Causal inference for complex observational data

Jianxuan Liu[1]

July 30, 2021

[1] Assistant Professor, Department of Mathematics, Syracuse University

Outline

- Causal inference
- Dimension reduction in Causal inference
- Measurement error issue in Causal inference

Causal inference

Causal questions: An action/treatment/intervention $\rightarrow$ an outcome. e.g. obesity $\rightarrow$ mortality

Notation: $A \rightarrow Y$

- Randomized experiments

Average treatment effect of treatment A on outcome Y

$$ATE = E(Y \mid A = 1) - E(Y \mid A = 0)$$

- Observational studies

$$ATE \neq E(Y \mid A = 1) - E(Y \mid A = 0)$$

Control variables $\mathbf{X}$ may have caused both Y and A, confounding the cause and effect relation between Y and A

Causal inference

The crux of establishing a causal relationship: ceteris paribus, i.e. holding all other factors fixed

Observational studies

- the ceteris paribus condition does not hold
- no guarantee that the change in Y is solely due to A

Causal inference

Potential outcomes framework (Neyman et al. 1990 and Rubin 1974)

A common assumption: no unmeasured confounding or ignobility (Rosenbaum & Rubin, 1983)

$$Y \perp A \mid \mathbf{X}$$

Propensity score

- conditional probability of A given the observed covariates
- a balancing score
- conditional on the propensity score, the distributions of the measured covariates are the same between treated and untreated subjects.
- adjust for propensity score can remove the confounding bias from the difference in covariates (Rosenbaum and Rubin, 1983)

Causal inference

Assumptions:

1. Positivity

- $\Pr(A \mid \mathbf{X})$ is bounded between 0 and 1
- ensure every subject has a nonzero probability to receive either treatment

2. Consistency

- Observed outcome $Y = AY(1) + (1 - A)Y(0)$ for $A \in \{0, 1\}$
- $Y(1) = Y(A = 1), Y(0) = Y(A = 0)$ are the potential outcomes

Causal inference

Estimation of the average causal effect

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{\pi_i} - \frac{(1 - A_i) Y_i}{1 - \pi_i} \right\}$$

where $\pi_i = \Pr(A_i \mid \mathbf{X_i})$ is the propensity score function.

High-Dimensional Data

One aspect of data complexity: large number of covariates, hard to interpret and visualize

Sparsity

- Only a few of the covariates have explanatory power, all the rest are redundant.
- Variable selection, often via penalization

Reducibility

- Only a few linear combinations of many covariates are useful
- (Sufficient) Dimension reduction

A **dimension reduction** $R(\mathbf{x})$ is said to be sufficient if the distribution of $y \mid R(\mathbf{x})$ is the same as that of $y \mid \mathbf{x}$

Dimension reduction

The distribution of Y relates to covariates $\mathbf{x}$ only through $\beta^T\mathbf{x}$, i.e.

$$Y \perp \mathbf{x} \mid \beta^T\mathbf{x}$$

Equivalently

$$\Pr(Y \leq y \mid \mathbf{x}) = \Pr(Y \leq y \mid \beta^T\mathbf{x}) \text{ for all } y$$

- Central space $S_{Y|\mathbf{x}}$: span of the columns in β
- Goal: estimate $S_{Y|\mathbf{x}}$

Estimation in Dimension Reduction Models

Target: find the column space of $\beta_{p\times d}$ with the smallest d

The smallest space exists and is uniquely defined (Cook, 2004)

Three classes of estimation approaches

1. Inverse regression based methods: Sliced inverse regression, **SIR** (Li, 1991), Sliced average variance estimation, **SAVE** (Cook and Weisberg, 1991) , direction regression **DR** (Li and Wang, 2007)
2. Nonparametric methods: Density based minimum average variance estimation, **dMAVE** (Xia, 2007), Sliced regression, **SR** (Wang and Xia, 2008)
3. Semiparametric methods (Ma and Zhu, 2012, Liu et al, 2018)

A New Robust estimator

- Not rely on the parametric specification of the propensity score model or the outcome regression model
- Data adaptive
- Handle many covariates simultaneously
- Covariates can be both continuous and discrete

The innovation

- only assume the treatment probability depends on the p-dimensional covariate vector $\mathbf{X}$ through several linear combinations $\beta^T\mathbf{X}$, $\beta \in \mathcal{R}^{p\times d}$, $d < p$
- Employ a nonparametric link function for the conditional probability

