Estimate structural parameter/function

機械学習の経済学への応用

川田恵介

## Goal

* **母集団における**条件付き平均差: の**特徴を推論**
  + 漸近正規性を根拠とした信頼区間推定
  + 機械学習を活用した統計モデル依存の緩和
  + セミパラメトリック推定を活用した収束性質の改善

## MainGoal

1. の推論
2. のNonparametric推定 (Tree系統)
3. の推論

* クロスセクション、および2時点パネル(かつ がカテゴリ変数)を想定

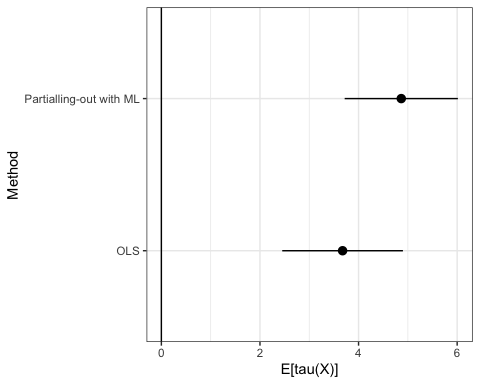
## とりあえずのゴール

* Partial Linear Model (Robinson 1988) の推定
* Partialling-out推定

1. の予測モデルを推定（機械学習も可）
2. 予測できなかった部分 を抽出し(partialling-out)、 OLSで回帰
3. Robust Standard Errorを計算し、(漸近)信頼区間を形成

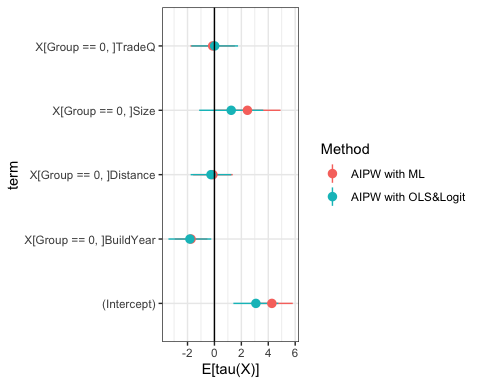
## Gallary: MarginalMeanDifference

* Chernozhukov et al. (2018), Chernozhukov, Escanciano, et al. (2022)



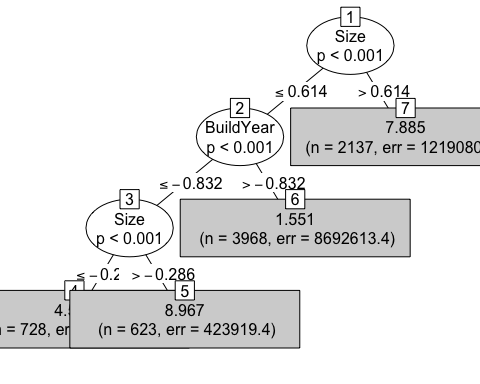
## Gallary: BLP

* Semenova and Chernozhukov (2021)



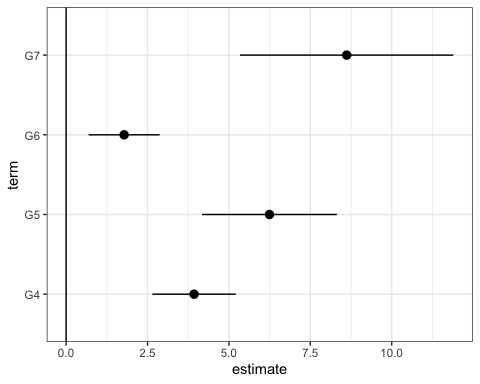
## Gallary: CausalTree

* Athey and Imbens (2016)



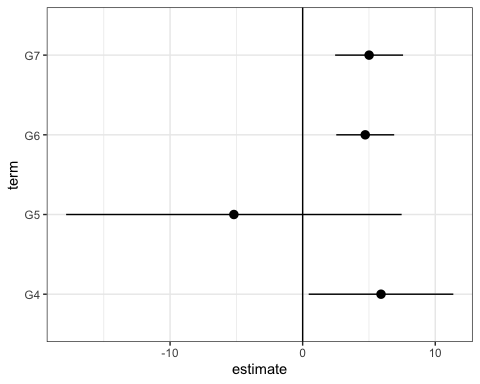
## Gallary: ConfirmTree

* Athey and Imbens (2016)



