Proximal Causal Inference with Proxies: Valid and Invalid

ACIC 2025

Instructors

Chan Park

Assistant Professor at UIUC

Prabrisha Rakshit

Postdoc at UPenn

Eric Tchetgen Tchetgen

University Professor at UPenn

Outline

- 1st half: 8:30am - 10:20am

  Instructor: Chan Park

  Proximal causal inference with valid proxy variables

- Break: 10:20am - 10:30am

- 2nd half: 10:30am - 12:30am

  Instructor: Prabrisha Rakshit

  Proximal causal inference with invalid proxy variables

Outline of the First Half

- Negative Controls
  - Brief review of standard causal inference
  - Concepts of negative control exposures and negative control outcomes
  - Usage of negative controls in the literature
- Proximal Causal Inference
  - Identification & estimation of causal effects in the presence of unmeasured confounding
  - Application to the SUPPORT study
- Extensions of Proximal Causal Inference to Various Settings
  - Longitudinal settings
  - Mediation, front-door formula
  - Network data
  - Time series data
- Demonstration using R

Available at
github.com/qkrcks0218/ACIC2025

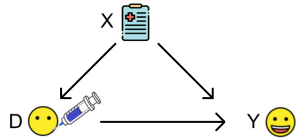
Negative Controls

Causal Inference

- Elements

  Treatment D Potential outcome $Y^{(d)}$

  Observed outcome Y (measured) Baseline covariate X

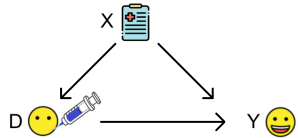


Causal Inference

- Elements

  Treatment D — Potential outcome $Y^{(d)}$

  Observed outcome Y — (measured) Baseline covariate X



- Flow of Causal Inference

Estimand

$$\tau^* = \mathrm{E}\big[Y^{(1)} - Y^{(0)}\big]$$

Define an estimand

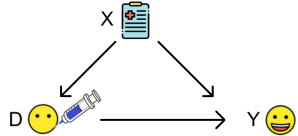
(eg. Average Treatment Effect (ATE); involves counterfactual, unobserved variables)

Causal Inference

- Elements

  Treatment D — Potential outcome $Y^{(d)}$

  Observed outcome Y — (measured) Baseline covariate X



- Flow of Causal Inference

Estimand

$$\tau^* = \mathrm{E}\left[Y^{(1)} - Y^{(0)}\right]$$

Define an estimand

(eg. Average Treatment Effect (ATE); involves counterfactual, unobserved variables)

- Fundamental problem of causal inference

  We only observe either $Y^{(1)}$ or $Y^{(0)}$, but not both

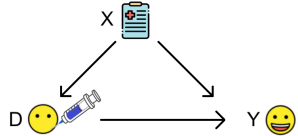
Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob	😀	1	30	😀	?
Sally	😫	0	25	?	😫

Causal Inference

- Elements

  Treatment D — Potential outcome $Y^{(d)}$

  Observed outcome Y — (measured) Baseline covariate X



- Flow of Causal Inference

Estimand

$$\tau^* = \mathrm{E}\big[Y^{(1)} - Y^{(0)}\big]$$

→

Identification

$$\tau^* = \mathrm{E}\big[\mathrm{E}(Y \mid D = 1, X) - \mathrm{E}(Y \mid D = 0, X)\big]$$

Define an estimand — (eg. Average Treatment Effect (ATE); involves counterfactual, unobserved variables)

Establish Identification — (eg. g-formula; find a representation of the ATE in terms of the observed data)

- Fundamental problem of causal inference

  We only observe either $Y^{(1)}$ or $Y^{(0)}$, but not both

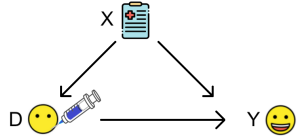
Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob	😀	1	30	😀	?
Sally	😫	0	25	?	😫

Causal Inference

- Elements

  Treatment D — Potential outcome $Y^{(d)}$

  Observed outcome Y — (measured) Baseline covariate X



- Flow of Causal Inference

Estimand

$$\tau^* = \mathrm{E}\big[Y^{(1)} - Y^{(0)}\big]$$

Identification

$$\tau^* = \mathrm{E}\big[\mathrm{E}(Y \mid D=1, X) - \mathrm{E}(Y \mid D=0, X)\big]$$

Estimation

$$\hat{\tau} = \texttt{aver}\big[\hat{\mathrm{E}}(Y \mid D=1, X_i) - \hat{\mathrm{E}}(Y \mid D=0, X_i)\big]$$

Define an estimand — (eg. Average Treatment Effect (ATE); involves counterfactual, unobserved variables)

Establish Identification — (eg. g-formula; find a representation of the ATE in terms of the observed data)

Construct an estimator — (eg. OLS; apply statistical methods to the observed data)

- Fundamental problem of causal inference

  We only observe either $Y^{(1)}$ or $Y^{(0)}$, but not both

Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob	😀	1	30	😀	?
Sally	😫	0	25	?	😫

Standard Assumptions

- Consistency: $Y = Y^{(D)}$

  The observed outcome $=$ one of the potential outcomes associated with the actually assigned treatment

Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob		1	30		?
Sally		0	25	?	

Standard Assumptions

- Consistency: $Y = Y^{(D)}$

  The observed outcome = one of the potential outcomes associated with the actually assigned treatment

- Positivity: $\Pr(D = d \,|\, X) > 0$

  Given X, each unit can be assigned to either treatment or control

Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob	😀	1	30	😀	?
Sally	😫	0	25	?	😫

Standard Assumptions

- Consistency: $Y = Y^{(D)}$

  The observed outcome = one of the potential outcomes associated with the actually assigned treatment

- Positivity: $\Pr(D = d \,|\, X) > 0$

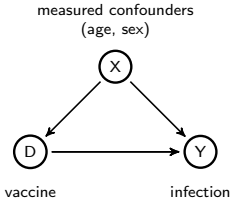
  Given X, each unit can be assigned to either treatment or control

- Most importantly, the assumption most relevant to this course:

- Ignorability / No Unmeasured Confounding: $Y^{(d)} \perp\!\!\!\perp D \,|\, X$

  All common causes of Y and D are measured
  Treatment is randomized within each stratum of X
  Not empirically verifiable

Unit	Y	D	X (age)	$Y^{(1)}$	$Y^{(0)}$
Bob	😀	1	30	😀	?
Sally	😫	0	25	?	😫



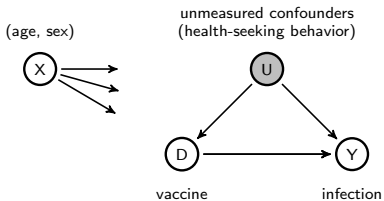
Unmeasured Confounding

- Hereafter all arguments are made implicitly conditional on X

- Unmeasured confounder U

  An **unmeasured** common cause of D and Y

  At the center of much skepticism about observational studies

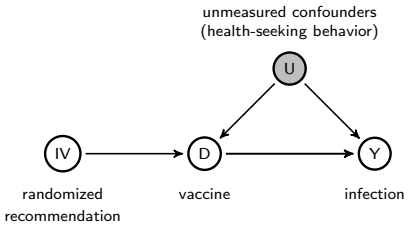


A Well-known Approach: Instrumental Variables

- An Instrumental Variable (IV) satisfies:

1. Relevance: $\text{IV} \not\perp\!\!\!\perp D$
2. Exclusion Restriction: $Y^{(\text{iv},d)} = Y^{(d)}$
3. Independence: $\text{IV} \perp\!\!\!\perp U$

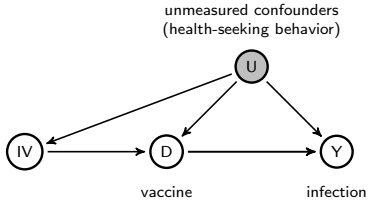
- An ideal IV can be obtained from a randomized assignment



A Well-known Approach: Instrumental Variables

- An Instrumental Variable (IV) satisfies:

1. Relevance: $\text{IV} \not\perp\!\!\!\perp D$
2. Exclusion Restriction: $Y^{(\text{iv},d)} = Y^{(d)}$
3. Independence: $\text{IV} \perp\!\!\!\perp U$



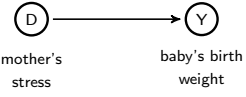
- An ideal IV can be obtained from a randomized assignment

- In observational studies, the independence assumption is often difficult to justify, as it is generally difficult to rule out the possibility that a given variable is correlated with U

- A hidden treasure: negative control variables and proxy variables

Motivating Example: Does stress during pregnancy affect birth weight?[6, 7][1]

- Observational study on effect of mother's stress on birth weight

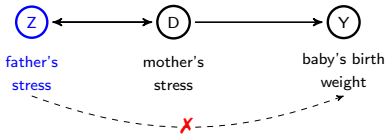


[1] Davey Smith (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*

Davey Smith (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies." *Epidemiology*

Motivating Example: Does stress during pregnancy affect birth weight?[6, 7][1]

- Observational study on effect of mother's stress on birth weight
- No causal effect of father's stress after adjusting for mother's stress
- $\texttt{lm}(Y \sim D + Z)$ and study the effect of Z



[1] Davey Smith (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*

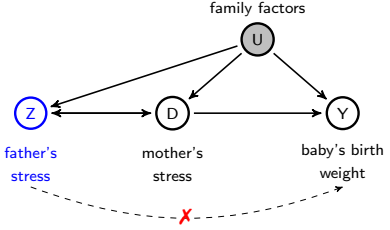
Davey Smith (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies." *Epidemiology*

Motivating Example: Does stress during pregnancy affect birth weight?[6, 7][1]

- Observational study on effect of mother's stress on birth weight

- No causal effect of father's stress after adjusting for mother's stress
- $\texttt{lm}(Y \sim D + Z)$ and study the effect of Z

  Nonzero effect of father's stress indicates hidden bias



- Family factors could be an unmeasured confounder

[1] Davey Smith (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*

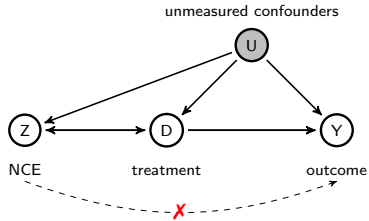
Davey Smith (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies." *Epidemiology*

Negative Control Exposure (NCE)

- Z is an NCE if $\boxed{Z \perp\!\!\!\perp Y \mid (U, D)}$

  Z does not causally affect Y

  Z is associated with Y only through U conditional on D

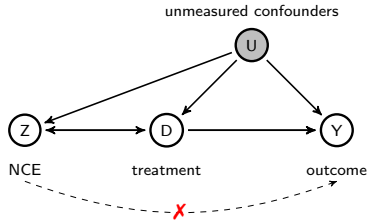


Negative Control Exposure (NCE)

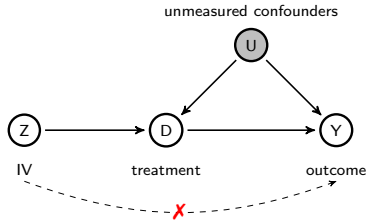
- Z is an NCE if $\boxed{Z \perp\!\!\!\perp Y \mid (U, D)}$

  Z does not causally affect Y

  Z is associated with Y only through U conditional on D



- A valid IV is a valid NCE

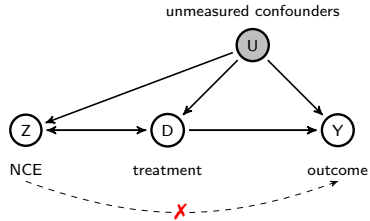


Negative Control Exposure (NCE)

- Z is an NCE if $\boxed{Z \perp\!\!\!\perp Y \mid (U, D)}$

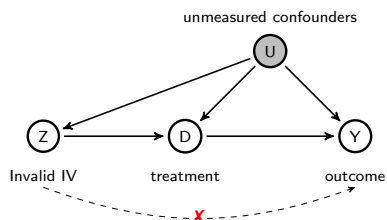
  Z does not causally affect Y

  Z is associated with Y only through U conditional on D



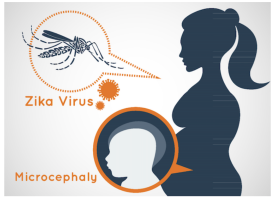
- A valid IV is a valid NCE

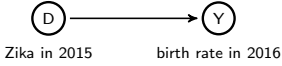
- An invalid IV also can be a valid NCE



Motivating Example: Did the 2015 Zika Virus Outbreak in Brazil Lead to a Decrease in Birth Rate? [23][1]

- Observational study on 2015 Zika virus outbreak in Brazil

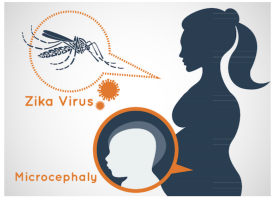




[1] Park, Richardson, TT (2024). Single proxy control. *Biometrics*

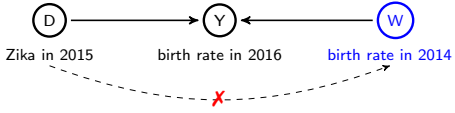
Motivating Example: Did the 2015 Zika Virus Outbreak in Brazil Lead to a Decrease in Birth Rate? [23][1]

- Observational study on 2015 Zika virus outbreak in Brazil



- No causal effect of 2015 Zika on 2014 birth rate

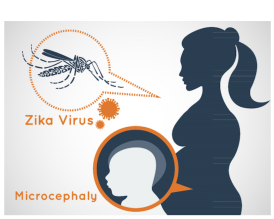
- $\texttt{lm}(W \sim D)$ and study the effect of D



[1] Park, Richardson, TT (2024). Single proxy control. *Biometrics*

Motivating Example: Did the 2015 Zika Virus Outbreak in Brazil Lead to a Decrease in Birth Rate? [23][1]

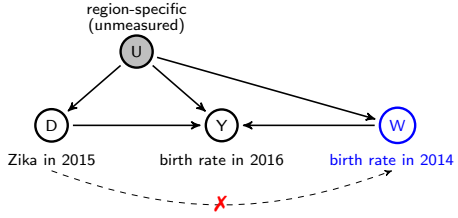
- Observational study on 2015 Zika virus outbreak in Brazil



- No causal effect of 2015 Zika on 2014 birth rate

- $\texttt{lm}(W \sim D)$ and study the effect of D

  nonzero effect of Zika indicates hidden bias



- Potential unmeasured confounding by region-specific variables

  healthcare infrastructure, socioeconomic status,
  cultural perspectives on childbirth

[1] Park, Richardson, TT (2024). Single proxy control. *Biometrics*

Negative Control Exposure (NCE) and Negative Control Outcome (NCO)

- Z is an NCE if $\boxed{Z \perp\!\!\!\perp Y \,|\, (U, D)}$

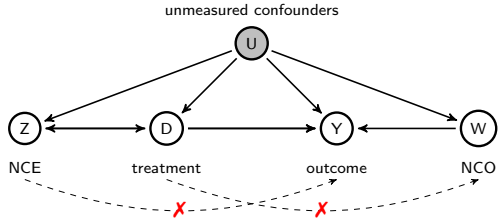
  Z does not causally affect Y

  Z is associated with Y only through U conditional on D

- W is an NCO if $\boxed{W \perp\!\!\!\perp (D, Z) \,|\, U}$

  W is not causally affected by D

  W is associated with (D, Z) only through U



Negative Control Exposure (NCE) and Negative Control Outcome (NCO)

- Z is an NCE if $\boxed{Z \perp\!\!\!\perp Y \,|\, (U, D)}$

  Z does not causally affect Y

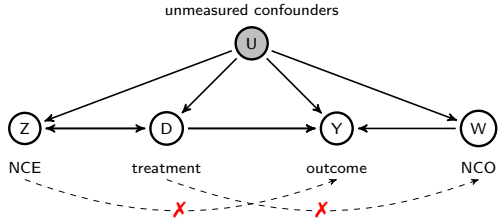
  Z is associated with Y only through U conditional on D

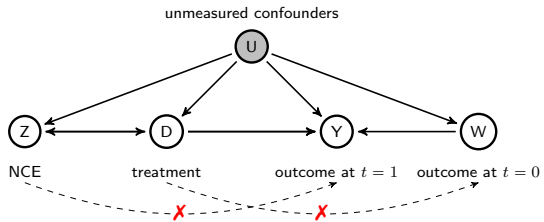
- W is an NCO if $\boxed{W \perp\!\!\!\perp (D, Z) \,|\, U}$

  W is not causally affected by D

  W is associated with (D, Z) only through U

- A pre-treatment outcome is a good NCO candidate

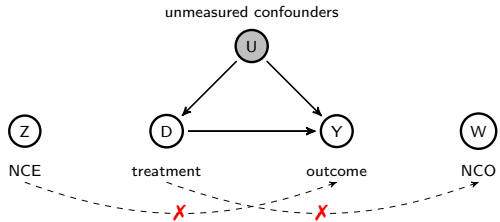




U-Comparability

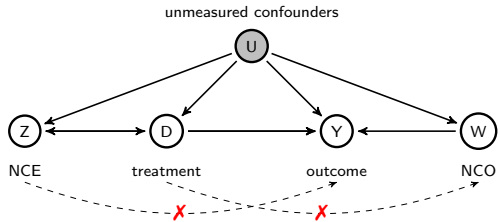
- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D)$
  W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, U$

- Toss a coin, and use the result as Z, W (?)



