補論: 平均差の特徴把握

機械学習

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

Table of contents

1 平均差の" 特徴把握"

- どのような" 層" で、特に差が大きいのか?
  - 価格上昇が激しい地域や物件の特徴は何か?
  - 親の属性に応じて、男女間賃金格差は異なるのか?

1.1 理想的な推定方法

- 全ての X の組み合わせについて、平均差 $E[Y|D=1,X]-E[Y|D=0,X]$ を計算する
  - X の値が一致している事例内で比較
    - X ごとに平均差を推定できる
- 問題点: X の組み合わせが増えると、サブグループ内での事例数が極端に小さくなり、実行不可能

1.2 実例

- 例: $X=\{$ Size, District $\}$

```
estimatr::lm_robust(Price ~ D, Data, subset = Size == 65 & District == "江東")
```

```
            Estimate Std. Error   t value     Pr(>|t|)  CI Lower CI Upper DF
(Intercept) 46.32000   2.358474 19.639817 2.635557e-27 41.599001  51.0410 58
D           13.96571   3.446884  4.051693 1.530831e-04  7.066027  20.8654 58
```

```
estimatr::lm_robust(Price ~ D, Data, subset = Size == 50 & District == "港")
```

```
            Estimate Std. Error   t value     Pr(>|t|)  CI Lower CI Upper DF
(Intercept) 63.55556   4.585942 13.858779 2.195337e-11 53.957069 73.15404 19
D           12.77778   8.229999  1.552586 1.370206e-01 -4.447808 30.00336 19
```

2 平均差の線形近似モデル

- 確立されている推定方法
- 平均差のシンプルなモデルを OLS で推定する
  - 予測モデルを補助的に用いつつ、近似的信頼区間を提供できる

2.1 線形近似モデル

- 平均差の線形近似モデル を推定

$$E[Y|D=1,X]-E[Y|D=0,X]\simeq\beta_0+\beta_1Z_1+..$$

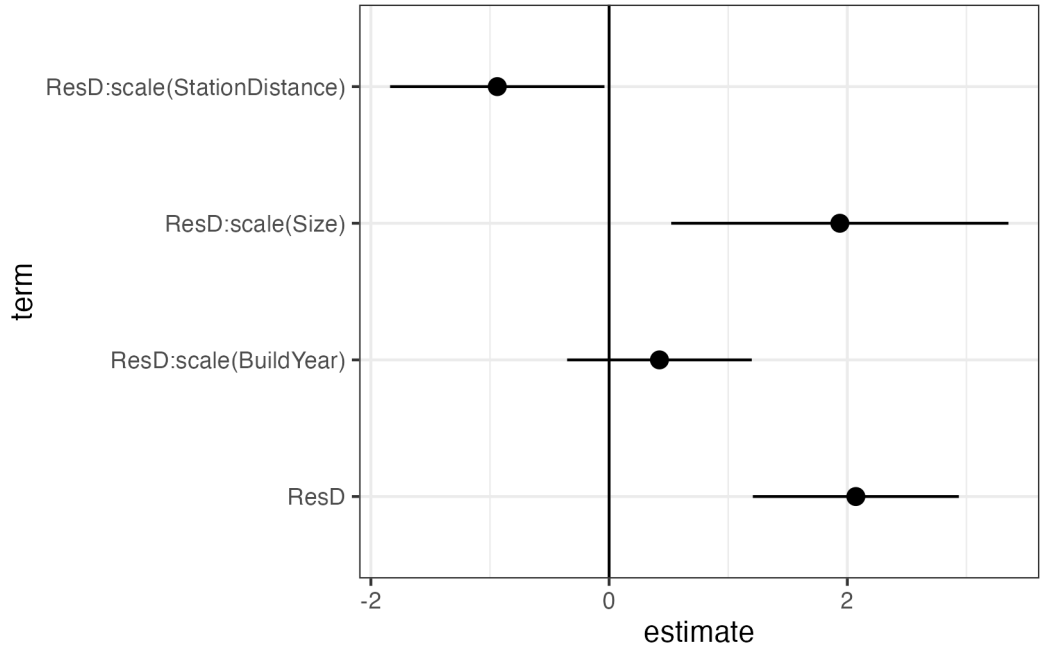
- 研究関心に応じて Z は、X の一部の変数のみでも良い
  - 標準化 $(Z-Z$の平均$)/Z$の標準偏差 を行うことで比較が容易となる

2.2 推定手順

1. データを訓練/テストにランダム分割
2. 訓練データのみを用いて、Y/D の予測モデル $g_Y(X)/g_D(X)$ を推定する
3. テストデータのみを用いて、$Y - g_Y(X)$ を $D - g_D(X)$ (主効果) と $Z \times (D - g_D(X))$ (交差項) で OLS 回帰

