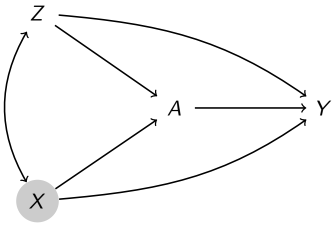
Covariate measurement error in propensity score analysis: Leveraging the covariate's posterior mean

(or, the inclusive factor score as proxy for a latent covariate)

Trang Quynh Nguyen, Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health

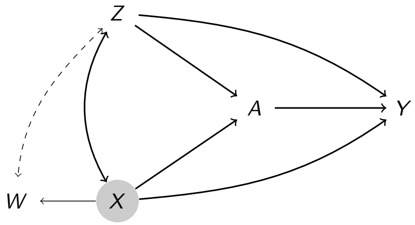
JSM, 2019/07/31

Motivating context: latent covariate in PS analysis



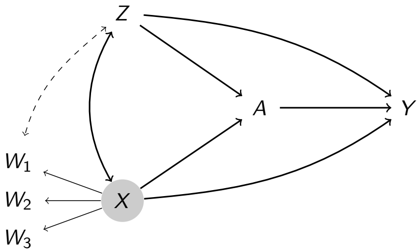
Ideal analysis: PS analysis based on (Z, X)

Motivating context: latent covariate in PS analysis



Problem: measurement error bias when using W as proxy for X

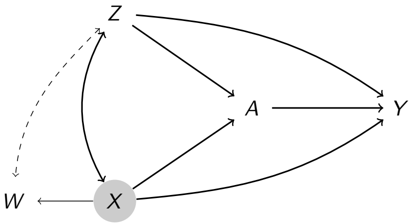
Motivating context: latent covariate in PS analysis



Setting characteristics:

- X latent construct that is never directly observed
- W consists of multiple measurements

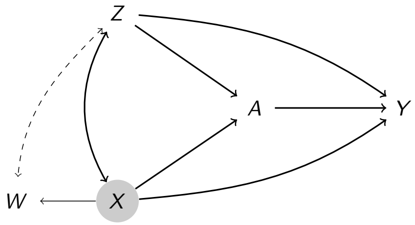
Motivating context: latent covariate in PS analysis



Add Health application:

- A suspension from school, Y problems with the law (police arrest)
- Z various individual and family characteristics
- X_1 violence tendency (measured via questions about fights and weapon use); X_2 academic achievement (measured via grades on several subjects)

Goal: Find a better proxy for X to be used in PS analysis



Once the proxy $(\tilde{X})$ obtained, analysis as usual

e.g., weighting/matching based on PS estimated with $(Z, \tilde{X})$

Idea

existing proxies suffer from measurement error bias

- W items
- sum/mean of W items (scale score)
- predicted value of X given W based on measurement model (Raykov 2012), aka the conventional factor score (cFS)

Idea

existing proxies suffer from measurement error bias

- W items
- sum/mean of W items (scale score)
- predicted value of X given W based on measurement model (Raykov 2012), aka the conventional factor score (cFS)

think imputation

- incompatibility of imputation model and analysis model
- $\tilde{X}$ should be informed by all variables in the PS model
  - leave Y out for design-analysis separation

Idea

existing proxies suffer from measurement error bias

- W items, scale score, cFS

think imputation

- incompatibility of imputation model and analysis model
- $\tilde{X}$ should be informed by all variables in the PS model
  - leave Y out for design-analysis separation

propose $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

- with latent X, estimated based on a SEM that combines the measurement model and the exposure assignment model
  - aka the inclusive factor score (iFS)

- theoretical support for this proxy
- how we estimate it
- simulations 1: models correctly specified, iFS estimates $\tilde{X}$ well
- simulations 2: iFS does not estimate $\tilde{X}$ as well

Quick connection to related work on weighting/matching functions

our work fits in the proxy variable approach, searching for a proxy for X

it implies using $H = (Z, \tilde{X})$ or $H = e(Z, \tilde{X})$ for matching and $Q = A/e(Z, \tilde{X}) + (1 - A)/[1 - e(Z, \tilde{X})]$ for weighting

