A fix for measurement error bias due to a latent covariate in propensity score weighting analysis: The factor score from a SEM combining the covariate's measurement model and the treatment assignment model

Trang Quynh Nguyen, Cyrus Ebnesajjad, Hwanhee Hong, Elizabeth Stuart
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Society for Prevention Research 24th Annual Meeting
San Francisco, CA – June 1, 2016

Summary

Methodological context: Propensity score methods are increasingly used to remove confounding to estimate the effect of a treatment/exposure/intervention. Key assumption is no unobserved confounding.

Problem: Measurement error bias due to a latent covariate

Our work: With multiple measurements of such variable, investigate bias and bias reduction when using factor scores to represent the variable, with several types of factor scores

Main finding: A simple fix is using the factor score from a SEM that includes the covariate's measurement model and the full treatment assignment model

Illustrative example: Analysis of Add Health data to examine out-of-school suspension as a risk factor for trouble with the law in young adulthood.

Propensity score and propensity score analysis

Consider a binary treatment. Interested in its effect on an outcome.

Propensity score = probability of being treated given a set of measured baseline covariates (conceptualized to be confounders of the treatment-outcome relationship)

The propensity score is a balancing score: conditioning on a level of the propensity score, we have balance on the covariates between treated and control.

Estimation of causal effect (using observational data) usually involves two steps

- Data pre-processing: to achieve balance on confounders (mimic RCT)
  - estimate the propensity score model
  - weight/match/subclassify based on the propensity score
    - for ATE, weight = inverse of probability of receiving the treatment received
- Analysis: use the pre-processed data to estimate the causal effect
    - either use the difference in means (or difference/ratio in proportions) of the outcome between the two conditions
    - or regress outcome on treatment, adjusting for covariates (*doubly robust*)

The no unobserved confounding assumption

In a randomized trial, randomization balances (in expectation) all characteristics, measured and not measured, realized and potential.

In observational studies, the propensity score balances only the variables used in estimating it, hence the assumption of no unobserved confounding.

This implies:

- No unobserved confounder: If a confounder is not accounted for in the propensity score model, we do not have balance on it, so it confounds the estimated treatment effect.
- No measurement error: If a confounder is measured with error, propensity score-based data processing may achieve balance on the mismeasured version but there is still some imbalance in the true covariate, and that residual imbalance confounds the estimated treatment effect.

Latent variable and measurement error

But some covariates may be latent variables, e.g., depression, anxiety, self-esteem, readiness-to-learn, substance dependence, stigma, etc.

Many are measured using multi-item scales.

A common practice in propensity score analysis is to use a summary score (sum/mean) of the items. This means there is measurement error bias.

With doubly robust propensity score analysis, measurement error bias affects both the propensity model and the outcome model.

A factor score strategy – prior work

Fit a latent factor model using measurement items, generate factor scores (FS) and use them to represent the latent variable in propensity score analysis.

Raykov (2012) suggested this strategy.

Jakubowski (2015) evaluated this strategy through simulation studies for propensity score matching.

It doesn't seem to work so well. My read: similarly biased to using the multiple measurements directly.

Our investigation

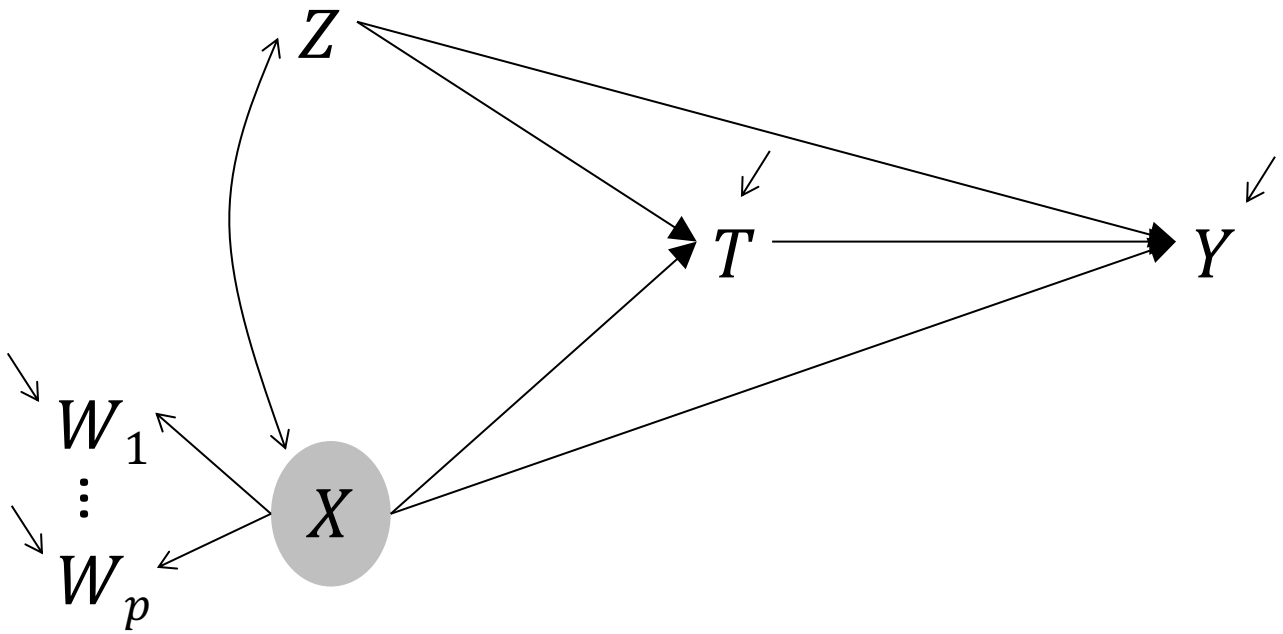
Analysis context: doubly robust propensity score weighting analysis

