Neyman's orthogonality condition

川田恵介

1 Double debiased machine learning

1.1 機械学習の応用

- 「社会やその仕組み(構造)」の特徴を把握するために
  - Moment 推定をベースに、補助的に機械学習を活用するアプローチが発展
    - Nuisance 関数を含むモーメント条件 with Neyman's orthogonality + 交差推定 + 機械学習 (Chernonozhukov et al., 2018; Victor Chernozhukov, Escanciano, et al., 2022)
- 多くの発展的議論 (Victor Chernozhukov, Newey, et al., 2022; Chernozhukov, Newey and Singh, 2022; Luedtke, 2024; Laan et al., 2025)

1.2 Get Startted: Return to education

```
library(tidyverse)

data("CPS1985", package = "AER")

Y <- log(CPS1985$wage)

D <- CPS1985$education

X <- model.matrix(~ age + ethnicity + gender, CPS1985)

X <- X[,-1]
```

1.3 Get Startted: Return to education

```
model <- ddml::ddml_plm(
  y = Y,
  D = D,
  X = X,
  learners = list(
    list(fun = ddml::ols),
    list(fun = ddml::mdl_ranger)
  ),
  shortstack = TRUE,
```

```r
  silent = TRUE
)

summary(model)
```

```
PLM estimation results:

, , nnls

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00039    0.01951     0.02 9.84e-01
D_r          0.07885    0.00789    10.00 1.58e-23
```

2 Nuisance 関数を含む Moment 法

2.1 Moment 法

- Moment 推定: 母集団上のモーメント条件を用いて推定対象を**定義**し、データ上で同じ条件を適用し**推定**する
  - 本講義の範囲内では、「"母集団上で仮想的に行われる計算結果"として推定対象を定義し、"データ上での同じ計算"の結果を推定値とする方法」と考えて概ね OK

2.2 例: Population OLS

- 関心のあるパラメタ β は、Population OLS の計算結果として**定義**
  - 以下を母集団で最小化するように決定

$$L = \min E[(Y - \beta X)^2]$$

  - モーメント条件としては、

$$\frac{\partial L}{\partial \beta} = 0 \Leftrightarrow 0 = E[X \times (Y - \beta X)]$$

2.3 例: 推定

- ある data 上で、同じ最小化問題を解く: 以下を最小化
  - $\mathbb{L}(data) = (Y - \beta X)^2 の data 上の平均値$
  - モーメント条件は、

$$\frac{\partial \mathbb{L}(data)}{\partial \beta} = 0$$

$$\Leftrightarrow 0 = X \times (Y - \beta X) の data 上の平均値$$

2.4 本スライドの推定対象

- 推定対象 $\beta_D(X)$ は partial linear model で定義される

$$E[Y \mid D, X] = \beta_D(X) \times D + f(X)$$

  - 本スライドでは $\beta_D(X) \simeq \beta_D$ の場合を議論
- $D = [0, 1]$ であれば、$\beta_D = X$ の分布を完全にバランスさせた後のYの平均値の比較
  - **識別の仮定** Conditional independence の元で、$\beta_D = D$ の効果

2.5 重要ポイント

- 同じ推定対象を算出する、母集団上での計算手順(定式化)は、一般に**無数**に存在する
  - 推定 (推定対象 → 推定値) が容易な定式化を採用すれば良い
- 経済学の類似例: 以下は全く同じ生産計画を導く
  - 投入財について利潤最大化問題を解く
  - 投入財について費用最小化問題を解いた後に、生産量について利潤最大が問題を解く
  - 計算/解釈しやすい方法を採用すれば良い

2.6 Single residual による定義

- β_D は、母集団上で以下の問題を解けば良い

$$E\left[(Y - f(X) - \beta_D \times D)^2\right]$$

- $f(X) = E[Y \mid D = 0, X]$ なので、

$$E\left[\left(\underbrace{Y - E[Y \mid D = 0, X]}_{=\hat{Y}} - \beta_D \times D\right)^2\right]$$

- $\hat{Y} \sim D$ を OLS 推定する

2.7 Single residual に基づく推定

- nuisance 関数 ($E[Y \mid D = 0, X]$) を事前に推定する必要がある

1. 予測モデル $g_Y(0, X) \simeq E[Y \mid D = 0, X]$ を Stacking などを用いて推定

- 例えば、$D = 0$の事例のみを用いて、Y を X から予測するモデルとして推定
- あるいは、 X, D から Y を予測するモデルを推定し、 $D = 0$ を代入

