Estimation with Partial Linear Model: Asymptotics

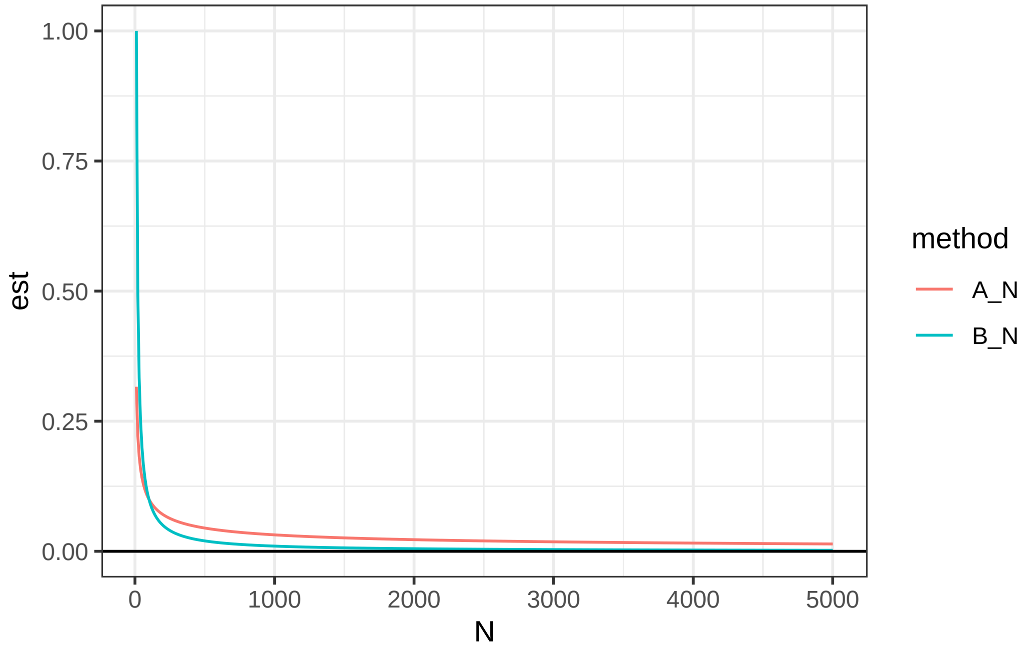
川田恵介

Table of contents

1 大標本性質: without nuisance

- 事例数が無限大に大きい時に成り立つ性質を、事例数が十分に大きことを前提に**近似**的に用いる
  - サンプリング方法に" 強い" 仮定 = "ランダムサンプリング"
  - 教科書的な最尤法やベイズ法に比べて、母集団への parametric assumption が少ない

1.1 イメージ: 近似に基づく議論



- 十分大きい N を前提に、近似的に"0" として議論
  - B_N の方が近似精度が良い

1.2 例: 平均値の推定

- Estimand: Y の母平均 $\theta_0 = E[Y] = \int Y f(Y) dY$
  - Estimator: サンプル平均 $\theta = \sum_i Y_i/N$
    - Moment 法 ("置き換え法")
- Estimator は、データ上の Y の分布に依存するので、研究者によって異なる
  - 一般に $\theta_0 \neq \theta$
  - 多くの実証研究では、点推定量だけでなく信頼区間 (ないし代替指標 (Imbens 2021)) を報告し、対処する

1.3 例: 平均値の推定

- 平均値の推定

```r
readr::read_csv("Public/Data.csv") |>
  estimatr::lm_robust(
    Price ~ 1,
    data = _)
```

```
            Estimate Std. Error  t value Pr(>|t|) CI Lower CI Upper    DF
(Intercept) 39.00496  0.2015849 193.4915        0 38.60984 39.40008 22138
```

- 何を根拠に、どのような解釈ができるのか?

