# Some Insights on Matching Methods in Estimating Energy Savings for an Opt-In, Behavioral-Based Energy Efficiency Program

*Bill Provencher, Navigant Consulting & University of Wisconsin-Madison*
*Bethany Vittetoe-Glinsmann, Navigant Consulting*
*Anne Dougherty, Opinion Dynamics Corporation*
*Katherine Randazzo, Opinion Dynamics Corporation*
*Phil Moffitt, Cape Light Compact*
*Ralph Prahl, Prahl and Associates*

## Abstract

In this paper we examine two statistical models in which participants in an opt-in behavioral program are matched to non-participants based on historical energy use to estimate program savings. The first model involves standard regression analysis, while the second uses regression analysis to modify a matching estimator. We discuss the extent to which matching on energy use is likely to address two potential sources of bias in estimating savings from the program: (a) specification bias arising when a regression model of energy use is misspecified; (b) selection bias arising if participants are different than non-participants in ways that affect energy use and are not observable in the available data. The application is to a cohort of customers participating in a small-scale opt-in residential behavioral program in eastern Massachusetts. The two models generate similar estimates of program savings; the first estimates that average household savings in the first year of the program were 1.49%, and the second estimates that average savings were 1.36%. Estimated savings are statistically significant despite a small sample size and low savings, likely reflecting the high quality of the matches. A pseudo-test indicates the estimates are not affected by selection bias.

### Introduction

Recent reviews of behavior-based programs have stressed the advantage of randomized controlled trials (RCTs) to identify energy savings from demand response programs. This method is ideal for generating unbiased estimates because it provides the opportunity to assure that control households and treatment households are balanced in the distribution of explanatory variables and assures the two groups are, on average in repeated trials, balanced with respect to unobservable variables (that is, no selection bias).

Nonetheless, programs with non-RCT designs are likely to remain common in behavior-based programs for a number of reasons: (a) programs may not be able to meet goals without accessing customers reserved for control groups; (b) implementers may fail to consider the evaluation advantages of RCT designs; (c) randomization is sometimes done erroneously; (d) for certain programs RCT designs may not be acceptable to stakeholders; and (e) using an RCT design to isolate the effects of opt-in portions of a program requires excluding, at least temporarily, customers who wish to enter the program, a step implementers often are not willing or able to take.

In this paper we examine two statistical models in which participants in an opt-in behavioral program are matched to non-participants based on historical energy use to estimate program savings. The first model involves standard regression analysis, while the second uses

regression analysis to modify a matching estimator. We discuss the extent to which matching on energy use is likely to address two potential sources of bias in estimating savings from the program: (a) specification bias arising when a regression model of energy use is misspecified; (b) selection bias arising if participants are different than non-participants in ways that affect energy use and are not observable in the available data. The application is to a small-scale opt-in residential behavioral program offered by Cape Light Compact (CLC), a utility in eastern Massachusetts. The two models generate very similar estimates of program savings. Estimated savings are statistically significant despite a small sample size and low savings, likely reflecting the high quality of the matches. A pseudo-test indicates the estimates are not affected by selection bias.

**Models used in the impact analysis**

Two models are used to estimate savings. The first follows the approach of Stewart (2010) and Ho et al. (2007), who essentially argue that matching a comparison group to the treatment group is a useful "pre-processing " step in a regression analysis to assure that the distributions of the covariates (i.e., the explanatory variables on which the output variable depends) for the treatment group are the same as those for the comparison group that provides the baseline measure of the output variable. This minimizes the possibility of model specification bias. The regression model is applied only to the post-treatment period, and the matching focuses on those variables expected to have the greatest impact on the output variable.

