Not Just Another Pretty Formula: Practical Methods for Mitigating Self-Selection Bias in Billing Analysis Regressions

Dr. Miriam L. Goldberg and G. Kennedy Agnew, DNV GL, Madison, WI
Dr. Meredith Fowlie, University of California Berkeley
Dr. Kenneth Train, NERA
Brian Arthur Smith, Pacific Gas and Electric Company

ABSTRACT

Several trends are focusing renewed attention on energy consumption data analysis for energy efficiency program impact evaluation incorporating treatment and comparison groups. The trends include changes related to program design, metering technology, analytic tools, public policy, and potential energy efficiency trading markets.

A key consideration in these contexts is how to design both the analysis and the comparison group specification to minimize self-selection bias. The potential for self-selection bias exists with any voluntary program. For energy efficiency programs, the concern is that customers who are interested in taking a program offer may tend to be changing consumption apart from the program in ways that are different from those who don't. Participants may be more inclined to adopt the program measures on their own, more inclined to be taking other energy-reducing actions, or conversely more likely to have life events that are increasing consumption.

This paper describes analytic methods to limit self-selection bias in program impact estimation using billing analysis. We identify key program conditions to consider in determining which types of self-selection correction are needed. A new self-selection correction method is presented that may be useful for analysis of many voluntary programs. Both the validity and statistical power of the proposed analytic methods can be enhanced by combining these methods with a random encouragement design.

The goal of the paper is to provide practical guidance on how to limit self-selection bias, and how to assess its potential, as well as to dispel perceptions that self-selection correction methods are too challenging to implement in most contexts.

Introduction

This paper describes methods to estimate the net savings of energy efficiency programs using customer-level consumption data analysis, also known as billing analysis for net savings. The specific focus is on mitigating self-selection bias when using a comparison group drawn from the program-eligible population. A key point of this discussion is that the use of a comparison group by itself is not necessarily sufficient to identify net savings.

Not addressed in this paper is the effect of spillover to non-participants. In some contexts, the existence of an efficiency program can affect the energy consumption of non-participants through various channels. Estimating or accounting for these spillovers is beyond the scope of this paper.

This paper is intended for use by evaluators who want to understand the techniques better, as well as by program administrators, regulators, and other stakeholders who want to understand what is and isn't possible. Most technical details are deferred to references. Key lessons are summarized in the Conclusions.

The paper considers a simple impact estimation structure as a framework for exploring the limitations of some common techniques, and the value of proposed alternatives. We consider the use of comparison groups with and without random assignment. In particular, we consider two random assignment designs: randomized control trials (RCT) and random encouragement designs (RED). We identify situations under which these designs can be used to estimate the net savings of interest, and delineate why they cannot always be used. We then describe a new alternative approach to address self-selection when the random assignment procedures and standard analysis are not applicable or not sufficient.

We review how "instrumental variable" strategies incorporating a model of program participation can eliminate certain sources of selection bias, and how this approach can be enhanced by use of an RED. The new estimation procedure proposed extends this use of the participation model, in a way that is both simpler and more robust compared to an earlier extension[1]. We show how, in situations where the RED design with a standard analysis does not by itself provide the net savings of interest, the new method can do so, subject to additional assumptions. We conclude by summarizing the applicability and limitations of each of the methods discussed for different situations, and identify some next steps for assessing the trade-offs empirically.

Background

Renewed Interest in Billing Analysis

The use of consumption data regression analysis for program net savings estimation is of increasing interest in California with the adoption of AB802 which emphasizes normalized metered usage data as the basis for savings estimates. Additional interest in these estimation approaches has been generated by the recent publication of the Uniform Methods Project Chapter 8, (National Renewable Energy Laboratory, 2013), the use of random assignment methods as the basis for ongoing savings estimation from Home Energy Reports programs, (e.g., Applied Energy Group 2014) as well as the increased use of random assignment methods for pilot programs and special studies (e.g. DNV GL 2015).

Gross and Net Savings

Net program savings is the difference between participants' consumption with versus without the program in place. As noted, nonparticipant spillover is not addressed in this paper, and is assumed for discussion purposes to be zero. The effect of the program on participants' consumption includes the effect of the program on the measure adoption, along with any incidental effect of the program on adoption of other measures or behavioral modifications outside the program (participant spillover) as well as any economic takeback effects.

Gross program savings is the difference between participants' consumption with versus without the measures targeted by the program in place. To the extent the measure adoption itself induces a customer to adopt other measures or to alter energy-using behavior in other ways, these effects are also part of the gross program savings. These are effects of the measure, regardless of how the program influenced its adoption.

