予測研究とデータ主導のモデル構築

Data visualization

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-07-31

1 予測研究

1.1 研究者によるモデリング

- 教科書的な OLS や最尤法は、モデルの大枠を研究者が詳細に決定し、細かいパラメタの値をデータが決める
  - 母平均や母分布を、比較的シンプルなモデルで表したい場合に有効

1.2 限界

- 研究対象と十分に関連する”比較的シンプルなモデル”をどのように設定するか？
  - 分析が不透明になりがち
- より明確な研究対象(予測/比較/因果効果等)を設定する研究が増えている

1.3 予測研究

- X から、Yの値を予測し、その精度を評価する
  - Y の予測値を計算する予測モデルの推定
- 機械学習の”得意分野”
  - よりデータ主導のモデリングが可能

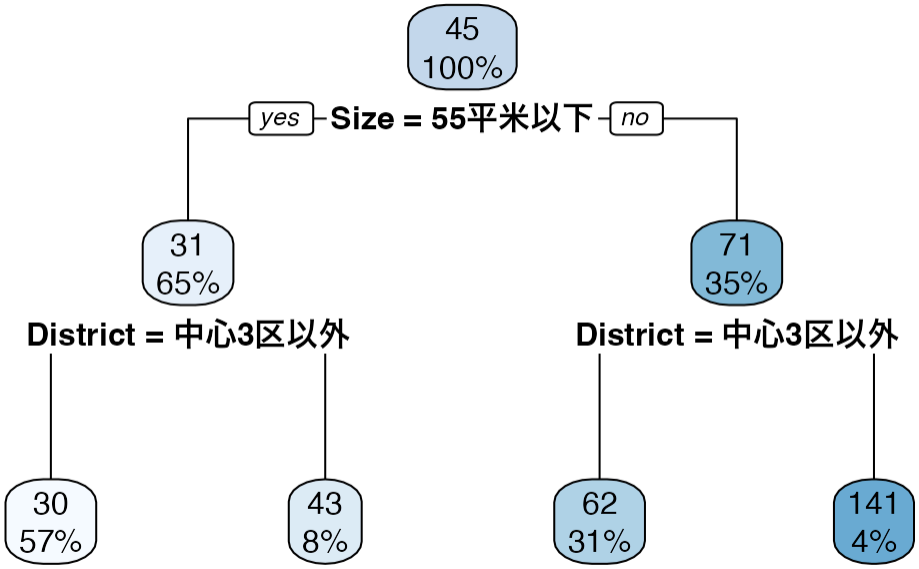
1.4 実習例: 不動産価格予測モデル

- **研究課題:** 物件の属性 $X =$ [広さ,立地] から中古マンションの取引価格(100 万円) (Y) を予測
  - 2024 年の東京 23 区を対象
  - 動機: 中古マンションの買取業務に関する意思決定の補助

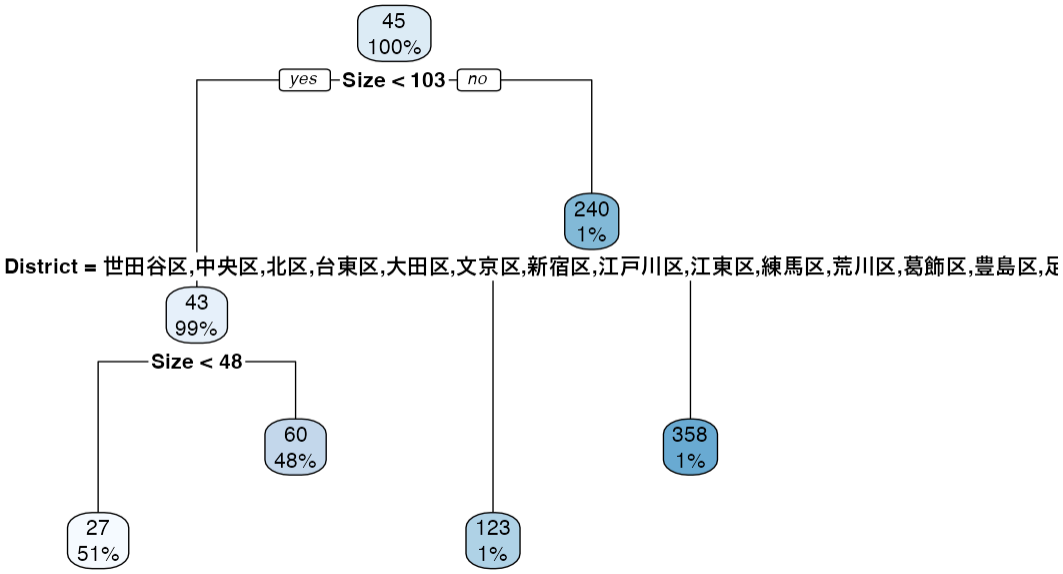
1.5 データ主導のモデリング

- 回帰木: X に応じて、サブグループに分け、サブグループごとの平均価格を予測値とする
- 問題はサブグループをどのように定義するか
  - **研究者主導:** 事前知識や”長年のかん”などに基づき決定
  - **データ主導:** データに最も適合するように決定
    - 例: greedy algorithm
    - 平均二乗誤差(後述)を最小化するように決定