Quick connection to related work on weighting/matching functions

our work fits in the proxy variable approach, searching for a proxy for X

it implies using $H = (Z, \tilde{X})$ or $H = e(Z, \tilde{X})$ for matching and $Q = A/e(Z, \tilde{X}) + (1 - A)/[1 - e(Z, \tilde{X})]$ for weighting

the weighting/matching function approach (McCaffrey et al. 2013; Lockwood & McCaffrey 2016) seeks functions of observed (and simulated) data that when used for weighting/matching lead to unbiased effect estimation

Quick connection to related work on weighting/matching functions

our work fits in the proxy variable approach, searching for a proxy for X

it implies using $H = (Z, \tilde{X})$ or $H = e(Z, \tilde{X})$ for matching and $Q = A/e(Z, \tilde{X}) + (1 - A)/[1 - e(Z, \tilde{X})]$ for weighting

the weighting/matching function approach (McCaffrey et al. 2013; Lockwood & McCaffrey 2016) seeks functions of observed (and simulated) data that when used for weighting/matching lead to unbiased effect estimation

Q and H are approximately unbiased weighting and matching functions: they target balance on the first moment of X while the exactly unbiased weighting/matching functions target full distribution balance

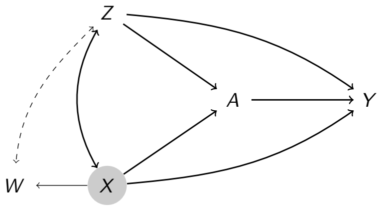
Assumptions – first layer

Causal inference assumptions

- SUTVA
- unconfoundedness: $A \perp\!\!\!\perp Y(a) \mid Z, X, \quad a = 0, 1$
- positivity: $0 < e(Z, X) < 1$

Measurement-related assumption

- strong surrogacy: $W \perp\!\!\!\perp A, Y(a) \mid Z, X$
- also weak surrogacy: $W \perp\!\!\!\perp Y(a) \mid Z, X, A$



Theoretical support for $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

because $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

$X - \tilde{X}$ has mean zero conditional on W, Z, A and $\tilde{X}$

which implies it has mean-balance (equality of the means between exposure conditions) before any data processing

and also has mean-balance after weighting by any positive bounded scalar function of $(Z, \tilde{X}, A)$, or matching on any function of $(Z, \tilde{X})$

Theoretical support for $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

because $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

$X - \tilde{X}$ has mean zero conditional on W, Z, A and $\tilde{X}$

which implies it has mean-balance (equality of the means between exposure conditions) before any data processing

and also has mean-balance after weighting by any positive bounded scalar function of $(Z, \tilde{X}, A)$, or matching on any function of $(Z, \tilde{X})$

since $X = \tilde{X} + (X - \tilde{X})$

this means PS analysis using $\tilde{X}$ as proxy for X can obtain mean-balance on X via obtaining balance on $\tilde{X}$, thus improve covariate balance

Theoretical support for $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

because $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

$X - \tilde{X}$ has mean zero conditional on W, Z, A and $\tilde{X}$

which implies it has mean-balance (equality of the means between exposure conditions) before any data processing

and also has mean-balance after weighting by any positive bounded scalar function of $(Z, \tilde{X}, A)$, or matching on any function of $(Z, \tilde{X})$

since $X = \tilde{X} + (X - \tilde{X})$

this means PS analysis using $\tilde{X}$ as proxy for X can obtain mean-balance on X via obtaining balance on $\tilde{X}$, thus improve covariate balance

balancing the distribution of Z and the mean of X allows unbiased effect estimation if the outcome is linear in X within each exposure condition

Identification and estimation of $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

challenge: the distribution of a latent variable is unidentified

to make progress requires additional assumptions

Identification and estimation of $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

challenge: the distribution of a latent variable is unidentified

to make progress requires additional assumptions

- sufficient conditional independence
  - W items (mostly) independent of Z and of one another given X
- selective distributional and functional form assumptions
  - X normal given Z
  - W normal-linear, or generalized linear, given X

(if X not latent, might use validation data and require fewer assumptions)

Identification and estimation of $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$

challenge: the distribution of a latent variable is unidentified

to make progress requires additional assumptions

- sufficient conditional independence
  - W items (mostly) independent of Z and of one another given X
- selective distributional and functional form assumptions
  - X normal given Z
  - W normal-linear, or generalized linear, given X

