Covariate Selection

Peter M. Steiner
University of Wisconsin–Madison

ACF/OPRE Methodological Advancement Meeting
Innovative Directions in Estimating Impact
Sep. 6-7, 2012

The research was supported by grants R305D100033 and R305D120005 from the Institute of Education Sciences, U.S. Department of Education.

Covariate Selection in Practice: When Should One Select Covariates?

Ex ante selection

- *Before* study is implemented and covariates are measured
- All covariates that are need to be measured in order to get an *ignorable* selection mechanism are determined in advance

Ex post selection

- *After* study was implemented
- Allows only for a selection among observed covariates
  → addresses only *overt bias* (due to observed covariates)
  → *hidden bias* (due to unobserved covariates) cannot be addressed

Covariate Selection Approaches Overview

For both ex ante and ex post covariate selection, decisions can be based on different *selection approaches*

- *Causal structural models* (*CSMs, DAGs*) derived from substantive theory (e.g., Pearl, 2009; Spirtes, Glymour & Scheines, 2000) → ex ante & ex post
- Select *covariates that are known to work in general*, i.e., successfully remove most of the bias → ex ante & ex post
- *Empirical tests* that rely on observed relations between covariates and the outcome (or treatment) → ex post
- *"Kitchen sink" approach*: consider all observed covariates → ex post

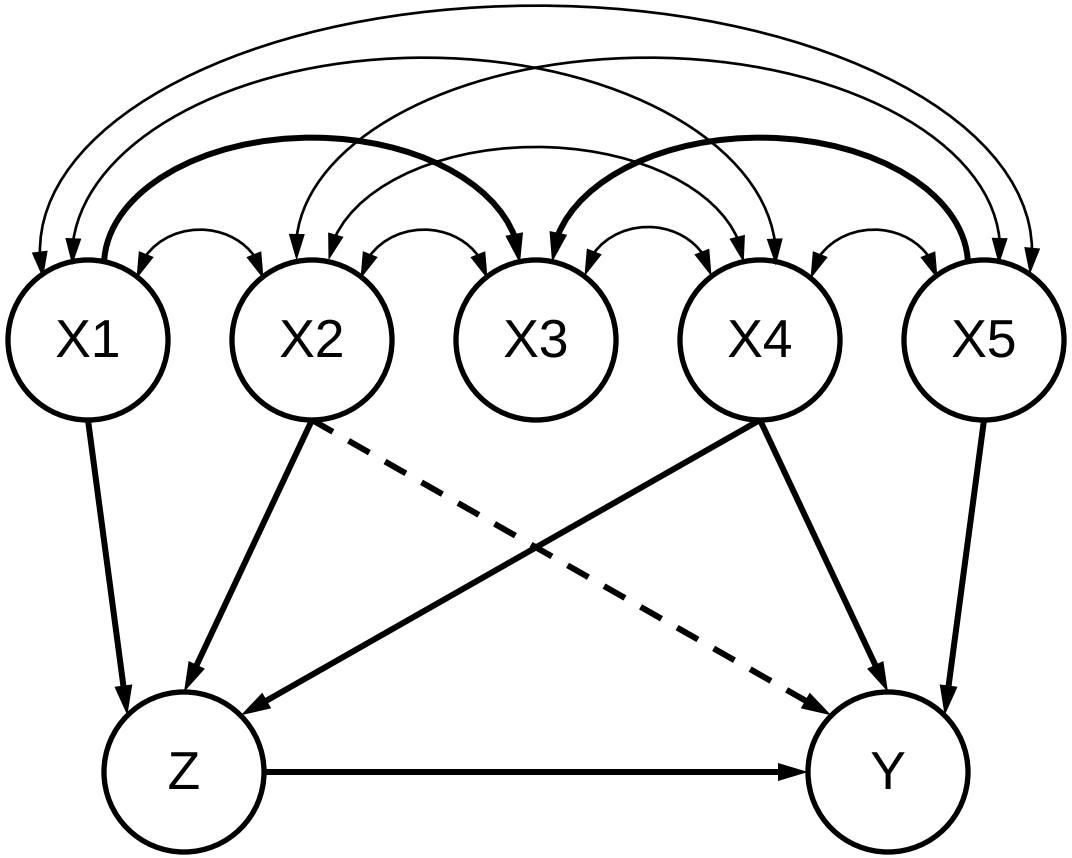
Causal Structural Model (CSM)

