Linear Model for Comparison

川田恵介

1 比較研究

1.1 比較研究

- 「何らかの集団を比較し、重要な特徴を把握する」研究目標
  - 質/量的研究問わず、社会科学研究の中心的課題の一つ
- 例:
  - 雇用形態(正規/非正規)間での賃金格差
  - 学位が賃金に与える因果効果を明らかにするために、学位取得者/非取得者を比較する
  - 旧西ドイツと東ドイツを比較する

1.2 社会の線型モデルに比べた利点

- OLS/最尤法ともに、社会の近似的モデルを推定するモデルとして解釈できる
  - 比較研究と同様に、社会の特徴理解が目的
- 比較研究の方が、研究対象と推定対象が明確になる傾向
  - “モデル"という曖昧さを含む言葉を使わずに、研究対象や推定対象を定義できる

2 単純比較研究

2.1 単純比較研究

- **研究課題:** グループ $(D = 1, 0)$ 間の差異を明らかにする
- **推定課題:** 母集団における平均差

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

- **推定値:** データ上の平均差 $\mu(D = 1) - \mu(D = 0)$
  - $+$ 信頼区間
  - OLS も活用可能

2.2 例: “人種"間平均賃金格差

- 研究課題 = 労働市場政策を議論する土台として、"人種"間格差の現状を知りたい
- 推定課題 = 男女間平均賃金格差

$$E[\textbf{賃金} \mid afam] - E[\textbf{賃金} \mid cauc]$$

- 推定値 = データ上の平均差 $\mu(afam) - \mu(cauc)$
  - ＋ 信頼区間
  - OLS(単回帰)による実装も可能:

    $\textbf{賃金} \sim afam$ダミー

2.3 例: “人種"間平均賃金格差

```
data("CPS1988", package = "AER")

lm(wage ~ ethnicity, CPS1988)
```

```
Call:
lm(formula = wage ~ ethnicity, data = CPS1988)

Coefficients:
  (Intercept)  ethnicityafam
        617.2         -170.4
```

2.4 別解釈

- データ上でも母集団上でも、繰り返し平均値の公式より、

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

$$= \sum_X \{ \underbrace{E[Y \mid D=1, X]}_{X\text{内での平均値}} \underbrace{f(X \mid D=1)}_{X\text{の分布}}$$

$$- E[Y \mid D=0, X] f(X \mid D=0) \}$$

- 単純差 = X 内での平均差 + X の分布の差

2.5 実例: シンプルな比較

E[Y|D,X]	f(X|D)	ethnicity	education
545.1	0.619	cauc	12
403.1	0.761	afam	12
784.6	0.237	cauc	16

E[Y|D,X]	f(X|D)	ethnicity	education
572.6	0.161	afam	16
961.0	0.144	cauc	18
832.6	0.078	afam	18

- cauc の平均賃金 = 661.7 / afam の平均賃金 = 463.9

3 バランス後の比較

3.1 What If 分析の一種

- 研究課題: **もし**X の分布に差がなかった場合のY の平均差は?
  - 例: 格差分析: もし”ethnicity”間で教育年数の分布に差がなかった場合の賃金格差は?
  - 例: 因果効果:「仮想的なランダム化実験結果」を再現するためには、Xの分布を揃える必要がある
- X の分布を、**研究者が事前に設定した**目標割合に調整する

3.2 例: もし学歴分布が同じであれば?

E_Y	f(X|D)	ethnicity	education	目標割合
545.1	0.619	cauc	12	0.6
403.1	0.761	afam	12	0.6
784.6	0.237	cauc	16	0.2
572.6	0.161	afam	16	0.2
961.0	0.144	cauc	18	0.2
832.6	0.078	afam	18	0.2

- cauc の(調整後)平均賃金 = 676.2 / afam の平均賃金 = 522.9

3.3 Balancing Weight による実装

- 以下の手順でバランス後の比較は実装できる

1. Balancing Weight $w(D,X)$ = 目標割合/実際の割合
2. Weighted mean difference を計算

$$D=1\text{における}[w(D=1,X)\times Y]\text{の平均値}$$

$$-D=0\text{における}[w(D=0,X)\times Y]\text{の平均値}$$

3.4 例: もし学歴分布が同じであれば?

E[Y|D,X]	f(X|D)	ethnicity	education	目標割合	w
545.1	0.619	cauc	12	0.6	0.969
403.1	0.761	afam	12	0.6	0.788
784.6	0.237	cauc	16	0.2	0.844
572.6	0.161	afam	16	0.2	1.242
961.0	0.144	cauc	18	0.2	1.389
832.6	0.078	afam	18	0.2	2.564

3.5 バランス後の比較の課題

- X の数が多い場合、完璧なバランスは難しい
  - $D = 1$ または $= 0$しか存在しない組み合わせが発生
  - 極端に大きな Weight を付与する事例が発生
    - 推定精度が悪化
- 近似的なバランスを目指す
  - 多くの発展的手法(含む機械学習の応用)は、近似的バランスの一つの手法であると解釈できる