- propensity score weighting to balance covariates
- regression of outcome on treatment and covariates using the weighted sample

Different representations of the latent covariate in analysis:

- factor scores based on
  - the measurement model for the latent covariate (*simple*)
  - SEM linking the latent factor to the treatment variable (*partially inclusive*)
  - SEM that includes the full treatment assignment model (*fully inclusive*)
- two comparison methods
  - using the multiple measurements directly
  - using a summary score (mean/sum) of the measurements

Notation



T: treatment (binary)

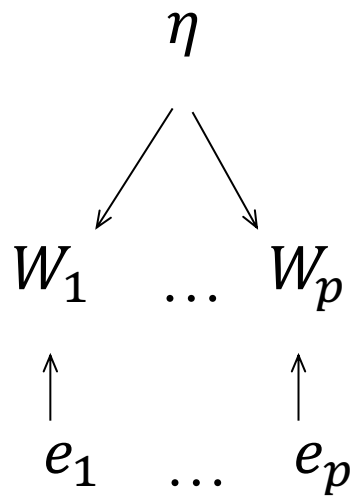
Y: outcome

Z: covariates measured without error

X: latent covariate, not observed

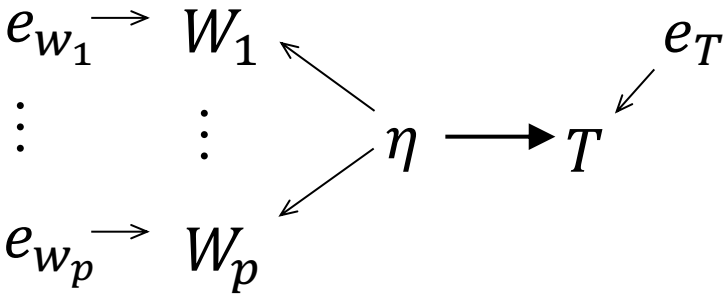
Ws: observed measurements, reflecting X and errors

Main models used to generate factor scores

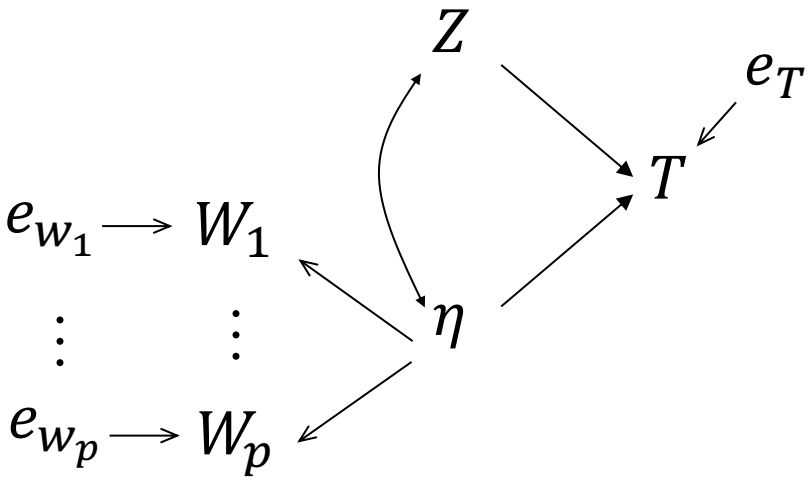


simple

factor score = predicted value of latent η given the observed variables and estimated model parameters



partially inclusive



fully inclusive

Model links and estimators, FS estimation methods

- Linear models: posterior mode (regression-based)
- Logit and probit models fit using ML: posterior mean
- Probit models also fit using WLSMV*: posterior mode

