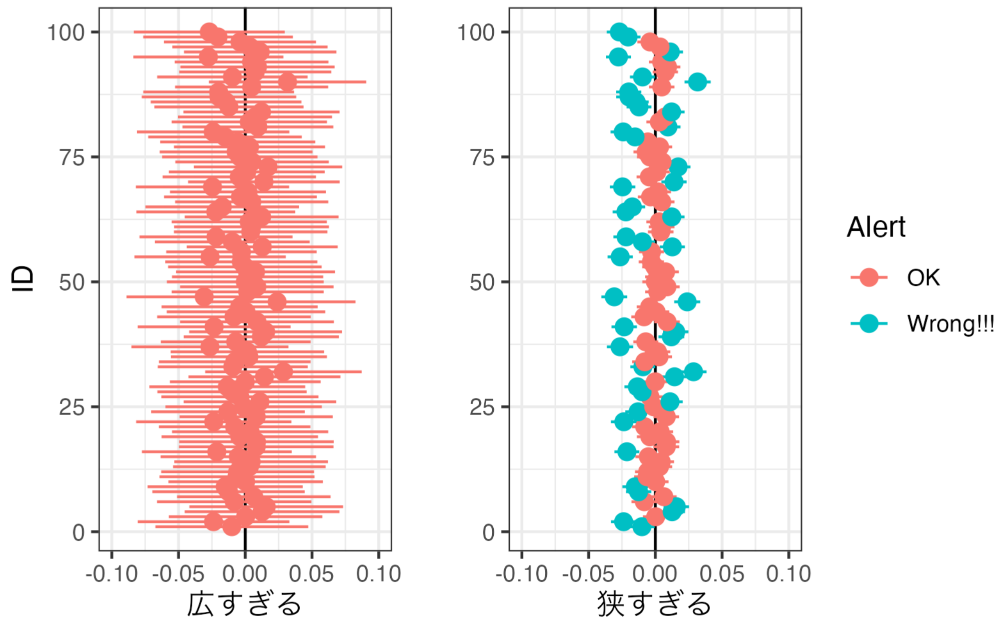
$$\underbrace{E_P[\beta] - \beta}_{\neq 0}$$

信頼区間

- 「推定値 = 真の値」を前提に議論を始めると、“100%” 間違う
  - 独立した研究者間での合意も不可能
- ハードルを下げる
  - 大多数 (典型的には 95%) の研究者について、真の値を含む区間 (信頼区間) を計算する

- 注意: 信頼区間自体は、独立した研究間で異なる
  - 合意可能なのは、「多くの研究者について、信頼区間は真の値を含む」のみ

不適切な区間



漸近正規性の活用

- 信頼区間を計算するには、推定値の分布 (研究者間の散らばり具合) への仮定が必要
  - 本当の分布は、母分布に依存
- 母分布に直接仮定を置くアプローチ: 教科書的な最尤法、ベイズ
- 近似性質を仮定するアプローチ: サンプリング方法への仮定 (ランダムサンプリング)“のみ” に基づいて導出される、漸近性質 (サンプルサイズがある程度大きければ、近似的になりたつ性質) を活用

漸近正規性

- 中心極限定理 + 不偏性から導ける漸近正規性を活用

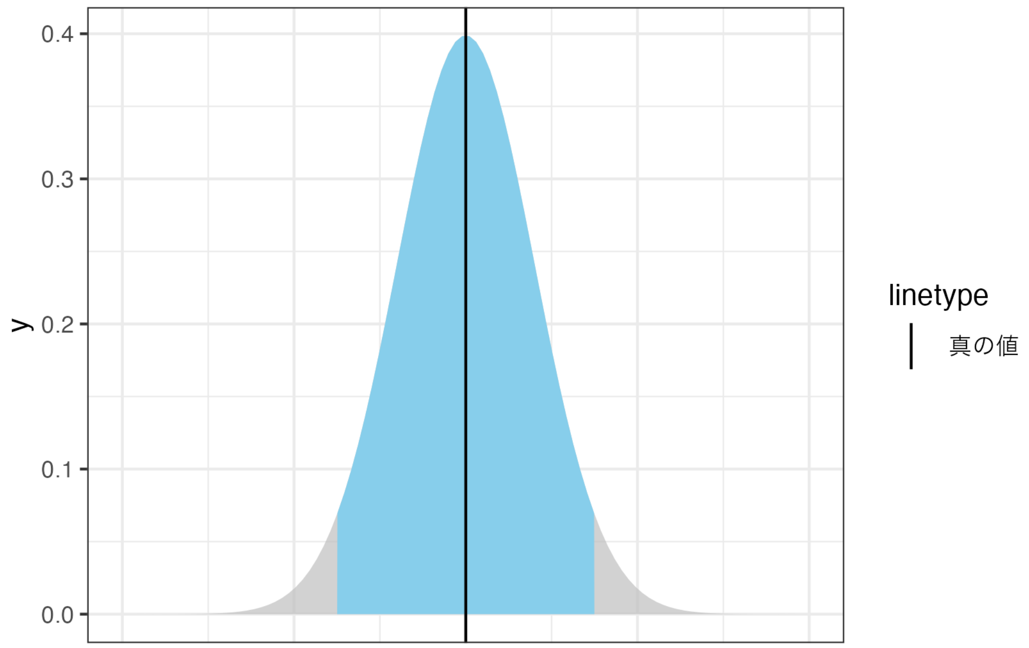
$$\lim_{N\to\infty} \sqrt{N} \times (\beta - \beta^P) \sim \mathcal{N}$$

