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)
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 inherit CEM’s properties)

⇝ A version of CEM: Last studied 40 years ago by Cochran
⇝ First used many decades before that

Matthew Blackwell (Harvard, IQSS)
Coarsened Exact Matching (CEM)

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

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 inherit CEM's properties

A version of CEM: Last studied 40 years ago by Cochran
First used many decades before that

Matthew Blackwell (Harvard, IQSS)
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 inherit CEM's properties.

\(\Rightarrow\) A version of CEM: Last studied 40 years ago by Cochran
\(\Rightarrow\) First used many decades before that.

Matthew Blackwell (Harvard, IQSS)
Matching without Balance Checking
Preprocess $(X, T)$ with CEM:

1. Temporarily coarsen $X$ as much as you’re willing
Preprocess \((X, T)\) with CEM:

1. **Temporarily coarsen** \(X\) as much as you’re willing
   - e.g., Education (grade school, high school, college, graduate)
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
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 \(X, C(X)\)
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 \(C(X)\)
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
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
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)
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 inherit CEM’s 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 inherit CEM’s properties)

\[ \rightarrow \] A version of CEM: Last studied 40 years ago by Cochran
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)

\[\Rightarrow\] A version of CEM: Last studied 40 years ago by Cochran

\[\Rightarrow\] First used many decades before that
Characteristics of Observational Data

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

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.)
Characteristics of Observational Data

- Lots of data

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

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.)
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

What to do?

Preprocess I: Eliminate extrapolation region (a separate step)

Preprocess II: Match (prune bad matches) within interpolation region

Model remaining imbalance
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
What to do?

Preprocess I: Eliminate extrapolation region (a separate step)

Preprocess II: Match (prune bad matches) within interpolation region

Model remaining imbalance

Matching without Balance Checking
What to do?
What to do?

- Preprocess I: Eliminate extrapolation region (a separate step)
What to do?

- Preprocess I: Eliminate extrapolation region (a separate step)
- Preprocess II: Match (prune bad matches) within interpolation region region
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 reduces model dependence, bias, and variance.

Matthew Blackwell (Harvard, IQSS)
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)

Before Matching

After Matching

Matching reduces model dependence, bias, and variance
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

Notation:

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

Treatment Effect for treated ($T_i = 1$) observation:

$$TE_i = Y_i(T_i = 1) - Y_i(T_i = 0) = \text{observed} - \text{unobserved}$$

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:

$$\text{SATT} = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i$$
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:

  $$TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)$$

  Estimated $Y_i(T_i = 0)$ 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:

  $$SATT = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i$$

Matthew Blackwell (Harvard, IQSS)

Matching without Balance Checking
The Goals, with some more precision

- Notation:
  - $Y_i$: Dependent variable

- Treatment Effect for treated ($T_i = 1$) observation:
  $$TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)$$

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:
$$SATT = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i$$
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

Treatment Effect for treated ($T_i = 1$) observation:

$$TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)$$

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:

$$SATT = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i$$

Matthew Blackwell (Harvard, IQSS)
Matching without Balance Checking
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 $i$:**

$$TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)$$
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 $i$:
  \[ \text{TE}_i = Y_i(T_i = 1) - Y_i(T_i = 0) \]
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 $i$:

  $$ TE_i = Y_i(T_i = 1) - Y_i(T_i = 0) $$
  $$ = \text{observed} - \text{unobserved} $$
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 $i$:**
  \[ TE_i = Y_i(T_i = 1) - Y_i(T_i = 0) = \text{observed} - \text{unobserved} \]

- **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 $i$:
  \[
  TE_i = Y_i(T_i = 1) - Y_i(T_i = 0) = \text{observed} - \text{unobserved}
  \]

- 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 $i$:**

  \[
  TE_i = Y_i(T_i = 1) - Y_i(T_i = 0)
  \]

  \[
  = \text{observed} - \text{unobserved}
  \]

- **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:**

  \[
  \text{SATT} = \frac{1}{n_T} \sum_{i \in \{T_i = 1\}} TE_i
  \]
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:

\begin{itemize}
\item choose \( n \)
\item match
\item check,
\item tweak
\item match
\item check,
\item tweak
\item match
\item check,
\item tweak
\item···
\end{itemize}

Actual practice:

\begin{itemize}
\item choose \( n \),
\item match,
\item publish,
\item STOP.
\end{itemize}

