機械学習

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

Table of contents

1 概念整理

1.1 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?
  - どのような事業領域が伸びているのか?

1.2 実務への影響

- すでに多くの活用事例がある
  - MicroSoft, CyberAgent, 日経センター
  - 金融機関 (Bank of England)
- 事例紹介 + 考察
  - 予測マシンの世紀　AI が駆動する新たな経済

1.3 データ分析の諸分野

- 統計学、計量経済学、医療統計、**機械学習**
  - 異なるコミュニテイ (例: 機械学習 = “AI の研究”) によって発展
  - 融合が急速に進む
  - 分析方法ごとの特徴を理解し、適した手法を採用することが重要に

1.4 到達目標

- 自身でデータ分析を行うための入門
  - 事例分析の抱える不透明性 (Uncertainly) への対処に焦点
- データ分析の結果を意思決定に活用するための必要知識の取得
- エントリーシートなどに、“AI を用いた予測モデル/レポート作成経験” と書けるようにする

2 不透明性 (Uncertainly)

- データ分析を含む事例分析の一般的課題
  - 独立した分析チームに**同じ市場/社会**の分析を依頼したとしても、同じ結論に到達し得ない
    - 誰がやっても、「水は 100 度で沸騰する」という結論に到達する理科の実験とは対照的

2.1 “個別事例分析”

- 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、2539 事例 (2021 年第 1 四半期) から含意を得る
- 最低価格で取引されている物件は

Price	District	Size
5	足立区	35

- “足立区の３５平米の物件の取引価格は 500 万円” と主張できるか?

2.2 “個別事例分析”

- 全く同じ属性を持つ物件は、他にも存在する

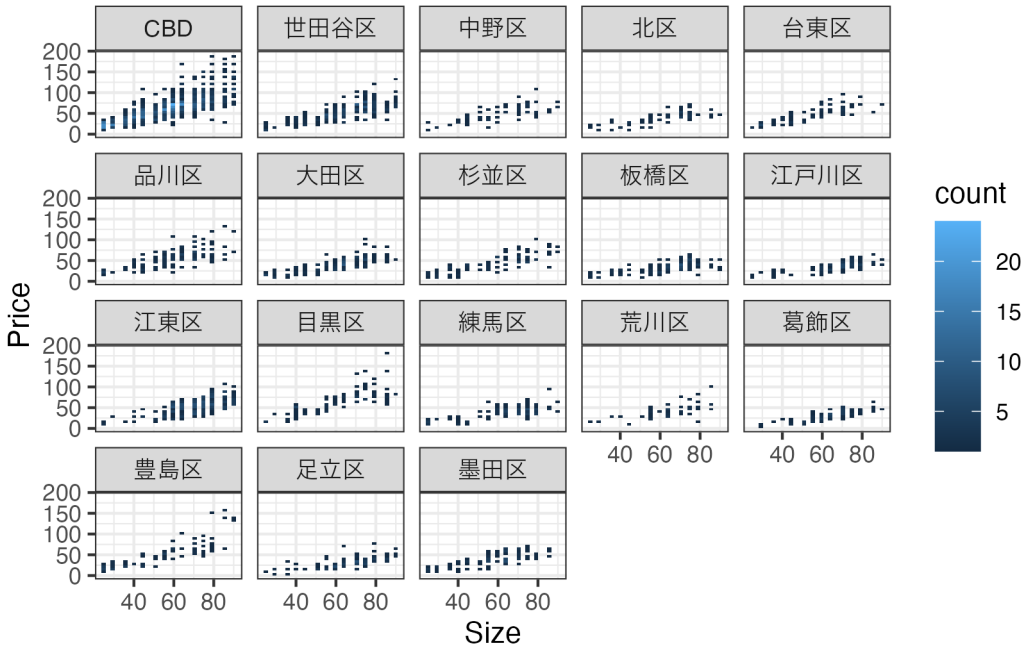
Price	District	Size
33	足立区	35
15	足立区	35
5	足立区	35

- データから観察できない要因で、この事例の取引価格が下振れているのでは?