1.4 大標本性質: ざっくり

- 事例数が無限大になると、Estimator の分布について、以下の性質が成り立つ
  - サンプル平均は、母平均 θ_0 に収束する

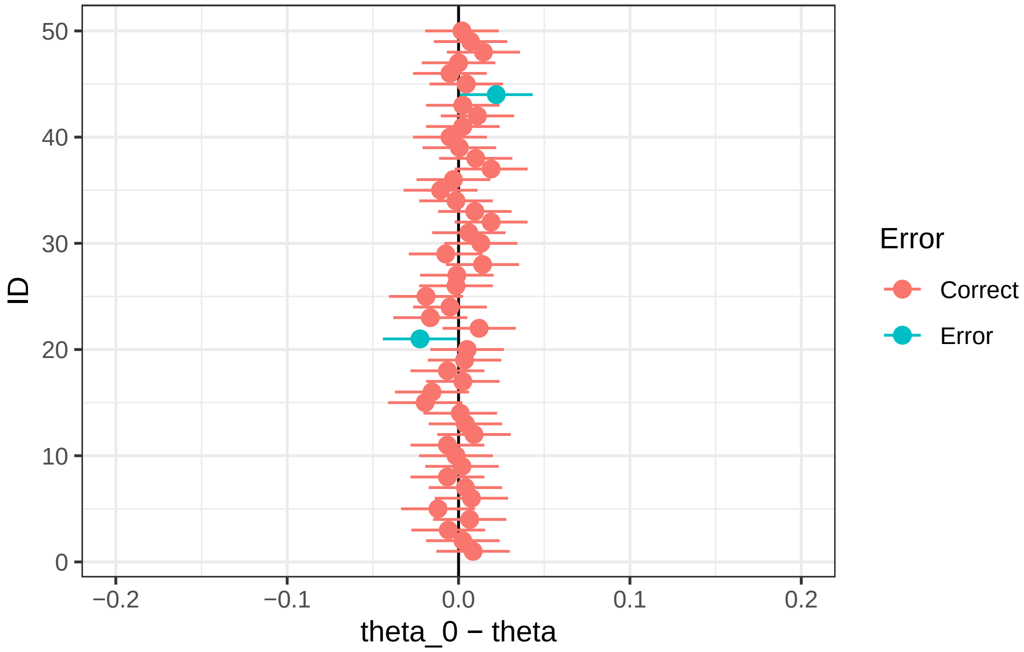
$$\theta_0 - \theta \rightarrow, N \rightarrow \infty$$

  - θ の分布は、正規分布 $N(\theta_0, \sigma^2/N)$ に収束する (中心極限定理)
    - $\sigma^2 = Y$ の母分散

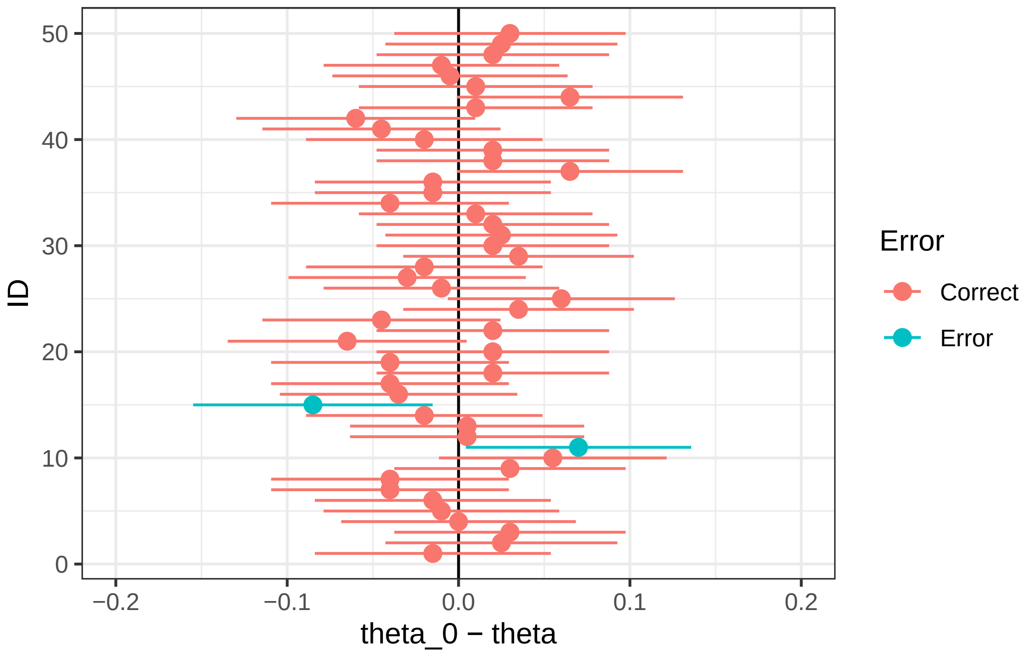
1.5 応用上の含意

- 事例数が**十分に大きい**と
  - 点推定量は、ほぼほぼ母平均と一致する
  - 信頼区間は、ほぼほぼ 95% の" 確率" で母平均を含む
- ただし、十分に大きい、の水準は違う