1.6 実例: 研究者主導のモデルの可視化



1.7 実例: データ主導のモデルの可視化



2 機械学習

2.1 機械学習

- 主に画像や音声、テキストデータを用いた予測研究への応用において、急速に普及した手法
  - 近年、比較研究等への応用も可能になり、社会/市場分析についても有力な手法に
- 回帰木はその代表例

2.2 背景: 機械学習への期待

- 機械学習や因果推論などの知見も取り込んだ「データ活用」の広がり
  - 例 Amazon, Cyber Agent, Microsoft, Mizuho, Netflix, Uber
- 「計量経済/生物医療統計 + 教師付き学習」が発展し、市場/社会の特徴把握 (含む因果効果の推定)への活用も大きく進展
  - 意思決定の”自動化”のみならず、人間による重要な意思決定支援にも活用可能

2.3 背景: 混乱

- 急速な普及とともに、混乱した議論が、肯定/否定を問わず見られる
- 私見: 実務への適切な実装には、概念的/実践的な議論を通じて、早めに”地に足をつける”ことが重要

- 参考資料: ハイプ・サイクル (wiki)
  - 予測マシンの世紀
  - AI Snake Oil (WIRED の紹介記事) “AI ツールに接する機会を得たなら、機械学習やニューラルネットワークといった AI の主要概念を理解することに、少しだけ時間を費やしてみてほしい”

3 予測研究の深掘り

3.1 研究目標

- 評価指標を、最初に設定する必要がある
- 典型的な指標は、二乗誤差

$$(Y - Y\text{の}\mathbf{予測値})^2$$

- 安定的な予測性能を目指す = 二乗誤差の”期待値”を減らす

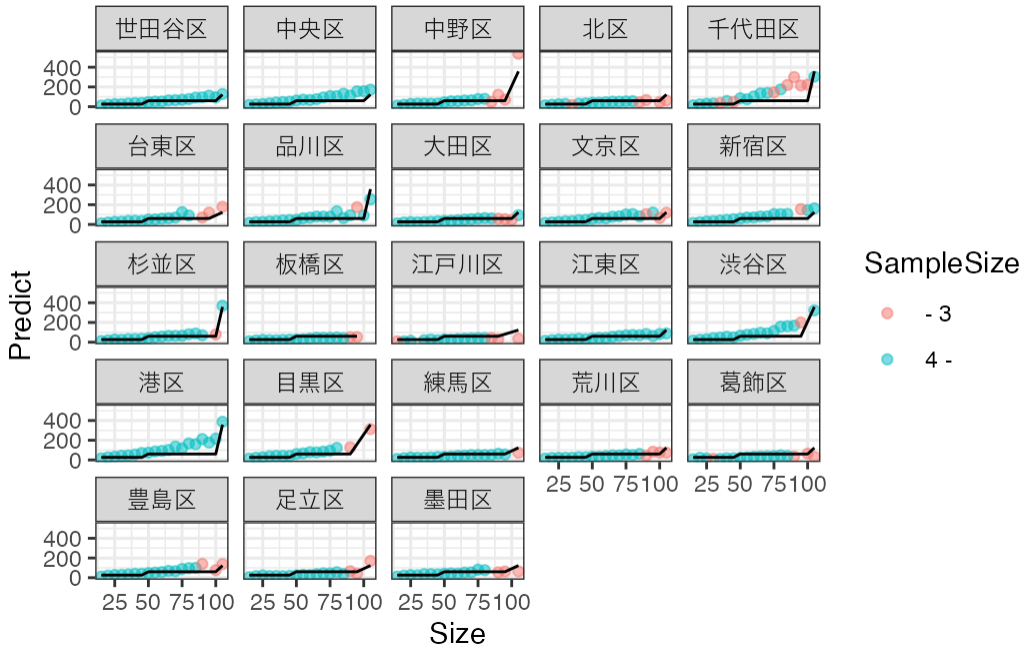
3.2 推定目標

- **定理:** 二乗誤差の母平均を最小化する予測モデルは、母平均 $E[Y \mid X]$
  - = 推定対象
  - $\simeq$ 無限大のデータを予測モデル推定に活用できるのであれば、「データ上の平均値」が推定対象
- 推定のチャレンジ: 限られた事例数のデータから、どうすれば”無限大のデータ”を用いた推定結果を近似できるか?