4 重回帰の別解釈

4.1 データ上での近似的な Balance

- $Y \sim D + X_1 + .. + X_L$ を OLS で推定した際の D の係数値 β_D は、以下の手順でも計算できる

1. データ上で、以下の性質を満たす$\omega(D, X)$を計算

- $(D_i = 1)$について$\omega(1, X) \times X_{i,l}$の平均

  $= (D_i = 0)$について$\omega(0, X) \times X_{i,l}$の平均

- 上記を満たす $\omega(d, x)$ のなかで、分散が最小

4.2 データ上での近似的な Balance

2. $\beta_D = (D_i = 1)$について$\omega(1, X) \times Y_i$の平均

   $-(D_i = 0)$について$\omega(0, X) \times Y_i$の平均

- X の**平均値**をバランスさせた上での、Y の平均差

- 近似的なバランス後の比較

- 詳細は、Chattopadhyay and Zubizarreta (2023)

4.3 例

```
data("CPS1988", package = "AER")

mean(CPS1988$wage[CPS1988$ethnicity == "cauc"])
```

```
[1] 617.2339
```

```
mean(CPS1988$wage[CPS1988$ethnicity == "afam"])
```

```
[1] 446.8526
```

- OLS を行っても OK

```
lm(wage ~ ethnicity, CPS1988)
```

```

Call:
lm(formula = wage ~ ethnicity, data = CPS1988)

Coefficients:
  (Intercept)  ethnicityafam
        617.2         -170.4
```

4.4 例

- lmw package を用いれば、OLS が算出する weight を計算できる

```
weight_ols <- lmw::lmw(~ ethnicity + education, CPS1988) |>
  magrittr::extract2("weights")
```

- weight を用いた平均

```
mean((weight_ols * CPS1988$wage)[CPS1988$ethnicity == "cauc"])
```

```
[1] 581.8683
```

```
mean((weight_ols * CPS1988$wage)[CPS1988$ethnicity == "afam"])
```

```
[1] 448.7225
```

4.5 例

- 重回帰の結果と一致

```
lm(wage ~ ethnicity + education, CPS1988)
```

```

Call:
lm(formula = wage ~ ethnicity + education, data = CPS1988)

Coefficients:
  (Intercept)  ethnicityafam      education
        9.888       -133.146         46.250
```

4.6 例

- weight を用いれると、学歴の平均値はバランス

```
mean(CPS1988$education[CPS1988$ethnicity == "cauc"])
```

```
[1] 13.1317
```

```
mean(CPS1988$education[CPS1988$ethnicity == "afam"])
```

```
[1] 12.32661
```

```
mean((weight_ols * CPS1988$education)[CPS1988$ethnicity == "cauc"])
```

```
[1] 12.38505
```

```
mean((weight_ols * CPS1988$education)[CPS1988$ethnicity == "afam"])
```

```
[1] 12.38505
```

4.7 さらなるバランス

- OLS を用いれば、"平均値"のみならず分散などもバランスできる
- $Y \sim D + X + X^2$ を推定すれば、Xの平均値と分散もバランス

- $Y \sim D + X_1 + X_2 + X_1^2 + X_2 + X1{*}X2$ を推定すれば、X_1, X_2の平均値と分散、共分散もバランス

4.8 補論: 分散/共分散

- 分散: X_1 のばらつきを捉える指標
  - $(X_1 - X_1$**の平均値**$)^2$ の平均値
- 共分散: X_1 と X_2 の相関関係を捉える指標
  - $(X_1 - X_1$**の平均値**$) \times (X_2 - X_2$**の平均値**$)$ の平均値

4.9 例: $Y \sim D + X$

E[Y|D,X]	f(X|D)	ethnicity	education	Omega	目標割合
545.1	0.619	cauc	12	1.210	0.749
403.1	0.761	afam	12	0.990	0.753
784.6	0.237	cauc	16	0.746	0.177
572.6	0.161	afam	16	1.026	0.165
961.0	0.144	cauc	18	0.514	0.074
832.6	0.078	afam	18	1.045	0.082

4.10 例: $Y \sim D + X + X^2$

E[Y|D,X]	f(X|D)	ethnicity	education	Omega	目標割合
545.1	0.619	cauc	12	1.216	0.753
403.1	0.761	afam	12	0.989	0.753
784.6	0.237	cauc	16	0.702	0.166
572.6	0.161	afam	16	1.030	0.166
961.0	0.144	cauc	18	0.563	0.081
832.6	0.078	afam	18	1.041	0.081

4.11 モデルの定式化

- 一般に分布をどこまでバランスさせれば十分なのか、よくわからない
  - 変数選択を活用しつつ、X の二乗項と交差項までをバランスさせる

4.12 母集団への含意

- データ上での OLS の推定結果は、「Population における OLS による平均値のバランス後の比較」の優れた推定値とみなせる

- OLS は母集団における OLS の優れた推定値であり、信頼区間も計算できる
- 事例数が、(組み合わせではなく) X の数に比べて、十分に多いことが前提
  - より多くの特徴をバランスさせようとすると、推定精度が悪化する

4.13 まとめ

- OLS は、X の分布の特徴を**バランス**する
- トレードオフ
  - より多くの特徴をバランスしようとすると、
  - 推定精度が悪化する
- 元々のXが多い場合、機械学習による変数選択の併用が有効 (次章)

5 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