1.6 平均値: $N = 2000$



1.7 平均値: $N = 200$



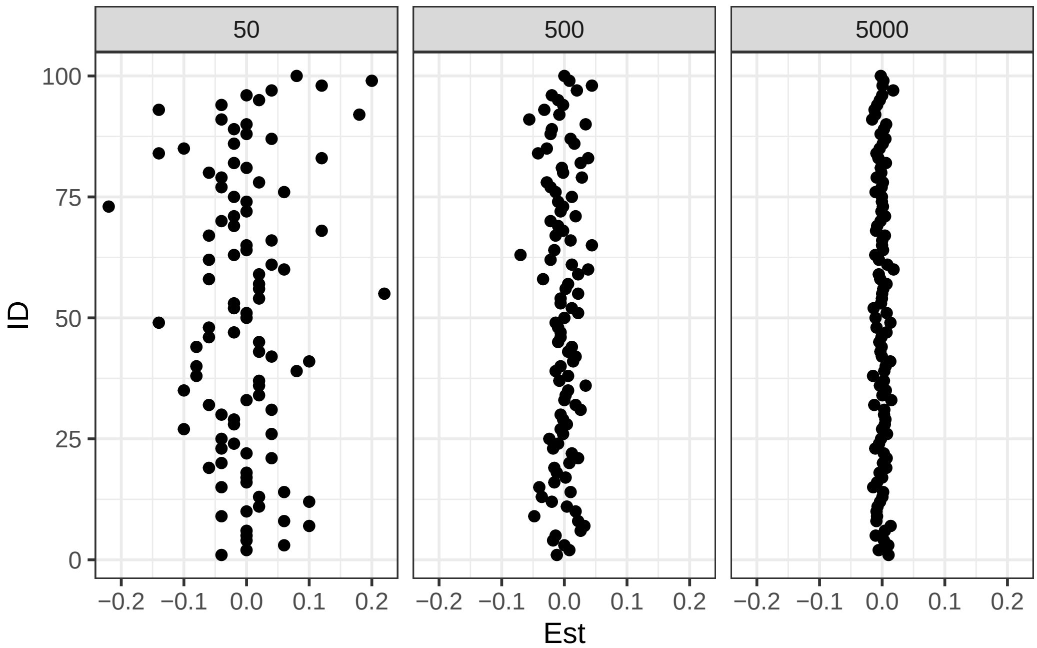
1.8 大標本性質

- $N^{a(<0.5)}(\theta_0 - \theta) \to 0, N \to \infty$
- $N^{a(>0.5)}(\theta_0 - \theta) \to ?, N \to \infty$
- $$N^{0.5}(\theta_0 - \theta) \to Normal(0, \sigma^2), N \to \infty$$
  - よって $\theta_0 - \theta \sim N(0, \sigma^2/N)$
    * σ を推定し、信頼区間を計算できる

