教師付き学習の概要と位置付け

機械学習の経済学への応用

川田恵介

事務連絡

- 講義資料 (https://github.com/tetokawata/TargetML)
  - スライドのノート版、Example Code, Data, など
- 評価: レポート（3 回予定）
  - 手法の説明 & R (Python も可) での実装

本講義

- 教師付き学習の入門と応用
  - 予測問題と母集団の” パラメタ” 推定の入門
- 対象: 機械学習の初学者 &| 経済学 ($\simeq$ 社会科学; Biomedical Science) への応用に関心がある人
- 便益: “統計モデルの定式化” 依存を減らす
  - ” 機械学習”, “データサイエンス” というキーワードの入った求人が大学内外から増えている
  - データ分析についての異なる” カルチャー” に馴染む

Motivation

- 伝統的アプローチ: データを見る前に設定した Parametric な統計モデルの推定: 例

$$Y_i = \beta_0 + \beta_1 X_{1i} + .. + \beta_L X_{Li} + u_i, u_i \sim N(0, \sigma^2)$$

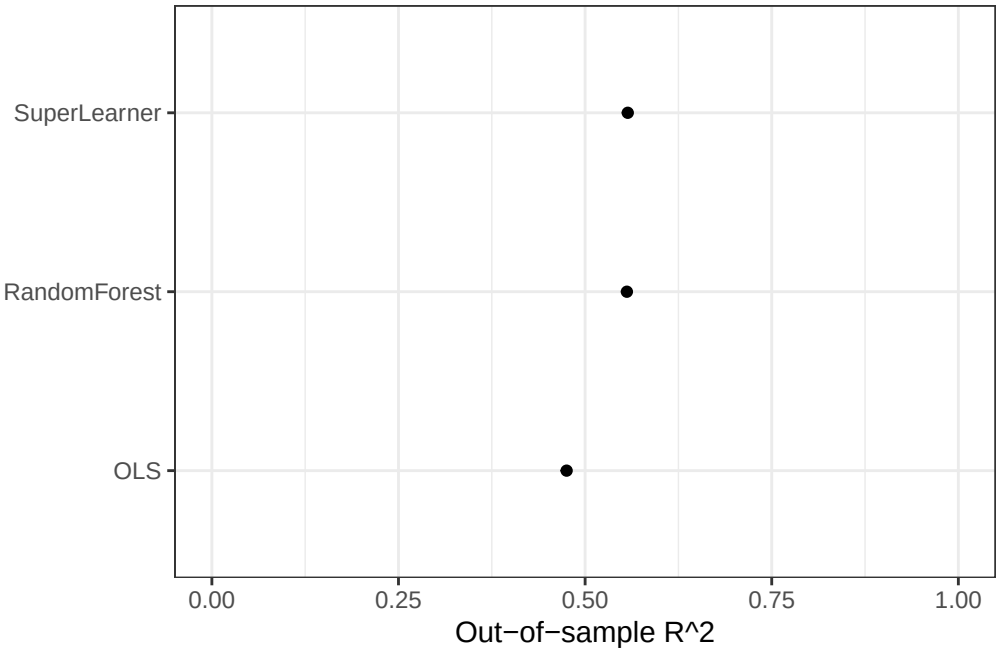
- 定式化依存問題
  - Statistics as a Science

- Let's Take the Con Out (Leamer 1983)
- Two culutures (Breiman 2001) (10 年経って)

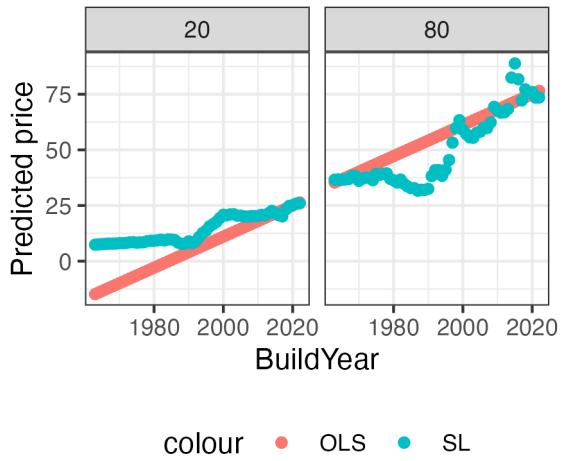
本講義の推奨

- 分析のゴールを明確に定め、それに適した手法を用いる
  - 「とりあえず統計モデルを推定する」をやめる
- 予測問題: 教師付き学習 (DataAdaptive な推定)
- 母集団の推論: Semiparametric 推定 with 教師付き学習