(Is balance even improved?)
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
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
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, \( \cdots \)
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, ···
  - 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, ···
  - Actual practice: choose $n$, 

Matthew Blackwell (Harvard, IQSS)
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, …
  - 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, ···
  - 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, ···
  - 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, ···
  - Actual practice: choose $n$, match, publish, STOP. (Is balance even improved?)
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 $X$,
(b) $X \sim \text{Normal}$
(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 matched original

$$\mathbb{E}(\bar{X}_m^T - \bar{X}_m^C) = \gamma \mathbb{E}(\bar{X}_T - \bar{X}_C)$$

When data conditions hold:

- Reducing mean-imbalance on one variable, reduces it on all
- 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)

Matthew Blackwell (Harvard, IQSS)
Matching without Balance Checking
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:

\[ X \text{ drawn randomly from a specified population } X, \]
\[ X \sim \text{Normal} \]
\[ \text{Matching algorithm is invariant to linear transformations of } X. \]
\[ Y \text{ is a linear function of } X. \]

EPBR Definition: Matched sample size set ex ante, and

\[ \text{balance calculated ex post} \]

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)
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 $X$

(b) $X \sim \text{Normal}$

(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 matched original

$$E(\bar{X}_m - \bar{X}_m) = \gamma E(\bar{X}_T - \bar{X}_C)$$

When data conditions hold:

Reducing mean-imbalance on one variable, reduces it on all

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)
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:
  1. \( X \) drawn randomly from a specified population \( X \),
  2. \( X \sim \text{Normal} \)

\[
\text{EPBR Definition: Matched sample size set ex ante, and matched original}
\]
\[
\mathbb{E}(\bar{X}_m^T - \bar{X}_m^C) = \gamma \mathbb{E}(\bar{X}_T - \bar{X}_C)
\]

When data conditions hold:
- Reducing mean-imbalance on one variable, reduces it on all
- 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)
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 $X$,
(b) $X \sim \text{Normal}$
(c) Matching algorithm is invariant to linear transformations of $X$. 

$E(\bar{X}_m^T - \bar{X}_m^C) = \gamma E(\bar{X}_T - \bar{X}_C)$
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 $X$,
  - (b) $X \sim \text{Normal}$
  - (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 matched original

$$E(\bar{X}_m^T - \bar{X}_m^C) = \gamma E(\bar{X}_T - \bar{X}_C)$$

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)
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
  \[
  E(\bar{X}_{mT} - \bar{X}_{mC}) = \gamma E(\bar{X}_T - \bar{X}_C)
  \]
  When data conditions hold:
  Reducing mean-imbalance on one variable, reduces it on all.
  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)
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 $X$,  
  - (b) $X \sim$ 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
  
  
  \[ E(\tilde{X}_{mT} - \tilde{X}_{mC}) = \gamma E(\tilde{X}_{T} - \tilde{X}_{C}) \]

- **When data conditions hold**:
• **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) $X \sim \text{Normal}$
  (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

\[
E(\bar{X}_{mT} - \bar{X}_{mC}) = \gamma E(\bar{X}_T - \bar{X}_C)
\]

• When data conditions hold:
  • Reducing mean-imbalance on one variable, reduces it on all
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 \( X \),
(b) \( X \sim \text{Normal} \)
(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

\[
E(\bar{X}_{mT} - \bar{X}_{mC}) = \gamma E(\bar{X}_T - \bar{X}_C)
\]

When data conditions hold:
- Reducing mean-imbalance on one variable, reduces it on all
- \( n \) set ex ante; balance calculated ex post
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 $X$,
  (b) $X \sim \text{Normal}$
  (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

\[
E(\bar{X}_{mT} - \bar{X}_{mC}) = \gamma E(\bar{X}_T - \bar{X}_C)
\]

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:
  1. $X$ drawn randomly from a specified population $X$,
  2. $X \sim \text{Normal}$
  3. Matching algorithm is invariant to linear transformations of $X$.
  4. $Y$ is a linear function of $X$.
- **EPBR Definition:** Matched sample size set ex ante, and
  \[
  E(\tilde{X}_{mT} - \tilde{X}_{mC}) = \gamma E(\tilde{X}_{T} - \tilde{X}_{C})
  \]
- 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

