Measurement Error in Causal Inference: A Review

Keith Barnatchez, Kevin Josey, Rachel Nethery

May 4th, 2023

Outline

Roadmap

Motivation

A necessary, but often unstated, assumption in causal inference is that all variables are measured without error

- Commonly violated; e.g. air pollution, self-reported health measures, gene expression levels

Motivation

A necessary, but often unstated, assumption in causal inference is that all variables are measured without error

- Commonly violated; e.g. air pollution, self-reported health measures, gene expression levels

Measurement error literature long-established...

- But work at the intersection of M.E. + causal inference is relatively new
- Growing set of methods; rationale behind them + their relative merits often unclear

Goals for Today's Talk

1. Give intuition for the problems M.E. can cause in associational/causal studies

Goals for Today's Talk

1. Give intuition for the problems M.E. can cause in associational/causal studies
2. Demonstrate the key role of study design in addressing M.E.
    - And argue for why one should always strive to collect validation data when possible

Goals for Today's Talk

1. Give intuition for the problems M.E. can cause in associational/causal studies
2. Demonstrate the key role of study design in addressing M.E.
   - And argue for why one should always strive to collect validation data when possible
3. Overview a few workhorse methods for addressing M.E. in parametric models commonly used in epi research
   - These methods have heavily influenced early work at intersection of M.E. + causal inference

Goals for Today's Talk

1. Give intuition for the problems M.E. can cause in associational/causal studies
2. Demonstrate the key role of study design in addressing M.E.
   - And argue for why one should always strive to collect validation data when possible
3. Overview a few workhorse methods for addressing M.E. in parametric models commonly used in epi research
   - These methods have heavily influenced early work at intersection of M.E. + causal inference
4. Review recent developments in the causal inference literature for addressing M.E.
   - Current gaps, connections to the missing data literature, and ways forward

Roadmap

Contents

Attenuation Bias

Measurement error in *parametric models* is well-studied

- Simplest example of M.E. induces *attenuation bias*

Attenuation Bias

Measurement error in *parametric models* is well-studied

- Simplest example of M.E. induces *attenuation bias*

Consider a simple scenario of *classical* measurement error

Attenuation Bias

Measurement error in *parametric models* is well-studied

- Simplest example of M.E. induces *attenuation bias*

Consider a simple scenario of *classical* measurement error

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$W = X + U, \qquad \underbrace{U \sim N(0, \sigma_U^2)}_{\text{Meas. error}}, \qquad X \perp\!\!\!\perp U$$

Researcher observes Y, and *error-prone* measurements of X: W

Attenuation Bias

Measurement error in *parametric models* is well-studied

- Simplest example of M.E. induces *attenuation bias*

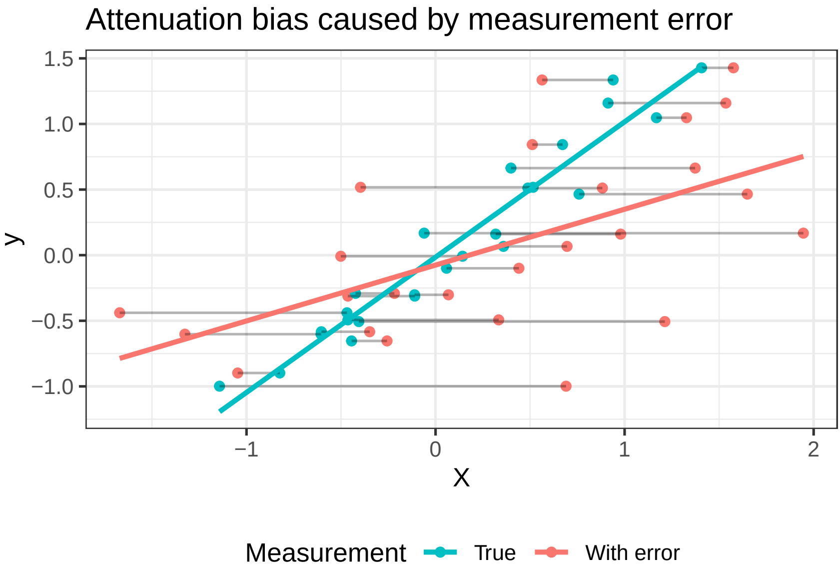
Consider a simple scenario of *classical* measurement error

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$W = X + U, \qquad \underbrace{U \sim N(0, \sigma_U^2)}_{\text{Meas. error}}, \qquad X \perp\!\!\!\perp U$$

Researcher observes Y, and *error-prone* measurements of X: W

- What happens here if we ignore measurement error?