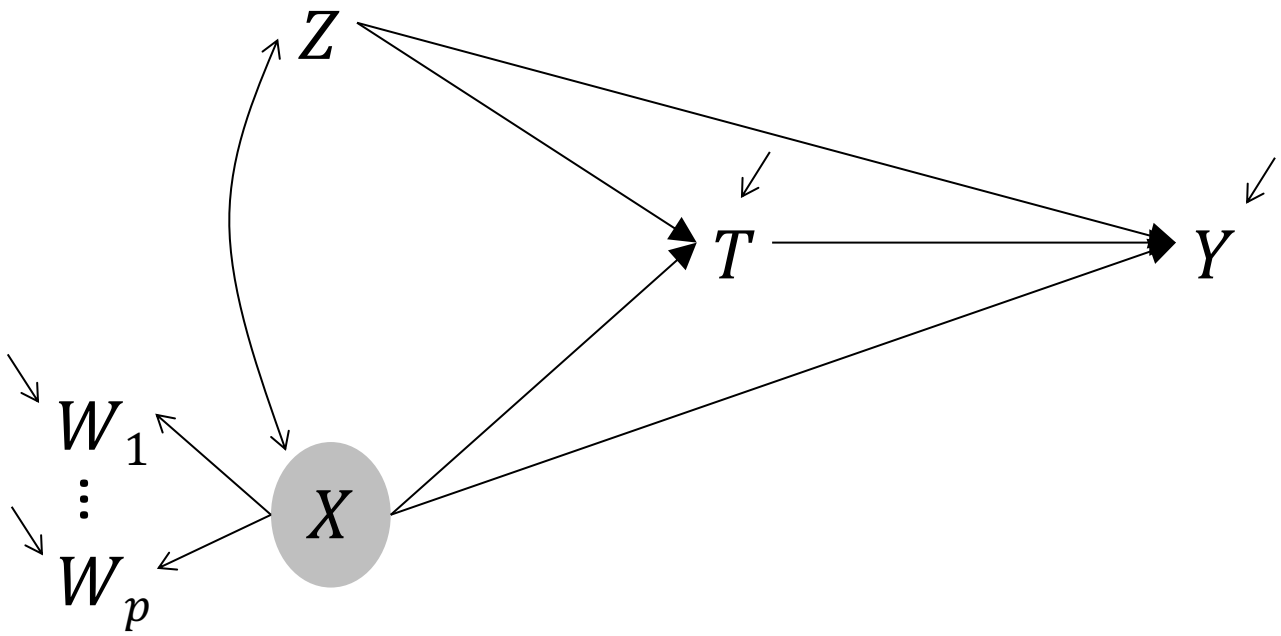
  * *fully inclusive* probit model using WLSMV requires a modification

We used Mplus 7.2.

But most methods can be implemented in other popular statistical packages.

Simulation studies

Data generating mechanism



Classical measurement error: non-differential w.r.t. X, Z, T, Y and independent across the measurements