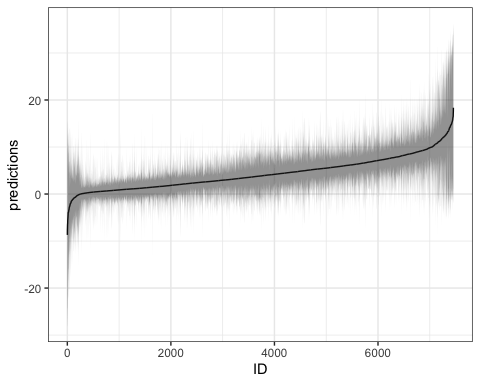
## Gallary: EvilTree

* Athey and Imbens (2016)



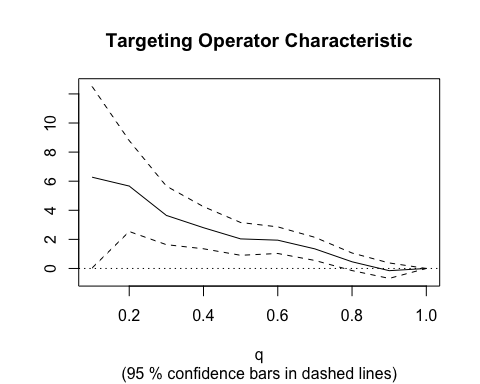
## Gallary: RandomForest

* Athey, Tibshirani, and Wager (2019), Wager and Athey (2018)



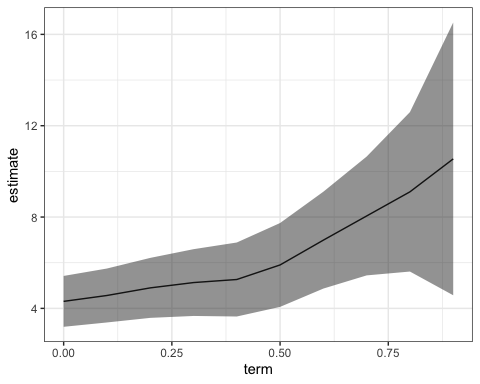
## Gallary: RATE

* Yadlowsky et al. (2021)



## Gallary: SortedCATE

* Kallus (2022)

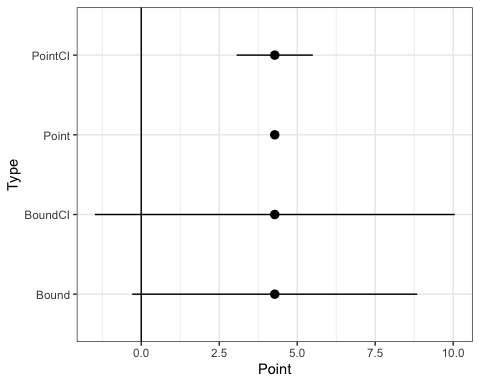


## Gallary: Sensitivity

* Chernozhukov, Cinelli, et al. (2022), Cinelli and Hazlett (2020)

[1] -0.2900195 8.8491779

[1] -1.489593 10.048752



## 基本方針

* 教師付き学習の目標:
  + OLS: の不偏推定
* 母分布全体の正確な推定は不可能
* “推論”も困難
  + 伝統的推定: 統計モデル定式化への依存
  + 機械学習: 漸近性質の悪化
* 分布の**部分的な特徴**のみを推論するのであれば？
  + 社会構造（格差、因果効果など）研究では”十分”

## なぜ十分か？

* “特定の”解釈が可能な母集団の特徴を推定
  + データの外で決定
* 例:
  + 使用OS間不平等として解釈可能？
  + 使用OSの因果効果として解釈可能？
* Age は**関心外** (Nuisance変数)

## なぜ困難か？

* 母平均を完璧に推定することは非現実的
  + 正しいモデルをどれだけ大きなデータ( 無限大)で推定したとしても、一般に正しい値と一致しない

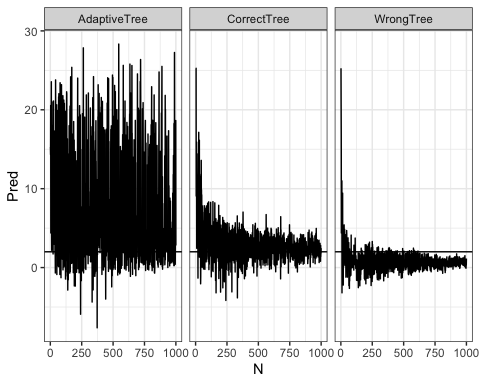
## Repeated Sampling Approach

* 諸々の仮定が”正しければ”、“信頼区間”は形成できる
  + 同じ手法で独立したサンプリングを行う大量の”並行世界”を想定
  + 多くの世界(95%など)で、真の値を含んだ区間を得られる
* Moving the GoalPost