例: 中古マンション取引価格予測



例: 中古マンション取引価格予測



教師付き学習概論

機械学習 (MachineLearning)

- “統計学” とは異なるルーツを持つ手法群: 大きく
  - 教師付き学習 (SupervisedLearning); 教師なし学習 (UnSupervisedLearning); 強化学習 (ReinforcementLearning)
- 本講義では教師付き学習とその応用を紹介
- おすすめ参考書
  - Introduction to Statistical Learning (with R code)
  - Understanding Machine Learning: From Theory to Algorithms

教師付き学習

- データ $Y, X = [X_1, .., X_L]$ から, 条件付き母分布の関数に最も fit する関数 $f(X)$ を推定 (学習、Fitting、近似)
- 典型的には母平均に fit する関数の学習を目指す

$$\min E[(\mu_Y(X) - f(X))^2]$$

- $\mu_Y(X) = E[Y|X]$

- 注: 最も一般的な課題設定は DensityFunction ($Pr[y \leq Y|X]$) への Fit だが、依然として難しい

ポイント: 教師付き学習

- 条件付き平均関数を推定する” 伝統的手法” は多数存在
  - Parametric 推定 (OLS、最尤法、ベイズ): 母集団を有限個のパラメータで**正しく記述できるとして**、推定
  - 伝統的 Nonparametric 推定 (KernelRegression): 大量の変数を取り扱えない
- 教師付き学習の利点: 大量の変数を扱う応用についても、統計モデルの定式化に強く依存せずに、母分布を近似可能
  - Data-adaptive にモデルを設定・推定

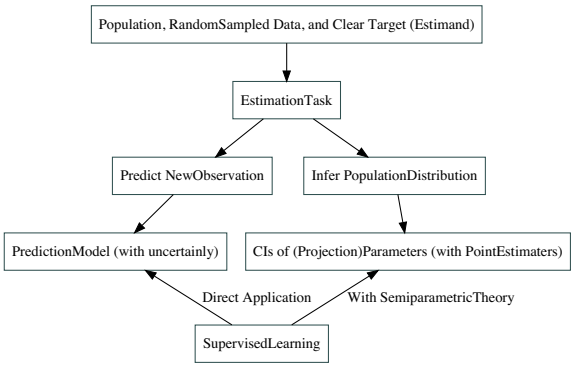
ポイント: 教師付き学習の応用

- 予測問題には、極めて有効
- 工夫すれば、母集団の特徴（要約）を適切に推定できる
  - 伝統的な Nonparametric 推定と同様に、収束速度が遅く、信頼区間が近似計算できない
  - Semiparametric 推定を応用
- 母分布そのものを” 適切” に推定することは依然として困難

データ分析概論

- 「分析のゴールの明確にし、推定手法をカスタマイズ」が**より**重要

データ分析のゴール



データ分析への具体的な要求

- 予測問題: $E[(\mu_Y(X_1, .., X_L) - f(X_1, .., X_L))^2]$ を最小化
  - 2 次的要求: 予測モデルの説明可能性、SamplingUncertainly の定量化
- 母集団の特徴の推定: 明確に解釈 **された**母集団の特徴について、意味のある信頼区間を形成
  - $\sqrt{N}$ 正規性 (Asymptotic Normality)
  - 漸近効率性