The formulation of a CSM requires *reliable domain-specific knowledge* of the true data-generating model (DGM)

- For each possible covariate, *whether it is related* to
  - > treatment selection (Z)
  - > outcome (Y)
  - > all other possible covariates (X)
- *Strength* and *direction* of relations (i.e., magnitude and sign of path coefficients)

CSM offers *clear guidelines* which covariates to select and which ones not to select (Pearl, 2009)

Causal Structural Model (CSM): Example



Y: outcome

Z: treatment

X1: instrument

X2: confounder: "near" instrument

X3: endogenous variable, unrelated to Z and Y

X4: confounder

X5: predictor of Y

Causal Structural Model (CSM)

If the CSM is known and all covariates are reliably measured, *covariate selection is straightforward*

Even in case of *unobserved covariates* the CSM tells us which covariates to include and which ones not

- Controlling for some confounding covariates might actually increase instead of reduce selection bias
- But need reliable knowledge about the strength and direction of relations;
  → but we rarely have such a detailed knowledge

Causal Structural Model (CSM)

Advantages

- CSMs provide *clear guidelines* for an ex ante selection of covariate
- Forces us to carefully *think about covariates* that might be potential confounders or bias-inducing or bias amplifying covariates

Disadvantages

- If the CSM does not correctly represent the true DGM, the resulting treatment effect is *biased* with respect to the underlying DGM
  → in practice we frequently *lack reliable knowledge* about the true DGM

Covariates That Work in General

Given the lack of strong substantive theories, are there *types of covariates* that, in general, remove at least most of the selection bias?

Two types of covariates:

- Covariates indexing the *selection process*
  → strongly related to selection
- *Pretests* and *proxy-pretest measures* of the outcome
  → strongly related to the outcome

Covariates That Work in General

Covariates indexing the selection process

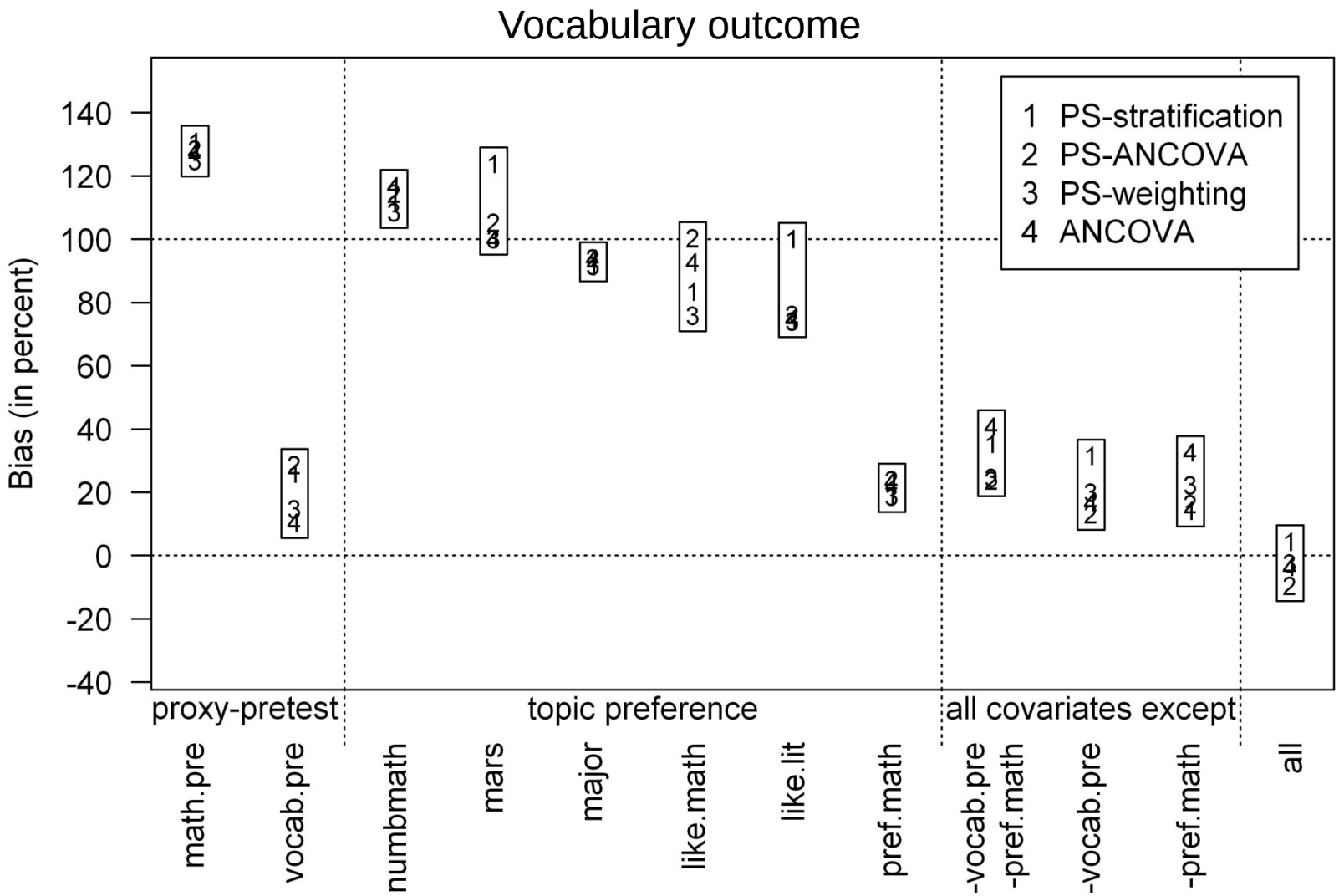
- Might be well known in case of *administrator selection* (if participants are selected according to some guidelines)
- In case of *self-selection*, selection processes might be more complex (involve many stakeholders) and less well known;
  → Thus, the identification of the crucial selection-relevant covariates is much more challenging

