序論

経済学各論 (機械学習)

川田恵介

- 講義資料: https://github.com/tetokawata/UnderGradEconML

1. データ分析周りのキーワード
2. “事例” から学ぶ難しさ
    - 機械学習でも、インタビュー調査でも、歴史分析でも直面
3. 適切に単純化されたモデルによる解決
    - 機械学習・計量経済学・統計学のきも
4. 講義の概要
5. 次回への準備

データ分析

- 不完全な事例集（経験、歴史、データ等）から人間が学ぶ方法
    - 経済学では、「“意思決定” に生かす知見」を伝統的に重視

- データ分析への注目はますます高まる
    - AI 人材
    - リスキリングにおける人気項目
- 学際的発展
    - 経済学　経営学　金融 (工) 学　生物・医学　政治学　社会学　統計学　計算機科学
    - 大学　企業　公的機関

機械学習

- “統計学” (計量経済学の土台) とは異なるルーツ (AI の開発) を有するデータ分析方法
  - 今では中核技術
- 様々な” バズ” 技術に活用
  - ChatGPT, AlphaGo などなど

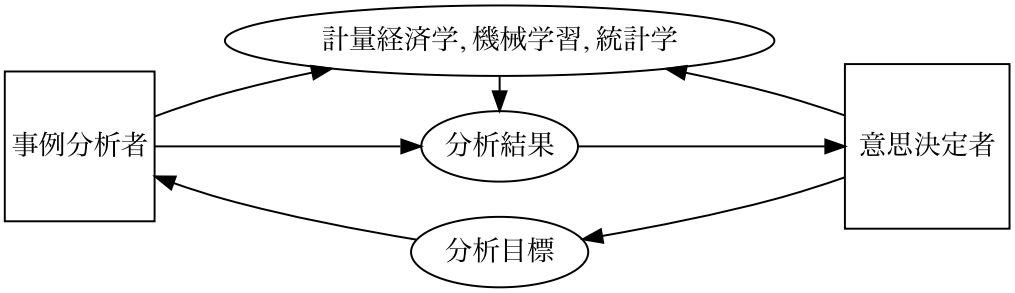
経済学部生と機械学習

- 志望キャリアをと会わず、経済学部生が知っておくべき教養となりつつある
  - データ分析自体はもちろん、正しく分析結果を活用する・“規制” する、ためにも重要
- 日本企業内経済学者
- 日本経済センター: 研修制度

機械学習 + (計量) 経済学

- 伝統的な計量経済学との融合が進む
- 特に重要なフィールドの一つ
- 例: マイクロソフトが進めるプロジェクト
  - EconML

イメージ図：ゴール像



例: 需要予測

- 過去の販売事例から、店舗レベルでの需要予測の精度が改善
  - 意思決定者が、物流・発注システムの改革も行うことで、食品破棄・売り逃しを減らせる
- 個人レベルでの予測精度が” 大幅” 改善
  - 意思決定者は、まったく新しい通販サービスの提供できる?
  - 注文を受ける” 前” に、予測された商品を発送、キャンセルしなければ料金を支払う (“注文 → 発注” から “発注 → 注文” へ)
- 予測マシンの世紀　AI が駆動する新たな経済

他のワード

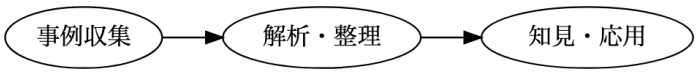
- 強化学習、深層学習 (Deep Learning), Generative Model = 機械学習の一手法
  - 本講義では教師付き学習を学び、その中で深層学習,Generative Mode にも触れる
- ビッグデータ：” 大きなデータ”
  - 元の定義は一つのコンピュータでは処理できないほど大きなデータ
- データサイエンス：厳密な定義はない
- ChatGPT 教師付き学習や強化学習を組みあせて、複雑な予測結果を表示 (Generative model)

事例から学ぶ技術

- 他者の事例・自身の経験・歴史を意思決定にどう生かすか？
  - 普遍的な課題
- （例）経験豊富な指導者が
  -「練習中の水分補給を禁止」
  -「筋肉を増やすために、超長距離遠泳を指示」
- アドバイスを聞くべきなのか、それとも” 老害” なのか？
- 意思決定の際に参照すべき事例はどれか、“偶然生じた例外的” 事例を間違えて採用してないか？

