Go for the Silver? Evidence from field studies quantifying the difference in evaluation results between "gold standard" randomized controlled trial methods versus quasi-experimental methods

Anna Spurlock, Peter Cappers, Ling Jin, Annika Todd, LBNL
Patrick Baylis, University of California at Berkeley

ABSTRACT

Randomized controlled trials (RCTs) are widely viewed as the "gold standard" of evaluation. However, analysis of the effect of energy pricing has largely been conducted through quasi-experimental methods. Using a rare set of large-scale randomized field experiments of time-based electricity pricing, we compare the estimates obtained from commonly used non-experimental methods against RCT estimates.

We demonstrate empirical evidence in favor of four stylized facts that highlight the importance of understanding two important sources of bias in this context: selection bias and spillover effects. First, difference-in-difference and propensity score methods tend to underestimate the true average treatment effect. Second, regression discontinuity methods tend to overestimate the effect. Third, selection biases in quasi-experimental methods tend to be more pronounced in opt-in treatments relative to opt-out treatments. Fourth, the three-in-five day baseline with an additive adjustment recommended by KEMA (2011) tends to underestimate the impact of the intervention, a pattern we attribute to intertemporal spillover effects.

Introduction

In this paper we scrutinize several methodologies commonly used in the evaluation of electricity demand response (DR) and pricing programs. We compare them to the "gold standard" randomized, controlled trial (RCT) experimental evaluation methodology, and find systematic evidence of selection and spillover effects that bias the non-experimental estimates.

RCTs have been widely used in fields such as public health and psychology. They are considered to be the "gold standard" in research design for empirical social science because the randomization process holds potential confounding factors equal across control and treatment groups, allowing the researcher to isolate the treatment effect of interest. However, some argue that obtaining the gold standard comes at a price; that RCTs tend to be expensive and time-consuming, can be challenging to implement correctly, are limited to settings where an experimental intervention is feasible, and are subject to concerns regarding external validity. While some of these barriers are important, many can easily be overcome with experience or sufficient advanced planning.

Meanwhile, a long history of empirical work has used an array of quasi-experimental research designs intended to simulate the experimental process, such as matching, propensity score weighting, regression discontinuity, and within-unit estimators. These research designs can often be applied after a program has taken effect, which can disincentivize the need to plan for evaluation carefully at the program implementation stage. However, these quasi-experimental techniques are potentially more likely to suffer from biases.

This paper builds on prior work in the peer-reviewed literature that compares results obtained using non-experimental research designs with experimental results.[1] Recent work has extended this type of analysis to residential electricity consumption data. In particular, a working paper by Jessoe, Miller, and Rapson (2015) examines the possibility of using high-frequency electricity data to recover causal effects without an experimental comparison group. An advantage of our approach is that we use an experiment with multiple treatment arms to validate trends in our results. Implementing quasi-experimental estimators across all treatment arms allow us to ascertain if consistent biases relative to experimental estimates exist. In particular, we provide evidence that selection biases and spillover effects drive the observed biases in the non-experimental results in our setting.

Empirical Context: Residential Electricity Pricing Programs

Accurate evaluations of DR and pricing programs in the electricity industry are important for several reasons. First, settlement and payment for incentive-based programs (such as peak time rebates) require an accurate evaluation of how consumption for a specific household changed on a single critical event window relative to their baseline (or counterfactual) consumption. In these programs, customers are paid for the amount of electricity they saved during a given critical event relative to this baseline. Second, utilities often claim savings and recover costs from ratepayers as authorized by regulators, and these savings need to be accurately measured through a program impact evaluation. Third, an assessment of how well a program is working is crucial for future program and portfolio planning, so that ratepayer dollars are spent on programs that achieve the highest savings at the lowest cost. Forth, accurate short- and long-term grid-level energy and capacity forecasts are necessary for maintaining reliability. These forecasts enter into resource planning efforts that inform the need for future infrastructure investment. Understanding the true savings resulting from a given program ensures that both utilities and ratepayers are being appropriately compensated for their efficiency efforts, while helping prevent the construction of unnecessary generation capacity.

Until recently, RCTs have been met with substantial resistance in the residential energy sector. Concerns that have been raised include: they require substantial planning up front at the program implementation phase in contrast to quasi-experimental techniques which typically require analysis only ex-post; they are seen as difficult to implement; and they are sometimes described as “unfair” because they restrict program participation to exclude the control group. As such, the majority of the evaluation and baseline methods used historically have been non-experimental. The specifics of several of these methods will be outlined later in the paper.