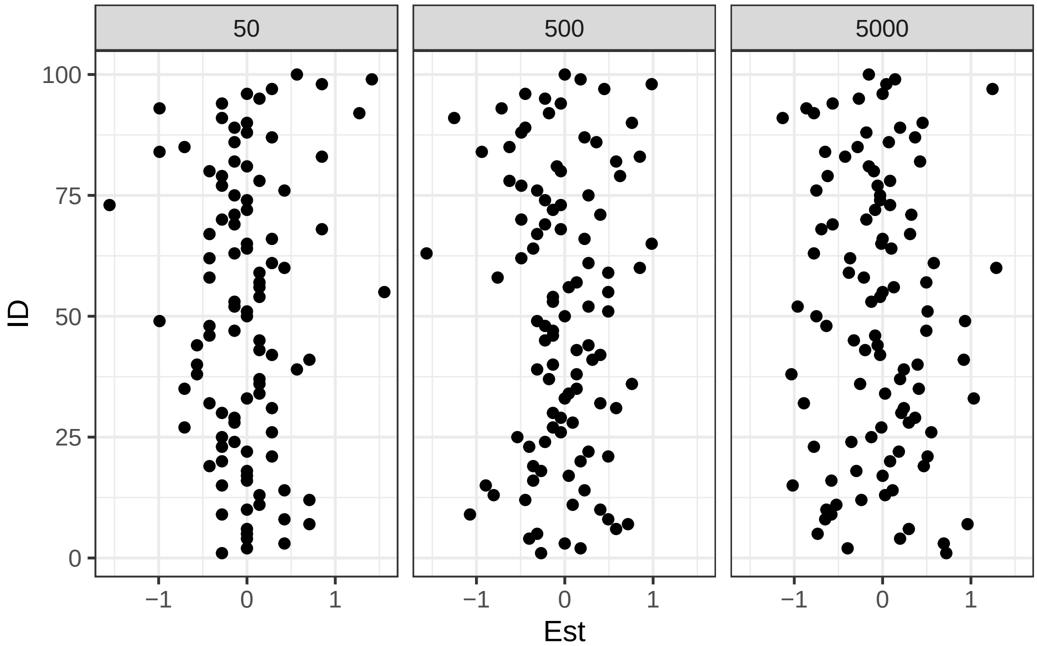
1.9 収束速度

- $\{a, b\} \to 0, N \to \infty$ である時に、
  $$\frac{a}{b} \to 0, N \to \infty$$
  であれば、"a は b よりも早く (確率) 収束する" と呼ぶ
  - N が十分に大きくなれば、$a < b$ が (高い確率) で成り立つ
- 平均値は、$N^{-a(<0.5)}$ よりも早く収束する

1.10 例: $\theta - \theta_0$



1.11 例: $\sqrt{N}(\theta - \theta_0)$



1.12 拡張: 合成指標

- Estimand: 複数の変数 $O = \{X_1, .., X_L\}$ によって、定義される指標 $m(O)$ の平均値 $\theta_0 = E[m(O)]$
  - サンプル平均値 $\theta = \sum m(O)/N$ で置き換える
  - ただし関数 $m(O)$ は既知であり、全ての研究者が同じ式を用いる必要がある
- 例: 国語 X と算数 Y の合計点の平均値

$$m(O = \{X, Y\}) = X + Y$$

1.13 拡張: Implicit function

- 隠関数の平均値として、Estimand は定義できるのであれば、以上の議論を適用できる
- Estimand: 以下の関数を満たす θ_0

$$E[m(\theta_0, O)] = 0$$

  - 一意に θ は定まり、微分可能性
- Estimator = サンプル平均 $0 = \sum m(\theta, O)$ を満たす θ

1.14 例

- サンプル平均: $m(\theta_0, O) = \theta_0 - Y$
- OLS: $m(O, \theta_0) = X(Y - \theta_0 X)$
  - Estimand $= \min E[(Y - \theta_0 X)^2]$ を達成する θ_0

1.15 注意点

- 以上の議論は同じ関数 m を Estimand の定義と推定に用いているが、分離できることに注意
- 同じ θ を定義する関数は、一般に" 無数" に存在する
  - 例: $m = \theta - E[Y]$ と $m = (\theta - E[Y])^2$ は同じ θ
  - 推定上、"便利" な定義を使えば良い

1.16 補論: 正規分布への収束

- Berry-Esseen's Centraol Limit Theorem
  - (see Chap 1 in CausalML)
- 任意の 標準化された X (平均 0, 分散 1) について

$$\underbrace{\sup_{x \in R}}_{\text{"最大値"}} |\Pr[X \leq x] - \Pr[N(0,1) \leq x]| \leq KE[|X|^3]/\sqrt{N}$$

- $K =$ 何らかのパラメタ (< 0.5)

2 大標本性質: with nuisance function

2.1 R-leaner

- Estimand = 以下を満たす θ_0

$$0 = E[m_R(O, \theta_0)]$$

  where

$$O = \{X, D, Y\}$$

$$\mu(X) = \{\mu_D(X), \mu_Y(X)\}$$

$$m_R = (D - \mu_D(X)) \times [Y - \mu_Y(X) - \theta_0(D - \mu_D(X))]$$

2.2 Single-leaner

- 一般に複数の m 関数が、同じ estimand の一致推定量を提供する。
- 例えば、

$$m_S = (D - \mu_D(X)) \times [Y - \theta_0(D - \mu_D(X))]$$

- どれを使えばいいのか?
  - 一つの指針は、大標本性質

2.3 Estimator

- データ上で置き換えると、$\sum m(O_i, g(X), \theta) = 0$ 、ただし $g(X) = \{g_D(X), g_Y(X)\}$ は Auxiliary data を用いて推定された関数
- 一見すると Moment 法がそのまま適用できそうだが、$g(X)$ に依存していることに注意
  - $\mu(X) \neq g(X)$ (AI のミス)