– 「データから観察できない要素が大量に存在する」、社会科学・実務において深刻な課題

イメージ図



例: 練馬 1L の取引価格は？

立地	部屋	取引価格（100 万円）
練馬	1DL	100
板橋	1L	10
板橋	1DL	8
板橋	1DL	12
板橋	1DL	20
板橋	1DL	30

課題

- “全く同じ” 事例が存在しない・極めて少数しか存在しない
  - なんとなくよく似た事例を参考にする？
- 一見同じに見える事例においても、“矛盾” が存在
  - なんとなく多数派の事例を採用？
- 例外的に見える事例が存在
  - なんとなく削除する？
- 事例集のもつ” 不完全性” にどのように対応するか？

  - 社会・経済データへの応用において普遍的な問題

データ分析

- ”大規模”な事例集 (= データ）から、検証可能な形で学ぶ方法
  - 業務の電子化により、容易に蓄積可能
  - ただし課題はそのまま
  - 多くの解決策が提案
- PC の性能改善により、さまざまな分析手法が実行可能
  - “誰にでも”使いやすいプログラムの開発
  - 無料の R や Python に多く実装

データのイメージ

- 整理 (tidy) されたデータ:

現状の問題点

- 事例集が巨大化: より多くの事例について、より多くの情報を取得できる
- 以前として、“誤解”が多い
  -「ボタンの掛け違い」
- 高校までの授業では、あまり重点を置かれていない”枠組み”への理解が必要
  - 人 (分析者) によって結果が違う
  - 確率的事象

“合意”の枠組み

- 最終的な目標は、“有益かつ、合意できる示唆を得る”
- “綺麗”な世界の現象については、厳密な合意が可能
  - 理科室の実験では、同じ結果を得られる
- 現実の世界では、厳密な合意は不可能
  - 同じやり方で収集したデータを、同じやり方で分析したとしても、結果は人によって異なる

  - データが異なるため

機械学習

- 統計学/計量経済学とは異なるルーツ (**AI** の開発) をもつデータ分析手法
  - 教師付き学習、教師なし学習、強化学習、等々
- 計量経済学とよく似た問題意識
  - 方言が大きく異なる
- 発展するにつれて、当初の目的を超えた価値を持つ
  - 特に変数が多い/複雑なデータ（Rich data）を用いた、予測問題で威力を発揮
  - 計量経済学や医療・生物統計との融合により、因果効果や格差推定においても有益

予測問題とは

- 事前に観察できる情報 X から、Y を予測する
  - 中古車買取マニュアル作り：ある車 X がいくら $= Y$ で再販できる？
  - 動画の suggestion システム：あるユーザー $= X$ がある動画をどのように” 評価” $= Y$ するか？
  - 退学、留年する可能性がある学生へのケア：ある学生が退学してしまう確率は？
  - ” マクロ” 政策：将来の人口、景気は?
- 機械学習が大きな比較優位をもつ

学習とは

- 煩雑な経験 (ヒアリング結果、歴史、個人的経験など) -> 学習-> 知見
  - 知見の一つ = 予測

素朴な方法

- よくある主張
  - “「モデルや理論」を用いずに、現実の事例をしっかりみるべき”
- 最も naive な学習法 = 事例の丸暗記 (learning by memorization)

  - 予測したい対象と同じ事例を思い出して、その結果を予測値とする
  - 全く同じ事例がなければ、“最も近い” 事例を参照
  - 予測したい事象が単純かつデータが十分あれば、機能しうる

丸暗記法が有効なケース

- 「限られた情報から整合的な結果」を生み出すように” 設計” されているのであれば有効
- 例
  - 判例
  - 同じ企業内の賃金
- 就活?、チャットでの質問？

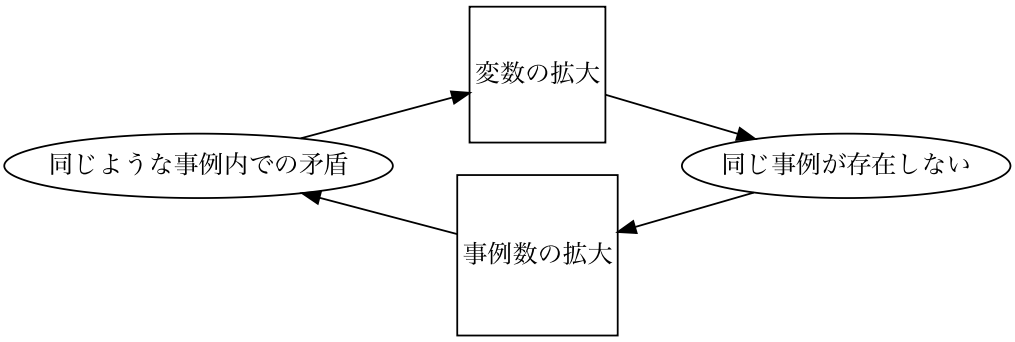
丸暗記のジレンマ

- 社会における事象 = 大量の要因が影響を与えている
  - データから観察できないものが多数
  - 本来的には、同じ事象は” 存在しない”
- 多くの属性を活用しようとすると、事例数が減少する

