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

Stefano Maria Iacus[1,*] , **Gary King**[2], **Giuseppe Porro**[3]

1. Department of Economics, Business and Statistics, University of Milan, Italy
2. Institute for Quantitative Social Science, Harvard University, USA
3. Department of Economics and Statistics, University of Trieste, Italy

* Contact author: stefano.iacus@unimi.it

Keywords: causal inference, matching methods, imbalance measures, evaluation, treatment effect estimation

We present an R package that implements a new matching method for causal inference in observational data (Iacus, King, Porro, 2009). Observational data are typically plentiful and common in the social sciences; as such, the main issue is reducing bias and only secondarily to keep the variance low. However, most matching methods seem designed for the opposite task, guaranteeing sample size ex ante (such as by choosing matching solutions of one-to-one) but achieving bias reduction (by reducing imbalance between treated and control groups in pre-treatment covariates) only sometimes and with a required extra ex post verification step.

Matching is a simple, intuitive technique of data preprocessing used to control for some or all of the potentially confounding influence of pretreatment control variables by reducing imbalance between the treated and control groups. After preprocessing in this way, any method of analysis that would have been used without matching can be applied to estimate causal effects. The resulting combination reduces model dependence and generally improves inferences with fewer assumptions.

CEM is a matching method with the property that the maximum imbalance between the treated and control groups is controlled by the user ex ante by clear and explicit choices rather than requiring it to be discovered ex post. With CEM, one can control the imbalance on one variable without affecting the maximum imbalance on any remaining variables. It is extremely easy to understand, teach, and use. Unlike many existing methods, it needs no distributional assumptions and so works with any data types. CEM also works on the original space of covariates and hence does not require the adoption of any distance or statistical model to perform the match, and eliminates the need for a separate prior procedure required for other methods that restrict data to common empirical support. It works well with multiple imputation methods for missing data, can be completely automated, and is extremely fast computationally even with very large data sets. CEM also works well for multicategory treatments, determining blocks in experimental designs, and evaluating extreme counterfactuals.

The package `cem` implements such matching method but introduces also a new tool to measure the imbalance in the whole multidimensional distribution of the data. This new index can be used to compare the solutions of different matching algorithms (and so is not specific to CEM). For a given data set the function `cem` returns a vector of weights, one per observation, which can be used later in any statistical model (i.e. `lm`, `glm`, etc) although the package provides also the `att` function for the estimation of the treatment effect (specifically the average treatment effect on the treated)

The package also introduces a diagnostic tool specifically designed for CEM which clearly shows which variable makes the match harder and a graphical tool to represent the distribution of the treatment effect along different strata of the sample rather than just the average treatment effect.

References

Iacus, S.M., King, G., Porro, G. (2009) Matching for Casual Inference Without Balance Checking `http://gking.harvard.edu/files/abs/cem-abs.shtml`