Each scenario: 1000 simulated n=1000 datasets

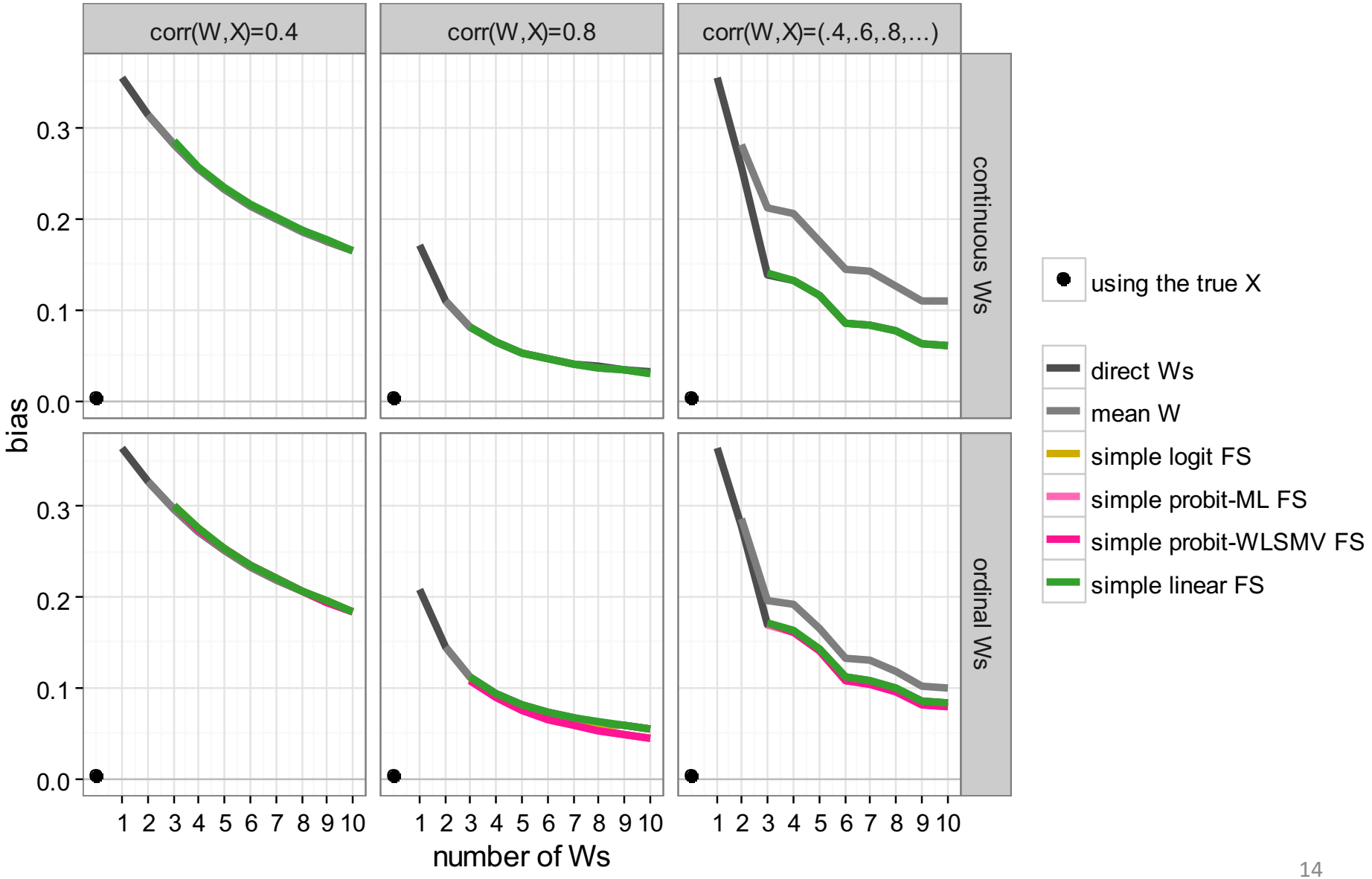
Simulation studies

	Data generating model	Parameter values
Z and X	$\begin{pmatrix} Z \\ X \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$	$\rho = 0, .4, -.4$
T	either logit: $T \sim \text{Binomial}\left(\frac{e^{\beta_0 + \beta_z Z + \beta_x X}}{1 + e^{\beta_0 + \beta_z Z + \beta_x X}} \right)$ or probit: $T^* \sim \text{Normal}(\beta_z Z + \beta_x X, 1.7^2), T = 1 \cdot (T^* > \tau_T),$ where β_0 and τ_T are set so that $\text{P}(T = 1) = p_T$.	$\beta_z = .5,1$ $\beta_x = .5,1$ $p_T = .5, .4, .3, .2$
Y	$Y \sim \text{Normal}\big((Z + \gamma X + 0T), 4\big).$	$\gamma = 1,2$

Simulation studies

	Data generating model	Parameter values
continuous $\boldsymbol{W}$	$\boldsymbol{W} \sim \text{Normal}(X, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is diagonal matrix with variance elements based on $\boldsymbol{\phi} = \text{cor}(\boldsymbol{W}, X)$.	$p = 2, \ldots, 10$ 3 cases of uniform W-X correlations: $\boldsymbol{\phi} = .\boldsymbol{4}$ (*low*), $.\boldsymbol{6}$ (*medium*), $.\boldsymbol{8}$ (*high*)
ordinal $\boldsymbol{W}$	$\boldsymbol{W}^* \sim \text{Normal}(X, \boldsymbol{\Sigma})$ as above then categorized into $\boldsymbol{W}$ with four equal mass categories	4 cases of mixed W-X correlations: $\boldsymbol{\phi} = (.4, .6, .4, .6, \ldots)$ (*lome*), $(.4, .8, .4, .8, \ldots)$ (*lohi*), $(.6, .8, .6, .8, \ldots)$ (*mehi*), $(.4, .6, .8, .4, .6, .8, \ldots)$ (*lomehi*)

Simple factor score, direct Ws, mean W



alpha=.27-.67; .79-.95; .40-.82

Bias

The *simple* FS method is similarly biased as the direct Ws method.

The mean W method is more biased than the simple FS and direct Ws methods when W-X correlations are not uniform.

All *simple* FSs perform similarly, except when Ws are ordinal and highly correlated with X – then logit/probit FSs are slightly less biased.

The more measurements and higher W-X correlations, the less bias.

The larger the effects of X on T and Y, the more bias.

Variance

All methods have underestimation of variance when treatment prevalence is far from .5. This problem gets worse for the direct Ws method with increasing number of measurements.

(relevant to high alpha situation when direct Ws looks appealing from a bias perspective)

Why *simple* FS not better than direct Ws?

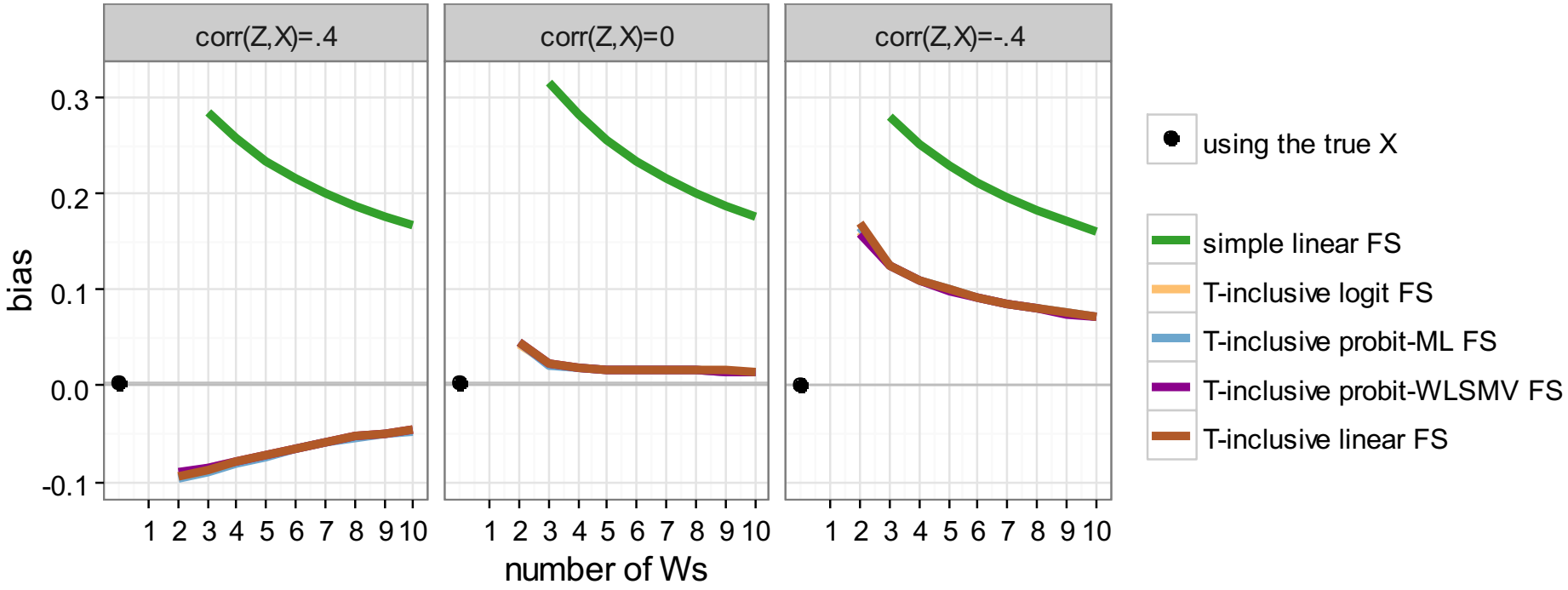
With continuous Ws, the two methods capture the same information about X based on the correlations among the Ws,