- サンプルサイズが大きければ、正規分布で近似できる

– 真の値よりも、“早め” に収束する

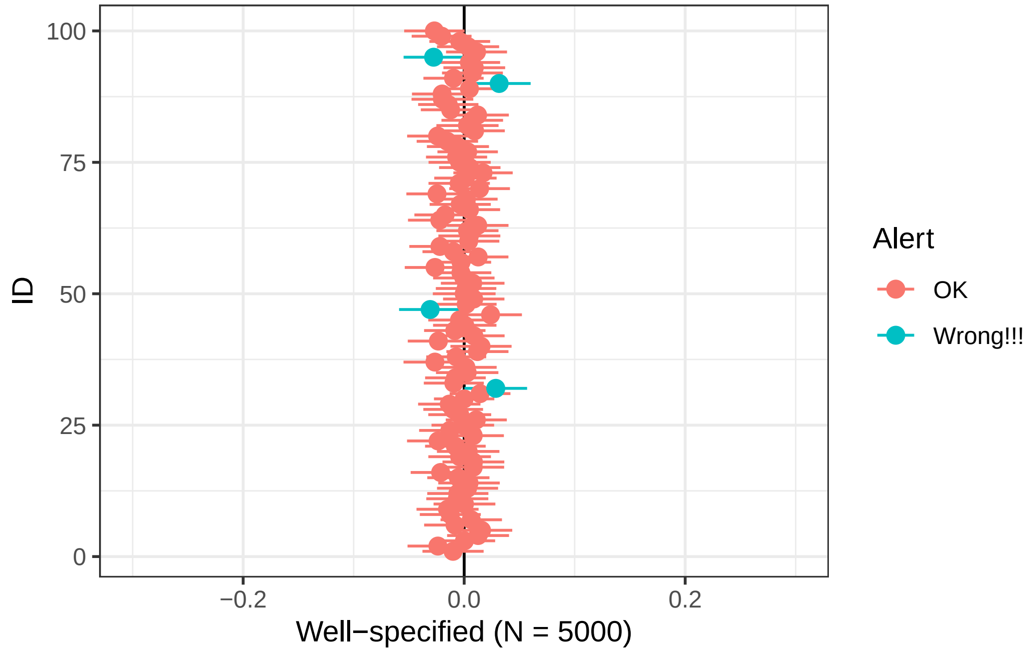
- $\sqrt{N}$ CAN estimator

漸近正規性

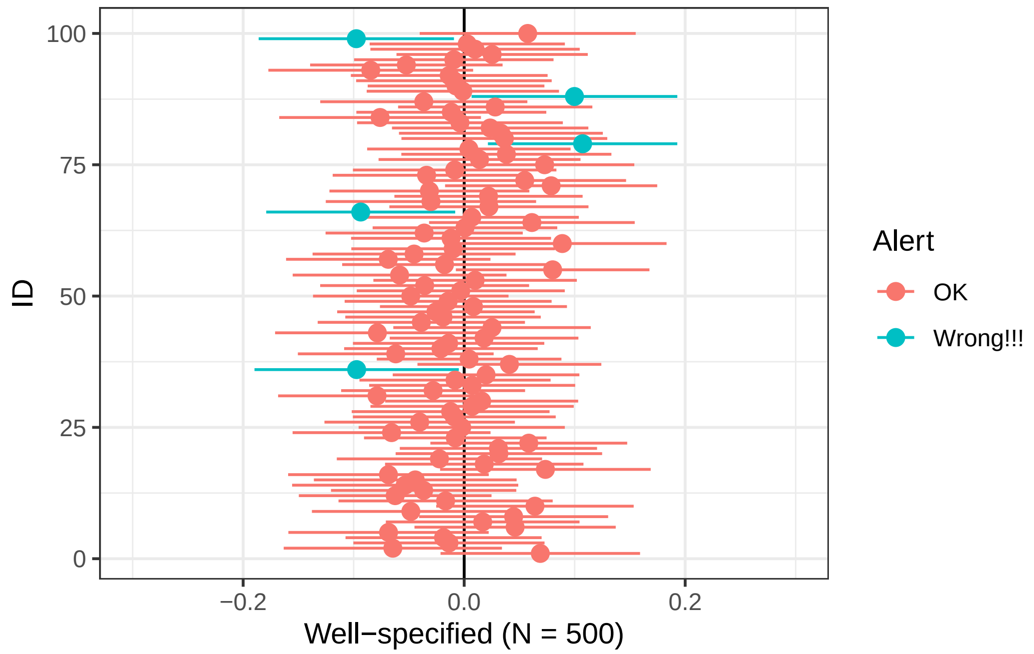


- 注意: 真の値からの” 距離” だけわかる

95% 信頼区間



サンプルサイズの影響



Misspecified model

- $g(D,X)=\beta_0+\beta_D D+\beta_1 X_1+..+\beta_L X_L$
  - β を推定
- β をどう選んでも、$E_P[Y|D,X]\neq g(D,X)$ (Approximation Error)
  - OLS で推定すると **BLP** β^{BLP} についての CAN estimator