Why Self-Selection Matters

Self-selection is a challenge for comparison group methods whenever customers are not randomly assigned to participate or not participate in the program. Self-selection means that, even if program participants can be matched with observationally similar non-participants, those who choose to join a program are different from those who don't, in ways that could affect changes in energy consumption apart from the participation choice. As a result, the analysis cannot separate the program or measure effect from the effect of being in the "inclined to join" group. The effects of self-selection in comparison group analyses can be substantial and meaningful. All methods to correct for self-selection in non-RCT contexts have some limitations. This paper describes the nature of the problem and how effective various methods can be for different situations.

Framework for the Discussion

Key Factors Affecting Participant-Comparison Group Differences

When we talk about the need for the comparison group to be similar to the participant group, we usually think about observable factors such as premise characteristics, equipment, and demographics/firmographics. In practice, we often use prior consumption to represent their combined effects. While these factors can all be important, there are other factors that can also determine energy consumption trajectories and are harder to observe directly:

1 The "Double Inverse Mills Ratio" method was explored in XENERGY, 1996.

Natural Adoption Rate, Free Ridership. Natural adopters are those who would have adopted the program measure on their own if the program didn't exist. Participants who are natural adopters, also called free riders, have zero net savings. For a comparison group to mirror participant consumption absent the program, the comparison group must have the same proportion of natural adopters (outside the program) as the program has free riders. That is, customers inclined to adopt the program measure(s) on their own must participate at the same rate as customers with no natural inclination to adopt.

For many programs, however, natural adopters who are aware of the program will be more likely to become participants. Even if the natural adoption rate is low across the population offered the program, and even if program awareness is low across the population, it's still likely that customers who would want to implement the measure in any case will be more likely to pay attention to program messages, and to take the program benefits, compared to customers who are not naturally interested in the measure and still need to be convinced to adopt it. As a result, the proportion of natural adopters among the comparison group will tend to be lower than the proportion among participants. Thus, even accounting for other customer characteristics, the comparison group will not by itself "net out" the effect of free ridership.

Non-Program-Measure Changes Different for Participants. Self-selection can also be a problem in the opposite direction. For many programs, customers with otherwise similar demographics may be more likely to participate when other events in the household are occurring that tend to increase consumption, such as adding a family member or undertaking a major renovation. Thus, the frequency of these other types of changes may be lower among the comparison group. Information on such life events are not typically available from demographic data or common customer surveys.

A variety of other factors lead some customers to participate in a particular year and others not to. Thus, the fact that a customer chose to participate in a given year is itself an indication that something was going on for that customer that's not explained by the variables we have, and that could be related to naturally occurring change. All of these are factors that can lead to naturally occurring change being systematically different for participants than for nonparticipants.

Measure Applicability. Many programs offer measures that make sense for only a limited portion of their customers in any given year. The consumption change for customers who have no need of or use for the program measures is likely to be different from that of customers who need the equipment upgrades or improvements related to the program measures. For example, participants in a typical (voluntary) HVAC program would mostly be replacing HVAC equipment with or without the program, while a general population comparison group would include a large proportion of customers who have no reason to be changing in their equipment. As a result, both the natural adoption rate and other natural changes would be very different among the general comparison group compared to those for the participants.

Analysis Framework

For simplicity, we assume that we are observing the change in annual consumption between the pre- and post-participation years. The methods described can be extended to models of monthly consumption fit across pre- and post-participation months.

We represent each customer's change in consumption as the combination of the naturally occurring change *noc*, which would have occurred without the program, and the potential net savings *pnet* the customer will have if they join the program. We use the convention that positive *noc* means an increase in consumption, and positive *pnet* means a decrease in consumption. That is, for customer j, the change in energy consumption ΔE_j is given by

$$\Delta E_j = noc_j - pnet_j * P_j$$

where P_j is a 0/1 dummy variable indicating customer j participated in the program during the period under study.

For a customer who would adopt the measure on their own without the program, the effect of the measure is included in the naturally occurring change noc_j, while the potential net savings $pnet_j = 0$. For a customer who would not adopt on their own, the potential net savings is the gross measure savings. For a

customer who does join the program, the net savings realized is the potential net savings. For a customer who doesn't join, the realized net savings is zero.

This framework is useful for identifying what effects are and aren't accounted for by different billing analysis tools. The next section describes the concerns that arise in billing analysis, and methods to address these concerns.

Concerns for Billing Analysis, and Tools to Address Them

Evaluators have many strategies available to them when using energy consumption data to construct estimates of average net energy savings. In what follows, these methods are reviewed with an emphasis on the extent to which selection bias is mitigated or addressed.

1. Use of a Comparison Group

If we start by looking at change in participants' consumption as the basis for determining the program impact, the first concern is that there may be underlying changes affecting all customers that are unrelated to the program. For example, such changes could include economic factors, social habits, or technology penetration. One common way to control for such changes is to use a comparison group. In the simplest analysis, the average the change for the comparison group is subtracted from the average change among the participants. This is the Difference in Differences (DID) calculation, where the average savings per participant is calculated as

$$S = \Delta E_P - \Delta E_C.$$

where ΔE_P and ΔE_C, respectively, are the average change in consumption for participants and the comparison group.