correlation of *simple* FS with X = multi-correlation of Ws with X.

The simple factor score does not use any additional information, such as information about the X-T association, or information about the X-Z-T joint distribution,

correlations of *simple* FS with T and Z **<** correlations of X with T and Z.

Partially inclusive vs. *simple* factor scores



Better, but still biased, especially when X and Z are correlated.

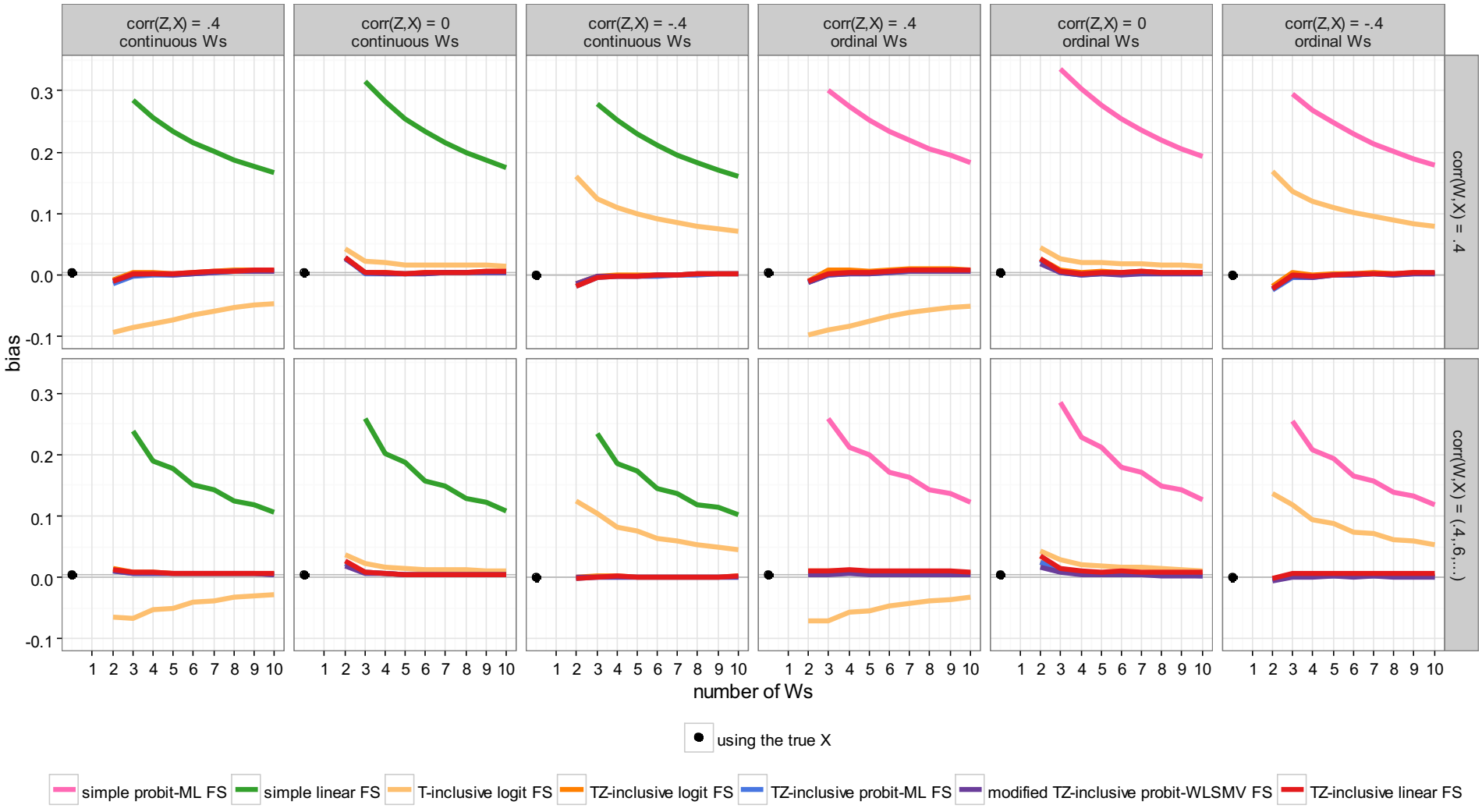
The sign of bias is a function of the sign of the X-Z correlation.