  - BLP について信頼区間を提供

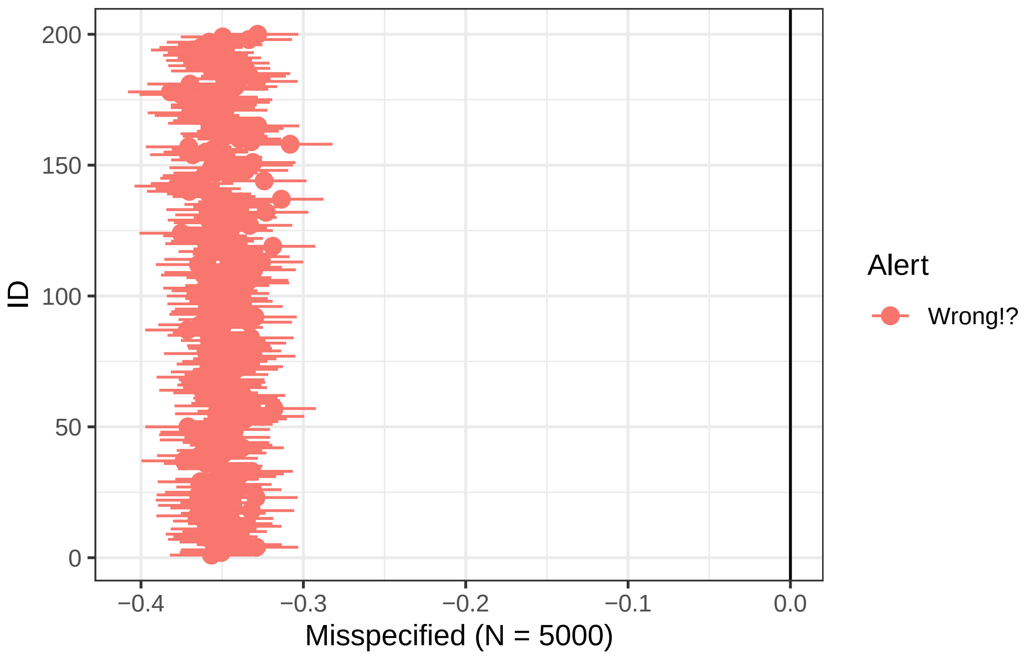
中間まとめ

$$\beta^{BLP}-\beta=\underbrace{\beta_{BLP}(X)-\beta_{\infty}(X)}_{=0}$$

$$+\underbrace{\beta_{\infty}(X)-E[\beta]}_{=0}$$

$$\underbrace{E[\beta]-\beta}_{\sim\mathcal{N}}$$

95% 信頼区間



典型的教師付き学習の応用

- $\beta^P = E_P[E_P[Y|D=1,X] - E_P[Y|D=0,X]]$ を推定できるか？
- 一致性を持つ予測モデル $g(D,X)$ を推定
  - $\beta = \sum_i [g(D+1,X_i) - g(D,X_i)]$