Attenuation Bias



Classical measurement error: more to the story

Consider a slightly more complex scenario of *classical* measurement error

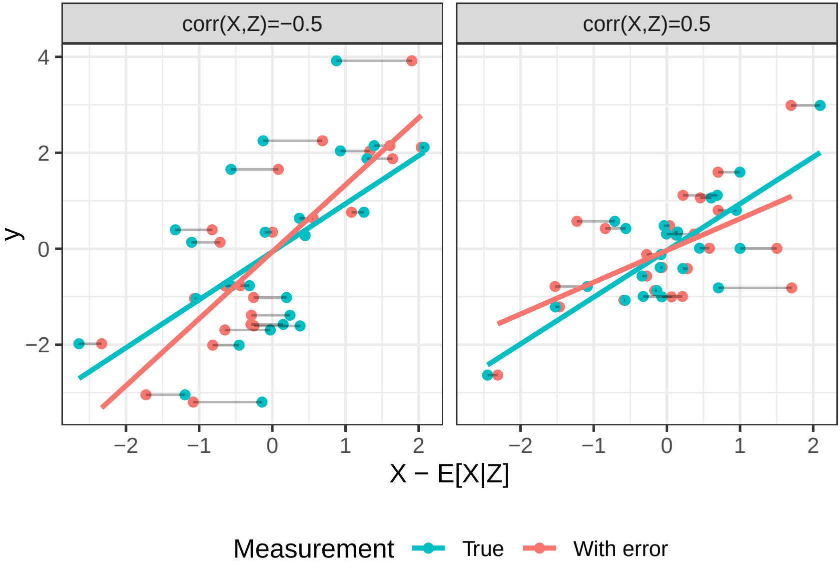
Classical measurement error: more to the story

Consider a slightly more complex scenario of *classical* measurement error

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$W = X + U, \qquad U \sim N(0, \sigma_U^2), \qquad X \perp\!\!\!\perp U$$

Researcher observes Y, Z and W

Classical measurement error: more to the story



More General Structure

In general, the problems caused by measurement error are much more complicated than the previous pictures imply

- **Error structure** can be complex and systematic:

$$\underbrace{W}_{\text{Meas.}} = \alpha_0 + \alpha_1 \underbrace{X}_{\text{True}} + \boldsymbol{Z}\boldsymbol{\beta} + \underbrace{U}_{\text{Error}}$$

- **Objects of interest** often extend beyond the parameters of a linear regression model, e.g.
  1. Parameters of non-linear models
  2. Distribution estimation
  3. Causal quantities

More General Structure

In general, the problems caused by measurement error are much more complicated than the previous pictures imply

- **Error structure** can be complex and systematic:

$$\underbrace{W}_{\text{Meas.}} = \alpha_0 + \alpha_1 \underbrace{X}_{\text{True}} + \boldsymbol{Z\beta} + \underbrace{U}_{\text{Error}}$$

- **Objects of interest** often extend beyond the parameters of a linear regression model, e.g.
  1. Parameters of non-linear models
  2. Distribution estimation
  3. Causal quantities

In response to these challenges, there's been a **lot** of work done on 1 and 2

- Work on 3 is more recent, heavily borrowing from work in 1

Contents

Identification

Return to our earlier scenario: we have data on an outcome Y, error-prone measurements W of continuous covariate X:

$$W = X + U, \quad U \sim N(0, \sigma_U^2) \text{ and } X \sim N(\mu_x, \sigma_X^2)$$
$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Identification

Return to our earlier scenario: we have data on an outcome Y, error-prone measurements W of continuous covariate X:

$$W = X + U, \quad U \sim N(0, \sigma_U^2) \text{ and } X \sim N(\mu_x, \sigma_X^2)$$
$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Joint distribution of (Y, W) characterized by 5 moment equations with 6 additional unknowns on RHS (Wang 2021):

$$\begin{aligned}
\mu_Y &= \beta_0 + \beta_1 \mu_X \\
\mu_X &= \mu_W \\
\sigma_Y^2 &= \beta_1 \text{Cov}(Y, W) + \sigma_\varepsilon^2 \\
\text{Cov}(Y, W) &= \beta_1 \sigma_X^2 \\
\sigma_W^2 &= \sigma_X^2 + \sigma_U^2
\end{aligned}$$

Identification + Study Design

Intuition: In order to adjust for measurement error, we need *some* information on the measurement error process

A few different ways to collect information on the M.E. process:

Identification + Study Design

Intuition: In order to adjust for measurement error, we need *some* information on the measurement error process

A few different ways to collect information on the M.E. process:

1. Obtain "gold-standard" measurements for a small subset of the main data (typically called the *validation data*)

Identification + Study Design

Intuition: In order to adjust for measurement error, we need *some* information on the measurement error process

A few different ways to collect information on the M.E. process:

1. Obtain "gold-standard" measurements for a small subset of the main data (typically called the *validation data*)
2. Obtain repeated measurements per subject of the error-prone variable (to estimate σ_U^2)

Identification + Study Design

Intuition: In order to adjust for measurement error, we need *some* information on the measurement error process

A few different ways to collect information on the M.E. process:

1. Obtain "gold-standard" measurements for a small subset of the main data (typically called the *validation data*)
2. Obtain repeated measurements per subject of the error-prone variable (to estimate σ_U^2)
3. Assume values/place priors on some of the M.E. params, ideally using information from previous studies

Identification + Study Design

Intuition: In order to adjust for measurement error, we need *some* information on the measurement error process

A few different ways to collect information on the M.E. process:

1. Obtain "gold-standard" measurements for a small subset of the main data (typically called the *validation data*)
2. Obtain repeated measurements per subject of the error-prone variable (to estimate σ_U^2)
3. Assume values/place priors on some of the M.E. params, ideally using information from previous studies

The resulting methods available for addressing M.E. are highly dependent on whether (1), (2) or (3) is used

