発展

Data visualization

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

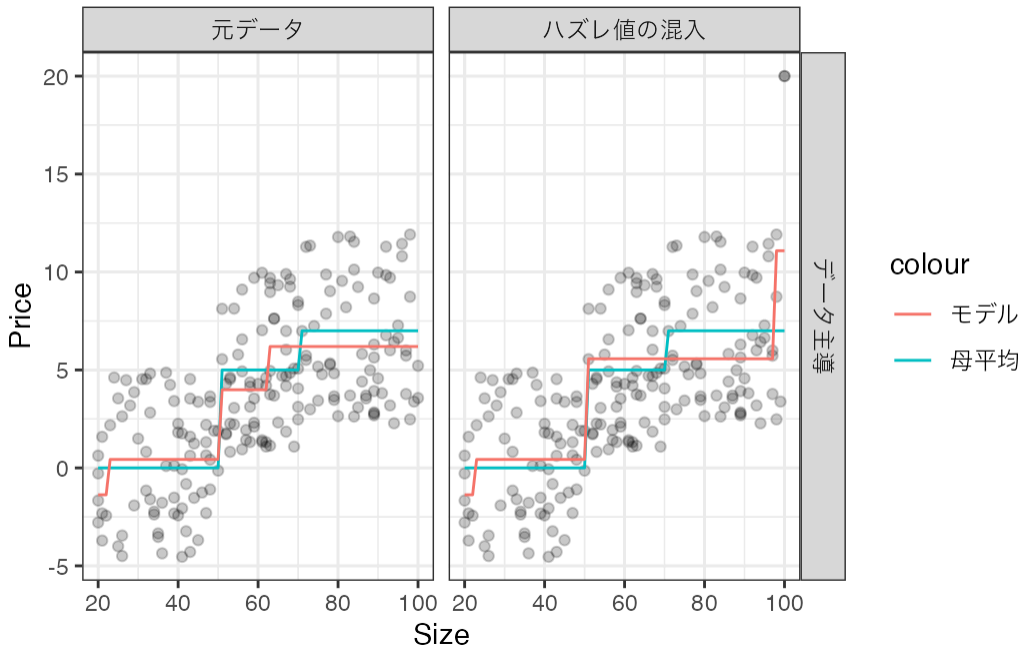
2025-08-05

1 予測モデルの改善

1.1 集計による解決

- データへの、平均値から極端に乖離した事例やその組み合わせの”混入”は、推定結果に大きな影響を与える
  - 特にデータ主導のグループ分けでは、グループの定義自体も変化し、データとの関係性が複雑化になる
- モデルの単純化も選択肢だが、回帰木については不十分な場合が多い

1.2 数値例 (200 事例)