2.8 Single residual に基づく推定

2. モデルの予測値を代入し、$\hat{Y} = Y - g_Y(0, X)$ を算出し、$\hat{Y} \sim D$ を OLS 回帰

2.9 Double residuals による定義

- β_D は以下を回帰しても得られる

$$E\left[\left(\underbrace{Y - E[Y \mid X]}_{\hat{Y}} - \beta_D \times \left(\underbrace{D - E[D \mid X]}_{\hat{D}}\right)\right)^2\right]$$

  - $\hat{Y} \sim \hat{D}$ を OLS 推定
- R-learner/Residuals regression とも呼ばれる

2.10 Double residuals による推定

- データ上で以下のように再現できる

1. 予測モデル $g_Y(X) \simeq E[Y \mid X]/g_D(X) \simeq E[D \mid X]$ を Stacking などを用いて推定
2. 推定値を代入し、$\hat{Y} = Y - g_Y(X)/\hat{D} = D - g_D(X)$ を算出し、$\hat{Y} \sim \hat{D}$ を OLS 回帰

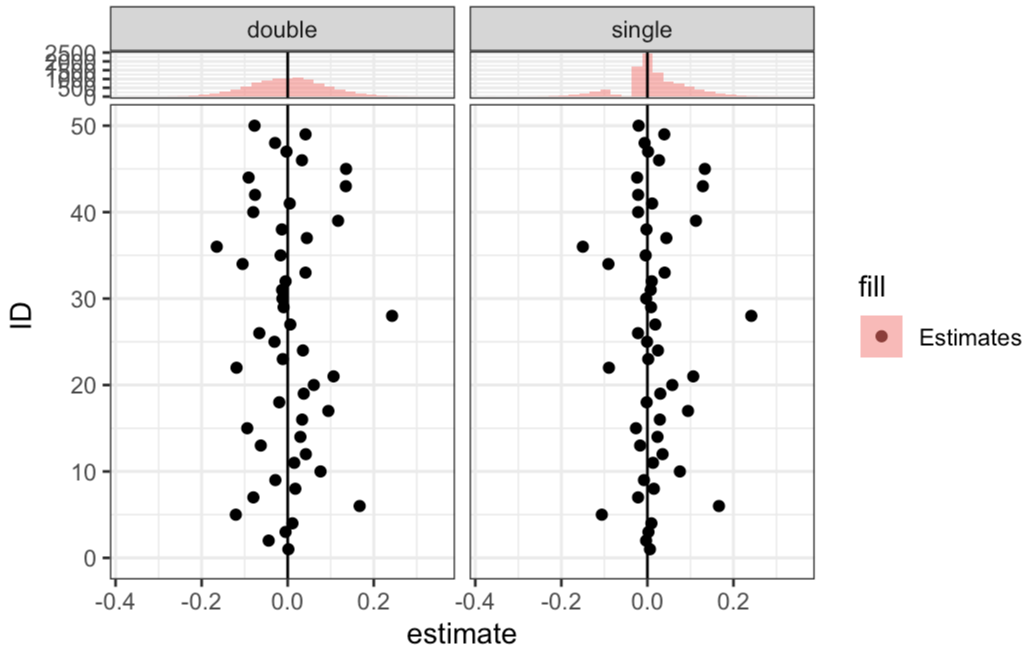
2.11 Single VS Double

- どちらが扱いやすい定式か?
- nuisance 関数 $g_Y(0, X)/g_Y(X)/g_D(X)$ の推定に、機械学習などのデータ主導の推定を活用したいのであれば、Double residuals が推奨される
  - β_D の分布を正規分布で近似**しやすくなり**、信頼区間の近似計算が容易
    - 実際には交差推定も活用する必要がある (次のスライド)
  - Single residual の場合、推定値の分布は一般にバイアスを持つ

2.12 数値例

- $Z_1, X \sim unif(0, 1), Z_2 \sim N(Z_1, 0.5)$
- $D \sim N(X, 0.5), Y \in N(X + Z_2, 1)$
- Nuisance はすべて LASSO で推定
  - スライド内では、500 事例
    - Appendix では、50000 事例

2.13 数値例



3 Oracle estimator を用いた理解

3.1 Single VS Double: 推奨

- Double residuals の定式化は、Neyman's orthogonality を満たすため、nuisance 関数の推定の自由度が高い
  - 機械学習などの推定誤差の性質が悪い/不透明な方法を用いても、緩やかな条件のもとで、最終的な推定結果 β_D の推定誤差に悪影響を与えにくい
    - 「機械学習の推定誤差が、最終的な推定結果に転嫁されにくい」
    - なぜ?

