CEM: Coarsened Exact Matching for Stata

Matthew Blackwell
Institute for Quantitative Social Science
Harvard University

joint work with

Stefano M. Iacus (Univ. of Milan), Gary King (Harvard) and Giuseppe Porro (Univ. of Trieste)

(Stata Conference Boston July 16, 2010)

Preview Slide: Coarsened Exact Matching (CEM)

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
  3. Pass on original (uncoarsened) units except those pruned

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
  3. Pass on original (uncoarsened) units except those pruned
- Analyze as without matching (adding weights for stratum-size)

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
  3. Pass on original (uncoarsened) units except those pruned
- Analyze as without matching (adding weights for stratum-size)
- (Or apply other matching methods within CEM strata & they inherert CEM's properties)

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
  3. Pass on original (uncoarsened) units except those pruned
- Analyze as without matching (adding weights for stratum-size)
- (Or apply other matching methods within CEM strata & they inherert CEM's properties)

$\rightsquigarrow$ A version of CEM: Last studied 40 years ago by Cochran

Preview Slide: Coarsened Exact Matching (CEM)

A simple (and ancient) method of causal inference, with surprisingly powerful properties

- Preprocess (X, T) with CEM:
  1. Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
     - Easy to understand, or can be automated as for a histogram
  2. Perform exact matching on the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
  3. Pass on original (uncoarsened) units except those pruned
- Analyze as without matching (adding weights for stratum-size)
- (Or apply other matching methods within CEM strata & they inherert CEM's properties)

⇝ A version of CEM: Last studied 40 years ago by Cochran

⇝ First used many decades before that

Characteristics of Observational Data

Characteristics of Observational Data

- Lots of data

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment:

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random,

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator,

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known
- Bias-Variance Tradeoff

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known
- Bias-Variance Tradeoff

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known
- Bias-Variance Tradeoff
- The idea of matching: sacrifice some data to avoid bias

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known
- Bias-Variance Tradeoff
- The idea of matching: sacrifice some data to avoid bias
- Removing heterogeneous data will often reduce variance too

Characteristics of Observational Data

- Lots of data
- Data is of uncertain origin. Treatment assignment: not random, not controlled by investigator, not known
- Bias-Variance Tradeoff
- The idea of matching: sacrifice some data to avoid bias
- Removing heterogeneous data will often reduce variance too
- (Medical experiments are the reverse: small-n with random treatment assignment; don't match unless something goes wrong)

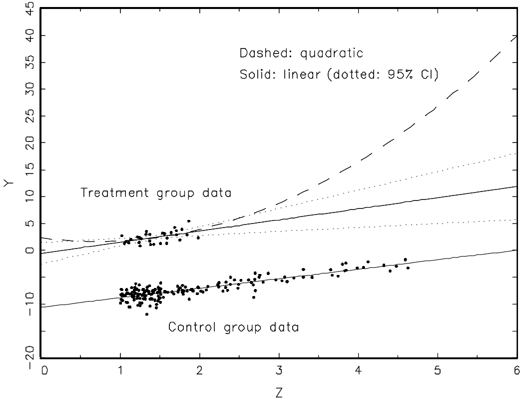
Model Dependence

Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

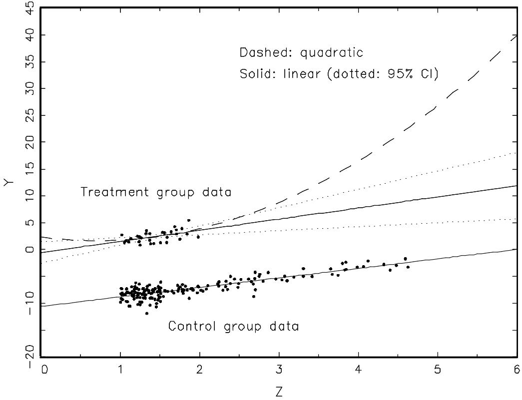
Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)