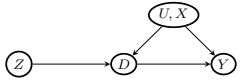
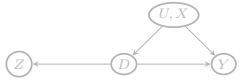
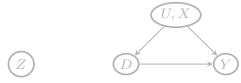
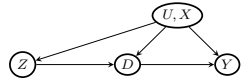
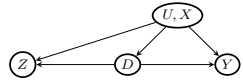
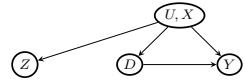
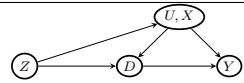
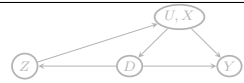
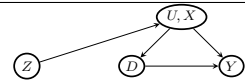
U-Comparability

- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D)$
  W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, U$
- Toss a coin, and use the result as Z, W (?)
- A variable completely irrelevant to the problem would not provide any useful information
- Unmeasured confounding mechanism of NCs should be comparable to that of D and Y
- **U-comparability**: $\boxed{Z \not\perp\!\!\!\perp U \,|\, D}$ and $\boxed{W \not\perp\!\!\!\perp U}$



Examples Encoding NC Assumptions [31][1]

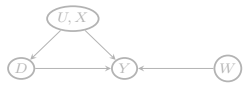
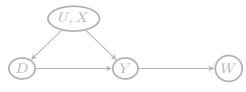
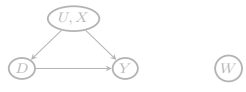
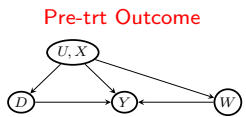
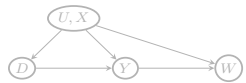
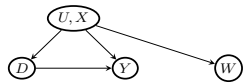
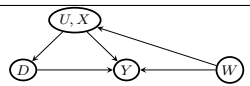
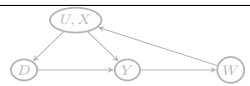
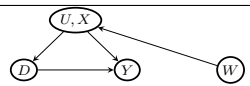
- Gray indicates violation of NC assumptions $Z \perp\!\!\!\perp Y \,|\, (U, D)$ and $Z \not\perp\!\!\!\perp U \,|\, D$

Examples of NCE			
	$Z \to D$ (pre-treatment)	$D \to Z$ (post-treatment)	$Z \perp\!\!\!\perp D \,|\, (U, X)$
No arrow between U and Z (may violate U-comparability)	Instrumental variable (IV) 		
$U \to Z$	Invalid IV 		
$Z \to U$	May violate Assumptions if there is $W \to U$ (collider bias) 		

[1] Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Examples Encoding NC Assumptions [31][1]

- Gray indicates violation of NC assumptions $W \perp\!\!\!\perp (D, Z) \,|\, U$ and $W \not\perp\!\!\!\perp U$

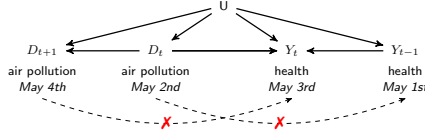
Examples of NCO			
	$W \to Y$ (pre-treatment)	$Y \to W$ (post-treatment)	$Y \perp\!\!\!\perp W \mid (D, U, X)$
No arrow between U and W (may violate U-comparability)			
$U \to W$	Pre-trt Outcome 		
$W \to U$	May violate Assumption if there is $Z \to U$ (collider bias) 		

[1] Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Negative Controls Are Widely Available

- Air pollution and health outcomes: the future ⇏ the past [10][1]

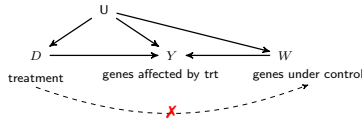
  NCE = future exposure; NCO = past outcome



- Genetics research and batch effect [16][2]

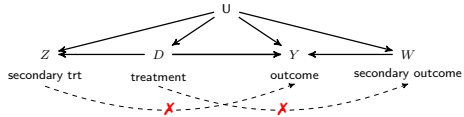
  Use control genes (W) to remove unwanted variation

$$W = \gamma U + \epsilon_W \ , \ Y = \beta D + \gamma U + \epsilon_Y$$



- Drug/vaccine comparative effectiveness and safety [29][3]

  Use secondary treatments or outcomes in electronic health records
  Obtain the p-values from $\texttt{lm}(Y \sim Z)$ and $\texttt{lm}(W \sim D)$
  Calibrate the p-values from $\texttt{lm}(Y \sim D)$ to correct for the deflation



[1] Flanders et al. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*

[2] Jacob et al. (2016) Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*

[3] Schuemie et al. (2014) Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*

Using Negative Controls to Detect, Reduce, and Correct Bias

Detect	[10][1]	Time-series study
	[7, 39][2]	invalid NCE
Reduce	[11, 22][3]	Time-series study
	[26][4]	Standardized mortality ratio in occupational cohort study
	[28, 29][5]	Drug–outcome pairs with no plausible causal effect
Debias	[27, 36][6]	Time-to-event outcome
	[15, 34][7]	Generalized difference-in-differences using NCO
	[23, 35][8]	Calibration using NCO
	[12, 16, 38][9]	Removing unwanted variation in gene-expression analysis
	[5, 20, 21] [10]	Nonparametric identification using double negative control/proximal causal inference

1 Flanders et al. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*

2 Davey Smith (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies." *Epidemiology*
Weisskopf, TT, Raz (2016). Commentary: on the use of imperfect negative control exposures in epidemiologic studies. *Epidemiology*

3 Flanders, Strickland, Klein (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *Am J Epi*
Miao, TT (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *Am J Epi*

4 Richardson et al. (2015). Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology*

5 Schuemie et al. (2014) Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*
Schuemie et al. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *PNAS*

6 Richardson et al. (2014). Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *Am J Epi*
TT, Sofer, Richardson (2015). NCO for unobserved confounding under a Cox proportional hazards model. *preprint*

7 Sofer et al. (2016). On negative outcome control of unobserved confounding as a generalization of DiD. *Statistical Science*, Glynn, Ichino (2019). Generalized Nonlinear DiDiD. *preprint*

8 TT (2014). The control outcome calibration approach for causal inference with unobserved confounding. *Am J Epi*, Park, Richardson, TT (2024). Single proxy control. *Biometrics*

9 Gagnon-Bartsch, Speed (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*
Wang et al. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics*
Jacob et al. (2016) Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*

10 Miao, Geng, TT (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*
Cui, Pu, Shi, Miao, TT (2024). Semiparametric proximal causal inference. *JASA*
Miao, Shi, Li, TT (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*

Proximal Causal Inference

Can Imperfect Measurements Still Be Useful?

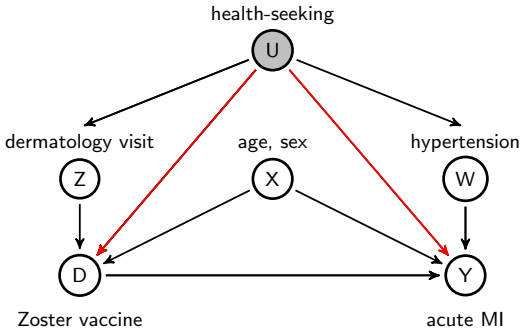
- Survey Studies [3][1]
- Obtaining a variable without measurement error often requires significant time and resources
  - This can be an extremely challenging task—perhaps impossible
- Noisy measures can still be highly informative about the quantity of interest, provided the measurement errors have the right structure

- NC/Proximal Causal Inference
- The no unmeasured confounding assumption depends on investigator's ability to accurately measure covariates capturing all potential sources of confounding
  - This assumption can be unrealistic
  - In practice, the most one can hope for is that measured covariates are at best proxies of the true underlying confounding mechanism
- Acknowledge that measured covariates are imperfect proxies of confounders
  - If these proxies satisfying certain assumptions, it may still be possible to infer causal quantities of interest

[1] Browning, Crossley (2009). Are two cheap, noisy measures better than one expensive, accurate one? *American Economic Review*

An Example in Vaccine Safety Study [17][1]

- Adverse effect of a new Zoster vaccine on acute Myocardial Infarction (MI; heart attack)
- Plan to adjust for the following baseline variables
  - age, sex (X)
  - dermatology visit (Z; check skin)
  - hypertension (W; high blood pressure)
- Naively adjusting (X, W, Z) does not eliminate confounding bias due to U
- Need to use X, W, Z differently!



[1] Li et al. (2024). Using Double Negative Controls to Adjust for Healthy User Bias in a Recombinant Zoster Vaccine Safety Study. *Am J Epi*

An Example in Vaccine Safety Study [17][1]

- Three types of measured covariates

- Confounder

  Common causes of the treatment and outcome

  age, sex (X)

- Treatment confounding proxy

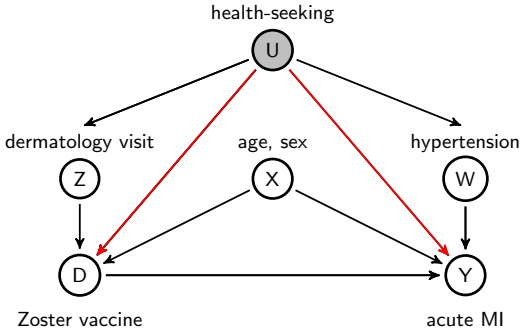
  $Y \perp\!\!\!\perp Z \,|\, (D, U, X)$ (NCE)

  dermatology visit (Z)

- Outcome confounding proxy

  $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$ (NCO)

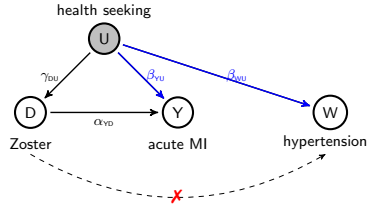
  hypertension (W)



[1] Li et al. (2024). Using Double Negative Controls to Adjust for Healthy User Bias in a Recombinant Zoster Vaccine Safety Study. *Am J Epi*

Intuitions for Identification

- $[V_1 \sim V_2]$ $= \texttt{coef}(\texttt{lm}(V_1 \sim V_2))$
  $=$ direct effect of V_2 on V_1
  $+$ confounding bias between V_2 and V_1



- $[Y \sim D] = \alpha_{YD} + [D \sim U] \times [Y \sim U] = \alpha_{YD} + \gamma_{DU}\beta_{YU}$
- $[W \sim D] = \alpha_{WD} + [D \sim U] \times [W \sim U] = \gamma_{DU}\beta_{WU}$

- If $[Y \sim U] = [W \sim U] \Leftrightarrow \beta_{YU} = \beta_{UW}$

$[Y \sim D]$	$=$	$\alpha_{YD} + \gamma_{DU}\beta_{YU}$	
$[W \sim D]$	$=$	$\gamma_{DU}\beta_{WU}$	
diff in coefs of D	$=$	α_{YD}	

- A Special Case: Difference-in-Differences (DiD)
  $[Y_0 \sim U] = [Y_1 \sim U]$ [27, 34, 36][1]

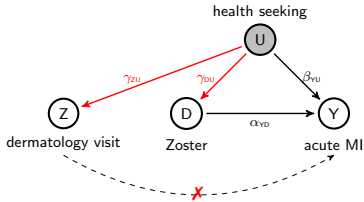
[1] Richardson et al. (2014). Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *Am J Epi*
TT, Sofer, Richardson (2015). NCO for unobserved confounding under a Cox proportional hazards model. *preprint*
Sofer et al. (2016). On negative outcome control of unobserved confounding as a generalization of DiD. *Statistical Science*

Intuitions for Identification

- $[V_1 \sim V_2] = \texttt{coef}(\texttt{lm}(V_1 \sim V_2))$

  $=$ direct effect of V_2 on V_1

  $+$ confounding bias between V_2 and V_1



- $[Y \sim D] = \alpha_{YD} + [D \sim U] \times [Y \sim U] = \alpha_{YD} + \gamma_{DU}\beta_{YU}$
- $[Y \sim Z] = \alpha_{YZ} + [Z \sim U] \times [Y \sim U] = \gamma_{ZU}\beta_{YU}$

- If $[D \sim U] = [Z \sim U] \Leftrightarrow \gamma_{DU} = \gamma_{ZU}$

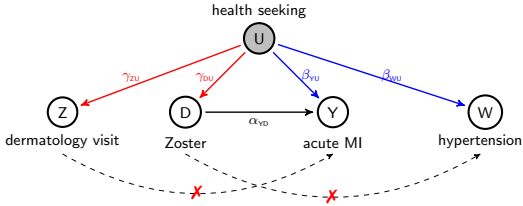
$[Y \sim D]$	$=$	α_{YD} $+$	$\gamma_{DU}\beta_{YU}$
$[Y \sim Z]$	$=$		$\gamma_{ZU}\beta_{YU}$
diff in coefs	$=$	α_{YD}	

- A Special Case: air pollution studies [10, 11, 22] [1]

[1] Flanders et al. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*
Flanders, Strickland, Klein (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *Am J Epi*
Miao, TT (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *Am J Epi*

Identification Strategy

- What if $[Y \sim U] \neq [W \sim U]$ and $[D \sim U] \neq [Z \sim U]$?
- Solution: Use both proxies [20, 21, 31, 32] [1]

