OLS の比較分析への応用

労働経済学 2

川田恵介

Table of contents

1 復習: バランス後の比較

1.1 バランス後の比較

- 推定目標の一つ: D 間で Y の平均値を比較する
  - D 間での X の分布の違いは、推定上の操作として排除する
    - X についてバランスする
- 格差/因果/比較研究の中心的な推定対象

1.2 例: 職種間格差研究

- 研究目標: 1985 年の米国で、職種 (D; $1 =$ 専門職 (technical) と $0 =$ 他職種) 間の賃金 (Y) 格差
- 推定目標: 母集団における平均賃金格差

$$E[Y \mid D = 1] - E[Y \mid D = 0]$$

- 推定値: データ上の平均差 + 信頼区間

1.3 例: データの導入

```
library(tidyverse)

data("CPS1985", package = "AER")

data <- mutate(
  CPS1985,
  Y = log(wage), # log of wage
  D = if_else(
    occupation == "technical",
    1,
    0
  ) # occupation dummy
)
```

- 以下、ランダムサンプルングデータであることを仮定

1.4 例: データ上の平均差 + 信頼区間

```
estimatr::lm_robust(Y ~ D, data)
```

```
              Estimate Std. Error   t value      Pr(>|t|)  CI Lower  CI Upper
(Intercept) 1.9812222 0.02486856 79.667742 7.375784e-298 1.9323696 2.0300749
D           0.3965148 0.05098282  7.777419  3.864442e-14 0.2963624 0.4966671
             DF
(Intercept) 532
D           532
```

1.5 批判的検討

- 「職種間格差」というよりは、労働市場に参入する以前の属性 X の違いを反映しているのでは?
- $X =$ [年齢, "人種", "性別", 学歴, 地域]
- 新たな推定目標

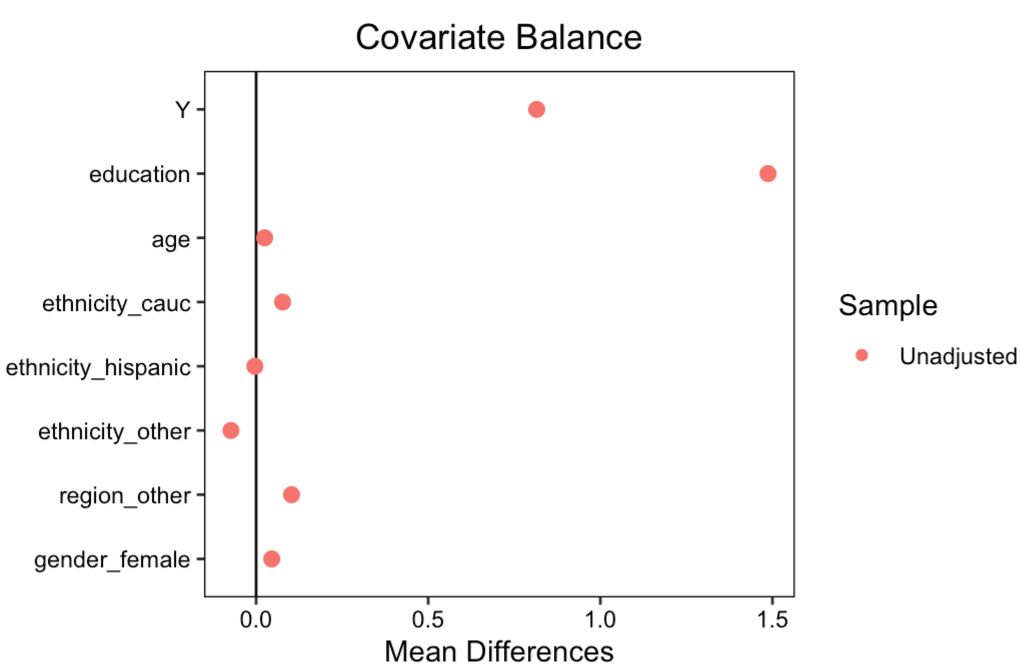
$$E[Y \mid D=1, X] - E[Y \mid D=0, X]$$

  の特徴把握

  - X の違いを **バランス** させる

1.6 例: 背景属性の違い

```
table <- cobalt::bal.tab(
  D ~ Y + education + age + ethnicity + region + gender,
  data)
cobalt::love.plot(table)
```



2 サブグループによる比較

2.1 サブグループによる比較

- X が全く同じ事例内で職種を比較
  - X の取りうる値の数が少ない場合有効

2.2 例: $X = gender$

```
      n      mean gender D
2   236  9.370805   male 0
429  53 12.773962   male 1
1   193  7.009637 female 0
432  52 11.105000 female 1
```