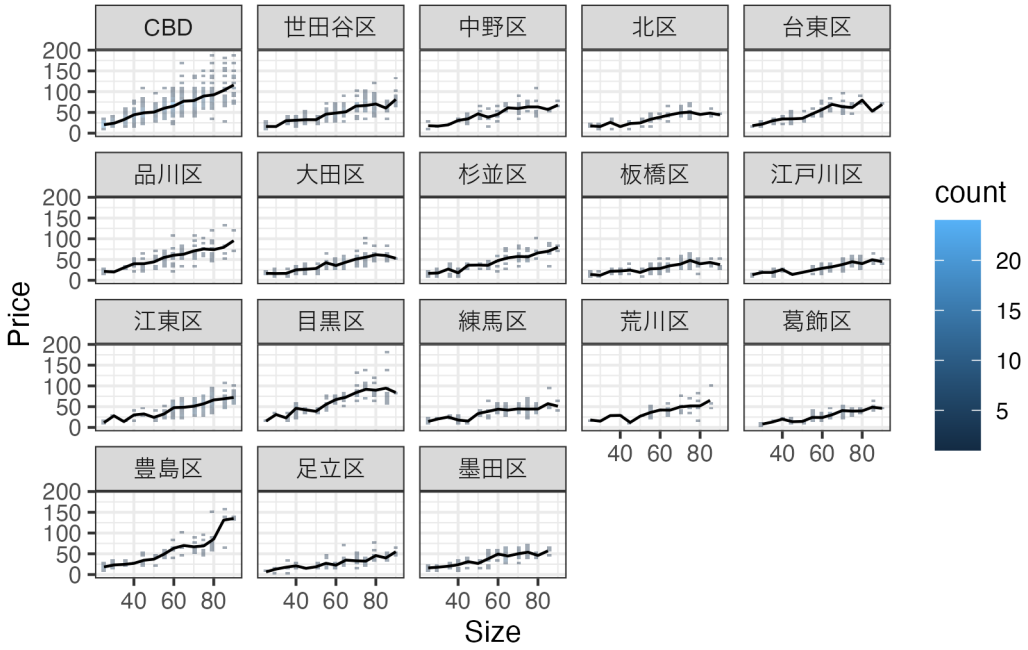
2.3 要約計画

- 大規模なデータになると、両親の学歴が同じ大量の事例
  - 何らかの要約を用いて、集団の” 傾向” を論じる
    - 典型的な事例の報告、研究者の所見/印象、平均値、中央値、分散
- 恣意的分析を避けるために、データを見る前に、要約計画を立てることが重要
  - 平均などの” 自動計算” できる方法が有効

2.4 実例



2.5 実例



2.6 Sampling Uncertainly の軽減

Price	District	Size
26.0	中野区	25
9.8	中野区	25

- 中野区 & 25 平米は 2 事例のみ = 平均取引価格は 17.9
- データを取りなおしたら、違う結果になるのでは?
  - データから観察できない価格決定要因の” 偏り” が、データ固有の特徴を生み出す
  - 観察する事例が人によって異なるため、結論が異なる (**Sampling uncertainly**)
- 対策: 多くの事例を集計 (モデル化) し、傾向把握を行う