例

- 38 歳香川県出身男性の化粧水への需要を予測したい
  - 事例の中に川田が含まれている $\rightarrow$ 高い需要を予測
- 事例数が増えると、同じ属性の事例が増え、川田の影響は薄まるが
- 年齢・性別・出身地以外にも、予測において重要な属性もありそう
  - 活用する変数を増やす (+ 学歴 + 職業 + 年収 +...) と、同一事例が減り、また川田のみになる

丸暗記



丸暗記の限界

- 丸暗記法がうまくいく前提: 化粧水の消費量を決める重要な要因が全て事例集に記録され、かつ十分な事例数がある
  - ほとんどの応用事例で不可能
- 例: 一卵性の双子
  - 生まれた時点では” ほとんど同じ” 属性を持つが、、、

単純化 & 一般化

- 過去の事例がないケースについても、予測する必要がある
  - 一般化 (generalization)
- 適切な単純化が必要
  - 観察できない属性が大きく偏っている事例の影響を抑えるため
  - 人間も (無意識的) に行っている

近似モデル

- 近似モデルをデータから推定
  - 近似モデル：複雑な現実を適切に単純化したモデル
- フェルミ推定

> **!** Important
>
> - *Truth is much too complicated to allow anything but approximations*
>   - (John von Neumann)

トレードオフ

- “現実は複雑なのだから、複雑なモデルを用いるべきでは?”
  - 複雑な現実を捉えられる
  - データから推定することが難しくなる

講義概要

- 前提知識: 四則演算、基礎的な統計知識（平均や分散など）
- 講義資料: 講義スライド、実習用データ
  - Github repository (リンク)) からダウンロード可能
  - R の実装については、Github repository(リンク) も参照
- 実習環境: **Posit cloud**
  - インターネット上で作業できる (自宅からでも)
  - 関心に応じて、R と Rstudio を自身の PC にインストール

成績評価

- レポート (100%)
  - 授業期間中に 3 回実施（1）各手法が正しく使えているか、（2）結果を正しく解釈できているか

ゴール

- 機械学習（教師付き学習）のコンセプトを理解し、自身でデータに応用できる
- レポートを通じて、国土交通省が提供するデータを用いて、以下を実装するプログラム作成
  - 中古マンション取引価格を予測するモデル
  - 中古マンションを改装する平均効果の推定
  - 改装の効果を予測するモデル
- “履歴書” に、“機械学習の分析用コードを作成し、実際のデータ分析” を行ったことがある、と書けるようにする。

無料ソフトの利用

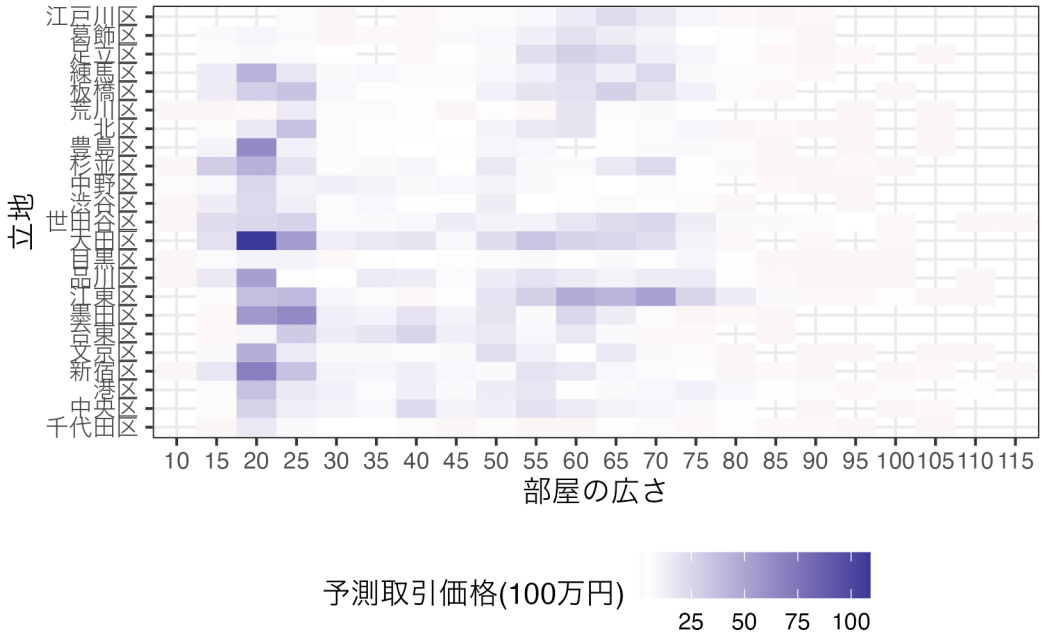
- R ないし Python を強く推奨
  - 企業、大学、公的研究機関の研究者が幅広く利用
- 本講義では R を使用
  - クラウド上の開発環境である Rcloud を併用

