MISQ Archivist

A Tree-Based Approach for Addressing Self-Selection in Impact Studies with Big Data

Inbal Yahav, Galit Shmueli, and Deepa Mani

Abstract

In this paper, we introduce a tree-based approach adjusting for observable self-selection bias in intervention studies in management research. In contrast to traditional propensity score (PS) matching methods, including those using classification trees as a subcomponent, our tree-based approach provides a standalone, automated, data-driven methodology that allows for (1) the examination of nascent interventions whose selection is difficult and costly to theoretically specify *a priori*, (2) detection of heterogeneous intervention effects for different pre-intervention profiles, (3) identification of pre-intervention variables that correlate with the self-selected intervention, and (4) visual presentation of intervention effects that is easy to discern and understand. As such, the tree-based approach is a useful tool for analyzing observational impact studies as well as for post-analysis of experimental data. The tree-based approach is particularly advantageous in the analyses of big data, or data with large sample sizes and a large number of variables. It outperforms PS in terms of computational time, data loss, and automatic capture of nonlinear relationships and heterogeneous interventions. It also requires less user specification and choices than PS, reducing potential data dredging. We discuss the performance of our method in the context of such big data and present results for very large simulated samples with many variables. We illustrate the method and the insights it yields in the context of three impact studies with different study designs: reanalysis of a field study on the effect of training on earnings, analysis of the impact of an electronic governance service in India based on a quasi-experiment, and performance comparison of contract pricing mechanisms and durations in IT outsourcing using observational data.

Keywords: Self-selection, classification and regression trees, intervention, decision-making, e-governance, outsourcing, analytics