2.7 サブサンプル分析

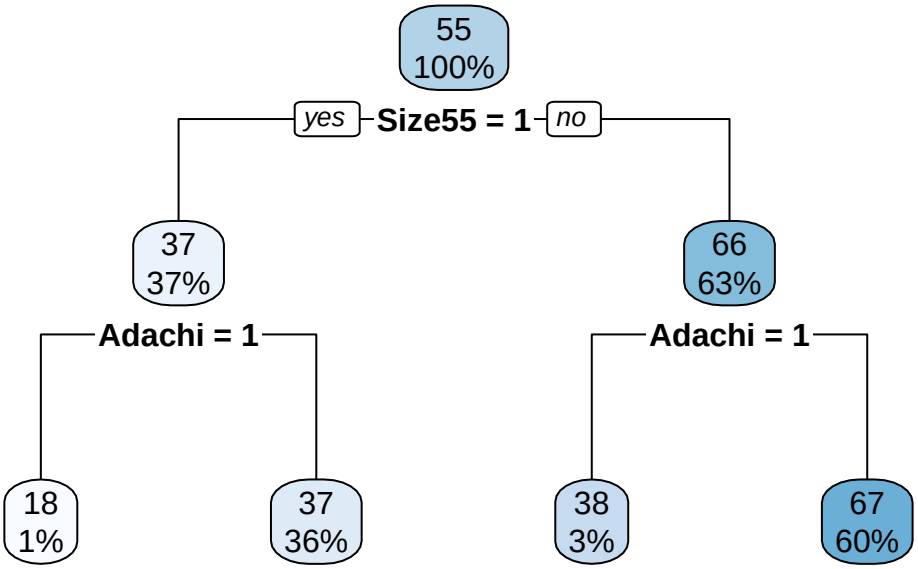
1. **分析者**が、サブグループ (モデル) を定義する

- 例: 足立区かどうか × 部屋の広さが 55 平米以下かどうか

2. サブグループの平均値を計算

- 集計により Sampling uncertainly を軽減する

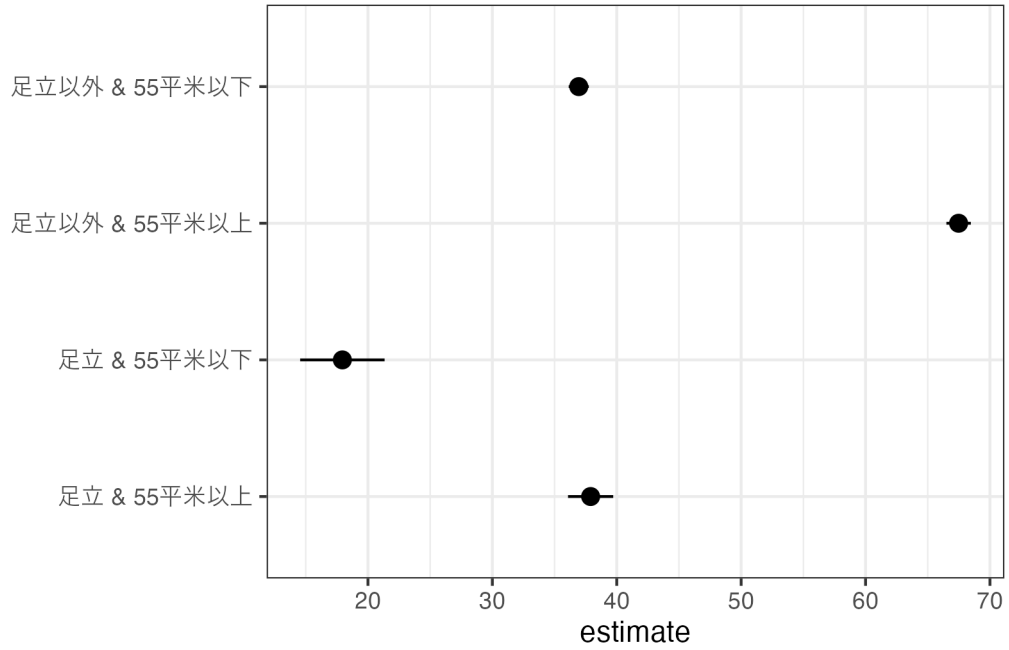
2.8 サブサンプル分析



2.9 サブサンプル分析の利点

- データが、関心となる集団 (日本全体など）からランダムサンプリングにより生成されていれば、以下の優れた理論的性質をもつ
  - 無限大の事例数で、Sampling Uncertainly は消失する
  - そこそこの事例数で、信頼区間を近似計算できる
    - 一定の確率で母平均を含む区間
    - 偶然生じた傾向か否かを区別でき、重大な意思決定への活用において、特に望ましい性質

2.10 例: 信頼区間



2.11 Model Uncertainly

- モデルの定義を変えると、推定結果が大きく変化する
- 事例の要約方法 (モデル化) が属人的で、不透明 (**Model uncertainly**)
  - 多くの教科書で直接的な言及を避けてきた問題
    - *「理論や背景をしっかり踏まえて、適切なモデル化を行うべし」以上の提案が難しい
  - 機械学習の適切な活用で、軽減できる

2.12 付論: Model Uncertainly の副作用

- モデルを色々試すことで、分析者にとって" 望ましい結果" (例: 明確な因果効果、存在しない格差の存在" 証明" など) を、"捏造" できる
  - Cherry-pick/p-huck などと呼ばれる
- 無実の証明が難しい
  - 疑われないようにするために、"複雑な分析" を避ける傾向

    * 例: どのような背景属性の組み合わさると、改装は大きな効果を持つか?