2.3 例: $X = gender$

```
# A tibble: 2 × 6
  gender mean_0 mean_1   n_0   n_1 difference
  <fct>   <dbl>  <dbl> <int> <int>      <dbl>
1 male     9.37   12.8   236    53       3.40
2 female   7.01   11.1   193    52       4.10
```

- 男性の中での差は、3.4
- 女性の中での差は、4.1

2.4 例: OLS の活用

```
estimatr::lm_robust(Y ~ D, data, subset = gender == "male") # 男性
```

```
             Estimate Std. Error   t value      Pr(>|t|)  CI Lower  CI Upper
(Intercept) 2.1021001 0.03447749 60.970213 2.698342e-166 2.0342393 2.1699609
D           0.3445401 0.07247817  4.753708  3.163793e-06 0.2018839 0.4871962
             DF
(Intercept) 287
D           287
```

```
estimatr::lm_robust(Y ~ D, data, subset = gender == "female") # 女性
```

```
             Estimate Std. Error   t value      Pr(>|t|)  CI Lower  CI Upper
(Intercept) 1.8334130 0.03284013 55.828428 1.345010e-140 1.7687253 1.8981006
D           0.4740957 0.06942158  6.829227  6.793686e-11 0.3373509 0.6108406
             DF
```

```
(Intercept) 243
D           243
```

2.5 サブグループの問題点

- 事例数が極端に少なくなり、比較ができなくなることが多い
  - 特に X に連続変数が含まれている/含まれている変数が多い
- X の組み合わせごとに、大量の推定値が算出され、人間が認識できなくなる

2.6 例: $X = education$

```
# A tibble: 17 × 6
   education mean_0 mean_1   n_0   n_1 difference
       <dbl>  <dbl>  <dbl> <int> <int>      <dbl>
 1         2   3.75  NA        1    NA     NA
 2         3   7     NA        1    NA     NA
 3         4   6     NA        1    NA     NA
 4         5  14     NA        1    NA     NA
 5         6   4.46  NA        3    NA     NA
 6         7   5.77  NA        5    NA     NA
 7         8   5.98  NA       15    NA     NA
 8         9   7.48   5.75    11     1     -1.73
 9        10   7.32  NA       17    NA     NA
10        11   6.58  NA       27    NA     NA
11        12   7.68   9.20   204    15      1.52
12        13   7.77   9.03    35     2      1.26
13        14  11.1   10.7     47     9     -0.392
14        15  10.9   10.4      7     6     -0.580
15        16   9.44  12.2     36    35      2.81
16        17  11.8   14.7      9    15      2.83
17        18  14.9   13.0      9    22     -1.94
```

3 重回帰によるバランス

3.1 伝統的アプローチ

- 伝統的な方法は、重回帰の活用
- 伝統的な定式化は、

  $Y \sim \beta_0 + \beta_D \times D + \beta_1 \times X_1 + \dots$

- 「適切な定式化」が前提

3.2 OLS の手順の復習

- OLS の手順は、

- Step 1. **研究者** が線型モデルの**定式化** $\beta_0 + \beta_1 X_1 + ..$ を設定
- Step 2. データへの不適合度を最小にするように、線型モデルのβを算出
  - データ上のY 平均値 に適合するように推定する

3.3 適切な定式化 (well-specification)

> ! Important
>
> - 仮定 (適切な定式化): β を適切に選べば、
>
> $E[Y \mid D, X] \approx \beta_0 + \beta_D \times D + \beta_1 \times X_1 + \dots$

- どんな X についても、

  $\beta_D \simeq E[Y \mid D = 1, X] - E[Y \mid D = 0, X]$

  - X が同じ回答者内での、職種間格差

3.4 不適切な定式化 (miss-specification)