We estimate the following model using participants and their matches:

**Model 1**

$$kWh_{kt} = \alpha_{0t} + \alpha_1 Treatment_k + \alpha_2 PREkWh_{kt} + \sum_{j=1}^{J} \beta^j EE_{kt}^j + \varepsilon_{kt},$$

where:

- $kWh_{kt}$ is the average daily electricity use by household *k* during month t;
- Greek characters denote coefficients to be estimated, and in particular $\alpha_{0t}$ is a monthly fixed effect.
- $Treatment_k$ is an indicator variable taking a value of 1 if customer *k* is in the program of interest, and 0 otherwise;[1]
- $PREkWh_{kt}$ is the average daily electricity use by household k during the most recent month before household *k* enrolled in the program that is also the same calendar month as month *t*. For instance, if household *k* enrolled in August 2011, the value of $PREkWh_{kt}$ for June 2012 is June 2011.
- $EE_{kt}^j$ is an indicator variable for energy efficiency program *j*, taking a value of 1if customer *k* is in the program in period *t* and 0 otherwise. *J* is the number of programs considered in the analysis.

[1] If program enrollment occurred during a bill cycle, the current bill cycle is not coded as the post period and the following bill cycle will be the first post period observation.

- $\varepsilon_{kt}$ is the error term.

In this model $\alpha_1$ indicates average daily savings by program participants.

The second model follows the approach summarized in Imbens and Wooldridge (2008). In this model the effect of the program in month *t* is the difference between the energy use of participant *k* and its estimated counterfactual (baseline) consumption. The estimated counterfactual consumption is the average consumption of its matched household amended to reflect differences between participants and their matches in the covariates ***X*** affecting energy use. Formally we have,

**Model 2**:

$$Savings_{kt} = kWh_{kt} - kWh_{kt}^{C}$$

$$kWh_{kt}^{C} = kWh_{kt}^{M} + \hat{\beta}\left(\mathbf{X}_{kt} - \mathbf{X}_{kt}^{M}\right)$$

where:

- $kWh_{kt}$ = the average daily electricity use by household *k* during month t;
- $kWh_{kt}^{C}$ = the estimated counterfactual energy use by household *k* during month *t*;
- $kWh_{kt}^{M}$ = the energy use by household *k*'s match during month *t*;
- $\mathbf{X}_{kt}$ = the values for household *k* in month *t* of the independent variables **X** affecting energy use;
- $\mathbf{X}_{kt}^{M}$ = the values of **X** in month *t* for household *k's* match.
- $\hat{\beta}$ = the factors used to adjust household *k*'s energy use to reflect differences between household *k* and its match in the value of **X.**

Following Abadie and Imbens (2011), the values of the adjustment factors $\hat{\beta}$ used in Model 2 are derived from a regression model applied to the post-program period, estimated using *only* the matched comparison households. In the current analysis the regression model used for adjustment purposes is identical to Model 1 except that the variable *Treatment* is excised, as the model is applied only to the matched comparison households

## The Application

The application is to CLC's Smart Home Energy Monitor Pilot (SHEMP) program. CLC is a public utility serving 200,000 customers on Cape Cod and Martha's Vineyard. The SHEMP program is an opt-in program that offers an integrated in-home monitoring and feedback system for customers on their household energy usage. Through this pilot, customers have access to nearly real-time data on their electric energy use. Customers receive the information through a website, where they can set goals and update their profile based on their home characteristics and any relevant household changes.

The SHEMP program has two primary cohorts: a small initial cohort ("Legacy" customers, N=83) that enrolled in the program in summer 2009 and a larger cohort ("Energize"

customers, N=277) that enrolled in the program in summer 2011. The analysis presented here is for the latter cohort, the Energize customers, because of the larger sample size.

Because the program is opt-in, and most customers enrolled in the program over a relatively short time span –most within four months –estimates of program savings rely on matched non-program comparison customers whose energy use provides a baseline against which the energy use of program participants is compared. In other words, the comparison group provides the "counterfactual" energy use of program households –the energy use of program households were they not enrolled in the program. The next section discusses the selection of the matched comparison households to be used in the models outlined above.