3.2 議論の方針

- CausalML 以外にも、しっかり理解するための教材は多数存在 (Fisher and Kennedy, 2021; Ichimura and Newey, 2022; Hines et al., 2022; Renson et al., 2025; Ahrens et al., 2025)
- ここでは理論的なベンチマーク (Oracle estimator)の分布性質をまず議論し、実際の推定値がベンチマークを近似できるかどうかを議論する
  - 正式な漸近性質は次回議論

3.3 復習: 推定誤差の原因

- 推定誤差(推定値の分布): 同じ母集団を対象としたとしても、データの特徴が異なるため生じる誤差
- 2 段階推定定において、推定誤差の源泉が 2 種類ある
  - Step 1 における、予測モデルの推定結果が異なる
  - **仮に**同じ予測モデルを用いても、Step 2 において、β_D の推定結果が異なる
    - 以上の**総和**が、推定誤差

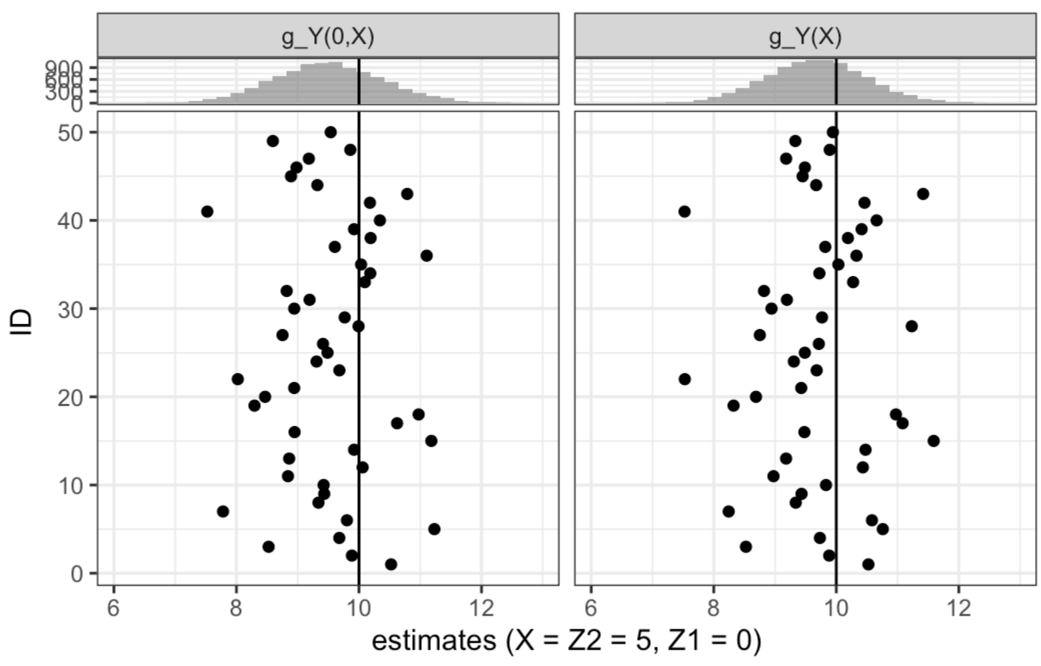
3.4 Step 1 における推定誤差の確認

- 説明のために、推定された$A \in [Y, D]$の予測モデルを以下のように表現する

$$g_A(X) = \underbrace{E[A \mid X]}_{\text{最善の予測値}} + \underbrace{e_A(X)}_{\text{推定誤差}}$$

- 様々なアルゴリズムを用いた Stacking 法などで$g_A(X)$ を推定すれば、**無限大**の事例数のもとで $e_A(X) \simeq 0 \Leftrightarrow g_A(X) = E[A \mid X]$ が期待できる
  - 一致性

3.5 数値例: Nuisance



- 500 事例では、推定誤差がマイナスに偏る (“Finite sample” bias)

3.6 Benchmark

- 論点整理のために、母平均が正確に推定できたケース ($e_A(X) = 0$)を”想像”する
  - Oracle estimator とも呼ばれる
- Step 2 における推定では、