1.3 解決策

- 伝統的なアプローチ: “ハズレ値”を人間が除外
  - 採用するのであれば、“細心の注意”が必要
  - (議論はあるが)、Y の値について行うのは、非推奨

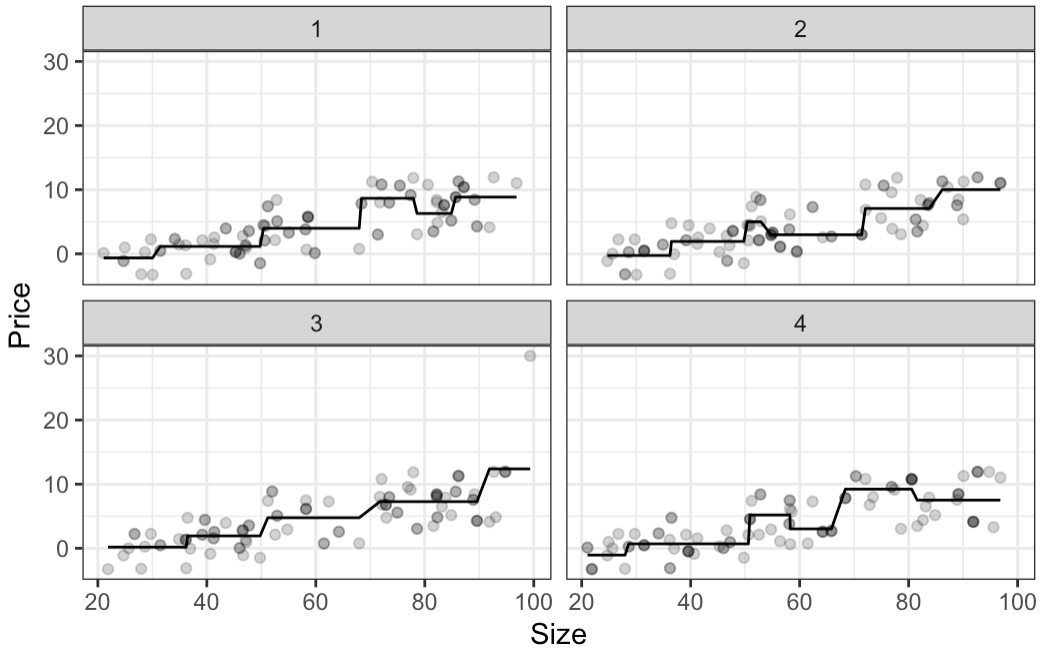
1.4 集計による解決

- モデルの集計
  - “異なる”データを用いた予測モデルの集計値(平均値)を最終予測とする
    - 特定のハズレ値の影響を緩和できる
- 問題点: 通常、データは一つしかない
  - 対応策: Boostrap 法により、データを複製する

1.5 シンプルな例

- データ $= [5, 6, 100]$
- 復元抽出により、同じ数(3)の事例をランダムに選ぶ
  - 複製データ $1 = [6, 6, 100] =$ の平均値 37.3
  - 複製データ $2 = [6, 6, 5] =$ の平均値 5.7
  - 複製データ $3 = [5, 5, 5] =$ の平均値 5
- 最終予測 $= 16$
- ハズレ値(“100”)を反映しない予測も活用される

1.6 数値例



1.7 利点

- 各複製データについて、ある事例が含まれる確率は 1/3 程度
  - 少数の事例に依存したモデルの比率は低い
    - より頑強なパターンの抽出が期待できる
- Random Forest: 回帰木を推定する際に、X からランダムに選ばれた変数を除外する
  - 計算速度が向上し、推定精度も改善することが多い

1.8 Takeaway

- モデル集計は、回帰木などのデータへの依存度が高い推定方法の改善に有効
- 注: OLS などに対して、有効な方法ではない

2 残差回帰への応用

2.1 OLS の限界

- $Y|D$ の予測モデルを OLS で推定する
  - 適切なモデルを研究者が設定する必要があるが、困難
- LASSO や Random Forest を活用する

2.2 残差回帰 with Random Forest

- β_D は以下の手順で推定できる

1. Y, D の予測モデルを RandomForest で推定
2. 残差 $Y -$ **予測値** を $D -$ **予測値** で回帰 (OLS)
3. 信頼区間を計算

2.3 小技

- Boostrap 法でモデル集計を行う場合、特定の事例を含まないデータで推定されたモデルのみを集計することが可能
  - 一回の推定で全ての事例に対して、予測値を算出できる

2.4 残差回帰の利点

- 予測モデルが、母平均 を"ある程度"近似できれば、信頼区間を推定できる
  - 高い予測精度を要求しない
    - 大きな個人差によって、予測が当たらなかったとしても、最善の予測値 (母平均)に近い予測を行うモデルが推定できれば OK

2.5 残差回帰の利点

- 機械学習が活用できるので、データ主導でモデル化できる
  - 研究者の分析時間の節約
  - 分析プロセスの透明化が可能
    - "X について、なぜそのような定式化を採用したのか?"という質問は、非常に回答しにくい
      - 機械学習を用いると、Y/D を最もうまく予測できるとデータとアルゴリズムが判断した、と回答できる

2.6 直感

- 機械学習を用いても、予測モデルと母平均の乖離を、一定以上削減するのは難しい
  - "AI(予測モデル)もミスを犯す"
- 残差回帰においては、2 種類の"AI" (Y/D を予測する AI) を推定
  - どちらかの AI の精度が低かったとしても、もう一つの AI が母平均をうまく近似すれば OK
    - Double Machine Learning とも呼ばれるアプローチ
    - "AI によるダブルチェック"

3 異質性の推定

3.1 実例

- 改築済み/未改築の取引価格格差は、背景属性 X によって異なる(異質性がある)と考えられる
  - 例: 古い物件の方が、改築済み/未改築の格差が大きい

3.2 一般化した線型モデル

- 部分線型モデルを一般化する

$$Y = \underbrace{\beta_D(X)}_{\text{非常に複雑な関数}} \times D + \underbrace{\beta_0 + \beta_1 X_1 + ..}_{\text{非常に複雑な関数}} + u$$