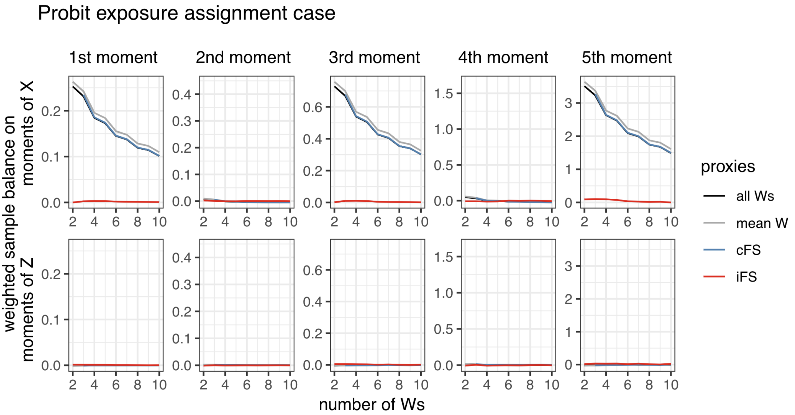
(if X not latent, might use validation data and require fewer assumptions)

estimation by Mplus (Muthen & Muthen 2016)

- SEM combining model components for $X|Z$, $W|X, Z$ and $A|Z, X$
- iFS computed using the EAP method (Bock & Aitkin 1981)

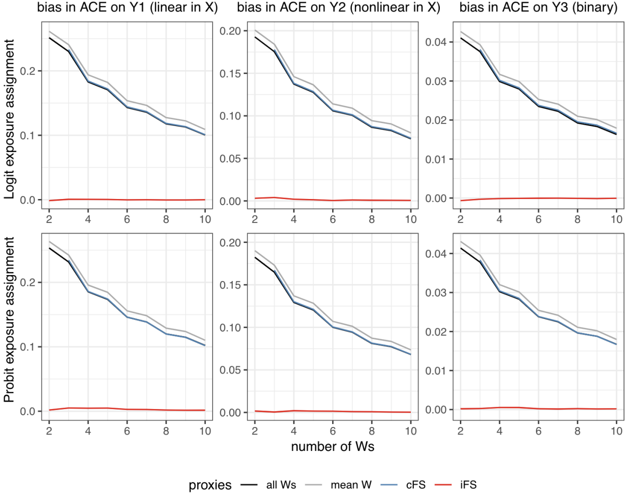
Sims 1: models correctly specified, iFS estimates $\tilde{X}$ well

balance on the covariates' first five moments obtained via PS weighting – centered at values obtained using the true X



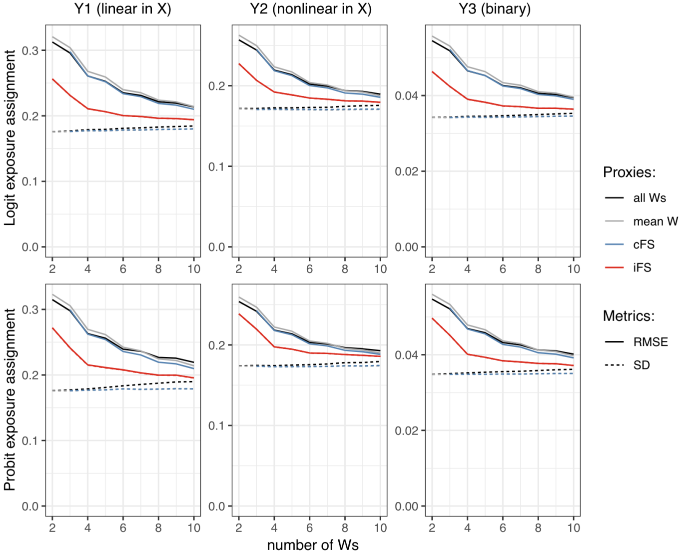
Sims 1: models correctly specified, iFS estimates $\tilde{X}$ well

bias in estimated ACE on 3 outcomes, linear and nonlinear in X – centered at values obtained using the true X