Data Structure: No Adjustments

Y	X	W	$\boldsymbol{Z}$
Y_1		W_1	$\boldsymbol{Z}_1$
Y_2		W_2	$\boldsymbol{Z}_2$
Y_3		W_3	$\boldsymbol{Z}_3$
Y_4		W_4	$\boldsymbol{Z}_4$
Y_5		W_5	$\boldsymbol{Z}_5$
Y_6		W_6	$\boldsymbol{Z}_6$
Y_7		W_7	$\boldsymbol{Z}_7$
Y_8		W_8	$\boldsymbol{Z}_8$
Y_9		W_9	$\boldsymbol{Z}_9$
Y_{10}		W_{10}	$\boldsymbol{Z}_{10}$

Data Structure: Ideal Scenario

Y	X	W	$\boldsymbol{Z}$
Y_1	X_1	W_1	$\boldsymbol{Z}_1$
Y_2	X_2	W_2	$\boldsymbol{Z}_2$
Y_3	X_3	W_3	$\boldsymbol{Z}_3$
Y_4	X_4	W_4	$\boldsymbol{Z}_4$
Y_5	X_5	W_5	$\boldsymbol{Z}_5$
Y_6	X_6	W_6	$\boldsymbol{Z}_6$
Y_7	X_7	W_7	$\boldsymbol{Z}_7$
Y_8	X_8	W_8	$\boldsymbol{Z}_8$
Y_9	X_9	W_9	$\boldsymbol{Z}_9$
Y_{10}	X_{10}	W_{10}	$\boldsymbol{Z}_{10}$

Data Structure: Internal Validation Data

Y	X	W	$\boldsymbol{Z}$
Y_1	X_1	W_1	$\boldsymbol{Z}_1$
Y_2		W_2	$\boldsymbol{Z}_2$
Y_3	X_3	W_3	$\boldsymbol{Z}_3$
Y_4		W_4	$\boldsymbol{Z}_4$
Y_5	X_5	W_5	$\boldsymbol{Z}_5$
Y_6		W_6	$\boldsymbol{Z}_6$
Y_7	X_7	W_7	$\boldsymbol{Z}_7$
Y_8		W_8	$\boldsymbol{Z}_8$
Y_9	X_9	W_9	$\boldsymbol{Z}_9$
Y_{10}		W_{10}	$\boldsymbol{Z}_{10}$

Data Structure: Repeated Measurements

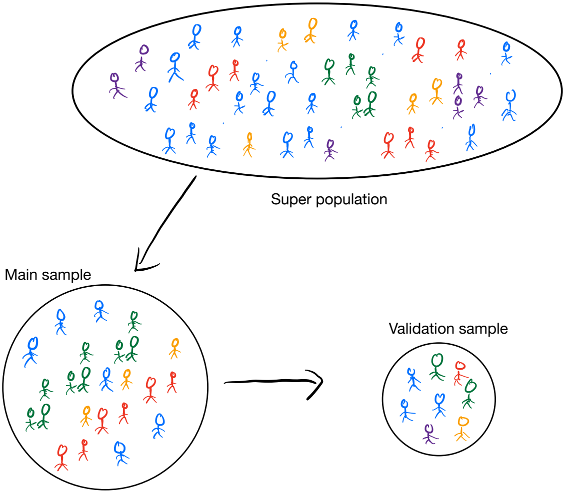
Y	X	W_1	W_2	$\boldsymbol{Z}$
Y_1		$W_{1,1}$	$W_{2,1}$	$\boldsymbol{Z}_1$
Y_2		$W_{1,2}$	$W_{2,2}$	$\boldsymbol{Z}_2$
Y_3		$W_{1,3}$	$W_{2,3}$	$\boldsymbol{Z}_3$
Y_4		$W_{1,4}$	$W_{2,4}$	$\boldsymbol{Z}_4$
Y_5		$W_{1,5}$	$W_{2,5}$	$\boldsymbol{Z}_5$
Y_6		$W_{1,6}$	$W_{2,6}$	$\boldsymbol{Z}_6$
Y_7		$W_{1,7}$	$W_{2,7}$	$\boldsymbol{Z}_7$
Y_8		$W_{1,8}$	$W_{2,8}$	$\boldsymbol{Z}_8$
Y_9		$W_{1,9}$	$W_{2,9}$	$\boldsymbol{Z}_9$
Y_{10}		$W_{1,10}$	$W_{2,10}$	$\boldsymbol{Z}_{10}$

Internal Validation Data

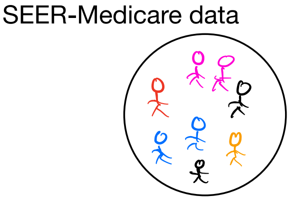
Mainly focus on methods that make use of *internal* validation data. Main reasons:

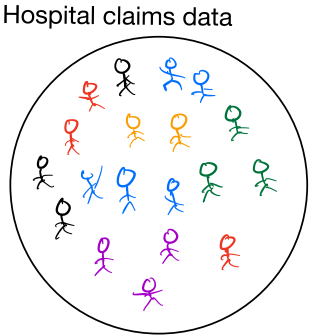
- Most M.E. adjustment methods are compatible with validation data (but many strictly require it)
- Allows for non-parametric identification of causal quantities like the average treatment effect (ATE)
  - Generally not possible without validation data
- Allows for us to use tools from the missing data literature
  - With internal validation data, M.E. becomes a missing data problem