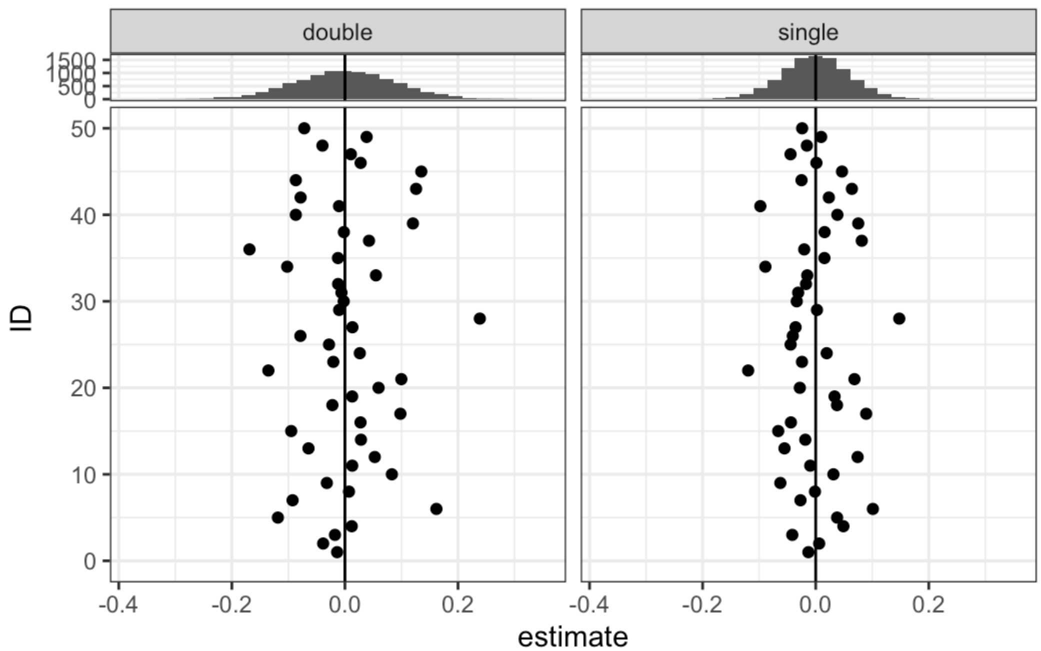
$$Y - E[Y \mid D = 0, X] \sim D$$

ないし

$$Y - E[Y \mid X] \sim D - E[D \mid X]$$

を OLS で推定すれば良い

3.7 数値例: Oracle 対決



- 推定誤差の分布は正規分布で近似できる

3.8 Oracle 推定の近似

- 十分に事例数が大きい場合
- Single residuals の推定誤差は、

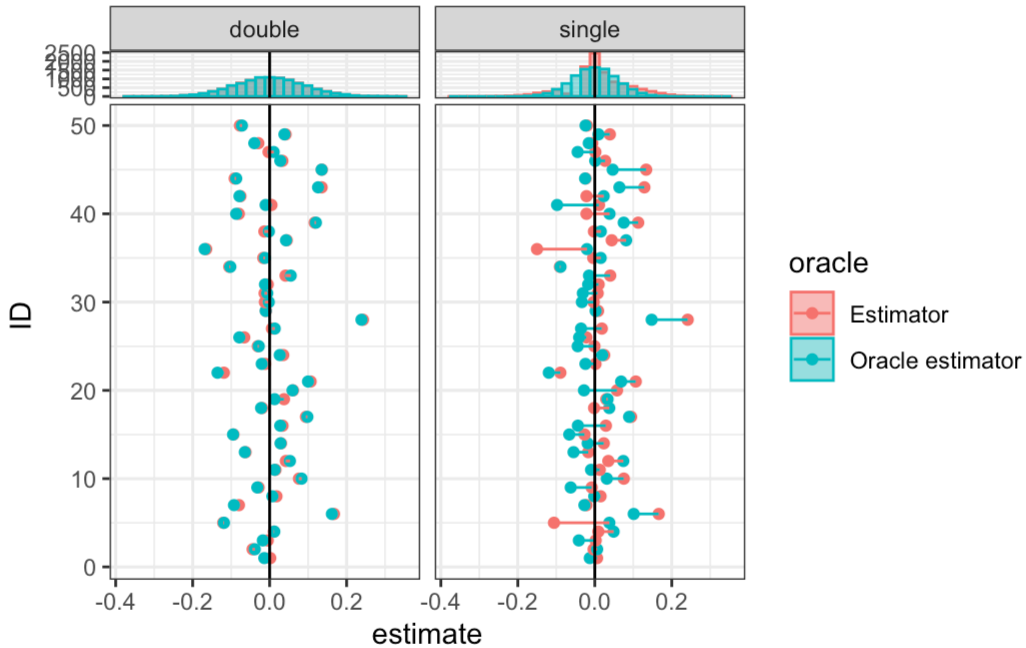
$$\underbrace{Step1\text{の誤差}}_{?} + \underbrace{Step2\text{の誤差}}_{\text{正規分布で近似}} \neq Oracle\text{推定}$$