Model Dependence

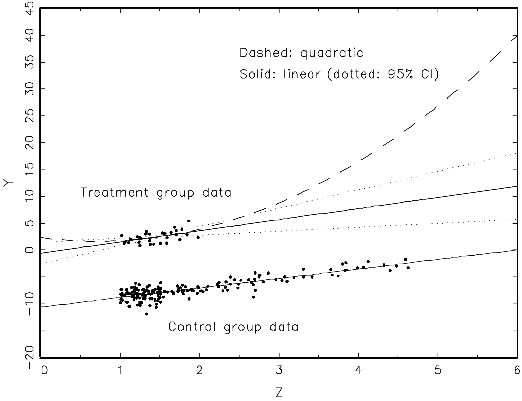
(King and Zeng, 2006: fig.4 *Political Analysis*)



What to do?

Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

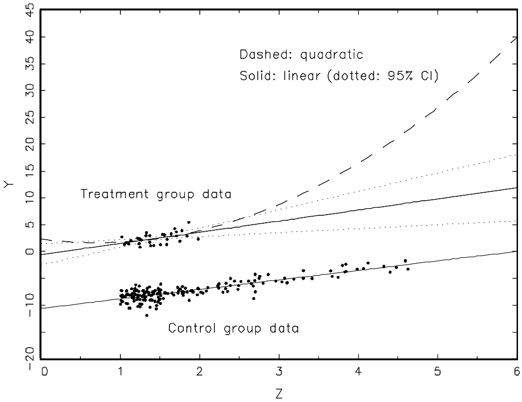


What to do?

- Preprocess I: Eliminate extrapolation region (a separate step)

Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

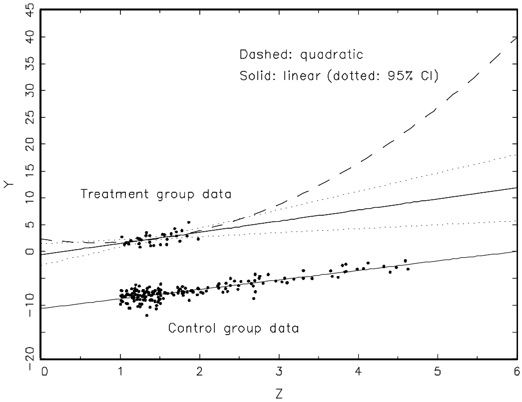


What to do?

- Preprocess I: Eliminate extrapolation region (a separate step)
- Preprocess II: Match (prune bad matches) within interpolation region

Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)



What to do?

- Preprocess I: Eliminate extrapolation region (a separate step)
- Preprocess II: Match (prune bad matches) within interpolation region
- Model remaining imbalance

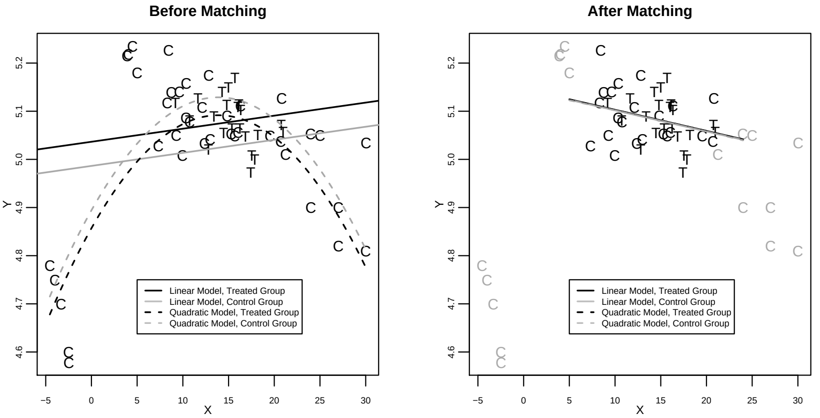
Matching within the Interpolation Region

Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

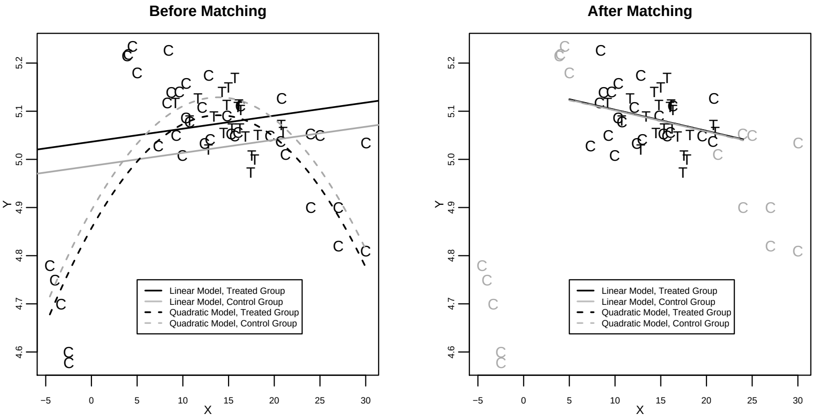
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



Matching reduces model dependence, bias, and variance

The Goals, with some more precision

The Goals, with some more precision

- Notation:

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

$$\text{TE}_i = Y_i(T_i = 1) - {\color{red} Y_i(T_i = 0)}$$

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

$$\begin{aligned} \mathrm{TE}_i &= Y_i(T_i = 1) - Y_i(T_i = 0) \\ &= \text{observed} \quad - \textit{unobserved} \end{aligned}$$

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

$$\begin{aligned} \mathrm{TE}_i &= Y_i(T_i = 1) - Y_i(T_i = 0) \\ &= \text{observed} \quad - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(T_i = 0)$ with Y_j from matched ($X_i \approx X_j$) controls

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

$$\begin{aligned} \text{TE}_i &= Y_i(T_i = 1) - Y_i(T_i = 0) \\ &= \text{observed} \quad - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(T_i = 0)$ with Y_j from matched ($X_i \approx X_j$) controls
- Prune unmatched units to improve balance (so X is unimportant)

The Goals, with some more precision

- Notation:
  - Y_i Dependent variable
  - T_i Treatment variable (dichotomous)
  - X_i Covariates
- Treatment Effect for treated ($T_i = 1$) observation i:

$$\begin{aligned} \text{TE}_i &= Y_i(T_i = 1) - Y_i(T_i = 0) \\ &= \text{observed} \quad - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(T_i = 0)$ with Y_j from matched ($X_i \approx X_j$) controls
- Prune unmatched units to improve balance (so X is unimportant)
- Sample Average Treatment effect on the Treated:

$$\text{SATT} = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} \text{TE}_i$$

Problems With Existing Matching Methods

Problems With Existing Matching Methods

- Don't eliminate extrapolation region

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
    - Least important (variance): matched n chosen ex ante

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice:

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice:

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice: choose n,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice: choose n, match,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice: choose n, match, publish,

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice: choose n, match, publish, STOP.

Problems With Existing Matching Methods

- Don't eliminate extrapolation region
- Don't work with multiply imputed data
- Not well designed for observational data:
  - Least important (variance): matched n chosen ex ante
  - Most important (bias): imbalance reduction checked ex post
- Hard to use: Improving balance on 1 variable can reduce it on others
  - Best practice: choose n-match-check, tweak-match-check, tweak-match-check, tweak-match-check, $\cdots$
  - Actual practice: choose n, match, publish, STOP. (Is balance even improved?)

Largest Class of Methods: Equal Percent Bias Reducing

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
    (a) X drawn randomly from a specified population $\mathbf{X}$,
    (b) $\mathbf{X} \sim$ Normal

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,
  (b) $\mathbf{X} \sim$ Normal
  (c) Matching algorithm is invariant to linear transformations of X.

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,
  (b) $\mathbf{X} \sim$ Normal
  (c) Matching algorithm is invariant to linear transformations of X.
  (d) Y is a linear function of X.

