Linear Model for Comparison

川田恵介

1 OLS の問題点

1.1 問題点一覧

1. 元々のXが多い場合に、十分に複雑なモデルを推定すると、推定精度が犠牲になる
2. 推定対象を定義する際に用いる Overlap weight の解釈が難しい
3. 負の Weight が生じ、非常にミスリーディングな結果が生じうる (Chattopadhyay and Zubizarreta, 2023)

- 問題 2 と 3 が、本日の論点

1.2 例: “昔の出席簿"データ

ID	Gender	TestScore
1	男性	60
2	男性	60
3	男性	60
4	女性	60
5	女性	60
6	女性	100

- 女性の平均点の方が高い
- 出欠番号は、"男性"からあいうえお順
  - “成績と関係ない”
  - 男女間で”バランスさせよう”がない

1.3 例: “昔の出席簿"データ

```
lm(TestScore ~ Gender + ID,
  data = Temp
)
```

```
Call:
lm(formula = TestScore ~ Gender + ID, data = Temp)

Coefficients:
(Intercept)   Gender男性           ID
      23.33        16.67        10.00
```

- ID をバランスさせると「男性の方が平均的が高くなる」
  - どうやってバランスさせているのか？

1.4 例: 問題点

ID	Gender	TestScore	Target
1	男性	60	−0.4166667
2	男性	60	0.3333333
3	男性	60	1.0833333
4	女性	60	1.0833333
5	女性	60	0.3333333
6	女性	100	−0.4166667

- マイナスの割合を目標 にする(???)ことで、平均値を 1.61 に「バランス」させている

1.5 例: 問題点

ID	Gender	TestScore
1	男性	10
2	男性	60
3	男性	60
4	女性	60
5	女性	60
6	女性	100

- 一番(男性)の成績が悪かったとする

1.6 例: 問題点

- 出席番号をバランスさせると、男性の成績が悪くなっているのに、男性の平均点が女性よりもさらに高くなる(!!?)

```r
lm(TestScore ~ Gender + ID,
  data = Temp
)
```

```
Call:
lm(formula = TestScore ~ Gender + ID, data = Temp)

Coefficients:
(Intercept)   Gender男性           ID
     -39.17        37.50        22.50
```

2 Direct balancing

2.1 Balancing weight の明示的な算出

1. 目標とする割合 $h(X)$ を明示的に指定
2. データ上のXの分布を$h(X)$と一致させる Weight$\omega(D, X)$を計算
   - 非負に限定
   - 何らかの基準に基づいて、散らばり方を最小化する
3. $\omega(D, X)$ を用いた平均差を計算する

2.2 例: Entropy weight (Hainmueller, 2012)

- 以下を最小化する

  $\omega(D, X) \times \log \omega(D, X)$**の平均値**

- ただし
  - $\omega(D, X) \geq 0$
  - $h(X) = \omega(d, X) \times f(X \mid D = d)$

2.3 代表的な目標割合

- データや母集団全体でのXの分布
  - 因果推論では、Average Treatment Effect を計算する際に使用
- $D = 1$ グループにおける X の分布
  - Average Treatment Effect on Treated を計算する際に使用

2.4 他の選択肢

- CBPS (Imai and Ratkovic, 2014), optitmal weight (Zubizarreta, 2015) など

- Entropy weight も含めて、WeightIt パッケージで容易に実装可能

2.5 実装: WeightIt

```
library(WeightIt)

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

WeightBalance <- weightit(
  married ~ education + age + ethnicity + gender, # G ~ X
  CPS1985, # Use DataClean
  method = "ebal", # Define EntropyWeight
  estimand = "ATE"
) # Define estimand
```

2.6 実装

```
WeightIt::lm_weightit(
  log(wage) ~ married,
  CPS1985,
  WeightBalance,
  vcov = "HC0"
) |>
  summary()
```

```


Call:
WeightIt::lm_weightit(formula = log(wage) ~ married, data = CPS1985,
    weightit = WeightBalance, vcov = "HC0")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.99614    0.04472  44.641   <1e-06 ***
marriedyes   0.09307    0.05271   1.766   0.0775 .
Standard error: HC0 robust
```

3 Reference

Bibliography

Chattopadhyay, A. and Zubizarreta, J. R. (2023) “On the implied weights of linear regression for causal inference,” Biometrika, 110(3), pp. 615–629

Hainmueller, J. (2012) “Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies,” Political analysis, 20(1), pp. 25–46

Imai, K. and Ratkovic, M. (2014) “Covariate balancing propensity score,” Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(1), pp. 243–263

Zubizarreta, J. R. (2015) “Stable weights that balance covariates for estimation with incomplete outcome data,” Journal of the American Statistical Association, 110(511), pp. 910–922