Flexible Estimation of the Propensity Score

Let $\pi(X) = P(A = 1 \mid \mathbf{X})$ be the propensity score function

$$\Pr(A = a \mid \mathbf{X} = \mathbf{x}) = \frac{\exp\{a\eta(\boldsymbol{\beta}^T\mathbf{x})\}}{1 + \exp\{\eta(\boldsymbol{\beta}^T\mathbf{x})\}}$$

- $\mathbf{X} \in \mathcal{R}^p$
- $\boldsymbol{\beta} \in {}^{p\times d}$, $\boldsymbol{\beta} = (\boldsymbol{I}_d, \boldsymbol{\beta}_l^T)^T$, $\boldsymbol{\beta}_l$ is an arbitrary $(p-d) \times d$ matrix
- η is an arbitrary unspecified function

Derivation of the efficient score function

The efficient score function: the residual after projecting the score vector w.r.t β onto the nuisance tangent space (Tsiatis 2006)

The estimating equation for β

$$\sum_{i=1}^{n} vecl\{\mathbf{x}_i - E(\mathbf{X}_i \mid \beta^T \mathbf{X}_i)\} \left[A_i - \frac{\exp\{\eta(\beta^T \mathbf{X}_i)\}}{1 + \exp\{\eta(\beta^T \mathbf{X}_i)\}} \right] \boldsymbol{\eta}'(\beta^T \mathbf{X}_i)^T = \mathbf{0}$$

Estimation of $E(\mathbf{X}_i \mid \beta^T \mathbf{X}_i)$

Nadaraya-Watson kernel estimator

$$\widehat{E}(\mathbf{X} \mid \beta^T \mathbf{X}) = \frac{\sum_{i=1}^n \mathbf{x}_i K_h(\beta^T \mathbf{X}_i - \beta^T \mathbf{X})}{\sum_{i=1}^n K_h(\beta^T \mathbf{X}_i - \beta^T \mathbf{X})}$$

- K: a multivariate kernel function, i.e. $K_h(\cdot) = K(\cdot/h)/h^d$.
- h: a bandwidth

Estimation of $\eta(\beta^T \mathbf{X}_i)$ and $\boldsymbol{\eta}'(\beta^T \mathbf{X}_i)$

Nonparametric kernel method

$$\sum_{i=1}^{n} \left[A_i - \frac{\exp\{b_0 + \mathbf{b}_1^T(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0)\}}{1 + \exp\{b_0 + \mathbf{b}_1^T(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0)\}} \right] K_h(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0) = 0$$

$$\sum_{i=1}^{n} \left[a_i - \frac{\exp\{b_0 + \mathbf{b}_1^T(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0)\}}{1 + \exp\{b_0 + \mathbf{b}_1^T(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0)\}} \right] (\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0) K_h(\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_0) = \mathbf{0}$$

The estimators $\widehat{b}_0$ and $\widehat{\mathbf{b}}_1$ are the estimators of η and $\boldsymbol{\eta}'$ at $\beta^T \mathbf{X}_0$, respectively. The efficient estimator of β solves the estimating equation

$$\sum_{i=1}^{n} vecl \left(\{\mathbf{x}_i - \widehat{E}(\mathbf{X}_i \mid \beta^T \mathbf{X}_i)\} \left[A_i - \frac{\exp\{\widehat{\eta}(\beta^T \mathbf{X}_i)\}}{1 + \exp\{\widehat{\eta}(\beta^T \mathbf{X}_i)\}} \right] \widehat{\boldsymbol{\eta}}'(\beta^T \mathbf{X}_i)^T \right) = \mathbf{0}$$

Simulation Studies

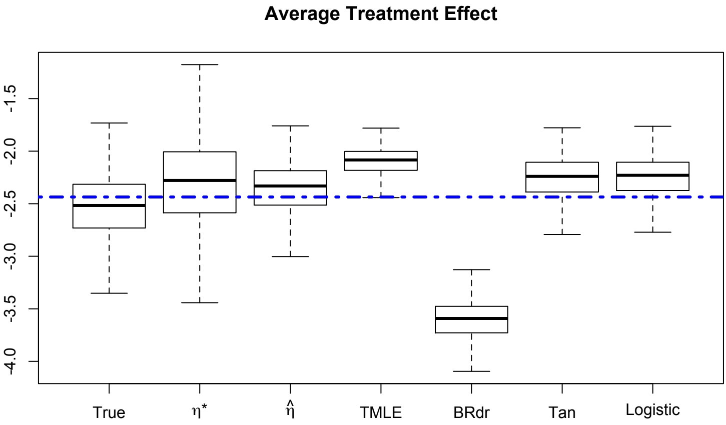

Figure 1: Scenario 1: No dimension reduction is available, $d = p$.

