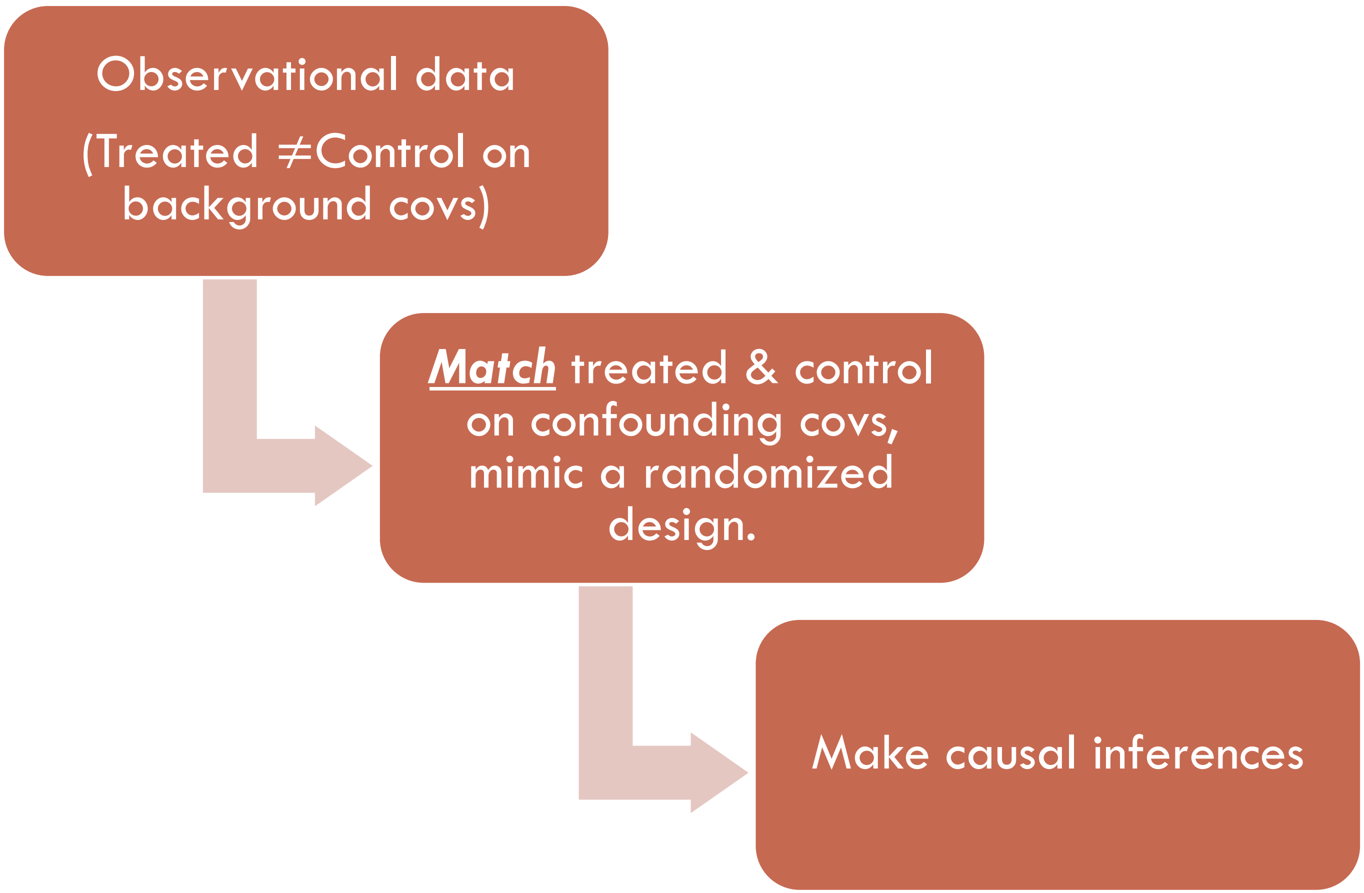
THE USE OF NONPARAMETRIC PROPENSITY SCORE ESTIMATION WITH DATA OBTAINED USING A COMPLEX SAMPLING DESIGN

Ji An & Laura M. Stapleton
University of Maryland, College Park

May, 2016

WHAT DOES A PROPENSITY SCORE METHOD DO?



WHAT IS A PROPENSITY SCORE?

- The conditional probability of a participant to be assigned to the treatment condition (Rosenbaum & Rubin, 1983)

$$\text{logit}(\pi_{\text{TREATMENT i}}) = \beta_0 + \sum_{p=1}^{P} \beta_p X_i$$

A GENERAL PS METHOD PROCESS

1
- Identification of critical covariates

2
Estimation of the propensity scores

3
- Conditioning (matching)

4
- Balance check

5
Treatment effect estimation

- Matching
- Subclassification
- Weighting
  - Inverse probability of treatment weighting (IPTW)
    $w_i = \frac{T_i}{\hat{e}_i} + \frac{1-T_i}{1-\hat{e}_i}$
  - Weighting by the odds
    $w_i = T_i + (1 - T_i)\frac{\hat{e}_i}{1-\hat{e}_i}$

INTRODUCTION: PRIMARY ISSUE

- The traditional PS method works well in SRS settings

- However, in reality… complex sampling (CS) design
  - Stage 1: The country → regions (strata)

    Select schools (PSUs)
  - Stage 2: School → demographic groups (strata)

    Sample students

  … … Disproportionate selection probabilities… …

- Consequence of ignoring the CS design
  - Bias in standard error estimates
  - Bias in parameter estimates