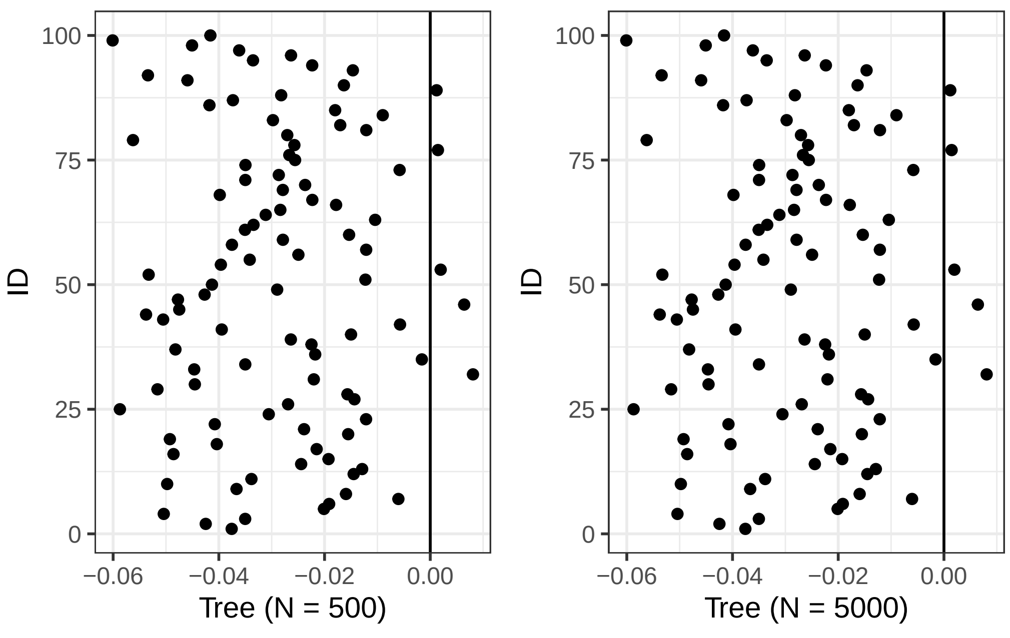
典型的教師付き学習の応用

$$\beta^P - \beta = \underbrace{\beta^P - \beta_\infty}_{\simeq 0}$$

$$+ \underbrace{\underbrace{\beta_\infty - E_P[\beta]}_{\neq 0} + \underbrace{E_P[\beta] - \beta}_{\neq 0}}_{?}$$

- **一般に**、母集団の明確な特徴について、CAN estimator にならない
  - 何を推定している？

Tree



Partialling Out

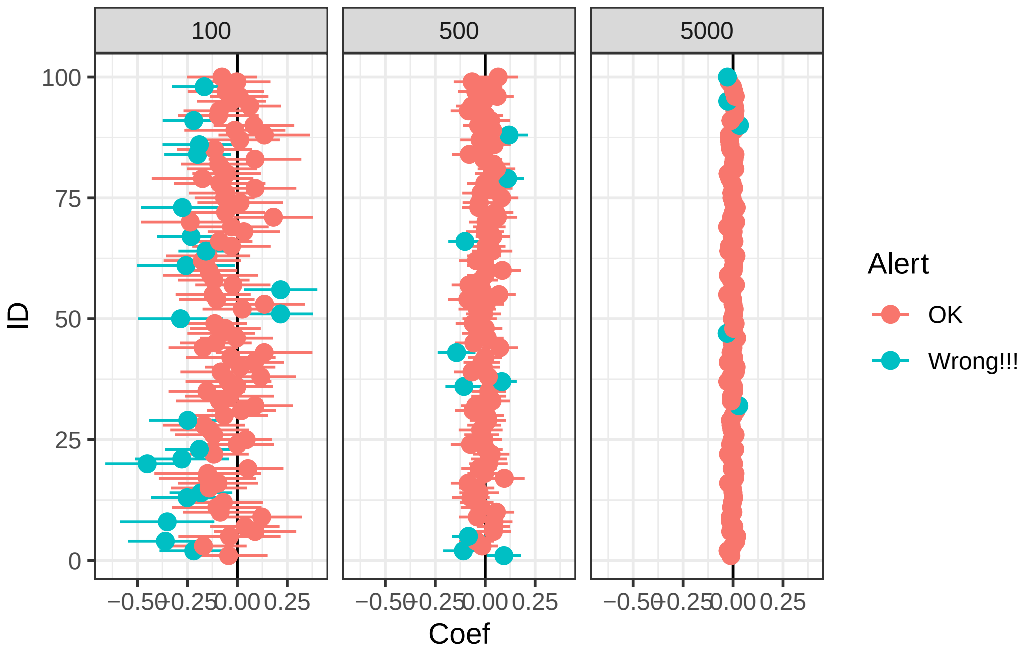
- 目指すは

$$\beta^P - \beta = \underbrace{\beta^P - \beta_\infty}_{\simeq 0}$$

$$+ \underbrace{\beta_\infty - E_P[\beta]}_{\simeq 0}$$

$$\underbrace{E_P[\beta] - \beta}_{\sim \mathcal{N}}$$

Partialling Out with Tree



まとめ

- CAN estimator であれば、近似的な信頼区間を用いて、母集団の性質について議論できる
- OLS は、**BLP** についての信頼区間を形成できるが、一般に周辺化された平均差についてではない
  - 特定の変数間の関係性を捉えるのには向いていない
- 教師付き学習の Naive な応用は、"何を推定しているのかわからない"

Partialling Out 推定量の特徴

- Partialling Out 推定量はどのような大表本性質を持つのか？
  - 緩やかな条件 (機械学習の推定値が $1/n^{1/4}$ よりも早い速度で、条件つき平均値に収束) の下で
  - 単回帰と同じ性質 ($\sqrt{N}$ CAN) !!!!

