Regression-Based Proximal Causal Inference

Jiewen Liu[1], Chan Park[2], Kendrick Li[3], Eric J. Tchetgen Tchetgen[2]

[1]Department of Biostatistics, Perelman School of Medicine & [2]Department of Statistics and Data Science, Wharton School, University of Pennsylvania; [3]Department of Biostatistics, St. Jude Children's Research Hospital

ARXIV Link

Background & Proxy Variable

- **Confounding proxies (e.g. negative controls)** are increasingly used to detect unmeasured confounding U in observational studies.
- **Outcome confounding proxy (W)** refers to a variable that shares the same potential source of confounding bias as a treatment (A) - outcome (Y) of primary interest but is not causally related to the treatment (A).
- **Treatment confounding proxy (Z)** refers to a variable that shares the same potential source of bias as the (A)-(Y) relationship of primary interest but is not causally related to the outcome (Y).

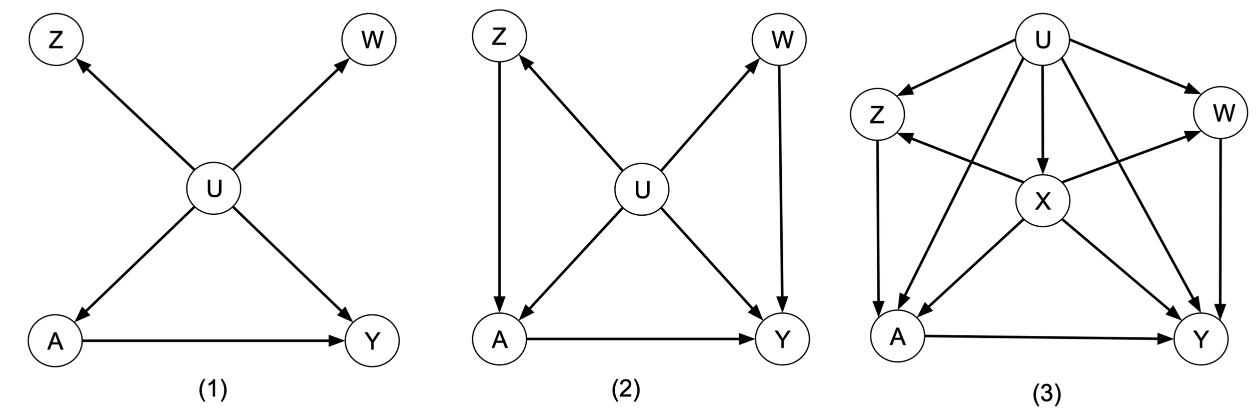

Figure 1: Three Common DAGs which PCI Applies to.

Previous Work of Proximal Causal Inference (PCI)

- Miao et al. (2018) studied the identification of causal effect with proxy variables. Tchetgen Tchetgen et al. (2023) developed proximal causal inference (PCI) to de-bias confounded causal effect estimates by leveraging a pair of proxy variables.
- However, implementing PCI involves solving complex integral equations that are typically ill-posed. Under linear models for outcome confounding proxy W and primary outcome Y, the proximal g-computation algorithm can be implemented by a two-stage OLS (see e.g. Tchetgen Tchetgen et al, 2023).

Continuous (Y, W) with Identity Links & Count (Y, W) with Log Links

Assumptions 1

$$E[Y|A,Z,U] = \beta_0 + \beta_a A + \beta_u U;\ E[W|A,Z,U] = \alpha_0 + \alpha_u U$$

Result 1

$$E[Y|A,Z] = \beta_0^* + \beta_a^* A + \beta_u^* E[W|A,Z],$$

$$\text{where } \beta_0^* = \beta_0 - \beta_u \frac{\alpha_0}{\alpha_u},\ \boldsymbol{\beta_a^* = \beta_a},\ \beta_u^* = \frac{\beta_u}{\alpha_u},\ \text{provided that } \alpha_u \neq 0.$$

Assumptions 2

$$\log(E[Y|A,Z,U]) = \beta_0 + \beta_a A + \beta_u U;\ \log(E[W|A,Z,U]) = \alpha_0 + \alpha_u U$$

$U|A,Z \sim E[U|A,Z] + \epsilon$; $E[\epsilon] = 0$; $\epsilon \perp\!\!\!\perp A, Z$; the marginal distribution of ϵ is unrestricted.

Result 2

$$\log(E[Y|A,Z]) = \beta_0^* + \beta_a^* A + \beta_u^* \log(E[W|A,Z]),$$

$$\text{where } \beta_0^* = \tilde{\beta}_0 - \beta_u \frac{\tilde{\alpha}_0}{\alpha_u},\ \boldsymbol{\beta_a^* = \beta_a},\ \beta_u^* = \frac{\beta_u}{\alpha_u},\ \text{provided that } \alpha_u \neq 0.$$

Results 1 and 2 suggest a two-stage linear regression and Poisson regression approach.