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  - (a) X drawn randomly from a specified population $\mathbf{X}$,
  - (b) $\mathbf{X} \sim$ Normal
  - (c) Matching algorithm is invariant to linear transformations of X.
  - (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\overbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \overbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  - (a) X drawn randomly from a specified population $\mathbf{X}$,
  - (b) $\mathbf{X} \sim$ Normal
  - (c) Matching algorithm is invariant to linear transformations of X.
  - (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\underbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \underbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,
  (b) $\mathbf{X} \sim$ Normal
  (c) Matching algorithm is invariant to linear transformations of X.
  (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\overbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \overbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:
  - Reducing mean-imbalance on one variable, reduces it on all

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,
  (b) $\mathbf{X} \sim$ Normal
  (c) Matching algorithm is invariant to linear transformations of X.
  (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\overbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \overbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:
  - Reducing mean-imbalance on one variable, reduces it on all
  - n set ex ante; balance calculated ex post

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  - (a) X drawn randomly from a specified population $\mathbf{X}$,
  - (b) $\mathbf{X} \sim$ Normal
  - (c) Matching algorithm is invariant to linear transformations of X.
  - (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\underbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \underbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:
  - Reducing mean-imbalance on one variable, reduces it on all
  - n set ex ante; balance calculated ex post
  - EPBR controls only expected (not in-sample) imbalance

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  - (a) X drawn randomly from a specified population $\mathbf{X}$,
  - (b) $\mathbf{X} \sim$ Normal
  - (c) Matching algorithm is invariant to linear transformations of X.
  - (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\underbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \underbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:
  - Reducing mean-imbalance on one variable, reduces it on all
  - n set ex ante; balance calculated ex post
  - EPBR controls only expected (not in-sample) imbalance
  - Methods are assumption-dependent & only potentially EPBR

Largest Class of Methods: Equal Percent Bias Reducing

- Goal: changing balance on 1 variable should not harm others
- For EPBR to be useful, it requires:
  (a) X drawn randomly from a specified population $\mathbf{X}$,
  (b) $\mathbf{X} \sim$ Normal
  (c) Matching algorithm is invariant to linear transformations of X.
  (d) Y is a linear function of X.
- EPBR Definition: Matched sample size set ex ante, and

$$\underbrace{E(\bar{\mathbf{X}}_{m_T} - \bar{\mathbf{X}}_{m_C})}^{\text{matched}} = \gamma \underbrace{E(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)}^{\text{original}}$$

- When data conditions hold:
  - Reducing mean-imbalance on one variable, reduces it on all
  - n set ex ante; balance calculated ex post
  - EPBR controls only expected (not in-sample) imbalance
  - Methods are assumption-dependent & only potentially EPBR
  - (In practice, we're lucky if univariate mean imbalance is reduced)

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types

- No restrictions on data types
- Designed for observational data (reversing EPBR):

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$$D(\mathbf{X}^{\epsilon}_{T}, \mathbf{X}^{\epsilon}_{C}) \leq \gamma(\epsilon) \qquad \textbf{vars to adjust}$$

$$D(X^{\epsilon}_{T}, X^{\epsilon}_{C}) \leq \gamma(\epsilon) \qquad \text{remaining vars}$$

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$$D(\mathbf{X}_T^{\epsilon}, \mathbf{X}_C^{\epsilon}) \leq \gamma(\epsilon) \qquad \textbf{vars to adjust}$$

$$D(X_T^{\epsilon}, X_C^{\epsilon}) \leq \gamma(\epsilon) \qquad \text{remaining vars}$$

Treated and control X variables to adjust

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$$D(\mathbf{X}^{\epsilon}_{T}, \mathbf{X}^{\epsilon}_{C}) \leq \gamma(\epsilon) \qquad \textbf{vars to adjust}$$
$$D(X^{\epsilon}_{T}, X^{\epsilon}_{C}) \leq \gamma(\epsilon) \qquad \text{remaining vars}$$

Remaining treated and control X variables

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$D(\mathbf{X}_T^\epsilon, \mathbf{X}_C^\epsilon) \leq \gamma(\epsilon)$	**vars to adjust**
$D(X_T^\epsilon, X_C^\epsilon) \leq \gamma(\epsilon)$	remaining vars