Partialling Out

- 議論を見やすくするために、データを 2 分割 (Auxiliary/Estimation データ) し、交差推定しない

1. Auxiliary データを用いて、$g_Y(X) \sim E_P[Y|X], g_D(X) \sim E_P[D|X]$ を推定
2. Estimation データを用いて、$Y - g_Y(X) \sim D - g_D(X)$ を OLS 推定し、β を推定

ばらつきの源泉は?

1. Estimation データにおける Y, D の分布
2. Auxiliary データから推定される予測関数 $g_Y(X), g_D(X)$

ポイント

- 一般に、$g_Y(X), g_D(X)$ の収束が遅いので、β の収束速度は $1/n^{1/2}$ に負けるが、、、
- $g_Y(X), g_D(X)$ が $1/n^{1/4}$ よりも早ければ、2 番目の源泉は漸近的に無視できる!!!
  - $g_Y(X), g_D(X)$ が" 推定されている" という事実を漸近的に無視できる!!!

Oracle 推定値

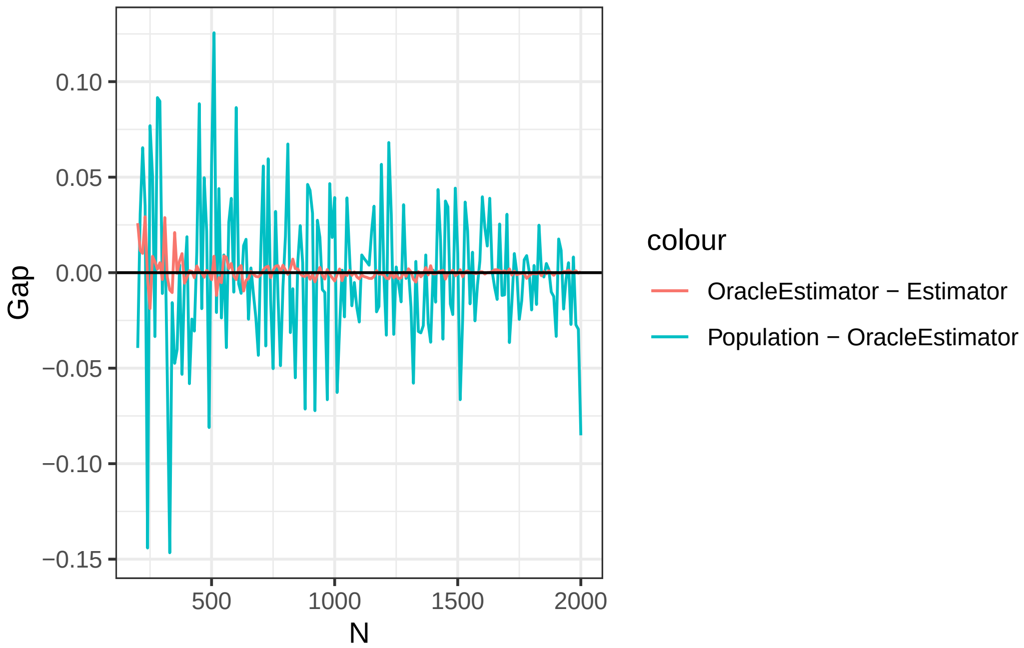
- 理解の助けとして、仮想的な推定値 (Oracle Estimator β^{Oracle}) を導入
- $Y - E_P[Y|X] \sim D - E_P[D|X]$ を OLS した結果得られる係数値
  - E_P は非確率変数なので、既知のルールで変換した変数同士を回帰しているだけ
  - ばらつきの厳選は、Estimation データにおける Y, D の分布のみ
  - $\sqrt{N}$ CAN estimator

分解

$$\beta^P - \beta = \underbrace{\beta^P - \beta^{Oralce}}_{\sim \mathcal{N}}$$

$$+ \underbrace{\beta^{Orable} - \beta}_{\rightarrow 0 \ !!!}$$

Estimation Error



例題: 母分散の推定

- 性質を理解しやすくするために、よりシンプルな状況を考える
- Estimand (母分散): $E_P[(Y - E_P[Y])^2]$
- Estimator $E[(Y - g)^2]$
  - g は Auxiliary データから推定