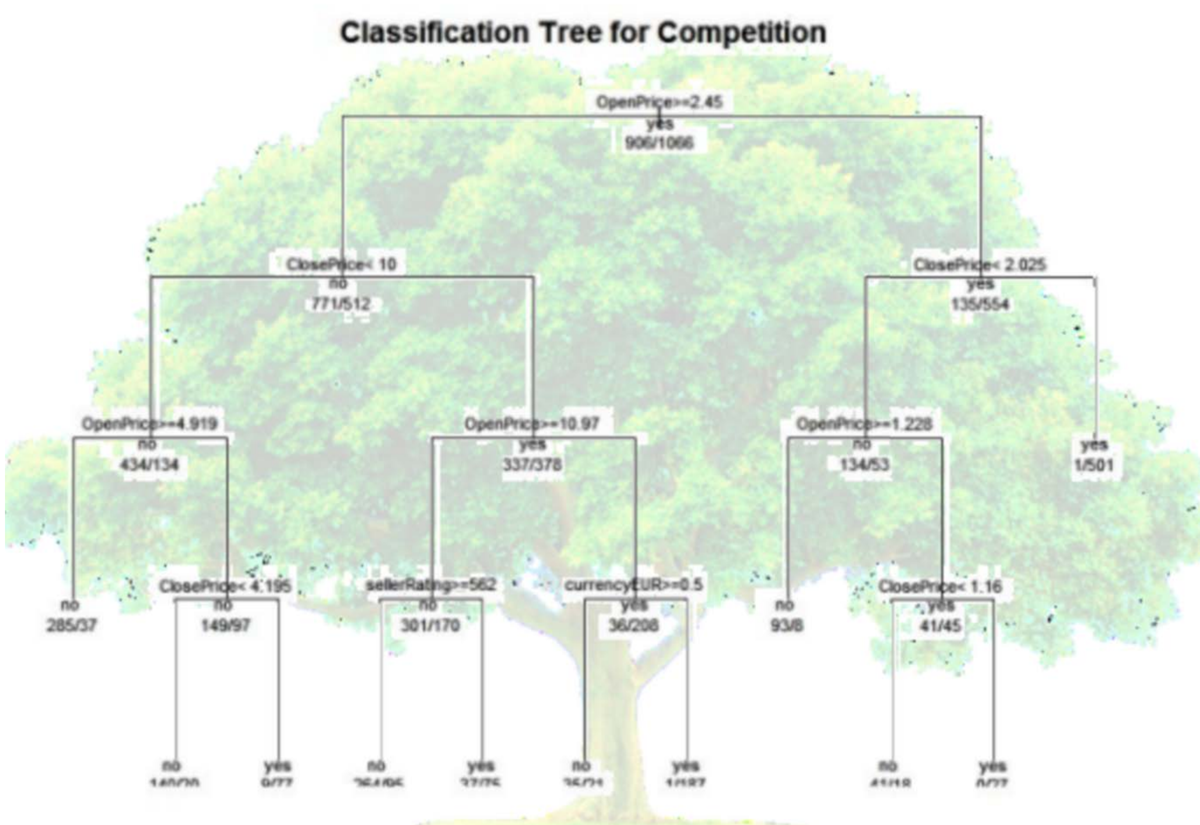
  Problematic *generalizability* to the population.

INTRODUCTION: PS ESTIMATION WITH CS DATA

- ***Model-based method***
  - Multilevel model
  - Fixed effects model (Thoemmes & West, 2011)
- ***Design-based method***
  - Sampling weighted regression
  - Incorporating sampling weight as a covariate

- ***Nonparametric methods*** (McCaffrey et al., 2004)
  - Classification and regression trees (CART)
  - Random forests
  - Boosted regression trees
  - Etc.

INTRODUCTION: NONPARAMETRIC METHODS



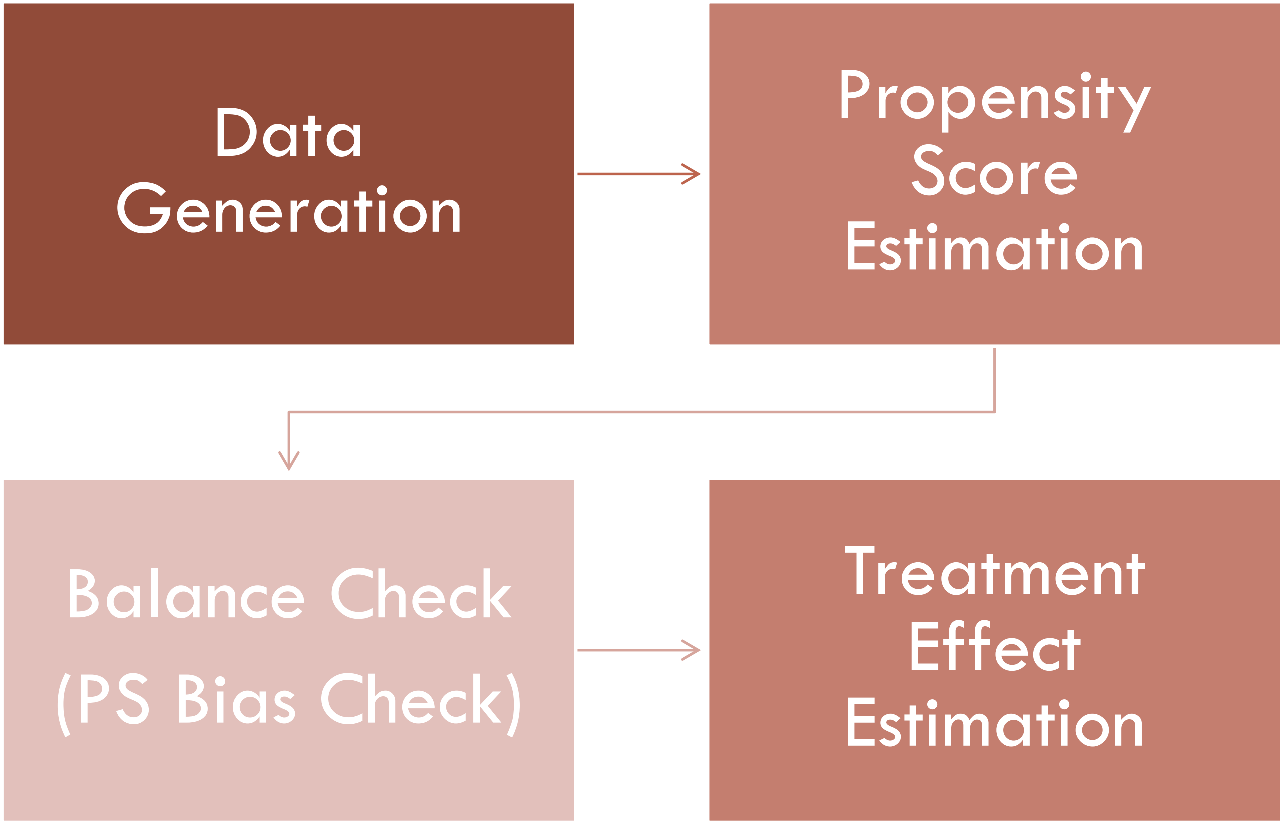
Picture from LinkedIn
by Jeffrey Strickland

OUR GOAL

- ***Do*** nonparametric PS methods outperform the other model-based or design-based methods?
  - Precision of PS estimates?
  - Quality of TE estimates?