Covariates That Work in General

(Proxy)-pretest measures of the outcome

- *Why* can (proxy)-pretests remove most of the bias?
  (a) pretest measures are *highly correlated* with the outcome
  (b) frequently serve as a *proxy* for unobserved confounders
  (c) hard to imagine that a selection mechanism only confounds the posttest but does not show up in the (proxy)-pretest (particularly if they are measured close in time)
- *Empirically validated* via within-study comparisons (reviews by Cook et al., 2008; Glazerman et al., 2003; but need more research)

Covariates That Work in General Example: Steiner et al. (2010)



Covariates That Work in General

Advantage

- *No strong knowledge* about the DGM required
- *Causal claims are better warranted* if (proxy)-pretest measures are available (particularly with multiple pretest measures)

Disadvantage

- Pretest and posttest measures need to be *reliably measured* and *highly correlated*
- *No guarantee* that it always works
- Pretest measures are *not always available*

Empirical Tests

Empirical tests on the observed data may be used to select covariates (i.e., model-selection)

Select covariates that are related to

- The *observed outcome* (Hill, 2008; Myers et al.; Steyer, in press) → violated design aspects when PS designs are used (i.e., covariate cannot be selected without looking at the outcome)
- The *pretest or proxy-pretest* of the outcome (Kelcey, 2011)
- The *treatment* — not recommended since bias-amplifying (near) instruments might get selected

Empirical Tests

What does *"related"* mean?

- *Partial correlations* (instead of bivariate correlations)
- Thus, empirical criteria strongly rely on the *correct specification* of the outcome or selection model

Selected covariates are then used for

- *Regression* or *propensity score* (PS) adjustments
- *Balance checks* (if PS methods are used) — imbalance in unselected covariates does not matter

Empirical Tests

Advantages

- Covariate selection is based on the *actually observed relations* between variables (instead of theoretically assumed ones)
- *No strong knowledge* about the DGM required

Disadvantages

- Only *overt bias* due to observed covariates can be addressed — ex post selection!
  → bias due to unobserved covariates remains hidden
- Covariate selection depends on the *correct specification* of the selection or outcome model
- *Violation of design requirements* (with posttest measures)