Simulation Studies

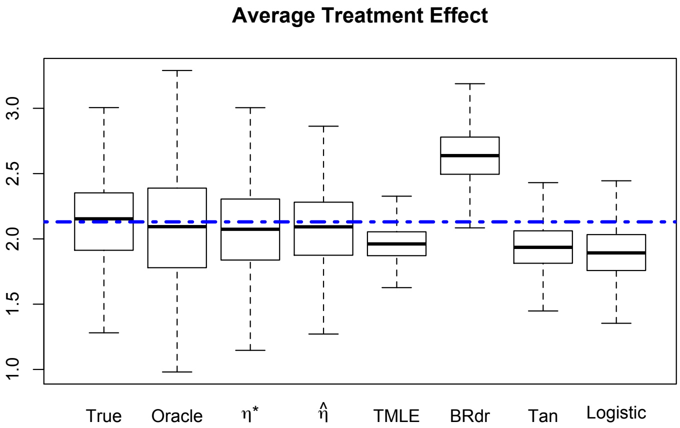

Figure 2: Scenario 2: Dimension can be reduced to $d = 1$.

Errors-in-variables

Propensity-score-based methods rely on the correct specification of propensity score modeling.

- have the correct model
- include all correctly measured covariates

Issue: some of the covariates are not measurable, or subject to measurement errors

Consequence: no unmeasured confounding assumption does not hold

Errors-in-variables

Data observed: $(Y_i, A_i, \mathbf{X}_i^*, \mathbf{Z}_i)$, iid $i = 1, \cdots, n$

$\mathbf{X}_i^*$ is the surrogate of the true covariates $\mathbf{X}_i$,

$$\mathbf{X}^* = \mathbf{X} + \mathbf{U}, \mathbf{U} \sim N(\mathbf{0}, \Omega)$$

Flexible propensity score model

$$\Pr(A_i \mid \mathbf{X}_i, \mathbf{Z}_i) = H\{\beta^T \mathbf{X}_i + \theta(\tilde{\gamma}^T \mathbf{Z}_i)\}$$

- $\tilde{\gamma} = (1, \gamma)$
- $H(\cdot)$ is the logistic distribution function
- $\theta(\cdot)$ is a nonparametric function of $\tilde{\gamma}^T \mathbf{Z}$.

Errors-in-variables

Direct use of error-prone covariates $\mathbf{X}^*$

- will not yield covariates balance on the underlying true covariates
- will not provide accurate treatment effect estimates
- McCaffrey et al. (2013)

Biased correction

$$\Pr(A = a \mid \Delta, \mathbf{Z}) = \frac{\exp[a\{\theta(\tilde{\gamma}^T\mathbf{Z}) + (\Delta - \Omega\beta/2)^T\beta\}]}{1 + \exp\{\theta(\tilde{\gamma}^T\mathbf{Z}) + (\Delta - \Omega\beta/2)^T\beta\}}$$

where $\Delta = \Delta(\mathbf{X}^*, A) = \mathbf{X}^* + A\Omega\beta$, a complete sufficient statistic of $\mathbf{X}$ (Stefanski and Carroll, 1987)

Derive the efficient estimating equations for β and γ

$$\sum_{i=1}^{n}[A_i - H\{(\Delta_i - \Omega\beta^T/2)\beta + \theta(\tilde{\gamma}^T\mathbf{Z})\}]E(\mathbf{X}_i \mid \Delta_i, \mathbf{Z}_i) = \mathbf{0}$$

$$\sum_{i=1}^{n}[A_i - H\{(\Delta_i - \Omega\beta^T/2)\beta + \theta(\tilde{\gamma}^T\mathbf{Z})\}]\mathbf{Z}_i^T\theta'(\tilde{\gamma}^T\mathbf{Z}_i) = \mathbf{0}$$

Unknown $E(\mathbf{X}_i \mid \Delta_i, \mathbf{Z}_i)$, $\theta(\cdot)$, and $\theta'(\cdot)$?