However, there has been a recent increase in interest in the application of RCT evaluation methods in the residential electricity sector. This trend has been enabled by increasingly available high-frequency data from smart meter infrastructure, and was spurred forward with the

[1] Much of the seminal work in this area was conducted in the labor economics literature (labor economics is the subfield of economics that studies work and employment, focusing on questions of labor supply and demand, wages, and the role of skill, motivation, education, and other factors on success outcomes). LaLonde (1986) conducts such a comparison in the context of an employment-training program, finding that the non-experimental estimates frequently fail to align with the experimental results. Heckman, Ichimura, and Todd (1997) analyze a separate program and find that non-experimental estimates can perform well so long as the comparison samples are drawn from a similar sample. Dehejia and Wahba (1997) find that propensity score estimates can outperform traditional econometric estimators, although Smith and Todd (2001) note that the former finding may be due to the sample selection imposed.

increased visibility and popularity of behavior-based programs, such as Opower's Home Energy Reports (e.g., Allcott 2011a). The average treatment effect sizes are quite small for behavior-based programs, and so regulators have tended to require a higher bar for their evaluation in terms of accuracy and robustness in order to accept the savings from these programs claimed by utilities. This is in contrast to savings claimed by energy efficiency programs, for example, for which average savings tend to be higher, and are assumed to be more lasting. The discussion around RCT in the context of behavior-based programs, however, facilitated the expansion of these methods beyond these programs alone.

In the context of time-based pricing programs, in 2009 the United States Department of Energy (DOE) issued a funding opportunity announcement for its Smart Grid Investment Grant (SGIG) that requested proposals from utilities seeking funding to expand their smart meter infrastructure. DOE required these proposals to include randomized pricing experiments, which were to be enabled by the new advanced metering infrastructure. Ten utilities were ultimately funded under SGIG and undertook Consumer Behavior Studies (CBS) that utilized randomized evaluation methodology for their pricing pilots.[2]

We use the opportunity offered by the randomized time-based rate pilots under the SGIG CBS in order to assess the accuracy of the non-experimental methods most commonly employed to evaluate DR and pricing programs historically. Building on the pioneering work by LaLonde (1986), we take a set of electricity pricing RCT experiments as the gold standard against which we compare our set of non-experimental estimates. Because electricity consumption is a data-rich context, we are able to implement a range of non-experimental techniques. Specifically, we estimate three quasi-experimental evaluation methods: (1) difference-in-differences (DID) (2) a propensity score estimator that reweights observations by their treatment likelihood, and (3) a regression discontinuity (RD) design that discontinuously influences treatment likelihood. We compare the estimates of the average treatment effects obtained using these quasi-experimental techniques to those obtained from the RCT. We additionally examine three common baseline-derived estimation methods and compare estimated average savings for each critical event day using these methods to those obtained from the RCT. The baseline methods we examine are: (1) a four-in-five day baseline, (2) this same four-in-five day baseline method with an additive adjustment (KEMA 2011), and (3) an individual customer regression estimator. Our primary purpose in this exercise is to document any systematic biases (e.g., selection bias and bias from spillover effects across days) present when these estimates were obtained using non-experimental methods.

We document empirical support for five results. First, difference-in-difference and propensity score methods tend to underestimate the true average treatment effect, suggesting the presence of selection bias when using these methods. Second, RD methods tend to overestimate the size of the true average treatment effect, underlining the limitation of RD to provide

[2] More information on the SGIG CBS studies can be found at https://www.smartgrid.gov/recovery_act/overview/consumer_behavior_studies.html. While time-based pricing for electricity has existed for a long time, and many evaluations of this type of pricing have been conducted and documented in white papers (some of which will be references below), academic researchers have typically focused on the fairly small set of programs that did happen to be implemented using experiments. Aigner (1984), Train and Mehrez (1994), and Jessoe and Rapson (2012) analyze the effect of separate time-of-use (TOU) experiments. Allcott (2011b) analyses a real-time pricing (RTP) experiment. Wolak (2007) examines the response to a critical peak pricing (CPP) program. The fact that past instances of randomized experiments are relatively limited is indicative of the resistance we've mentioned to these methods in this industry historically.

externally valid estimates. Third, biases in non-experimental research designs tend to be more pronounced in opt-in treatments relative to opt-out treatments, further confirming the selection effect interpretation. Fourth, the baseline methods, particularly the four-in-five day baseline both with and without the adjustment, tend to underestimate the impact of the intervention, pointing to the importance of intertemporal spillover effects.