"Imbalance" given chosen distance metric

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$D(\mathbf{X}_T^{\epsilon}, \mathbf{X}_C^{\epsilon}) \leq \gamma(\boldsymbol{\epsilon})$	**vars to adjust**
$D(X_T^{\epsilon}, X_C^{\epsilon}) \leq \gamma(\epsilon)$	remaining vars

Bounds (maximum imbalance)

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$$D(\mathbf{X}_T^{\boldsymbol{\epsilon}}, \mathbf{X}_C^{\boldsymbol{\epsilon}}) \leq \gamma(\boldsymbol{\epsilon}) \qquad \textbf{vars to adjust}$$

$$D(X_T^{\epsilon}, X_C^{\epsilon}) \leq \gamma(\epsilon) \qquad \text{remaining vars}$$

One tuning parameter ϵ_j, one for each X_j

A New Class of Methods: Monotonic Imbalance Bounding

- No restrictions on data types
- Designed for observational data (reversing EPBR):
  - Most important (bias): degree of balance chosen ex ante
  - Least important (variance): matched n checked ex post
- Balance is measured in sample (like blocked designs), not merely in expectation (like complete randomization)
- Covers all forms of imbalance: means, interactions, nonlinearities, moments, multivariate histograms, etc.
- One adjustable tuning parameter per variable
- Convenient monotonicity property: Reducing maximum imbalance on one X: no effect on others

MIB Formally (simplifying for this talk):

$$D(\mathbf{X}_T^{\boldsymbol{\epsilon}}, \mathbf{X}_C^{\boldsymbol{\epsilon}}) \leq \boldsymbol{\gamma}(\boldsymbol{\epsilon}) \qquad \textbf{vars to adjust}$$

$$D(X_T^{\epsilon}, X_C^{\epsilon}) \leq \gamma(\epsilon) \qquad \text{remaining vars}$$

If $\boldsymbol{\epsilon}$ is reduced, $\boldsymbol{\gamma}(\boldsymbol{\epsilon})$ decreases & $\gamma(\epsilon)$ is unchanged

What's Coarsening?

What's Coarsening?

- Coarsening is intrinsic to measurement

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/$\{$neutral,no opinion$\}$/disagree

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers
  - Religion, Occupation, SEC industries, ICD codes, etc.

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers
  - Religion, Occupation, SEC industries, ICD codes, etc.
- Temporary Coarsening for CEM; e.g.:

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers
  - Religion, Occupation, SEC industries, ICD codes, etc.
- Temporary Coarsening for CEM; e.g.:
  - Education: grade school, middle school, high school, college, graduate

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers
  - Religion, Occupation, SEC industries, ICD codes, etc.
- Temporary Coarsening for CEM; e.g.:
  - Education: grade school, middle school, high school, college, graduate
  - Income: poverty level threshold, or larger bins for higher income

What's Coarsening?

- Coarsening is intrinsic to measurement
  - We think of measurement as clarity between categories
  - But measurement also involves homogeneity within categories
  - Examples: male/female, rich/middle/poor, black/white, war/nonwar.
  - Better measurement devices (e.g., telescopes) produce more detail
- Data analysts routinely coarsen, thinking grouping error is less risky than measurement error. E.g.:
  - 7 point Party ID $\rightsquigarrow$ Democrat/Independent/Republican
  - Likert Issue questions $\rightsquigarrow$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\rightsquigarrow$ winner/losers
  - Religion, Occupation, SEC industries, ICD codes, etc.
- Temporary Coarsening for CEM; e.g.:
  - Education: grade school, middle school, high school, college, graduate
  - Income: poverty level threshold, or larger bins for higher income
  - Age: infant, child, adolescent, young adult, middle age, elderly

CEM as an MIB Method

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for:

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles,

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set . . .

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
    - too large?

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched *n* is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set . . .
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\implies$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances
  - too small?