- 緩やかな仮定 (次回議論) のもとで、Double residuals の推定誤差は、

$$\underbrace{Step1\text{の誤差}}_{\simeq 0} + \underbrace{Step2\text{の誤差}}_{\text{正規分布で近似}} \simeq Oracle\text{推定}$$

- Step 1 における推定誤差を、近似的に無視できる

3.9 数値例: VS Oracle 対決



3.10 まとめ

- Double residuals による定義 + 緩やかな仮定 (nuisance 関数についての推定精度 + 交差推定; 次回議論)のもとで、推定値は oracle estimator と近似的に一致
  - 十分な事例数であれば、step 1 の推定誤差を無視できる
    - step 1 に機械学習を活用し、推定誤差の詳細な性質が不明でも、最終的な推定結果が影響されにくくできる

4 Neyman's orthogonality

4.1 微分の活用

- なぜ Step 1 の誤差の影響が、定式化で異なるのか?
  - "微分"の応用で、確認できる

4.2 復習: 微分の使い方

- $f(X)$ を微分 $\partial f(X)/\partial X = X$ を少し変化させた場合の$f(X)$ の変化率
  - $\partial f(X)/\partial X\ |_{X=0} = 0 \Leftrightarrow X$ の$(X = 0$ からの)局所的な変化について $f(X)$ の値は頑強

- local robust

4.3 推定

- データ上で以下をβ_Dについて最小化

$$L_S(data) = (Y - g_Y(D=0,X) - \beta_D D)^2\text{の平均}$$

$$L_D(data) = (Y - g_Y(X) - \beta_D(D - g_D(X)))^2\text{の平均}$$

- $\partial L_S(data)/\partial\beta_D = \partial L_D(data)/\partial\beta_D = 0$ が、β_D の推定値の決定式

4.4 推定誤差の影響

- β_Dの決定式が、予測モデル g_Y の推定誤差 e_Y について local robust であること確認したい
- $g_Y(X) = E[Y \mid X] + e_Y(X)$ と書き換え、$e_Y(X)$ で交差微分を計算し、$e_Y(X) = 0$ で評価

4.5 シンプルな例

- 本スライドでは、X の組み合わせは少数であると想定する
  - $X = [\text{板橋}, \text{文京}]$
  - 板橋/文京区についての予測誤差 $e_Y(\text{板橋})/e_Y(\text{文京})$ をパラメタとして扱えるため、通常の微分が可能
- より一般のケースでは、関数微分を用いる必要がある

4.6 Single rediauls

- $$\partial\left(\frac{\partial\mathbb{L}_S(data)}{\partial\beta_D}\right)/\partial e_Y(0,\text{板橋})\ |_{e_Y(0,\text{板橋})=0}$$

$$= \frac{\partial(D[Y - E[Y \mid 0,X] - e_Y(0,X) - \beta_D D])\text{の平均}}{\partial e_Y(0,\text{板橋})}\ |_{e_Y(0,\text{板橋})=0}$$

$$= -D\text{の板橋内での平均} \neq 0$$

- step 1 の推定誤差について、β_D は robust ではない

4.7 Double rediauls

- $$\partial\left(\frac{\partial\mathbb{L}_D(data)}{\partial\beta_D}\right)/\partial e_Y(\text{板橋})\ |_{e_Y(\text{板橋})=0}$$

$$= -(D - g_D(\text{板橋}))\text{の板橋内での平均}$$

$$= -[D\text{の板橋内での平均値} - g_D(\text{板橋})]$$

4.8 Double rediauls

- step 1 で用いる事例数が十分にあれば、g_D(板橋) $\simeq$ 母平均
- step 2 で用いる事例数が十分にあれば、データ上での平均 $\simeq$ 母平均
- よって、$\partial\left(\frac{\partial L_D(data)}{\partial \beta_D}\right)/\partial e_Y(\text{板橋}) \mid_{e_Y(\text{板橋})=0} \simeq 0$