古典的推定

- 母集団の推定は、古典的な推定手法においても、中心的課題
  - “Parametric な統計モデルをデータを見る前に設定し、推定する”

$$E[Y|X] = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$$

- **推定された統計モデルを用いて**、研究課題に答える
- Estimand の定式化, 識別・推定上の仮定を**全て**統計モデル上で議論する

理想の世界

- 母集団を、単純なモデルで記述でき、かつ大部分は既知
- 例: 母平均 $\mu_Y(D, X)$ は以下に従うことを知っている

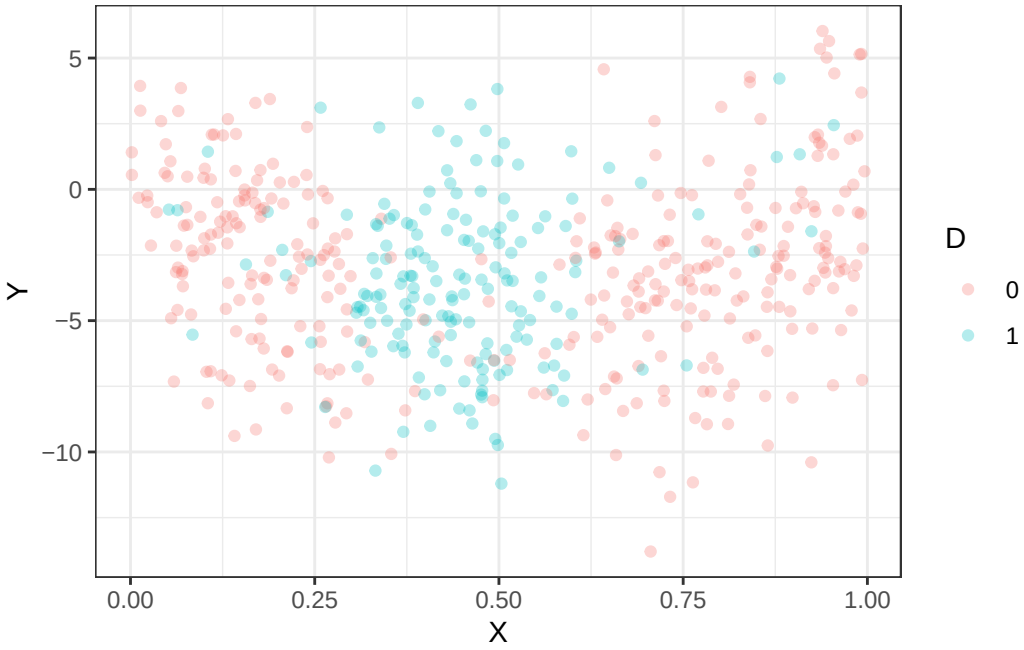
$$\mu_Y(X_1, X_2) = \beta_0 + \beta_D D + \beta_1 X + \beta_2 X^2$$