**Selecting Matched Comparison Households**

Whether the estimate of savings is accurate –statistically speaking, efficient and unbiased—depends on selecting comparison households that accurately represent the counterfactual behavior of program participants. We take the perspective that the best matches for program household *k* are those households whose monthly energy consumption during a period before household *k*'s enrollment in the program most closely matches household *k*'s consumption during the same period. The underlying logic is that households with energy consumption closely matched over an extended period demonstrate that they respond the same to the many exogenous factors –weather in particular, but also the prices of related goods and services, broader economic conditions, and social influences—that drive energy consumption.

From a statistical perspective, an argument to include other observable variables in the match must follow from the logic that these other variables are correlated with any separation in the match during the post-enrollment period that is not due to the effect of the program, nor to other variables included in the analysis, and that the values of these other observable variables are different on average for the program and comparison households. With this in mind, we also account for electric heat in the development of the matches.

The matching method used to develop the comparison group is a two-stage process. For each program participant, energy consumption in the 24 months before program enrollment (i.e., July 2009 to June 2011) was compared to *all* CLC residential customers with billing data over the same 24 months –approximately 162,000 customers. The basis of the comparison is the difference in monthly energy use between a participant and a potential match, $D_{PM}$ (**D**ifference between **P**articipant and potential **M**atch). The quality of a match is the Euclidean distance to the participant over the 24 values of monthly $D_{PM}$; that is, letting SSD denote the sum of squared $D_{PM}$ over the matching period, it is $SSD^{1/2}$. The ten CLC non-program residential customers with the shortest Euclidean distance to a participant were chosen as "match finalists" for the participant (first stage). From the ten finalists, three customers were chosen to be included in the analysis (second stage). Typically these three were the matches with the lowest Euclidean distance *and* the same heat type as the participant. If there were not at least three finalists with the same heat type, the three matches included in the analysis were chosen sequentially as follows: (a) all finalists with the same heat type; (b) the remaining finalist(s) with the lowest Euclidean distance.

The energy use by participants and their matches during the matching period is presented in Figure 1. Another view of the quality of the match is presented in Figure 2, which shows the average difference between participants and their matches before the start of the program, with a linear trend indicating that over the 41 months before the start of the program there is a slight

average difference between treatment households and their matches, and no statistically significant trend in the difference.