2.13 Model Uncertainly の軽減

- モデルの根拠の明示することが重要

1. 分析課題が詳細に決まっているケース
    - 例: “足立区と物件と他を比較することが重要”
        – 詳細まで決まっていないケースはまれ
2. 何らかの一般基準に基づくモデル生成
    - 機械学習活用に比較優位

2.14 機械学習の活用

- 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
        * 選定基準の明確化による Model Uncetainly の” 削減”
            · 予測研究であれば、説得的な基準

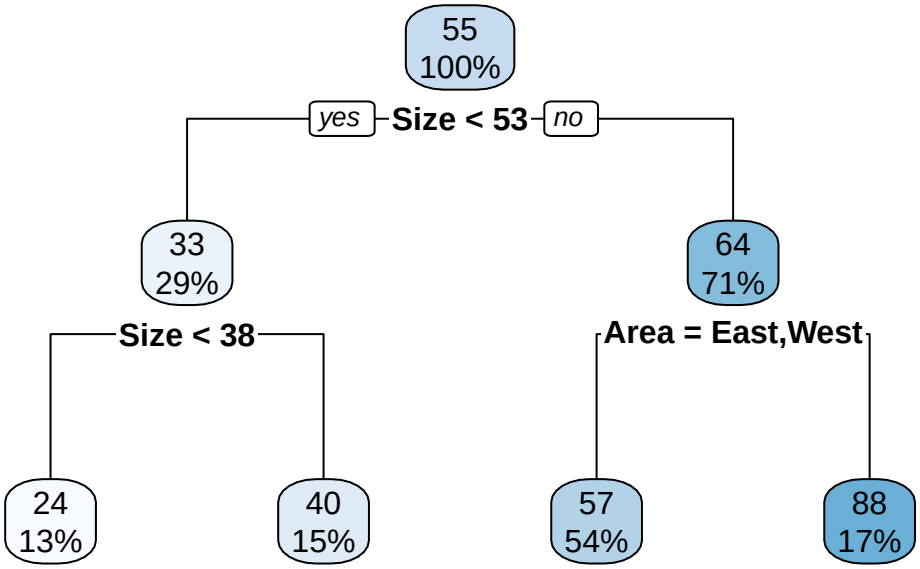
2.15 例

- 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)

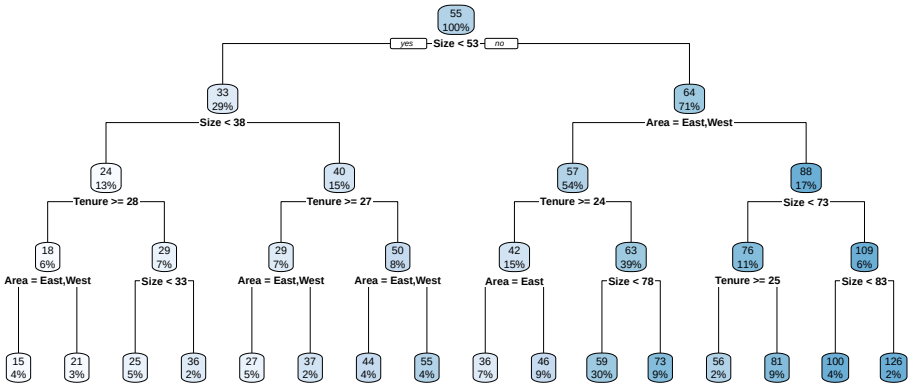
2.16 実例



2.17 例. 複雑なモデル

- 機械学習を用いれば、より複雑なモデルも容易に生成できる
- 例. 最大 32 グループに分ける

2.18 複雑なモデル



3 意思決定問題への応用

3.1 意思決定問題

- 意思決定 = 状況把握/予測 + 判断
  - 今日、傘を持って出かけるか?
    - 状況把握: 雨が降るか/人に会うか/体調はどうか
    - 判断: 雨に濡れることは、どのくらい問題か?
  - どの事業に経営資源を収集するのか?
    - 状況把握: 各事業の現状、将来性
    - 判断: 企業の経営目標に寄り添うのは?

3.2 “ミクロな” 意思決定

- 個別” 事例” のみに影響を与える意思決定
  - 医療データ: 患者ごとの治療方針

  - 不動産データ: 中古マンションの買取
  - 労働市場: 労働者ごとの 30 年後の賃金
- “事例ごとの” 状況把握が必要
  - 複雑なモデルを用いた、個別予測が有効
    - 機械学習の活用に大きな利点

3.3 実例

- 転職の意思決定
  - 状況把握: 自身の市場価値
    - 例: レバテックのサイト
  - 判断: どのような働き方をしたいか

3.4 実習例: 取引価格予測モデル

Size	Distance	Tenure	District	Predict
80	9	14	杉並区	64
85	11	17	世田谷区	71
70	5	41	品川区	47
40	5	43	板橋区	20
45	5	9	CBD	62