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances
  - too small? $\rightsquigarrow$ n may be too small

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances
  - too small? $\rightsquigarrow$ n may be too small
  - as large as you're comfortable with, but n is still too small?

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances
  - too small? $\rightsquigarrow$ n may be too small
  - as large as you're comfortable with, but n is still too small? $\rightsquigarrow$ No magic method of matching can save you;

CEM as an MIB Method

- Define: ϵ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)
- Setting ϵ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  $\Longrightarrow$ Setting ϵ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched n is determined ex post)
- By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence
- What if ϵ is set ...
  - too large? $\rightsquigarrow$ You're left modeling remaining imbalances
  - too small? $\rightsquigarrow$ n may be too small
  - as large as you're comfortable with, but n is still too small?
    $\rightsquigarrow$ No magic method of matching can save you;
    $\rightsquigarrow$ You're stuck modeling or collecting better data

Other CEM properties

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit):

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit): would bin college drop out with 1st year grad student;

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit): would bin college drop out with 1st year grad student; and not bin Bill Gates & Warren Buffett

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit): would bin college drop out with 1st year grad student; and not bin Bill Gates & Warren Buffett
- Approximate invariance to measurement error:

	CEM	pscore	Mahalanobis	Genetic
% Common Units	96.5	70.2	80.9	80.0

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit): would bin college drop out with 1st year grad student; and not bin Bill Gates & Warren Buffett
- Approximate invariance to measurement error:

	CEM	pscore	Mahalanobis	Genetic
% Common Units	96.5	70.2	80.9	80.0

- Fast and memory-efficient even for large n; can be fully automated

Other CEM properties

- Automatically eliminates extrapolation region (no separate step)
- Bounds model dependence
- Bounds causal effect estimation error
- Meets the congruence principle
  - The principle: data space = analysis space
  - Estimators that violate it are nonrobust and counterintuitive
  - CEM: ϵ_j is set using each variable's units
  - E.g., calipers (strata centered on each unit): would bin college drop out with 1st year grad student; and not bin Bill Gates & Warren Buffett
- Approximate invariance to measurement error:

	CEM	pscore	Mahalanobis	Genetic
% Common Units	96.5	70.2	80.9	80.0

- Fast and memory-efficient even for large n; can be fully automated
- Simple to teach: coarsen, then exact match

CEM in Stata – An example

```
. cem age education black nodegree re74, tr(treated)

Matching Summary:
-----------------
Number of strata: 205
Number of matched strata: 67

             0     1
      All  425   297
  Matched  324   228
Unmatched  101    69

Multivariate L1 distance: .46113967

Univariate imbalance:

               L1       mean   min   25%      50%      75%      max
      age  .13641   -.17634     0     0        0        0       -1
education  .00687    .00687     1     0        0        0        0
    black  3.2e-16  -2.2e-16    0     0        0        0        0
 nodegree  5.8e-16   4.4e-16    0     0        0        0        0
     re74  .06787    34.438     0     0   492.23   39.425   96.881
```

Imbalance Measures

Imbalance Measures

Variable-by-Variable Difference in Global Means

$$I_1^{(j)} = \left| \bar{X}_{m_T}^{(j)} - \bar{X}_{m_C}^{(j)} \right|, \quad j = 1, \ldots, k$$

Imbalance Measures

Variable-by-Variable Difference in Global Means

$$I_1^{(j)} = \left| \bar{X}_{m_T}^{(j)} - \bar{X}_{m_C}^{(j)} \right|, \quad j = 1, \ldots, k$$

Multivariate Imbalance: difference in histograms (bins fixed ex ante)

$$\mathcal{L}_1(f, g) = \sum_{\ell_1 \cdots \ell_k} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}|$$

Imbalance Measures

Variable-by-Variable Difference in Global Means

$$I_1^{(j)} = \left| \bar{X}_{m_T}^{(j)} - \bar{X}_{m_C}^{(j)} \right|, \quad j = 1, \ldots, k$$

