When You Can't Go for the Gold: What is the best way to evaluate a non-RCT demand response program?

Although randomized control trials (RCTs) are widely acknowledged as the gold standard of impact evaluation, they cannot always be implemented due to legal, financial, or other considerations. These problems are particularly acute for demand response programs, which are often created as a result of capacity constraints. Excluding customers to create a control group could result in a failure to remove enough demand from the grid during an event. Consequently, evaluators must frequently estimate program impacts without a true control group. It is therefore critical to understand how the choice of baseline estimation methods affects saving estimates, and which estimation method is best to use under different circumstances.

This paper uses data from a 2015 evaluation of a pilot smart thermostat program to evaluate several common (and not so common) approaches to estimating demand response program impacts. Because the pilot thermostat program was an RCT, we compare the results from the non-RCT approaches to the unbiased savings estimates. The pilot program included 1,440 participants. Although participants opted into the program, they were randomly assigned to treatment and control status during individual demand response events.

Specifically, this paper compares the performance of common impact evaluation methods such as X-in-Y baselines, linear fixed effects regression, and matched comparison groups. It also explores less common but potentially useful methods such as synthetic baselines and Bayesian models. In addition to evaluating the general performance of these analytic approaches, this paper pays particular attention to their relative performance under a very common condition: when event days are much hotter than most or all comparison days.

This has important implications not only for evaluators tasked with calculating impacts for non-RCTs, but also for utilities and program implementers who must decide when running a program as an RCT is worth the cost. It also gives us some insight into how much error, both from bias and accuracy, we can expect from these non-RCT-based methods. A better understanding of the performance of non-RCT evaluations methods will help ensure that evaluators' results and recommendations truly help "make reductions real".