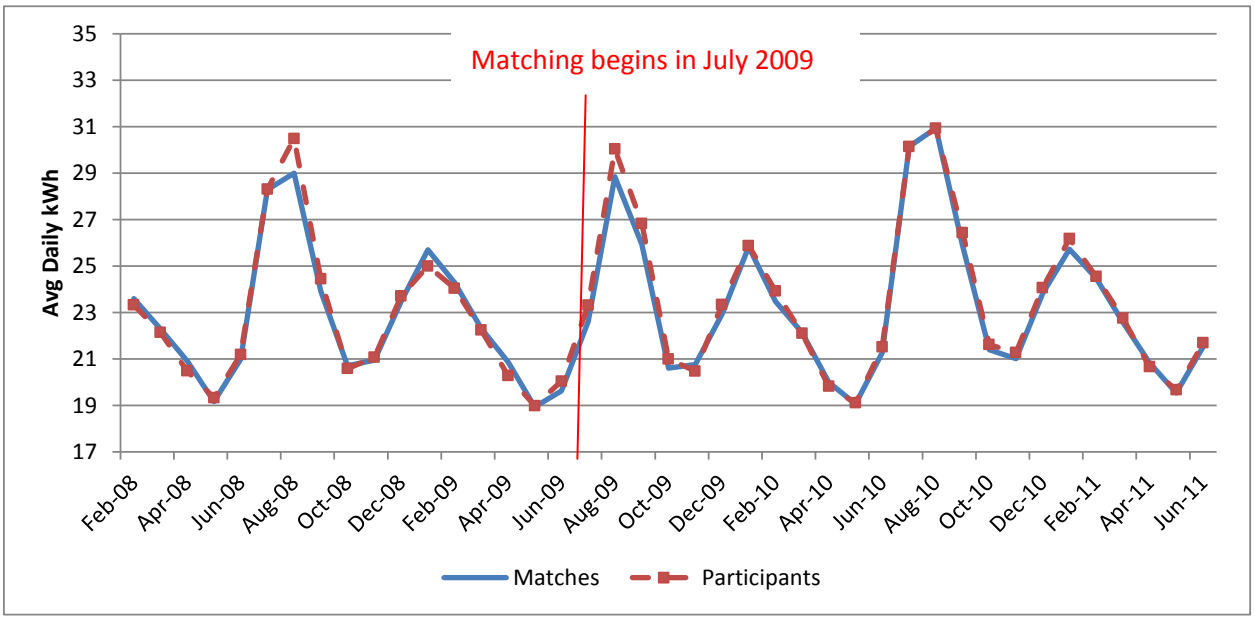


**Figure 1.** Comparison of the average daily kWh consumption of participant households and their 24-month matches in the 41 months before program enrollment

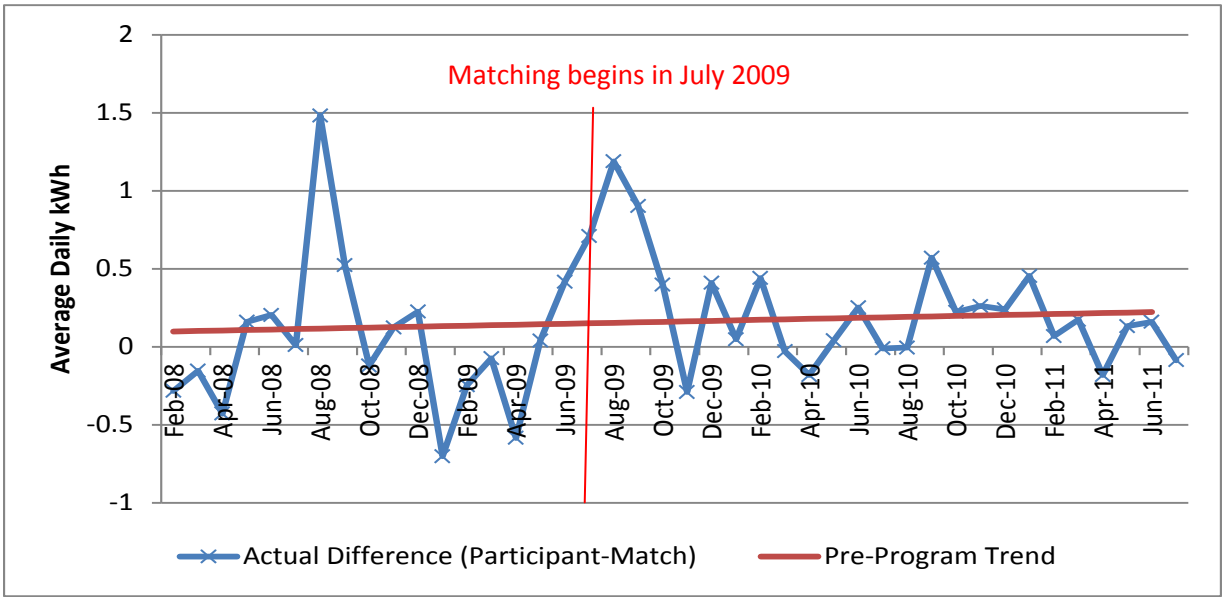


**Figure 2.** Comparison of the difference in average daily kWh consumption between participant households and their 24-month matches in the 41 months before program enrollment begins

## The Issue of Selection Bias in the Estimate of Program Savings

The analysis described above attempts to estimate the average program effect on program participants. The purpose of the matched comparison households is to provide an estimate of the counterfactual (baseline) energy use by participants –the energy use by participants if they were not in the program. As noted previously, matching estimators are designed to eliminate model specification bias, by assuring that the distribution of covariates X conditioning the counterfactual estimate is the same as that under treatment. With respect to energy use, by far the most important conditioning variable is pre-program energy use in the same billing period of the previous year. This variable, along with monthly fixed effects, accounts for about 95% of the variation in energy use over a 1-year period. The implication is that given a model that matches on pre-program energy use, with regression correction as advocated by Imbens and Wooldridge (2008) and used in Model 2, we are highly likely to generate an excellent counterfactual for participants with respect to observable variables.

