比較分析への活用

機械学習

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

Table of contents

1 シンプルな比較分析

- 本講義では、集計情報として ($D-Y$ 間での) 比較分析を紹介する
  - 伝統的な推定方法が有効
    - 平均以外 (分散、分位点等) にも適用可能

1.1 実例

- １年前と比べた労働市場/社会の状況変化
- 雇用形態間の所得格差

1.2 平均の比較

- ２種類の方法が考えられる
  - “Y の特徴を調べる”: $D=1/0$ それぞれについて、Y の平均値を計算し、差をとる
    - データと推定結果の関連性について近似的な性質 (中心極限定理など) が成り立つ
    - 信頼区間が計算でき、強く推奨
  - “予測モデルの特徴を調べる”: $D=1/0Y$ それぞれについて、Y の平均値を計算し、差をとる
    - データと推定結果の関連性が複雑で、Blackbox
    - バイアスが生じ、信頼区間の計算も難しく、非推奨

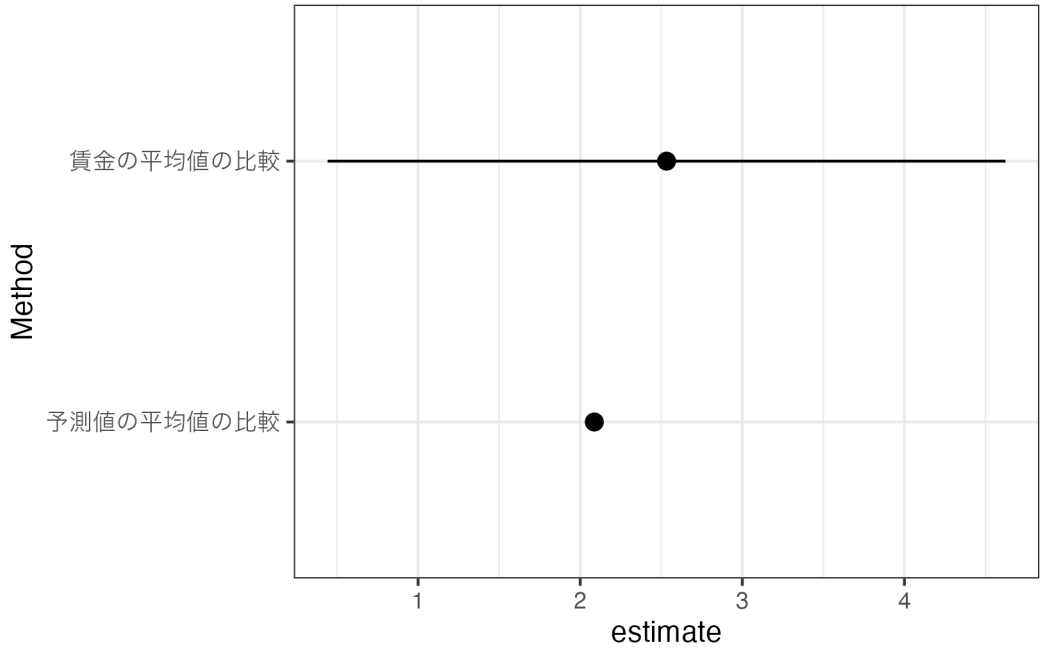
1.3 信頼区間

- データ上の (単純) 平均値であったとしても、データの偶然の偏り (Sampling Uncertainly) の影響を受ける

  - 一般に、データから得た推定値 $\neq$ 母平均

- 信頼区間 = 一定の確率 (多くの初期設定で 95 %) で、母平均を含む区間
  - Y の平均値を推定値とするのであれば、事例数がある程度大きい ($\simeq$ 200 以上) であれば、近似計算が可能
  - Y の予測値の平均値を用いる場合、計算方法が確立されていない