For policy-makers, this work contributes to our understanding of the usefulness of non-experimental methods in ex-post measurement of changes in consumption as a result of electricity rate design. Many utilities and public utilities commissions are considering a broader implementation of time-based pricing of electricity in the next decade.[3] Policymakers may want to test the effects of these changes, but may not have the resources or time to implement a full RCT.[4] Our results suggest the following: (1) while RD is viewed favorably by the empirical economics community, it may in fact be more biased relative to the population average than properly constructed difference-in-differences or propensity score estimates due to the limitations associated with using it to estimate average treatment effects across a broad spectrum of the population; (2) practitioners considering an implementation of a time-based electricity rate should note that a non-experimental evaluation of default or opt-out rates is less likely to be biased due to selection considerations than a non-experimental evaluation of an opt-in rate; and (3) four-in-five day baseline methods, if used, must include a adjustment for weather and usage patterns, but are likely to be bias due to spillover effects. The individual customer regression baseline method we assess performs slightly better, but also suffers from spillover effects.

Overview of the Time-Based Pricing Experiments

The set of experiments analyzed in this paper come from the SGIG CBS studies conducted by the Sacramento Municipal Utility District (SMUD). SMUD's customer base has approximately 530,000 residential households. After many were excluded based on pre-defined eligibility criteria, approximately 174,000 households remained in the study.[5]

There were three pricing treatments tested: a time-of-use (TOU) program where customers faced higher prices 4pm to 7pm on non-holiday weekdays, but lower prices off-peak; a Critical Peak Pricing (CPP) program where they faced very high prices during the peak period of 24 critical event days in total called a day in advance over the course of two summers, but lower prices during all other hours; and a combined rate where a CPP was applied on top of a TOU. The experimental prices were in effect between June 1st and September 30th for the two summers in the study (2012 and 2013). In addition, there was an enabling technology associated with some of the treatment groups in which customers were offered in-home displays (IHD). There were two forms of recruitment: opt-in, where households were encouraged to enroll in the rate program, but would not be enrolled unless they actively opted in; or opt-out, where households were notified that they were enrolled by default and were encouraged to stay in the rate program, but had the opportunity to leave the program if they wished.

As a result of different combinations of these parameters, households in the experimental population were randomly assigned into ten groups; in this paper, we examine eight of those

[3] For example California is moving towards TOU as the default rate (CPUC 2015).

[4] We note that the existence of the present set of RCTs is due to a large DOE grant, which also funds this study.

[5] Households were excluded from the experiment if: they did not have interval meters to capture hourly electricity usage installed prior to June 2011; they were participating in one of SMUDs other concurrent programs; or if they had master metered accounts.

groups, seven of which were encouraged to participate in a time-based pricing treatment, while the eighth group was the control group, which received no encouragement and remained on the standard rate.[6] Specifically, the treatment arms included: (1) CPP opt-in with IHD, (2) CPP opt-in with no IHD, (3) CPP opt-out with IHD, (4) TOU opt-in with IHD, (5) TOU opt-in with no IHD, (6) TOU opt-out with IHD, and (7) TOU-CPP opt-out with IHD.

Data

The data consist of hourly energy consumption, in kilowatt-hours (kWh) for each household in our control group, as well as for each household in our seven treatment groups, regardless of whether or not they ended up enrolled on the treatment pricing, or whether or not they opted out at any point in the pilot period. These energy consumption data were collected for one year prior to the start of the pilot period (June 1st, 2011 - May 31th, 2012) and for two years during the pilot period (June 1st, 2012 - September 30th, 2013).[7]

We also use hourly weather data, including dry and wet bulb temperature as well as humidity. There is only one weather station in close proximity to all participants in the SMUD service area, so the weather data does not vary across households, only over time.

Theory

The fundamental problem of causal inference is that it is impossible to simultaneously observe units in both treated and untreated states. In the context of estimating the effect of electricity pricing treatments, this means that researchers cannot observe how much electricity a control customer would have demanded had she been exposed to the treatment or how much a treatment customer would have demanded had she not been treated. Experimental methods circumvent this problem by randomizing, while quasi-experimental methods use a variety of techniques to claim that treatment is “as good as random.” Econometrically, the goal in any evaluation is to ensure that the error term (capturing any and all unobserved factors) is uncorrelated with the independent variable of interest. For example, in an electricity pricing setting, it must be assumed that households who participate in a new pricing program are not systematically different in ways that affect their electricity consumption and ability to respond to the treatment compared to households that do not participate. In a randomized setting, this assumption is known to be true, by virtue of the randomization itself. In quasi-experimental settings, this assumption cannot be proved, but must be claimed. The following section provides an overview of the evaluation methods we employ, with an emphasis on the assumptions required to overcome the fundamental problem of causal inference.[8]

[6] The final two treatment arms were alternative control groups used for an evaluation based on a recruit and delay RCT experimental design, and did not face time-based prices during the 2012-2013 timeframe.