## 数値例: 収束

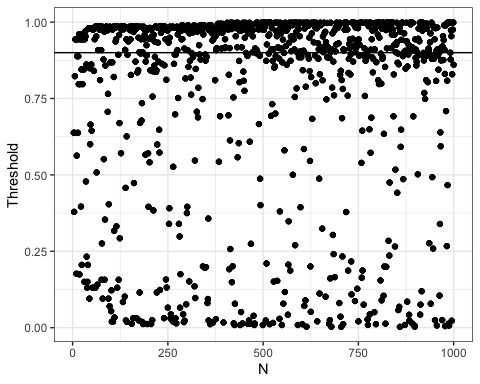
* シナリオ: 最初のサンプルで上振れ( の上位99%)
* アルゴリズム: CorrectTree ( を に回帰), WrongTree ( を に回帰), AdaptiveTree (最大2分割, 最小 サンプルサイズ = 5)

## 数値例: 収束

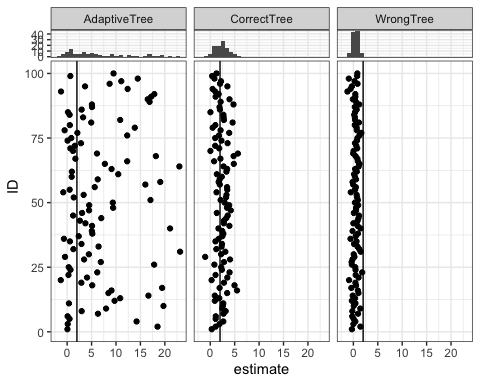


## なぜ？

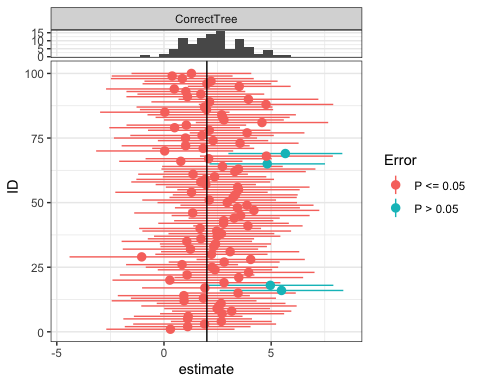
* 推定するモデルそのものがデータの影響を受けてしまう
  + 予測木において、ハズレ値を少数グループに”隔離”した方が、Fitが改善する



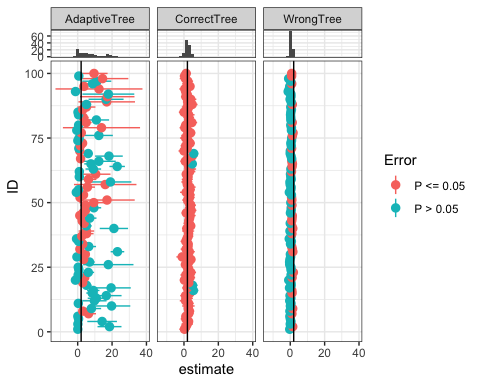
## 分布: 500サンプル



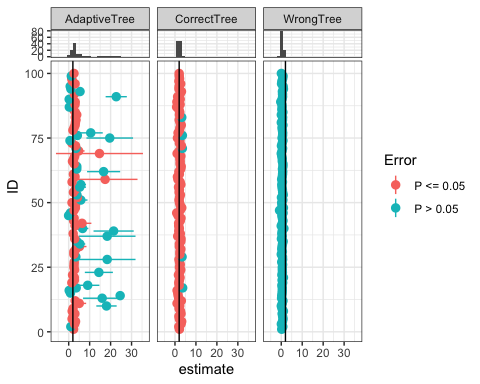
## 漸近正規性を用いた信頼区間



## 漸近正規性を用いた信頼区間



## 分布: 2000サンプル



## Target Estimation

* 母分布の部分的特徴を推論
* 典型例: where
  + 因果推論: 内でDがランダムに決定されているのであれば、平均効果と解釈可能
  + 格差研究: 内平均格差
* セミパラ推定 with 機械学習を用いることでより頑強に推定可能
  + Partialling-outはその一例