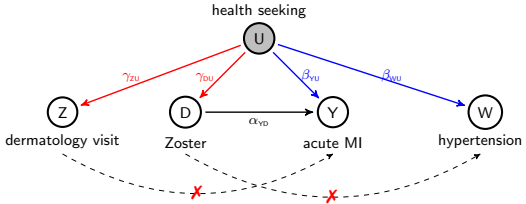


[1] Miao, Shi, Li, TT (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*
Miao, Geng, TT (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*
Shi, Miao, TT (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *JRSS:B*
Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Identification Strategy

- What if $[Y \sim U] \neq [W \sim U]$ and $[D \sim U] \neq [Z \sim U]$?
- Solution: Use both proxies [20, 21, 31, 32] [1]
- Note that

$$
\begin{array}{rrcl}
 & [Y \sim D] & = & \alpha_{YD} + \gamma_{DU}\beta_{YU} \\
 & [W \sim D] & = & \gamma_{DU}\beta_{WU} \\
\hline
\Rightarrow & [Y \sim D] & = & \alpha_{YD} + \underbrace{\frac{\beta_{YU}}{\beta_{WU}}[W \sim D]}_{=\text{bias}}
\end{array}
$$

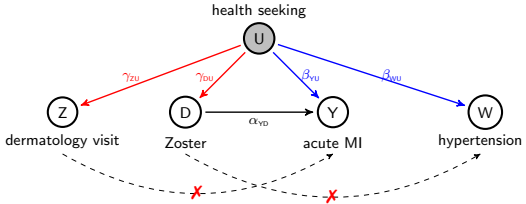


[1] Miao, Shi, Li, TT (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*
Miao, Geng, TT (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*
Shi, Miao, TT (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *JRSS:B*
Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Identification Strategy

- What if $[Y \sim U] \neq [W \sim U]$ and $[D \sim U] \neq [Z \sim U]$?
- Solution: Use both proxies [20, 21, 31, 32] [1]
- Note that

$$
\begin{aligned}
[Y \sim D] &= \alpha_{YD} + \gamma_{DU}\beta_{YU} \\
[W \sim D] &= \gamma_{DU}\beta_{WU} \\
\hline
\Rightarrow [Y \sim D] &= \alpha_{YD} + \underbrace{\frac{\beta_{YU}}{\beta_{WU}}[W \sim D]}_{=\text{bias}}
\end{aligned}
$$



$$
\begin{aligned}
[Y \sim Z] &= \gamma_{ZU}\beta_{YU} \\
[W \sim Z] &= \gamma_{ZU}\beta_{WU}
\end{aligned}
\quad \Rightarrow \quad
\frac{[Y \sim Z]}{[W \sim Z]} = \underbrace{\frac{\beta_{YU}}{\beta_{WU}}}_{=\text{scale}}
$$

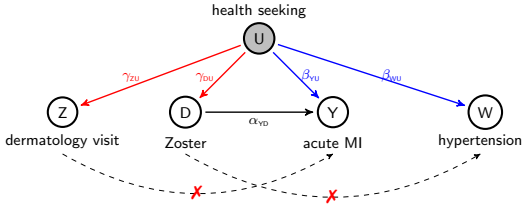
- W recovers bias up to a scale
- Z recovers that scale

[1] Miao, Shi, Li, TT (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*
Miao, Geng, TT (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*
Shi, Miao, TT (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *JRSS:B*
Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Identification Strategy

- What if $[Y \sim U] \neq [W \sim U]$ and $[D \sim U] \neq [Z \sim U]$?
- Solution: Use both proxies [20, 21, 31, 32] [1]
- Note that

$$\begin{aligned} [Y \sim D] &= \alpha_{YD} + \gamma_{DU}\beta_{YU} \\ [W \sim D] &= \gamma_{DU}\beta_{WU} \\ \hline \Rightarrow \quad [Y \sim D] &= \alpha_{YD} + \underbrace{\frac{\beta_{YU}}{\beta_{WU}}[W \sim D]}_{=\text{bias}} \end{aligned}$$



$$\begin{aligned} [Y \sim Z] &= \gamma_{ZU}\beta_{YU} \\ [W \sim Z] &= \gamma_{ZU}\beta_{WU} \end{aligned} \quad \Rightarrow \quad \frac{[Y \sim Z]}{[W \sim Z]} = \underbrace{\frac{\beta_{YU}}{\beta_{WU}}}_{=\text{scale}}$$

$$\Rightarrow \quad \alpha_{YD} = [Y \sim D] - \frac{[Y \sim Z]}{[W \sim Z]}[W \sim D]$$

- W recovers bias up to a scale
- Z recovers that scale

[1] Miao, Shi, Li, TT (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*
Miao, Geng, TT (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*
Shi, Miao, TT (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *JRSS:B*
Shi, Miao, TT (2020). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*

Linear Additive Model

- Consider

$$\begin{aligned} \mathrm{E}(Y \,|\, D, Z, U) &= \beta_0 + \beta_D D + \beta_U U \\ \mathrm{E}(W \,|\, D, Z, U) &= \gamma_0 + \gamma_U U \end{aligned}$$

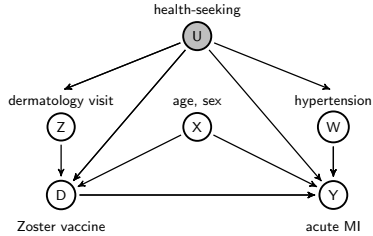
$$\beta_D = \mathsf{ATE} = \mathrm{E}[Y^{(1)} - Y^{(0)}] = \mathrm{E}[Y^{(1)} - Y^{(0)} \,|\, U]$$

- One can establish that

$$\mathrm{E}(Y \,|\, D, Z) = \widetilde{\beta}_0 + \beta_D D + \widetilde{\beta}_U \mathrm{E}(W \,|\, D, Z)$$

- Two-stage Regression Procedure [37][1]

1. $\widehat{W} \leftarrow \texttt{lm}(W \sim D + Z)$
2. $\beta_D \leftarrow \texttt{coef}(\texttt{lm}(Y \sim D + \widehat{W}))$



[1] TT et al. (2024). An Introduction to Proximal Causal Inference. *Statistical Science*

Generalized Linear Additive Models

- Liu et al. [18][1]: Generalization of the two-stage regression procedure to count/binary/categorical Y and W
- The coefficient of D in the 2nd stage is consistent for the causal effect (on the scale defined by the corresponding Y link function)

		Y		
		Continuous (Identity Link)	Count (Log Link)	Binary/Categorical (Logit Link)
W	Continuous (Identity Link)	Linear $W \sim D + Z$ $S = \mathrm{E}(W\|D, Z)$ Linear $Y \sim D + S$	Linear $W \sim D + Z$ $S = \mathrm{E}(W\|D, Z)$ Poisson $Y \sim D + S$	Linear $W \sim D + Z + Y$ $S = \mathrm{E}(W\|D, Z, Y = 0)$ Logistic $\mathbb{1}(Y = t) \sim D + S$
	Count (Log Link)	Poisson $W \sim D + Z$ $S = \log(\mathrm{E}(W\|D, Z))$ Linear $Y \sim D + S$	Poisson $W \sim D + Z$ $S = \log(\mathrm{E}(W\|D, Z))$ Poisson $Y \sim D + S$	Poisson $W \sim D + Z + Y$ $S = \log(\mathrm{E}(W\|D, Z, Y = 0))$ Logistic $\mathbb{1}(Y = t) \sim D + S$
	Binary/Categorical (Logit Link)	Logistic $\mathbb{1}(W = k) \sim D + Z$ $S_k = \mathrm{logit}(P(W = k\|D, Z))$ Linear $Y \sim D + S + W$	Logistic $\mathbb{1}(W = k) \sim D + Z$ $S_k = \mathrm{logit}(P(W = k\|D, Z))$ Poisson $Y \sim D + S + W$	Logistic $\mathbb{1}(W = k) \sim D + Z + Y$ $S_k = \mathrm{logit}(P(W = k\|D, Z, Y = 0))$ Logistic $\mathbb{1}(Y = t) \sim D + S + W$

[1] Liu, Park, Li, TT (2024). Regression-Based Proximal Causal Inference. *Am J Epi*

Review: Nonparametric Identification Under No Unmeasured Confounding

- Assuming no unmeasured confounding $Y^{(d)} \perp\!\!\!\perp D \,|\, L$ where L = pre-treatment covariates

$$\begin{aligned} \mathrm{E}[Y^{(d)}] &= \mathrm{E}\big[p(d,L)\mathbb{1}(D=d)Y\big] && \text{IPW} \\ &= \mathrm{E}\big[\mu(d,L)\big] && \text{g-formula} \\ &= \mathrm{E}\big[p(d,L)\mathbb{1}(D=d)\{Y-\mu(d,L)\}+\mu(d,X)\big] && \text{AIPW} \end{aligned}$$

where

$$p(d,L) = \frac{1}{\Pr(D=d \,|\, L)} \quad = \text{inverse propensity score}$$

$$\mu(d,L) = \mathrm{E}(Y \,|\, D=d, L) \quad = \text{outcome regression}$$

- These identifying formulas are nonparametric

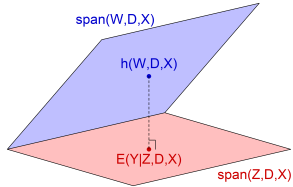
- **Question**: Can we establish similar identifying formulas in the presence of U using proximal causal inference?

Bridge Function

- Bridge function:

  Find h such that $\mathrm{E}[h(W,D,X)|Z,D,X] = \mathrm{E}(Y|Z,D,X)$

  Find q such that $\mathrm{E}[q(Z,D,X)|W,D,X] = \dfrac{1}{\Pr(D\,|\,W,X)}$

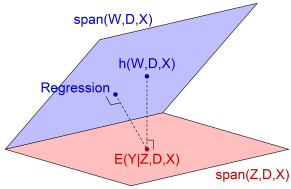


Bridge Function

- Bridge function:

  Find h such that $\mathrm{E}[h(W,D,X)|Z,D,X] = \mathrm{E}(Y|Z,D,X)$

  Find q such that $\mathrm{E}[q(Z,D,X)|W,D,X] = \dfrac{1}{\Pr(D\,|\,W,X)}$



- h and q cannot be obtained from standard regression!

  $h \neq$ projection of $E(Y|Z,D,X)$ onto (W,D,X)

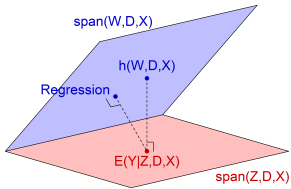
  $h =$ a function whose projection onto (Z,D,X) is $E(Y\,|\,Z,D,X)$

Bridge Function

- Bridge function:

  Find h such that $\mathrm{E}[h(W,D,X)|Z,D,X] = \mathrm{E}(Y|Z,D,X)$

  Find q such that $\mathrm{E}[q(Z,D,X)|W,D,X] = \dfrac{1}{\Pr(D\,|\,W,X)}$



- h and q cannot be obtained from standard regression!

  $h \neq$ projection of $E(Y|Z,D,X)$ onto (W,D,X)

  $h =$ a function whose projection onto (Z,D,X) is $E(Y\,|\,Z,D,X)$

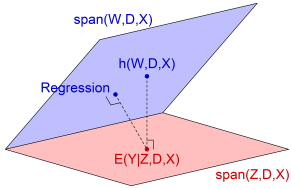
- Ill-posed inverse problems, Fredholm integral equations of the first kind

Bridge Function

- Bridge function:

  Find h such that $\mathrm{E}[h(W,D,X)|Z,D,X] = \mathrm{E}(Y|Z,D,X)$

  Find q such that $\mathrm{E}[q(Z,D,X)|W,D,X] = \dfrac{1}{\Pr(D\,|\,W,X)}$



- h and q cannot be obtained from standard regression!

  $h \neq$ projection of $E(Y|Z,D,X)$ onto (W,D,X)

  $h =$ a function whose projection onto (Z,D,X) is $E(Y\,|\,Z,D,X)$

- Ill-posed inverse problems, Fredholm integral equations of the first kind

- $h =$ outcome confounding bridge function, $\quad q =$ treatment confounding bridge function

Nonparametric Identification [5, 37][1]

- If h and q exist,

$$\begin{aligned}
\mathrm{E}[Y^{(d)}] &= \mathrm{E}[q(Z,d,X)\mathbb{1}(D=d)Y] && \text{Proximal IPW} \\
&= \mathrm{E}[h(W,d,X)] && \text{Proximal g-formula} \\
&= \mathrm{E}[q(Z,d,X)\mathbb{1}(D=d)\{Y-h(W,d,X)\}+h(W,d,X)] && \text{Proximal AIPW}
\end{aligned}$$

- Under no unmeasured confounding:

$$\begin{aligned}
\mathrm{E}[Y^{(d)}] &= \mathrm{E}\big[p(d,L)\mathbb{1}(D=d)Y\big] && \text{IPW} \\
&= \mathrm{E}\big[\mu(d,L)\big] && \text{g-formula} \\
&= \mathrm{E}\big[p(d,L)\mathbb{1}(D=d)\{Y-\mu(d,L)\}+\mu(d,X)\big] && \text{AIPW}
\end{aligned}$$

[1] Cui, Pu, Shi, Miao, TT (2024). Semiparametric proximal causal inference. *JASA*, TT et al. (2024). An Introduction to Proximal Causal Inference. *Statistical Science*

Example: Binary Z, W

- Suppressing X, $h(w,d)$ solves

$$\mathrm{E}[h(W,D)|Z,D] = \mathrm{E}(Y|Z,D)$$

$$\Leftrightarrow \quad \underbrace{\begin{bmatrix} \Pr(W=1 \,|\, Z=1,D) & \Pr(W=0 \,|\, Z=1,D) \\ \Pr(W=1 \,|\, Z=0,D) & \Pr(W=0 \,|\, Z=0,D) \end{bmatrix}}_{=B(D)} \begin{bmatrix} h(W=1,D) \\ h(W=0,D) \end{bmatrix} = \underbrace{\begin{bmatrix} \mathrm{E}(Y \,|\, Z=1,D) \\ \mathrm{E}(Y \,|\, Z=0,D) \end{bmatrix}}_{=C(D)}$$

$$\overset{\text{if the inverse exists}}{\Rightarrow} \quad \begin{bmatrix} h(W=1,D) \\ h(W=0,D) \end{bmatrix} = B^{-1}(D)C(D)$$

- $\mathrm{E}[Y^{(d)}]$ is identified by

$$\mathrm{E}[Y^{(d)}] = \mathrm{E}[h(W,d)] = \Pr(W=1)h(W=1,d) + \Pr(W=0)h(W=0,d)$$

Review: Estimation Under No Unmeasured Confounding

- Posit parametric models for p and μ[1]

  $p(D, X; \theta_p) = 1/\Pr(D \,|\, X; \theta_p); \quad \mu(D, X; \theta_\mu) = \mathrm{E}(Y \,|\, D, X; \theta_\mu)$

- Estimate θ_p and θ_μ from regression models

  $\texttt{glm}(D \sim X, \texttt{family="binomial"}); \quad \texttt{lm}(Y \sim D + X)$

- Obtain estimators using the ID formula

$$\text{AIPW estimator} = \texttt{aver}\Big[\widehat{p}(D, X)\mathbb{1}(D = d)\{Y - \widehat{\mu}(d, X)\} + \widehat{\mu}(d, X)\Big]$$