Ways to obtain validation data: Double sampling

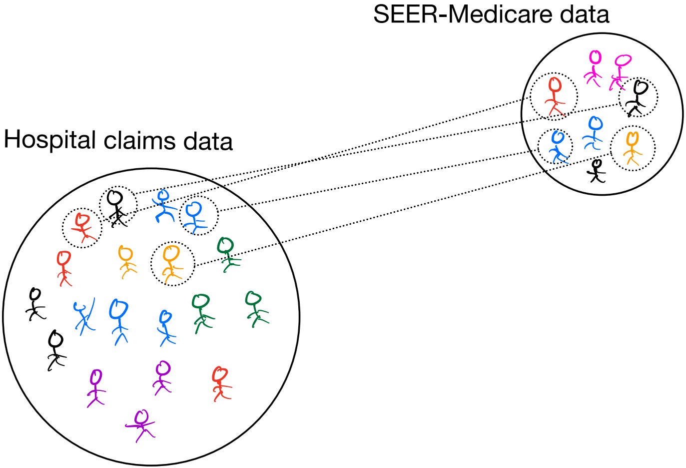


Other ways: Cleverness (Braun et al. 2017)





Other ways: Cleverness (Braun et al. 2017)



Contents

Addressing M.E. in Parametric Models

To fix ideas, suppose we'd like to estimate the parameters of the following model:

$$g(\mathbb{E}(Y|X, \boldsymbol{Z})) = \beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}$$

Addressing M.E. in Parametric Models

To fix ideas, suppose we'd like to estimate the parameters of the following model:

$$g(\mathbb{E}(Y|X, \boldsymbol{Z})) = \beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}$$

where we observe

$$(Y_i, \color{red}{W_i}, \boldsymbol{Z}_i),\ i \in \{1, \dots, N\}$$

Addressing M.E. in Parametric Models

To fix ideas, suppose we'd like to estimate the parameters of the following model:

$$g(\mathbb{E}(Y|X, \boldsymbol{Z})) = \beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}$$

where we observe

$$(Y_i, W_i, \boldsymbol{Z}_i),\ i \in \{1, \ldots, N\}$$

and for a subset of subjects (say the first n) we observe

$$(Y_j, W_j, X_j, \boldsymbol{Z}_j),\ j \in \{1, \ldots, n\},\ n < N$$

Addressing M.E. in Parametric Models

To fix ideas, suppose we'd like to estimate the parameters of the following model:

$$g(\mathbb{E}(Y|X, \boldsymbol{Z})) = \beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}$$

where we observe

$$(Y_i, W_i, \boldsymbol{Z}_i),\ i \in \{1, \ldots, N\}$$

and for a subset of subjects (say the first n) we observe

$$(Y_j, W_j, X_j, \boldsymbol{Z}_j),\ j \in \{1, \ldots, n\},\ n < N$$

Typical to assume some parametric form for the M.E. model, e.g.

$$W = X + U, \quad U \perp\!\!\!\perp X$$

Addressing M.E. in Parametric Models

To fix ideas, suppose we'd like to estimate the parameters of the following model:

$$g(\mathbb{E}(Y|X, \boldsymbol{Z})) = \beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}$$

where we observe

$$(Y_i, W_i, \boldsymbol{Z}_i),\ i \in \{1, \ldots, N\}$$

and for a subset of subjects (say the first n) we observe

$$(Y_j, W_j, X_j, \boldsymbol{Z}_j),\ j \in \{1, \ldots, n\},\ n < N$$

Typical to assume some parametric form for the M.E. model, e.g.

$$W = \alpha_0 + \alpha_X X + \boldsymbol{Z}\boldsymbol{\alpha}_Z + U, \quad (X, \boldsymbol{Z}) \perp\!\!\!\perp U$$

Regression Calibration

Natural place to start: use information in the validation data to impute missing values in the main data

One way to "impute" is to just replace W with $\hat{\mathbb{E}}(X|W, \boldsymbol{Z})$

- where $\hat{\mathbb{E}}(X|W, \boldsymbol{Z})$ is estimated in the validation data

Regression Calibration: the main idea

Notice

$$\begin{aligned}
\mathbb{E}(Y|W,\boldsymbol{Z}) &= \mathbb{E}_{X|W,\boldsymbol{Z}}\mathbb{E}(Y|W,X,\boldsymbol{Z}) \\
&= \mathbb{E}_{X|W,\boldsymbol{Z}}\mathbb{E}(Y|X,\boldsymbol{Z}) && (Y \perp\!\!\!\perp W|X,\boldsymbol{Z}) \\
&= \mathbb{E}_{X|W,\boldsymbol{Z}}[\beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta_Z}] \\
&= \beta_0 + \beta_X \mathbb{E}(X|W,\boldsymbol{Z}) + \boldsymbol{Z}\boldsymbol{\beta_Z}
\end{aligned}$$