For $E(\mathbf{X}_i \mid \Delta_i, \mathbf{Z}_i)$

- directly propose a working model
- $E(\mathbf{X}_i^* \mid \Delta_i, \mathbf{Z}_i) = \Delta_i - E(A_i \mid \Delta_i, \mathbf{Z}_i)\Omega\beta$
- the conditional moment of $A_i \mid (\Delta_i, \mathbf{Z}_i)$ can serve as a practical guide choosing a working model

For $\theta(\cdot)$, and $\theta'(\cdot)$

- Solve nonparametrically

$$\sum_{i=1}^{n}[A_i - H\{(\Delta_i - \Omega\beta^T/2)\beta + \theta_0 + \theta_1\}][1\ \tilde{\gamma}^T\mathbf{Z}_i - \tilde{\gamma}^T\mathbf{Z}_j]^T K_h(\tilde{\gamma}^T\mathbf{Z}_i - \tilde{\gamma}^T\mathbf{Z}_j) = \mathbf{0}$$

$\widehat{\theta}_0, \widehat{\theta}_1$ are the estimator of θ and θ' at $\tilde{\gamma}^T\mathbf{Z}_j, j = 1, \cdots, n$.

$K_h(\cdot) = K(\cdot/h)/h$ is a multivariate kernel function with bandwidth h.

ATE estimation

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\widehat{\Pr}(A_i = 1 \mid \mathbf{X}_i^*, \mathbf{Z}_i, \widehat{\beta}, \widehat{\gamma}, \widehat{\theta}, E^*)} - \frac{(1 - A_i) Y_i}{\widehat{\Pr}(A_i = 0 \mid \mathbf{X}_i^*, \mathbf{Z}_i, \widehat{\beta}, \widehat{\gamma}, \widehat{\theta}, E^*)}$$

$\widehat{\Pr}(A_i = 1 \mid \mathbf{X}_i^*, \mathbf{Z}_i, \widehat{\beta}, \widehat{\gamma}, \widehat{\theta}, E^*)$ is the estimated propensity score.

Simulation studies

Setting 1:

- the true covariate **X** is univariate, $X \sim N(-1, 1)$
- with both low ($\sigma_U = 0.3$) and high ($\sigma_U = 0.9$) degree of measurement error.
- error-free $Z_j \sim \text{Uniform}(0, \pi)$ for $ j =1, 2, 3$
- the true parameters: $\beta = 0.7$, $\widetilde{\gamma} = (1.0, 0.3, 0.4)^T$

Simulation results - Setting 1

		Setting 1	$\sigma_U = 0.3$		
$\beta = 0.7$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.0602	-0.0043	0.0260	-0.0093	0.0148
emp.se	0.1254	0.1371	0.1500	0.1113	0.0906
est.se	0.1263	0.1375	0.1470	0.1330	0.0975
mse	0.0194	0.0188	0.0232	0.0125	0.0084
95% CI	0.9220	0.9400	0.9360	0.9600	0.9540
		Setting 1	$\sigma_U = 0.9$		
$\beta = 0.7$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.3260	-0.0352	-0.0425	-0.0868	-0.1493
emp.se	0.0954	0.1765	0.1796	0.1362	0.1474
est.se	0.0927	0.1676	0.1692	0.1514	0.1960
mse	0.1154	0.0324	0.0340	0.0261	0.0440
95% CI	0.0900	0.9340	0.9340	0.9420	0.9520

Simulation studies

Setting 2:

- $\mathbf{X}$ are multivariate, $X_1 \sim N(-1.1, 1), X_2 \sim N(1, 1)$
- degrees of measurement errors $\sigma_{U_1} = 0.5, \sigma_{U_2} = 0.9$,
- error-free $Z_j \sim \text{Uniform}(0, \pi)$ for $j = 1, \cdots, 8$
- the true parameters: $\boldsymbol{\beta} = (1.2, -0.8)^T$,
  $\widetilde{\gamma} = (1.0, 0.3, 0.4, 0.1, 0.2, -0.1, -0.2, 0.2)^T$.

