Estimating Causal Effects With Error-Prone Exposures Using Control Variates

Keith Barnatchez [1] Kevin Josey [2] Rachel Nethery [1] Giovanni Parmigiani [1] Bryan Shepherd [3]

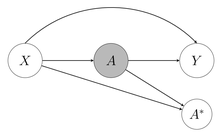
[1]Harvard T.H. Chan School of Public Health [2]Colorado School of Public Health [3]Vanderbilt University Medical Center

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Exposures of interest are often measured with error

In practice, we are often only able to obtain *error prone* measurements, A^*, of the exposure of interest, A

Y	A	A^*	X
Y_1	?	A_1^*	X_1
⋮	⋮	⋮	⋮
Y_n	?	A_n^*	X_n



Suppose interest lies in estimating the average treatment effect:

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

Well-established literature that using A^* in place of A produces substantial **bias** that scales with the severity of the measurement error

Addressing measurement error through study design

In many settings, it is possible to **validate** a random subset of error-prone observations

- **EHR data**: Manual chart review
- **Surveys**: Intensive follow-up

Validation procedure induces a **missing data** structure:

Y	A	A^*	X	S
Y_1	A_1	A_1^*	X_1	1
Y_2		A_2^*	X_2	0
Y_3		A_3^*	X_3	0
Y_4	A_4	A_4^*	X_4	1
Y_5		A_5^*	X_5	0
Y_6	A_6	A_6^*	X_6	1

(a) Main dataset

Y	A	A^*	X	S
Y_1	A_1	A_1^*	X_1	1
Y_4	A_4	A_4^*	X_4	1
Y_6	A_6	A_6^*	X_6	1

(b) Validation dataset

Large set of **imputation-based** methods for addressing measurement error with validation data

Challenges with current approaches

- Multiple imputation / regression calibration
  - Consistency of estimator relies on consistency of the imputation model
- Doubly robust approaches
  - Difficulty of implementation, hampering their use in applied practice
  - *Instability* in smaller samples due to multiplicative weighting terms

There remains a critical need flexible, straightforward to implement methods that possess desirable theoretical properties under minimal modeling assumptions

Assumptions

1. Consistency: $Y = AY(1) + (1-A)Y(0)$
2. Positivity: $0 < \mathbb{P}(A = 1|\boldsymbol{X} = x) < 1$ for all x with positive support
3. No unmeasured confounding: $(Y(1), Y(0)) \perp\!\!\!\perp A|\boldsymbol{X}$

Initially assume (for simplicity) that $S \perp\!\!\!\perp (Y, A, A^*, \boldsymbol{X})$, where $\mathbb{P}(S = 1) = \rho \in (0, 1)$

The control variates method enjoys the **simplicity** and **flexibility** of imputation-based approaches, while inheriting many properties possessed by current **doubly-robust** approaches

Our proposal

Intuition: Improve the efficiency of an initial unbiased (but inefficient) estimator of τ by **augmenting** it with a variance reduction term formed from the full data

Our approach, adapted from [1] who focused on settings with partially unmeasured confounders, can be carried out in 4 steps:

Step 1: Obtain validation data only estimator

Y	A	A^*	$\boldsymbol{X}$
Y_1	A_1	A_1^*	X_1
Y_3	A_3	A_3^*	X_3
Y_5	A_5	A_5^*	X_5

Validation data only estimate: $\hat{\tau}_{\text{val}}$

Step 2: Construct the control variate

Y	A	A^*	$\boldsymbol{X}$
Y_1	A_1	A_1^*	X_1
Y_2		A_2^*	X_2
Y_3	A_3	A_3^*	X_3
Y_4		A_4^*	X_4
Y_5	A_5	A_5^*	X_5
Y_6		A_6^*	X_6

Error-prone estimate: $\hat{\tau}^{\text{e.p.}}_{\text{main}}$

Y	A	A^*	$\boldsymbol{X}$
Y_1	A_1	A_1^*	X_1
Y_3	A_3	A_3^*	X_3
Y_5	A_5	A_5^*	X_5

Error-prone estimate: $\hat{\tau}^{\text{e.p.}}_{\text{val}}$

Step 3: Compute the variance reduction term

Obtain $\hat{\Gamma} = \widehat{\text{Cov}}(\hat{\tau}_{\text{val}}, \hat{\tau}^{\text{e.p.}}_{\text{val}} - \hat{\tau}^{\text{e.p.}}_{\text{main}})$ and $\hat{V} = \widehat{\text{Var}}(\hat{\tau}^{\text{e.p.}}_{\text{val}} - \hat{\tau}^{\text{e.p.}}_{\text{main}})$

Step 4: Form the final estimator

Obtain final estimate: $\hat{\tau}_{\text{CV}} = \hat{\tau}_{\text{val}} - \hat{\Gamma}\hat{V}^{-1}(\hat{\tau}^{\text{e.p.}}_{\text{val}} - \hat{\tau}^{\text{e.p.}}_{\text{main}})$