Regression Calibration: the main idea

Notice

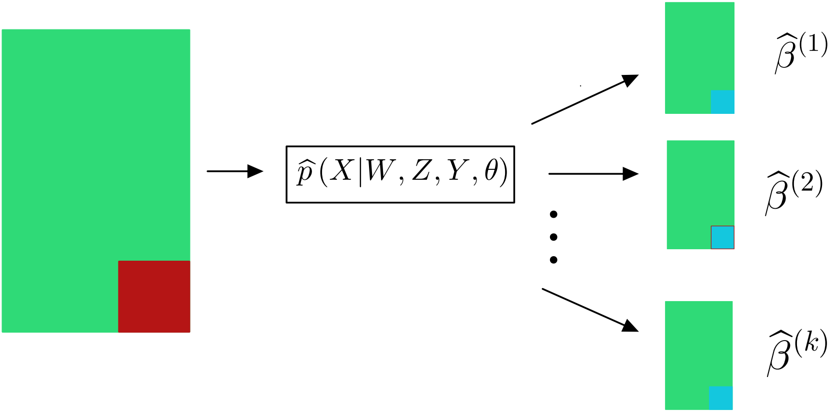
$$\begin{aligned}
\mathbb{E}(Y|W,\boldsymbol{Z}) &= \mathbb{E}_{X|W,\boldsymbol{Z}}\mathbb{E}(Y|W,X,\boldsymbol{Z}) \\
&= \mathbb{E}_{X|W,\boldsymbol{Z}}\mathbb{E}(Y|X,\boldsymbol{Z}) && (Y \perp\!\!\!\perp W|X,\boldsymbol{Z}) \\
&= \mathbb{E}_{X|W,\boldsymbol{Z}}[\beta_0 + \beta_X X + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}] \\
&= \beta_0 + \beta_X \mathbb{E}(X|W,\boldsymbol{Z}) + \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{Z}}
\end{aligned}$$

Under a linear outcome model, if we 1) regress X on W and $\boldsymbol{Z}$ in the validation data, and 2) replace W with $\hat{\mathbb{E}}(X|W,\boldsymbol{Z})$ in our outcome regression, then

- we're estimating the same reg. parameters we'd estimate if we had complete information on X

Consistency hinges upon linearity of the outcome model + meas. errors not depending on Y (given X and $\boldsymbol{Z}$)

Multiple Imputation for Measurement Error (MIME)



Likelihood Approach

Hinges on 3 assumptions

1. The density of Y given $\boldsymbol{Z}, X$ is from an exponential family with dispersion parameter ϕ

Likelihood Approach

Hinges on 3 assumptions

1. The density of Y given $\boldsymbol{Z}, X$ is from an exponential family with dispersion parameter ϕ
2. The M.E. variance is known (!!) to be a fixed value σ_U^2

Likelihood Approach

Hinges on 3 assumptions

1. The density of Y given $\boldsymbol{Z}, X$ is from an exponential family with dispersion parameter ϕ
2. The M.E. variance is known (!!) to be a fixed value σ_U^2
3. The measurement error is additive/normally distributed:
    - $W = X + U, \quad X \perp\!\!\!\perp U, \quad U \sim N(0, \sigma_U^2)$

Likelihood Approach

Hinges on 3 assumptions

1. The density of Y given $\boldsymbol{Z}, X$ is from an exponential family with dispersion parameter ϕ
2. The M.E. variance is known (!!) to be a fixed value σ_U^2
3. The measurement error is additive/normally distributed:
    - $W = X + U, \quad X \perp\!\!\!\perp U, \quad U \sim N(0, \sigma_U^2)$

Then, it turns out the variable

$$\Delta = W + Y\sigma_U^2\beta_X/\phi$$

is sufficient for X

- Can condition on $\boldsymbol{Z}$ and Δ, solve set of score equations
- Consistent if above assumptions hold

Likelihood Approach

I.e. instead of solving score equations for

$$\prod_{i=1}^{n} f(y_i | \boldsymbol{z}_i, x_i, \boldsymbol{\beta})$$

Likelihood Approach

I.e. instead of solving score equations for

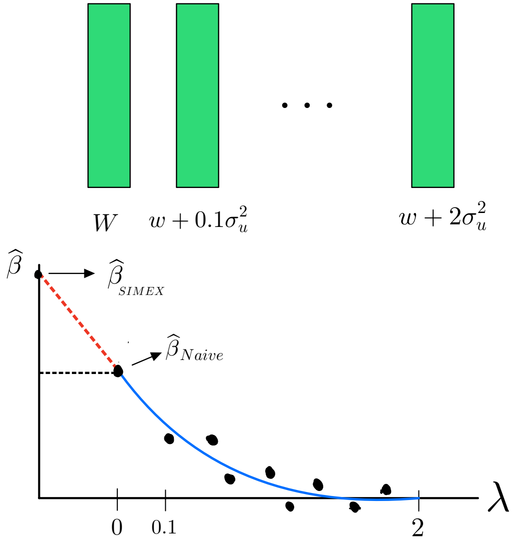
$$\prod_{i=1}^{n} f(y_i|\boldsymbol{z}_i, x_i, \boldsymbol{\beta})$$

Solve the score equations for

$$\prod_{i=1}^{n} f(y_i|\boldsymbol{z}_i, x_i, \Delta_i, \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{z}_i, \Delta_i, \boldsymbol{\beta})$$