Simulation results - Setting 2

$\beta_1 = 1.2$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.2865	-0.0654	0.0014	-0.2542	-0.4613
emp.se	0.1732	0.2178	0.2589	0.3078	0.2942
est.se	0.1529	0.1947	0.2221	0.6939	0.8520
mse	0.1121	0.0517	0.0670	0.1594	0.2993
95% CI	0.5220	0.8900	0.8780	0.9620	0.9120
$\beta_2 = -0.8$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	0.3860	0.0680	0.0763	0.3765	0.3702
emp.se	0.1394	0.2616	0.2595	0.2595	0.3121
est.se	0.1155	0.2130	0.2216	0.6533	0.8072
mse	0.1684	0.0731	0.0731	0.2091	0.2344
95% CI	0.2120	0.8700	0.8520	0.9620	0.9480

Simulation studies

Setting 3:

- **X** are multivariate,
  $X_1 \sim N(-0.9, 1), X_2 \sim N(-1.0, 1), X_3 \sim N(0.5, 1)$
- degrees of measurement errors, i.e.,
  $\sigma_{U_1} = 0.9, \sigma_{U_2} = 0.8, \sigma_{U_3} = 0.8$, respectively
- error-free $Z_j \sim \text{Uniform}(0, \pi)$ for $j = 1, \cdots, 4$
- the true parameters are $\boldsymbol{\beta} = (1.0, 0.8, -0.7)^T$,
  $\widetilde{\gamma} = (1.0, 0.3, 0.4, 0.5)^T$.

Simulation results - Setting 3

$\beta_1 = 1.0$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.4925	-0.0903	-0.0911	-0.2020	-0.1788
emp.se	0.1313	0.2405	0.2555	0.1825	0.1703
est.se	0.1131	0.2052	0.2141	0.2398	0.2108
mse	0.2598	0.0660	0.0736	0.0741	0.0610
95% CI	0.0980	0.8700	0.8540	0.9640	0.9380

$\beta_2 = 0.8$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.3520	-0.0739	-0.0722	-0.4169	-0.4402
emp.se	0.1440	0.2410	0.2503	0.2632	0.2190
est.se	0.1186	0.1932	0.1993	0.6696	0.7629
mse	0.1447	0.0635	0.0679	0.2431	0.2418
95% CI	0.2780	0.8300	0.8260	0.9760	0.9920

$\beta_3 = -0.7$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	0.3174	0.0765	0.0819	0.5776	0.6283
emp.se	0.1301	0.2143	0.2209	0.1303	0.1400
est.se	0.1174	0.1915	0.1988	0.6648	0.8728
mse	0.1177	0.0518	0.0555	0.3506	0.4144
95% CI	0.3120	0.8820	0.8760	0.9440	0.9580

Simulation studies

Setting 4:

- $\mathbf{X}$ are multivariate,
  $X_1 \sim N(-0.9, 1), X_2 \sim N(-1.0, 1), X_3 \sim N(0.5, 1)$, and
  $(X_4, Z_4) \sim \mathrm{MVN}\left(\begin{bmatrix}-0.5\\0\end{bmatrix}, \begin{bmatrix}1 & 0.2\\0.2 & 1\end{bmatrix}\right)$.
- degrees of measurement errors: $\sigma_{\mathbf{U}} = [0.9, 0.8, 0.8, 0.3]$.
- error-free $Z_j \sim \text{Uniform}(0, \pi)$ for $j = 1, \cdots, 3$
- the true parameters are $\boldsymbol{\beta} = (1.0, 0.8, -0.7, 0.7)^T$,
  $\widetilde{\gamma} = (1.0, 0.3, 0.4, 0.5)^T$.

Simulation results - Setting 4

$\beta_1 = 1.0$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.5028	-0.2221	-0.2428	0.1028	-0.5308
emp.se	0.2387	0.4290	0.4453	0.6324	0.4247
est.se	0.2111	0.3777	0.3975	0.5034	1.0862
mse	0.3098	0.2334	0.2572	0.4106	0.4621
95% CI	0.4600	0.8486	0.8486	0.9714	0.9571

$\beta_2 = 0.8$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.3758	-0.2123	-0.2420	-0.0241	-0.4183
emp.se	0.2502	0.4146	0.4149	0.6372	0.4164
est.se	0.2167	0.3630	0.3647	1.1909	1.4497
mse	0.2308	0.2169	0.2307	0.4066	0.3484
95% CI	0.5943	0.8629	0.8429	0.9886	0.9429