[7] Coverage of the hourly energy consumption and billing data was quite complete. While there are a handful of missing observations (less than one percent), they do not differ systematically across treatment groups, nor across those who did and did not end up in treatment. A comparison of pre-treatment energy usage shows that there is no statistical difference between the control group and each of the six experimental treatment groups (including average kWh per day, peak hours, and peak to off peak ratio)

[8] For a detailed explanation of different types of impact evaluations see Cappers, Todd, Perry, Neenan & Boisvert (2013) for energy savings impact evaluations, and Imbens and Wooldridge (2009) for a comprehensive econometric discussion.

Experimental Design

The key feature of RCTs is that units are assigned randomly between control and treatment groups. Proper randomization and sufficient sample size should ensure that these two groups are similar across both observable and unobservable attributes. If this is the case, then any differences in the average outcome between the control and treatment groups should be entirely attributable to the treatment itself.

In our context, customers were randomly assigned to treatment and control groups. Each treatment group is then comparable to the control group. To account for statistically insignificant but slight differences in the two groups, we estimate the difference between the average change in electricity usage in the pre-treatment period and the post-treatment period between the treatment and control groups. Because not all customers from the treatment groups enrolled in the program, we are actually using a Randomized Encouragement Design (RED), which allows us to estimate the average effect of taking up the treatment. This design requires the additional assumption that treatment status (i.e., being encouraged to enroll) did not affect energy usage except by causing enrollment. The treatment effect from an RED can be estimated using two-stage least squares.

Quasi-Experimental Average Treatment Effect Estimators

Difference-in-differences (DID). The difference-in-differences estimate compares the difference in before- and after-treatment electricity usage between customers who chose to enroll in the treatment and the randomly selected control group that was not informed about the pilot. Customers that were encouraged to enroll in treatment but did not are omitted from the analysis. For this estimate to be unbiased, we require that the selection into treatment choice be as good as random, i.e., customers' decisions to enroll are uncorrelated with their electricity usage. This is a strong assumption, since it is natural to expect that individuals who choose to enroll in a time-based electricity pricing program, such as a TOU or CPP, may have different electricity use patterns as compared to those who choose not to enroll. For example, customers with lower peak consumption may be more likely to enroll than those with higher peak consumption because they may anticipate being able to save money with the time-based rate. Because of the potential for self-selection, we anticipate that the DID estimate will show evidence of selection bias. In addition, we note that there may be important differences between the opt-in and opt-out treatments. Since the opt-in treatments enrolled at most 20% of treated customers, the selection effect for these groups is likely to be substantial. However, since the opt-out treatments enrolled at least 90%, selection is likely to be more muted.

Propensity score matching. Our third quasi-experimental technique theoretically improves upon the DID approach by trying to control for selection using observable characteristics. It does this by using a standard propensity-score matching approach to account for selection into treatment. We construct estimates of each customer's enrollment likelihood based on their pre-treatment electricity usage. We then estimate a regression that adjusts for differences due to selection into treatment using the propensity score. The required assumption in implementing this methodology is that the variables used to construct the propensity fully account for unobserved differences between the control and treatment groups. If this assumption does not hold, then we would expect the propensity score estimate to also be subject to selection bias.

Regression discontinuity (RD). RD designs take advantage of cutoffs that affect selection into treatment. In the electricity context, a relevant cutoff might be generated if a program offers time-based pricing to any customers with total pre-treatment summer electricity usage above a given threshold but not to those below. The underlying assumption is that customers above and below the eligibility threshold are similar except in their ability to join the pricing program. In essence the assumption is that customers cannot anticipate the cutoff and manage their consumption such that they are able to orchestrate their qualification, or not, for treatment. It is reasonable to assume that this is the case, as in order for a household to manage their consumption so that they landed just above or below the cutoff, they would have to pay close attention to their aggregated electricity consumption over the course of several bill cycles. The RD approach, therefore, assumes customers just above or below the cutoff are as good as randomly assigned, and similar in most other ways. However, in order to ensure this similarity, the treatment effect is estimated using only customers existing in a narrow band (in our case 10%) right below and above the cutoff. This means that RD methods are likely to suffer from reduced external validity, and may not be able to recreate the average treatment effect for the entire population if the treatment effect is highly correlated, particularly in a non-linear way, with the variable used to define the cutoff.