書き換え

$$E[(Y-g)^2] = E[(\underbrace{Y - E_P[Y]}_{IrreducibleError} + \underbrace{E_P[Y] - g}_{ReducibleError})^2]$$

書き換え

$$E[(Y-g)^2] = \underbrace{E[(Y - E_P[Y])^2]}_{Oracle}$$

$$+2 \underbrace{\times E[(E_P[Y] - g) \times (Y - E_P[Y])]}_{Reducible \times Irreducible}$$

$$+ \underbrace{E[(E_P[Y] - g) \times (E_P[Y] - g)]}_{Reducible \times Reducible}$$

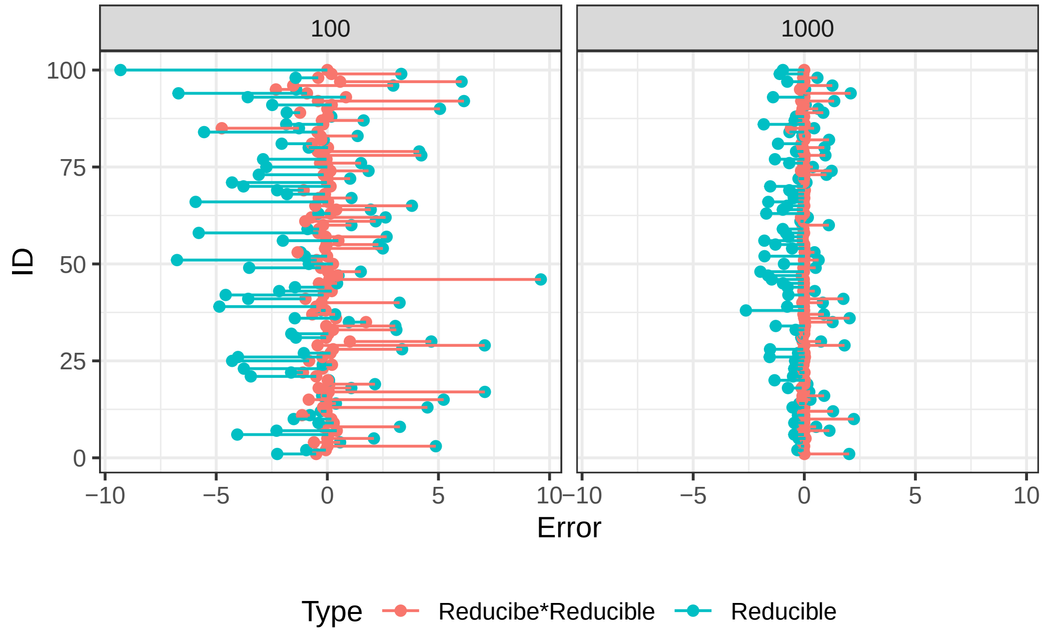
Reducible × Reducible

$$E[(E_P[Y] - g) \times (E_P[Y] - g)]$$

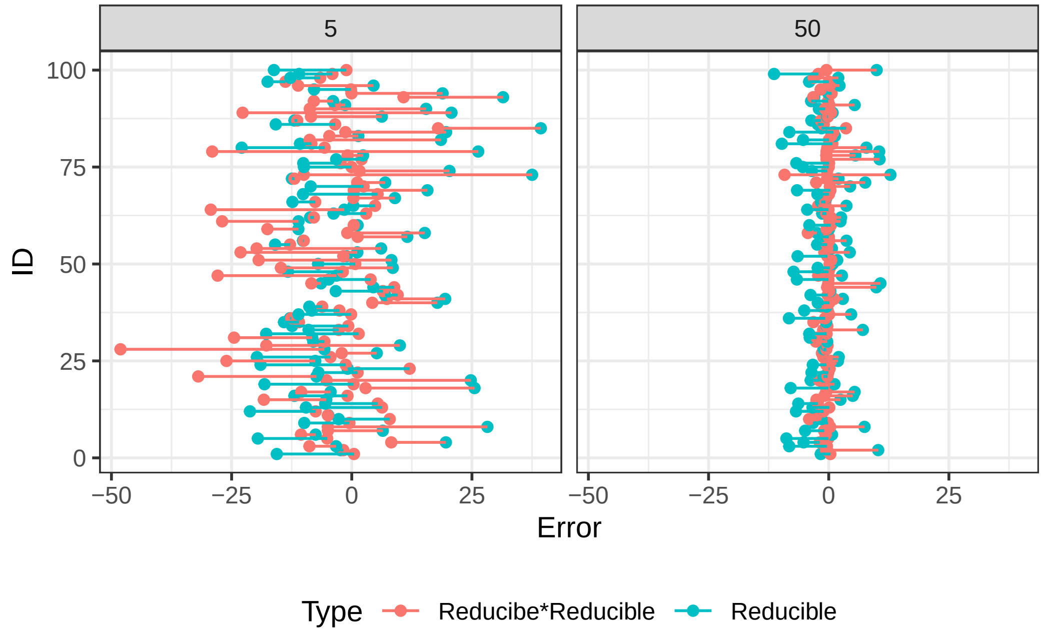
$$= \underbrace{(E_P[Y] - g) \times (E_P[Y] - g)}_{Reducible \times Reducible}$$