Binary (Y, W) with Logit Links

Assumptions 3

$$\text{logit}(\Pr(Y=1|A,Z,W,U)) = \beta_0 + \beta_a A + \beta_u U + \beta_w W$$

$$\text{logit}(\Pr(W=1|A,Z,Y,U)) = \alpha_0 + \alpha_u U + \alpha_y Y$$

$$U|A,Z,Y=0,W=0 \sim E[U|A,Z,Y=0,W=0] + \epsilon;\ E[\epsilon] = 0;$$

$\epsilon \perp\!\!\!\perp (A,Z)|Y=0, W=0$; the distribution of $\epsilon|Y=0, W=0$ is unrestricted.

Result 3

$$\text{logit}(\Pr(Y=1|A,Z,W)) = \beta_0^* + \beta_a^* A + \beta_u^* \text{logit}(\Pr(W=1|A,Z,Y=1)) + \tilde{\beta}_w W,$$

$$\text{where } \beta_0^* = \tilde{\beta}_0 - \beta_u \frac{(\tilde{\alpha}_0 + \tilde{\alpha}_y)}{\alpha_u},\ \boldsymbol{\beta_a^* = \beta_a},\ \beta_u^* = \frac{\beta_u}{\alpha_u},\ \text{provided that } \alpha_u \neq 0.$$

Result 3 suggests a two-stage logistic regression approach.

Implement PCI through Two-Stage Generalized Linear Models (GLMs)

We develop a two-stage regression approach to implement PCI

- (i): Applicable to continuous, count, and binary outcomes cases, when identity, log, logit link functions, or their combinations are applied. Relevant to a wide range of real-world applications.
- (ii): Easy to implement using off-the-shelf software for GLMs.

Y Data Type / W Data Type	Continuous (Identity Link)	Count (Log Link)	Binary (Logit Link)
Continuous (Identity Link)	Linear $W \sim A+Z$ $S = E[W\|A,Z]$ Linear $Y \sim A+S$	Linear $W \sim A+Z$ $S = E[W\|A,Z]$ Poisson $Y \sim A+S$	Linear $W \sim A+Z+Y$ $S = E[W\|A,Z,Y=1]$ Logistic $Y \sim A+S$
Count (Log Link)	Poisson $W \sim A+Z$ $S = \log(E[W\|A,Z])$ Linear $Y \sim A+S$	Poisson $W \sim A+Z$ $S = \log(E[W\|A,Z])$ Poisson $Y \sim A+S$	Poisson $W \sim A+Z+Y$ $S = \log(E[W\|A,Z,Y=1])$ Logistic $Y \sim A+S$
Binary (Logit Link)	Logistic $W \sim A+Z$ $S = \text{logit}(\Pr(W=1\|A,Z))$ Linear $Y \sim A+S+W$	Logistic $W \sim A+Z$ $S = \text{logit}(\Pr(W=1\|A,Z))$ Poisson $Y \sim A+S+W$	Logistic $W \sim A+Z+Y$ $S = \text{logit}(\Pr(W=1\|A,Z,Y=1))$ Logistic $Y \sim A+S+W$

Figure 2: S denotes the proximal control variable for U.

Application: Right Heart Catheterization (RHC) Treatment Effect

As error-prone snapshots of the underlying physiological state over time, physiological measurements (ph1, hema1) and (pafi1, paco21) are considered as confounding proxies (W) and (Z), respectively.

(W): ph1, hema1 encoded by 1 if greater than the median; $W=0$ if (ph1=0,hema1=0); $W=1$ if (ph1=1,hema1=0); $W=2$ if (ph1=0,hema1=1); $W=3$ if (ph1=1,hema1=1). **(Z)**: pafi1, paco21. (Y): 1 if the patient alive at 30th day. (A): 1 if the RHC is performed.

Two-stage logistic regression estimation:

$$\text{logit}(\Pr(W=k|A,Z,X,Y)) = \alpha_{0k}^* + \alpha_{ak}^* A + \alpha_{zk}^* Z + \alpha_{xk}^* X + \tilde{\alpha}_{yk} Y,\ \text{where } k \in \{1,2,3\},$$

$$\text{logit}(\Pr(Y=1|A,Z,X,W)) = \beta_0^* + \beta_a^* A + \beta_x^* X + \beta_u^* \sum_{k=1}^{3} \text{logit}(\Pr(W=k|A,Z,Y=1))$$

$$+ \sum_{k=1}^{3} \tilde{\beta}_{wk} I(W=k),\ \text{where } \boldsymbol{\beta_a^* = \beta_a}.$$

Estimates: $\hat{\beta}_a(Proximal) = -0.40\ (-0.56, -0.26)$, $\hat{\beta}_a(MLE) = -0.36\ (-0.51, -0.21)$.

Reference

- W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," Biometrika, vol. 105, no. 4, pp. 987–993, 2018.
- E. J. Tchetgen Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao, "An introduction to proximal causal learning," Statistical Science (2023).