2.4 分解

- 肝は、$\sqrt{N}(\theta_0 - \theta), N \to \infty$ の保証
- 仮想的な Estimator θ^* を考える

$$\sum m(O_i, \mu(X), \theta^*) = 0$$

- AI がミスを犯さないケースの推定値

2.5 分解

- $\sqrt{N}(\theta_0 - \theta) = \underbrace{\underbrace{\sqrt{N}(\theta_0 - \theta^*)}_{O\text{に依存}}}_{\to N(0,\sigma^2), N\to\infty} + \underbrace{\underbrace{\sqrt{N}(\theta^* - \theta)}_{AI\text{のミスに起因}}}_{\to ?, N\to\infty}$
- 一項目に対しては、中心極限定理を適用できる
- 二項目は、
  - Single learer であれば発散する恐れがある
  - R learner であれば、AI のミスの影響が削減できる
    - $* \to 0, N \to \infty$

2.6 イメージ

- Auxiliary data

```
# A tibble: 4 x 3
      X      D      Y
  <int>  <dbl>  <dbl>
1    -1  2.27   3.25
2     1  1.41   1.82
3    -1 -0.540  0.325
4     0 -0.929 -1.09
```

- Main data with prediction (by random forest)

```
# A tibble: 4 x 11
      X      D        Y PredictY PredictD TrueY TrueD   ResY   ResD ResTrueY
  <int>  <dbl>    <dbl>    <dbl>    <dbl> <dbl> <dbl>  <dbl>  <dbl>    <dbl>
1     1  0.401 -0.562       1.06    0.538     1     1 -1.63  -0.137   -1.56
2    -1  1.29  -0.0376      1.06    0.538     1     1 -1.10   0.757   -1.04
3     0  0.390  0.523       1.06    0.538     0     0 -0.540 -0.148    0.523
4     1 -0.208 -0.314       1.06    0.538     1     1 -1.38  -0.746   -1.31
# i 1 more variable: ResTrueD <dbl>
```

2.7 イメージ

- データは、$X = U(-1, 1), D = X^2 + N(0, 1), Y = 2D - X^2 + N(0, 1)$
- $\theta_0 = 2$
- $\theta = (Y - g_Y(X)) \sim (D - g_D(X))$
- $\theta^* = (Y - \mu_Y(X)) \sim (D - \mu_D(X))$
  - $\sqrt{4} * (\theta_0 - \theta^*) = 1.5741579$
  - $\sqrt{4} * (\theta^* - \theta) =$ -0.751546