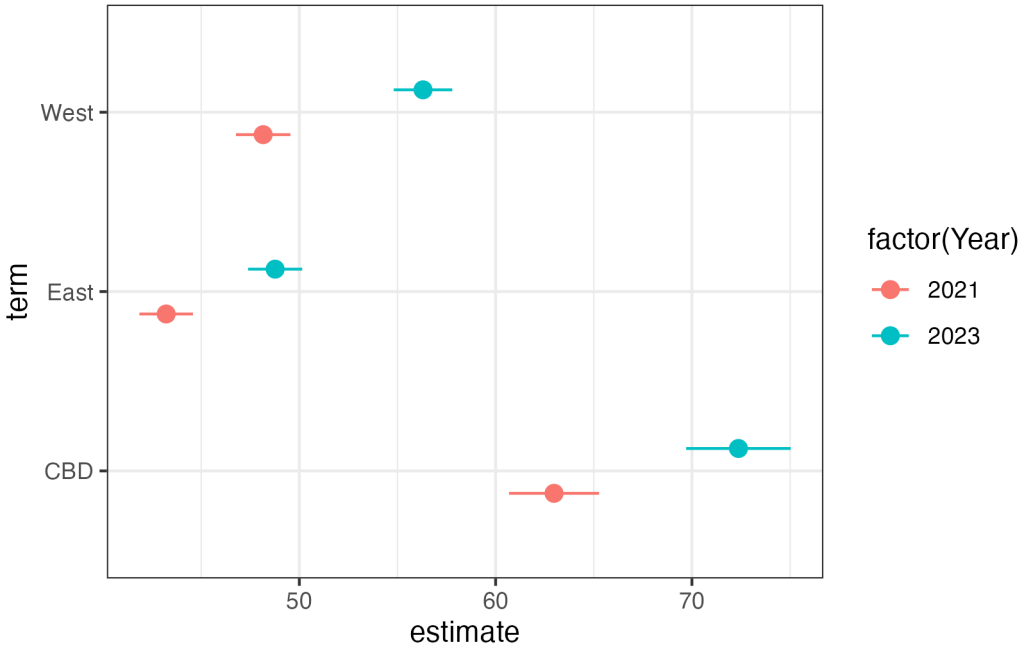
3.5 “マクロな” 意思決定への応用

- 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
- 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
  - 幅広い合意形成に向けた、幅広い情報伝達が難しい
- 事例のシンプルな要約であれば、伝統的な統計手法に大きな比較優位

3.6 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、現状分析/経営方針を説明
  - 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/政策方針を説明
  - 労働経済白書

3.7 実習例: 平均取引価格の推移



3.8 What-if 分析

- 伝統的な手法 + 機械学習によって、より柔軟な集計が可能
- 本講義では、Balanced comparison を議論する。
  - 因果推論、格差、現状分析において、非常に重要な分析方法となっている。
  - 伝統的な統計学/機械学習の教科書では、紹介されていない

3.9 実例. 既存店ベースの比較

- あるコンビニチェーンで、店舗あたりの平均売り上げが 1000 万円増大した
- 去年から今年にかけて、新規出店が大きく増加した
  - 売り上げが大きくなる傾向のある新規店の割合が大きく、結果、平均売り上げが増大したのではないか?
- 既存店に絞って、比較する