Why doesn't the *partially inclusive* FS method work?

Model imcompatibility:

- the factor score model (imputation model) does not include Z
- the propensity score model (analysis model) includes Z

Fully inclusive vs. *partially inclusive* & *simple* FSs



Fully inclusive factor scores effectively remove bias!

Why does the *fully inclusive* FS method work?

Moving from the *simple* FS to the *fully inclusive* FS

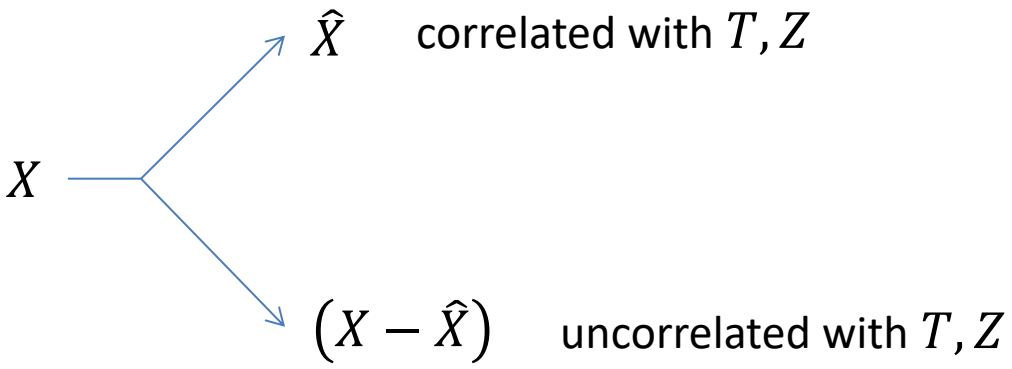
takes us only a tiny bit closer to the true X (the correlation of the FS with X increases very slightly),

but brings us substantially closer to T and Z (the *fully inclusive* FS is more correlated with T and Z than X is).

The *fully inclusive* FS approximates the predicted value of X from a regression model using the true X

$$\hat{\eta}\left|\boldsymbol{W},T,Z \longrightarrow \hat{X}\right|\boldsymbol{W},T,Z$$

and this predicted value of X is sufficient to remove confounding.



The *fully inclusive* FS method: other findings

Bias – differentiating the *fully inclusive* factor scores

When treatment prevalence is far from .5, the FS consistent with the true treatment assignment model performs slightly better. This is more pronounced when the effect of X on T increases. However, the differences are small compared to the magnitude of bias of the *simple* FS method.

When Ws are ordinal and some Ws are highly correlated with X, the linear FS is slightly more biased than the non-linear ones.

Variance and coverage

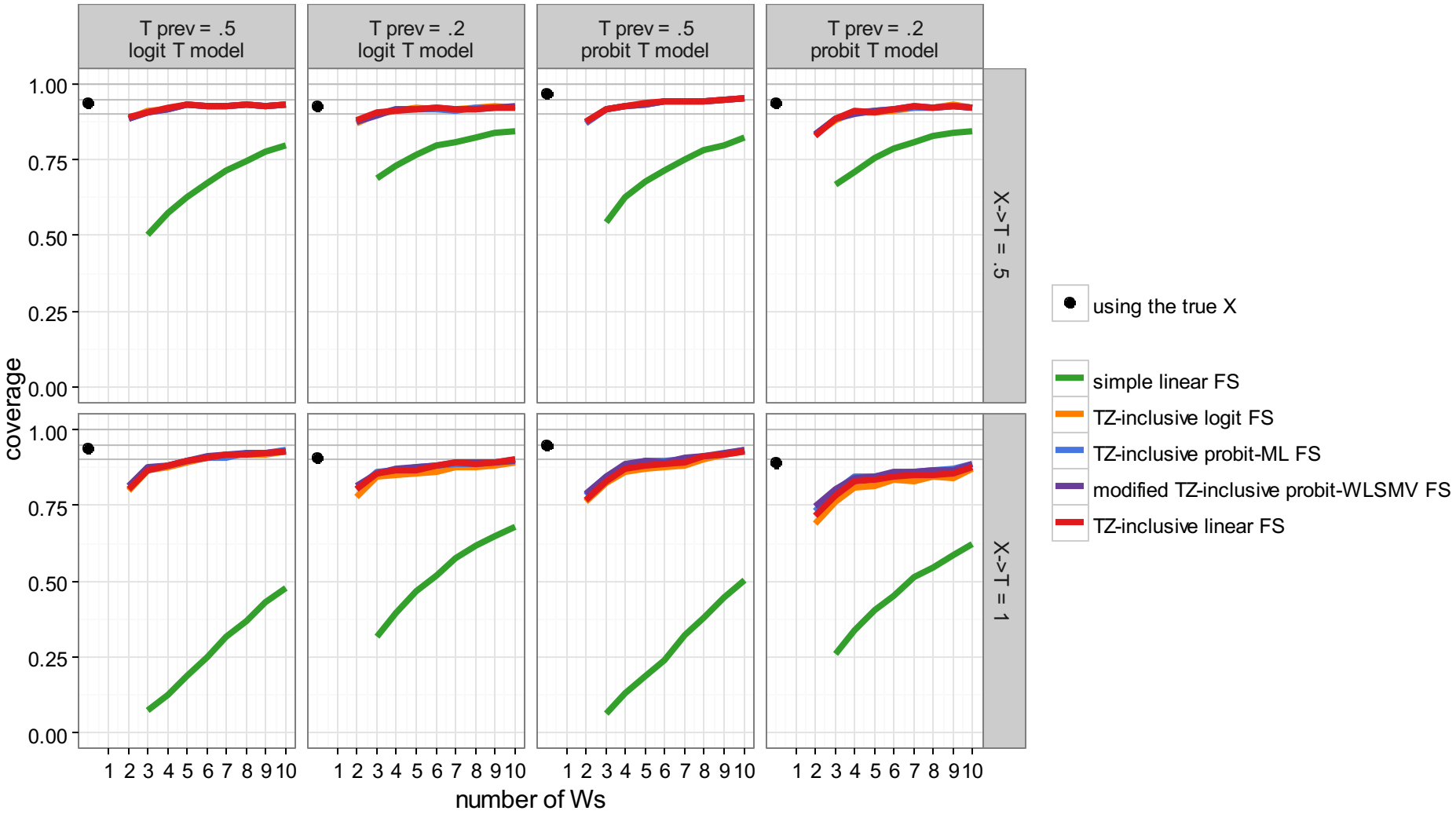
Variance is underestimated, more so when the number of Ws is small, and more so when treatment prevalence is far from .5.