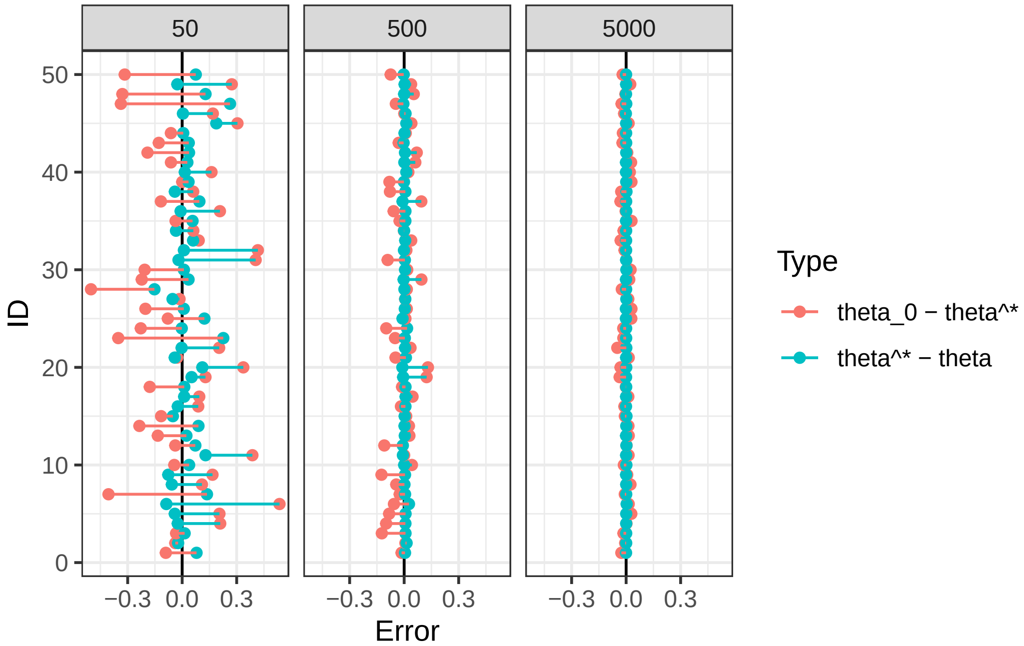
2.8 仮定

- $\sqrt{N}(\theta_0 - \theta) = \underbrace{\sqrt{N}(\theta_0 - \theta^*)}_{\to N(0,\sigma^2), N\to\infty} + \underbrace{\sqrt{N}(\theta^* - \theta)}_{\to 0, N, N\to\infty}$ を保証したい

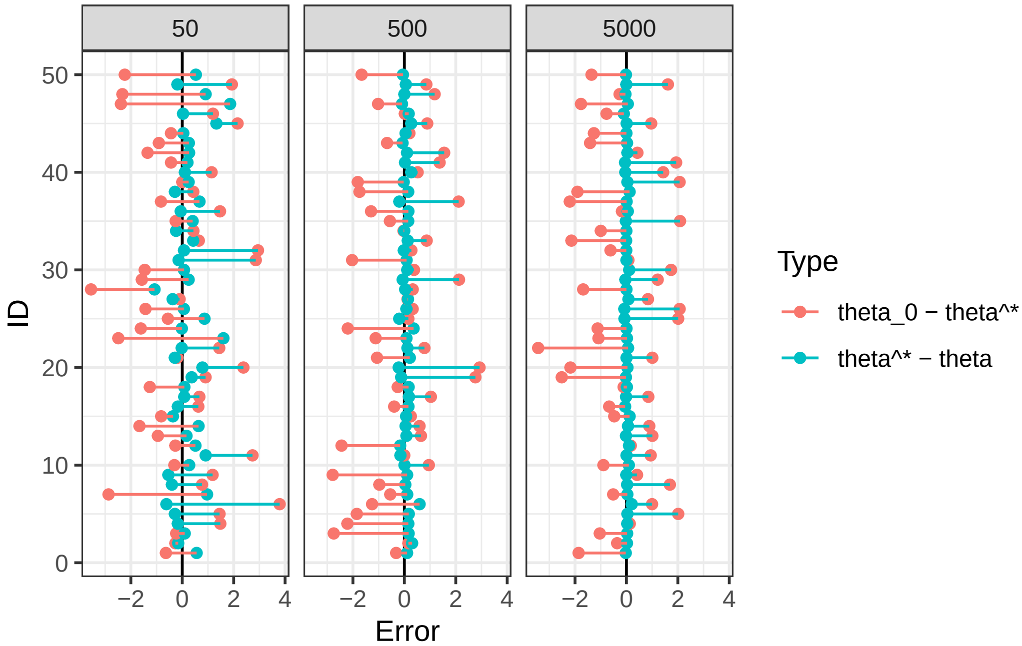
– 第 2 項 (AI のミス) の影響は、(信頼区間を計算できる程度に) 事例数が十分に大きければ、無視できる

- R-learner を前提とした場合、主要な十分条件は
  - サンプル分割
  - g が十分な速度で収束する

2.9 イメージ: R learner



2.10 イメージ: Normalized



2.11 AI のミスの影響への保障: Recap

- AI のミスの影響が $N^{1/4}$ 以上の速度で減少
- 例: $g_D(X = I)$ のミスの影響

$$\left(\sum_{i|X_i=I} Y_i/N_M(I) - g_Y(I)\right) \times \underbrace{e_D(I)}_{\mu_D(I)-g_D(I)}$$