## Plug-in 推定

* 最も直接な推定方針は以下

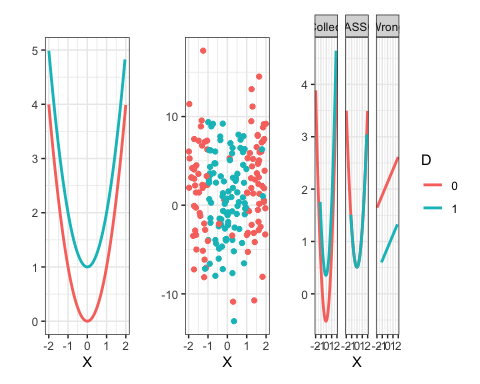
1. を推定: ”予測モデル”
2. 各事例に対して予測モデルを適用し、 をの推定値とする

* のデータ上での分布(Empirical Distribution)をPlug-inしている
  + OLSもこの一種: をOLSで推定

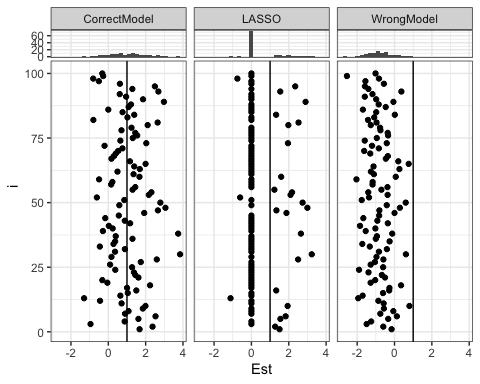
## Pluginの問題点

* の推論結果は、 の推定精度に決定的に依存する
  + の数 サンプルサイズ、かつ推定モデルが正しければ、OLSでOK
  + ハードルが下がっていない
* 誤定式モデルをOLS推定: 一般に は一致性すら満たさない
* 機械学習: は一致性を満たしうるが、一般に漸近正規性が成り立たない
* どちらも通常の信頼区間が形成できない