- 全ての推定結果について、信頼区間を近似計算できるので、そこも含めて結果を判断

2.3 実例



- 駅から遠くても広い物件の方が、価格上昇が大きい

3 機械学習を用いたモデル推定

3.1 Model Uncertainly の削減

- 平均差 $E[Y|D=1,X] - E[Y|D=0,X]$ の近似モデルを線形 $\beta_0 + \beta_1 X_1 + ..$ に決め打ちにすると、Model Uncertainly は発生
- $E[Y|D=1,X] - E[Y|D=0,X]$ の近似モデル $g_\tau(X)$ を機械学習で推定する
  - Model Uncertainly の削減が期待でき、数多くの手法 (T, X, S, R, DR learners) が提案されている

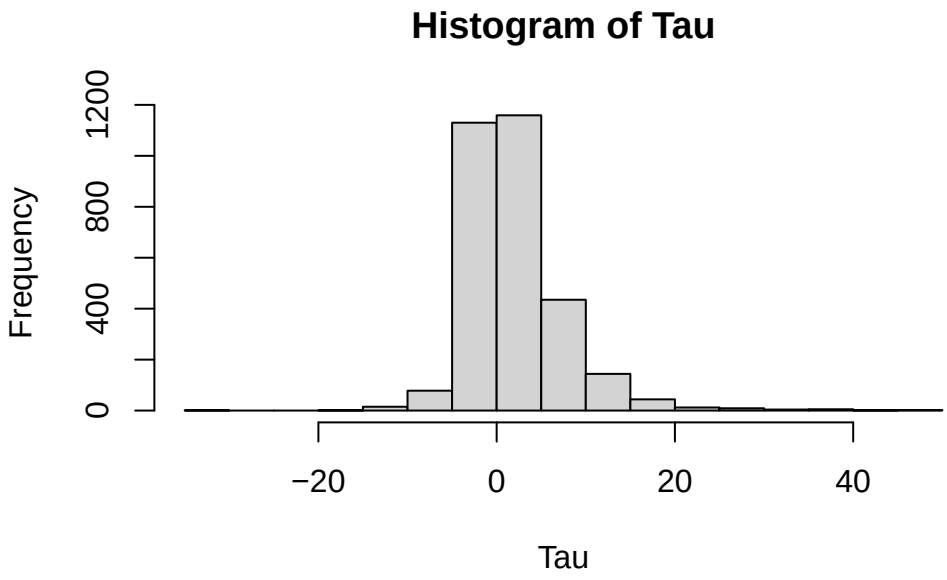
3.2 T-learner

1. 訓練データ $\&D = 1/0$ の事例を用いて、Y の予測モデル $g_Y(1, X)/g_Y(1, X)$ を推定
2. $E[Y|D = 1, X] - E[Y|D = 0]$ の近似モデル: $\tau(X) = g_Y(1, X) - g_Y(0, X)$. をテストデータについて計算

3.3 問題点

- 一般に、$E[Y|X]$ と同等以上に、推定誤差が大きくなりがち
- 一般に信頼区間計算はできない

3.4 例. T learner



3.5 異質性の探索への活用

- 「平均差が特に大きく $+$ / $-$ となるグループがあるか」のみを知りたいのであれば、より頑強な推定ができる

1. 訓練/テストデータに分割
2. 訓練データのみを用いて、予測モデル $g_Y(1, X), g_Y(1, X), g_Y(X), g_D(X)$ を推定

3. テストデータについて平均差の予測値 $\tau(X) = g_Y(1, X) - g_Y(0, X)$ を計算し、大小に応じて、テストデータをサブグループ分け

4. サブグループごとに R-learner にて、平均差を推定

3.6 異質性の探索への活用

```
High = Tau >= quantile(Tau, probs = 0.5) # 中央値の計算

estimatr::lm_robust(
  ResY ~ ResD,
  subset = High) # 上位 50% グループ
```

```
             Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper
(Intercept) -0.505509  0.3661766 -1.380506 1.676342e-01 -1.223775 0.2127567
ResD         3.175939  0.7331108  4.332141 1.573572e-05  1.737922 4.6139563
              DF
(Intercept) 1518
ResD        1518
```

```
estimatr::lm_robust(
  ResY ~ ResD,
  subset =! High) # 下位 50% グループ
```

```
            Estimate Std. Error  t value   Pr(>|t|)     CI Lower  CI Upper   DF
(Intercept) 0.476732  0.2458010 1.939504 0.05262507 -0.005413533 0.9588774 1518
ResD        1.028922  0.5000971 2.057445 0.03981390  0.047967913 2.0098767 1518
```

3.7 Reference