Non-Experimental Baseline Methods

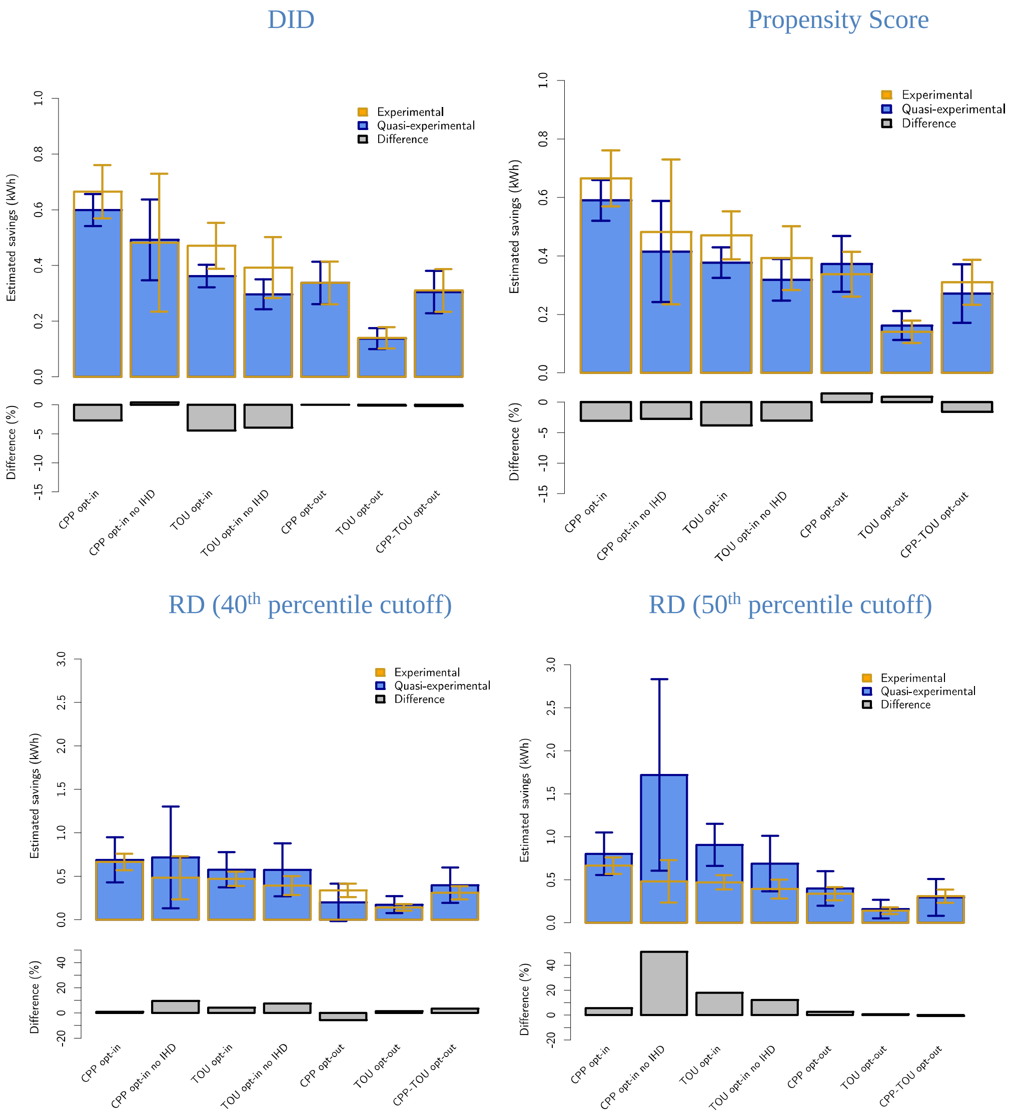
Four-in-five day baseline. To estimate the average energy savings for each critical event day, we create a baseline by using the average consumption from the four highest consumption days out of the last 5 non-event business days, to which the customer's usage on a critical event day is compared. This is referred to as the four-in-five day baseline method. In addition, we implemented the additive adjustment to this baseline as recommended by KEMA (2011). The purpose of this adjustment is to help control for weather and underlying usage pattern differences across days. We chose this baseline method because it performed the best in KEMA's review of a variety of baseline methods (KEMA 2011). This estimate may be biased if the non-event days are substantially cooler, or if there are spillover effects (i.e., customers who are enrolled in a CPP treatment have started reducing their consumption, through habit formation or technological investments, outside of the critical event days).