- Asymptotically normal for $\mathrm{E}[Y^{(d)}]$ under certain conditions

[1] One can use nonparametric/machine learning approaches with cross-fitting

Estimation for Proximal Causal Inference

- Outcome/Treatment confounding bridge function:

$$\mathrm{E}[h(W,D,X)|Z,D,X] = \mathrm{E}(Y|Z,D,X) \quad \Rightarrow \quad \mathrm{E}[g_h(Z,D,X)\{Y - h(W,D,X)\}] = 0 \text{ for any } g_h$$

- Posit parametric models for $h = h(w,d,x;\theta_h)$ and $q = q(z,d,x;\theta_q)$

- Estimate h and q from generalized method of moments (GMM)[1]

$$\widehat{\theta}_h \leftarrow \texttt{aver}\Big[g_h(Z,D,X)\{Y - h(W,D,X;\theta_h)\}\Big] = 0,$$

```
ivreg::ivreg(Y~W+D+X | Z+D+X)
gmm::gmm(g=Y~W+D+X, x=~Z+D+X)
```

- Obtain estimators using the ID formula

$$\text{Proximal AIPW estimator} = \texttt{aver}\Big[\widehat{q}(Z,d,X)\mathbb{1}(D=d)\big\{Y - \widehat{h}(W,d,X)\big\} + \widehat{h}(W,d,X)\Big]$$

- Asymptotically normal for $\mathrm{E}[Y^{(d)}]$ under certain conditions

- One can use nonparametric approaches [14, 19, 33][2]

[1] Can use existing instrumental variable software packages; R: gmm, ivreg, sem, ivpack, AER; SAS: SYSLIN; Stata: ivregress, ivreg, ivreg2

[2] Singh, Sahani, Gretton (2019). Kernel IV regression. *NeurIPS*, Mastouri et al. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. *ICML* Ghassami, Ying, Shpitser, TT (2022). Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. *AISTAT*

Comparison to Standard Causal Inference

- Standard Causal Inference
- Usage of covariates L

  Treat equally as confounders

  $L = \{\text{age, sex, hypertension, dermatology visit}\}$

- Identifying formula for $\mathrm{E}[Y^{(d)}]$

  IPW $\quad \mathrm{E}\big[p(d,L)\mathbb{1}(D=d)Y\big]$

  g-formula $\quad \mathrm{E}\big[\mu(d,L)\big]$

  AIPW $\quad \mathrm{E}\Big[\begin{array}{l} p(d,L)\mathbb{1}(D=d)\{Y-\mu(d,L)\} \\ +\mu(d,X)\end{array}\Big]$

- Estimators for $\mathrm{E}[Y^{(d)}]$

  OLS $\quad \texttt{lm}(Y \sim D + L)$

  AIPW $\quad \texttt{aver}\Big[\begin{array}{l} \hat{p}(d,L)\mathbb{1}(D=d)\{Y-\hat{\mu}(d,L)\} \\ +\hat{\mu}(a,X)\end{array}\Big]$

- Proximal Causal Inference

  Divide into three buckets

  $X = \{\text{age, sex}\}$, $W = \{\text{hypertension}\}$, $Z = \{\text{dermatology visit}\}$

  Proximal IPW $\quad \mathrm{E}[q(Z,d,X)\mathbb{1}(D=d)Y]$

  Proximal g-formula $\quad \mathrm{E}[h(W,d,X)]$

  Proximal AIPW $\quad \mathrm{E}\Big[\begin{array}{l} q(Z,d,X)\mathbb{1}(D=d)\{Y-h(W,d,X)\} \\ +h(W,d,X)\end{array}\Big]$

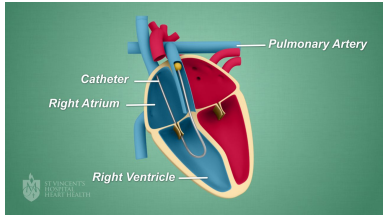
  2 Stage $\quad \texttt{lm}(Y \sim D + X + \widehat{W})$, $\widehat{W} \leftarrow \texttt{lm}(W \sim D + X + Z)$

  Proximal AIPW $\quad \texttt{aver}\Big[\begin{array}{l} \hat{q}(Z,d,X)\mathbb{1}(D=d)\{Y-\hat{h}(W,d,X)\} \\ +\hat{h}(W,d,X)\end{array}\Big]$

Application to the SUPPORT study

- Right heart catheterization (RHC) procedure

  A catheter (a thin, flexible tube) is inserted into a vein and guided through the blood vessels to reach the right side of the heart and pulmonary artery
  Performed to measure blood flow and heart pressure
  Common belief: measurements from the RHC can help doctors assess heart function and diagnose conditions



source: St Vincent's Hospital

- The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) [4][1]

  Evaluate the effectiveness of RHC among patients admitted to the intensive care unit (ICU)
  2184 patients with RHC, 3551 without RHC

- The SUPPORT study found that RHC is harmful (associated with a lower chance of survival)

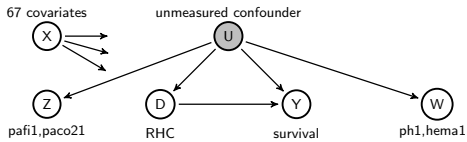
[1] Connors et al. (1996). The effectiveness of right heart catheterization in the initial care of critically Ill patients. *Jama*

Application to the SUPPORT study

- Y = the number of days between admission and death or censoring at 30 days

- 71 covariates

  demographics, comorbidity, vital signs, functional status
  Physiological status measured from a blood test during the first day in the ICU (proxy candidates)



- Determine Z and W

  (pafi1, paco21, ph1, hema1) are strongly correlated with D and Y
  Variables most strongly associated with D and Y were selected as Z and W, respectively
  Z = (pafi1,paco21); W = (ph1,hema1)

Application to the SUPPORT study

		Standard Causal Inference	Proximal Causal Inference
Estimator	OLS / 2 Stage [37][2]	-1.25 (-1.80,-0.70)	-1.80 (-2.64,-0.96)
	AIPW [5][2]	-1.17 (-1.79,-0.55)	-1.66 (-2.50,-0.83)

- Estimation methods show little difference
- Standard and proximal causal inference yield substantially different results
- RHC consistently exhibits a negative impact on survival

[2] TT et al. (2024). An Introduction to Proximal Causal Inference. *Statistical Science*

[2] Cui, Pu, Shi, Miao, TT (2024). Semiparametric proximal causal inference. *JASA*

Extensions of Proximal Causal Inference to Various Settings

Extensions

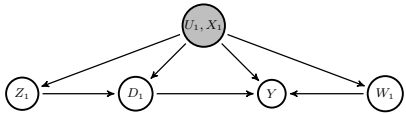
- Longitudinal Settings
- Mediation analysis with hidden confounders
- Mediation analysis and front-door model with hidden mediators
- Interference with homophily driven by hidden factors
- Proximal synthetic controls

Longitudinal Setting

- Discussed the point exposure setting

  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$



Longitudinal Setting

- Discussed the point exposure setting

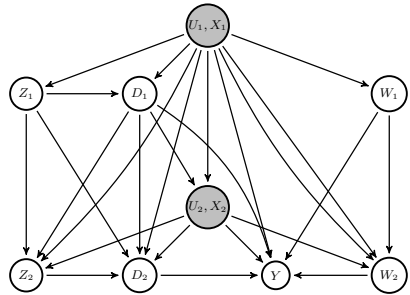
  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$

- Extension to longitudinal settings

  $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(\vec{d})}) \,|\, (U_1, X_1)$

  $(\vec{Z}_2, \vec{D}_2) \perp\!\!\!\perp (\vec{W}_2, Y^{(\vec{d})}) \,|\, (D_1 = d_1, \vec{U}_2, \vec{X}_2)$



Longitudinal Setting

- Discussed the point exposure setting

  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

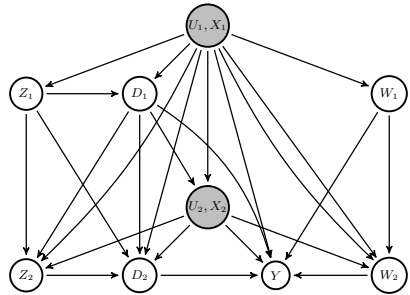
- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$

- Extension to longitudinal settings

  $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(\vec{d})}) \,|\, (U_1, X_1)$

  $(\vec{Z}_2, \vec{D}_2) \perp\!\!\!\perp (\vec{W}_2, Y^{(\vec{d})}) \,|\, (D_1 = d_1, \vec{U}_2, \vec{X}_2)$

- Estimand: $\mathrm{E}[Y^{(\vec{d})}]$



Longitudinal Setting

- Discussed the point exposure setting

  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$

- Extension to longitudinal settings

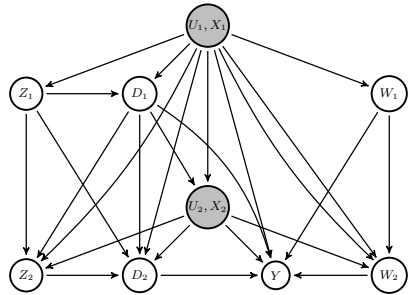
  $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(\vec{d})}) \,|\, (U_1, X_1)$

  $(\vec{Z}_2, \vec{D}_2) \perp\!\!\!\perp (\vec{W}_2, Y^{(\vec{d})}) \,|\, (D_1 = d_1, \vec{U}_2, \vec{X}_2)$

- Estimand: $\mathrm{E}[Y^{(\vec{d})}]$

- Nested bridge functions:

  $\mathrm{E}[h_2(\vec{D}_2, \vec{W}_2, \vec{X}_2) \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2] = \mathrm{E}(Y \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2)$



Longitudinal Setting

- Discussed the point exposure setting

  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$

- Extension to longitudinal settings

  $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(\vec{d})}) \,|\, (U_1, X_1)$

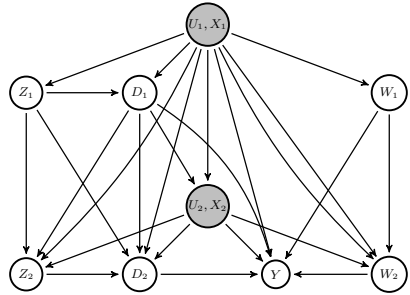
  $(\vec{Z}_2, \vec{D}_2) \perp\!\!\!\perp (\vec{W}_2, Y^{(\vec{d})}) \,|\, (D_1 = d_1, \vec{U}_2, \vec{X}_2)$

- Estimand: $\mathrm{E}[Y^{(\vec{d})}]$

- Nested bridge functions:

  $\mathrm{E}[h_2(\vec{D}_2, \vec{W}_2, \vec{X}_2) \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2] = \mathrm{E}(Y \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2)$

  $\mathrm{E}[h_1(\vec{D}_2, W_1, X_1) \,|\, \vec{D}_2, Z_1, X_1] = \mathrm{E}[h_2(\vec{D}_2, \vec{W}_2, \vec{X}_2) \,|\, \vec{D}_2, Z_1, X_1]$



[1] Ying, Miao, Shi, TT (2023). Proximal causal inference for complex longitudinal studies. *JRSS:B*

Longitudinal Setting

- Discussed the point exposure setting

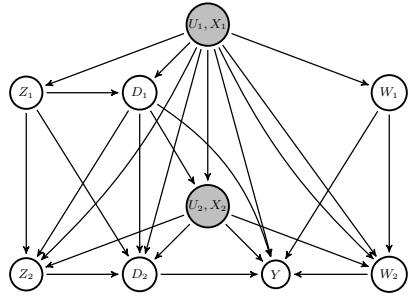
  $Z_1 \perp\!\!\!\perp Y \,|\, (U_1, D_1, X_1)$ and $W_1 \perp\!\!\!\perp (D_1, Z_1) \,|\, (U_1, X_1)$

- Suffices to write as $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(d)}) \,|\, (U_1, X_1)$

- Extension to longitudinal settings

  $(Z_1, D_1) \perp\!\!\!\perp (W_1, Y^{(\vec{d})}) \,|\, (U_1, X_1)$

  $(\vec{Z}_2, \vec{D}_2) \perp\!\!\!\perp (\vec{W}_2, Y^{(\vec{d})}) \,|\, (D_1 = d_1, \vec{U}_2, \vec{X}_2)$



- Estimand: $\mathrm{E}[Y^{(\vec{d})}]$

- Nested bridge functions:

$$\mathrm{E}[h_2(\vec{D}_2, \vec{W}_2, \vec{X}_2) \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2] = \mathrm{E}(Y \,|\, \vec{D}_2, \vec{Z}_2, \vec{X}_2)$$

$$\mathrm{E}[h_1(\vec{D}_2, W_1, X_1) \,|\, \vec{D}_2, Z_1, X_1] = \mathrm{E}[h_2(\vec{D}_2, \vec{W}_2, \vec{X}_2) \,|\, \vec{D}_2, Z_1, X_1]$$

- **Identification**: $\mathrm{E}[Y^{(\vec{d})}] = \mathrm{E}\big[h_1(\vec{d}, W_1, X_1)\big]$; Proximal AIPW identification is also available [40][1]

[1] Ying, Miao, Shi, TT (2023). Proximal causal inference for complex longitudinal studies. *JRSS:B*

Longitudinal Setting

- Suppressing X

- Nested bridge functions:

$$\mathrm{E}[h_2(\vec{D}_2, \vec{W}_2) \,|\, \vec{D}_2, \vec{Z}_2] = \mathrm{E}(Y \,|\, \vec{D}_2, \vec{Z}_2) \quad \Rightarrow \quad \mathrm{E}\big[g_2(\vec{D}_2, \vec{Z}_2)\{Y - h_2(\vec{D}_2, \vec{W}_2)\}\big] = 0$$

$$\mathrm{E}[h_1(\vec{D}_2, W_1) \,|\, \vec{D}_2, Z_1] = \mathrm{E}[h_2(\vec{D}_2, \vec{W}_2) \,|\, \vec{D}_2, Z_1] \quad \Rightarrow \quad \mathrm{E}\big[g_1(\vec{D}_2, Z_1)\{h_2(\vec{D}_2, \vec{W}_2) - h_1(\vec{D}_2, W_1)\}\big] = 0$$

- $\mathrm{E}[Y^{(\vec{d})}] = \mathrm{E}\big[h_1(\vec{d}, W_1)\big]$

- Moment equation:

$$(\widehat{\theta}_1, \widehat{\theta}_2) \quad \leftarrow \quad \texttt{aver} \left[\begin{array}{l} g_2(\vec{D}_2, \vec{Z}_2)\{Y - h_2(\vec{D}_2, \vec{W}_2; \theta_2)\} \\ g_1(\vec{D}_2, Z_1)\{h_2(\vec{D}_2, \vec{W}_2; \theta_2) - h_1(\vec{D}_2, W_1; \theta_1)\} \end{array} \right] = 0$$

- $\widehat{\mathrm{E}}[Y^{(\vec{d})}] = \texttt{aver}\big[h_1(\vec{d}, W_1; \widehat{\theta}_1)\big]$

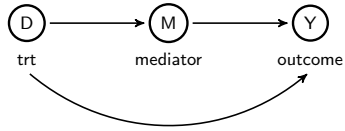
- Asymptotically normal under certain conditions