3.3 モデル化

- 推定精度を確保するためには、“適度に単純化したモデル” が必要
- 全ての X の組み合わせについて、 Y の平均を計算する方法は、丸暗記法 (Learning-by-memorization) と呼ばれる
  - 予測値が非常に大きく変動することを許容した”複雑なモデル”
- 回帰木: 「サブグループ内では予測は同じ」になることを要求
  - より単純化されたモデル

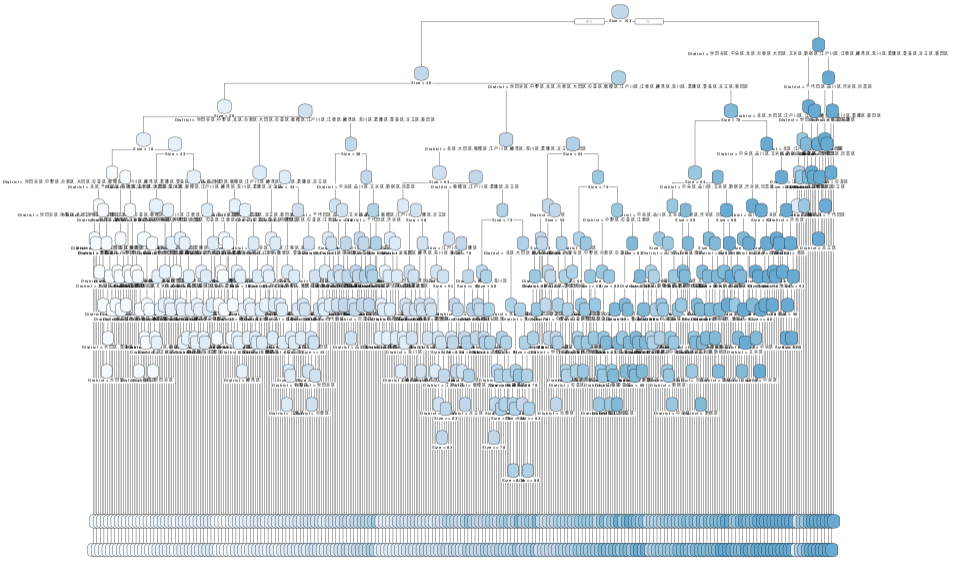
3.4 例



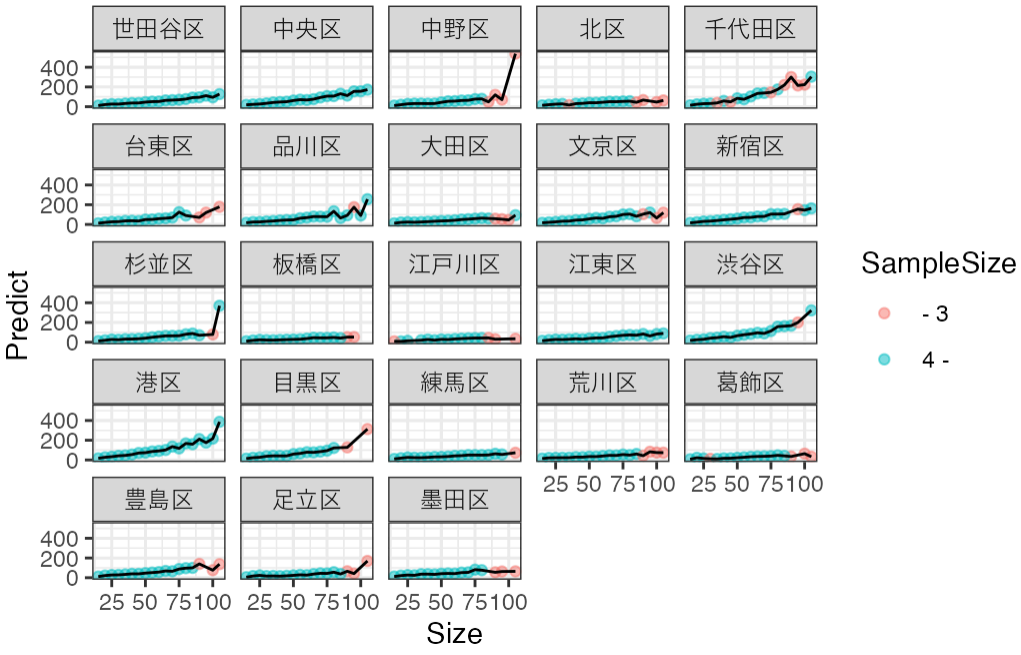
3.5 モデルの複雑化

- 機械学習の手法を用いれば、複雑なモデルを少ない労力で推定できる
- 回帰木の場合: 最大分割回数を増やすなど
  - 予測性能が悪化する可能性もある
    - 平均値の計算に用いる事例数が減少し、観察できない要因の偏りの影響を強く受ける

