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(入門書的な最尤法/ベイズ法)は、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、社会の特徴理解を目標とする研究に活用できる
    - 「社会のモデル」を推定し、「モデル」の特徴を理解する
- 比較研究は、研究対象と推定対象が明確に定義できる傾向
  - “モデル"という曖昧さを含む言葉を使わずに、多様な研究課題(因果効果や格差)について、推定目標を定義できる

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:

$$\textbf{賃金} \sim afam\textbf{ダミー}$$

2.3 例: “人種"間平均賃金格差

```r
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、

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

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

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

- 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

E[Y|D,X]	f(X|D)	ethnicity	education
403.1	0.761	afam	12
784.6	0.237	cauc	16
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 研究課題: What If?

- 研究課題: **もし**X の分布に差がなかったら?
  - X の分布を、**研究者が事前に設定した**割合 $h(X)$ に調整
- 幅広い研究目標における実質的な推定目標となる
  - 例: 格差分析: グループ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」を再現するためには、Xの分布を揃える
  - 例: 物価指数: もし消費する財の分布が変化しなかった場合の、生活費の差は?

3.2 例: もし学歴分布が同じならば?

E[Y|D,X]	f(X|D)	ethnicity	education	h(X)
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 $h(X)$/実際の割合 $f(X \mid D,X)$
2. Weighted mean difference を計算

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

$-D = 0$における$[w(D = 0, X) \times Y]$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	h(X)	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 まとめ: バランス後の比較研究

- **研究対象:** 因果効果/格差の推定等
- **推定対象:** X の分布を目標割合に調整した後の平均差

$$E[Y \mid D = 1, X] - E[Y \mid D = 0, X] \times \underbrace{h(X)}_{\text{目標割合}} \text{の和}$$

- **推定値:** Balancing weight $w(d, X)$ を利用し、

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

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

3.6 まとめ: 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
  - **推定値** が異なる

3.7 近似的なバランスの手法群

- 予想モデルの活用:

- **(Augmented) inverse probability weights/Residualized regression** (Van Der Laan and Rubin, 2006; Chernozhukov et al., 2018; 2022)

- 直接的な Balance:
  - **OLS**/LASSO (Chattopadhyay and Zubizarreta, 2023; Bruns-Smith et al., 2025), Entropy/Stable weight (Hainmueller, 2012; Zubizarreta, 2015), Auto debiased machine learing (Chernozhukov, Newey and Singh, 2022; Bruns-Smith, Dukes, Feller and Ogburn, 2025)

3.8 補論: 近似的なバランスの手法群

- バランス後の比較を行う方法として、$E[Y \mid d, X]$ を推定し、$E[Y \mid 1, X] - E[Y \mid 0, X]$ を計算、比較する手法が議論されてきた (Varian, 2014)
- 統計的性質を改善するために、Balancing weights と組み合わせる手法が有力視されている (Ben-Michael et al., 2021; Chernozhukov et al., 2024)
  - 直近では、$E[Y \mid d, X]$ を推定する手法は、「暗黙のうちに」Balancing weight を使用したバランス後の比較となることが議論されている (Chattopadhyay and Zubizarreta (2023); Bruns-Smith, Dukes, Feller and Ogburn (2025), Jared Murray (Youtube))

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

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

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