Accepting that the analysis approach addresses model specification bias, we turn to the question of selection bias. In the current context, selection bias is the result of the counterfactual being derived from matches that systematically overstate/understate the true counterfactual energy use by participants during the program year due to unobservable differences between the two groups. It implies, in other words, that even though the participants and their matches behave the same on average for 24 months before the start of the program, in the absence of the program their energy use would not continue to be the same on average because unobservable factors cause the development of systematic differences in the energy use between the two groups.

For behavioral programs it is difficult to develop a convincing argument for selection bias given good matches based on pre-program billing history. A standard narrative concerning unobserved differences between participants and comparison households does not support the argument for selection bias. This narrative is that the participants are more likely than the typical household to behave like "energy hawks" –always on the lookout for ways to save energy—and that this behavioral characteristic is what drove them into the program. Given good matches over a long horizon, though, this argument is unpersuasive because the matches are observationally equivalent; the matches act *as if* they have a similar behavioral propensity with respect to energy savings. Note in Figure 2, for instance, that there is no trend in the difference between participants and their matches over a 3 ½ year period before the start of the program.

More generally, matches based on the energy use history account for selection bias that is due to "stable" differences between participants and the general non-participant population with respect to energy use. Suppose an underlying set of unobservable variables **Z** reflect a household's behavioral propensity to save energy, and these variables are correlated with participation in the program. One can reasonably expect that close matching on the energy use history will, on average, generate the same distribution of **Z** among the matched households as among the participant households. As observed by Stuart (2010),

> "This assumption [nonconfoundedness] is often more reasonable than it may sound at first since matching on or controlling for the observed covariates also matches on or controls for the unobserved covariates, in so much as they are correlated with those that are observed" (pg. 3).

In other words, the *behavioral* narrative for selection bias is necessarily reflected in a parallel *statistical* narrative. The statistical argument has to be that in the regression model there are

unobservable variables affecting energy use at or after the start of the program that are correlated with the participation decision.[2] Note, though, that to the extent these same variables affect energy use in the pre-program year, their effect is absorbed by the preconsumption variable $PREkWh_{kt}$, thereby mitigating against the associated selection bias.

The claim that longer matching horizons do a better job of driving selection bias from the analysis implies the assumption of greater stability of **Z**. There is no right/wrong answer to the question of the correct matching horizon, though to account for seasonal effects it is clear that the minimum match horizon should be 12 months. It is worth mentioning that, with respect to the claim that matching addresses selection bias, matching on demographic variables implies that **Z** is invariant over time –perfect stability—and relatively highly correlated with the matched demographic variables.

## A Pseudo-Test for Selection Bias

It is not possible to statistically test for selection bias, but Imbens and Wooldridge (2008) present a test that is suggestive. In the current context the logic of the test is that in the absence of selection bias there should be no difference between participants and matches in average energy use outside of the matching period and outside of the program period. A simple implementation of the test is to determine whether, given matching based on months $t=3,...M$ before the start of the matching period, average $D_{PM}$ in months t=1,2 is not statistically different than zero, and, similarly average $D_{PM}$ in months t=M+1, M+2, …M+n before the start of the program period is not statistically different than zero (where M+n months before the start of the program denotes the first month of the study period).

In the preliminary analysis we implemented this test by matching participants to nonparticipants using the 12 month period covering t=3,…14 before the start of the program, leaving for testing the two months immediately before the start of the program, as well as many months (27 months) before the start of the matching period. The test failed to reject the null hypothesis that treatment and control households had the same mean kWh/day in all months available for testing, with p-values of .87 in the month just before the start of the program and .95 two months before the start of the program. This result provides some reassurance that selection bias is not a significant issue when matching on past energy use, at least in the particular application examined here.