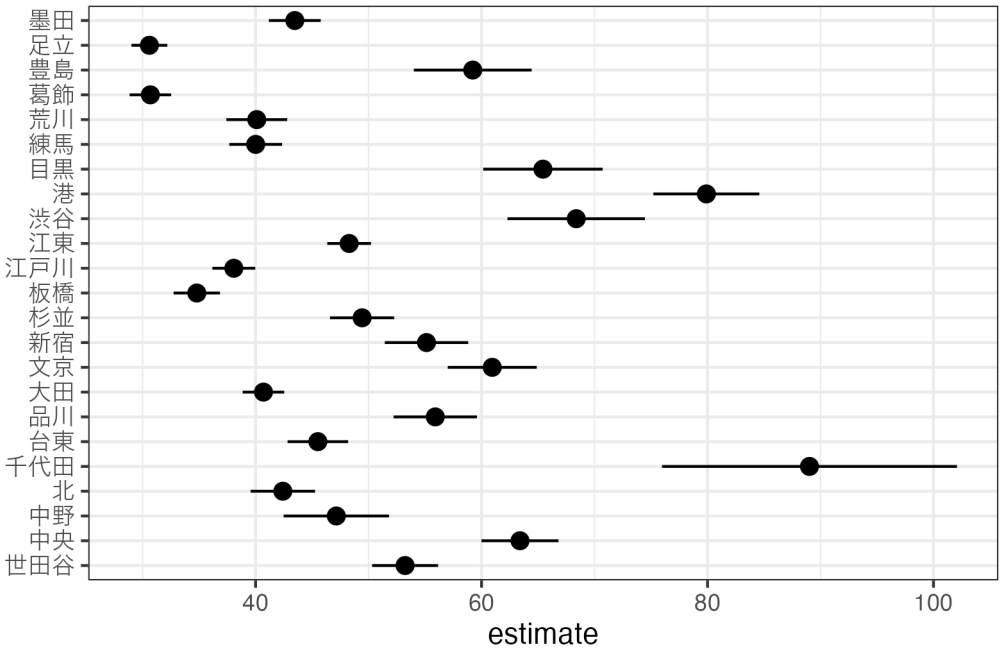
1.4 実例. 2021/2022 年の平均取引価格変化



1.5 信頼区間の重要性

- マクロな意思決定は、取り返しがつかない/幅広い層に影響を与えるものが多い
- 判断材料となる情報にも、高い信頼性/透明性が求められる
  - Sampling Uncertainly のせいで、点推定値は “100%” 間違えている
    - とくに少数グループについて、母平均からかけ離れた値が推定されがち
    - “推定誤差” を常に考慮する必要がある

1.6 実例: 23 区内価格格差



1.7 まとめ

- ある程度の事例数を活用し、シンプルな平均値やその差を推定したのであれば、計量経済学の入門書で紹介される「平均値の推定」法が実用的
  - 事例数が非常に少ないケース (30 以下など) であれば、ベイズ法なども有力

1.8 補論: 統計モデルの推定

- 平均差による母平均の差の推定結果は、以下の母平均のモデルを OLS で推定した結果と一致する

$$Y = \beta_0 + \underbrace{\beta_1}_{=E[Y|D=1]-E[Y|D=0]} D + \underbrace{u}_{E[u|D=1]=E[u|D=0]=0}$$

  - β_1 = Parameter of interest
- 「母集団の信頼できる統計モデルのパラメタを推定する」方法として、教科書では紹介されてきた
  - 研究関心が複雑になると信頼できるモデルを構築するのが難しくなる

2 バランス後の比較

- D 間で X についての偏りを均した上で、Y についてどの程度差が存在するのか?

  - 因果効果、格差分析のキモ

- 注: D, Y, X は分析に先立って、分析者が政策的関心等に応じて設定する必要がある

2.1 アプローチ

- 伝統的には、重回帰分析を用いた分析 $Y \sim D + X_1 + ..$ が行われてきた
  - 今でも有効だが、モデルの定式化に結果が依存してしまう問題がある
- 機械学習と OLS の併用が有効
  - Target/Debiased learning(Van der Laan, Rose, et al. 2011; Chernozhukov et al. 2018, 2022)
- CausalMLBook の 1,4,10 章に相当

2.2 変化の分解

- $E[Y|D=1] - E[Y|D=0] = Y$ の平均値の差
- 通常、Y 以外の変数 X についても、D 間で差が存在する
- X について差がなかった (バランスさせた) 場合に、Y についてどの程度差が存在するのか？

2.3 例. 出生率

- 出生の動向を把握する上で、新生児数を年次比較する
  - 成人の年齢構造の変化を無視している
    - 50 年前よりも、高齢者の比率が高い
- 合計特殊出生率
  -「年齢構造が同じであった場合の」出生率はどのように変化を議論できる
- 学歴や家族構成などもバランスさせることは可能か？

