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 ( 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

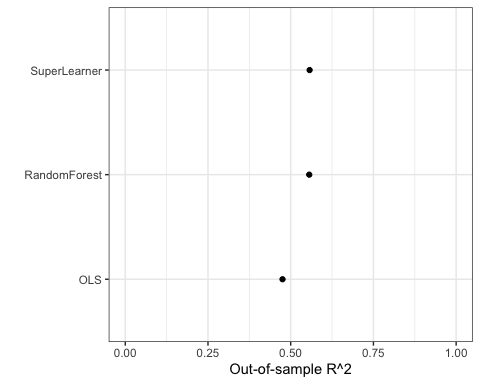
## Motivation

* 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
* 定式化依存問題
  + [Statistics as a Science](https://magazine.amstat.org/blog/2015/02/01/statscience_feb2015/)
  + Let’s Take the Con Out (Leamer 1983)
  + Two culutures (Breiman 2001) ([10年経って](https://muse.jhu.edu/issue/45147))

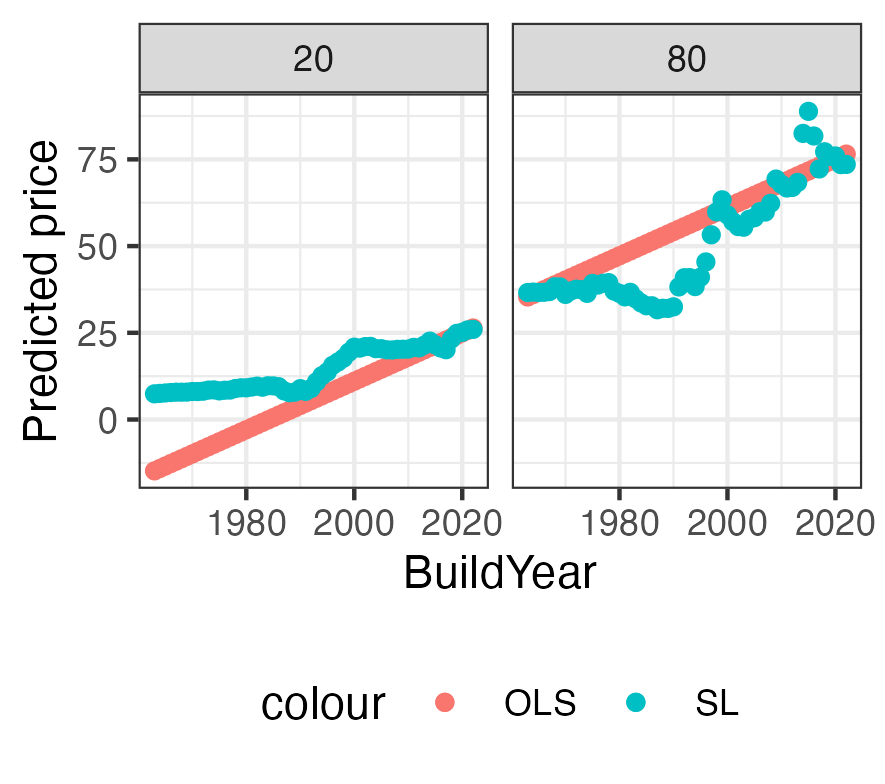
## 本講義の推奨

* 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)](https://www.statlearning.com/)
  + [Understanding Machine Learning: From Theory to Algorithms](https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html)

## 教師付き学習

* データ から, 条件付き母分布の関数に最もfitする関数 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
* 注: 最も一般的な課題設定はDensityFunction 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