Extensions

- Longitudinal Settings
- Mediation analysis with hidden confounders
- Mediation analysis and front-door model with hidden mediators
- Interference with homophily driven by hidden factors
- Proximal synthetic controls

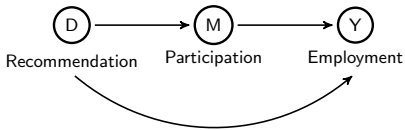
Mediation

- Mediator (M) explains part of the relationship between D and Y



Mediation

- Mediator (M) explains part of the relationship between D and Y



- Example: Job Corps program

  Education and career training program administered by the U.S. Department of Labor

  D: Randomized, recommendation to participate in Job Corps

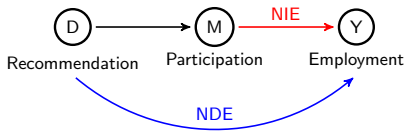
  M: Actual participation

  Y: Employment

Natural Direct/Indirect Effects

- The average treatment effect of D on Y

$$
\begin{aligned}
&\mathrm{E}[Y^{(1)} - Y^{(0)}] \\
&= \mathrm{E}[Y^{(1,M^{(1)})} - Y^{(1,M^{(0)})}] \\
&= \underbrace{\mathrm{E}[Y^{(1,M^{(1)})} - Y^{(1,M^{(0)})}]}_{\text{Natural Indirect Eff.}} + \underbrace{\mathrm{E}[Y^{(1,M^{(0)})} - Y^{(0,M^{(0)})}]}_{\text{Natural Direct Eff.}}
\end{aligned}
$$



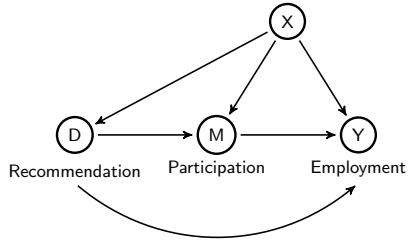
- Interpretation:
  - NDE: Effect of the JC recommendation (D) on employment (Y) that is not mediated by the participation (M)
  - NIE: Effect of the JC recommendation (D) on employment (Y) that is mediated by the participation (M)

- $\mathrm{E}[Y^{(1,M^{(0)})}]$ = **mediation functional**

Review: Identification of the Mediation Functional

- Under no unmeasured confounding (and certain assumptions), the mediation functional is identified by the **mediation formula** [25][1]:

$$\mathrm{E}[Y^{(1,M^{(0)})}] = \mathrm{E}_X\left[\sum_m \mathrm{E}(Y \mid D=1, M=m, X)\,\Pr(M=m \mid D=0, X)\right]$$

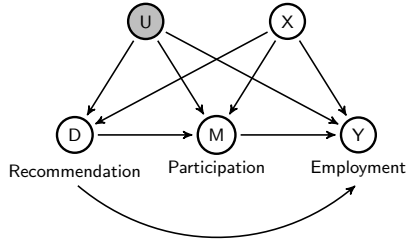


[1] Pearl (2001). Direct and indirect effects. *UAI*

Review: Identification of the Mediation Functional

- Under no unmeasured confounding (and certain assumptions), the mediation functional is identified by the **mediation formula** [25][1]:

$$\mathrm{E}[Y^{(1,M^{(0)})}] = \mathrm{E}_X\left[\sum_m \mathrm{E}(Y \mid D=1, M=m, X)\,\Pr(M=m \mid D=0, X)\right]$$



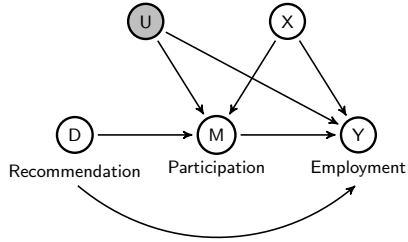
- In the presence of U, the mediation formula is no longer valid

[1] Pearl (2001). Direct and indirect effects. *UAI*

Review: Identification of the Mediation Functional

- Under no unmeasured confounding (and certain assumptions), the mediation functional is identified by the **mediation formula** [25][1]:

$$\mathrm{E}[Y^{(1,M^{(0)})}] = \mathrm{E}_X\left[\sum_m \mathrm{E}(Y \mid D=1, M=m, X)\,\Pr(M=m \mid D=0, X)\right]$$



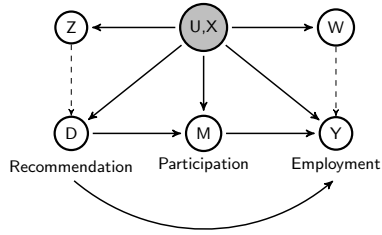
- In the presence of U, the mediation formula is no longer valid
- Even in randomized trials, common causes of M and Y may exist

[1] Pearl (2001). Direct and indirect effects. *UAI*

Proximal Identification of the Mediation Functional

- Suppose that there exist Z and W satisfying

$$Z \perp\!\!\!\perp (Y, M) \,|\, (U, X, D) \qquad\qquad W \perp\!\!\!\perp (D, M, Z) \,|\, (U, X)$$



- Define two outcome confounding bridge functions

$$\mathrm{E}\big[h_1(W, M, X) \,|\, Z, D = 1, M, X\big] = \mathrm{E}(Y \,|\, Z, D = 1, M, X)$$
$$\mathrm{E}\big[h_0(W, X) \,|\, Z, D = 0, X\big] = \mathrm{E}[h_1(W, M, X) \,|\, Z, D = 0, X\big]$$

- Proximal mediation formula [8][1]

$$\mathrm{E}\big[Y^{(1, M^{(0)})}\big] = \mathrm{E}\big[h_0(W, X)\big]$$

- An asymptotically normal estimator can be constructed

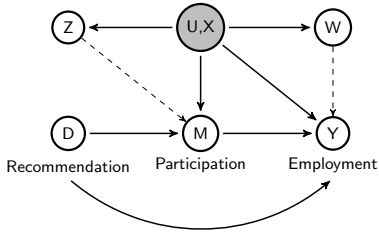
[1] Dukes, Shpitser, TT (2023). Proximal mediation analysis. *Biometrika*

Proximal Identification of the Mediation Functional

- Suppose that there exist Z and W satisfying

$$Z \perp\!\!\!\perp Y \quad \big| (U, X, D) \qquad\qquad W \perp\!\!\!\perp (D, M, Z) \,\big|\, (U, X)$$

D is randomized



- Define two outcome confounding bridge functions

$$\mathrm{E}\big[h_1(W, M, X) \,\big|\, Z, D = 1, M, X\big] = \mathrm{E}(Y \,\big|\, Z, D = 1, M, X)$$

$$\mathrm{E}\big[h_0(W, X) \,\big|\, Z, D = 0, X\big] = \mathrm{E}[h_1(W, M, X) \,\big|\, Z, D = 0, X\big]$$

- Proximal mediation formula [8][1]

$$\mathrm{E}\big[Y^{(1, M^{(0)})}\big] = \mathrm{E}\big[h_0(W, X)\big]$$

- An asymptotically normal estimator can be constructed

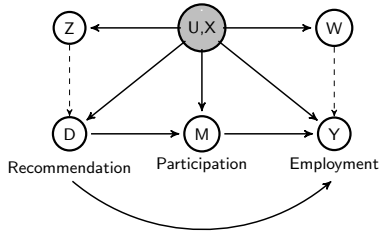
[1] Dukes, Shpitser, TT (2023). Proximal mediation analysis. *Biometrika*

Extensions

- Longitudinal Settings
- Mediation analysis with hidden confounders
- Mediation analysis and front-door model with hidden mediators
- Interference with homophily driven by hidden factors
- Proximal synthetic controls

Hidden Mediators

- So far, we allow for U as long as certain proxies are available

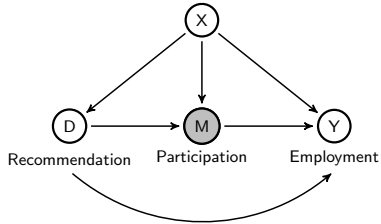


Hidden Mediators

- So far, we allow for U as long as certain proxies are available

- What if there is no U, but M is not observed?

  Respondents may inaccurately report their true participation M due to pressure to comply, confusion with other programs, concerns about not receiving an incentive, etc



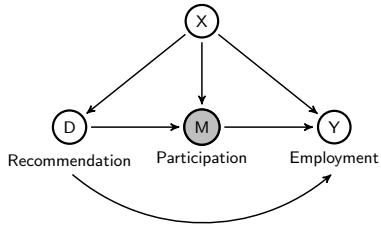
Hidden Mediators

- So far, we allow for U as long as certain proxies are available

- What if there is no U, but M is not observed?

  Respondents may inaccurately report their true participation M due to pressure to comply, confusion with other programs, concerns about not receiving an incentive, etc

- Idea: if we have proxies of M, we can identify $\mathrm{E}[Y^{(1,M^{(0)})}]$



Identification of the Mediation Functional with Hidden Mediators

- Suppose that there exist Z and W satisfying

$$Z \perp\!\!\!\perp Y \,|\, (X, D, M) \qquad\qquad W \perp\!\!\!\perp (D, Z) \,|\, (X, M)$$

- Define an outcome confounding bridge functions

$$\mathrm{E}\big[h(W, X) \,|\, Z, D = 1, X\big] = \mathrm{E}(Y \,|\, Z, D = 1, X)$$

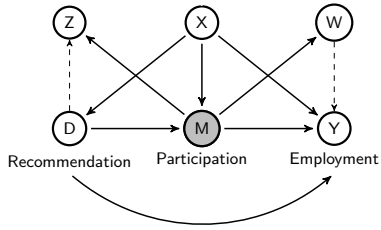
- The mediation functional is identified [13][1]

$$\mathrm{E}\big[Y^{(1, M^{(0)})}\big] = \mathrm{E}_X\left[\sum_w h(w, X) \Pr(W = w \,|\, D = 0, X)\right]$$

Original mediation formula:

$$\mathrm{E}[Y^{(1, M^{(0)})}] = \mathrm{E}_X\left[\sum_m \mathrm{E}(Y \,|\, D = 1, M = m, X) \Pr(M = m \,|\, D = 0, X)\right]$$

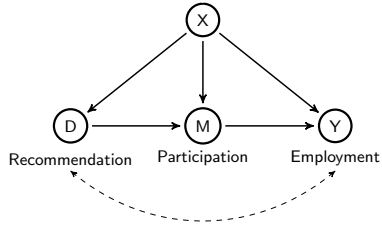
- An asymptotically normal estimator can be constructed



[1] Ghassami, Yang, Shpitser, TT (2024). Causal inference with hidden mediators. *Biometrika*

Front-Door Model

- Suppose $Y^{(d,m)} = Y^{(m)}$ (exclusion restriction)
  and allow for unmeasured confounding between D and Y



- **Front-door Model** [24][1]:
  the ATE can be identified even though D-Y are confounded

$$\mathrm{E}[Y^{(0)}] = \mathrm{E}[(1-D)Y] + \mathrm{E}_X\left[\sum_m \mathrm{E}(Y \,|\, D=1, M=m, X)\,\Pr(M=m \,|\, D=0, X)\,\Pr(D=1 \,|\, X)\right]$$

- What if M is hidden?

[1] Pearl (2000). *Causality: Models, Reasoning, and Inference*

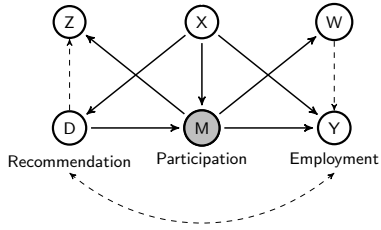
Identification of the Front-Door Model with Hidden Mediators

- Suppose that there exist Z and W satisfying

$$Z \perp\!\!\!\perp Y \,\big|\, (X, D, M) \qquad\qquad W \perp\!\!\!\perp (D, Z) \,\big|\, (X, M)$$

- Define an outcome confounding bridge functions

$$\mathrm{E}\big[h(W, X) \,\big|\, Z, D = 0, X\big] = \mathrm{E}(Y \,\big|\, Z, D = 0, X)$$



- The ATE is identified [13][1]

$$\mathrm{E}[Y^{(0)}] = \mathrm{E}[(1 - D)Y] + \mathrm{E}_X\left[\sum_w h(w, X)\,\Pr(W = w \,\big|\, D = 1, X)\,\Pr(D = 0 \,\big|\, X)\right]$$

- An asymptotically normal estimator can be constructed

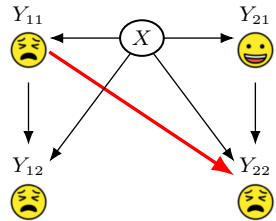
[1] Ghassami, Yang, Shpitser, TT (2024). Causal inference with hidden mediators. *Biometrika*

Extensions

- Longitudinal Settings
- Mediation analysis with hidden confounders
- Mediation analysis and front-door model with hidden mediators
- Interference with homophily driven by hidden factors
- Proximal synthetic controls

Causal Inference with Dependence/Interference

- Consider a married couple & flu transmission
- Y_{jt} = unit j in time t
- $Y_{22}^{(y_{11})}$ = pot. outcome of unit 2 in time 2 provided that $Y_{11} = y_{11}$
- Consistency: $Y_{22} = Y_{22}^{(Y_{11})}$
- X = Couple's measured covariate



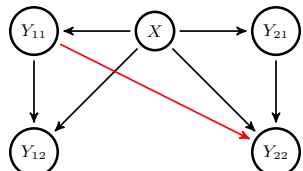
- Average Causal Peer Effect = $\mathrm{E}[Y_{22}^{(y_{11})} - Y_{22}^{(y'_{11})}]$
- Can be identified under no unmeasured confounding assumption $Y_{22}^{(y_{11})} \perp\!\!\!\perp Y_{11} \,|\, (Y_{21}, X)$

$$\texttt{ACPE} = \mathrm{E}[\mathrm{E}(Y_{22} \,|\, Y_{11} = y_{11}, Y_{21}, X) - \mathrm{E}(Y_{22} \,|\, Y_{11} = y'_{11}, Y_{21}, X)]$$

- What are we missing?