2.4 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

$$E[Y|D=1] - E[Y|D=0] = 1000\text{万円}$$

- 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 ($= X$) に絞って、比較する

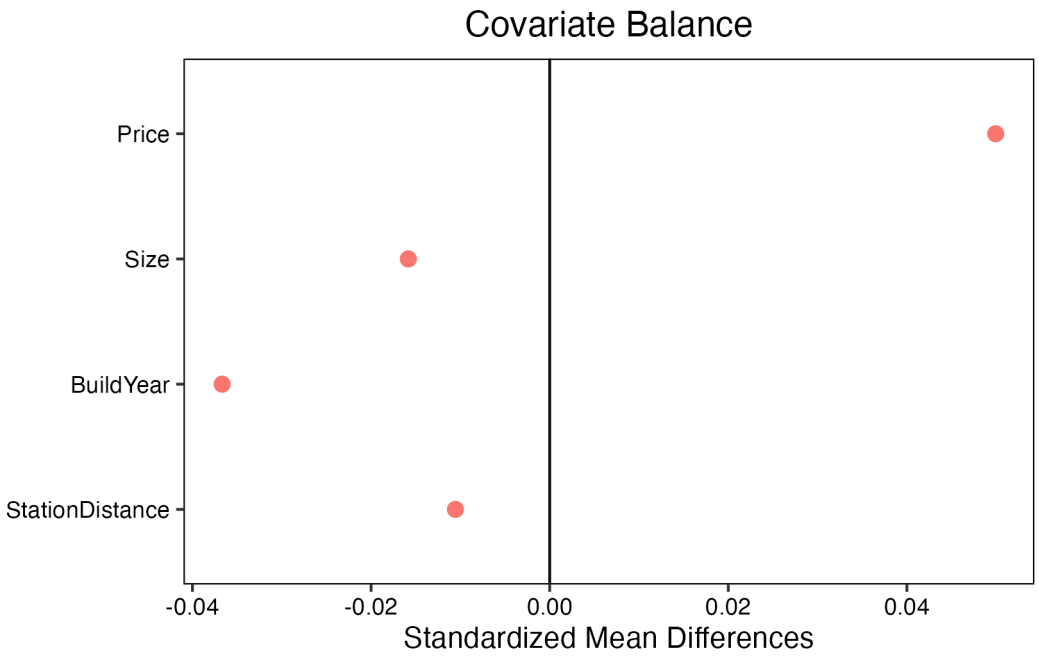
2.5 例. 客層

- 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
- 来客の年齢や職業 ($= X$) の分布を仮想的に揃えて、比較する

2.6 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 D 間で偏る
  - 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
  - (発展) 観察できない Confounders への対処は?
    - 統計的因果推論にて多くの議論が蓄積

2.7 実例. 2021-2022



2.8 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
- 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
  - 多くの応用で非現実的

2.9 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
```

2.10 R(esidual)-learner

- X を D 間でバランスさせた元での比較を行うために、以下の手順を実行する

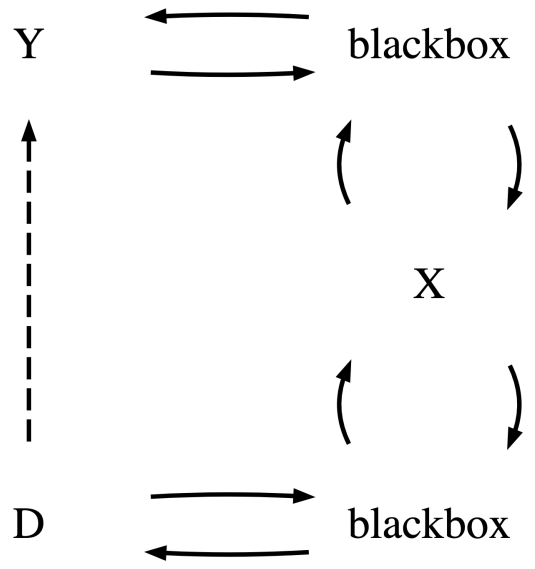
1. データを 2 分割する (訓練/テスト)
2. 訓練データを用いて、X から Y/D を予測するモデル $g_Y(X)/g_D(X)$ を推定する
3. テストデータを用いて、Y/D の予測誤差を OLS で回帰し係数 β_D を推定値とする

$$Y - g_Y(X) \sim D - g_D(X)$$

- 後日、より効率的なサンプル分割方法 (交差推定) を紹介

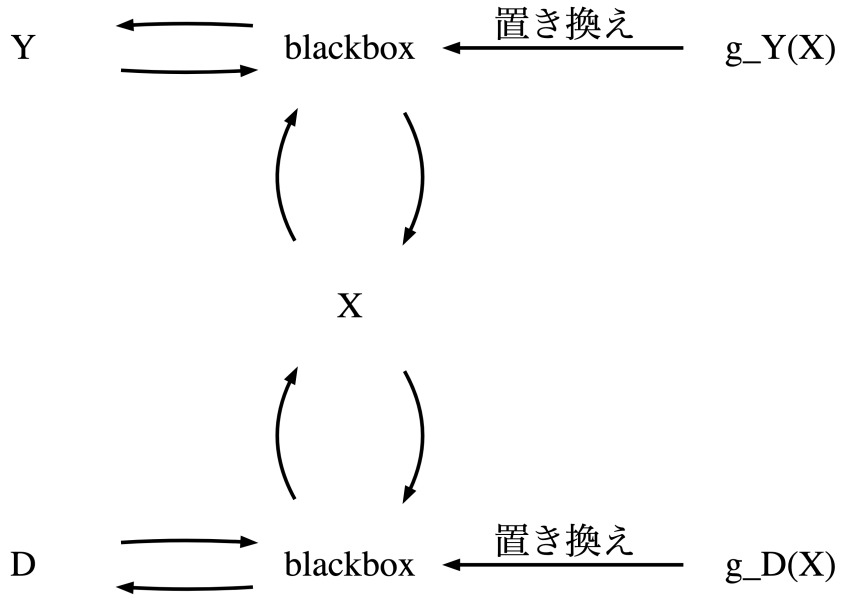
2.11 直感

- X の偏りを排除したい



2.12 直感

- 予測モデルにより、X の偏りの影響を評価する



2.13 例: 基本アイディア

