CEM: A Matching Method for Observational Data in the Social Sciences

S.M. Iacus (Univ. of Milan) & G. King (Harvard Univ.) & G. Porro (Univ. of Trieste)

Rennes, useR! 2009, July 8th - 10th
We consider an observational study with \( n \) observations. For each unit \( i \)

\[
Y_i = \text{outcome} \quad T_i = \text{treatment indicator} \quad X_i = \text{covariates}
\]

**ESTIMATION GOAL:** the treatment effect

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

but \( Y_i(0) \) is not observed. For the treated unit \( i \) with covariates \( X_i \), it is natural to look for another unit \( j \) in the sample for which \( Y_j(0) \) is observed and such that \( X_j \sim X_i \)

**MATCHING GOAL:** for each treated unit \( i \) find the “twin” control unit \( j \) (i.e. with \( X_j \sim X_i \)) in order to reduce bias in the estimation of \( \text{TE}_i \)
Matching solutions in R (incomplete list)

- **MatchIt**: (pscore, mahalanobis, etc)
- **Matching**: (genetic matching, pscore, etc)
- **optmatch**: (full optimal matching)
- **rrp**: (random recursive partitioning)
- **arm**: (single nearest neighbour)
- **SpectralGEM**: (spectral graph theory)
- **analogue**: (analogue matching, nearest neighbour)
- **PSAgraphics** (diagnostic)
- **RItools** (diagnostic)
Coarsened Exact Matching (CEM), is a simple (and ancient) method of causal inference, with unexplored powerful properties. CEM is as simple as
Coarsened Exact Matching (CEM), is a simple (and ancient) method of causal inference, with unexplored powerful properties. CEM is as simple as

1. Temporarily *coarsen* $X$ as much as you're willing (e.g., for education: grade school, high school, college, graduate);
Coarsened Exact Matching (CEM), is a simple (and ancient) method of causal inference, with unexplored powerful properties. CEM is as simple as

1. Temporarily \textit{coarsen} $X$ as much as you’re willing (e.g., for education: grade school, high school, college, graduate);

2. Perform \textit{exact matching} on the coarsened data $C(X)$, sort observations into strata and prune any stratum with 0 treated or 0 control units, i.e. set weight=0 for pruned observations and CEM weights to matched units;
Coarsened Exact Matching (CEM), is a simple (and ancient) method of causal inference, with unexplored powerful properties. CEM is as simple as

1. Temporarily *coarsen* $X$ as much as you’re willing (e.g., for education: grade school, high school, college, graduate);

2. Perform *exact matching* on the coarsened data $C(X)$, sort observations into strata and prune any stratum with 0 treated or 0 control units, i.e. set weight=0 for pruned observations and CEM weights to matched units;

3. use the **original uncoarsened** data $X$ (with appropriate weights) in your analysis, except those units pruned.

Maximum imbalance is controlled ex-ante by the choice of coarsening
COARSEN THE DATA X INTO C(X)

DO EXACT MATCHING ON COARSENE Data C(X)

pass original uncoarsened data X to the analysis stage

CEM weights

THE ANALYSIS STAGE

lm
glm
randomForest
coxph
etc
cem offers standard 1-dim as well as a new multidimensional measure of imbalance $L_1 \in [0, 1]$: the distance between multidimensional histograms of the distributions of treated and control units.

```
R> library(cem)
R> data(LL) # The Lalonde(1986) benchmark data
R> # initial imbalance
R> imb <- imbalance(LL$treated,LL,drop=c("re78","treated"))
R> imb

Multivariate Imbalance Measure: L1=0.735
Percentage of local common support: LCS=17.8%

Univariate Imbalance Measures:

<table>
<thead>
<tr>
<th>statistic</th>
<th>type</th>
<th></th>
<th>L1</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></td>
<td></td>
<td>1.792038e-01</td>
<td>4.705882e-03</td>
<td>0</td>
<td>1</td>
<td>0.0000</td>
<td>-1.0000</td>
</tr>
<tr>
<td>education</td>
<td></td>
<td></td>
<td>1.922361e-01</td>
<td>9.811844e-02</td>
<td>1</td>
<td>0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>black</td>
<td></td>
<td></td>
<td>1.346801e-03</td>
<td>1.346801e-03</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>married</td>
<td></td>
<td></td>
<td>1.070311e-02</td>
<td>1.070311e-02</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>nodegree</td>
<td></td>
<td></td>
<td>-8.347792e-02</td>
<td>8.347792e-02</td>
<td>0</td>
<td>-1</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>re74</td>
<td></td>
<td></td>
<td>-1.014862e+02</td>
<td>5.551115e-17</td>
<td>0</td>
<td>0</td>
<td>69.73096</td>
<td>584.9160</td>
</tr>
<tr>
<td>re75</td>
<td></td>
<td></td>
<td>3.941545e+01</td>
<td>5.551115e-17</td>
<td>0</td>
<td>0</td>
<td>294.18457</td>
<td>660.6865</td>
</tr>
<tr>
<td>hispanic</td>
<td></td>
<td></td>
<td>-1.866508e-02</td>
<td>1.866508e-02</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>u74</td>
<td></td>
<td></td>
<td>-2.009903e-02</td>
<td>2.009903e-02</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>u75</td>
<td></td>
<td></td>
<td>-4.508616e-02</td>
<td>4.508616e-02</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
```
After matching with CEM