```
[1] 617.2339
```

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

```
[1] 446.8526
```

- OLS を行っても OK

```r
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

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

- weight を用いた平均

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

```
[1] 581.8683
```

```r
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 Moment のバランス

- OLS を用いれば、"平均値"のみならず分散なども(データ上で)バランスできる
  - 研究者が指定する、分布の特性値(Moment) 、をバランスさせる
- $Y \sim D + X + X^2$ を推定すれば、Xの平均値と分散もバランス
- $Y \sim D + X_1 + X_2 + X_1^2 + X_2^2 + X_1 \times X_2$ 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 Moment のバランス

- 非常に複雑にすれば、より多くの特性がバランスし、X の分布がほぼバランス
- ただしデータ上の目標割合は、"勝手に"Overlap weight が選ばれてしまう
  - Overlap weight $h^O(X) = E[D = 1 \mid X] \times E[D = 0 \mid X]$
  - 詳細は後述

4.9 例: $Y \sim D + X$

E[Y|D,X]	f(X|D)	ethnicity	education	w	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	w	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 母集団への含意

- これまでと同様に、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

4.12 まとめ: トレードオフ

- **推定値:** OLS は一般に

$$Y \sim D + \underbrace{f(X)}_{=\beta_0+\ldots+\beta_L X_L}$$

- **本来の推定目標:** X の分布を目標割合 $h(X)$ に完全に調整した後の、平均差