Matthew Blackwell (Harvard, IQSS)
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) $X \sim \text{Normal}$
(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

$$E(\bar{X}_{m_T} - \bar{X}_{m_C}) = \gamma E(\bar{X}_T - \bar{X}_C)$$

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

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):

$$D(X \in T, X \in C) \leq \gamma(\epsilon)$$

vars to adjust

$$D(X \in T, X \in C) \leq \gamma(\epsilon)$$

remaining vars
A New Class of Methods: Monotonic Imbalance Bounding

- 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
- 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
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 \textit{ex ante}
  - Least important (variance): matched \textit{n} checked \textit{ex post}
- Balance is measured \textit{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):

\[
D(X^e_T, X^e_C) \leq \gamma(\epsilon) \quad \text{vars to adjust}
\]
\[
D(X^e_T, X^e_C) \leq \gamma(\epsilon) \quad \text{remaining vars}
\]
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):

$$D(X^e_T, X^e_C) \leq \gamma(\epsilon)$$  vars to adjust

$$D(X^e_T, X^e_C) \leq \gamma(\epsilon)$$  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):

\[
D(X^{\varepsilon}_T, X^{\varepsilon}_C) \leq \gamma(\varepsilon) \quad \text{vars to adjust}
\]

\[
D(X^{\varepsilon}_T, X^{\varepsilon}_C) \leq \gamma(\varepsilon) \quad \text{remaining vars}
\]

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):

$$D(X^e_T, X^e_C) \leq \gamma(\epsilon)$$  
$$D(X^e_T, X^e_C) \leq \gamma(\epsilon)$$

“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):

$$D(X^\epsilon_T, X^\epsilon_C) \leq \gamma(\epsilon)$$

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 \textit{ex ante}
  - Least important (variance): matched \( n \) checked \textit{ex post}
- Balance is measured \textit{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):

\[
D(X^\epsilon_T, X^\epsilon_C) \leq \gamma(\epsilon) \quad \text{vars to adjust}
\]

\[
D(X^\epsilon_T, X^\epsilon_C) \leq \gamma(\epsilon) \quad \text{remaining vars}
\]

One tuning parameter \( \epsilon_j \), 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):

\[
D(X_T^\epsilon, X_C^\epsilon) \leq \gamma(\epsilon) \quad \text{vars to adjust}
\]
\[
D(X_T^\epsilon, X_C^\epsilon) \leq \gamma(\epsilon) \quad \text{remaining vars}
\]

If \( \epsilon \) is reduced, \( \gamma(\epsilon) \) decreases & \( \gamma(\epsilon) \) is unchanged
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 → Democrat/Independent/Republican
- Likert Issue questions → agree/neutral/no opinion/disagree
- Multiparty voting → 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

Matthew Blackwell (Harvard, IQSS)
What’s Coarsening?

- Coarsening is **intrinsic to measurement**

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**
  - Democrat/Independent/Republican
- **Likert Issue questions**
  - agree/neutral/no opinion/disagree
- **multiparty voting**
  - 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
What’s Coarsening?

- Coarsening is intrinsic to measurement
- We think of measurement as clarity between 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  
  \[\rightarrow\]  Democrat/Independent/Republican

- Likert Issue questions  
  \[\rightarrow\]  agree/\{neutral, no opinion\}/disagree

- Multiparty voting  
  \[\rightarrow\]  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
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 $\mapsto$ Democrat/Independent/Republican
- Likert Issue questions $\mapsto$ agree/neutral/no opinion/disagree
- Multiparty voting $\mapsto$ 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
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.
Coarsening is intrinsic to measurement

1. We think of measurement as clarity between categories
2. But measurement also involves homogeneity within categories
3. Examples: male/female, rich/middle/poor, black/white, war/nonwar.
4. Better measurement devices (e.g., telescopes) produce more detail
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
  \[\rightarrow\] Democrat/Independent/Republican

- Likert Issue questions
  \[\rightarrow\] agree/\{neutral, no opinion\}/disagree