```
# A tibble: 4 x 9
  Price  Size District StationDistance BuildYear PredY   ResY     D    ID
  <dbl> <dbl> <chr>              <dbl>     <dbl> <dbl>  <dbl> <dbl> <int>
1    92    55 港                    11      2015  92.0 0.0481     1     1
2    63    65 杉並                   4      1996  62.9 0.0504     1     2
3    22    25 板橋                   4      2013  21.9 0.0571     0     3
4    19    20 練馬                   4      2008  19.0 0.0332     0     4
```

- 事例 1 と 4 の間で、700 万円近くの 以上の価格格差
- X が大きく異なる
  - (X からの) 予測値についても、同程度の格差
- 実際の格差から、(X の影響を捉える) 予測値の差を引くとほとんど格差はない

2.14 直感

- 価格 = 取引年 + X の影響 + その他
- 予測モデルが X の影響を完全に捉え、その他の影響がなければ、

$$\underbrace{Y_1 - Y_2}_{\text{事例1と2の賃金格差}} - \underbrace{g_Y(X_1) - g_Y(X_2)}_{\text{予測格差}}$$

$$= \underbrace{Y_1 - g_Y(X_1)}_{1についての予測誤差} - \underbrace{(Y_2 - g_Y(X_2))}_{2についての予測誤差}$$

は、バランス後の賃金格差を捉える

2.15 直感

- その他の影響は、当然存在する
- 個別事例について、影響を排除することは困難
- 大量の事例の平均をとることで、平均格差を明らかにする

$$\underbrace{E[Y - g_Y(X)|D=1]}_{D=1についての平均予測誤差} - \underbrace{E[Y - g_Y(X)|D=0]}_{D=0についての平均予測誤差}$$

- $g_Y(X) = E[Y|X]$ であれば、バランス後の平均格差の優れた推定値

2.16 直感

- 限られた事例から推定された予測モデルは、理想的なものと乖離する: $g_Y(X) \neq E[Y|X]$
  - 格差の推定値に、予測困難な悪影響を与える
- D の予測モデル $g_D(X)$ も併用する
  - X が D に与える影響も極力排除する
- Y の予測誤差 $Y - g_Y(X)$ を D の予測誤差 $D - g_D(X)$ で回帰する

2.17 Double Debaised Machine Learning

- R learner は、Double/Debiased Machine Learning と呼ばれる手法の一つ
- 最終的には OLS や平均値などで推定する
  - 予測モデルの不完全性 (AI のミス) の悪影響を、複数の予測モデルを組み合わせることで減らす
  - 予測モデルが**同時に**母平均から大きく乖離しなければ OK (“AI による Double Check”)

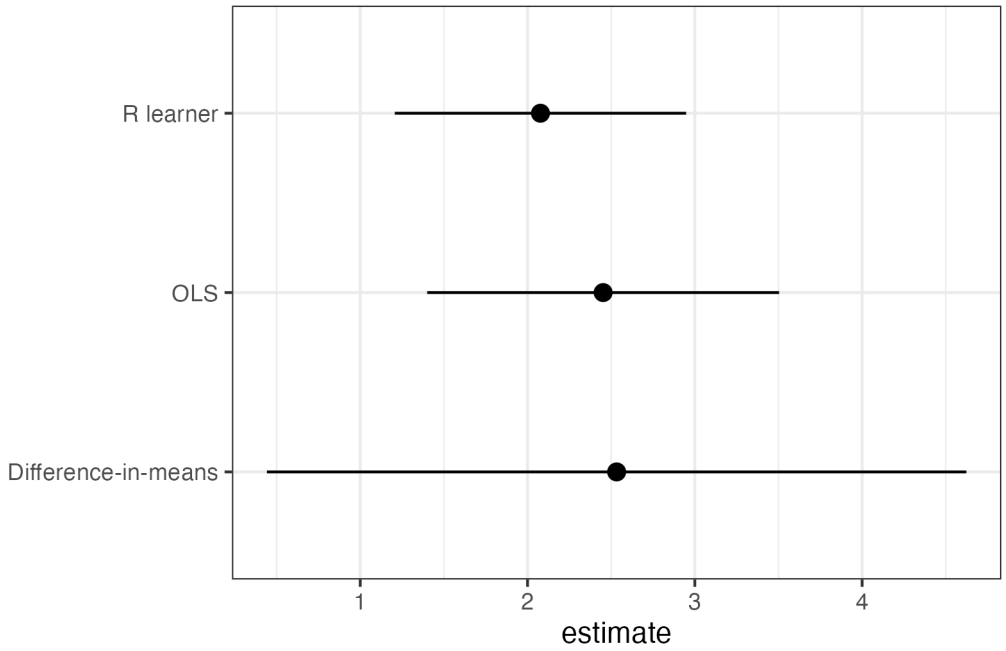