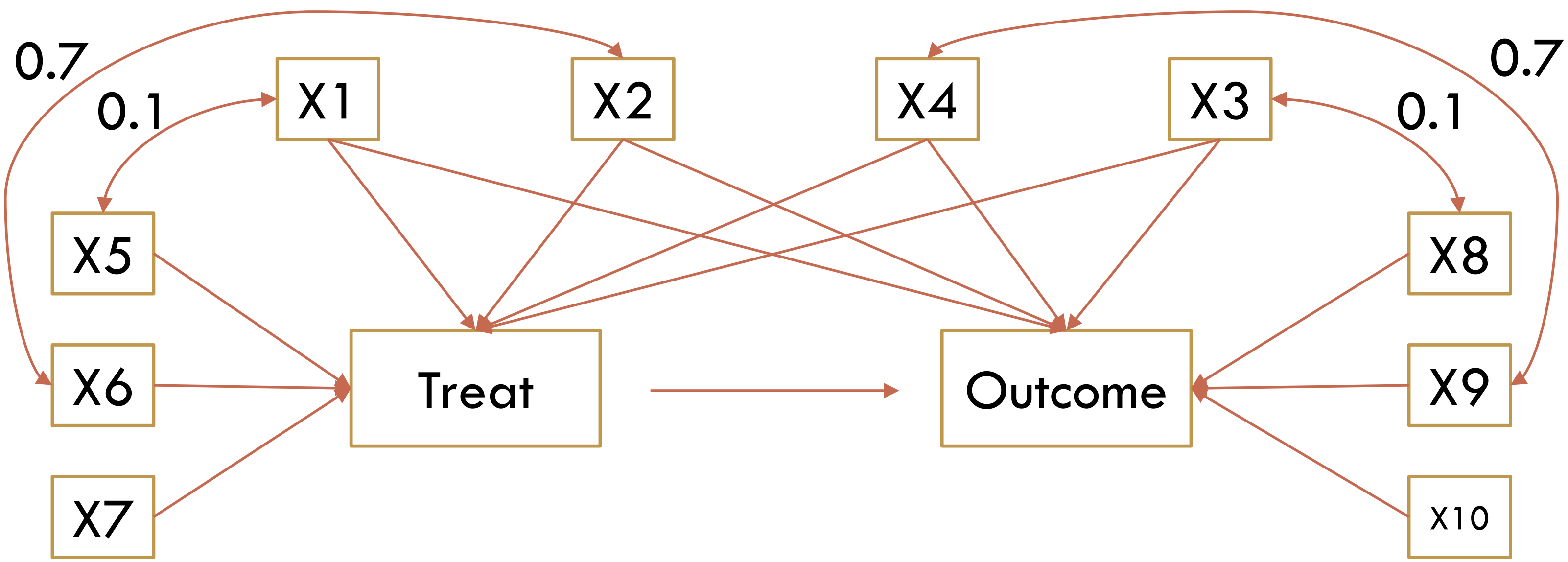
- ***What*** is the best way to accommodate CS design in the PS analyses?

METHODS


Data Generation
Propensity Score Estimation
Balance Check (PS Bias Check)
Treatment Effect Estimation

COVARIATES

Data Generation



Dummy: X1, X3, X5, X6, X8, X9, Treat; others: continuous
(Setoguchi et al., 2008; Lee et al., 2010)

POPULATION

Data Generation

- About 75, 000 students in the finite population

- 50 counties
  30 schools per county **(private & public)**
  ave. 50 students per school **(ELL & non-ELL)**

- ICC around 0.25 (Hedges & Hedberg, 2007)

- Pop1: main effects only (additivity and linearity)
  $$logit(e|Z = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7$$

- Pop7: moderate non-additivity and non-linearity with 3 quadratic terms &10 interactions (Setoguchi et al, 2008; Lee et al., 2010)

SAMPLE

Data Generation

Two-Stage Sampling:

	Stage 1 (Select schools within each county)		Stage 2 (Select students within each school)	
	Private	**Public**	**ELL**	**Non-ELL**
Pop	33%	67%	25%	75%
Sel rate	50%	25%	≤50%	25%

About 9000-10000 students in each sample

100 replications

PS MODELS

Propensity Score Estimation

7 PS models (5 parametric, 2 nonparametric)

- M1: ***SL*** on the baseline covariates.
- M2: ***SL*** on the baseline covariates **+ *the survey weight.***
- M3: ***SL weighted by the survey weight.***
- M4: ***Fixed effects model.***
- M5: ***ML*** with random intercepts.

- M6: ***Random forests.***
- M7: ***Boosted regression trees.***

BALANCE CHECK

Balance Check

(PS Bias Check)

Accuracy of PS

- Absolute bias

Balance

- SMD = $\frac{\lfloor \bar{X}_{pT} - \bar{X}_{pC} \rfloor}{\sigma_{pT}}$
- Balance weighted by ***IPTW*** (consistent with the ***PS-adjusted*** TE)
- Balance weighted by ***IPTW*SAMPWT*** (consistent with the ***PS&CS-adjusted*** TE)

