機械学習

川田恵介 (keisukekawata@iss.u-tokyo.ac.jp)

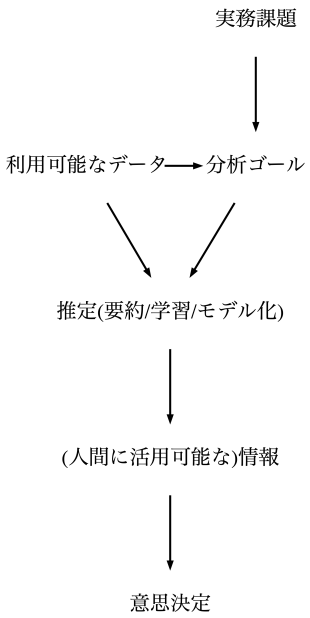
Table of contents

1 序論

1.1 データ分析への注目

- 研究者のみならず、実務者からの関心も高まっている
  - データを活用し、意思決定の参考となる” 情報” を得たい
- 例: ALICE, 日本経済センター研修事業, サーバーエージェント AIlab

1.2 データ活用の流れ



1.3 例

- 過去の診療データ → 患者の病状を予測
  - → 医者と患者による治療方針決定
- 利用者調査 → サービス選好の推移
  - → 企業によるサービス開発戦略決定
- 家計調査 → 日本社会全体での所得と支出推移
  - → 政府による政策/有権者による投票決定

1.4 到達目標

- 自身でデータ分析を行うための入門
- データ分析の結果を活用するために必要な知識の取得
  - 特に今後さらに増加することが予想される機械学習を用いた分析結果の活用
- 学生のうちに、機械学習 (AI) を用いたデータ分析を経験する

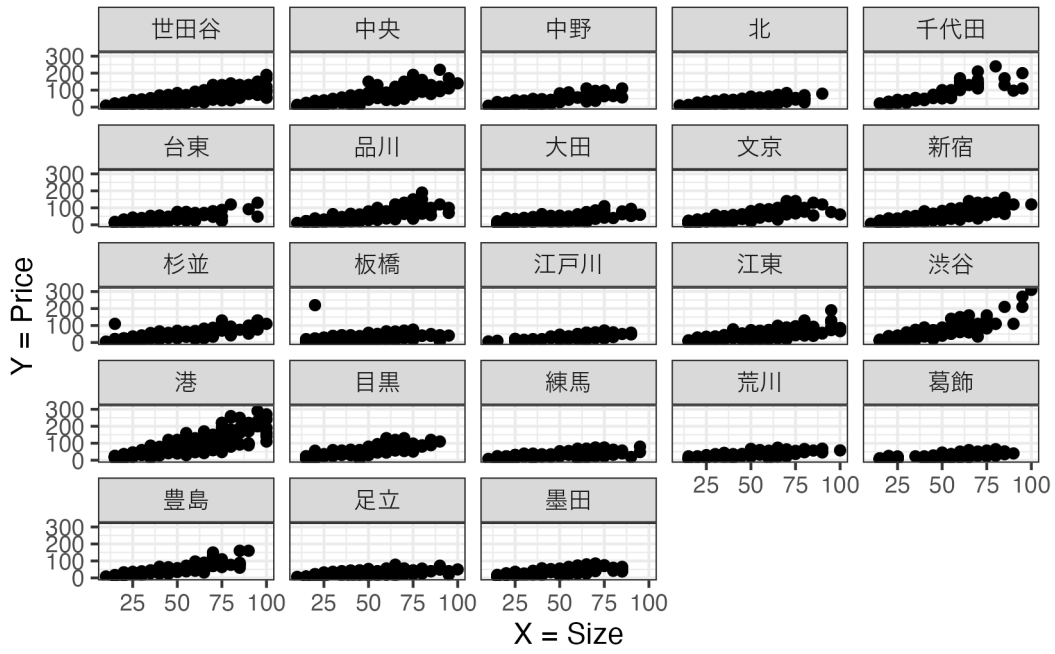
2 推定方法

2.1 データ分析法

- = 事例から学ぶ方法
- 過去の経験や事例、歴史（データ）を活用し, 意思決定に役立つ情報提供
  - **各**顧客は、どのようなサービスを好みのか?
  - 事業全体で価格を上げると、どの程度需要が下がるのか?
  - どのような領域に資源を集中すべきか?