## 数値例

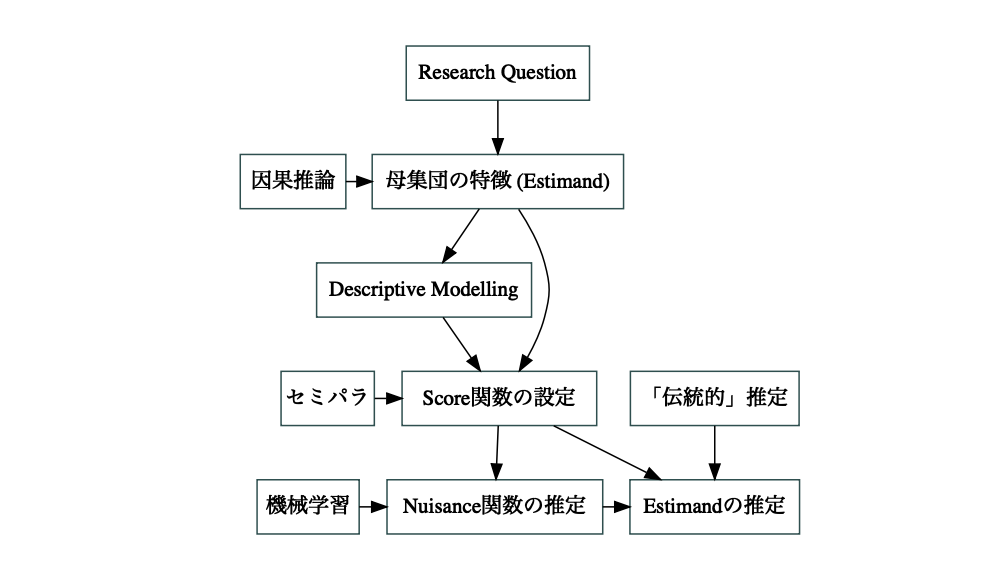


## 数値例

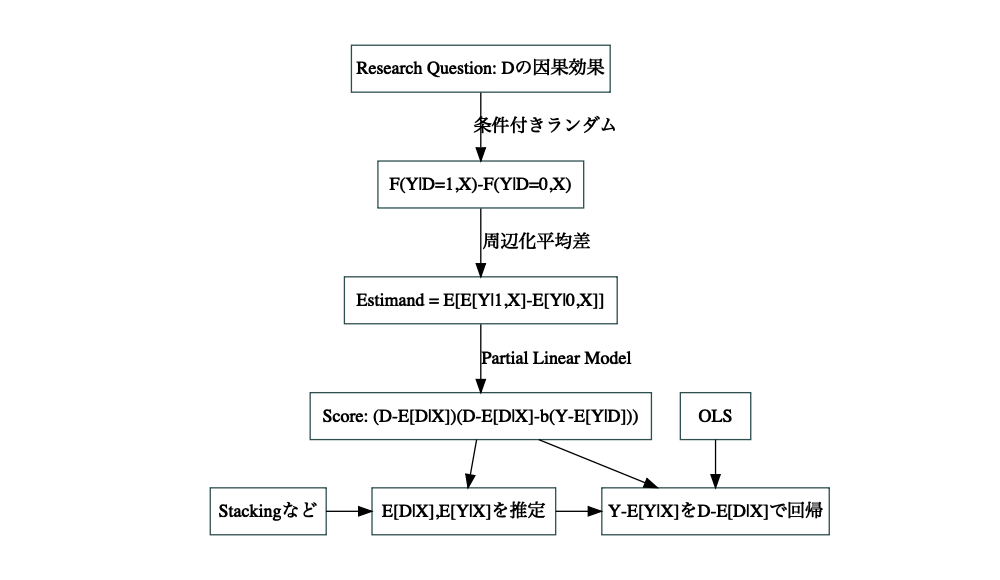


## Research PipeLine

Gruber et al. (2022)



## 例: Research PipeLine



## まとめ

* 限定的なデータから複雑な社会を推論するために工夫を重ねる
* 母集団を用いた論点整理: Estimandは母集団上で定義 (データが変わっても、不変)
* Principle (原理原則)のある選択: 第３者に説明・再現可能 (Urminsky, Hansen, and Chernozhukov, n.d.)
  + 機械学習 + 事前知識 (例 VanderWeele (2019))
* 教師付き学習 条件つき平均値をBlackBoxな関数として学習するための道具
  + 解釈可能なモデリングを行うためのツールとしては使わない

## Reference

Athey, Susan, and Guido Imbens. 2016. “Recursive Partitioning for Heterogeneous Causal Effects.” *Proceedings of the National Academy of Sciences* 113 (27): 7353–60. <https://doi.org/10.1073/pnas.1510489113>.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. “Generalized Random Forests.” *The Annals of Statistics* 47 (2): 11481178. <https://doi.org/10.1214/18-AOS1709>.

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): C1C68. <https://doi.org/10.1111/ectj.12097>.

Chernozhukov, Victor, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. 2022. “Long Story Short: Omitted Variable Bias in Causal Machine Learning.” <https://doi.org/10.3386/w30302>.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. 2022. “Locally Robust Semiparametric Estimation.” *Econometrica* 90 (4): 1501–35. <https://doi.org/10.3982/ECTA16294>.

Cinelli, Carlos, and Chad Hazlett. 2020. “Making Sense of Sensitivity: Extending Omitted Variable Bias.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (1): 39–67.

Gruber, Susan, Hana Lee, Rachael Phillips, Martin Ho, and Mark van der Laan. 2022. “Developing a Targeted Learning-Based Statistical Analysis Plan.” *Statistics in Biopharmaceutical Research* 0 (0): 1–8. <https://doi.org/10.1080/19466315.2022.2116104>.

Kallus, Nathan. 2022. “Treatment Effect Risk: Bounds and Inference.” *arXiv Preprint arXiv:2201.05893*.

Robinson, P. M. 1988. “Root-n-Consistent Semiparametric Regression.” *Econometrica* 56 (4): 931954. <https://doi.org/10.2307/1912705>.

Semenova, Vira, and Victor Chernozhukov. 2021. “Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions.” *The Econometrics Journal* 24 (2): 264289. <https://doi.org/10.1093/ectj/utaa027>.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. n.d. “The Double-Lasso Method for Principled Variable Selection.” <https://doi.org/10.31234/osf.io/2pema>.

VanderWeele, Tyler J. 2019. “Principles of Confounder Selection.” *European Journal of Epidemiology* 34 (3): 211–19. <https://doi.org/10.1007/s10654-019-00494-6>.

Wager, Stefan, and Susan Athey. 2018. “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.” *Journal of the American Statistical Association* 113 (523): 1228–42. <https://doi.org/10.1080/01621459.2017.1319839>.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2021. “Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects.” *arXiv Preprint arXiv:2111.07966*.