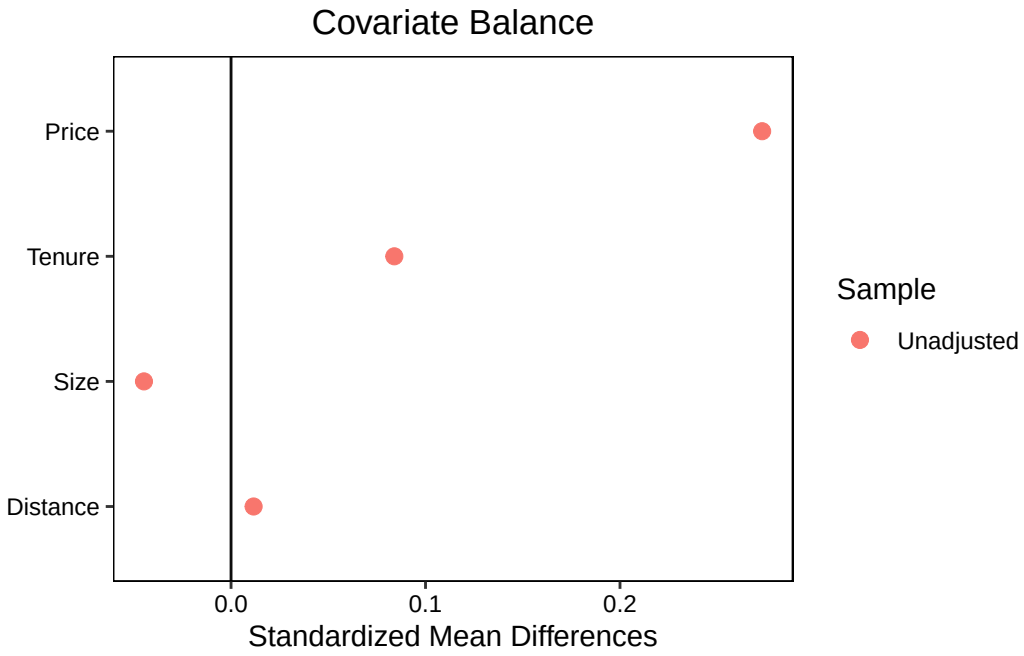
3.10 実例. 客層

- あるコンビニチェーンで、大手町店と本郷三丁目店で、客単価が大きく異なる
  - 客層の違いに起因しているのではないか?
    - 本郷三丁目の方が、大学生が多い等
- 来客の年齢の分布を仮想的に揃えて、比較する

3.11 実例. 因果推論

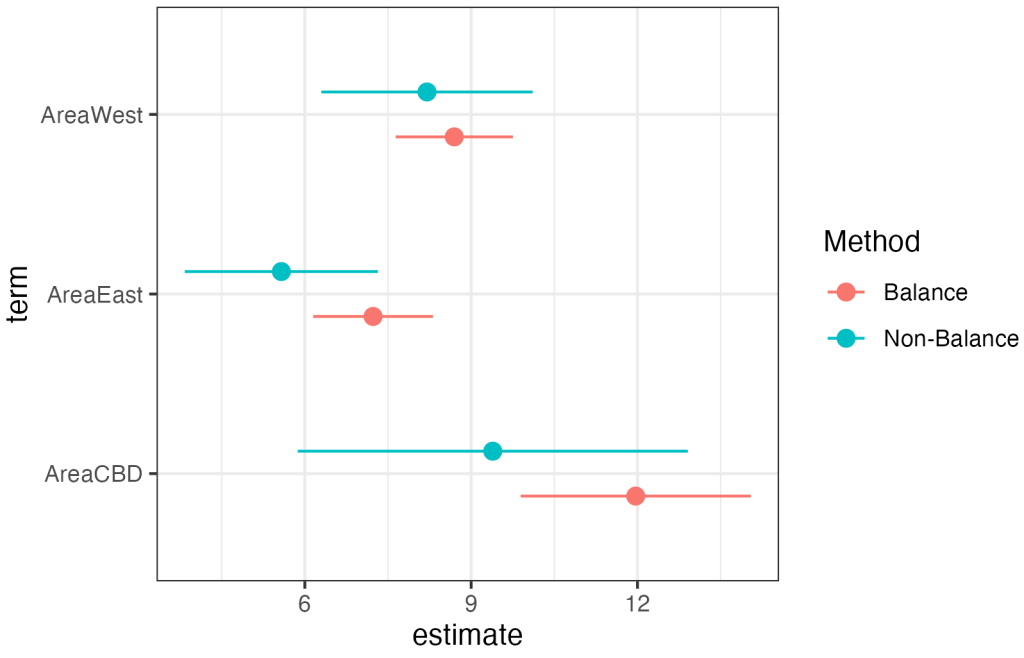
- コンビニの改装が平均的にどの程度、平均売上を上昇させるのか
  - 「改装するかどうかの意思決定、および価格に影響を与える変数」(Confounders) が偏る
  - データから観察できる Confounders については、分布を揃える必要がある

3.12 実習例: 2021-2023 年比較



- 取引価格のみならず、取引物件の性質も変化している

3.13 実習例: 取引価格の変化



3.14 まとめ

- 意思決定に応じて、必要な情報の” 細かさ” が異なる
- “細かさ” が異なれば、最適な手法が異なる可能性がある
  - 事例ごとの予測を提供するのであれば、現状、機械学習には大きな優位性がある
  - シンプルな集計情報であれば、伝統的な推定方法 (平均値など) の実用性が高い
  - 複雑な集計情報であれば、伝統的な推定方法と機械学習との併用が有効

4 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
  - 多様な統合開発環境 (IDE)

4.1 準備

- Posit cloud への登録
  - クラウド環境で R を使用できる
  - ただし時間制限あり
- 関心がある人は、自身の PC へ Rstudio をインストール
  - 時間無制限

4.2 Example code

- 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

4.3 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