2.2 散布図: 広さと立地と価格

- 乱雑であり、同じ X でも Y が異なる事例が多い



2.3 推定法

- 要約の方法として、大きく**統計学**と**機械学習/AI** がキーワードとして流布している
  - 伝統的な計量経済学は、統計学と密接
  - 機械学習は、統計学とは異なるルーツ (AI の開発) を持つ

2.4 機械学習

- 機械学習の一分野である**教師付き学習**は、伝統的な統計学と多くの議論を共有
  - OLS や logit は必ず紹介される/母分布を用いて議論を整理
  - 伝統的手法との融合が進む
- 新しい推定のアイディアを、データ分析が持ち込まれている
  - 伝統的な方法と比べて、よりデータ主導のモデル化を行う傾向がある

3 推定の例

- 東京 23 区の中古マンションにおける” 取引価格の特徴” について、2732 事例 (2023 年大三四半期) から含意を得る
  - どのような物件が高/低価格で取引される傾向があるのか?

3.1 “個別事例分析”

- 最も低価格の物件は、

```
# A tibble: 1 x 5
  Price District StationDistance  Size BuildYear
  <dbl> <chr>              <dbl> <dbl>     <dbl>
1   0.2 練馬                  15    15      1990
```

- 練馬の狭く古い物件が、最も安い傾向?
  - データを取りなおしたら、違う結果が出るのでは?
  - データから観察できない要因で、この事例の価格が下ぶれているだけでは?