4.9 事例数の影響

- Single residuals: 事例数が増えれば、step 1 の推定誤差 $e_A(X)$ を削減できる
  - 現実的な事例数のもとでは、step 1 の推定誤差の影響は、0 にはならない
- Double residuals: step 1 の推定誤差の削減 + step 1 の推定誤差の”影響”の削減
  - 現実的な事例数のもとで、ほぼ 0 になることが期待できる
    - 「収束の速度が改善する」(次回議論)

4.10 Neyman’s orthogonality

- 推定対象となるパラメタ θ は、nuisance 関数 g_0 を用いて、以下の最適化問題の解として定義する

$$\min L(\theta, g_0)$$

- Neyman’s orthogonality: 母集団上で定義されるモーメント条件が以下の条件を満たす

$$\frac{\partial^2 L(\theta, g)}{\partial \beta \partial g} \underset{g=g_0}{\mid} = 0$$

4.11 別表現

- θは母集団上でのモーメント条件 $E[m(\theta, g_0)] = 0$ として定義する
- Neyman’s orthogonality: 母集団上で定義されるモーメント条件が以下の条件を満たす
-

$$E\left[\frac{\partial m(\theta, g)}{\partial g} \mid_{g=g_0}\right] = 0$$

4.12 まとめ

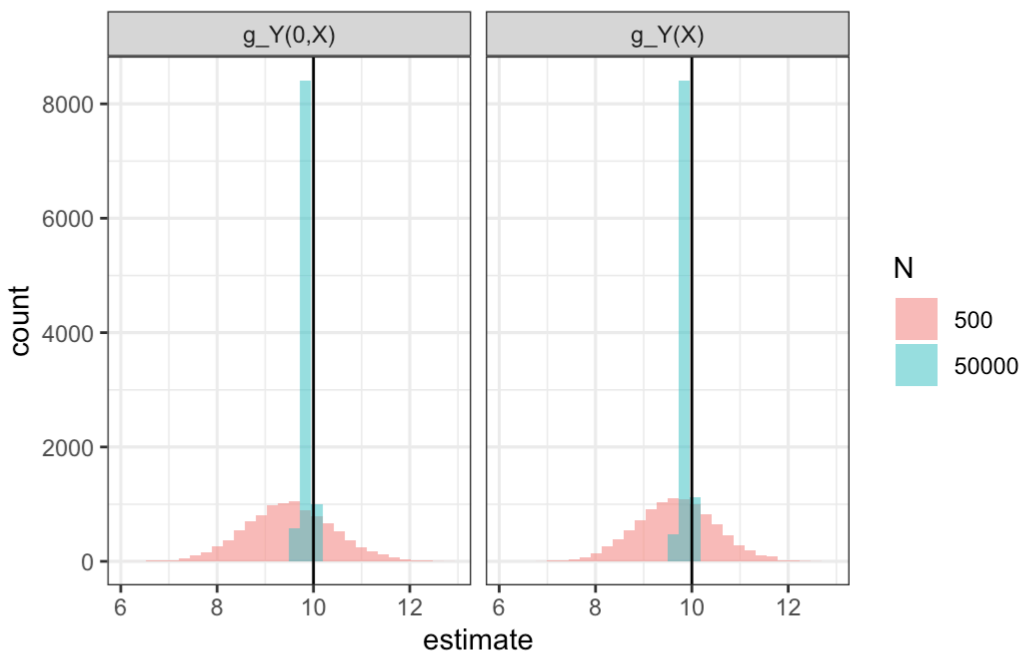
- 機械学習を用いて推定した予測モデル $g(X)$ は、一般に母平均からの推定誤差 $E[Y \mid X] - g(X)$ を減らすようにデザインされる
  - 推定誤差をうまく削減したとしても、その詳細な性質が black box
    - 一般にバイアスを持ち、正規分布で近似できない