3.3 FWL 定理の応用

- $\beta_D(X)$ を以下の手順で推定

1. Y, D の予測モデル $g_Y(X), g_D(X)$ を交差推定で推定
2. $Y - \textbf{予測値} \sim \underbrace{\beta_D(X)}_{\beta_0+\beta_1 X_1+..} \times (D - \textbf{予測値})$ を OLS 推定

3.4 補論: 標準化

- 近似モデルにおいて、通常 β_0 の解釈は難しい
  - 全てのZ が"0"であった場合の"値"
- Z を 標準化 $(Z - Z\textbf{の平均値})/Z\textbf{の標準偏差}$ すれば、
  - 全てのZ が平均値であった場合の"値"となり、より解釈しやすい

4 補論: Stacking

4.1 予測モデルの選択

- OLS や Random Forest 等で推定した予測値のうち、どれを使用するのか?
  - 理論的に常に優れた方法は存在しない
- 方法 1. 予測性能を評価し、最善のモデルを利用する
- 方法 2. 予測値を集計
  - 代表的な方法は、Stacking

4.2 Stacking

- 最終予測モデル

$$= \beta_{OLS} \times OLS\textbf{の予測}$$

$$+\beta_{RF} \times RandomForestの予測 + ...$$

- β : 各予測結果を反映させる度合い
- 各予測値を”X”として用いた、線型モデル

4.3 推定方法

1. データをサブデータ $\{1, .., G\}$ にランダム分割
2. 第 l サブデータ**以外**で予測モデルを推定し、第 l サブデータを予測
3. 第 2 サブデータ**以外**で予測モデルを複数推定し、第 2 サブデータを予測
4. 以上を全てのデータについて繰り返す
5. 予測対象 Y に対して、各予測値で回帰して β を推定

4.4 数値例: 3 分割

```
# A tibble: 9 × 3
  StationDistance    Price Group
            <int>    <dbl> <fct>
1               9  6.05    3
2               4  3.94    2
3               7 31.0     3
4               1  8.64    1
5               2 -5.99    3
6               7 -4.48    1
7               2 -0.895   1
8               3  0.00785 2
9               1 -3.12    2
```

4.5 数値例: Step 1

```
# A tibble: 9 × 5
  StationDistance    Price Group   OLS RandomForest
            <int>    <dbl> <fct> <dbl>        <dbl>
1               9  6.05    3     NA           NA
2               4  3.94    2     NA           NA
3               7 31.0     3     NA           NA
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3     NA           NA
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2     NA           NA
9               1 -3.12    2     NA           NA
```

- Group 2,3 を Training データとして活用

4.6 数値例: Step 2

```
# A tibble: 9 × 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    3     NA           NA
2               4  3.94    2      4.86        -0.189
3               7 31.0     3     NA           NA
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3     NA           NA
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2      3.55        -0.189
9               1 -3.12    2      0.938        1.91
```

- Group 1,3 を Training データとして活用

4.7 数値例: Step 3

```
# A tibble: 9 × 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    3     -4.88        -1.84
2               4  3.94    2      4.86        -0.189
3               7 31.0     3     -3.03        -1.84
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3      1.61         0.945
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2      3.55        -0.189
9               1 -3.12    2      0.938        1.91
```

- Group 1,2 を Training データとして活用

4.8 数値例: Stacking

```
lm(Price ~ OLS + RandomForest, PopData)
```

```

Call:
lm(formula = Price ~ OLS + RandomForest, data = PopData)

Coefficients:
 (Intercept)           OLS  RandomForest
       5.056        -1.248         0.243
```

- ω を非負、総和を 1 に基準化することも有効

4.9 Takeaway

- 機械学習伝統的な推定手法ではあまり用いられてこなかった、アイディアを用いた多くの手法が存在
- バランス後の比較など、純粋な予測研究以外への応用法も確立されている

4.10 継続学習用推奨資料

- 機械学習: An Introduction to Statistical Learning
- 機械学習 + 計量経済学: Applied Causal Inference Powered by ML and AI
- 講師作成の資料
  - 母平均の「補助線」の推定
  - 格差/因果/比較分析のためのデータ分析

4.11 Reference

Bibliography