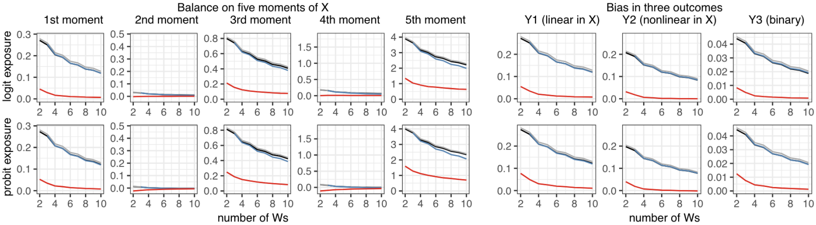


Sims 1: models correctly specified, iFS estimates $\tilde{X}$ well

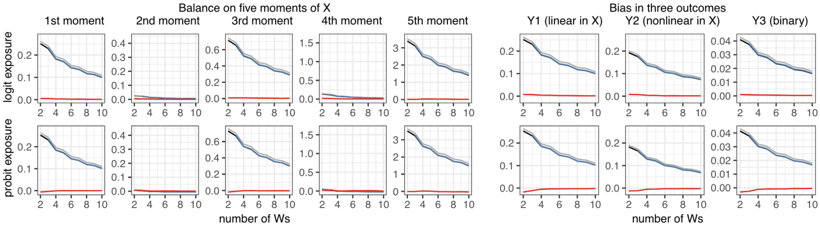
RMSE and SD



Sims 2: iFS does not estimate $\tilde{X}$ as well: ordinal W

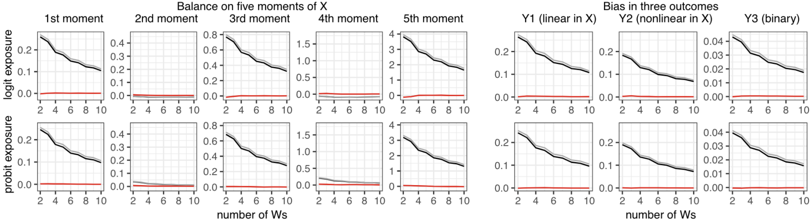


Sims 2: iFS does not estimate $\tilde{X}$ as well: linear iFS



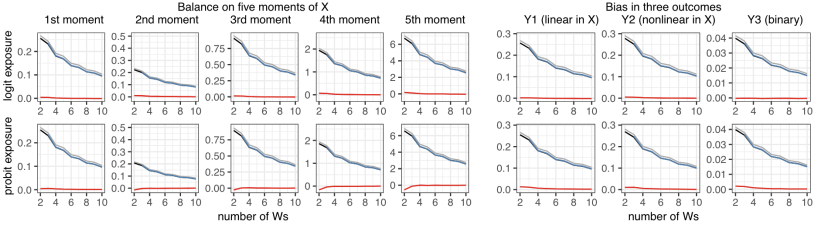
Sims 2: iFS does not estimate $\tilde{X}$ as well: wrong link

metrics compare to using the true X with the wrong link function – centered at values obtained when using the true X



Sims 2: iFS does not estimate $\tilde{X}$ as well: asymmetric dist's

$X \mid Z$ and $W \mid X, Z$ skewed



Summary

- propose $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$ as proxy for X in PS analysis
- theoretical result: balance on the first moment of X; unbiased effect estimation if outcome is linear in X
- simulation results
  - correct models, iFS estimates $\tilde{X}$ well: balance on first five moments of X, bias removal even with outcome nonlinear in X
  - iFS does not estimate $\tilde{X}$ as well: also performs well
    - w/ continuous W, balance and bias comparable to using the true X

Summary

- propose $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$ as proxy for X in PS analysis
- theoretical result: balance on the first moment of X; unbiased effect estimation if outcome is linear in X
- simulation results
  - correct models, iFS estimates $\tilde{X}$ well: balance on first five moments of X, bias removal even with outcome nonlinear in X
  - iFS does not estimate $\tilde{X}$ as well: also performs well
    - w/ continuous W, balance and bias comparable to using the true X
- while iFS specific to latent covariate, $\tilde{X} = \mathrm{E}[X \mid W, Z, A]$ relevant to any unobserved X measured indirectly through W