Individual customer regression baseline. We also estimate a within-customer treatment effect for the critical events, which uses non-event hours across the entire treatment period for each customer, controlling for temperature, as a baseline for comparison to the event hours. The temperature adjustment is achieved by running regressions for individual customers separately while using the same model specification. The key assumption in this method is that there are no treatment spillovers as described above (i.e., that customers do not systematically adjust their behavior outside of the critical event days as a result of being in a CPP treatment). This method also assumes that the temperature dependence (determined from the individual customer regressions) of customer's usage does not change over the range observed between non-event versus event periods.

Results

Comparing Average Treatment Effect Estimates

Figure 1 summarizes the differences between the average treatment effect estimated using the experiment and those obtained using the quasi-experimental approaches described in the previous section. The comparison is show for the DID (top left panel), propensity score (top right panel), and two versions of the RD (bottom two panels). The two RD estimates differ in the location of the eligibility cutoff used in constructing the simulated RD. In the bottom left panel the cutoff limited enrollment to those above the 40th percentile of pre-period summer usage, while the bottom right panel shows results when the cutoff was limited to those above the 50th percentile of pre-period summer usage. In all graphs the experimental estimate of average per-household hourly kWh peak period savings are show by the blue solid bars, the corresponding quasi-experimental estimated treatment effect is show in the outlined orange bars. The difference as a percent of average hourly peak consumption is shown in the lower scale in grey. The whiskers on the bars indicate the 95 percent confidence intervals of the two sets of estimates. We document three stylized facts related to the biases in the average treatment effects estimated using the quasi-experimental methods shown in Figure 1.


DID

Propensity Score

RD (40th percentile cutoff)

RD (50th percentile cutoff)

Figure 1. Comparison of treatment effect estimates. *Source*: Author calculations.

Difference-in-difference and propensity score methods underestimate the treatment effects. The DID approaches and the propensity-score method underestimate (in absolute value) the effect of the treatment relative to the randomized design. If an evaluation of these rates were done using one of the quasi-experimental approaches, the bias would have reduced the estimated treatment effect by as much as 5 percentage points. So, if the true average treatment effect were 20% of hourly peak consumption for example, the quasi-experimental estimates would have generated an estimate of only 15% in some cases. To interpret the result, we recall that the treatment group in this design consists entirely of customers who deliberately select into treatment. This group is observationally different from the control group and is likely to have different electricity usage patterns. We interpret the difference between the DID estimates and the propensity score estimates as driven by this selection effect: customers who actively chose to participate in the time-based pricing program had different underlying trends, biasing the result downwards. We note that this bias could have been either towards or away from zero, depending on the nature of the selection effect or trends in weather.

RD methods tend to overestimate the treatment effects. As discussed above, the simulated regression discontinuity method that we construct avoids the selection bias present in the other two methods by design. In contrast to the DID estimate, the RD estimates tend to overestimate (in absolute value) the true effect for the opt-in groups. Empirically, the only difference between these two estimates is that the RD method excludes treated customers below the threshold and control customers above the threshold, while the RCT method includes all treatment and control customers above and below the threshold. The magnitude of the difference is as much as 20 percentage points in some cases. So, a true treatment effect of 20% of hourly consumption, for example, would have been estimated to be as much as 40% in some cases if evaluated using the RD method. The overestimation is more pronounced the higher the eligibility cutoff. This highlights the fact that, if the variable used to define the cutoff is correlated, particularly in a nonlinear way, with the likelihood of selecting into treatment and the treatment response, the RD is likely to be biased, which is what we see here; because the treatment groups have, by design, higher pre-period usage, they are able to reduce more in the post-period.

The biases are more pronounced in opt-in versus opt-out designs. In all designs, estimation of the average effect of the opt-out treatments is less biased than the opt-in treatments. We interpret this finding as strong evidence that selection bias is driving the larger differences for the opt-in treatment arms: because at-most 20% of individuals chose to opt-in to treatment when offered, the sample obtained using an opt-in enrollment method is likely to be more heavily selected than that obtained using an opt-out enrollment method, which achieved 90% enrollment. Because the quasi-experimental methods are potentially subject to sample selection bias, using a less-selected sample to begin with naturally improves the quality of the quasi-experimental estimate.

Comparing Event Day Treatment Effects

Figure 2 documents average differences in estimates of event-day treatment effects, comparing the four-in-five (with and without adjustment) and individual customer regression baseline-based estimators to the experimental estimates. These estimates were generated using the CPP opt-in and CPP opt-out treatment groups only. The differences in kWh, while the bottom set of panels shows the difference as a percent of average electricity consumption by the control group during those events.