2.18 推定値の性質

- $g_Y(X), g_D(X)$ が $E[Y|X], E[D|X]$ をある程度近似できるのであれば (後述)、近似的な信頼区間を計算可能
  - 十分な事例数 (私見では 500 事例以上)

  - 適した推定方法の採用 (後日議論)

- 注: Y, D を高い精度で予測する必要はない
  - 個人差が大きく予測しにくい現象であったとしても、比較分析への活用においては、予測モデルは有益

2.19 実例



2.20 補論: OLS との比較

- 「X の分布を均した後の比較」を行う伝統的方法は、以下のモデルを重回帰する

$$E[Y|D, X] = \beta_0 + \beta_D D + \beta_1 X_1 + ...$$

- FWL 定理 (wiki) より、R-leaner の特殊ケースと見做せる

2.21 補論: OLS との比較

- FWL 定理より、OLS の推定結果 β_D は、以下の手順により得られる推定結果と一致

1. **全データ** を用いて、**OLS** で予測モデル $g_Y(X) = \beta_0 + \beta_1 X_1 + ..., g_D(X) = \alpha_0 + \alpha_1 X_1 + ..$ を推定
2. 以下は R-learner と同じ手順

- 予測モデルの定式化が正しく、かつ事例数に比べて十分に単純 (β の数が少ない) のであれば、優れた推定方法だが、そうではない応用例も多い

2.22 補論: 統計モデルとしての書き換え

- R-learner の推定結果は、以下のモデルを推定した場合の結果と一致 (Robinson 1988)

$$Y = \tau D + \underbrace{f(X)}_{\text{何らかの}(blackbox\text{な})\text{関数}} + \underbrace{u}_{E[u|D,X]=0}$$

  - 教科書的な線形モデル:「$f(X) = \beta_0 + \beta_1 X_1 + ..$ と特定化しても $E[u|D, X] = 0$ が成り立つ」という仮定を**追加**

2.23 まとめ

- シンプルな比較分析であれば、伝統的な推定方法が有効
- 複雑な比較分析、ここでは X をバランスさせた後での比較、においては機械学習も併用することで、分析の信頼性を改善できる
  - OLS などと比べて、推定結果が大きく変わらなかったとしても、分析の頑強性の確認ができる

Reference

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21 (1): C1–68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. 2022. "Locally Robust Semiparametric Estimation." *Econometrica* 90 (4): 1501–35.

Robinson, Peter M. 1988. "Root-n-Consistent Semiparametric Regression." *Econometrica: Journal of the Econometric Society*, 931–54.

Van der Laan, Mark J, Sherri Rose, et al. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Vol. 4. Springer.