[2] It is possible that this correlation starts just before the start of the program, in which case matches may appear to be good but start to "separate". But this case is easily detected, per the pseudo-test discussed in the next section.

## Correcting for Selection Bias

The available evidence strongly supports the assumption that the analysis does not suffer from selection bias. Still, it is worth considering taking steps to correct for selection bias, because whether selection bias exists is not knowable. The standard correction for selection bias involves two-stage instrumental variables (IV) analysis. This approach requires identifying variables correlated with the participation decision but assumed to be *not* correlated with the error term of the regression model of monthly energy use used to estimate program savings (in this case, the regression model of Model 1). IV analysis necessarily involves a loss of efficiency in the estimate of program savings because the participation decision is replaced by a prediction of the participation decision. Moreover, in small samples such as used in this analysis, weak instruments –instrumental variables not highly correlated with the participation decision –can generate biased estimates of savings. IV analysis can be, in other words, a cure worse than the disease.

In principle, a survey of participants and their matches provides an outstanding opportunity to develop instrumental variables for program participation, as the analyst can define questions believed to be orthogonal to energy use but highly correlated with program participation. This approach requires that survey responses are representative of the population of participants and their matches, and that response rates are sufficiently high to justify the benefit of IV estimation in eliminating potential selection bias.

We conducted a small survey of participants and their matches, where matches were drawn from the list of ten first-round matches (see previous section, "Selecting Matched Comparison Households"). A total of 54 pairs of surveys for Energize households and matched comparison households were completed. Only three of the survey questions generated responses that were sufficiently correlated with the participation decision to warrant consideration as IV variables (absolute value of the correlation, $r$, in parentheses):

- "I always try new technologies before other people do" ($r$=.282);
- "I trust my utility" ($r$=.161);
- "I am more likely to change my actions if people I respect have already taken action" ($r$=.104).

Regressing these variables along with the covariates in Model 2 that vary across customers (in particular, $PREkWh_{kt}$ and the $EE_{kt}$ variables) on the participation decision –the first stage of IV estimation—generated a Wald statistic of 5.30. This is a very low value, indicating that instrumental variable analysis is highly problematic. The second stage of the IV analysis generated an estimate of program savings that was the wrong sign, wildly disproportionate (net savings over *negative* 10%), and not statistically significant.

The upshot is that although in principle, and in certain situations, a survey of participants and matches could be an ideal means of correcting for selection bias the via instrumental variables method, in practice it may be too difficult, too expensive and too risky to be attempted in an analysis unless selection bias is strongly suspected. This is an issue that warrants further research.

## Analysis Results and Discussion

Figure 3 presents average $D_{PM}$ for participants and their matches over the period February 2008 to September 2012. The month of July 2012 is dropped from the figure because that is the month when most participants transitioned into the program. The figure makes clear two features:

- During the pre-program period the difference in energy use between participants and their matches is small on average, especially in the year before the start of the program, and there is no trend in the difference;
- There is a sharp drop in the difference at the start of the program.