- Multiparty voting
  \[\rightarrow\] 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
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
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 $\leadsto$ Democrat/Independent/Republican
- Likert Issue questions $\leadsto$ agree/{neutral,no opinion}/disagree
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 $\rightarrow$ Democrat/Independent/Republican
- Likert Issue questions $\rightarrow$ agree/{neutral,no opinion}/disagree
- multiparty voting $\rightarrow$ 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 $\mapsto$ Democrat/Independent/Republican
  - Likert Issue questions $\mapsto$ agree/{neutral,no opinion}/disagree
  - multiparty voting $\mapsto$ 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.:
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 $\leadsto$ Democrat/Independent/Republican
  - Likert Issue questions $\leadsto$ agree/\{neutral,no opinion\}/disagree
  - multiparty voting $\leadsto$ 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 $\mapsto$ Democrat/Independent/Republican
  - Likert Issue questions $\mapsto$ agree/\{neutral,no opinion\}/disagree
  - multiparty voting $\mapsto$ 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
Define: \( \epsilon \) as largest (coarsened) bin size (\( \epsilon = 0 \) is exact matching). Setting \( \epsilon \) bounds the treated-control group difference, within strata and globally, for:

- means,
- variances,
- skewness,
- covariances,
- comoments,
- coskewness,
- co-kurtosis,
- quantiles,
- and full multivariate histogram.

\( \Rightarrow \) Setting \( \epsilon \) 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 \( \epsilon \) is set too large?

\( \Rightarrow \) You're left modeling remaining imbalances too small?

\( \Rightarrow \) \( n \) may be too small as large as you're comfortable with, but \( n \) is still too small?

\( \Rightarrow \) No magic method of matching can save you; \( \Rightarrow \) You're stuck modeling or collecting better data.
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ bounds the treated-control group difference, within strata and globally, for:
- means,
- variances,
- skewness,
- covariances,
- comoments,
- coskewness,
- co-kurtosis,
- quantiles,
- and full multivariate histogram.

$\epsilon \Rightarrow$ Setting $\epsilon$ 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 $\epsilon$ is set too large?

$\Rightarrow$ You're left modeling remaining imbalances too small?

$\Rightarrow$ $n$ may be too small as large as you're comfortable with, but $n$ is still too small?

$\Rightarrow$ No magic method of matching can save you;

$\Rightarrow$ You're stuck modeling or collecting better data.
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ bounds the treated-control group difference, within strata and globally, for:

- means,
- variances,
- skewness,
- covariances,
- comoments,
- coskewness,
- co-kurtosis,
- quantiles,
- and full multivariate histogram.

Setting $\epsilon$ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level ($\epsilon$ 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 $\epsilon$ is set too large?

$\Rightarrow$ You’re left modeling remaining imbalances

What if $\epsilon$ is set too small?

$\Rightarrow$ $n$ may be too small

What if $n$ is still too small?

$\Rightarrow$ No magic method of matching can save you;

$\Rightarrow$ You’re stuck modeling or collecting better data
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ bounds the treated-control group difference, within strata and globally, for: means,
CEM as an MIB Method

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

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

- Setting $\epsilon$ bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances,
- = $\Rightarrow$
- Setting $\epsilon$ controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level ($n$ 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 $\epsilon$ is set . . .

Too large? $\Rightarrow$ You're left modeling remaining imbalances
Too small? $\Rightarrow$ $n$ may be too small as large as you're comfortable with, but $n$ is still too small?
$\Rightarrow$ No magic method of matching can save you; $\Rightarrow$ You're stuck modeling or collecting better data
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

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

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

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

Setting \( \epsilon \) bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles,

What if \( \epsilon \) is set too large?

You're left modeling remaining imbalances too small?

\( n \) may be too small as large as you're comfortable with, but \( n \) is still too small?

No magic method of matching can save you; you're stuck modeling or collecting better data
CEM as an MIB Method

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

What if $\epsilon$ is set too large? 
You're left modeling remaining imbalances too small? 
$n$ may be too small as large as you're comfortable with, but $n$ is still too small? 
No magic method of matching can save you; 
You're stuck modeling or collecting better data.
CEM as an MIB Method

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

Setting \( \epsilon \) bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.

\[ \Rightarrow \] Setting \( \epsilon \) controls all multivariate treatment-control differences, interactions, and nonlinearities, up to the chosen level (matched \( n \) is determined \textit{ex post})

By default, both treated and control units are pruned: CEM estimates a quantity that can be estimated without model dependence

What if \( \epsilon \) is set . . .
- too large?
Define: \( \epsilon \) as largest (coarsened) bin size (\( \epsilon = 0 \) is exact matching)

Setting \( \epsilon \) 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 \( \epsilon \) 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 \( \epsilon \) is set . . .