- Δ_i only depends on observed data/parameters to be estimated

SIMEX



W
$w + 0.1\sigma_u^2$
$w + 2\sigma_u^2$
$\widehat{\beta}$
$\widehat{\beta}_{SIMEX}$
$\widehat{\beta}_{Naive}$
λ
0
0.1
2

Roadmap

Contents

Background

Work on M.E. adjustment for parametric models (1980s) long pre-dates M.E. adjustment in causal inference (2010s)

- In turn, early M.E. + causal inference work has heavily borrowed from work on M.E. adjustment in parametric models

Background

Work on M.E. adjustment for parametric models (1980s) long pre-dates M.E. adjustment in causal inference (2010s)

- In turn, early M.E. + causal inference work has heavily borrowed from work on M.E. adjustment in parametric models

One problem...

- Growing consensus in causal inference to avoid parametric assumptions wherever possible
- By necessity, many M.E. methods *need* to make parametric assumptions
  - Essential when no validation data available
  - Development/application of modern semi-parametric methods has been slow

Goal: Discuss current approaches to addressing M.E. in causal research

Background: Main Approaches

- Regression calibration
  - Used a lot *in practice*, but not much methods development since estimators are generally inconsistent

Background: Main Approaches

- Regression calibration
  - Used a lot *in practice*, but not much methods development since estimators are generally inconsistent
- Likelihood-based approaches
  - A **lot** of methods development. Can accommodate no val. data, but requires strict parametric/distributional assumptions

Background: Main Approaches

- Regression calibration
  - Used a lot *in practice*, but not much methods development since estimators are generally inconsistent
- Likelihood-based approaches
  - A **lot** of methods development. Can accommodate no val. data, but requires strict parametric/distributional assumptions
- Multiple Imputation
  - Can flexibly model M.E. process, easy to implement with packages like `mice` and `AIPW`

Background: Main Approaches

- Regression calibration
  - Used a lot *in practice*, but not much methods development since estimators are generally inconsistent
- Likelihood-based approaches
  - A **lot** of methods development. Can accommodate no val. data, but requires strict parametric/distributional assumptions
- Multiple Imputation
  - Can flexibly model M.E. process, easy to implement with packages like `mice` and `AIPW`
- SIMEX (basically jackknife for measurement error)
  - Good properties when M.E. magnitude is small, but performs poorly for large magnitudes and...
  - Doesn't handle complex error structures well

Problem Setting

Three key pieces to any observational causal inference problem:

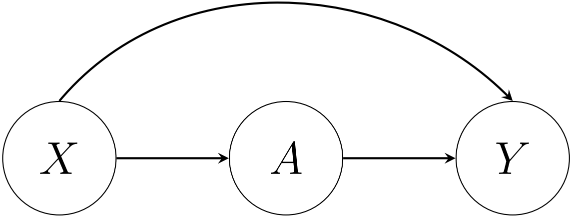
1. Outcome Y
2. Treatment A
3. Confounders $\boldsymbol{X}$

Measurement error can occur in any/all of them

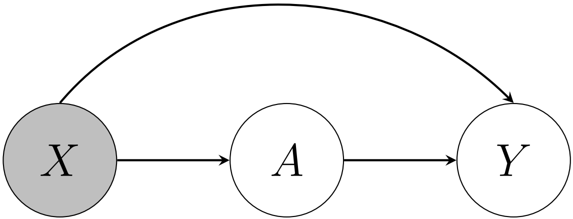
Will focus on scenario where error occurs in an important confounder

- Specifically, we'll continue to suppose we have a vector of correctly-measured confounders $\boldsymbol{Z}$ and one mis-measured confounder X (with measurements W)

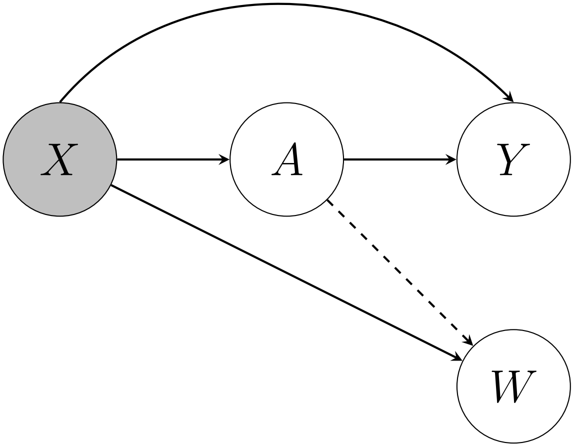
Confounder Measurement Error



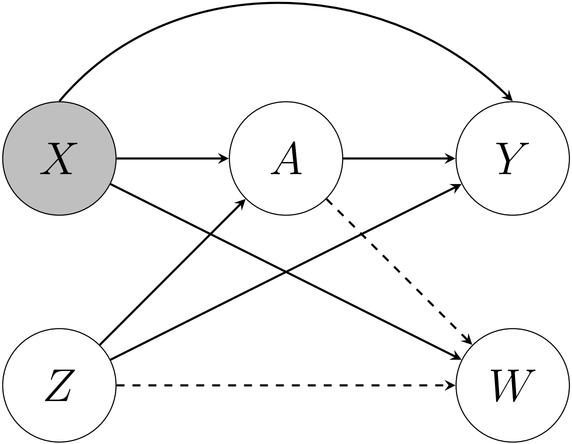
Confounder Measurement Error



Confounder Measurement Error



Confounder Measurement Error



Problem Setting

To fix ideas, suppose we observe

$$(Y_i, A_i, \boldsymbol{Z}_i, W_i, S_i) \sim \mathbb{P},\ i \in 1, \ldots, N$$

and for a subset of subjects we observe

$$(Y_j, A_j, \boldsymbol{Z}_j, W_j, X_j, S_j = 1),\ j \in \{1, \ldots, n\}, n < N$$