- β の値は未知

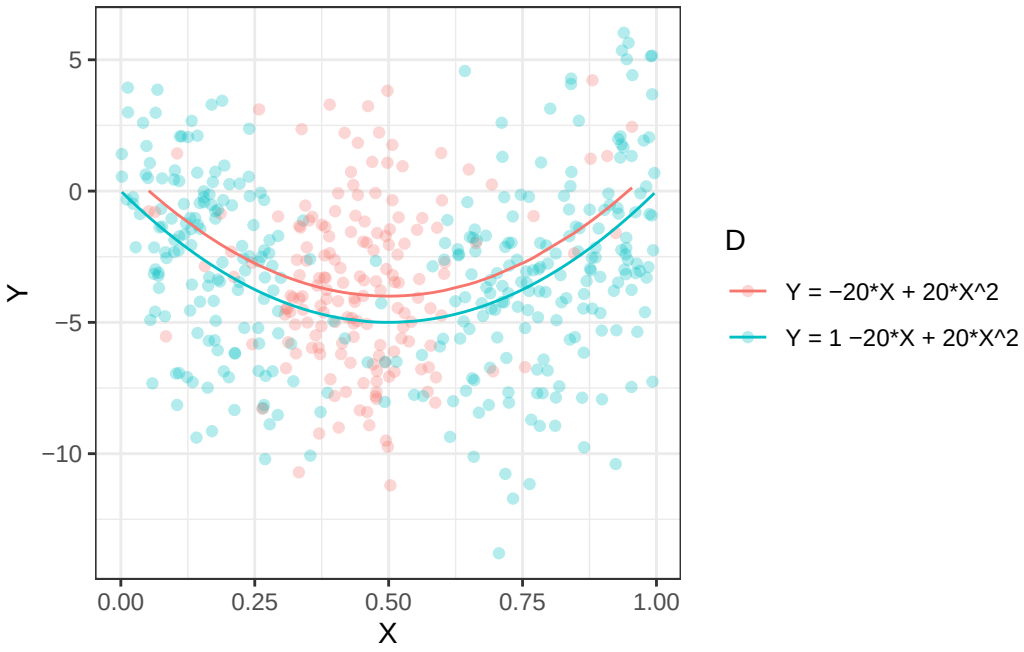
理想の世界

- 十分なサンプルサイズがあり、**モデルがただしければ**、OLS で $\mu_Y(D, X)$ を**高い**精度で推定できる
  - 誤差項に ParametricAssumption を追加できれば、最尤法|ベイズ推定も可能
- 推定された関数 $f(D, X)$ は、
  - $E[(\mu_Y(D, X) - f(D, X))^2]$ を実用可能な水準まで削減 (予測達成)
  - β について信頼区間計算が可能 (パラメータ推定達成)
  - そもそも単純なので、人間がそのまま理解できる (記述達成)