Coverage worsens when the effect of X on T increases.

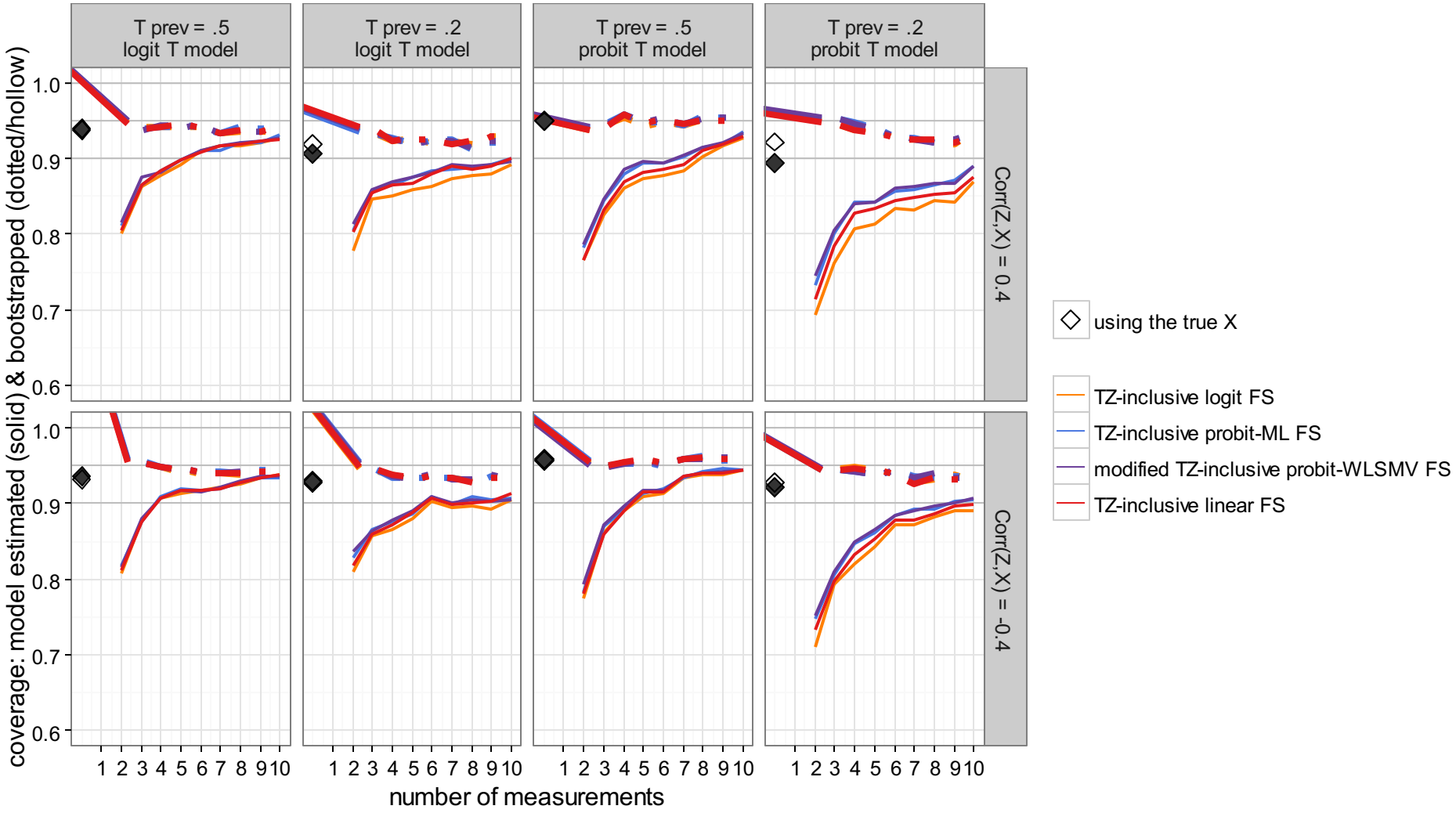
Coverage is worse when treatment prevalence is far from .5 (also when using X).