Problem Setting

To fix ideas, suppose we observe

$$(Y_i, A_i, \boldsymbol{Z}_i, W_i, S_i) \sim \mathbb{P},\ i \in 1, \ldots, N$$

and for a subset of subjects we observe

$$(Y_j, A_j, \boldsymbol{Z}_j, W_j, X_j, S_j = 1),\ j \in \{1, \ldots, n\}, n < N$$

We'd like to estimate the ATE:

$$\tau \overset{\text{def}}{=} \mathbb{E}(Y(1) - Y(0))$$

where $Y_i(a)$ is unit i's *potential outcome* had they been given treatment level a

Assumptions

Will make the following standard causal inference assumptions:

- **Consistency**: $Y = AY(1) + (1 - A)Y(0)$
- **Unconfoundedness**: $Y(a) \perp\!\!\!\perp A|X, \boldsymbol{Z}$ (implied by DAG)
- **Positivity**: $\mathbb{P}(\boldsymbol{z}, x) > 0 \implies 0 < \mathbb{P}(A = 1|\boldsymbol{z}, x) < 1$

Note: Unconfoundedness will not hold in observed data due to confounder M.E.

- $Y(a) \not\perp\!\!\!\perp A|W, \boldsymbol{Z}$
- But it *will* hold in the validation data

Causal Identification

Under 1) consistency , 2) unconfoundedness and 3) positivity, the ATE can be identified in the validation data via

$$\begin{aligned}
\mathbb{E}(Y(a)|S=1) &= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y(a)|X,\boldsymbol{Z},S=1) \\
&= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y(a)|X,\boldsymbol{Z},A=a,S=1) \\
&= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y|X,\boldsymbol{Z},A=a,S=1)
\end{aligned}$$

Causal Identification

Under 1) consistency , 2) unconfoundedness and 3) positivity, the ATE can be identified in the validation data via

$$\begin{aligned}\mathbb{E}(Y(a)|S=1) &= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y(a)|X,\boldsymbol{Z},S=1)\\ &= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y(a)|X,\boldsymbol{Z},A=a,S=1)\\ &= \mathbb{E}_{X,\boldsymbol{Z}|S=1}\mathbb{E}(Y|X,\boldsymbol{Z},A=a,S=1)\end{aligned}$$

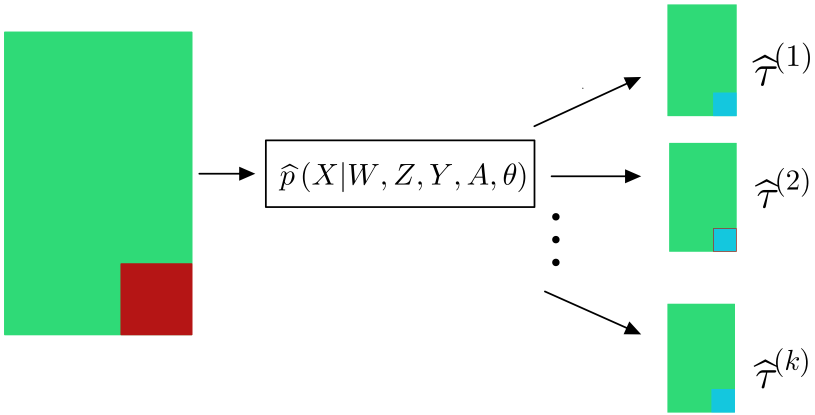
If our validation data is a random sample of the main data, then

$$\mathbb{E}(Y(a)|S=1) = \mathbb{E}(Y(a)) = \mathbb{E}_{X,\boldsymbol{Z}}\mathbb{E}(Y|X,\boldsymbol{Z},A=a,S=1)$$

- There are more useful/general identifying expressions than this (see e.g. Levis 2022)
- Without access to val. data, non-parametric identification generally not possible

Contents

Multiple Imputation



Multiple Imputation

One possible implementation:

1. Estimate imputation model with flexible approach, e.g. predictive mean matching
2. Estimate the treatment effect via augmented inverse probability weighting
    - Estimate the **outcome** and **propensity score models** non-parametrically
    - Good statistical rates ($\sqrt{n}$ consistency) despite estimating nuisance functions with flexible ML models

Multiple Imputation

One possible implementation:

1. Estimate imputation model with flexible approach, e.g. predictive mean matching
2. Estimate the treatment effect via augmented inverse probability weighting
   - Estimate the **outcome** and **propensity score models** non-parametrically
   - Good statistical rates ($\sqrt{n}$ consistency) despite estimating nuisance functions with flexible ML models