Key idea: While $\hat{\tau}^{\text{e.p.}}_{\text{val}}$ and $\hat{\tau}^{\text{e.p.}}_{\text{main}}$ will be biased for τ, notice that $\hat{\tau}^{\text{e.p.}}_{\text{val}} - \hat{\tau}^{\text{e.p.}}_{\text{main}} \to 0$, implying

$$\hat{\tau}_{\text{val}} - b(\hat{\tau}^{\text{e.p.}}_{\text{val}} - \hat{\tau}^{\text{e.p.}}_{\text{main}}) \to \tau \quad \text{for any } b \in \mathbb{R}$$

Setting $b = \Gamma V^{-1}$ yields the largest possible variance reduction relative to $\hat{\tau}_{\text{val}}$

Properties

- **Flexibility**: Method is *general* – it can accommodate *any* choice for the component estimators $\hat{\tau}_{\text{val}}$, $\hat{\tau}^{\text{e.p.}}_{\text{val}}$ and $\hat{\tau}^{\text{e.p.}}_{\text{main}}$ so long as they are regular asymptotically linear
  - Recommendation: Doubly-robust methods (e.g. AIPW, TMLE)
- **Efficiency gain**: $\text{Var}(\hat{\tau}_{\text{CV}}) = \text{Var}(\hat{\tau}_{\text{val}}) - \Gamma^2/V$
- **Double robustness**: $\sqrt{n}$ rates of convergence if outcome and propensity score models are both $o_{\text{P}}(n^{-1/4})$, consistency if at least one nuisance model is consistent

Extensions

Control variates method can account for...

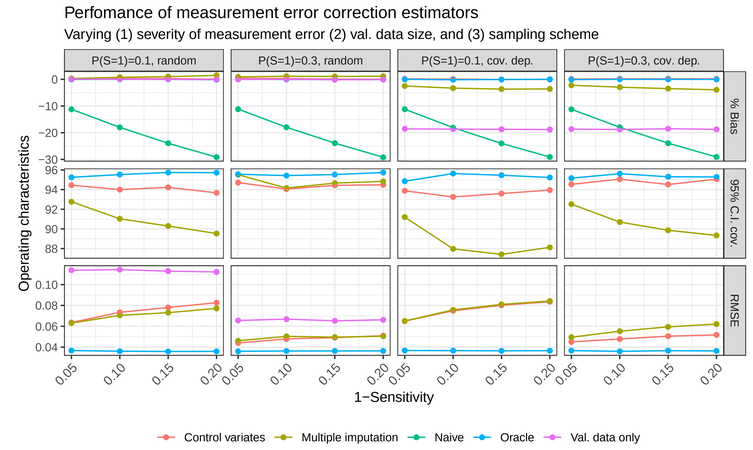
- More general validation data sampling schemes / account for multiple study sites
- Simultaneous error in the outcome of interest
- Other causal/non-causal estimands estimands
  - E.g. **local average treatment effects** if one has access to an instrumental variable
  - Stochastic intervention effects

References

[1] Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2019.

Simulation

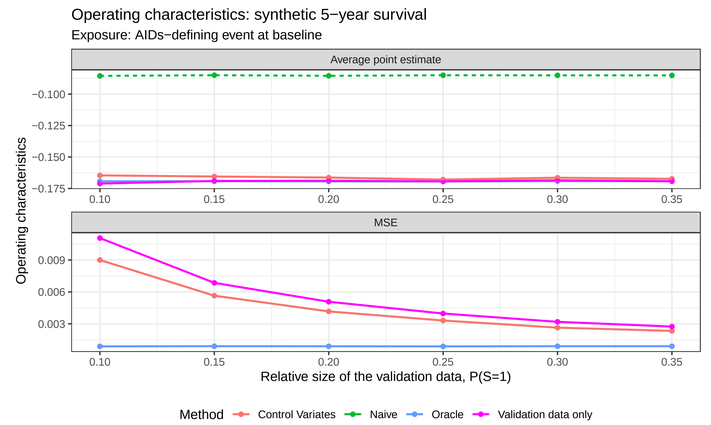
Goal: Assess performance of the control variance estimator under (1) varying levels of measurement error severity, (2) increasingly large shares of validation data, and (3) different validation data sampling schemes



Main takeaways: (1) Ignoring measurement error can generate severe **bias**, (2) the control variates method can provide substantial **variance reduction**, competing well with current standard approaches, while (3) possessing additional safeguards against model misspecification

Real Data: Vanderbilt Comprehensive Care Clinic (VCCC)

- EHR database with ≈ 1900 patients living with HIV receiving care from the VCCC
- Substantial error in key variables, including occurence of an AIDs-defining event (ADE) at baseline
- Team of researchers validated **every** observation
- **Causal estimand**: Average causal effect of ADE at baseline on (synthetic) 5-year survival
- **Goal**: Investigate performance of control variates method when revealing increasingly larger shares of the validation data



- High false positive rate of ADE (≈ 10%) leads to large **bias** (> 40%) in estimate of ATE
- Control variates method provides moderate **efficiency gains** for smaller validation sizes

https://www.kbarnatchez.com keithbarnatchez@g.harvard.edu