In cases where coverage is poor, bootstrapping substantially improves coverage.

Coverage by outcome-model-estimated CI: *fully inclusive* vs. *simple* factor scores



Coverage by bootstrap interval (dotted/hollow) and by outcome-model-estimated CIs (solid)



Simulation conclusions

A simple fix for bias due to a latent covariate in propensity score weighting analysis is the *fully inclusive* factor score method.

Low coverage may be an issue that results from several factors: large effect of latent variable on treatment assignment, treatment prevalence far from .5, and few measurements.

Bootstrapping improves coverage.

Illustrative example

Question of interest: Does out-of-school suspension increase the risk of problems with the law in young adulthood?

Data: The National Longitudinal Study of Adolescent to Adult Health (Add Health)

- representative sample of US adolescents
- 1994-95 (w1, grades 7-12), 1996 (w2), 2001-02 (w3), 2008-09 (w4, ages 24-32)
- public use data, males only

2 sub-samples: using complete data

- those who had never been suspended by wave 1: n=961, estimate ATT
- those who had been suspended prior to wave 1: n=468, estimate ATE

Exposure: suspended between waves 1 and 2: 6.0% and 32.9%

Outcome: arrested by wave 4: 30.9% and 60.7%

Covariates measured at wave 1:

- age, race/ethnicity, parent education, parent marital status
- delinquent behavior, academic achievement

Factor analysis: 3 factors:

- general delinquent (D1, 8 items) and violent delinquent (D2, 4 items)
- academic achievement (AA, 4 items)

suspended-before sub-sample – ATE results

Adjustment, ignoring measurement error

		OR (95% CI)	RR	RD	p1	p0
D1, D2, AA	unadjusted	2.010 (1.252,3.227)	1.256	0.148	0.725	0.577
AA	adjusted	1.681 (1.046,2.700)	1.183	0.109	0.704	0.595
D1 & D2	adjusted	1.942 (1.213,3.110)	1.235	0.138	0.724	0.586
D1, D2 & AA	adjusted	1.629 (1.020,2.602)	1.166	0.100	0.699	0.600

Correction for measurement error (all three adjusted)

		OR (95% CI)	RR	RD	p1	p0
none	corrected	1.629 (1.020,2.602)	1.166	0.100	0.699	0.600
AA	corrected	1.423 (0.897,2.257)	1.116	0.071	0.682	0.611
D1 & D2	corrected	1.532 (0.966,2.430)	1.142	0.086	0.693	0.607
D1, D2 & AA	corrected	1.364 (0.869,2.139)	1.100	0.062	0.678	0.616

never-suspended-before sub-sample – ATT results

Adjustment, ignoring measurement error

		OR (95% CI)	RR	RD	p1	p0
D1, D2, AA	unadjusted	2.552 (1.400,4.651)	1.683	0.200	0.493	0.293
AA	adjusted	2.344 (1.232,4.459)	1.555	0.176	0.493	0.317
D1 & D2	adjusted	2.022 (1.034,3.953)	1.424	0.147	0.493	0.346
D1, D2 & AA	adjusted	1.864 (0.916,3.794)	1.341	0.125	0.493	0.368

Correction for measurement error (all three adjusted)

		OR (95% CI)	RR	RD	p1	p0
none	corrected	1.864 (0.916,3.794)	1.341	0.125	0.493	0.368
AA	corrected	1.763 (0.878,3.540)	1.304	0.115	0.493	0.378
D1 & D2	corrected	1.604 (0.768,3.347)	1.237	0.095	0.493	0.399
D1, D2 & AA	corrected	1.550 (0.757,3.175)	1.218	0.088	0.493	0.405

Thoughts about application

When is this kind of correction useful?

- When the measurement error results in significant confounding, i.e., the latent covariate is an important confounder AND measurement error is large

  -> less useful with scales with good psychometric properties (high reliability), more useful when piecing together a measure using ad hoc items

  -> more useful with multiple latent confounders

  -> maybe more useful when the latent covariate is a higher order factor (to be investigated)

- When the estimated treatment effect is borderline significant

Also, to help adjust for confounding due to measurement error

- it may help to include covariates that are correlated w/ the latent covariate (to be investigated) -- but beware of bad-behaving triangles

Extensions

- Investigate/extend to more complicated error structures
  - e.g., correlated errors, differential error, non-normal error
- Simulation with more complicated treatment assignment models
- Compare the factor score approach to the full latent modeling approach
- Investigate performance with propensity score matching

Acknowledgements

The Drug Dependence Epidemiology Training Program,
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

Our measurement error research team,
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health

NIH funding:

NIDA grant T32DA007292 (Debra Furr-Holden and Renee M. Johnson)

NIMH grant R01MH099010 (Elizabeth Stuart)

Thank you!

tnguye28@jhu.edu