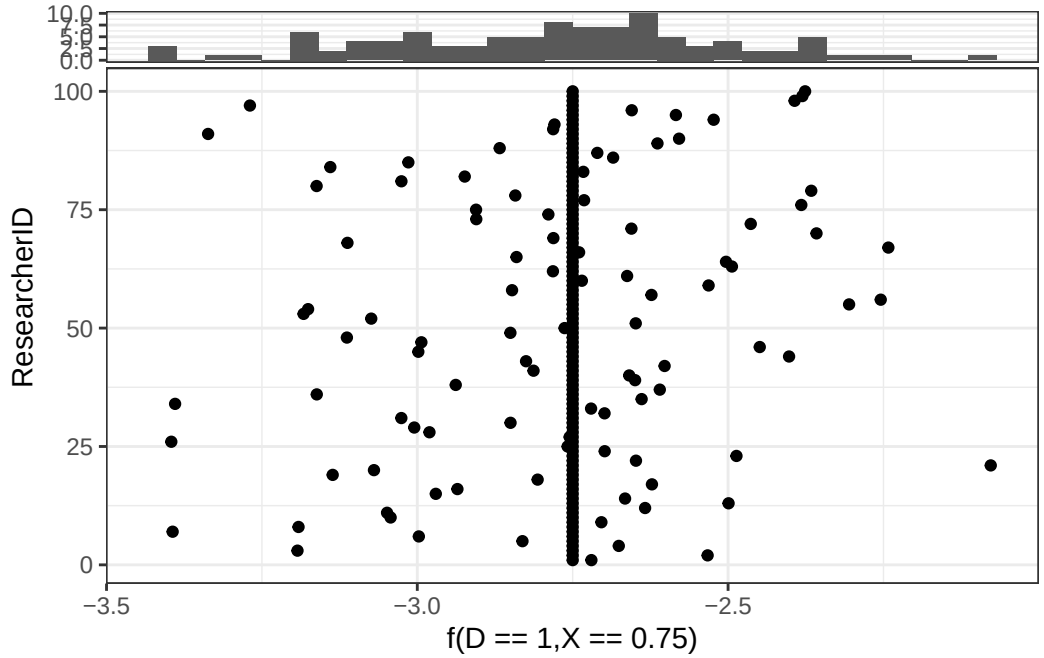
数値例: データ



数値例: 母平均



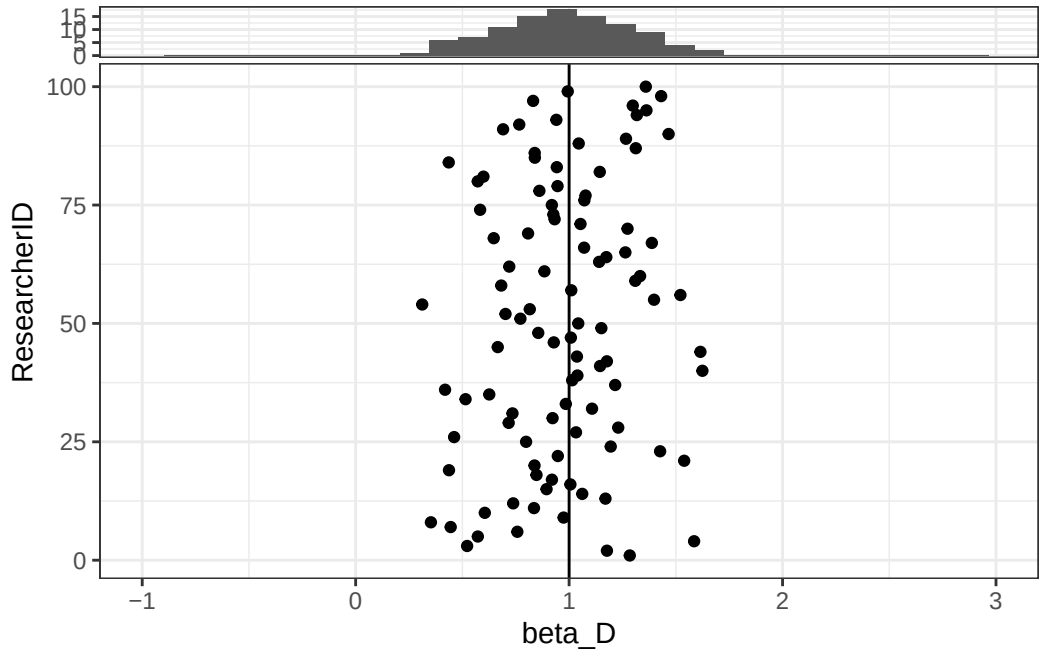
数値例: 予測



Sampling Uncertainly

- 確率的に選ばれた、母集団の一部を観察
  - 同じ手法・母集団・研究課題に挑む研究者であったとしても、異なる結論が出てくる
  - たくさんの工夫

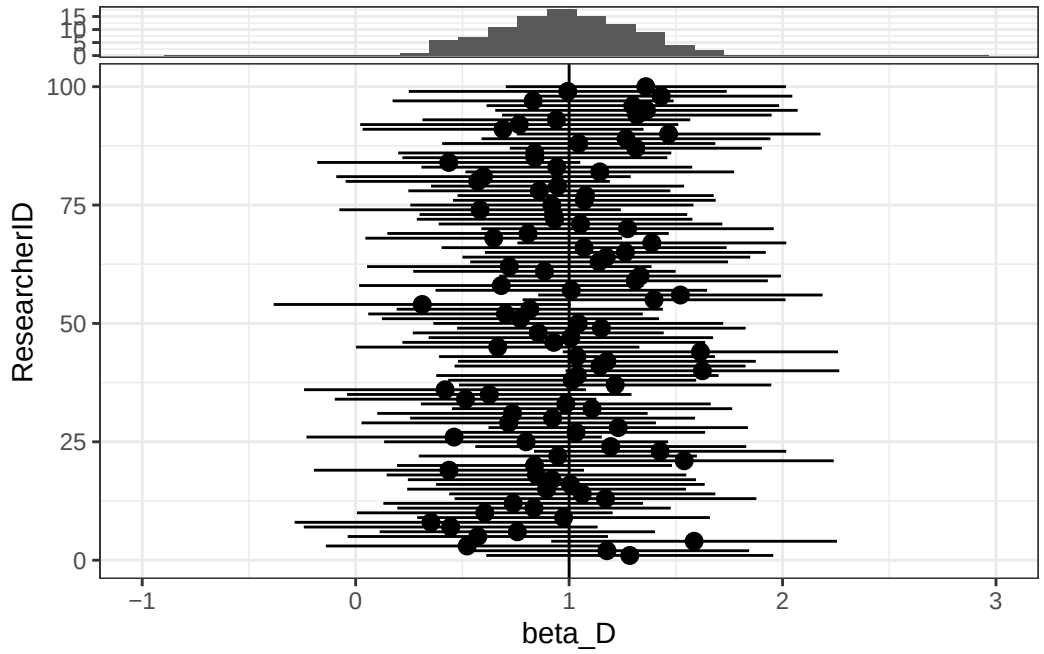
数値例: $beta_1$



漸近性質の活用

- ”サンプルサイズが大きくなれば、推定値は真の値の近くに分布する”性質を活用
- 一致性: 無限大に大きくなれば、全員真の値に収束
  - 経済学的では非実用的
- 漸近正規性: より早い速度で、正規分布に収束
  - 信頼区間の計算が可能