- Auxiliary データが研究者によって異なるので、ばらつきをもつ
  - g がバイアスを持つのであれば、掛け算もバイアスを持つ
- ただし掛け算なので、g の推定精度が十分に高ければ (乖離が”1” より小さければ)、掛け算の方が乖離は小さくなる
  - Reducible 単体よりも、収束速度は早くなる

例



不適切な例: 小規模サンプル



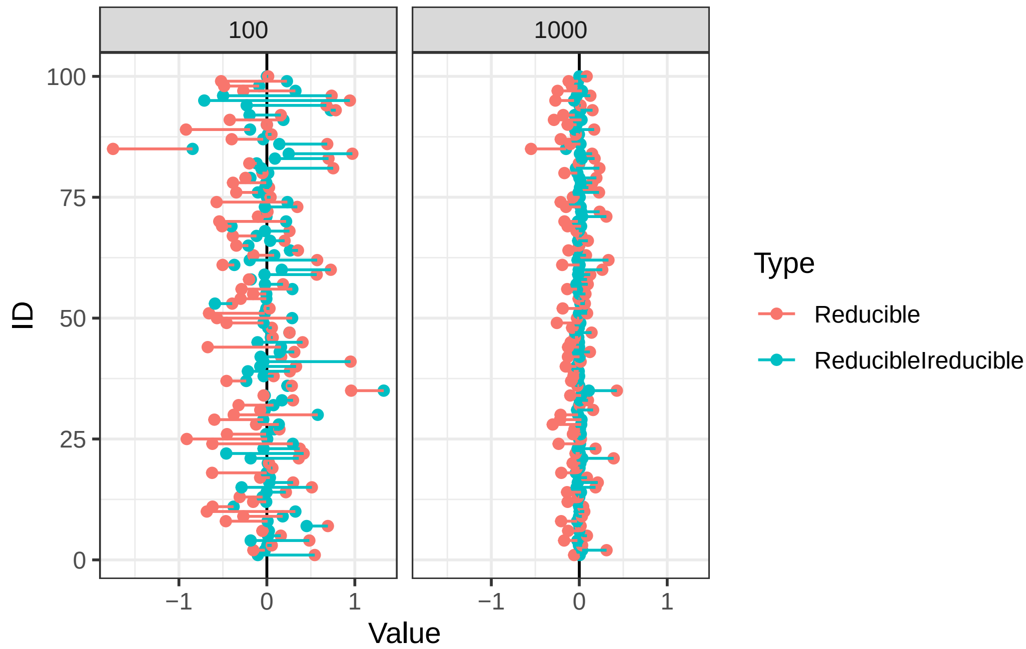
Reducible $\times$ Irreducible

$$2 \times E[\underbrace{(E_P[Y] - g) \times (Y - E_P[Y])}_{\text{無相関}}]$$

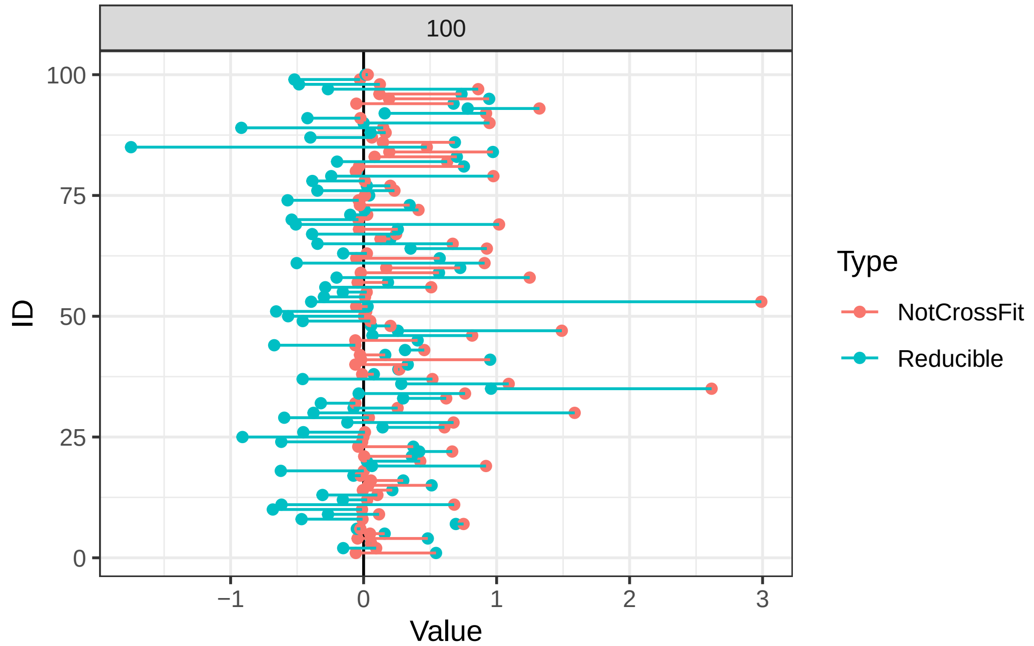
$$= 2 \times (E_P[Y] - g) \times \underbrace{[E[Y] - E_P[Y]]}_{1/\sqrt{N}\text{で収束}}$$