- $f(X)$**が十分に複雑な OLS の推定目標:** X の分布を overlap weight $h^O(X)$に完全調整した後の平均差
- **OLS の実質的な推定目標:** X の特徴(Moment)のみをバランスした後の平均差

4.13 まとめ: トレードオフ

- $f(X)$を(事例数に比べて)非常に複雑にすると
- $$\text{本来の推定目標} - \text{推定値} =$$
- $$\underbrace{\text{本来の推定目標} - f(X)\text{が十分な複雑な推定目標}}_{\text{不変}}$$
- $$\underbrace{f(X)\text{が十分な複雑な推定目標} - \text{実質的な推定目標}}_{\text{減少}}$$
- $$\underbrace{\text{実質的な推定目標} - \text{推定値}}_{\text{増加}}$$

4.14 まとめ: OLS の問題点

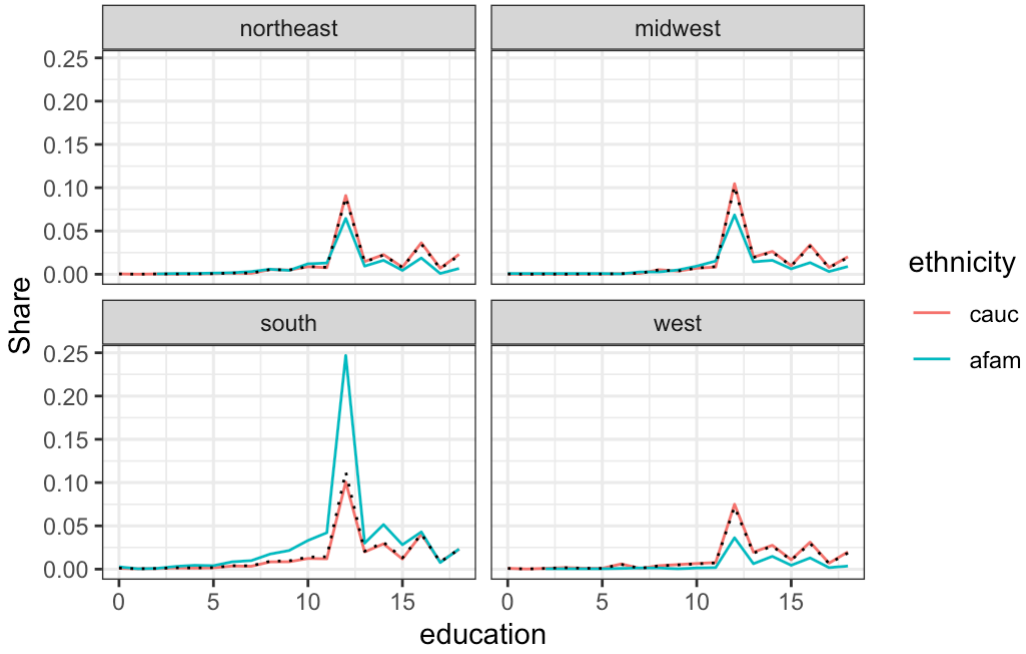
- 問題点 1. 元々のXが多い場合に、十分に複雑なモデルを推定すると、推定精度が犠牲になる
- 問題点 2. 推定対象を定義する際に用いる Overlap weight の解釈が難しい
- 問題点 3. 負の Weight が生じ、非常にミスリーディングな結果が生じうる (Chattopadhyay and Zubizarreta, 2023)

4.15 次回以降

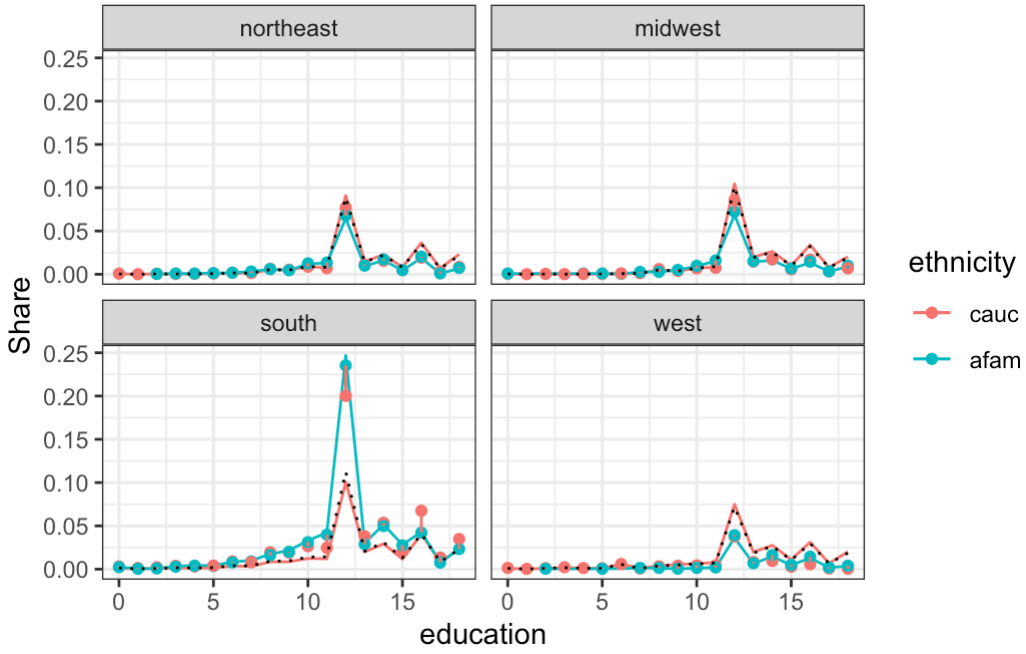
- 機械学習を応用することで、元々のXが多い場合に、十分に複雑なモデルを、推定精度を犠牲にせずに推定する手法を議論
  - 問題点 1,3 を緩和
  - OLS が持つ良い統計的性質 (中心極限定理や Bootstrap に基づく近似的な信頼区間計算) を保持するも目的

5 補論

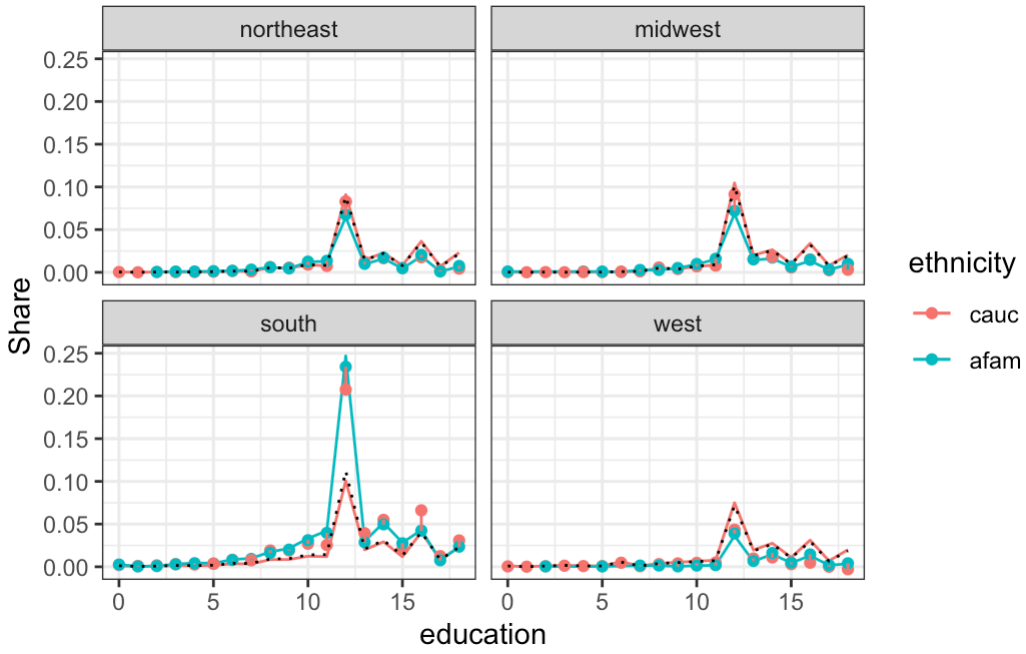
5.1 CPS1988 における”人種別”X の分布



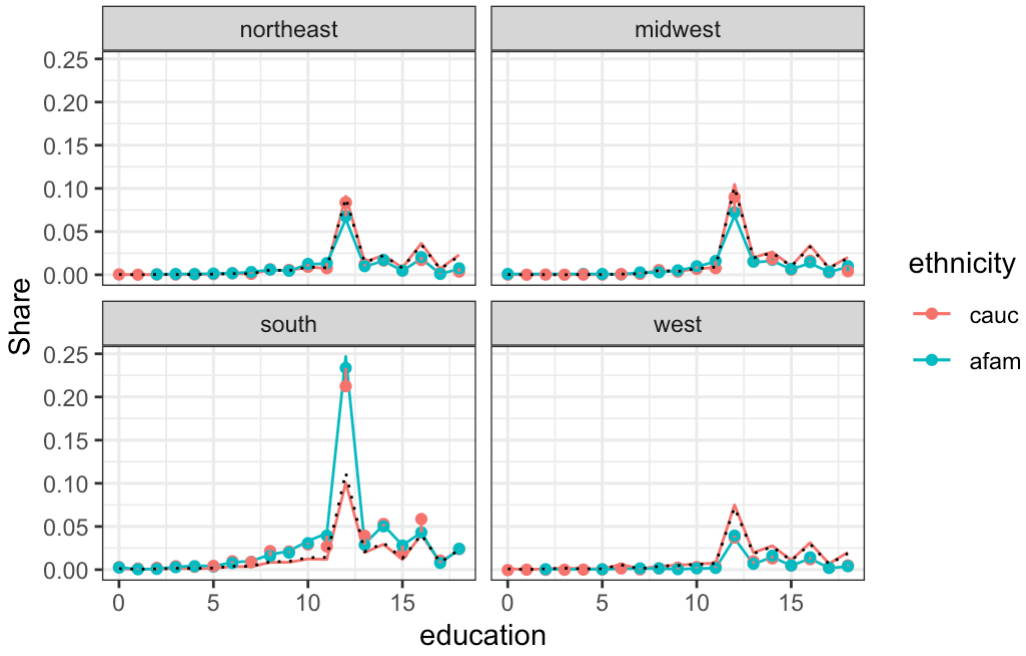
5.2 $Y \sim D + education + region$



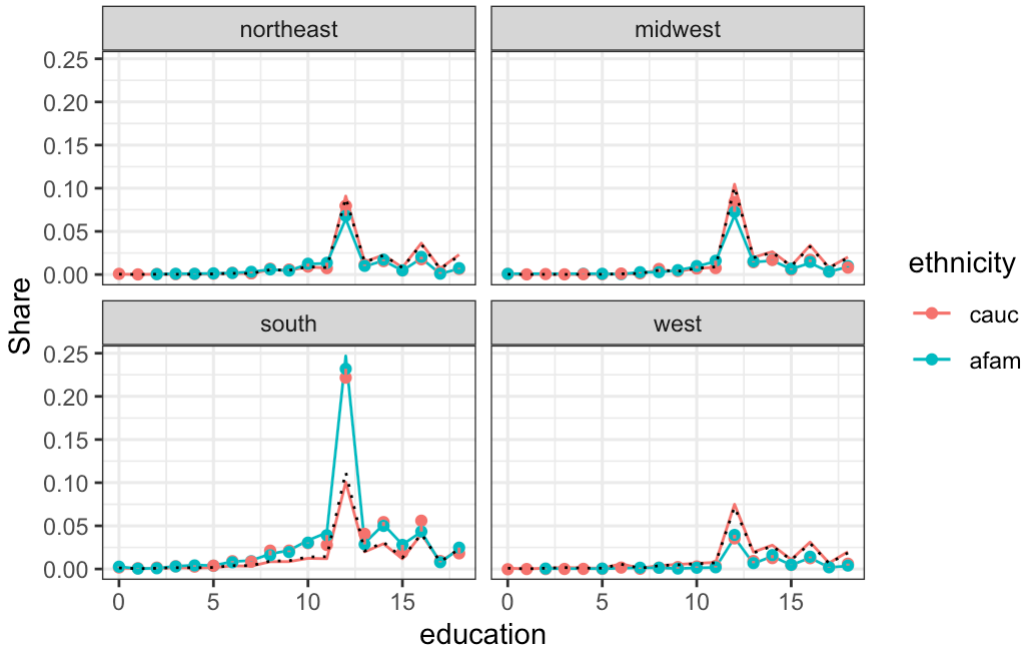
5.3 education の二乗項追加



5.4 交差項追加



5.5 二乗項との交差項追加



6 Reference

Bibliography

Ben-Michael, E. et al. (2021) “The balancing act in causal inference,” arXiv preprint arXiv:2110.14831 [Preprint]

Bruns-Smith, D. et al. (2025) “Augmented balancing weights as linear regression,” Journal of the Royal Statistical Society Series B: Statistical Methodology, p. qkaf19. Available at: https://doi.org/10.1093/jrsssb/qkaf019

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

Chernozhukov, V. et al. (2018) Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India

Chernozhukov, V. et al. (2022) “Locally robust semiparametric estimation,” Econometrica, 90(4), pp. 1501–1535

Chernozhukov, V. et al. (2024) Applied Causal Inference Powered by ML and AI. Available at: https://arxiv.org/abs/2403.02467

Chernozhukov, V., Newey, W. and Singh, R. (2022) “Automatic debiased machine learning of causal and structural effects,” Econometrica, 90(3), pp. 967–1027

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

Van Der Laan, M. J. and Rubin, D. (2006) “Targeted maximum likelihood learning,” The international journal of biostatistics, 2(1)

Varian, H. R. (2014) “Big data: New tricks for econometrics,” Journal of economic perspectives, 28(2), pp. 3–28

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