$\beta_3 = -0.7$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	0.3197	0.1324	0.1427	0.6487	0.7120
emp.se	0.2508	0.4104	0.4227	0.6209	0.3711
est.se	0.2184	0.3571	0.3678	1.6195	1.5186
mse	0.1651	0.1860	0.1990	0.8062	0.6446
95% CI	0.6743	0.8514	0.8429	0.9171	0.9314

$\beta_4 = 0.7$	Naive	RC-Norm	RC-Unif	semi-1	semi-2
bias	-0.1254	-0.0919	-0.0558	0.6280	0.5324
emp.se	0.2798	0.2997	0.3383	0.8308	1.2812
est.se	0.2773	0.2982	0.3357	1.1645	1.7563
mse	0.0940	0.0983	0.1176	1.0846	1.9250
95% CI	0.8600	0.8886	0.8914	0.9114	0.9514

Simulation results: ATE

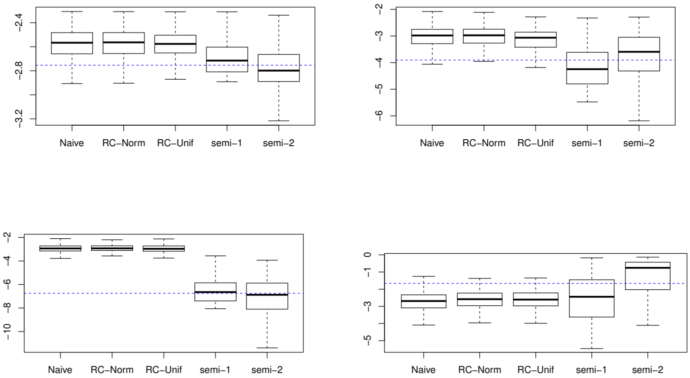

Figure 3: $\widehat{\tau}$ in simulation setting 1 (upper left), 2 (upper right), 3 (lower left) , and 4 (lower right). The dashed line is the true average treatment effect.

Conclusions

- A gentle introduction of causal inference
- Causal inference with sufficient dimension reduction
  - propose a new robust estimator
  - parametric models suffer from the risk of model misspecification, semiparametric models are more flexible
  - less prone to propensity score model misspecification
  - does not rely on the specification of the outcome regression model
  - attractive when a reliable outcome regression model is hard to obtain
  - capable to handle high dimension complex data
- Causal inference with errors-in-variables
  - propose a flexible semiparametric solution to evaluating the causal effects
  - errors-in-variables and subject to confounding
  - the dimension of the confounding variables can be large and correlated with the error-prone covariates.
  - the resulting estimators are locally efficient
  - estimators are efficient if the working model is correctly specified

References

1. Liu, J. and Li, W. (2021) A Semiparametric Method for Evaluating Causal Effects in the Presence of Error-Prone Covariates. *Biometrical Journal*. In press
2. Liu, J., Ma, Y. and Wang, L.(2018) An Alternative Robust Estimator of Average Treatment Effect in Causal Inference. *Biometrics*. Vol 74(3), 910-923
3. Neyman, J., Dabrowska, D., & Speed, T. (1990). On the application of probability theory to agricultural experiments: Essay on principles. *Statistical Science*, 5, 465–480.
4. Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
5. Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66, 688–701.
6. Stefanski, L., & Carroll, R. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74, 703–716.
7. Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Questions & comments?

Jianxuan Liu

jliu193@syr.edu

https://jliu193.expressions.syr.edu/home/


Շնորհակալություն
Teşekkür ederim
Terima kasih
Merci
ขอบคุณ
Gratiam ago
Cảm ơn
Grazie
धन्यवाद।
شكرًا
Спасибо
நன்றி
Mulţumesc
多謝
To-siā
Danke
धन्यवाद।
Asante
Děkuji
Ďakujem
Dankie
naṉṟi
Köszönöm
Tak
Gràcies
Thank you
Tack
唔該
Dankon
謝謝
Hvala
Takk
Diolch
dhanyavaad
감사합니다
Kiitos
谢谢
Faleminderit
Obrigado
Paldies
Bedankt
Dziękujemy
Mèsi
ありがとう
Ви благодарам
Gracias
Lóo-lȧt
Σας ευχαριστούμε