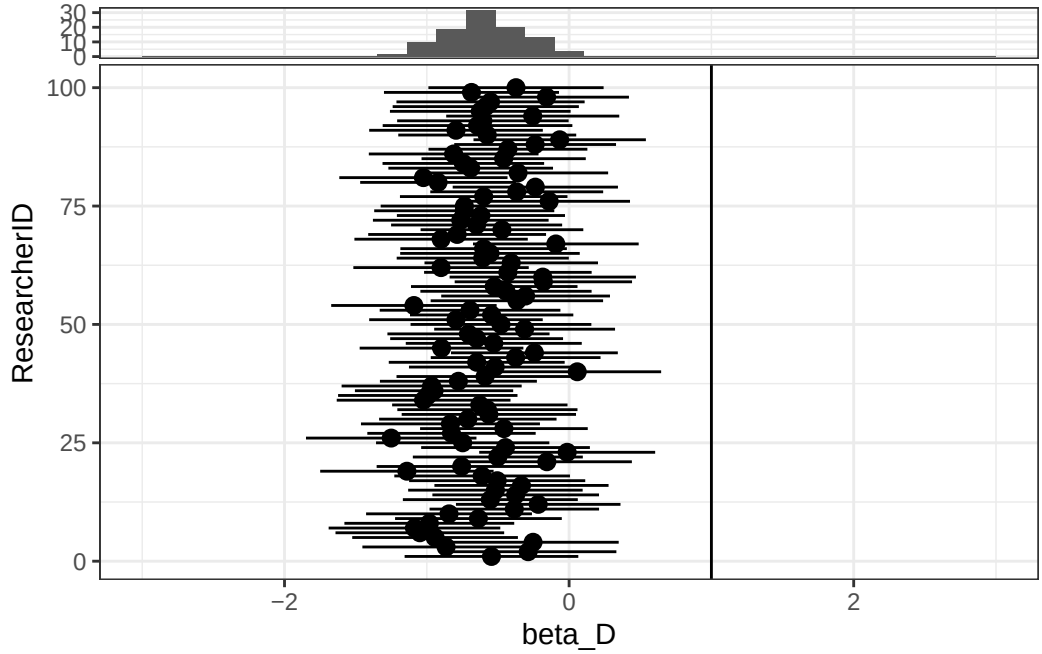
数値例: $beta_1$



誤定式

- 経済学のほぼ全ての応用で、正しいモデルを設定することは不可能
  - 任意の β について、$\mu_Y(X) \neq f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- 一般に予測性能が悪化し、母集団についての信頼区間が” 信頼できなくなる”

数値例: $beta_1$



誤定式化の避け方

- モデルを柔軟にする (推定するパラメータを増やす) と誤定式のリスクは必ず減る

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_i + .. + \beta_L X_i^L + u_i$$

- 過剰適合が生じ、推定が" できなくなる"
- 教師付き学習の代表的アイディア: モデルを適切に単純化する
  - 変数を"Shrink, Chop, and Throw out!!!"