TE MODELS

Treatment Effect Estimation

IPTW implemented to achieve ***ATE***

TE Models

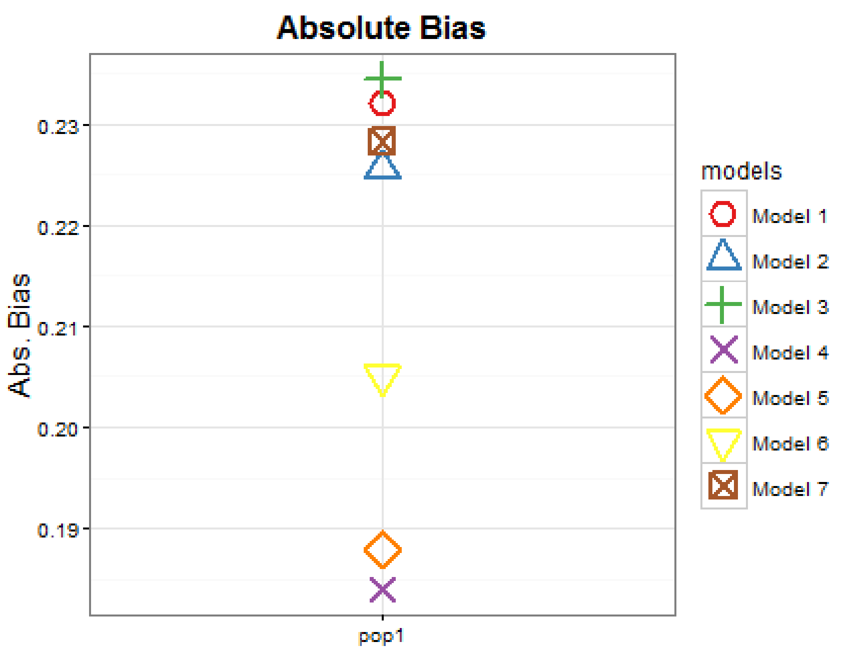
- Naïve (no adjustment at all)
- CS adj. (weighted by ***SAMPWT***)
- PS adj. (weighted by ***IPTW***) (via 7 PS models)
- PS & CS adj. (weighted by ***IPTW*SAMPWT***) (via 7 PS models)

$$\hat{Y} \sim Treat$$

(Absence of a fully specified TE model)

RESULTS: ABSOLUTE BIAS

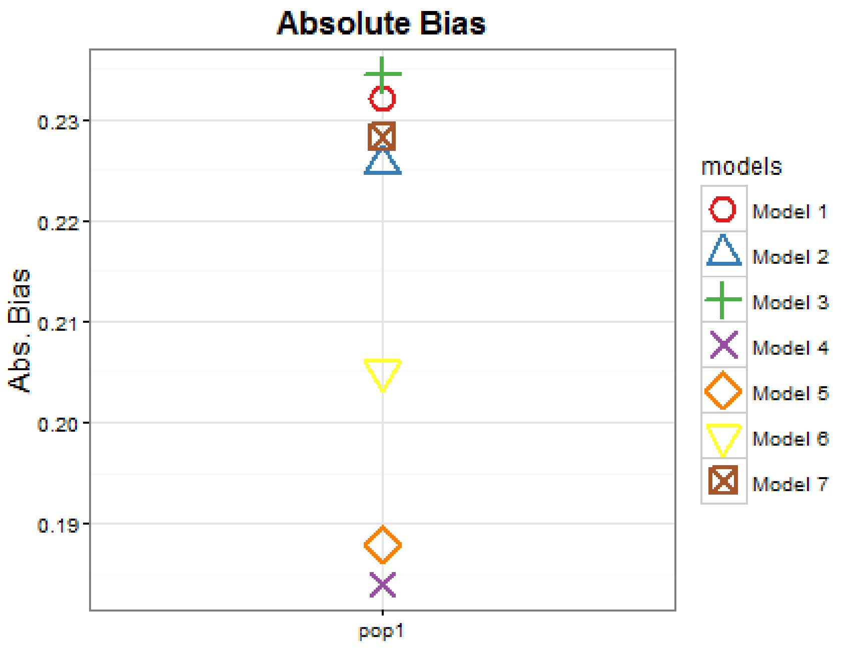
- Models 4 and 5 (**<u>the fixed effects and multilevel models</u>**) have the best performance in PS accuracy



RESULTS: ABSOLUTE BIAS

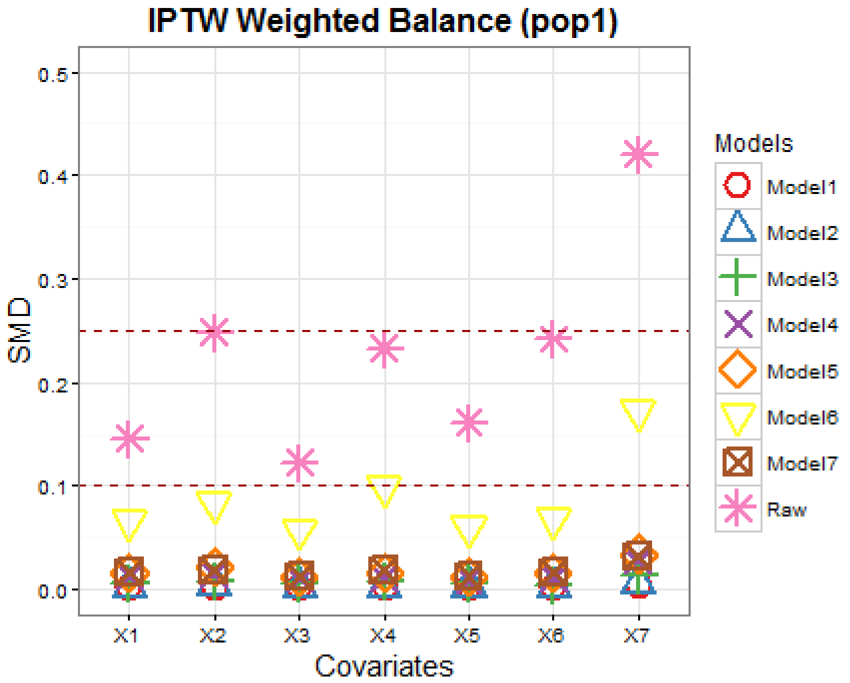
- Model 6 (**random forests**) did not outperform Models 4 and 5 but is better than the others.