3.2 サブサンプル分析

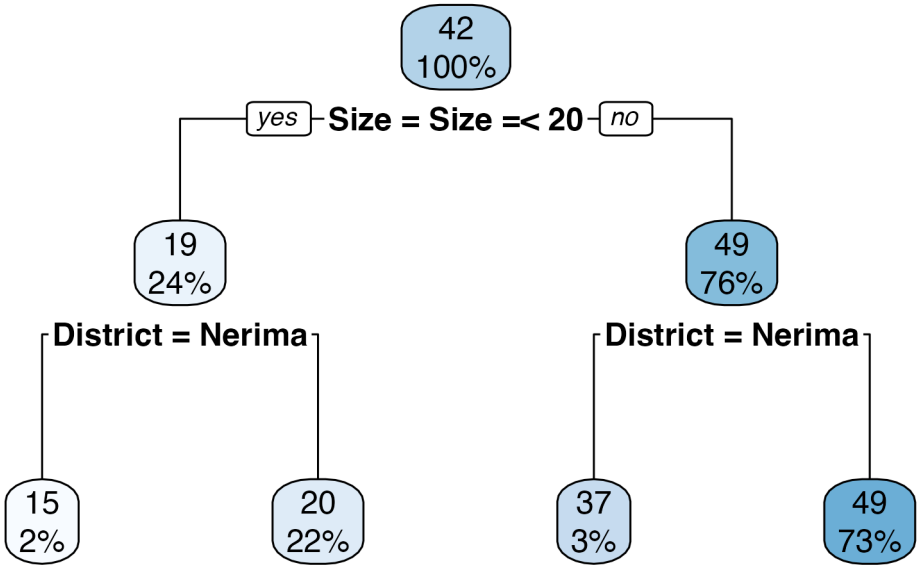
1. **分析者**が、サブグループ (モデル) を定義する

- 例: 練馬区かどうか × 部屋の広さが 20 以下

2. サブグループの平均値を計算

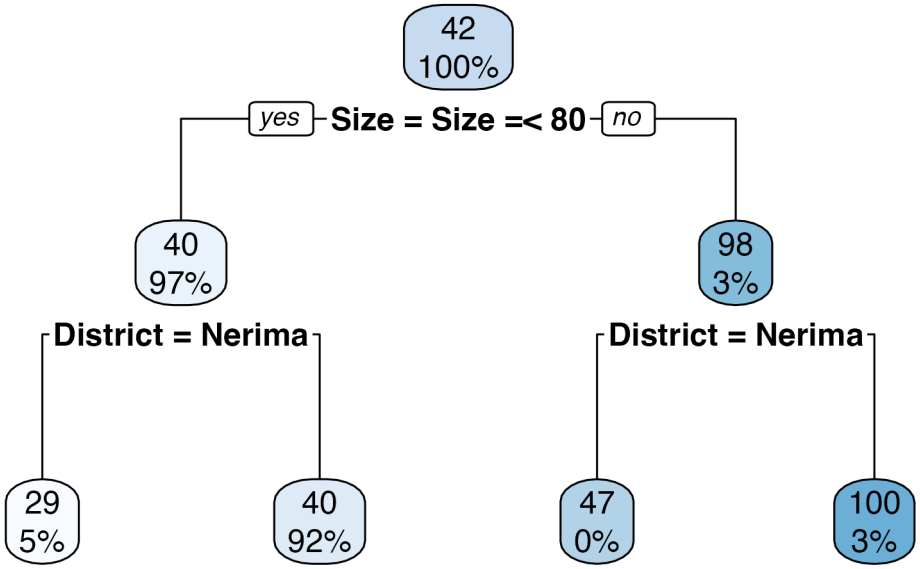
- 集計により事例固有の観察できない要因の影響を軽減する

3.3 サブサンプル分析



3.4 代替的サブサンプル分析

- モデルの定義を変えると、推定結果が大きく変化する



3.5 決定木アルゴリズム

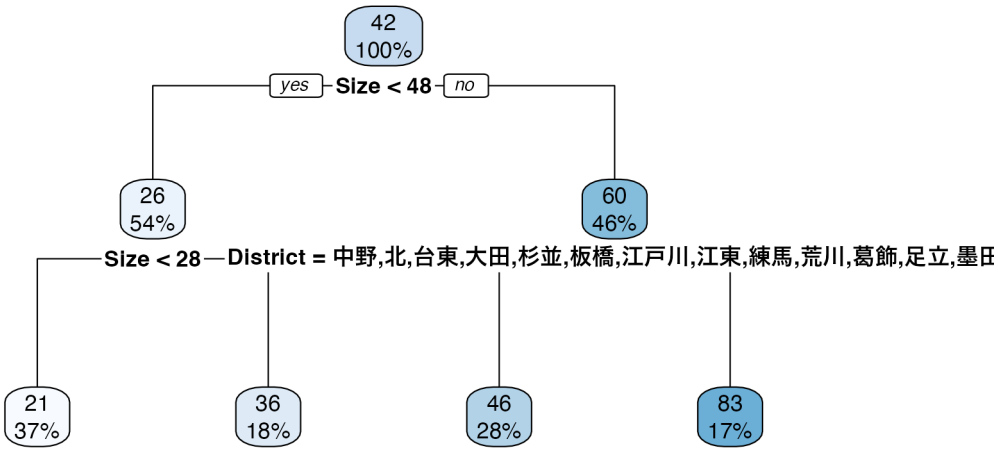
- 決定木アルゴリズム: **データ**に最も適合するように、サブグループを定義する
  - 明確な基準 (“データへの適合”) のもとで、要約方法を決定
- 例. 最大４グループに分けることは前提に、平均二乗誤差

$$(Y - \text{モデルの予測値})^2 \text{のデータ上での平均値}$$

を可能な限り削減するようにグループ分けを行う

  - 近似的に削減する (Greedy-algorithm)