応用: 母集団の推定

- 母集団の推定にも応用可能だが、工夫が必要

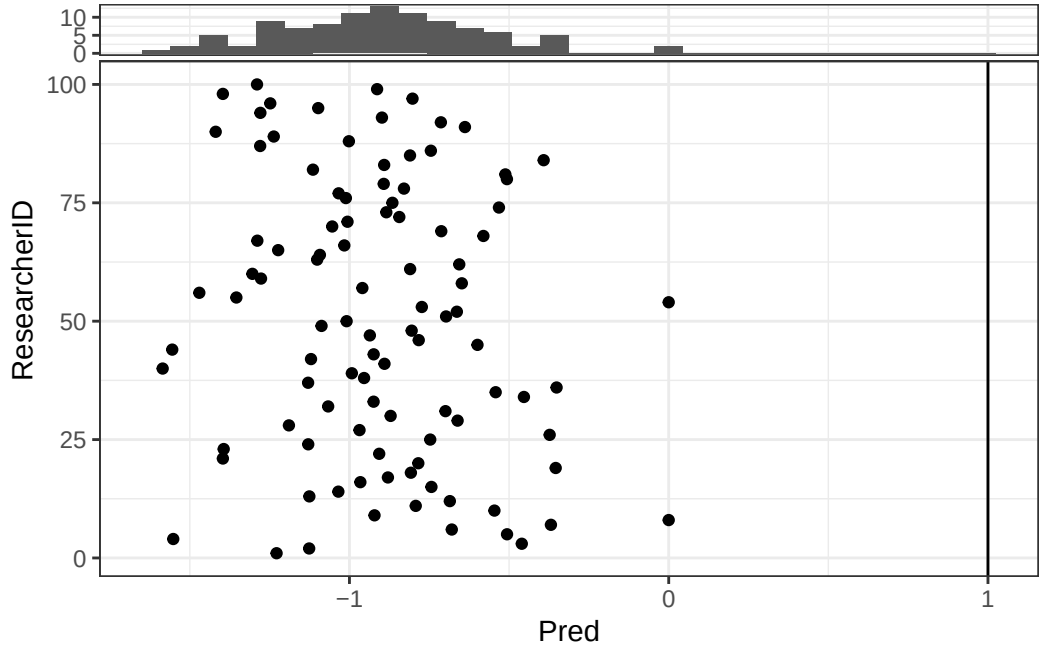
$$\tau = E[\mu_Y(1, X) - \mu_Y(0, X)]$$

$$= f(D = 1, X) - f(D = 0, X)$$

として推定

- 収束度が遅く、信頼区間の近似計算ができない
  - Variance を低下させるために、Bias を導入しているため

数値例: Naive Method



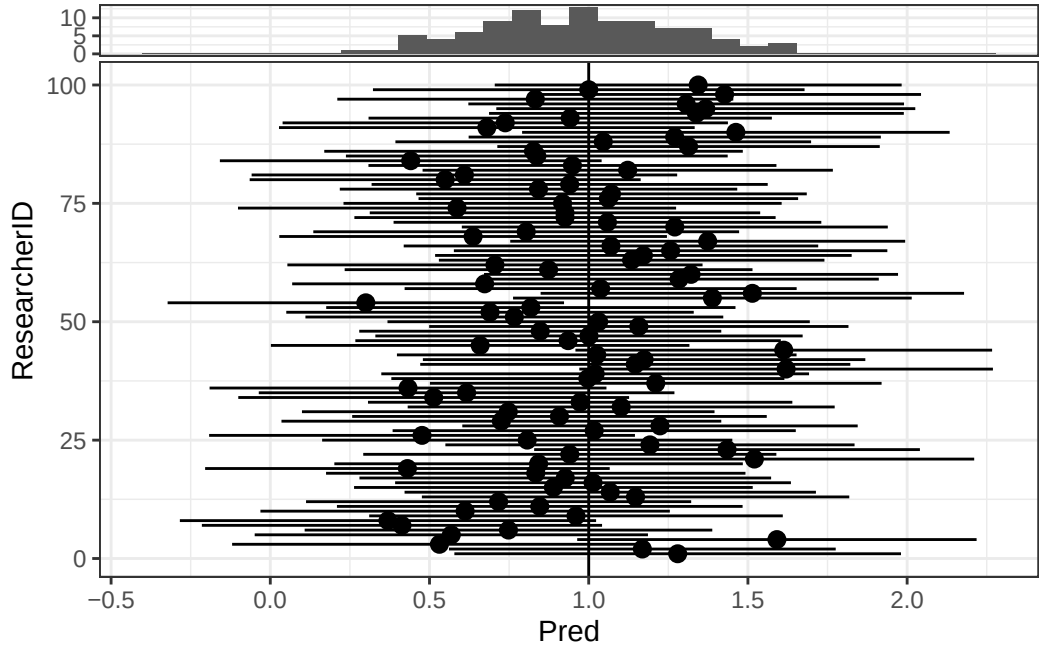
応用: SemiParametric 推定

- SemiParametric 推定の手法を応用することで、収束の遅さを保管する (TargetedLearning | DebiasedMachineLearning) 。
- Score の設定: m の設定;