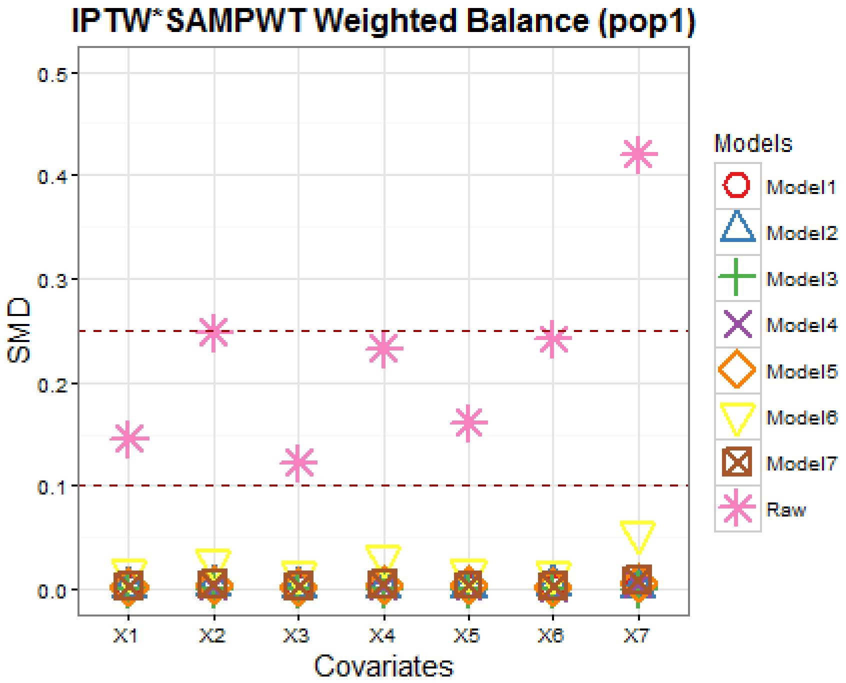
(However, when we get to a complex PS model they do!)



RESULTS: BALANCE

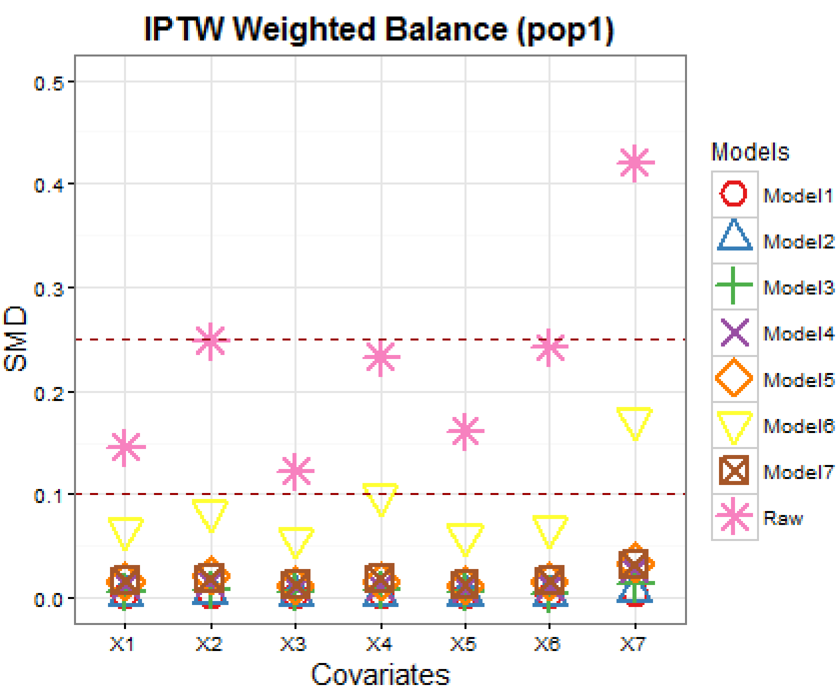
- All PS models achieved very good covariate balance.