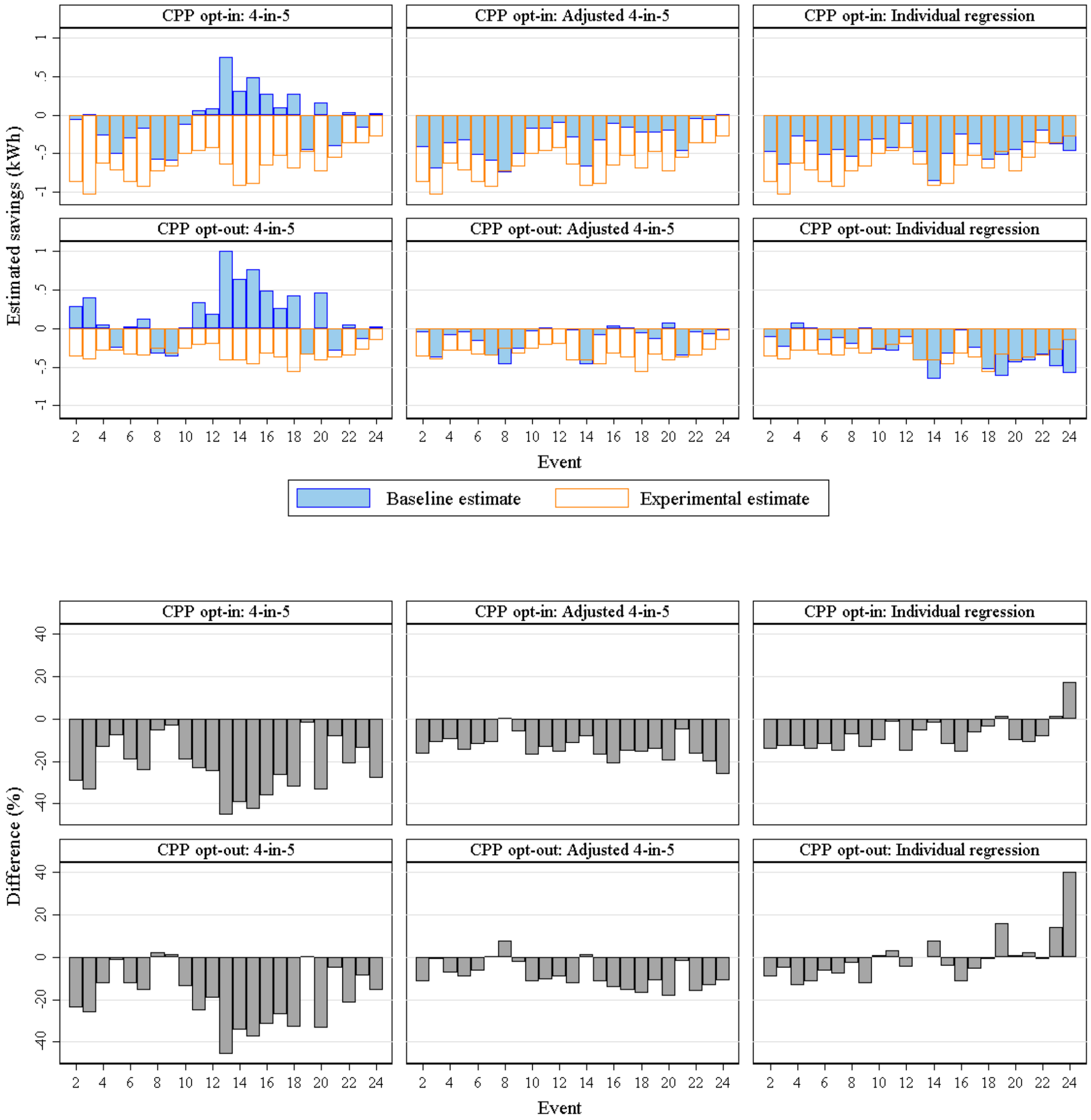


Figure 2. Comparison of event day treatment effect estimates using baseline methods. *Source*: Author calculations.

All baseline-based methods tend to underestimate the treatment effects. We find that the baseline approaches tends to underestimate (in absolute value) the effect of event days on the electricity usage of the treatment group on average. In the case of the four-in-five day baseline method with the adjustment, the difference is as much as 20 percentage points in some cases, meaning a true effect of 25%, for example, might be estimated to be as low as 5%. The underestimation using the four-in-five day baseline without the additive adjustment is substantially worse, while the individual customer regression baseline performs slightly better on average. This systematic bias is not correlated strongly with temperature on event days, we therefore interpret the bias to be due in large part to spillover effects: through habit formation or technology changes customers are reducing their usage during non-event hours as a result of being in treatment, this means that event savings compared to this reduced baseline are underestimated. Note that these differences are averaged across all the households. This means that the difference is likely to be even worse for some households individually. Therefore, if the four-in-five day baseline method were being used to determine repayment for a peak time rebate program, for example, the variation present, both in terms of the direction of bias (underestimating savings) and magnitude of the inaccuracy (as much as completely washing out the average treatment effect) would mean that the utility would likely be significantly undercompensating households if using this method.