Homophily

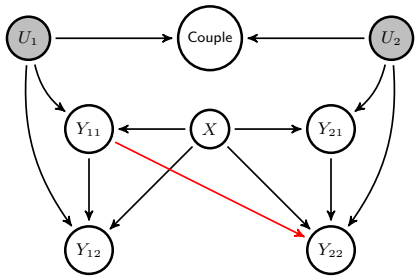
- Why did they become a couple/dyad?
- Measured variables X (age, gender) may not contain the key reason



Homophily

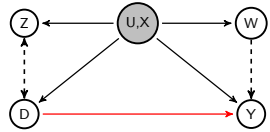
- Why did they become a couple/dyad?
- Measured variables X (age, gender) may not contain the key reason
- Instead, it may be driven by unmeasured characteristics U

  (cat/dog person, hometown, hobbies, overlapping schools, use of dating apps, etc)
- **Homophily Bias** arises when people become connected due to unobserved characteristics
- Proximal causal inference can be used to address homophily bias



Connecting the Dyad Example to Proximal Causal Inference

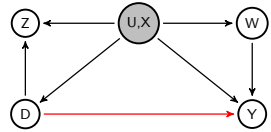
- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D, X)$
- W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Connecting the Dyad Example to Proximal Causal Inference

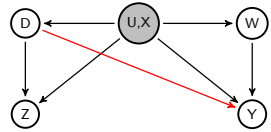
- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D, X)$
- W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

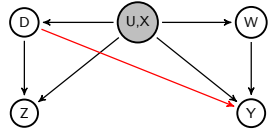
Connecting the Dyad Example to Proximal Causal Inference

- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D, X)$
- W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$

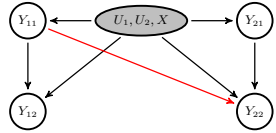


[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Connecting the Dyad Example to Proximal Causal Inference



- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D, X)$
- W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$



- Consistent with the dyad model under homophily bias

  NCE: Y_{12} = peer's outcome at time 2

  NCO: Y_{21} = focal unit's outcome at time 1

- Suppose $h(Y_{21}, Y_{11}, X)$ satisfies

$$\mathrm{E}[h(Y_{22}, Y_{11}, X) \,|\, Y_{12}, Y_{11}, X] = \mathrm{E}(Y_{22} \,|\, Y_{12}, Y_{11}, X)$$

- [9][1]: The ACPE is identified by

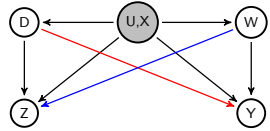
$$\texttt{ACPE} = \mathrm{E}[Y_{22}^{(y_{11})} - Y_{22}^{(y'_{11})}] = \mathrm{E}[h(Y_{22}, y_{11}, X) - h(Y_{22}, y'_{11}, X)]$$

- An asymptotically normal estimator can be constructed

[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Connecting the Dyad Example to Proximal Causal Inference

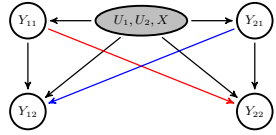
- Z is an NCE if $Z \perp\!\!\!\perp Y \,|\, (U, D, X)$
- W is an NCO if $W \perp\!\!\!\perp (D, Z) \,|\, (U, X)$



- Consistent with the dyad model under homophily bias

  NCE: Y_{12} = peer's outcome at time 2

  NCO: Y_{21} = focal unit's outcome at time 1



- Suppose $h(Y_{21}, Y_{11}, X)$ satisfies

$$\mathrm{E}[h(Y_{22}, Y_{11}, X) \,|\, Y_{12}, Y_{11}, X] = \mathrm{E}(Y_{22} \,|\, Y_{12}, Y_{11}, X)$$

- [9][1]: The ACPE is identified by

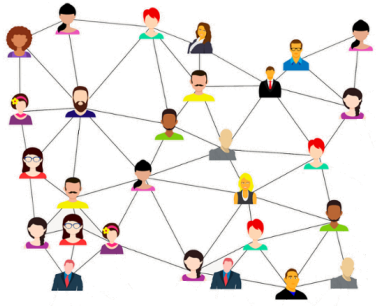
$$\texttt{ACPE} = \mathrm{E}[Y_{22}^{(y_{11})} - Y_{22}^{(y'_{11})}] = \mathrm{E}[h(Y_{22}, y_{11}, X) - h(Y_{22}, y'_{11}, X)]$$

- An asymptotically normal estimator can be constructed
- If $Y_{21} \to Y_{12}$, we need another NCE or NCO

[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extension to Network Data

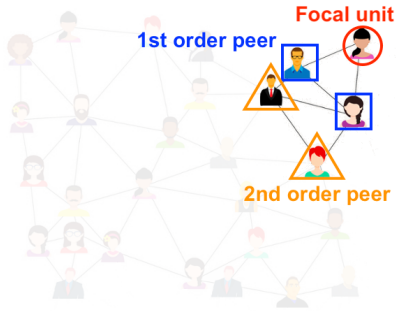
- Consider a social network



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers



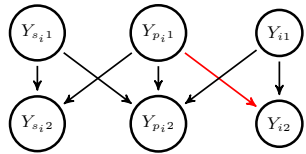
[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers

- Average Causal Peer Effect

$$\texttt{ACPE} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{Y_{i2}^{(y_{p_i1})} - Y_{i2}^{(y'_{p_i1})}\right\}\right]$$



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

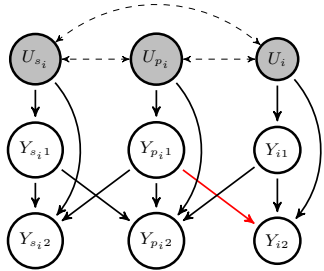
Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers

- Average Causal Peer Effect

$$\texttt{ACPE} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{Y_{i2}^{(y_{p_i1})} - Y_{i2}^{(y'_{p_i1})}\right\}\right]$$

- Unmeasured characteristics may exist



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

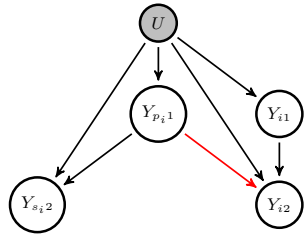
Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers

- Average Causal Peer Effect

$$\texttt{ACPE} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{Y_{i2}^{(y_{p_i 1})} - Y_{i2}^{(y'_{p_i 1})}\right\}\right]$$

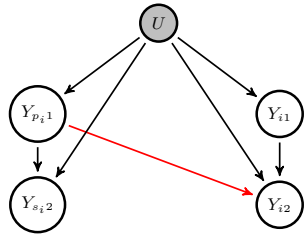
- Unmeasured characteristics may exist



[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers



- Average Causal Peer Effect

$$\texttt{ACPE} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{Y_{i2}^{(y_{p_i1})} - Y_{i2}^{(y'_{p_i1})}\right\}\right]$$

- Unmeasured characteristics may exist

[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extension to Network Data

- Consider a social network
- Given a focal unit, find its 1st and 2nd-order peers

- Average Causal Peer Effect

$$\texttt{ACPE} = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{Y_{i2}^{(y_{p_i1})} - Y_{i2}^{(y'_{p_i1})}\right\}\right]$$

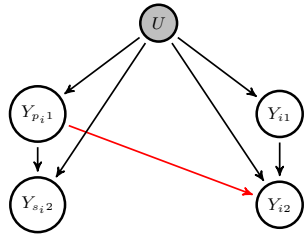
- Unmeasured characteristics may exist

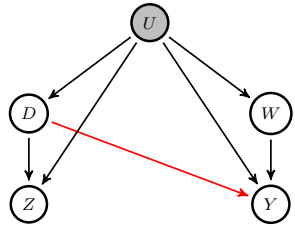
- Proximal causal inference for network data:

  NCE $= Y_{s_i2} =$ 2nd-order peers' outcome at time 2

  NCO $= Y_{i1} =$ focal unit's outcome at time 1

- Identification & estimation can be established [9][1]





[1] Egami, TT (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *JRSS:B*

Extensions

- Longitudinal Settings
- Mediation analysis with hidden confounders
- Mediation analysis and front-door model with hidden mediators
- Interference with homophily driven by hidden factors
- Proximal synthetic controls

Synthetic Control Approach

- California tobacco control program in Abadie et al. [1][1]

Treatment:	Proposition 99 (a large-scale tobacco control program) in 1988
Pre- and post-intervention periods:	$T_0 = 19$ (1970-1988) and $T_1 = 12$ (1989-2000) years
Outcome Y_t:	Per-capita cigarette sales of California in year t
Outcome W_{it}:	Per-capita cigarette sales of other 38 states in year t

[1] Abadie, Diamond, Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *JASA*

Synthetic Control Approach

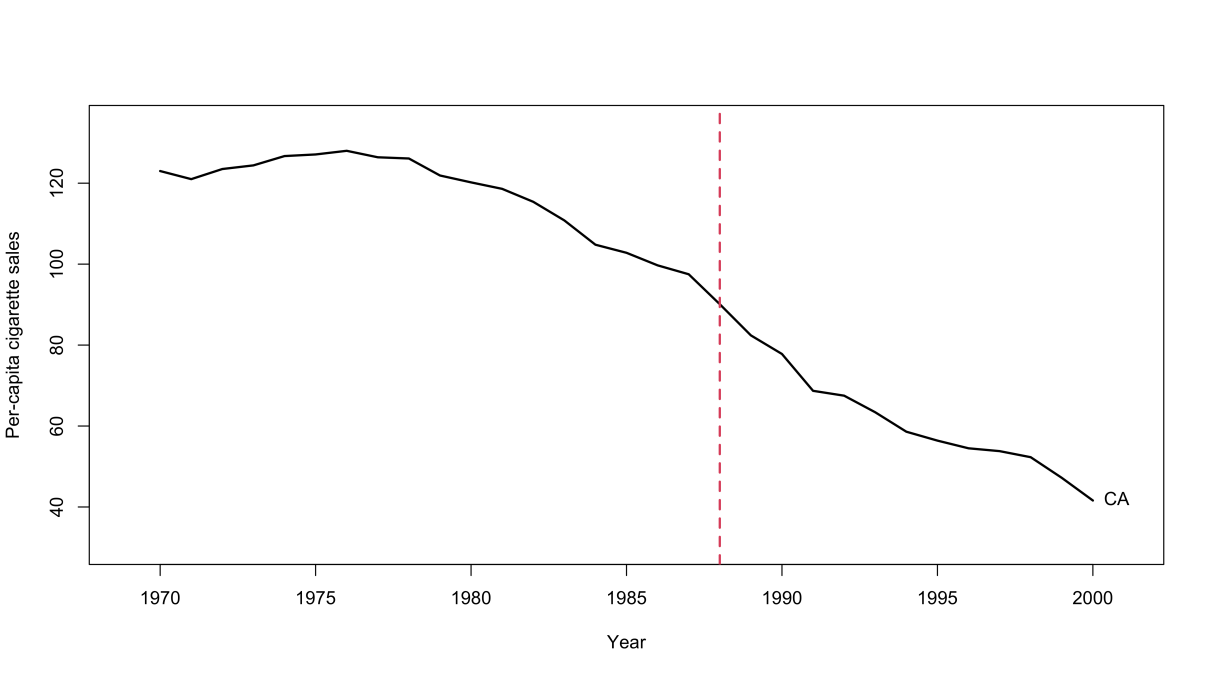
- California tobacco control program in Abadie et al. [1][1]

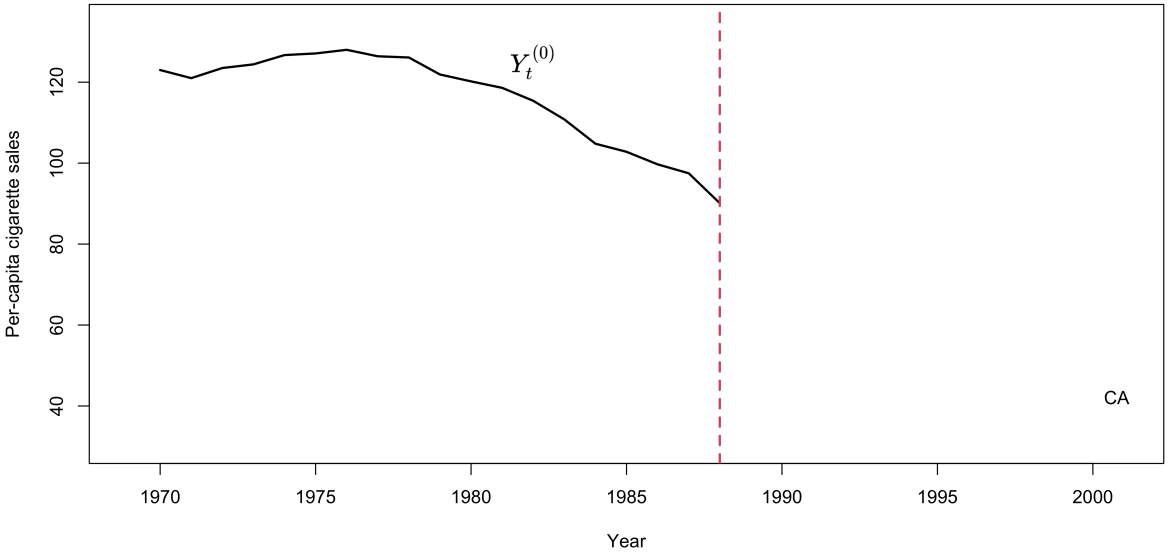
Treatment:	Proposition 99 (a large-scale tobacco control program) in 1988
Pre- and post-intervention periods:	$T_0 = 19$ (1970-1988) and $T_1 = 12$ (1989-2000) years
Outcome Y_t:	Per-capita cigarette sales of California in year t
Outcome W_{it}:	Per-capita cigarette sales of other 38 states in year t

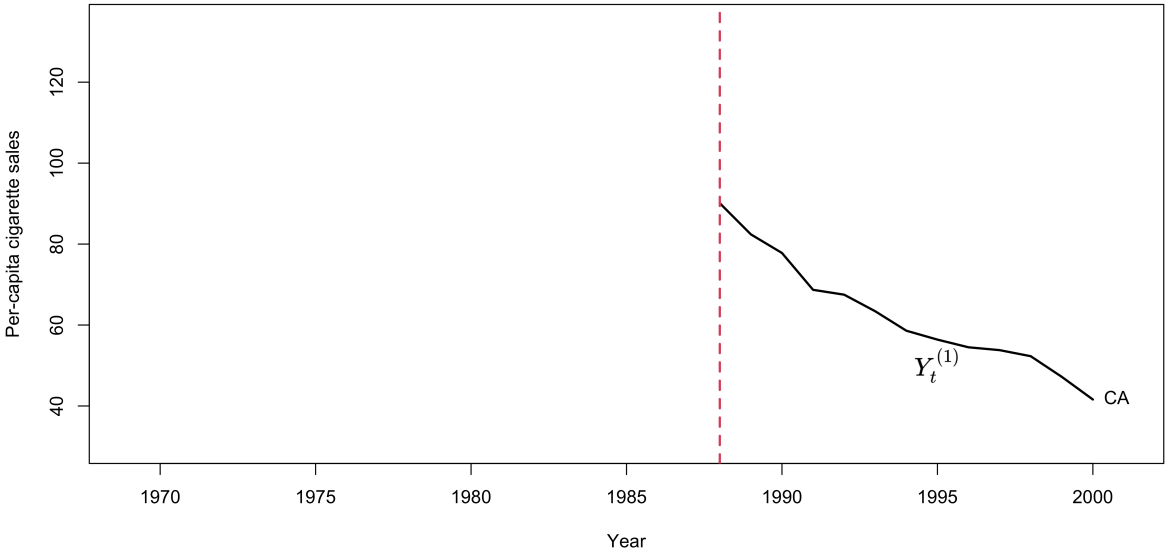
- Estimand: Effect of California's Proposition 99 on California's per-capita cigarette sales