R> mat <- cem("treated", LL, drop="re78",L1.breaks=imb$L1$breaks)
R> mat
   G0  G1
All  425  297
Matched 222 163
Unmatched 203 134

Multivariate Imbalance Measure: L1=0.432
Percentage of local common support: LCS=44.7%

Univariate Imbalance Measures:

<table>
<thead>
<tr>
<th>statistic</th>
<th>type</th>
<th>L1</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td></td>
<td>5.551115e-17</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>1.000</td>
</tr>
<tr>
<td>education</td>
<td></td>
<td>1.022495e-02</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>black</td>
<td></td>
<td>-1.110223e-16</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>married</td>
<td></td>
<td>0.000000e+00</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>nodegree</td>
<td></td>
<td>-1.110223e-16</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>re74</td>
<td></td>
<td>7.197514e+00</td>
<td>0</td>
<td>0</td>
<td>-70.85522</td>
<td>416.416</td>
</tr>
<tr>
<td>re75</td>
<td></td>
<td>1.220698e+01</td>
<td>0</td>
<td>234.4843</td>
<td>140.79126</td>
<td>-852.252</td>
</tr>
<tr>
<td>hispanic</td>
<td></td>
<td>0.000000e+00</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>u74</td>
<td></td>
<td>0.000000e+00</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>u75</td>
<td></td>
<td>0.000000e+00</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
The choice of coarsening affects the matching solution. Due to high computationally efficiency of \texttt{cem}, the function \texttt{relax.cem} allows for automatic coarsening relaxations.

\begin{verbatim}
R> relax.cem(mat,LL)
Executing 42 different relaxations

Pre-relax: 163 matched (54.9 %)
\end{verbatim}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{chart.png}
\caption{Pre-relax: 163 matched (54.9 %)}
\end{figure}
ATT estimation on the matched data only

R> att(mat, re78 ~ treated, LL) -> TE
R> TE

<table>
<thead>
<tr>
<th></th>
<th>G0</th>
<th>G1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>425</td>
<td>297</td>
</tr>
<tr>
<td>Matched</td>
<td>222</td>
<td>163</td>
</tr>
<tr>
<td>Unmatched</td>
<td>203</td>
<td>134</td>
</tr>
</tbody>
</table>

Linear regression model on CEM matched data:

SATT point estimate: 550.962564 (p.value=0.368242)
95% conf. interval: [-647.777701, 1749.702830]

ATT estimation on all treated observations via extrapolation

R> att(mat, re78 ~ treated, LL, extrapolate=TRUE)

<table>
<thead>
<tr>
<th></th>
<th>G0</th>
<th>G1</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>425</td>
<td>297</td>
</tr>
<tr>
<td>Matched</td>
<td>222</td>
<td>163</td>
</tr>
<tr>
<td>Unmatched</td>
<td>203</td>
<td>134</td>
</tr>
</tbody>
</table>

Linear regression model with extrapolation:

SATT point estimate: 1290.247549 (p.value=0.062168)
95% conf. interval: [391.886467, 2188.608631]

The distribution of the treatment effect accross CEM strata can be further visualized

R> plot(TE,mat,LL,vars=c("re75","re74","education","age","hispanic"))
ATT estimation and visualization

Linear regression model on CEM matched data

- Negative
- Zero
- Positive
For the latest version of the manuscript, R and Stata software, visit

http://GKing.Harvard.edu/cem