- too large? \( \rightsquigarrow \) You’re left modeling remaining imbalances
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ is set... 
- too large? $\rightsquigarrow$ You’re left modeling remaining imbalances
- too small?
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ is set . . .

- too large? $\implies$ You’re left modeling remaining imbalances
- too small? $\implies$ $n$ may be too small
CEM as an MIB Method

- Define: \( \epsilon \) as largest (coarsened) bin size (\( \epsilon = 0 \) is exact matching)
- Setting \( \epsilon \) bounds the treated-control group difference, within strata and globally, for: means, variances, skewness, covariances, comoments, coskewness, co-kurtosis, quantiles, and full multivariate histogram.
  \[ \Rightarrow \] Setting \( \epsilon \) 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 \( \epsilon \) is set ...
  - too large? \( \leadsto \) You’re left modeling remaining imbalances
  - too small? \( \leadsto \) \( n \) may be too small
  - as large as you’re comfortable with, but \( n \) is still too small?
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ is set . . .

- too large? $\leadsto$ You’re left modeling remaining imbalances
- too small? $\leadsto$ $n$ may be too small
- as large as you’re comfortable with, but $n$ is still too small?
  $\leadsto$ No magic method of matching can save you;
Define: $\epsilon$ as largest (coarsened) bin size ($\epsilon = 0$ is exact matching)

Setting $\epsilon$ 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 $\epsilon$ 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 $\epsilon$ is set ... 

- too large? $\leadsto$ You’re left modeling remaining imbalances
- too small? $\leadsto$ $n$ may be too small
- as large as you’re comfortable with, but $n$ is still too small?
  $\leadsto$ No magic method of matching can save you;
  $\leadsto$ You’re stuck modeling or collecting better data
Other CEM properties

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

CEM: $\epsilon_j$ 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:

<table>
<thead>
<tr>
<th>Method</th>
<th>Common Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEM pscore</td>
<td>96.5</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>70.2</td>
</tr>
<tr>
<td>Genetic</td>
<td>80.9</td>
</tr>
</tbody>
</table>
- Fast and memory-efficient even for large $n$; can be fully automated
- Simple to teach: coarsen, then exact match
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

Estimators that violate it are nonrobust and counterintuitive

CEM: $\epsilon_j$ 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
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 \((\text{no separate step})\)
- 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: \(\epsilon_j\) is set using each variable’s units

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
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: $\epsilon_j$ 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: $\epsilon_j$ 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: $\epsilon_j$ 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: $\epsilon_j$ 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:

<table>
<thead>
<tr>
<th></th>
<th>CEM</th>
<th>pscore</th>
<th>Mahalanobis</th>
<th>Genetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Common Units</td>
<td>96.5</td>
<td>70.2</td>
<td>80.9</td>
<td>80.0</td>
</tr>
</tbody>
</table>
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: $\epsilon_j$ 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:

<table>
<thead>
<tr>
<th></th>
<th>CEM</th>
<th>pscore</th>
<th>Mahalanobis</th>
<th>Genetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Common Units</td>
<td>96.5</td>
<td>70.2</td>
<td>80.9</td>
<td>80.0</td>
</tr>
</tbody>
</table>

- 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: \( \epsilon_j \) 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:
  
<table>
<thead>
<tr>
<th>Method</th>
<th>CEM</th>
<th>pscore</th>
<th>Mahalanobis</th>
<th>Genetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Common Units</td>
<td>96.5</td>
<td>70.2</td>
<td>80.9</td>
<td>80.0</td>
</tr>
</tbody>
</table>

- Fast and memory-efficient even for large \( n \); can be fully automated
- Simple to teach: coarsen, then exact match
. cem age education black nodegree re74, tr(treated)

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

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>425</td>
<td>297</td>
</tr>
<tr>
<td>Matched</td>
<td>324</td>
<td>228</td>
</tr>
<tr>
<td>Unmatched</td>
<td>101</td>
<td>69</td>
</tr>
</tbody>
</table>

Multivariate L1 distance: .46113967

Univariate imbalance:

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>mean</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>.13641</td>
<td>-.17634</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>education</td>
<td>.00687</td>
<td>.00687</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>black</td>
<td>3.2e-16</td>
<td>-2.2e-16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nodegree</td>
<td>5.8e-16</td>
<td>4.4e-16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>re74</td>
<td>.06787</td>
<td>34.438</td>
<td>0</td>
<td>0</td>
<td>492.23</td>
<td>39.425</td>
<td>96.881</td>
</tr>
</tbody>
</table>
Imbalance Measures