- 注意: g の推定を別のデータで行っているので、書き換えができる
  - バイアスの除去
  - 収束の改善

例



不適切な例: 交差推定なし



PartiallingOut

- ほぼ同じ議論が適用できる

$$\beta = \frac{E[(D - E[D|X])(Y - E[Y|X])]}{E[(D - E[D|X])^2]}$$

$$= \frac{E[(D - E_P[D|X] + E_P[D|X] - E[D|X])(Y - E_P[Y|X] + E_P[Y|X] - E[Y|X])]}{E[(D - E[D|X])^2]}$$

分解

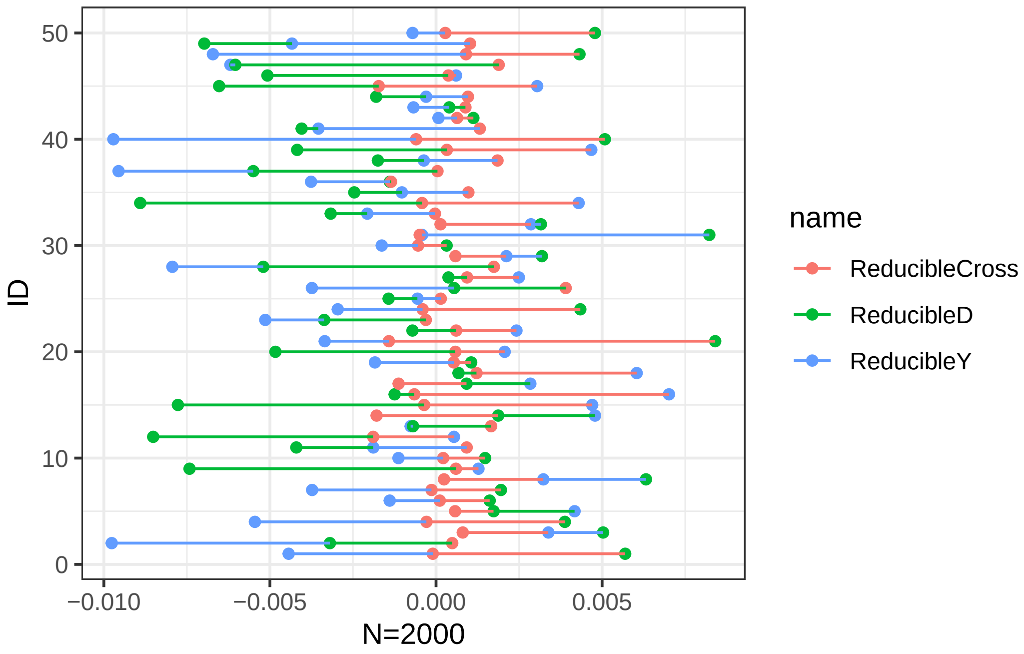
- Oracle 推定料 + Reducible×Reducible + Reducible× Irreducible
- 大きな違いは、$Reducible \times Reducible$

$$= E[(E_P[Y|X] - g_Y(X)) \times (E_P[D|X] - g_D(X))]$$

- Yを予測する”AI” と D を予測する”AI” の間違いの掛け算
  - 十分に精度が高ければ、“AI” 同士の” ダブルチェック” により、間違いを減らせる

– 精度が悪いと、間違いが増加

数値例



まとめ

- Partialling out 推定は、g_Y, g_D が $n^{-1/4}$ よりも早く収束すれば、漸近正規性
  - サンプルサイズがある程度大きくなれば、バイアスを無視できる (Debiased)
- AI 同士の” 掛け算” (Double) にすることで、収束速度を上げている
  - 交差推定も活用
- より一般化可能 (後述)

Reference

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68.