- 例えば、

$$E[Y \mid D, age] = 10 \times D + age^2$$

  が本当母平均であるにもかかわらず

  - $\beta_0 + \beta_D \times D + \beta_1 \times age$ を推定すると、β_0, β_1 をどのように選んでも、

$$E[Y \mid D, age] \neq \beta_0 + \beta_D \times D + \beta_1 \times age$$

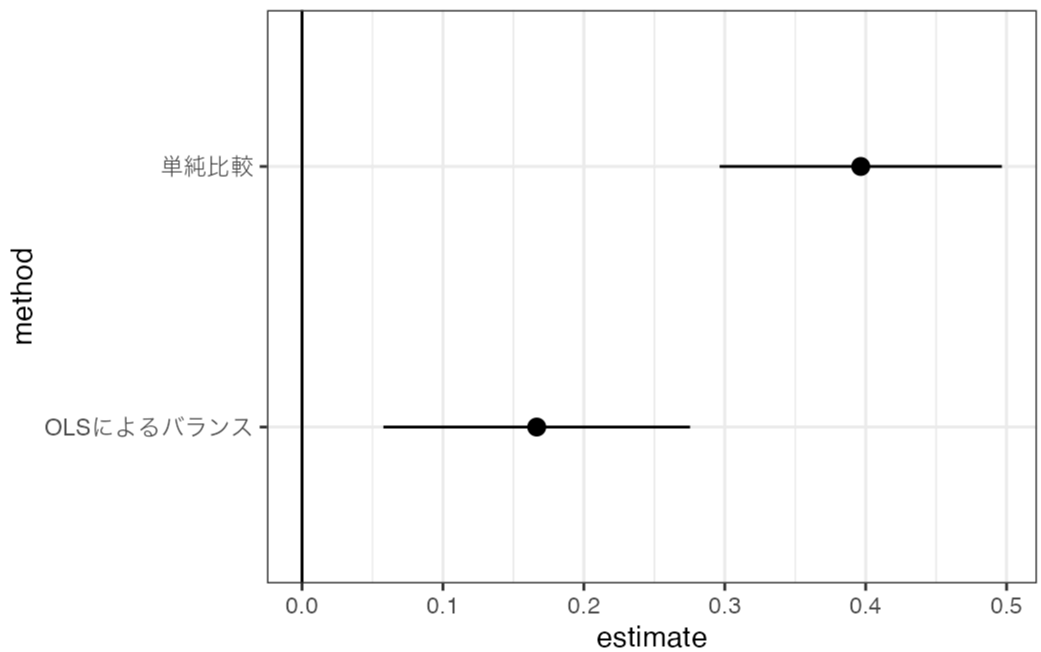
3.5 例

```
estimatr::lm_robust(
  Y ~ D + education + age + ethnicity + region + gender,
  data)
```

```
                     Estimate  Std. Error   t value     Pr(>|t|)    CI Lower
(Intercept)        0.76088407 0.135934777  5.597420 3.506784e-08  0.49384235
D                  0.16655554 0.055388034  3.007067 2.763930e-03  0.05774662
education          0.06777298 0.008716557  7.775200 3.995165e-14  0.05064944
age                0.01179519 0.001768916  6.668032 6.570682e-11  0.00832018
ethnicityhispanic -0.09990256 0.085345968 -1.170560 2.423058e-01 -0.26756337
ethnicityother    -0.07069243 0.056929066 -1.241763 2.148773e-01 -0.18252869
regionother        0.11712747 0.044214351  2.649083 8.313701e-03  0.03026907
genderfemale      -0.26200564 0.038766362 -6.758582 3.709207e-11 -0.33816154
                     CI Upper  DF
(Intercept)        1.02792580 526
```

```
D                   0.27536446 526
education           0.08489651 526
age                 0.01527019 526
ethnicityhispanic   0.06775825 526
ethnicityother      0.04114382 526
regionother         0.20398586 526
genderfemale       -0.18584973 526
```

3.6 例



3.7 Takeaway

- 適切な定式化の仮定のもとで、OLS はX のバランスを達成できる
- 例: “Y ~ D + education + age + ethnicity + region + gender” が適切な定式化であれば、 $X = [$ education,age,ethnicity,region,gender $]$ をバランスすると、職種間賃金格差は 40 % 程度から 18 % 割程度まで小さくなる

3.8 補論

- $Y \sim D + X_1 + X_2 + ..$ を OLS 推定する
- Random Sampling であれば、母平均の**線型近似モデル(BLP)** は推定できる
- 母平均**を**推定できるかどうかは、モデルの定式化に依存
- 不適切な定式化の下では

$$\text{推定されたモデル} \approx BLP \neq \underbrace{E[Y \mid D, X]}_{\text{推定対象}}$$

4 複雑化による解決と課題

4.1 複雑化による改善

- 二乗項や交差項などを加えると、適切な定式化に近づく
  - X と D の交差項も導入する
- 注: 実戦では、Entropy / Balance weight を用いるとことも強く推奨

4.2 例: 複雑化したモデルの推定

```
model <- estimatr::lm_robust(
  Y ~ D +
    D:(
      (education + age + ethnicity + region + gender)^2 + # Xとその交差項
    I(education^2) + I(age^2)
    ) + # DとXの交差
    (education + ethnicity + region + gender)^2 + # Xとその交差項
    I(education^2) + I(age^2),  # Xの二乗項
  data,
  se_type = "stata")
```