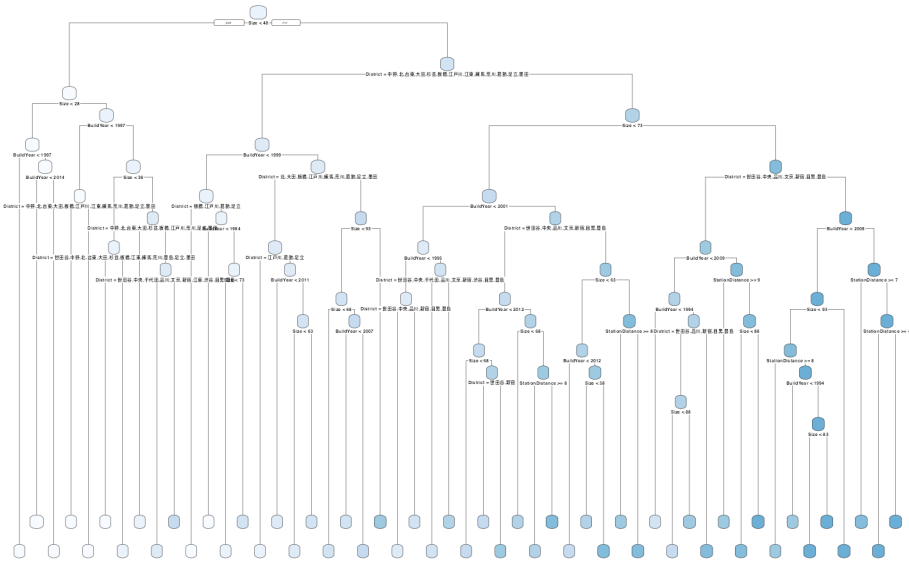
3.6 例. 決定木



- データへの適合度 (平均二乗誤差) は、45 % ほど改善

3.7 機械学習の活用例

- 複雑なモデルも推定できる



- データへの適合度 (平均二乗誤差) は、80 % ほど改善

4 分析のゴール

- 分析のゴールに応じて、データ主導 (決定木) と人間主導 (サブグループ分析)、どちらが優れているのかが異なる
  - 分析のゴールを大きく 予測 VS 記述 に分類

4.1 分析のゴール: 予測

- 新しい事例ごとに、欠損情報 Y を X から予測する
  - 本講義: 建物の属性 $(= X)$ から 取引価格 $(= Y)$ を予測
  - 他の例: 個人属性 $(= X)$ から 転職後の賃金 $(= Y)$ を予測する
    - ビズリーチ
- データ主導のアプローチ (機械学習) を用いて、適度に複雑な予測モデルを推定することが有効