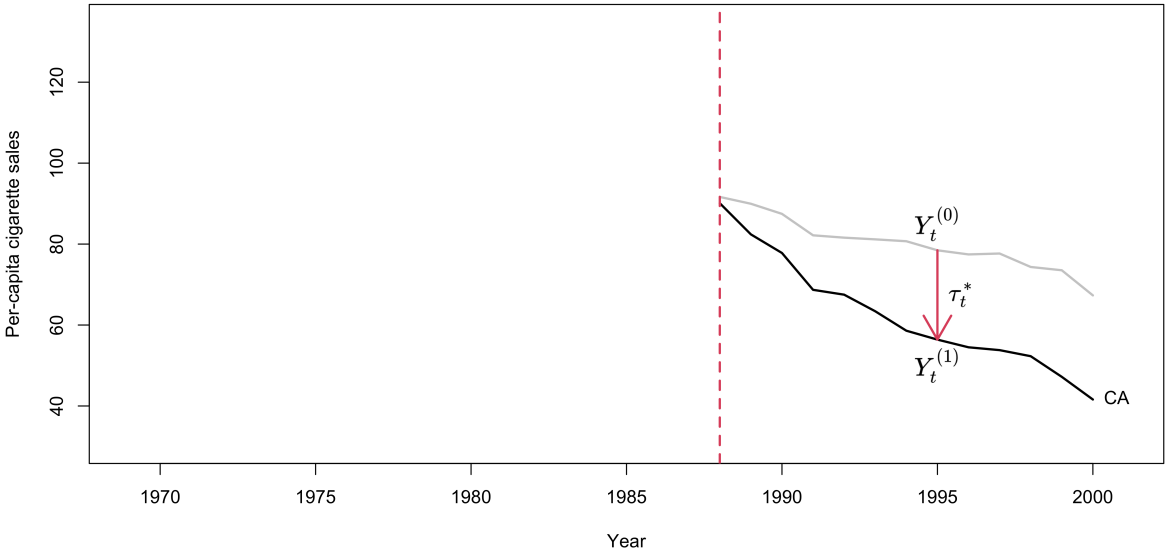
$$\tau_t^* = \mathrm{E}\big[Y_t^{(1)} - Y_t^{(0)}\big] \,, \quad 1988 < t$$

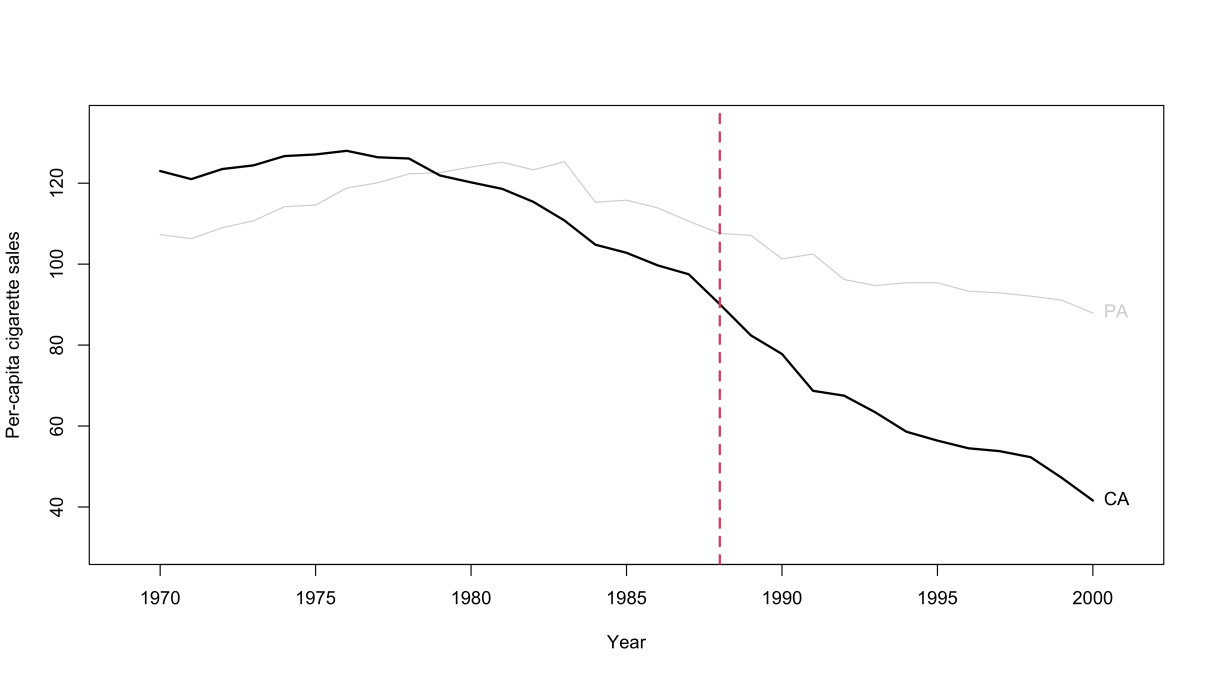
[1] Abadie, Diamond, Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *JASA*

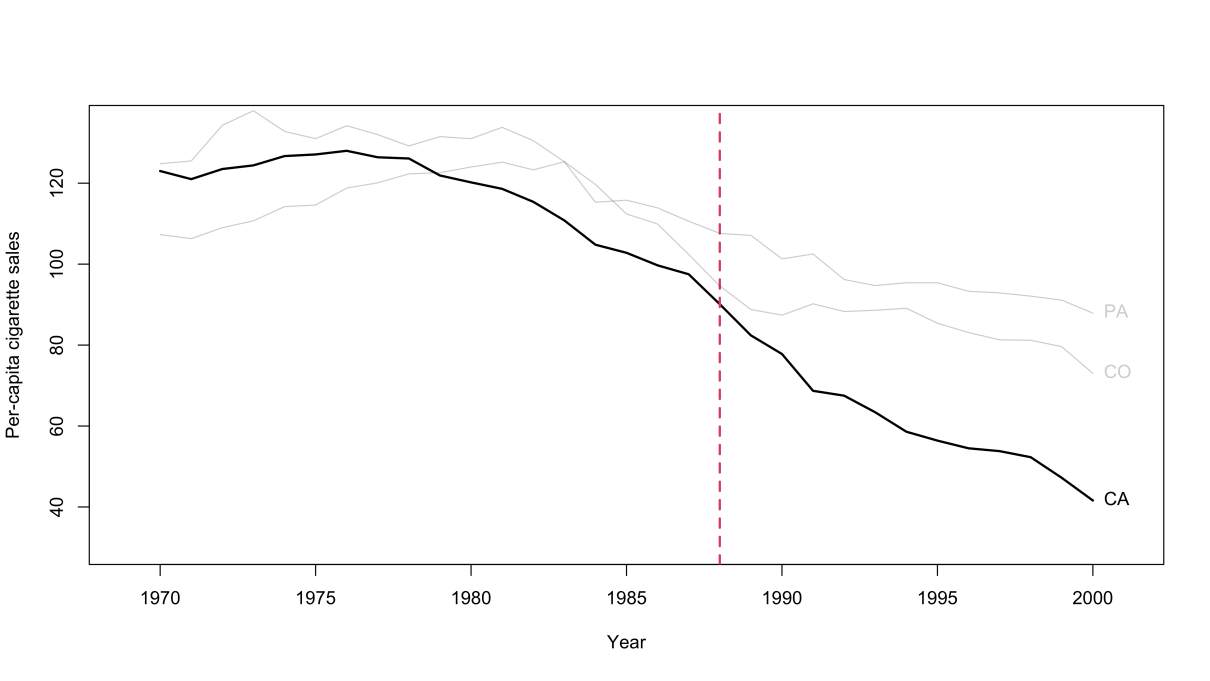

Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
CA

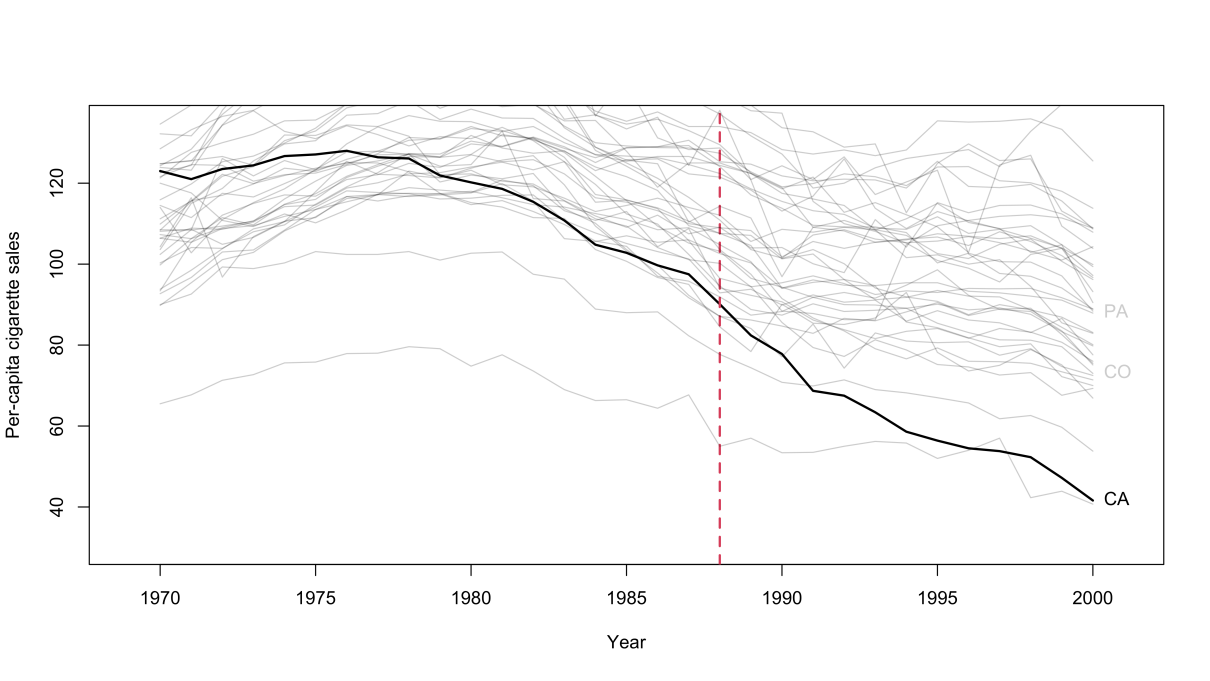

$Y_t^{(0)}$
CA
Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year

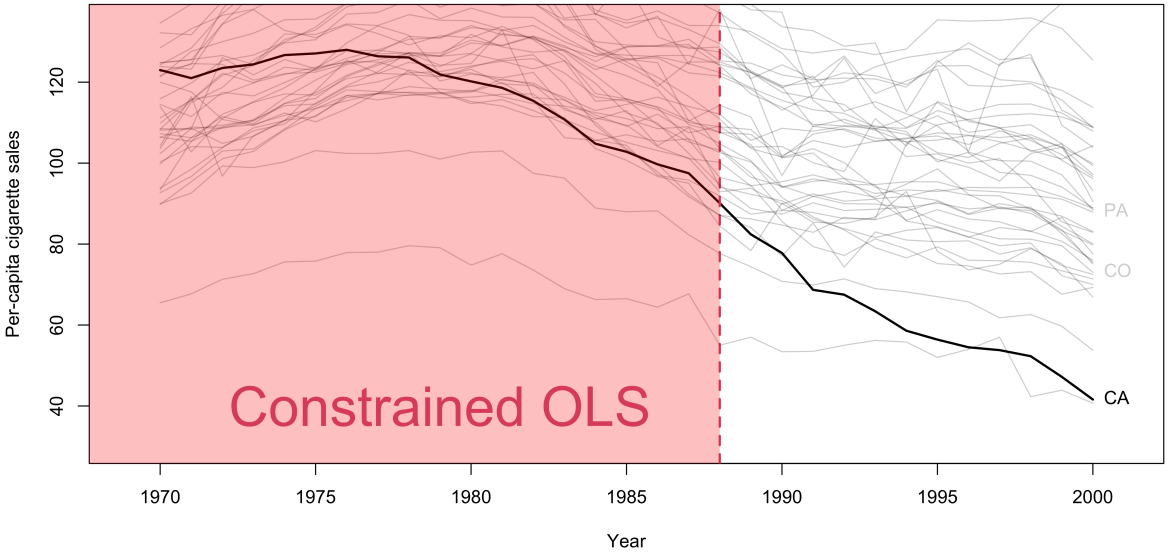

Per-capita cigarette sales
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year
$Y_t^{(1)}$
CA

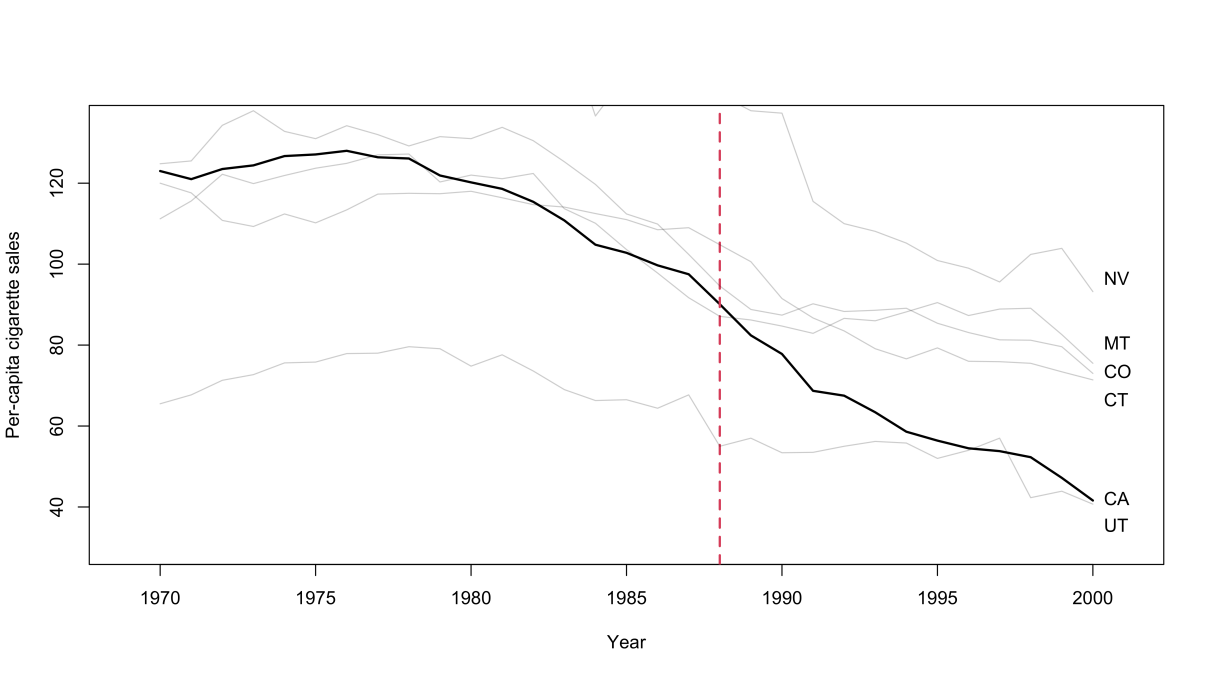

Per-capita cigarette sales
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year
$Y_t^{(0)}$
τ_t^*
$Y_t^{(1)}$
CA