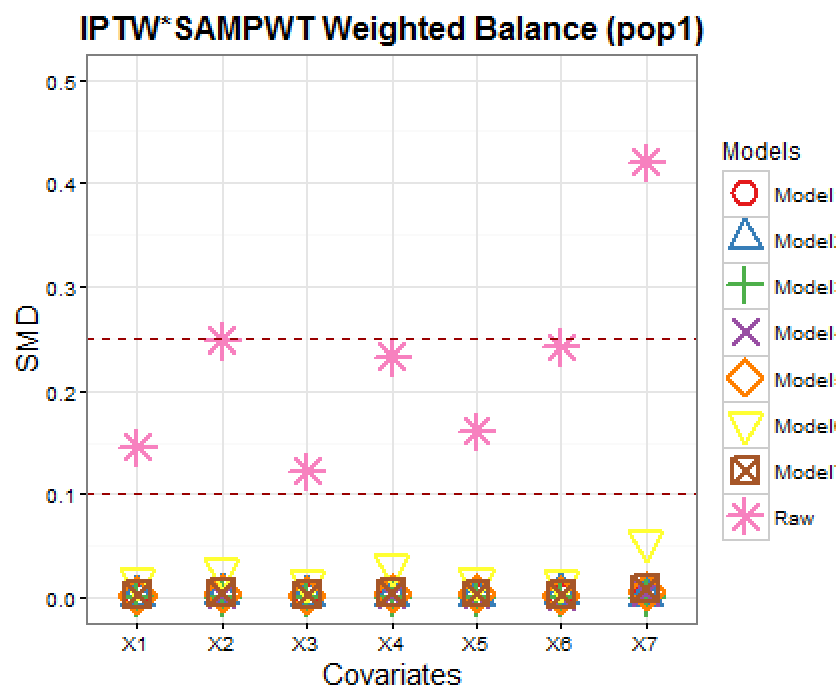




RESULTS: BALANCE

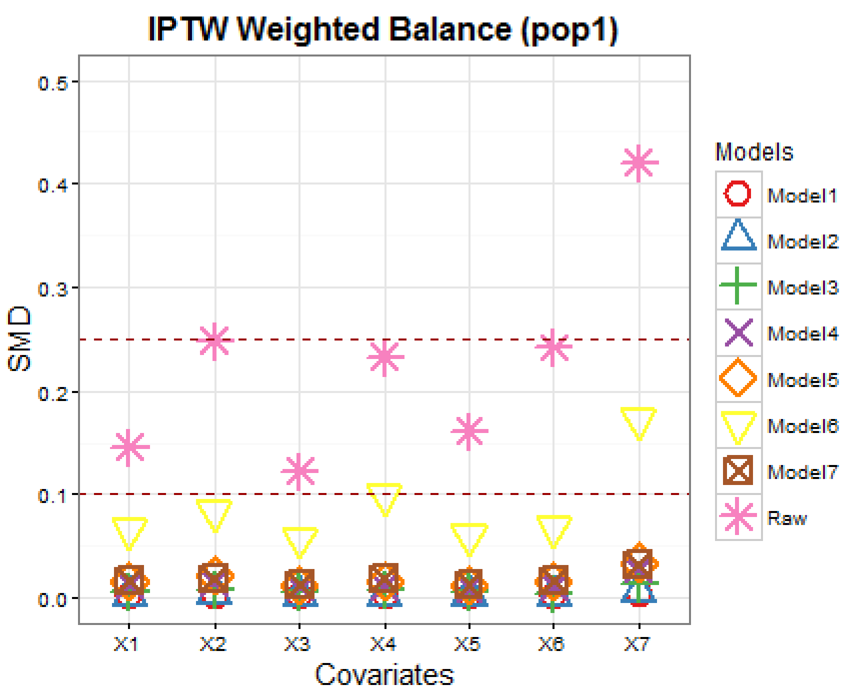
- Combining CS and PS adjustment (IPTW*SAMPWT) produced better balance than using PS adjustment only (IPTW).

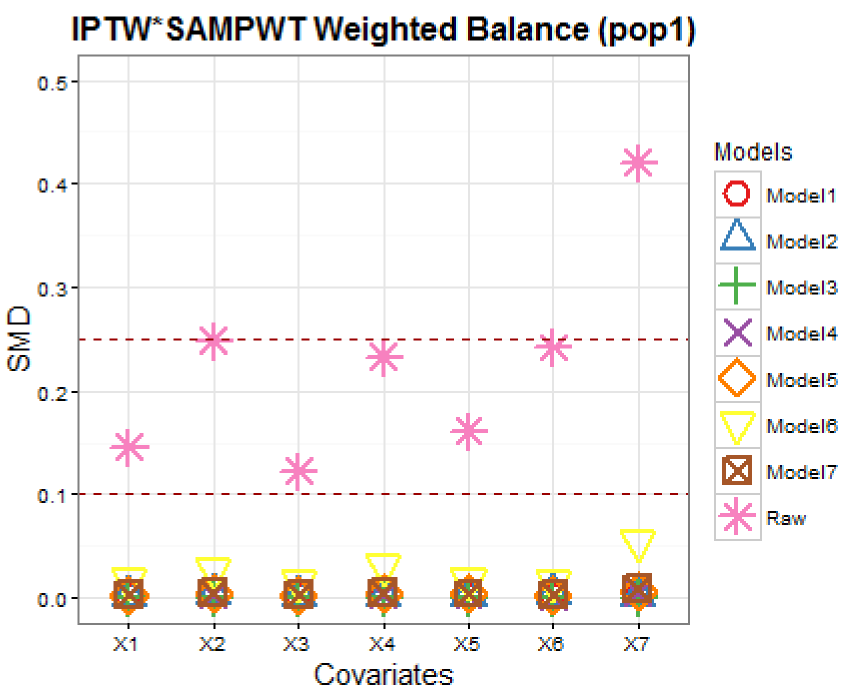




RESULTS: BALANCE

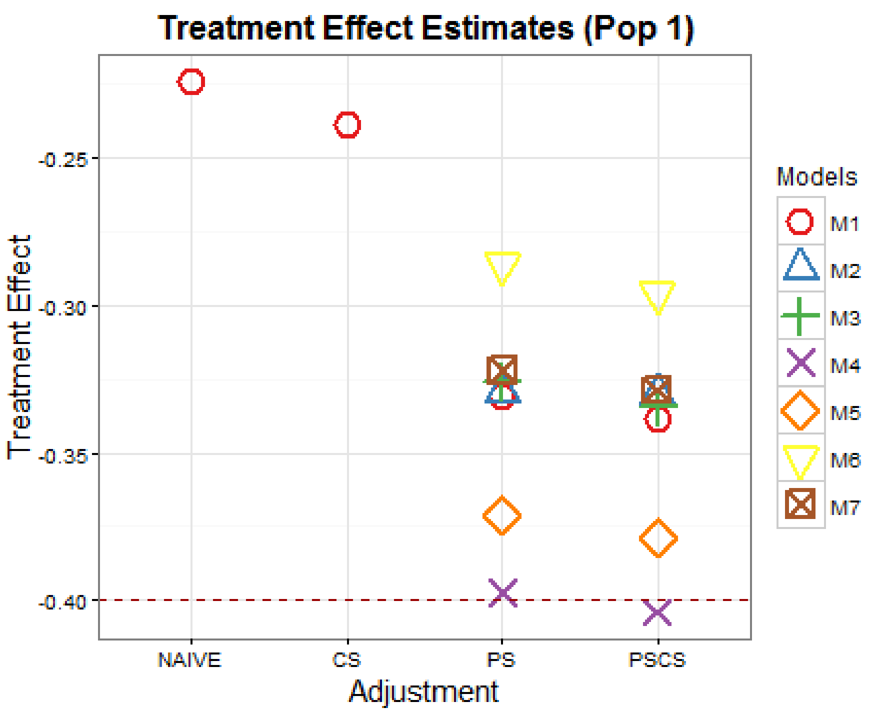
- Random forests yielded worse balance than the other models, yet still good.