4.2 分析結果例

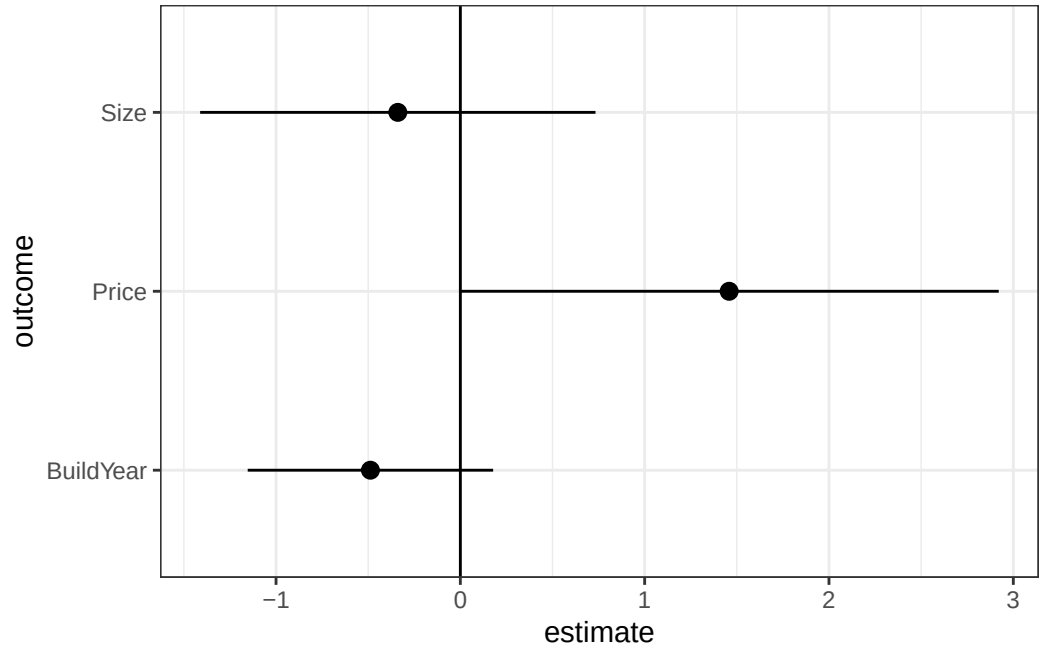
PredY	District	Size	StationDistance	BuildYear
59.265641	江東	65	10	2003

28.300147	杉並	25	15	2022
58.875302	世田谷	65	13	2005
101.033753	中央	75	3	2005
44.642129	墨田	55	9	1991
19.851346	品川	20	6	2002
32.172202	千代田	20	2	2006
23.320493	渋谷	25	8	1985
4.409519	練馬	25	10	1976
25.947103	足立	60	7	1992
24.535517	品川	45	4	1973
35.053441	板橋	40	2	2008
96.298159	千代田	55	5	2018
23.413771	荒川	45	5	1986
53.022208	江東	65	8	2003
72.239402	品川	50	6	2010
35.350336	江東	55	12	1997
8.683508	杉並	10	8	1995
179.199265	港	80	3	2011
37.265744	世田谷	50	6	1971

4.3 分析のゴール: 記述分析

- 特定の D と Y の関係性を集計/記述する
  - 本講義: 2021-2022 年にかけて、取引価格がどのように変化したのか?
  - 他の例: 改装済み/前の店舗間で、平均収益はどの程度異なるのか?
  - 政策例: 過去３０年間で出生率はどの程度変化したのか?
- 研究者主導のアプローチ (伝統的推定) に優位性
  - 真の値を 95% の確率で含む区間 (信頼区間) を計算できる