– ネットにさえ繋がれば、異なる PC で作業できる

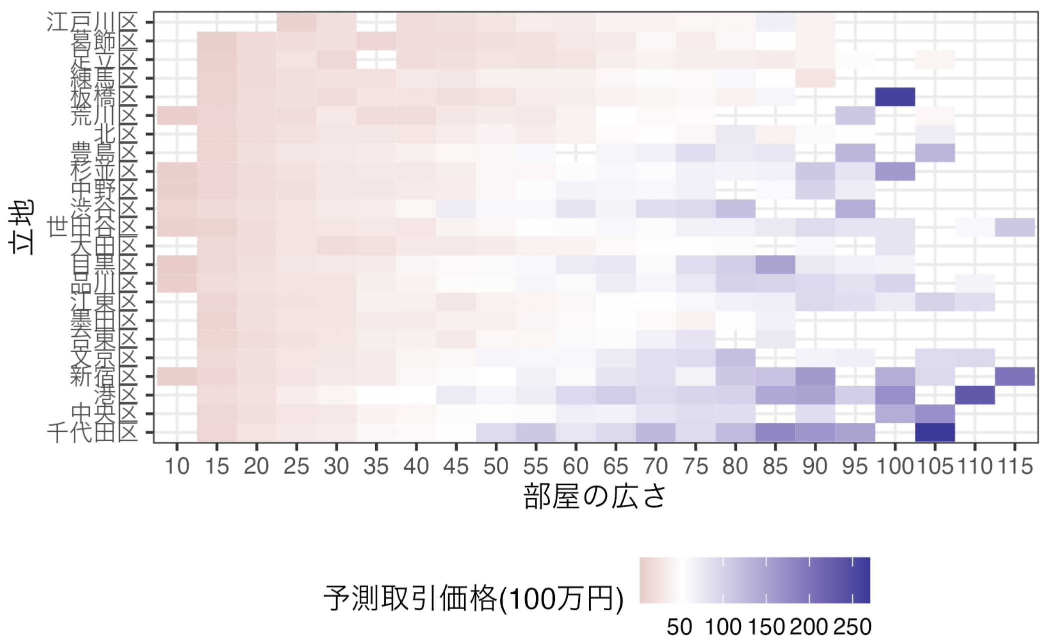
例

- 講義で用いるデータを使って、取引価格の予測モデルを構築

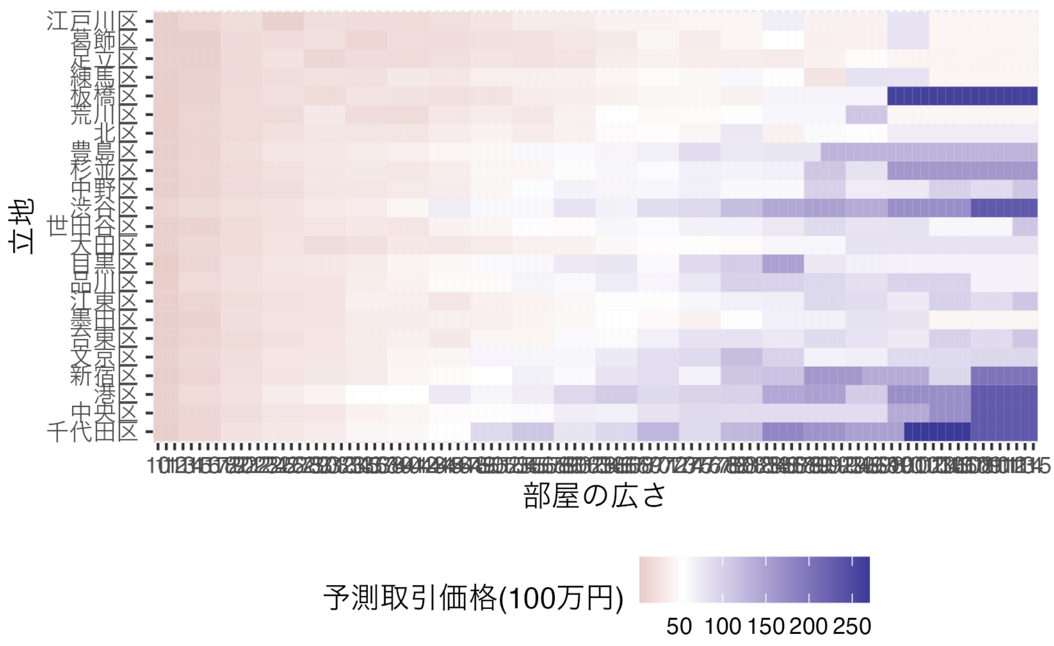
例: サンプルサイズ



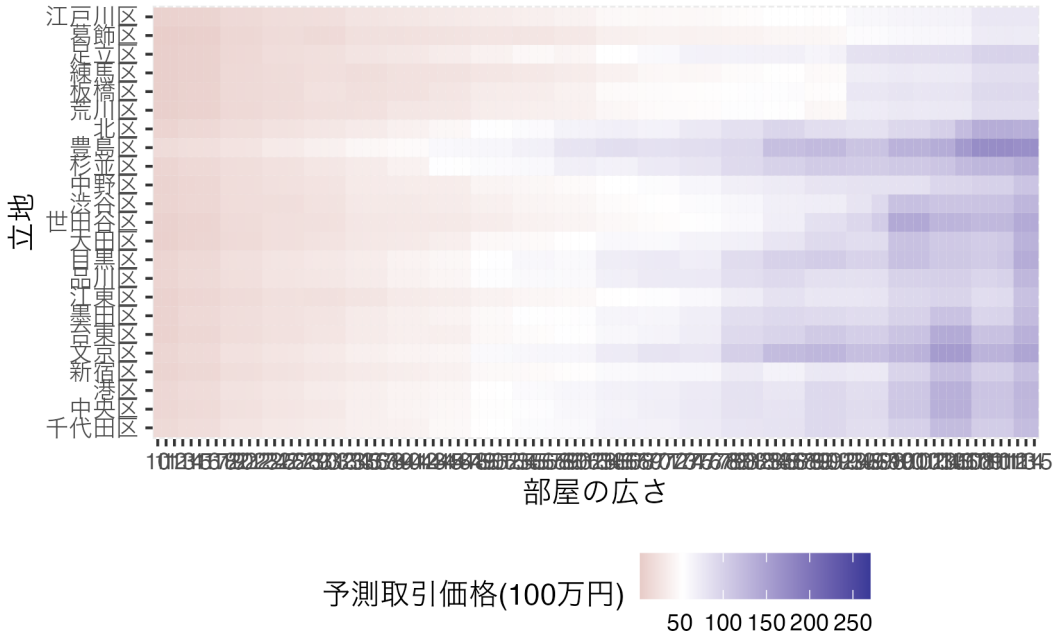
例: Learning by memoraization



例: Learning by memoraization もどき

