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

The problem of matching

We consider an observational study with n observations. For each unit i

$$Y_i = \text{outcome} \quad T_i = \text{treatment indicator} \quad \underline{X}_i = \text{covariates}$$

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 $\underline{X}_j \simeq \underline{X}_i$

MATCHING GOAL: for each treated unit i find the "twin" control unit j (i.e. with $\underline{X}_j \simeq \underline{X}_i$) in order to reduce bias in the estimation of 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` (diagnotic)
- `RItools` (diagnostic)

CEM Overview

Coarsened Exact Matching (CEM), is a simple (and ancient) method of causal inference, with unexplored powerful properties. CEM is as simple as

CEM Overview

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);

CEM Overview

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;

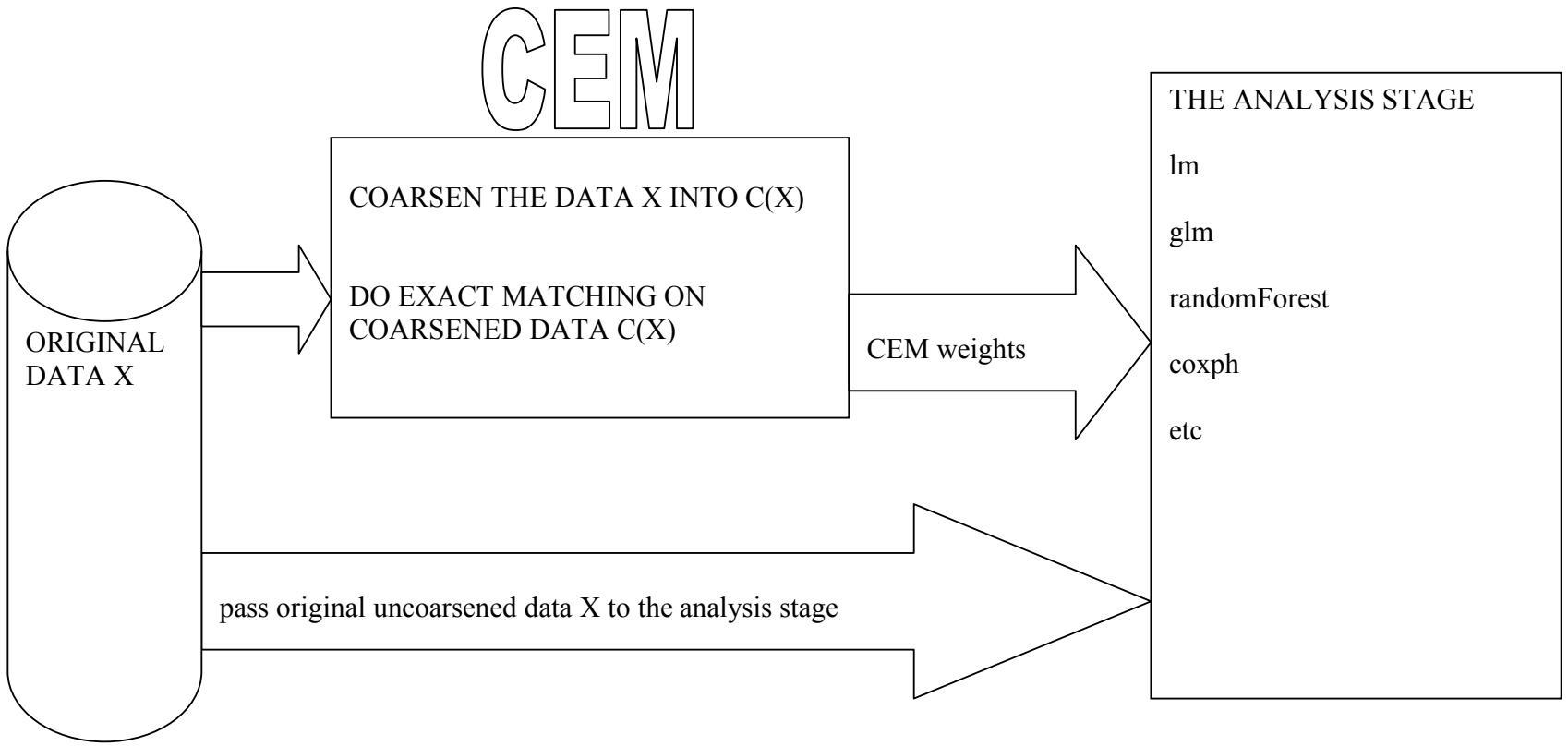
CEM Overview

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

CEM Overview



CEM package

`cem` offers standard 1-dim as well as a new multidimensional measure of imbalance $\mathcal{L}_1 \in [0, 1]$: 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:

             statistic   type           L1 min 25%       50%      75%        max
age        1.792038e-01 (diff) 4.705882e-03   0   1   0.00000  -1.0000    -6.0000
education  1.922361e-01 (diff) 9.811844e-02   1   0   1.00000   1.0000     2.0000
black      1.346801e-03 (diff) 1.346801e-03   0   0   0.00000   0.0000     0.0000
married    1.070311e-02 (diff) 1.070311e-02   0   0   0.00000   0.0000     0.0000
nodegree  -8.347792e-02 (diff) 8.347792e-02   0  -1   0.00000   0.0000     0.0000
re74      -1.014862e+02 (diff) 5.551115e-17   0   0  69.73096 584.9160 -2139.0195
re75       3.941545e+01 (diff) 5.551115e-17   0   0 294.18457 660.6865   490.3945
hispanic  -1.866508e-02 (diff) 1.866508e-02   0   0   0.00000   0.0000     0.0000
u74       -2.009903e-02 (diff) 2.009903e-02   0   0   0.00000   0.0000     0.0000
u75       -4.508616e-02 (diff) 4.508616e-02   0   0   0.00000   0.0000     0.0000
```

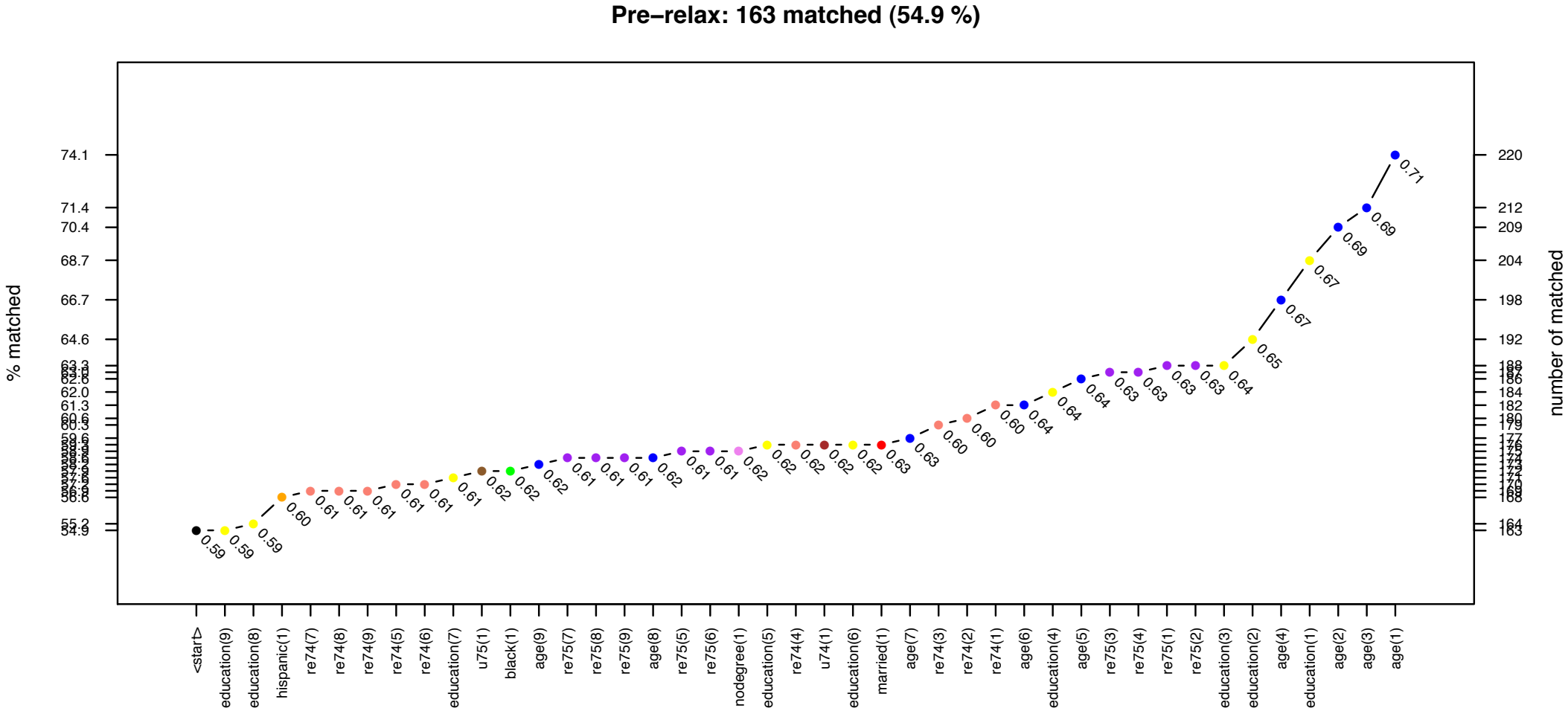
After matching with CEM