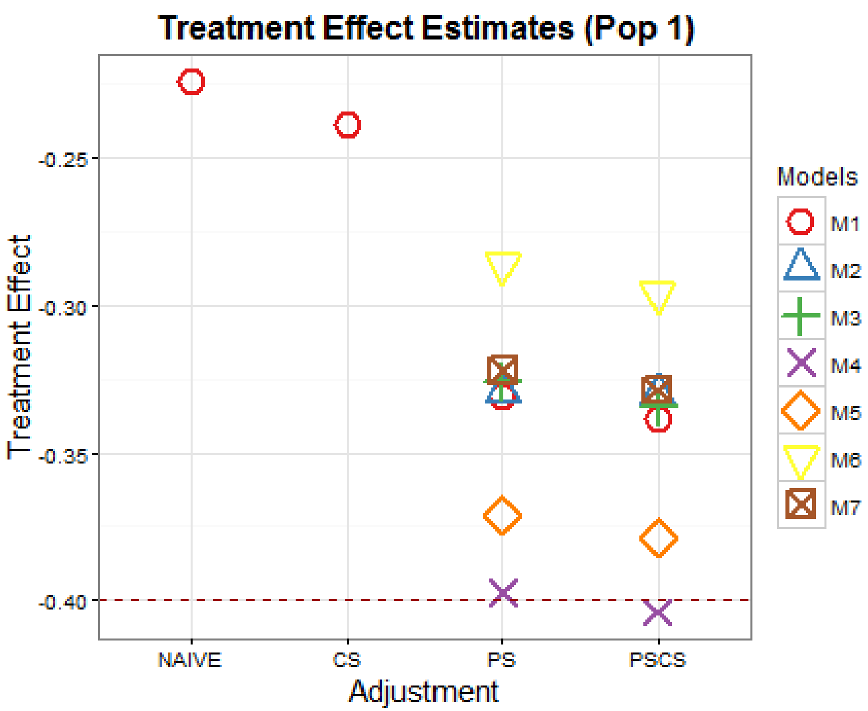
RESULTS: TREATMENT EFFECT

- Models 4 and 5 had the best performance for estimating (the PS and therefore) the TE in the absence of a fully specified TE model.



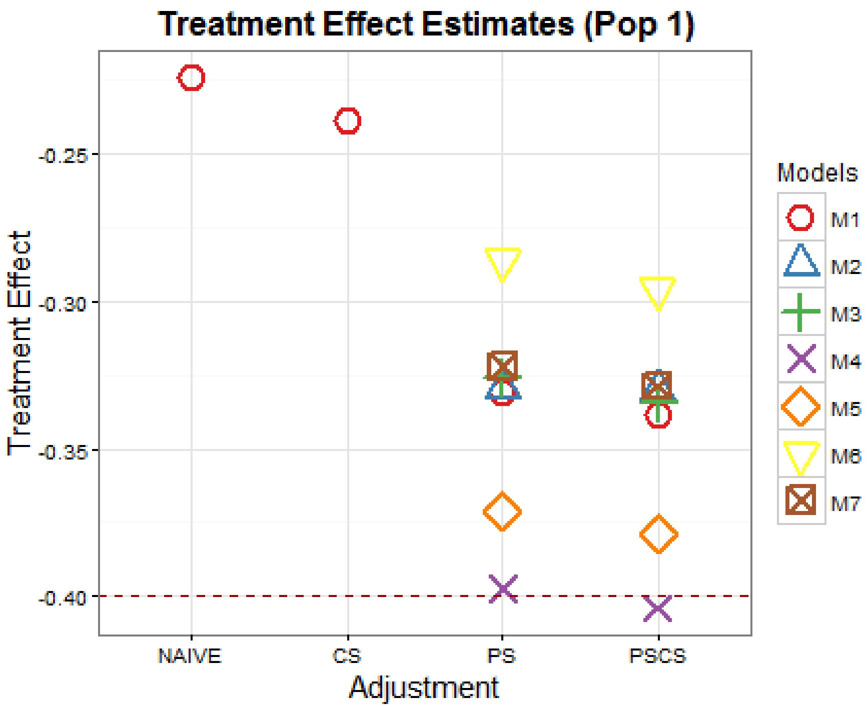
RESULTS: TREATMENT EFFECT

- The nonparametric methods **did NOT outperform** Models 4 and 5 when the PS model is correctly specified.



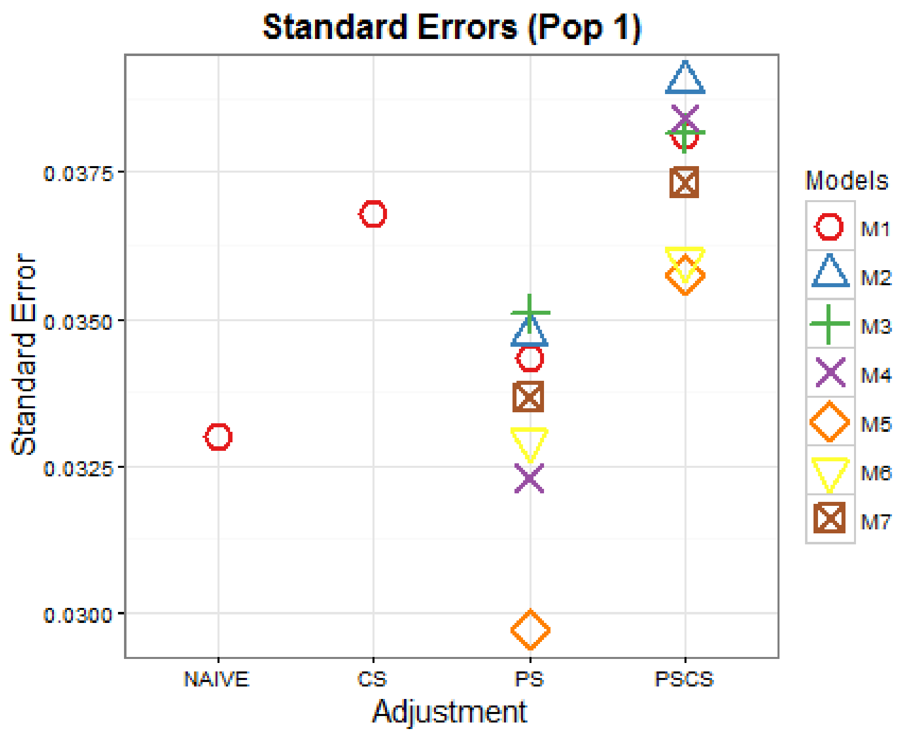
RESULTS: TREATMENT EFFECT

- Adjustment for CS does make a difference in the *accuracy of TE* (although in this simulation it's relatively small)! ☺