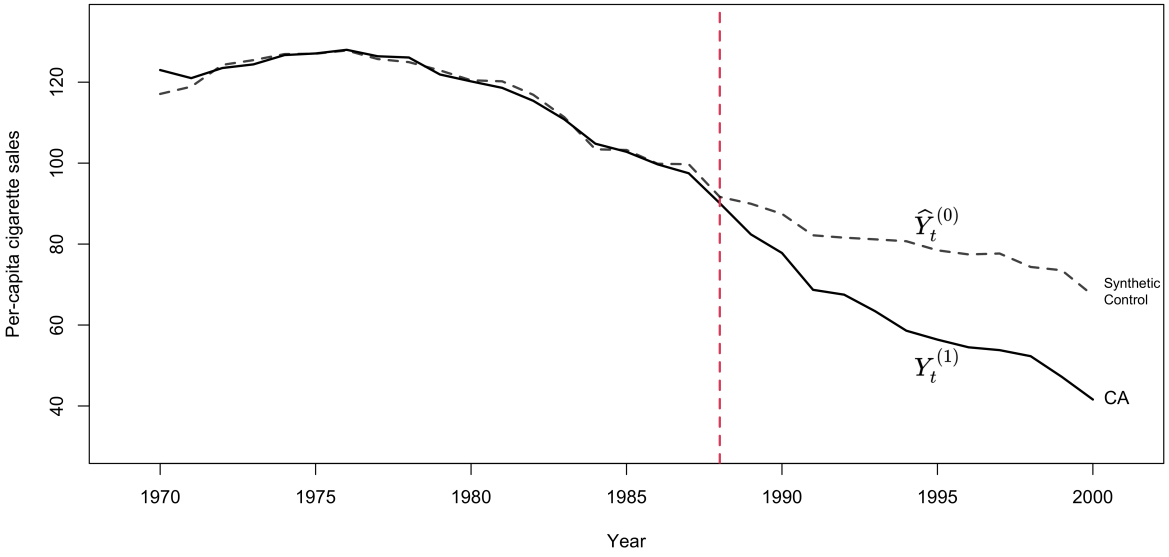

Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
PA
CA

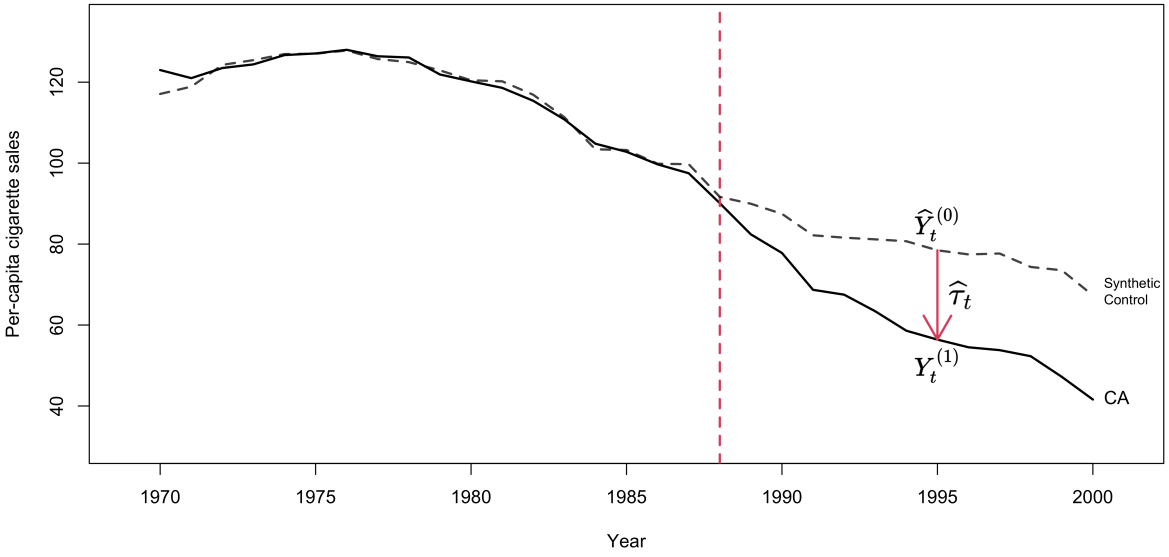

Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
PA
CO
CA


Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
PA
CO
CA


Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
Constrained OLS
PA
CO
CA


Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
NV
MT
CO
CT
CA
UT


Per-capita cigarette sales
40
60
80
100
120
1970
1975
1980
1985
1990
1995
2000
Year
$\widehat{Y}_t^{(0)}$
Synthetic
Control
$Y_t^{(1)}$
CA


Per-capita cigarette sales
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year
$\widehat{Y}_t^{(0)}$
$\widehat{\tau}_t$
$Y_t^{(1)}$
Synthetic
Control
CA

Synthetic Control Approach

- California tobacco control program in Abadie et al. [1][1]

Treatment:	Proposition 99 (a large-scale tobacco control program) in 1988
Pre- and post-intervention periods:	$T_0 = 19$ (1970-1988) and $T_1 = 12$ (1989-2000) years
Outcome Y_t:	Per-capita cigarette sales of California in year t
Outcome W_{it}:	Per-capita cigarette sales of other 38 states in year t

- Estimand: Effect of California's Proposition 99 on California's per-capita cigarette sales

$$\tau_t^* = \mathrm{E}\big[Y_t^{(1)} - Y_t^{(0)}\big] \,, \quad 1988 < t$$

- Synthetic control: find a linear combination of other states

$$\widehat{Y}_t^{(0)} \leftarrow \texttt{constrained OLS}(Y_t \sim W_t, \texttt{data} = \texttt{pre-trt data}) \quad \Rightarrow \quad \widehat{\tau}_t = Y_t - \widehat{Y}_t^{(0)}$$

[1] Abadie, Diamond, Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *JASA*

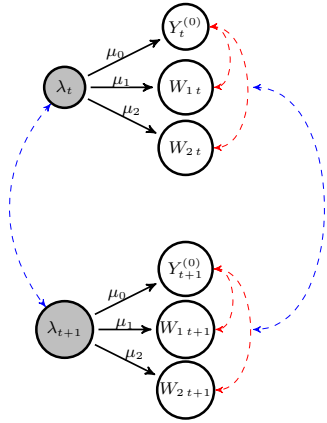
Interactive Fixed Effects Model

- Consider an IFEM [2][1]

$$\begin{aligned} Y_t^{(0)} &= \mu_0^\intercal \lambda_t + \epsilon_{0t} \\ W_{it} &= \mu_i^\intercal \lambda_t + \epsilon_{it} \end{aligned}$$

λ_t: latent factors $\qquad$ μ_i: factor loadings

ϵ_{it}: error term with $\mathrm{E}(\epsilon_{it} \,|\, \lambda_t) = 0$



- Suppose there exists γ such that $\mu_0 = \sum_i \gamma_i \mu_i$

$$\mathrm{E}[Y_t^{(0)} - \textstyle\sum_i \gamma_i W_{it}] = \mathrm{E}[\underbrace{(\mu_0 - \textstyle\sum_i \gamma_i \mu_i)}_{=0}{}^\intercal \lambda_t + \underbrace{\epsilon_{0t} - \textstyle\sum_i \gamma_i \epsilon_{it}}_{\text{error}}] = 0$$

- $W_t^\intercal \gamma = \sum_i \gamma_i W_{it}$ is a **synthetic control** with $\mathrm{E}[Y_t^{(0)}] = \mathrm{E}[W_t^\intercal \gamma]$

- $\tau_t = \mathrm{E}[Y_t^{(1)} - Y_t^{(0)}] = \mathrm{E}[Y_t - W_t^\intercal \gamma]$ for $t \in \{\text{post-trt}\}$

[1] Bai (2009). Panel Data Models with Interactive Fixed Effects. *Econometrica*

Naive OLS

- IFEM with a SC

$$\begin{array}{rcccl} Y_t^{(0)} & = & \mu_0^\intercal \lambda_t & + & \epsilon_{0t} \\ W_{it} & = & \mu_i^\intercal \lambda_t & + & \epsilon_{it} \end{array} \quad , \quad \mu_0 = \sum_i \gamma_i \mu_i$$

- A linear model over the pre-treatment periods

$$Y_t = \sum_i \gamma_i W_{it} + \underbrace{\left(\epsilon_{0t} - \sum_i \gamma_i \epsilon_{it} \right)}_{\text{error}} \quad \Rightarrow \quad Y_t = \sum_i \gamma_i W_{it} + \eta_t$$

- Naive OLS: $\texttt{lm}(Y_t \sim 0 + W_{it})$ to estimate γ using $t \in$ {pre-trt}

- The OLS estimator is **inconsistent**! because the regressors W_{it} and the error are correlated

- An exception: noise-less setting $(\epsilon_{it} = 0)$, but nearly impossible in real-world applications

Proximal Synthetic Control Approach [30][1]

- IFEM with a SC

$$\begin{aligned} Y_t^{(0)} &= \mu_0^\intercal \lambda_t + \epsilon_{0t} \\ W_{it} &= \mu_i^\intercal \lambda_t + \epsilon_{it} \end{aligned} \quad , \quad \mu_0 = \sum_i \gamma_i \mu_i$$

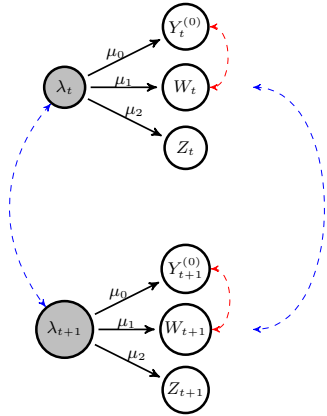
- Suppose we have Z_t satisfying $Z_t \perp\!\!\!\perp (Y_t^{(0)}, W_t) \,|\, \lambda_t$

- The SC weight γ can be identified via a moment condition:

$$\mathrm{E}\big[g(Z_t) \times (Y_t - W_t^\intercal \gamma)\big] = 0 \ , \quad \forall g(Z_t), \quad t \in \{\text{pre-trt}\}$$

- Estimate γ, and construct a consistent SC estimator:

$$\begin{aligned} \widehat{\gamma} \ &\leftarrow \ \texttt{aver}\big[g(Z_t) \times (Y_t - W_t^\intercal \gamma)\big] = 0 \ , && t \in \{\text{pre-trt}\} \\ \widehat{\tau}_t \ &\leftarrow \ Y_t - W_t^\intercal \widehat{\gamma} \ , && t \in \{\text{post-trt}\} \end{aligned}$$



[1] Shi, Li, Miao, Hu, TT (2023). Theory for identification and inference with synthetic controls: A proximal causal inference framework. *arXiv*

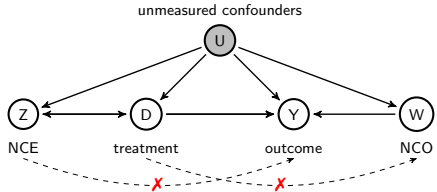
Summary

- Negative Controls

  NCE: $Z \perp\!\!\!\perp Y \,|\, (U, D)$ and $Z \not\perp\!\!\!\perp U \,|\, D$

  NCO: $W \perp\!\!\!\perp (Z, D) \,|\, U$ and $W \not\perp\!\!\!\perp U$

  Can be used to detect, reduce, and correct confounding bias



- Proximal Causal Inference

  Divide covariates into three buckets

  confounder, treatment confounding proxy (NCE), outcome confounding proxy (NCO)

  Identification and estimation

GLM	→	2-stage regression procedure
bridge functions	→	proximal IPW, proximal g-formula, proximal AIPW

- Extensions of Proximal Causal Inference to Various Settings

  Longitudinal settings

  Mediation, front-door formula

  Network data

  Time series data

Demonstration using R

github.com/qkrcks0218/ACIC2025

Reference

[1] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

[2] Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

[3] Browning, M. and Crossley, T. (2009). Are two cheap, noisy measures better than one expensive, accurate one? *American Economic Review*, 99(2):99–103.

[4] Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897.

[5] Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

[6] Davey Smith, G. (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*, 102(2):245–256.

[7] Davey Smith, G. (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies". *Epidemiology*, 23(2):350–351.

[8] Dukes, O., Shpitser, I., and Tchetgen Tchetgen, E. J. (2023). Proximal mediation analysis. *Biometrika*. asad015.

[9] Egami, N. and Tchetgen Tchetgen, E. J. (2023). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):487–511.

[10] Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., and Tolbert, P. E. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*, 22(1):59.

[11] Flanders, W. D., Strickland, M. J., and Klein, M. (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *American Journal of Epidemiology*, 185(10):941–949.

[12] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

Reference

[13] Ghassami, A., Yang, A., Shpitser, I., and Tchetgen Tchetgen, E. (2024). Causal inference with hidden mediators. *Biometrika*, 112(1):asae037.

[14] Ghassami, A., Ying, A., Shpitser, I., and Tchetgen Tchetgen, E. (2022). Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7210–7239. PMLR.

[15] Glynn, A. and Ichino, N. (2019). Generalized nonlinear difference-in-difference-in-differences. *preprint*, 90.

[16] Jacob, L., Gagnon-Bartsch, J. A., and Speed, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28.

[17] Li, K., Emerman, I., Cook, A. J., Fireman, B. H., Sundaram, M., Tseng, H.-F. X., Weintraub, E. S., Yu, O., Nelson, J. L., and Shi, X. (2024). Using double negative controls to adjust for healthy user bias in a recombinant zoster vaccine safety study. *American Journal of Epidemiology*, page kwae439.

[18] Liu, J., Park, C., Li, K., and Tchetgen Tchetgen, E. J. (2024). Regression-based proximal causal inference. *American Journal of Epidemiology*, page kwae370.

[19] Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7512–7523. PMLR.

[20] Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

[21] Miao, W., Shi, X., Li, Y., and Tchetgen Tchetgen, E. J. (2024). A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*, 8(4):262–273.

[22] Miao, W. and Tchetgen Tchetgen, E. J. (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *American Journal of Epidemiology*, 185(10):950–953.

[23] Park, C., Richardson, D. B., and Tchetgen Tchetgen, E. J. (2024). Single proxy control. *Biometrics*, 80(2):ujae027.

[24] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Reference

[25] Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420. Morgan Kaufmann Publishers Inc.

[26] Richardson, D. B., Keil, A., Tchetgen Tchetgen, E. J., and Cooper, G. S. (2015). Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology*, 26(5):727–732.

[27] Richardson, D. B., Laurier, D., Schubauer-Berigan, M. K., Tchetgen Tchetgen, E. J., and Cole, S. R. (2014). Assessment and indirect adjustment for confounding by smoking in cohort studies using relative hazards models. *American Journal of Epidemiology*, 180(9):933–940.

[28] Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., and Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577.

[29] Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., and Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33(2):209–218.

[30] Shi, X., Li, K., Miao, W., Hu, M., and Tchetgen Tchetgen, E. (2023). Theory for identification and inference with synthetic controls: A proximal causal inference framework. *Preprint arXiv:2108.13935*.

[31] Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2020a). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, 7:190–202.

[32] Shi, X., Miao, W., and Tchetgen Tchetgen, E. J. (2020b). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540.

[33] Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32.

[34] Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, 31(3):348–361.

Reference

[35] Tchetgen Tchetgen, E. (2014). The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640.

[36] Tchetgen Tchetgen, E. J., Sofer, T., and Richardson, D. (2015). Negative outcome control for unobserved confounding under a cox proportional hazards model. *preprint*.

[37] Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2024). An introduction to proximal causal inference. *Statistical Science*, 39(3):375 – 390.

[38] Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics*, 45(5):1863–1894.

[39] Weisskopf, M. G., Tchetgen Tchetgen, E. J., and Raz, R. (2016). Commentary: on the use of imperfect negative control exposures in epidemiologic studies. *Epidemiology*, 27(3):365–367.

[40] Ying, A., Miao, W., Shi, X., and Tchetgen Tchetgen, E. J. (2023). Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. qkad020.