$$0 = E[m(Y_i, D_i, X_i, g, \tau)]$$

- g : 未知の Nuisance 関数 (例: 条件付き母平均、傾向スコア)
- m を g の推定誤差について Robust な関数として設定すれば、g の推定に機械学習を応用可能

数値例: Semipara



応用: 経済学

- 経済学における” 典型的” な実証課題は、**特定の解釈ができる**母集団の特徴を推定する
  - 識別: 特定の解釈ができるかどうかを議論
  - データのみから判断不可能|困難

応用: 差別

- 例: 母集団における「WindowsOS User と Mac|LinuxOS User との間での賃金差」は、両 OS Users 間での差別と解釈できるか？
  - 差別 = 規範的判断を、データのみからは判断不可能|困難 (ヒュームの法則)

応用: 因果効果

- 例: 母集団における「WindowsOS User と Mac|LinuxOS User との間での賃金差」は、使用する OS からの因果効果と言えるか？
  - OS 間での観察できない要因も揃えた上で比較したい

まとめ

- 「事前に正しい統計モデルを設定する」推定法には、多くの問題点
  - 仮定が見にくい
  - 細かい定式化に推定結果が依存
- 教師付き学習の活用: 推定目標を明示的に定めれば、非常に有益
  - 予測問題: 多くの優れたアルゴリズム
  - 母集団の推論: セミラパ推定 + 既存アルゴリズム
- 因果効果の識別 (事前解釈) は、別問題

まとめ

- 「事前に正しい統計モデルを設定する」推定法には、多くの問題点
  - 仮定が見にくい
  - 細かい定式化に推定結果が依存
- 教師付き学習の活用: 推定目標を明示的に定めれば、非常に有益
  - 予測問題: 多くの優れたアルゴリズム
  - 母集団の推論: セミラパ推定 + 既存アルゴリズム
- 因果効果の識別 (事前解釈) は、別問題

混乱した議論

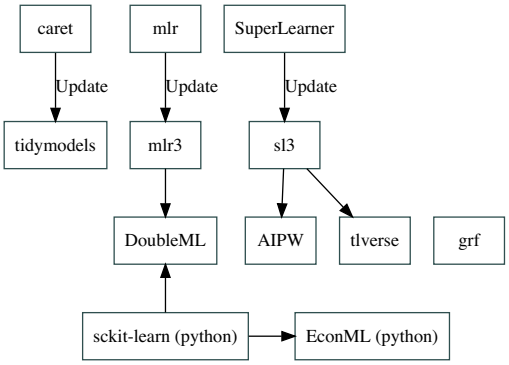
- 機械学習 | 傾向スコアを使えば、因果効果を**識別**できる
- 機械学習を使っても因果効果は**識別**できないので、因果効果の**推定**には役に立たない

RoadMap

- 代表的教師付き学習アルゴリズムの紹介:
  - Tree/LinearModel 系
  - Stacking (実用上の推奨)

- 条件付き平均差の推定
  - Double/DebiasedMachineLearning
- 他の応用
  - GeneralizedPartialLinearModel
  - SensitivityAnalysis
  - NonParametricPrediction

Core Packages: Semipara + ML



次回に向けた準備

- R と RStudio とパッケージのインストール
  - R と Rstudio のインストール
  - Packages のインストール
  - ProjectFolder の作成
  - Data のアップロード

- 必要パッケージ: mlr3verse; tidyverse; rpart.plot

Reference

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Leamer, Edward E. 1983. “Let’s Take the Con Out of Econometrics.” *The American Economic Review* 73 (1): 31–43. https://www.jstor.org/stable/1803924.