```
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:

              statistic   type           L1 min 25%      50%       75%      max
age        1.862046e-01 (diff) 5.551115e-17   0   0   0.0000   1.00000    1.000
education  1.022495e-02 (diff) 1.022495e-02   0   0   0.0000   0.00000    0.000
black     -1.110223e-16 (diff) 6.245005e-17   0   0   0.0000   0.00000    0.000
married    0.000000e+00 (diff) 5.898060e-17   0   0   0.0000   0.00000    0.000
nodegree  -1.110223e-16 (diff) 5.551115e-17   0   0   0.0000   0.00000    0.000
re74       7.197514e+00 (diff) 5.551115e-17   0   0   0.0000 -70.85522  416.416
re75       1.220698e+01 (diff) 5.551115e-17   0   0 234.4843 140.79126 -852.252
hispanic   0.000000e+00 (diff) 5.551115e-17   0   0   0.0000   0.00000    0.000
u74        0.000000e+00 (diff) 2.775558e-17   0   0   0.0000   0.00000    0.000
u75        0.000000e+00 (diff) 5.551115e-17   0   0   0.0000   0.00000    0.000
```

Diagnostic tool

The choice of coarsening affects the matching solution. Due to high computationally efficiency of `cem`, the function `relax.cem` allows for automatic coarsening relaxations

```
R> relax.cem(mat,LL)
Executing 42 different relaxations
.......[20%]....[40%].....[60%]....[80%]....[100%]
```



ATT estimation and extrapolation

ATT estimation on the matched data only

```
R> att(mat, re78 ~ treated, LL) -> TE
R> TE
            G0  G1
All        425 297
Matched    222 163
Unmatched  203 134

Linear regression model on CEM matched data:

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

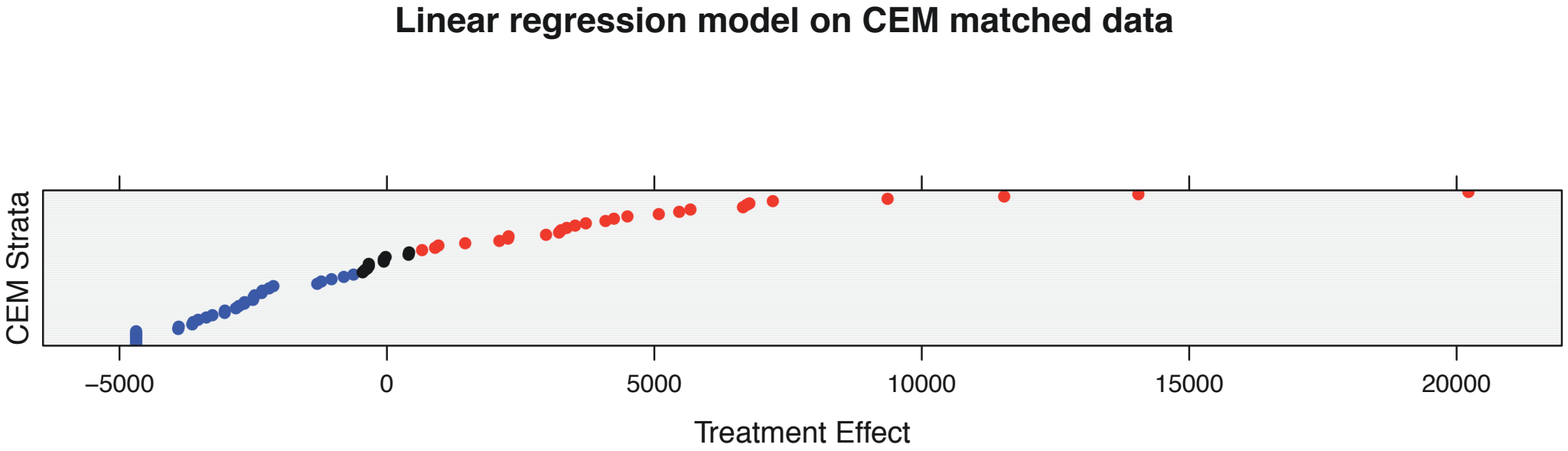
ATT estimation on all treated observations via extrapolation

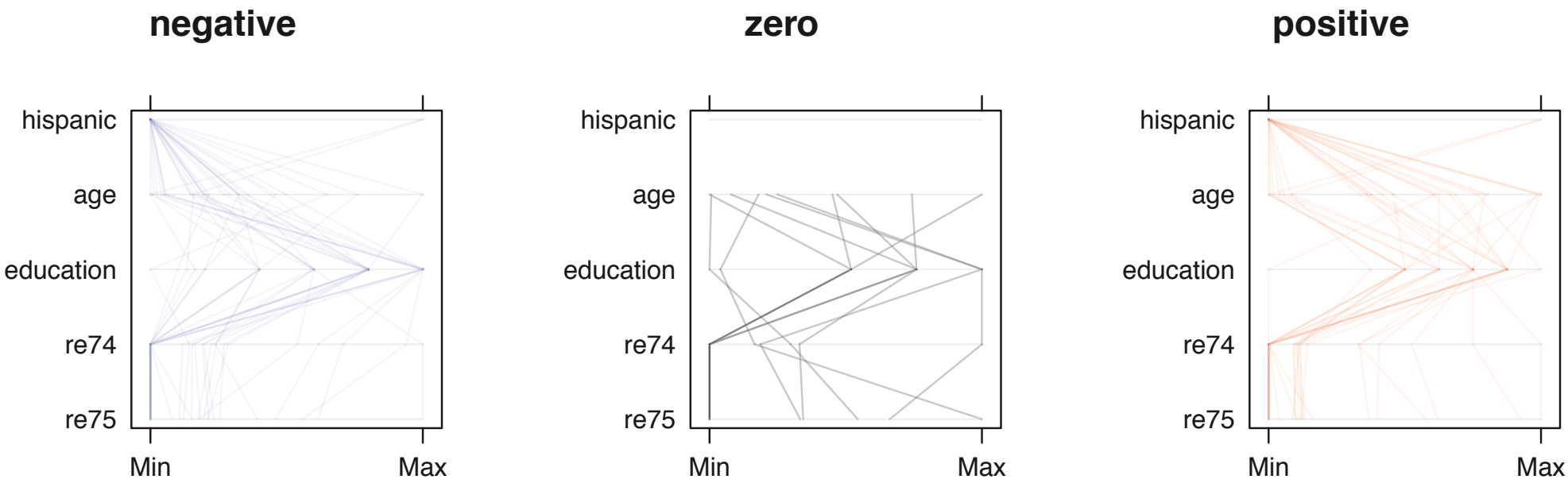
```
R> att(mat, re78 ~ treated, LL, extrapolate=TRUE)

            G0  G1
All        425 297
Matched    222 163
Unmatched  203 134

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





For more information

For the latest version of the manuscript, R and **Stata** software, visit

`http://GKing.Harvard.edu/cem`