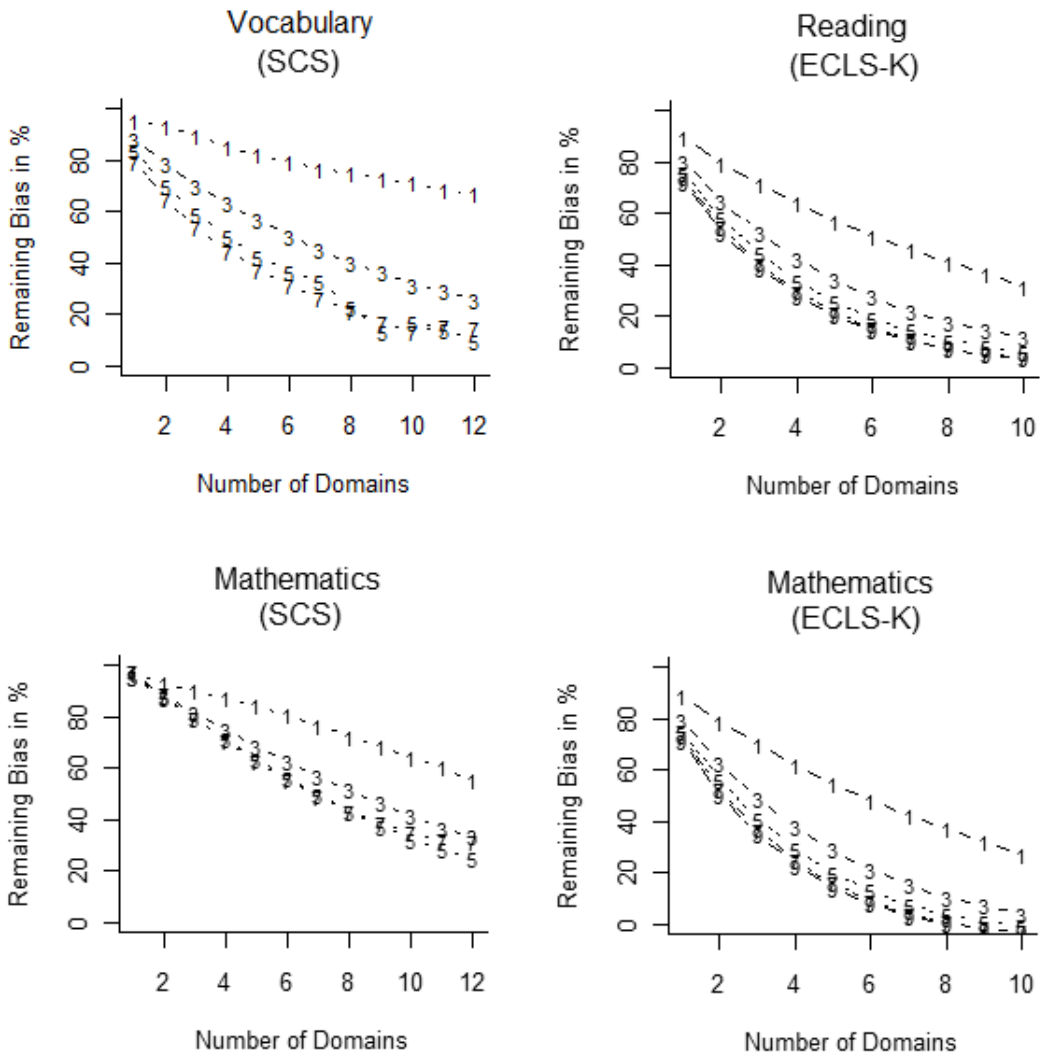
“Kitchen Sink” Approach

Consider *all observed covariates* “unless a variable can be excluded because there is consensus that it is unrelated to the outcome” (Rubin & Thomas, 1996)

- Tries to *mimic a randomized experiment* by balancing *all* observed covariates (instead of some pre-selected covariates only)
- Mostly relevant for matching techniques
- For a successful removal of bias, the number and *heterogeneity of covariates* matters: multiple construct domains with multiple measures per domain (domains & constructs are ideally selected on grounds of substantive knowledge)