Approach is consistent if imputation, complete-data ATE estimators are consistent (Nguyen and Stuart 2023)

Likelihood-based Methods

Same idea as earlier:

1. Assume outcome, treatment models come from exponential families, and simple M.E. structure with known (!!) variance σ_U^2
2. Using σ_U^2 and W, can construct a variable Δ that is sufficient for the unknown X
3. Condition on $\boldsymbol{Z}$ and Δ, solve score equations

Lots of papers taking version of this approach[1]

- Val. data not an option/infeasible for many applied examples
- Methods development/sensitivity analysis along these lines still important

[1] E.g. see Shu and Yi (2019); Blette (2021); McCaffrey et al. (2013)

Connection to Work on Missing Data

With validation data, M.E. is really just a missing data problem

- Implying we can use tools developed for missing data problems in causal inference
- Multiple imputation is one example
- But can also take advantage of estimators developed with semi-parametric efficiency in mind
  - I.e. estimators based on efficient influence functions

Connection to Work on Missing Data

A few examples:

- M.E. in the exposure: Kennedy (2020)
- M.E. in outcomes: Kallus and Mao (2020)
- M.E. in confounders: Levis (2022)

Connection to Work on Missing Data

A few examples:

- M.E. in the exposure: Kennedy (2020)
- M.E. in outcomes: Kallus and Mao (2020)
- M.E. in confounders: Levis (2022)
- M.E. in outcome in high dimensions: Sankaranarayanan (soon)

Connection to Work on Missing Data

A few examples:

- M.E. in the exposure: Kennedy (2020)
- M.E. in outcomes: Kallus and Mao (2020)
- M.E. in confounders: Levis (2022)
- M.E. in outcome in high dimensions: Sankaranarayanan (soon)

Basic idea: under partial missingness + causal assumptions, 1) find identifying expression for ATE, 2) derive eff. influence function (EIF), 3) propose estimator based on EIF

- These estimators have nice properties/theoretical guarantees
  - Good statistical rates, even when nuisance models estimated with flexible ML methods that themselves have slower rates
- But implementation can be quite involved

Roadmap

Discussion

- Work in causal inference for addressing M.E. still in relatively early stages
- Study design is crucial
  - To avoid heavy reliance on parametric restrictions, strive for validation data designs
- With val. data, can frame M.E. as a missing data problem
  - Can use existing tools from missing data literature, but need to keep colloborative aspect of M.E. work in mind
  - i.e. multiple imputation + AIPW easier to implement; simulation studies needed to compare with existing DR estimators
- Double sampling not always possible; continued work only assuming repeated measurements/known M.E. variance needed
  - In particular, methods for sensitivity analysis under different M.E. mechanisms

References

Blette, B. S. (2021). *Causal Inference for Error-Prone Exposures*. PhD thesis, The University of North Carolina at Chapel Hill.

Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F., and Zigler, C. (2017). Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. *Biostatistics*, 18(4):695–710.

Josey, K. P., Berkowitz, S. A., Ghosh, D., and Raghavan, S. (2021). Transporting experimental results with entropy balancing. *Statistics in Medicine*, 40(19):4310–4326.

Kallus, N. and Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*.

Kennedy, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *The international journal of biostatistics*, 16(1).

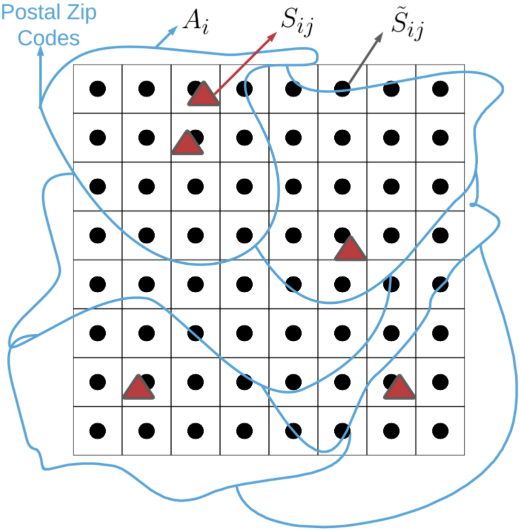
McCaffrey, D. F., Lockwood, J., and Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, 100(3):671–680.

Nguyen, T. Q. and Stuart, E. A. (2023). Multiple imputation for propensity score analysis with covariates missing at random: some clarity on within and across methods. *arXiv preprint arXiv:2301.07066*.

Shu, D. and Yi, G. Y. (2019). Weighted causal inference methods with mismeasured covariates and misclassified outcomes. *Statistics in medicine*, 38(10):1835–1854.

Wang, L. (2021). Identifiability in measurement error models. In *Handbook of Measurement Error Models*, pages 55–70. Chapman and Hall/CRC.

Other ways: Cleverness (Josey et al. 2021)



	Notations
▲	Air quality monitor locations (S_{ij} is known)
S_{ij}	Measured $PM_{2.5}$ using monitors
●	(i, j): grid cells
$\tilde{S}_{ij}$	Predicted grid-level $PM_{2.5}$ values
A_i	Unobserved ZIP-level $PM_{2.5}$ values