Application: sample, variables, and estimand

- Add Health is a nationally representative cohort of youth recruited in 1994-95 school year (when in grades 7-12)
- analysis sample restricted to individuals who at wave 1 had experienced school suspension
- exposure: additional suspension in the approximately one-year period between waves 1 and 2
- outcome: subsequent (up until wave 4 in 2008) police arrest
- estimand: ACEE, specifically for the group of exposed individuals in the sample
- covariates: baseline (wave 1) individual and family characteristics, including two latent variables violence tendency (ordinal alpha 0.81) and academic achievement (ordinal alpha 0.67)
- assume full conditional independence (few items)

Application: covariate balance

	Exposed group	Unexposed group before PS weighting		Unexposed group after PS weighting based on mean scores		Unexposed group after PS weighting and based on iFSs	
	mean (%)	mean (%)	SMD	mean (%)	SMD	mean (%)	SMD
Observed covariates							
Age	15.9	16.1	-0.16	15.9	0.02	15.8	0.08
Race (%)							
White	62.9	64.6	-0.04	64.3	-0.03	66.7	-0.08
Black/African-American	33.6	27.8	0.13	32.6	0.02	30.2	0.07
Native American	7.9	4.7	0.14	8.7	-0.03	8.4	-0.02
Asian	2.1	3.6	-0.08	1.4	0.05	1.1	0.08
Hispanic ethnicity (%)	8.6	10.8	-0.08	12.0	-0.11	11.5	-0.10
Parent education (%)							
Less than high school	18.6	17.0	0.04	21.2	-0.06	22.8	-0.10
High school	38.6	26.4	0.27	35.8	0.06	35.9	0.05
Business/vocational training	15.0	11.9	0.09	13.4	0.05	13.4	0.05
Some college (not graduated)	12.1	25.6	-0.33	11.4	0.02	10.4	0.06
College graduate or higher	15.7	19.1	-0.09	18.2	-0.06	17.5	-0.05
Parent marital status (%)							
Married	59.3	66.1	-0.14	56.7	0.05	56.8	0.05
Single	15.0	4.3	0.40	16.5	-0.04	15.7	-0.02
Widowed	3.6	4.3	-0.04	3.2	0.02	2.8	0.05
Divorced	15.0	20.6	-0.14	16.0	-0.03	16.9	-0.05
Separated	7.1	4.7	0.11	7.6	-0.02	7.9	-0.03
Proxies of latent covariates							
Violence							
Mean score (range 0-3)	0.65	0.43	0.39	0.66	-0.02	0.70	-0.09
Inclusive factor score	-0.86	-1.32	0.55	-0.94	0.08	-0.84	-0.03
Academic achievement							
Mean score (range 1-4)	1.18	1.52	-0.50	1.17	0.01	1.02	0.23
Inclusive factor score	-0.21	0.40	-0.75	0.01	-0.28	-0.20	-0.01

Application: changes in estimated ACEE due to measurement error bias correction using the iFS method

		WEIGHTING-ONLY ESTIMATOR			WEIGHTING-PLUS ESTIMATOR		
	outcome proportion in the exposed group	weighted outcome proportion in the unexposed	ACEE point estimate	95% confidence interval	mean predicted potential outcome probability under non-exposure for the exposed	ACEE point estimate	95% confidence interval
neither corrected	70.7	59.7	11.0	(3.1, 18.3)	59.1	11.6	(3.7, 18.5)
violence corrected	70.7	61.1	9.7		60.5	10.3	
acad. achiev. corrected	70.7	62.5	8.2		61.4	9.3	
both corrected	70.7	63.2	7.5	(-0.8, 15.9)	62.2	8.6	(1.7, 18.2)

Acknowledgements

- JEBS Editor Dan McCaffrey and three anonymous reviewers for pointing us to related work and pushing us to be better over the last few years
- Dan Scharfstein and Ilya Shpitser for their tough questions which are illuminating
- Betsy Ogburn for conversations about error propagation and collider bias
- Abhirup Datta for conversations about probability and the function Q
- National Institute of Mental Health grant R01MH099010 (PI Stuart)
- the Add Health study and cohort

Thank you!

tnguye28@jhu.edu