Variable-by-Variable Difference in Global Means

$$I_j = \left| \bar{X}_j^T - \bar{X}_j^C \right|, j = 1, \ldots, k$$

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

$$L_1 (f, g) = \sum_{\ell_1}^{\ell_k} |f_{\ell_1} \cdots f_{\ell_k} - g_{\ell_1} \cdots g_{\ell_k}|$$

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

$$I_j = \sum_{s=1}^{S} \left| \bar{X}_j^{m_s} - \bar{X}_j^{m_s} \right|, j = 1, \ldots, k$$
Imbalance Measures

Variable-by-Variable Difference in Global Means

\[ \ell_1^{(j)} = \left| \bar{X}_{mT}^{(j)} - \bar{X}_{mC}^{(j)} \right|, \quad j = 1, \ldots, k \]
Imbalance Measures

Variable-by-Variable Difference in Global Means

\[ l_1^{(j)} = \left| \bar{X}_{mT}^{(j)} - \bar{X}_{mC}^{(j)} \right|, \quad j = 1, \ldots, k \]

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

\[ L_1(f, g) = \sum_{\ell_1 \cdots \ell_k} \left| f_{\ell_1 \cdots \ell_k} - g_{\ell_1 \cdots \ell_k} \right| \]
Imbalance Measures

Variable-by-Variable Difference in Global Means

\[ l_1^{(j)} = \left| \bar{X}_{mT}^{(j)} - \bar{X}_{mC}^{(j)} \right|, \quad j = 1, \ldots, k \]

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)

\[ l_2^{(j)} = \frac{1}{S} \sum_{s=1}^{S} \left| \bar{X}_{mT}^{(j)} - \bar{X}_{mC}^{(j)} \right|, \quad j = 1, \ldots, k \]
### Estimating the Causal Effect from `cem` output

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>128314324</td>
<td>1</td>
<td>128314324</td>
</tr>
<tr>
<td>Residual</td>
<td>2.2420e+10</td>
<td>550</td>
<td>40764521.6</td>
</tr>
<tr>
<td>Total</td>
<td>2.2549e+10</td>
<td>551</td>
<td>40923414.2</td>
</tr>
</tbody>
</table>

|       | Number of obs = 552 | F( 1, 550) = 3.15 | Prob > F = 0.0766 | R-squared = 0.0057 | Adj R-squared = 0.0039 | 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 to 2063.306 |
| _cons | 4919.49  | 354.7061  | 13.87 | 0.000 | 4222.745 to 5616.234 |
Choosing a custom coarsening

```
. table education

<table>
<thead>
<tr>
<th>education</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>110</td>
</tr>
<tr>
<td>10</td>
<td>162</td>
</tr>
<tr>
<td>11</td>
<td>195</td>
</tr>
<tr>
<td>12</td>
<td>122</td>
</tr>
<tr>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
</tr>
</tbody>
</table>
```

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)

Matthew Blackwell (Harvard, IQSS)
Matching without Balance Checking
### Choosing a custom coarsening

#### . table education

<table>
<thead>
<tr>
<th>education</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>9</td>
<td>110</td>
</tr>
<tr>
<td>10</td>
<td>162</td>
</tr>
<tr>
<td>11</td>
<td>195</td>
</tr>
<tr>
<td>12</td>
<td>122</td>
</tr>
<tr>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
</tr>
</tbody>
</table>

#### 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)
```

Matthew Blackwell (Harvard, IQSS)  Matching without Balance Checking
CEM Extensions I

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 \( \geq 1 \) 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
CEM Extensions I

- CEM and Multiple Imputation for Missing Data

- Histogram bin size calculations

- Improve Existing Matching Methods
  - Applying other methods within CEM strata

- Automating user choices
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 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 \( \geq 1 \) 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 $\geq 1$ 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 $\geq 1$ unit having each value of $T$

- **Blocking in Randomized Experiments**: no modification needed; randomly assign $T$ within CEM strata

- Automating user choices
- **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 $\geq 1$ 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 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 $\geq 1$ 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 \( \geq 1 \) 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
http://GKing.Harvard.edu/cem