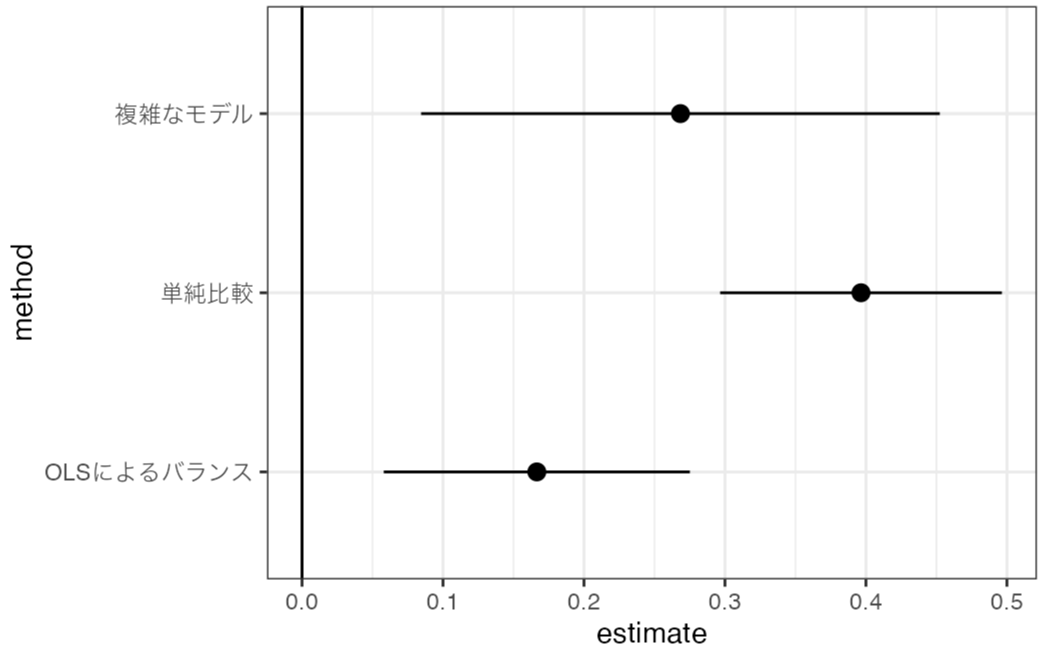
4.3 例: 複雑化したモデルの推定

- margins package を用いた推論

```
model |>
  margins::margins(variables = "D") |>
  summary()
```

```
 factor    AME     SE      z      p  lower  upper
      D 0.2684 0.0939 2.8599 0.0042 0.0845 0.4524
```

4.4 例



4.5 複雑化の弊害

- X の数が多くなれば、複雑化したモデルのβ の数は爆発的に増える
  - OLS での推定が難しくなる
    - Entropy / Balance weight を用いても同じ問題が生じる
- 実害: β の推定結果について
  - 推定誤差の拡大/信頼区間が信頼できなくなる
    - 推定値の分布が、正規分布へ十分に収束しない

4.6 複雑な推定の例

- K. Vafa, S. Athey, and D. M. Blei [1]
  - 詳細な職歴をバランスさせた後に、男女間賃金を比較
- A. Dube, J. Jacobs, S. Naidu, and S. Suri [2]
  - 詳細な職務内容をバランスさせた後に、求人の提示賃金と充足速度を相関を推定

4.7 Takeaway: OLS が適した場面

- 「研究者が設定した**単純なモデル**が、母平均の**適切な定式化**なのであれば」
  - 実用的な推定精度
  - + 信頼区間を用いた統計的推論

  - 母集団におけるバランス後の差は、“概ねこの範囲”という主張ができる
- X の数が少ないのであれば、交差項や二乗項を導入して利用できる

4.8 Takeaway: OLS の問題点

- X の数が多い場合、OLS での推定は難しい
- OLS はデータへの適合**のみ**を目指して推定するので
  - 過剰適合を抑えるには、Step 1 の段階で**研究者**が適切なモデルのを設定する必要がある
  - 実践では難しい

4.9 Reference

Bibliography

[1] K. Vafa, S. Athey, and D. M. Blei, “Estimating wage disparities using foundation models,” Proceedings of the National Academy of Sciences, vol. 122, no. 22, p. e2427298122, 2025.

[2] A. Dube, J. Jacobs, S. Naidu, and S. Suri, “Monopsony in online labor markets,” American Economic Review: Insights, vol. 2, no. 1, pp. 33–46, 2020.