3.6 例. [分割=30、最小事例数=1]



3.7 例. [分割=30、最小事例数=1]



4 予測モデルの評価

4.1 Train/Test への分割

- 予測モデル推定に用いた事例について算出した予測値は、“予測ではない”
- 標準的な手順では、推定を始める前に、テストに用いるデータ(Test data)をランダムに抽出し、“封印”する
- Training data のみを用いて、モデルを推定し、テストデータへの予測性能を確かめる
  - 事例数が十分あれば、テストデータにおける平均二乗誤差 = 母集団における平均二乗誤差

4.2 実例

```
set.seed(111) # シード値固定

split <- rsample::initial_split(data, prop = 4 / 5) # データ分割

train <- rsample::training(split)

test <- rsample::testing(split)

tree <- rpart::rpart(
  Price ~ Size + District,
  data = train
) # 回帰木 with default parameter

ols <- lm(
  Price ~ Size + District,
  data = train
) # OLS
```

4.3 性能

```
pred_tree <- predict(tree, test) # 予測値

pred_ols <- predict(ols, test)
```

4.4 性能

```
mean((test$Price - pred_tree)^2) # 評価値
```

```
[1] 868.3134
```

```r
mean((test$Price - pred_ols)^2)
```

```
[1] 803.8644
```

4.5 Takeaway

- “データ”から、“モデル(単純化されたパターン)”を推定し、活用する
  - 観察できない要因の偏りの影響を緩和
  - 予測モデルは、その一種
- 機械学習は、よりデータ主導で、モデルを推定できる
- 予測と”当てはめ”を区別
  - 予測対象を、予測モデルを推定するデータに含めることは原則できない
  - 予測性能を評価する手段として、サンプル分割を活用

4.6 補論: 交差推定の活用

- 交差推定を用いれば、データ上の全ての事例について、“予測値”を得ることも可能

1. データを 2 **以上**のサブデータ(サブデータ 1,2…)に分割
2. サブデータ 1 以外のデータを用いて、予測モデルを推定し、サブデータ 1 のYを予測
3. サブデータ 2,3..について、同じ手順を繰り返し、予測値を算出

4.7 数値例: 3 分割

StationDistance	Price	Group
9	6.05	3
4	3.94	2
7	31.05	3
1	8.64	1
2	−5.99	3
7	−4.48	1
2	−0.89	1
3	0.01	2
1	−3.12	2

4.8 数値例: Step 1

StationDistance	Price	Group	OLS	Tree
9	6.05	3	NA	NA
4	3.94	2	NA	NA
7	31.05	3	NA	NA
1	8.64	1	−4.12	5.32
2	−5.99	3	NA	NA
7	−4.48	1	12.88	5.32
2	−0.89	1	−1.29	5.32
3	0.01	2	NA	NA
1	−3.12	2	NA	NA

- Group 2,3 を Training データとして活用

4.9 数値例: Step 2

StationDistance	Price	Group	OLS	Tree
9	6.05	3	NA	NA
4	3.94	2	4.86	5.73
7	31.05	3	NA	NA
1	8.64	1	−4.12	5.32
2	−5.99	3	NA	NA
7	−4.48	1	12.88	5.32
2	−0.89	1	−1.29	5.32
3	0.01	2	3.55	5.73
1	−3.12	2	0.94	5.73

- Group 1,3 を Training データとして活用

4.10 数値例: Step 3

StationDistance	Price	Group	OLS	Tree
9	6.05	3	−4.89	0.68
4	3.94	2	4.86	5.73

StationDistance	Price	Group	OLS	Tree
7	31.05	3	−3.03	0.68
1	8.64	1	−4.12	5.32
2	−5.99	3	1.61	0.68
7	−4.48	1	12.88	5.32
2	−0.89	1	−1.29	5.32
3	0.01	2	3.55	5.73
1	−3.12	2	0.94	5.73

- Group 1,2 を Training データとして活用

4.11 Reference

Bibliography