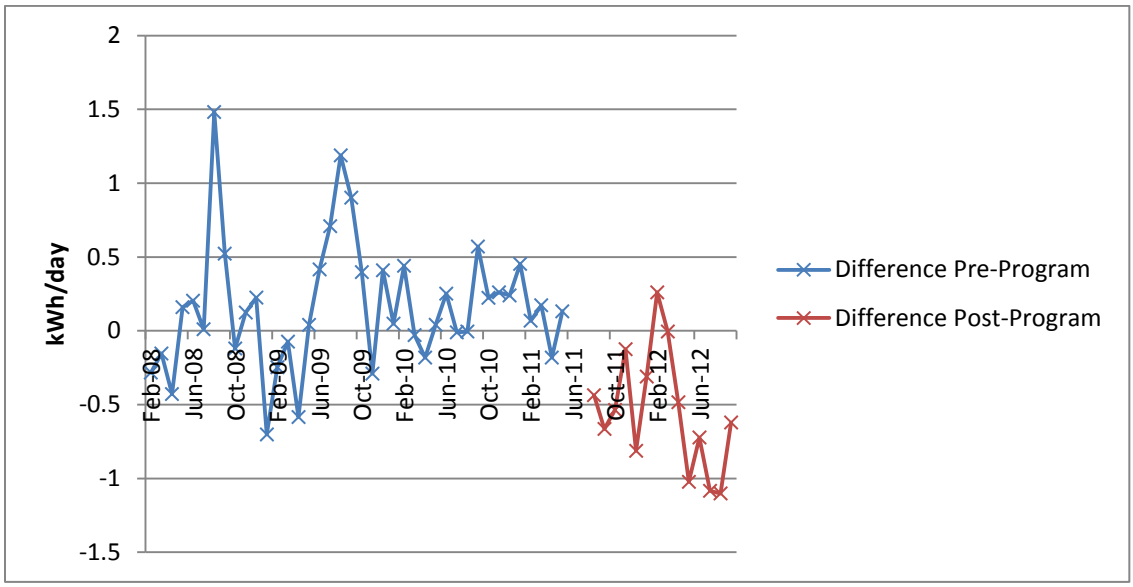


**Figure 3.** Difference between participants and matches in average kWh/day ($D_{PM}$) over the study period (24-month matches, Energize customers)

Model 1 generated an estimate of average program savings of 1.49% (standard error=0.63%, errors clustered at the customer level), whereas Model 2 generated an estimate of 1.36%. An issue with Model 2 is the difficulty of generating standard errors; because the estimate uses the regression equation to adjust the matching estimator, standard errors are not readily produced. Using an approximation method we estimated a standard error for Model 2 of 0.24%.

Models 1 and 2 were also estimated using 12-month matches using the same matching approach as used to generate the 24-month matches. Estimated average program savings for these matches were 1.93% for Model 1(standard error=0.64%) and 1.99% for Model 2. For Model 1, the savings estimates for the 12-month and 24-month matches are not statistically different at any reasonable confidence level, but nonetheless deserve additional investigation. For Model 2 the issue of whether 24-month and 12-month matches generate statistically different estimates of program savings awaits calculation of standard errors, though the similarity of results for Models 1 and 2 suggests not. In the discussion here we focused on the 24-month matches because they appeared to provide a good fit to participants, with no trend in energy use

during the pre-program period. The issue of differential trends in energy use between participants and their matches during the pre-program period is generally problematic and perhaps not well recognized by practitioners. The use of lagged energy consumption in the post-program regression models ($PREkWh_{kt}$ in the models used here) will not remove the trend entirely.

## Conclusion

This paper makes five points about attempts to identify energy savings from opt-in programs using a comparison group:

1. Developing a comparison group based on historical energy use is likely to minimize model specification bias in regression analysis.
2. Although statistical estimates of savings from behavioral opt-in programs are vulnerable to selection bias, matching based on historical energy use is likely to substantially reduce selection bias.
3. Matching on historical energy use provides the opportunity for a pseudo-test that is *suggestive* of the presence/absence of selection bias.
4. When matching on historical energy use, the analyst should check that there is no trend in the difference in the average energy use of participants and their matches during the pre-program period. Such a trend reflects differences between the two groups and time-varying unobservables and is thus an indication of selection bias.
5. In the case where the analyst is concerned about selection bias even in the case where a comparison group is developed via matching on historical energy use, surveys of participants and their matches may provide an avenue for correcting for selection bias via instrumental variables regression.

## References


Abadie, A. and G.W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects". *Journal of Business and Economic Statistics* 29(1):1-11.

Stuart, E.A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward". *Statistical Science,* 25(1):1-21.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199-236.

Imbens, G.M. and J.M. Wooldridge. 2008. "Recent Developments in the Econometrics of Program Evaluation." NBER paper 14251.