Multivariate Imbalance: difference in histograms (bins fixed ex ante)

$$\mathcal{L}_1(f, g) = \sum_{\ell_1 \cdots \ell_k} |f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k}|$$

Local Imbalance by Variable (given strata fixed by matching method)

$$I_2^{(j)} = \frac{1}{S} \sum_{s=1}^{S} \left| \bar{X}_{m_T^s}^{(j)} - \bar{X}_{m_C^s}^{(j)} \right|, \quad j = 1, \ldots, k$$

Estimating the Causal Effect from cem output

```
. reg re78 treated [iweight=cem_weights]

      Source |       SS       df       MS              Number of obs =     552
-------------+------------------------------           F(  1,   550) =    3.15
       Model |   128314324     1   128314324           Prob > F      =  0.0766
    Residual |  2.2420e+10   550  40764521.6           R-squared     =  0.0057
-------------+------------------------------           Adj R-squared =  0.0039
       Total |  2.2549e+10   551  40923414.2           Root MSE      =  6384.7

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     treated |   979.1905   551.9132     1.77   0.077    -104.9252    2063.306
       _cons |    4919.49   354.7061    13.87   0.000     4222.745    5616.234
------------------------------------------------------------------------------
```

Choosing a custom coarsening

```
. table education

----------------------
education |      Freq.
----------+-----------
        3 |          1
        4 |          6
        5 |          5
        6 |          7
        7 |         15
        8 |         62
        9 |        110
       10 |        162
       11 |        195
       12 |        122
       13 |         23
       14 |         11
       15 |          2
       16 |          1
----------------------
```

Choosing a custom coarsening

```
. table education

----------------------
education |      Freq.
----------+-----------
        3 |          1
        4 |          6
        5 |          5
        6 |          7
        7 |         15
        8 |         62
        9 |        110
       10 |        162
       11 |        195
       12 |        122
       13 |         23
       14 |         11
       15 |          2
       16 |          1
----------------------
```

Grade school	0–6
Middle school	7–8
High school	9–12
College	13–16
Graduate school	>16

Choosing a custom coarsening

```
. table education

----------------------
education |      Freq.
----------+-----------
        3 |          1
        4 |          6
        5 |          5
        6 |          7
        7 |         15
        8 |         62
        9 |        110
       10 |        162
       11 |        195
       12 |        122
       13 |         23
       14 |         11
       15 |          2
       16 |          1
----------------------
```

Grade school	0–6
Middle school	7–8
High school	9–12
College	13–16
Graduate school	$>$16

```
. cem age education (0 6.5 8.5 12.5 17.5) black nodegree re74, tr(treated)
```

CEM Extensions I

CEM Extensions I

- CEM and Multiple Imputation for Missing Data

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
    1. put missing observation in stratum where plurality of imputations fall

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
    1. put missing observation in stratum where plurality of imputations fall
    2. pass on uncoarsened imputations to analysis stage

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T
- Blocking in Randomized Experiments: no modification needed; randomly assign T within CEM strata

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T
- Blocking in Randomized Experiments: no modification needed; randomly assign T within CEM strata
- Automating user choices

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T
- Blocking in Randomized Experiments: no modification needed; randomly assign T within CEM strata
- Automating user choices Histogram bin size calculations

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T
- Blocking in Randomized Experiments: no modification needed; randomly assign T within CEM strata
- Automating user choices Histogram bin size calculations
- Improve Existing Matching Methods

CEM Extensions I

- CEM and Multiple Imputation for Missing Data
  1. put missing observation in stratum where plurality of imputations fall
  2. pass on uncoarsened imputations to analysis stage
  3. Use the usual MI combining rules to analyze
- Multicategory treatments: No modification necessary; keep all strata with ≥ 1 unit having each value of T
- Blocking in Randomized Experiments: no modification needed; randomly assign T within CEM strata
- Automating user choices Histogram bin size calculations
- Improve Existing Matching Methods Applying other methods within CEM strata

For papers, software, tutorials, etc.

http://GKing.Harvard.edu/cem