- ⇔ 通常の OLS については、母平均の”近似モデル”についての推定誤差はバイアスのない正規分布で近似できる

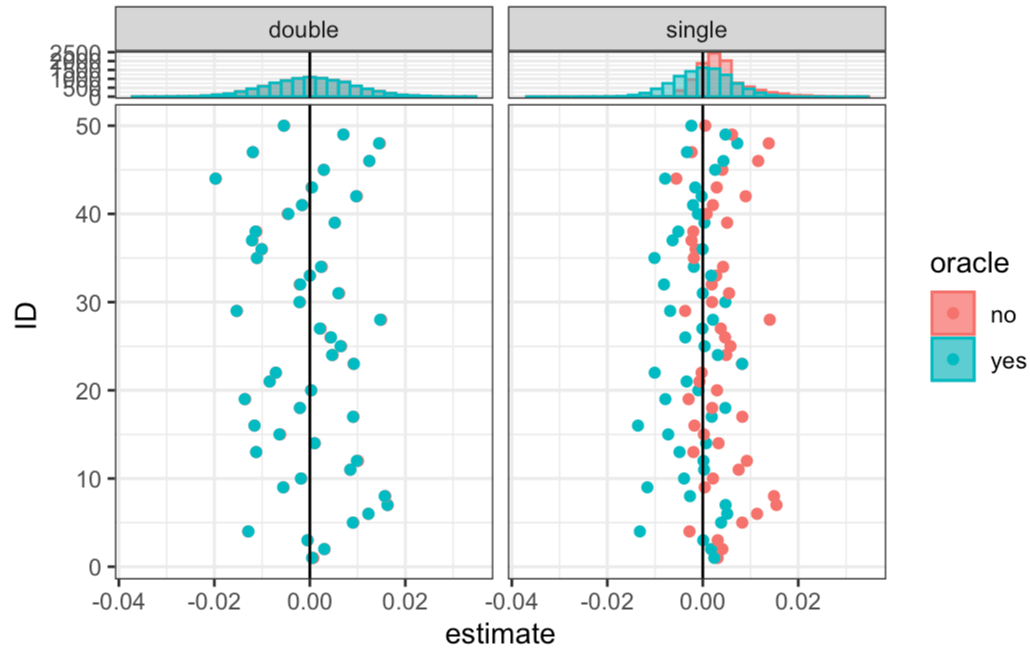
4.13 まとめ

- Debiased machine learning のアイディア: 機械学習の推定誤差を”改善”ではなく、最終的な推定値への影響を削減する
  - Neyman’s orthogonality を満たすように推定対象を定義する
- 経済学における類似例: 景気変動や地震への対応
  - 一般に景気変動の要因や地震そのものを”消失”させることは困難
    - 悪影響が家計に及ばないような(金融/財政/社会保障/減災)政策を議論

4.14 Appendix: Large sample example



4.15 Appendix: Large sample example



4.16 Reference

Bibliography

Ahrens, A. et al. (2025) “An introduction to double/debiased machine learning,” arXiv preprint arXiv:2504.08324 [Preprint].

Chernonozhukov, V. et al. (2018) “Double/debiased machine learning for treatment and structural parameters,” The Econometrics Journal, 21(1), pp. C1–C68.

Chernozhukov, Victor, Escanciano, et al. (2022) “Locally robust semiparametric estimation,” Econometrica, 90(4), pp. 1501–1535.

Chernozhukov, V., Newey, W. and Singh, R. (2022) “Automatic debiased machine learning of causal and structural effects,” Econometrica, 90(3), pp. 967–1027.

Chernozhukov, Victor, Newey, et al. (2022) “Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests,” in International Conference on Machine Learning, pp. 3901–3914.

Fisher, A. and Kennedy, E.H. (2021) “Visually communicating and teaching intuition for influence functions,” The American Statistician, 75(2), pp. 162–172.

Hines, O. et al. (2022) “Demystifying statistical learning based on efficient influence functions,” The American Statistician, 76(3), pp. 292–304.

Ichimura, H. and Newey, W.K. (2022) “The influence function of semiparametric estimators,” Quantitative Economics, 13(1), pp. 29–61.

Laan, L. van der et al. (2025) “Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands,” arXiv preprint arXiv:2501.11868 [Preprint].

Luedtke, A. (2024) “Simplifying debiased inference via automatic differentiation and probabilistic programming,” arXiv preprint arXiv:2405.08675 [Preprint].

Renson, A. et al. (2025) “Pulling back the curtain: the road from statistical estimand to machine-learning based estimator for epidemiologists (no wizard required),” arXiv preprint arXiv:2502.05363 [Preprint].