Conclusion

Using a rich set of field experiments designed to test customer response to time-based pricing, we compare a set of established non-experimental evaluation methods to their corresponding experimental estimates. By comparing across multiple treatment arms we are able to provide support for a set of stylized facts, each of which has important policy implications for ex post estimation of time-based pricing programs.

First, we document that the DID and propensity score estimates are likely to generate bias due to selection. In our setting, weather variation between the pre- and post- period and the nature of the self-selection caused these estimates to be biased towards zero. Second, we show that even well-constructed RD estimates can be biased away from the treatment estimate due to energy use level differences between the treatment and control groups. Third, we observe that selection biases are more pronounced in all designs under opt-in treatments as compared to opt-out treatments. This finding strongly suggests that policy-makers should take this into account when designing the enrollment mechanism for a time-based pricing program: in addition to being less costly and more effective at reducing total electricity usage, ex post estimation of opt-out designs using quasi-experimental designs are less likely to be unbiased. Finally, we find that non-experimental baseline-based savings estimates tend to underestimate the actual savings, which we interpret to be likely a result of intertemporal spillover effects.

We caution that these results are limited to a set of treatment arms in a single experimental setting, and we emphasize that the direction of the biases in the quasi-experimental estimates is not necessarily likely to be stable in other contexts. Instead we suggest that careful consideration be given to underlying trends in treatment and control groups when interpreting quasi-experimental results, and that, when possible, opt-out enrollment mechanisms should be implemented. We also find evidence in favor of using within-customer estimation strategies instead of a four-in-five day baseline approach to estimate the effect of individual critical event days, but caution that practitioners should carefully consider the effect of spillovers in this context.

This work was supported by the U.S. Department of Energy’s Office of Electricity Delivery and Energy Reliability under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

References

Aigner, D. J. 1984. "The welfare econometrics of peak-load pricing for electricity: Editor's Introduction." *Journal of Econometrics* 26(1): 1-15.

Allcott, H. 2011a. "Social norms and energy conservation." *Journal of Public Economics* 95(9): 1082-1095.

———. 2011b. "Rethinking real-time electricity pricing." *Resource and Energy Economics* 33(4): 820-842.

Cappers, P., A. Todd, M. Perry, B. Neenan, and R. Boisvert. 2013. “Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines.” Lawrence Berkeley National Laboratory Report LBNL-6301-E.

CPUC (California Public Utilities Commission). 2015. *Proposed Decision Agenda ID 13928 Rulemaking 12-06-013 Order Instituting Rulemaking on the Commission’s Own Motion to Conduct a Comprehensive Examination of Investor Owned Electric Utilities’ Residential Rate Structures, the Transition to Time Varying and Dynamic Rates, and Other Statutory Obligations.* http://delaps1.cpuc.ca.gov/CPUCProceedingLookup/f?p=401:56:15072416273515::NO:RP,57,RIR:P5_PROCEEDING_SELECT:R1206013

Imbens, G. W., and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5-86.

Jessoe, K., and D. Rapson. 2012. *Knowledge is (less) power: Experimental evidence from residential energy use*. National Bureau of Economic Research Working Paper W18344.

Jessoe, K., D. Miller, and D. Rapson. 2016. “Can high-frequency data and non-experimental research designs recover causal effects? Validation using an electricity usage experiment.” Unpublished working paper.

KEMA. 2011. *PJM Empirical Analysis of Demand Response Baseline Methods.* https://www.pjm.com/~/media/markets-ops/dsr/pjm-analysis-of-dr-baseline-methods-full-report.ashx

LaLonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review* 76 (4): 604-20

Smith, J. A., and P. E. Todd. 2001. "Reconciling conflicting evidence on the performance of propensity-score matching methods." *The American Economic Review* 91(2): 112-118.

Train, K., and G. Mehrez. 1994. "Optional time-of-use prices for electricity: econometric analysis of surplus and pareto impacts." *The RAND Journal of Economics*: 263-283.

Wolak, F. A. 2007. "Residential customer response to real-time pricing: The anaheim critical peak pricing experiment." Center for the Study of Energy Markets Working Paper CSEM WP 151.