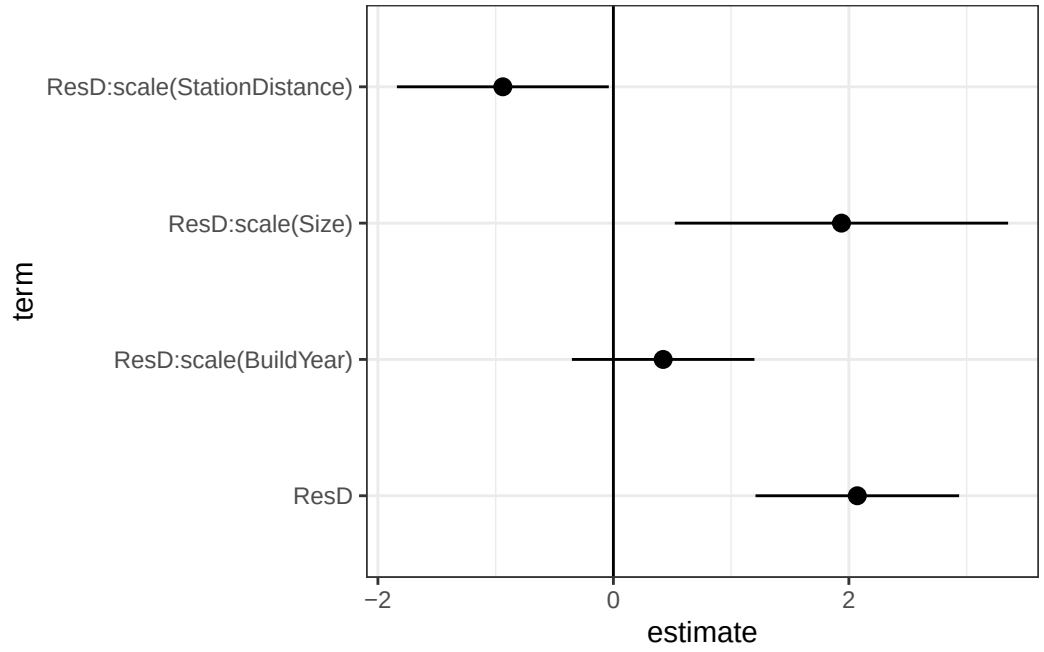
4.4 分析結果例



4.5 分析のゴール: 複雑な比較分析

- D 間で存在する X についての偏りを均した後の Y の平均格差
  - 合計特殊出生率
  - 既存店ベース
- 機械学習と伝統的アプローチのハイブリットが有効

4.6 分析結果例



5 意思決定問題

- その背後にある意思決定問題に応じて、適切な分析のゴールを設定する必要がある
  - ミクロ VS マクロに分類

5.1 “ミクロな” 意思決定への活用

- 影響の範囲が狭い意思決定に有益
  - 影響を与える事例について、欠損情報を予測できる
- 容易だが、めんどくさい意思決定の自動化に注目
  - 写真判定、迷惑メール判定、等々

5.2 “マクロな” 意思決定への応用

- 影響の範囲が広い (“マクロな”) 意思決定に対しては、個別事例の予測値” そのもの” の便益は限定的
  - 例. 政策決定、企業の戦略決定

- 通常、複数の人間による集団的意思決定が必要
  - 幅広い合意形成に向けて、影響を与える (大量の) 事例の特徴について、意思決定者が理解できる情報提供が必要
    - 大量の予測値を示されても、理解できない

5.3 実例

- 中期経営計画: 就業者や株主、世間に伝えやすい、集計情報に基づいて、説明を行なっている
  - セブン&アイ
- 白書: 有権者等に向けて、集計情報に基づいた、現状分析結果を説明している
  - 労働経済白書

5.4 参考書

- Applied Causal Inference Powered by ML and AI
  - 機械学習の入門と格差/因果分析への応用までカバーした優れた教科書
- Introduction to Statistical Learning
  - 定評のある機械学習の入門書

6 R

- Python と並ぶ、データ分析の人気言語
  - 高い透明性と拡張性、再現可能性、無料

6.1 準備

- 必要な package (追加プログラム) をインストール (Youtube 動画)
  - 所内 PC では不要
- Project folder の作成 (Youtube 動画)
  - File -> New project -> New directly -> New project から、任意の場所に folder を作成
  - すべての成果物が保存される

6.2 Example code

```
Data = readr::read_csv("Data/Public.csv")

Model = rpart::rpart(
  Price ~ Size + BuildYear,
  Data
  ) # rpart パッケージ内の rpart 関数を使用し、決定木を推定

rpart.plot::rpart.plot(Model) # rpart.plot 関数を使用し、可視化
```

- コードを実行する際には、(慣れるまでは)、以下の手順を徹底

1. ctr + a を押し、全ての行を選択する
2. ctr + enter を押し、実行する

6.3 Error が出たら

- 「error は必ず起きる」、という心構えをもつ
- 再現性の確認：全ての行を再度実行
  - コード実行しわすれ、がエラーの原因となることが多い
- よくあるミス (大文字/小文字の区別、コンマ) を確認
  - 極力予測変換を活用し、タイポを減らす
- 解決できない場合は、コード全体をチェット欄にコピペしてください

6.4 Reference