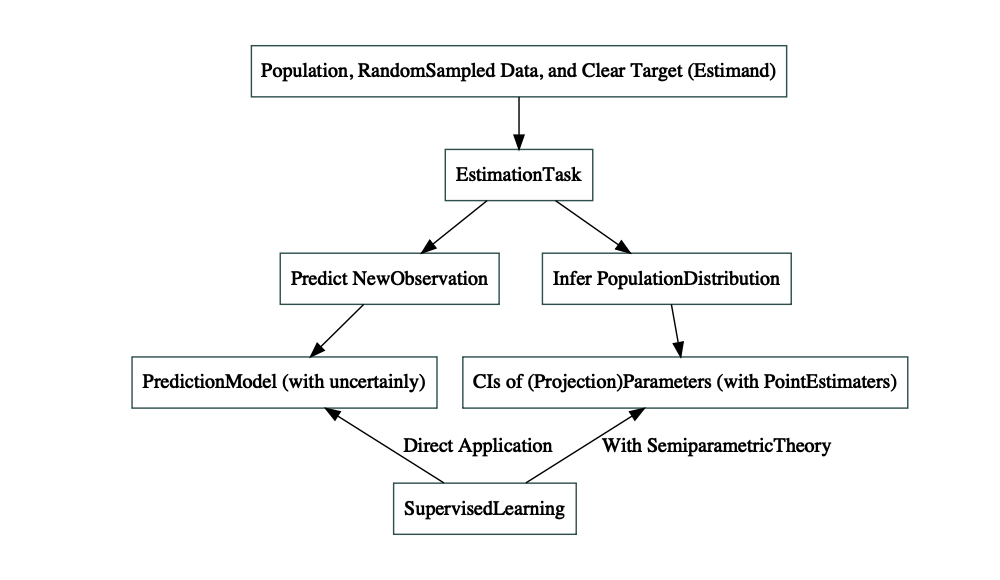
## ポイント: 教師付き学習の応用

* 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: 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
  + 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”
* **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

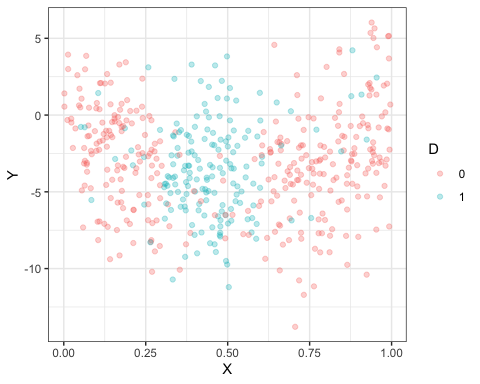
## 理想の世界

* 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 は以下に従うことを知っている
* の値は未知

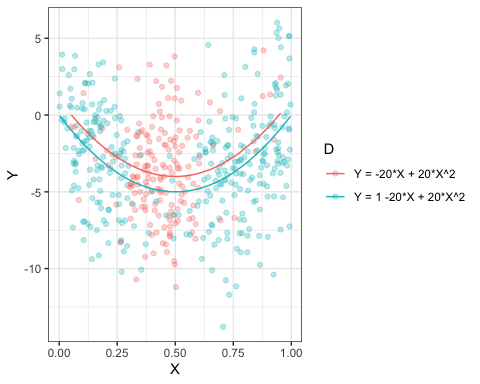
## 理想の世界

* 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 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 は、
  + を実用可能な水準まで削減 (予測達成)
  + 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)

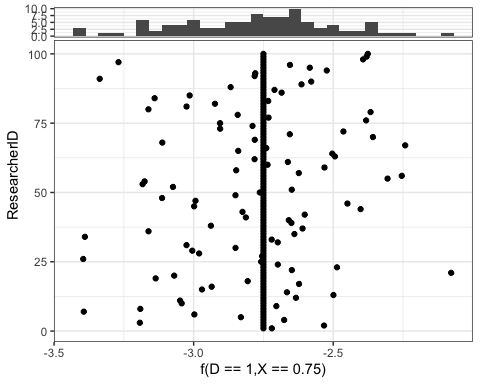
## 数値例: データ



## 数値例: 母平均



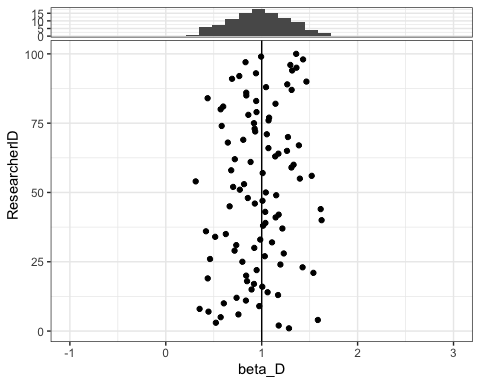
## 数値例: 予測



## 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

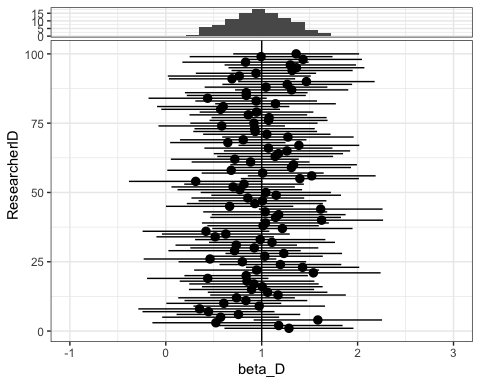
## 数値例:



## 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

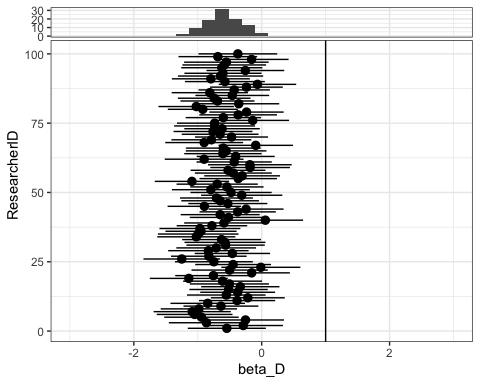
## 数値例:



## 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 について、
* 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:



## 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
* 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!!!”

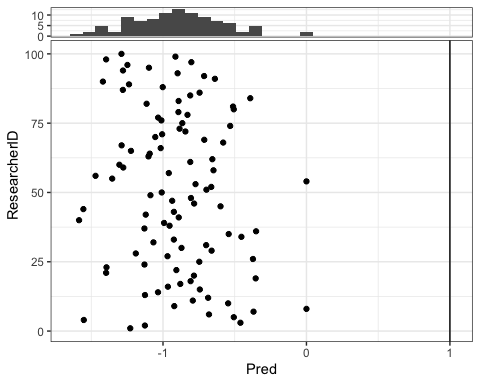
## 応用: 母集団の推定

* 母集団の推定にも応用可能だが、工夫が必要
* 例えば、 を

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

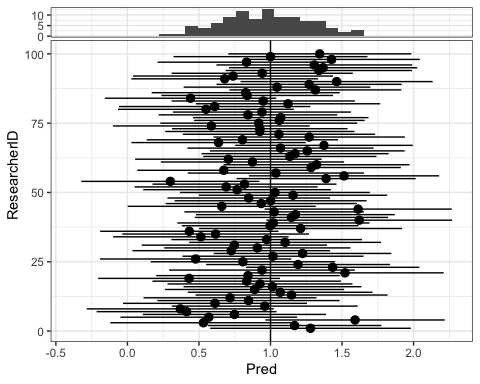
## 数値例: Naive Method



## 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: の設定;
* : 未知のNuisance関数 (例: 条件付き母平均、 傾向スコア)
* を の推定誤差についてRobustな関数として設定すれば、 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

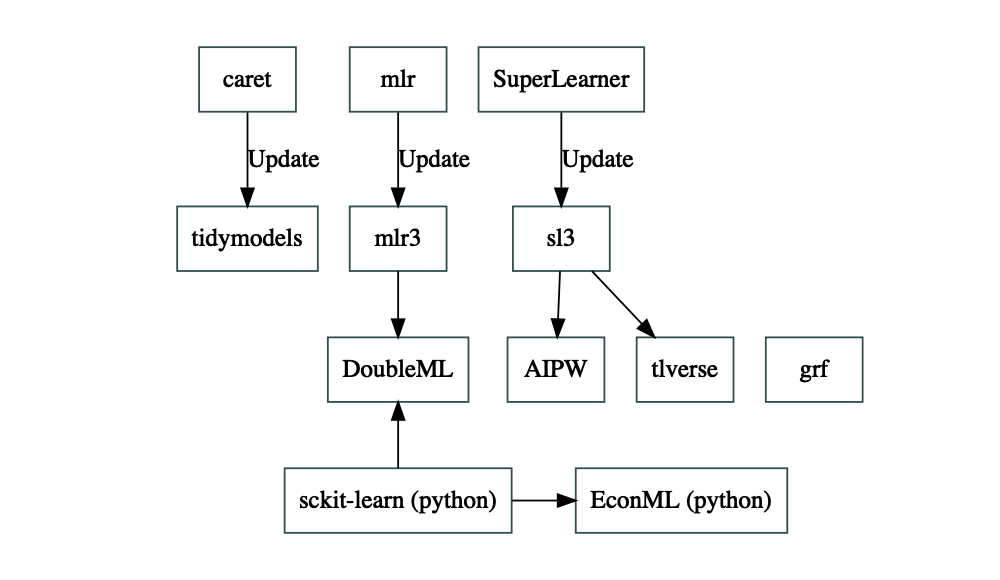
## 混乱した議論

* 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](https://arxiv.org/abs/1608.00060)
* 他の応用
  + [GeneralizedPartialLinearModel](https://arxiv.org/abs/2006.08402)
  + [SensitivityAnalysis](https://www.nber.org/papers/w30302)
  + [NonParametricPrediction](https://arxiv.org/abs/1610.01271https://arxiv.org/abs/1610.01271)

## 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](https://youtu.be/fDlXx8e5W78)
  + [Packagesのインストール](https://youtu.be/2fmThveX7_s)
  + [ProjectFolderの作成](https://youtu.be/rqQP4jsF0oQ)
  + [Dataのアップロード](https://youtu.be/f2EU44WFyQM)
* 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>.