  - g_Y, g_D が十分な速度で μ_Y, μ_D に収束
  - g_D と $\sum_{i|X_i=I} Y_i/N_M(I)$ が無相関

2.12 仮定: 収束速度

- 十分条件の一つは、

$$\left\{N^{1/4}\sqrt{E[(\mu_Y(X) - g_Y(X))^2]},\right.$$

$$\left.N^{1/4}\sqrt{E[(\mu_D(X) - g_D(X))^2]}\right\} \to 0, N \to \infty$$

  - $N^{1/4}$ よりも収束速度が速い

2.13 補論: 収束速度

- g を正しいモデルで OLS 推定できれば、

$$N^{a(<0.5)}\sqrt{E[(\mu(X)-g(X))^2]}\rightarrow 0, N\rightarrow\infty$$

  - $N^{1/4}$ よりも 確実に収束速度が速い

- 誤定式化を犯している OLS では、そもそも収束しない
- 多くの機械学習は、$N^{1/2}$ よりも収束速度が**遅い**
  - R learner は、機械学習 (含む Nonparametric estimation) の収束の遅さを補完

2.14 前提: サンプル分割

- サンプル分割しないと予測モデルと Main data の平均値との間に相関が生じ、収束速度が低下する
  - 相関の影響は (個人的に) わかりにくい
    * 個人的おすすめは、外れ値がデータに紛れ込んだ時の影響を想像する

2.15 数値例

- 同じデータで g を (random forest で) 推定する

```
# A tibble: 8 x 6
      X        D         Y PredictY PredictD    MeanY
  <int>    <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
1    -1 -0.540   -2.37        0.674   1.02     0.849
2     1  0.0714  -1.16       -0.135   0.480  -0.00968
3    -1  0.705   -0.000952    0.674   1.02     0.849
4     0 -0.00577  0.241      -0.725  -0.0520 -1.14
5    -1  3.40     4.92        0.674   1.02     0.849
6     1  1.76     2.96       -0.135   0.480  -0.00968
7     1  0.201   -1.84       -0.135   0.480  -0.00968
8     0 -1.15    -2.52       -0.725  -0.0520 -1.14
```

2.16 数値例: Add outliear

- D (例: 部屋の広さ) が非常に大きな事例が混入
  - Y (例: 取引価格) も同時に大きい

```
# A tibble: 9 x 6
      X        D          Y PredictY PredictD    MeanY
  <dbl>    <dbl>      <dbl>    <dbl>    <dbl>    <dbl>
1    -1 -0.540   -2.37          5.56     3.30  0.849
2     1  0.0714  -1.16          4.53     2.75 -0.00968
3    -1  0.705   -0.000952      5.56     3.30  0.849
4     0 -0.00577  0.241        18.4      9.31 25.9
5    -1  3.40     4.92          5.56     3.30  0.849
6     1  1.76     2.96          4.53     2.75 -0.00968
7     1  0.201   -1.84          4.53     2.75 -0.00968
8     0 -1.15    -2.52         18.4      9.31 25.9
9     0 40       80            18.4      9.31 25.9
```