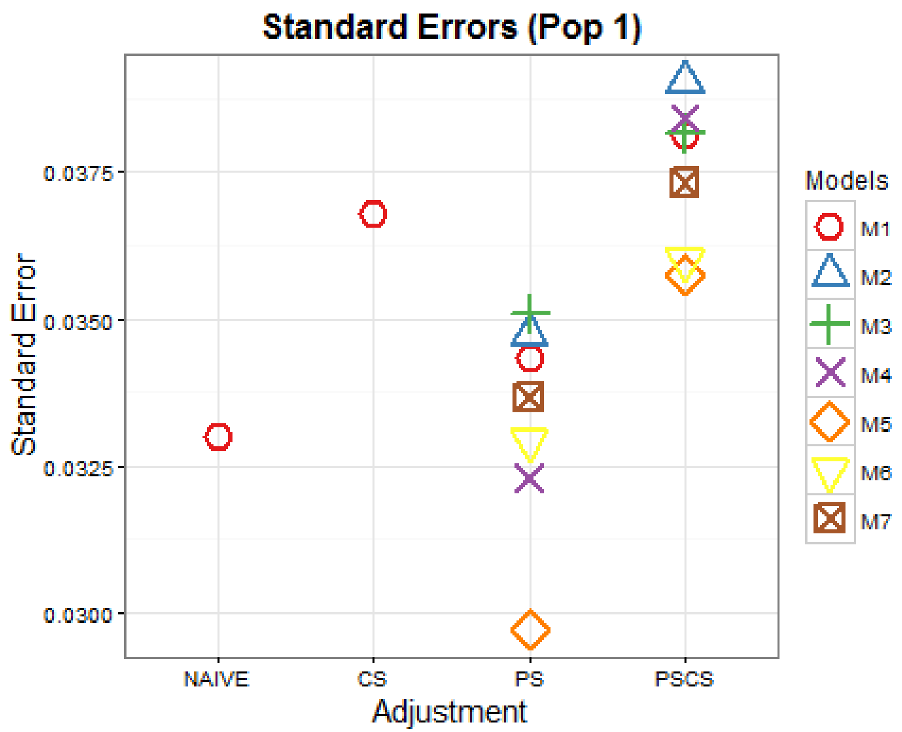
RESULTS: TREATMENT EFFECT (SE)

- Adjustment for CS does make a difference in the *precision of TE*! ☹



RESULTS: TREATMENT EFFECT (SE)

- Good news: the nonparametric methods ranked better in terms of the *precision of TE*! ☺



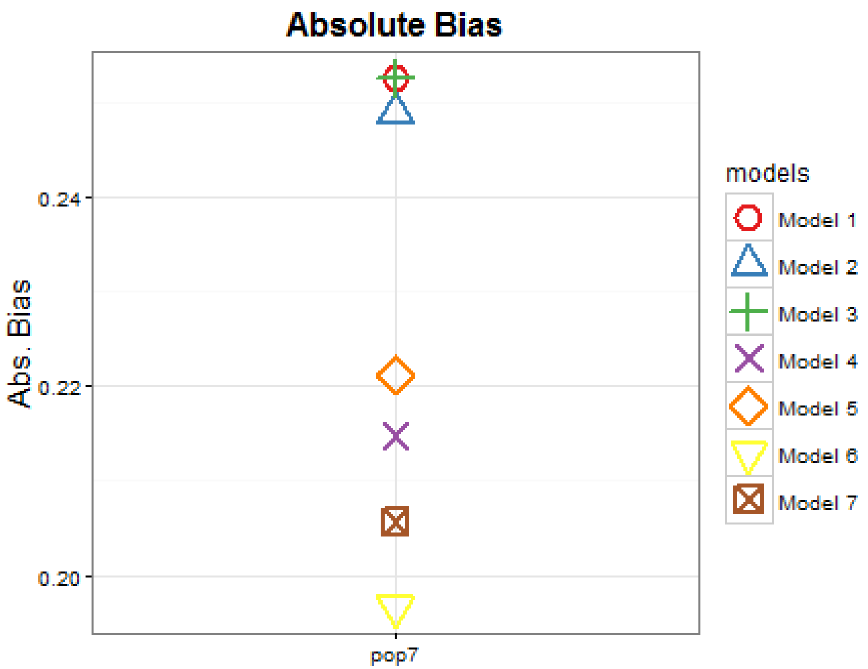
CONCLUSION (1)

- ***Do*** nonparametric PS methods outperform the other model-based or design-based methods?

    No.

    However… when the PS model is unknown and thus is misspecified by parametric models…

(POP7)



CONCLUSION (2)

- ***What*** is the best way to accommodate CS design in the PS analyses?

PS Estimation	Conditioning	Effect	TE Estimation
M1: SL	IPTW	ATE	Naïve (no adjustment)
M2: SL+SAMPWT			CS adj. (***SAMPWT***)
M3: SL(SAMPWT)			PS adj. (***IPTW***)
M4: Fixed effect			PS & CS adj. (***IPTW*SAMPWT***)
M5: Multilevel			
M6: Random forests			
M7: Boosted regression			

CONCLUSION (2)

- ***What*** is the best way to accommodate CS design in the PS analyses?

PS Estimation	Conditioning	Effect	TE Estimation
M1: SL	IPTW	ATE	Naïve (no adjustment)
M2: SL+SAMPWT			CS adj. (***SAMPWT***)
M3: SL(SAMPWT)			PS adj. (***IPTW***)
M4: Fixed effect			PS & CS adj. (***IPTW*SAMPWT***)
M5: Multilevel			
M6: Random forests			
M7: Boosted regression			

FUTURE RESEARCH

- Other matching methods…

PS Estimation	Conditioning	Effect	TE Estimation
M1: SL	IPTW	ATE	Naïve (no adjustment)
M2: SL+SAMPWT	Matching	ATT	CS adj. (***SAMPWT***)
M3: SL(SAMPWT)	Subclassification	ATE/ATT	PS adj. (***IPTW***)
M4: Fixed effect	WBO	ATT	PS & CS adj. (***IPTW*SAMPWT***)
M5: Multilevel			
M6: Random forests			
M7: Boosted regression			

FUTURE RESEARCH

- Other matching methods..
- Misspecified PS models…

PS Estimation	Conditioning	Effect	TE Estimation
M1: SL	IPTW	ATE	Naïve (no adjustment)
M2: SL+SAMPWT	Matching	ATT	CS adj. (***SAMPWT***)
M3: SL(SAMPWT)	Subclassification	ATE/ATT	PS adj. (***IPTW***)
M4: Fixed effect	WBO	ATT	PS & CS adj. (***IPTW*SAMPWT***)
M5: Multilevel			
M6: Random forests			
M7: Boosted regression			

Thank you!

jian12@umd.edu