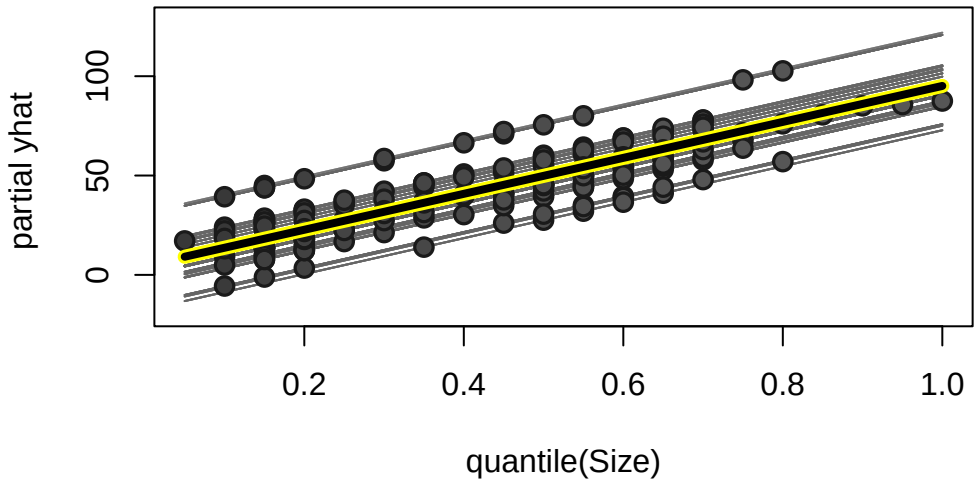


例: SuperLearner (170% **程度改善**)



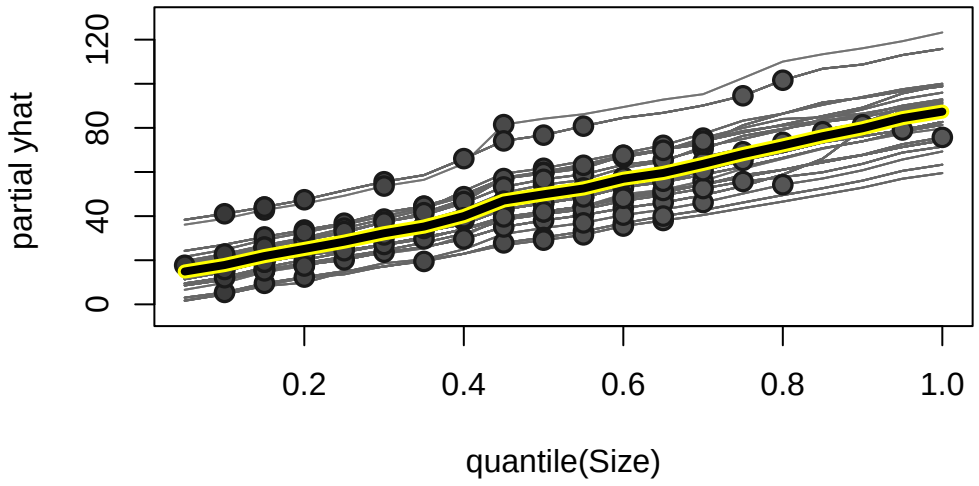
ICE: OLS

```
y not passed, so range_y is range of ice curves and sd_y is sd of predictions on real observations
```



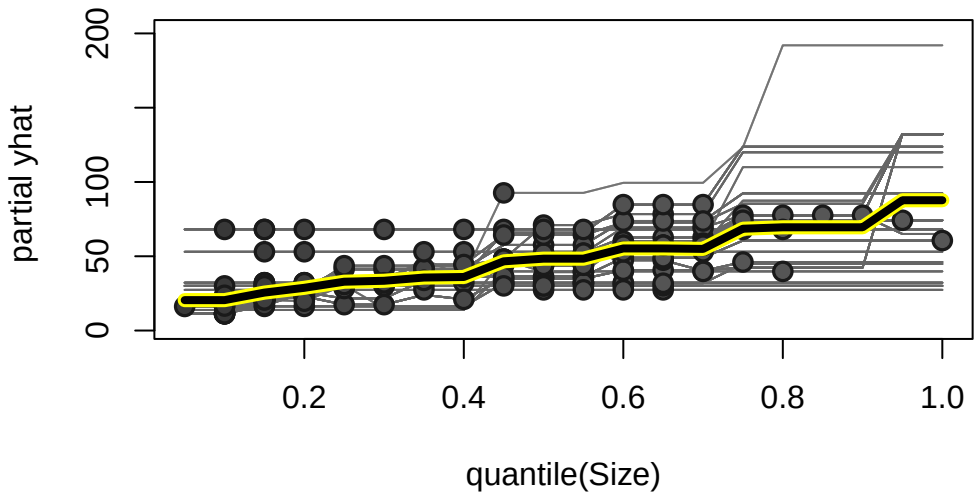
ICE: SuperLearner

```
y not passed, so range_y is range of ice curves and sd_y is sd of predictions on real observations
```



ICE: **丸暗記**

```
y not passed, so range_y is range of ice curves and sd_y is sd of predictions on real observations
```



次回までに

- Positcloud の設定 (Youtube へのリンク)
  - 時間がある人は本日中に