- $\sum Y/N - g_Y$ も $g_Y - \mu_Y$ も同時増加

2.17 補論: 収束速度

- g を正しいモデルで OLS 推定できれば、

$$N^{a(<0.5)}\sqrt{E[(\mu(X)-g(X))^2]} \rightarrow 0, N \rightarrow \infty$$

  - $N^{1/4}$ よりも 確実に収束速度が速い

- 誤定式化を犯している OLS では、そもそも収束しない

- 多くの機械学習は、$N^{1/2}$ よりも収束速度が**遅い**

  - R learner は、機械学習 (含む Nonparametric estimation) の収束の遅さを補完

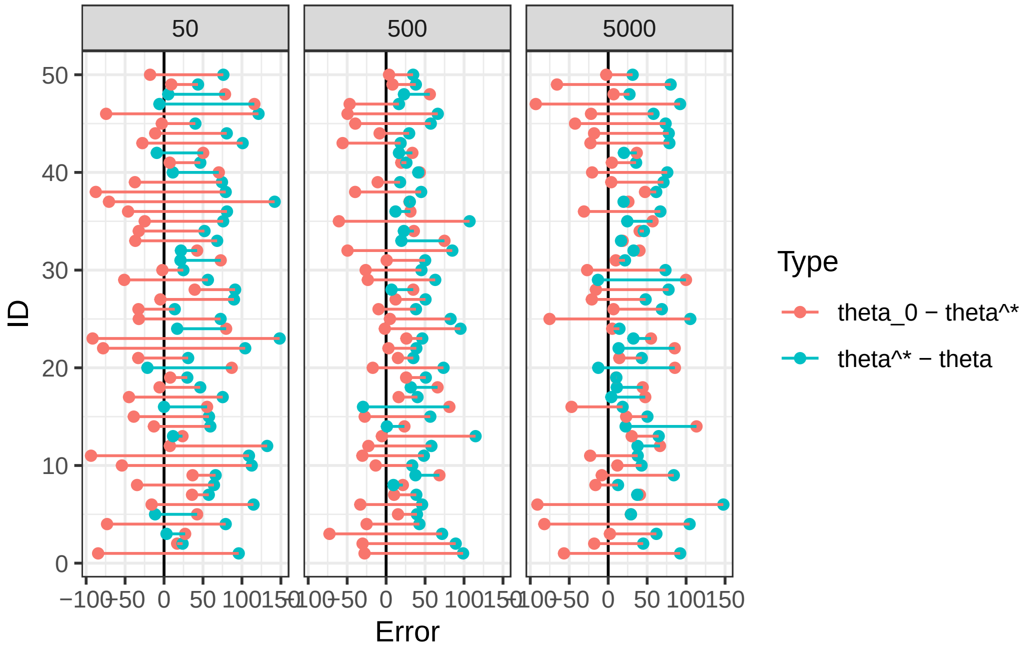
2.18 Single learner

- $g_D(X)$ が、$N^{1/2}$ よりも速い速度で収束する必要がある

  - 正しいモデルを OLS 推定する必要がある

    * 実質" 不可能"

2.19 イメージ: Single Model



3 Neyman's ohtogonal condition

- R learner への議論は、より一般的な状況に適用可能

3.1 Estimand

- $$E[m(\theta_0, O, \mu)]$$

  として、Estimand θ_0 を定義

- m については、
  - θ について一意に定まり、かつ微分可能
  - Neyman の直行条件を満たす
  - サンプル分割、$N^{-1/4}$ よりも収束速度が速いのであれば、μ は機械学習の推定結果で置き換えられる
    - 機械学習で推定できる必要はある

3.2 Neyman's ohthgonal condition

- AI の微妙なミスに対して、estimator が影響を受けない
  - $\partial m/\partial \mu = 0$
    - 関数で微分するとは？？？
- $$\left.\frac{\partial E[m(\theta_0, X, g(t))]}{\partial t}\right|_{t=0} = 0$$
  - $g_Z(t) = tg_Z(X) + (1-t)\mu_Z(X), t \in [0, 1]$
    - 母平均を、何らかの関数に少し移動させる
    - ガトー微分

3.3 実装

- Neyman の直行条件を満たす m 関数は、以下の方法で導出できる
  - テイラー近似 (手計算) (Hines et al. 2022 がわかりやすい入門)
  - データから" 自動計算" する (Chernozhukov, Newey, and Singh 2022)
    - 現状、大衆的な実装方法はない

3.4 仮定の検討

- $n^-1/4$ よりも速い収束の保証は、現状強い仮定
  - X の数が多い場合に特に怪しい
    - 現状は、「正しいモデルを仮定」するよりもマシなので、とりあえず目をつぶって応用している印象
    - Best practice として、Stacking を利用
  - 本質的な代替案としては、高次近似の利用 (Bonvini et al. 2024 とその引用文献) だが、まだ基礎的理論研究が続いている印象

Reference

Bonvini, Matteo, Edward H Kennedy, Oliver Dukes, and Sivaraman Balakrishnan. 2024. "Doubly-Robust Inference and Optimality in Structure-Agnostic Models with Smoothness." *arXiv Preprint arXiv:2405.08525*.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh. 2022. "Automatic Debiased Machine Learning of Causal and Structural Effects." *Econometrica* 90 (3): 967–1027.

Hines, Oliver, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. 2022. "Demystifying Statistical Learning Based on Efficient Influence Functions." *The American Statistician* 76 (3): 292–304.

Imbens, Guido W. 2021. "Statistical Significance, p-Values, and the Reporting of Uncertainty." *Journal of Economic Perspectives* 35 (3): 157–74.