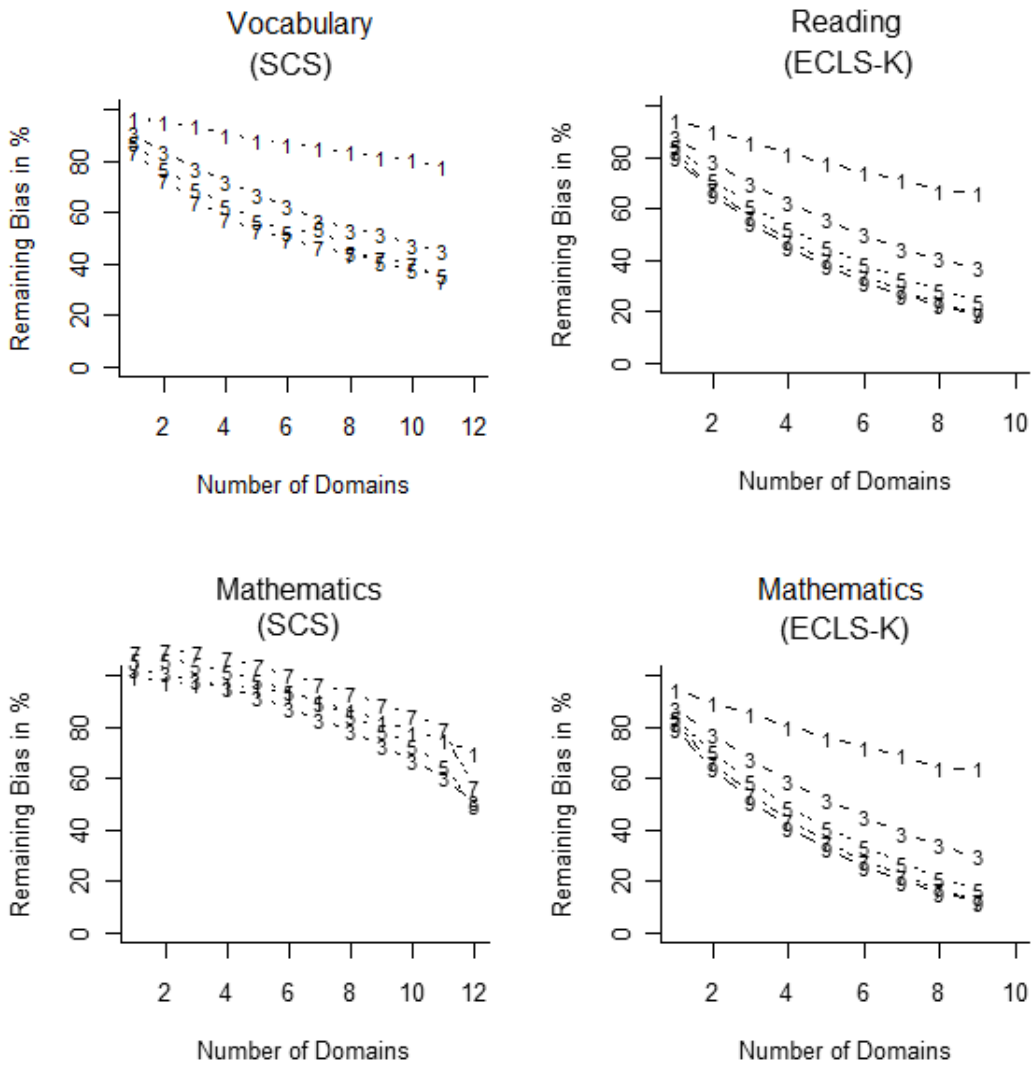
"Kitchen Sink" Approach Example (Steiner, Cook & Li, in prep.)

All covariates



“Kitchen Sink” Approach Example

Without critical Covariates (pretests)



“Kitchen Sink” Approach

Advantages

- *No strong substantive theories* are required
- Might even work if most effective covariates are unobserved—given a broad range of observed covariates

Disadvantage

- Only *overt bias* due to observed covariates can be addressed — ex post selection!
- If *not all confounders are observed*, balancing all observed covariates might not necessarily be a good idea (→ near-instruments, M-bias)
  Hypothesis: if a reliable pretest and a broad range of covariates is available, bias-inducing and bias-amplifying effects are rather small

Conclusions

CSMs are helpful for covariate selection if strong substantive theories are available

Pretest measures might work very well, but there is no guarantee

Empirical tests rely on the correct specification of the model and may violate design aspects

"Kitchen sink" approach might be the best we can do without strong substantive theory, particularly if pretest measures are available

→ No uniformly best approach

Conclusions

Best strategy for selecting covariates might be to begin with an ex ante selection by

- Using a hypothesized *CSM*
- Measuring multiple waves of *pretests* and *proxy-pretest* measures
- Measuring a *broad range of additional covariates* covering covariate domains relevant to the CSM

And then an ex post selection

- *Model-based*: select covariates on the basis of their relation to the outcome (or the hypothesized CSM)
- *Design-based*: achieve balance on all covariates using a matching approach

Conclusion

Need *more research* about what works in practice (without assuming unrealistic CSMs) – within-study comparisons helps here

Discussed only the 'simple' case of estimating the treatment effect from single-level data

With *multilevel data* everything gets much more complex since selection processes may operate at different levels in opposite direction

For estimating *mediation effects,* covariates selection is probably even more critical and complex (sequential ignorability!)

Contact:
psteiner@wisc.edu