Equivalently, we use a simple regression formula representing the consumption change ΔE_j for customer j as a function of the participation dummy:

$$\Delta E_j = a - bP_j$$

and the estimated savings per participant is given by

$$S = b.$$

Whether we use the DID calculation or the simple Ordinary Least Square regression on the participation dummy, the savings estimate S will have the same value.

Limitations. The limitation of using either of these simple forms is that the general population of nonparticipants may not be similar to the participants absent the program. Considering the DID form, we have

$$\begin{aligned} S &= \Delta E_P - \Delta E_C. \\ &= (nocP - pnetP) - nocC \\ &= (nocP - nocC) - pnetP \end{aligned}$$

The savings we want is the average potential net savings for the participants, pnetP. The savings estimate S from the DID calculation or the simple regression is this net savings of interest plus the difference in naturally occurring change between participants and the comparison group. If the average naturally occurring change isn't the same for the two groups, we have a biased estimate of net savings. Thus, the use of the comparison group partially nets out the naturally occurring change among the participants, but only partially if the two groups' naturally occurring change isn't the same. This same bias will be present with the regression form of the estimate.

2. Randomized Control Trials (RCT)

Randomized Control Trials (RCT) are often referred to as the "gold standard" for study design. If customers can be assigned randomly to be program participants or not, there is no role for self-selection, and no potential for self-selection bias in the DID or equivalent regression analysis. Because of the random

assignment, average naturally occurring change is expected to be the same for participants and the control group, and the estimated savings is expected to equal participant average net savings, as desired.

Limitations. Random assignment to be in a program or not is inconsistent with the way most programs are delivered, outside of pilots, and certain behavior programs that don't provide direct tangible benefits. Usually, customers cannot be forced to participate in a program. And even when participation can be required for some customers, denying participation to other customers is often politically or ethically difficult. Situations where net savings estimation is a challenge are precisely those situations where program participation is voluntary.

3. Matching and Additional Variables

If there's a prevailing trend toward being more frugal or more expansive in energy consumption, the contribution of that trend to the change in consumption is likely to be different for homes with different characteristics. Larger homes might have a larger magnitude change. Different end uses will change in different ways. Certain kinds of customers will respond to the prevailing trend in different ways. For all these reasons, we'd like the mix of home and household characteristics in the comparison group to be similar to that among the participants. This similarity can be achieved to a certain extent by selecting one or more comparison group customers to match each of the participants. Common matching approaches match on pre-period consumption. Examples are given in Churchwell 2013.

Differences between the comparison group and the participant group can also be controlled for by including household characteristics as additional terms in the regression. Neighborhood average characteristics from Census data can be incorporated without requiring supplemental survey information. When additional explanatory terms X_k are added, the regression takes the form

$$\Delta E_j = a_0 + a_1 X_{1j} + a_2 X_{2j} + \ldots + a_m X_{mj} - bP_j.$$

We represent this expanded regression equation in the more compact matrix form

$$\Delta E_j = \mathbf{X}_j \boldsymbol{\alpha} - bP_j$$

where the bold term $\mathbf{X}_j$ indicates the set of predictors 1, X_{1j}, X_{2j}, ..X_{mj} and the bold symbol $\boldsymbol{\alpha}$ represents the corresponding set of coefficients a_0, a_1, .. a_m.

Of course, a variety of more complicated structures can be used, including nonlinear forms. We keep to the linear regression form for illustration of the methods. Matching and the expanded regression form may each be used alone or they may be used together.

Limitations. The limitation of controlling for observable characteristics by matching or by including them in the regression is that the characteristics we can observe and match on typically don't account for all factors that affect both naturally occurring change and participation. It's not necessary to have a tight prediction of the naturally occurring change in total, but it is important that any unaccounted for change is expected to be the same for participants as for the comparison group. If the participant's change is likely to be systematically different from the comparison group, even after including the additional explanatory variables, the estimation bias remains. Such a systematic difference arises if there are factors that affect both the likelihood of participation and the naturally occurring change.

As described above, key factors that can lead to differences in naturally occurring change that aren't accounted for by observable characteristics include measure applicability, life events that trigger taking action, and natural adoption rates. For the moment, we assume natural adoption rates are negligible, and consider other ways that naturally occurring change could be different between the two groups.

4. Instrumental Variables (IV-Only)

Whenever there are factors that determine both the participation decision and naturally occurring changes in energy consumption (noc), we're left with an unaccounted for difference between participant and non-participant naturally occurring change. This difference creates the potential for bias in our estimate of net savings. Even with a matched comparison group or additional explanatory variables in the regression equation, this

selection bias will not be mitigated if unobserved determinants differ systematically between participants and non-participants in the analysis.

A common way to address this type of problem is through the use of *instrumental variables.* We denote the method here as "IV-Only" to distinguish it from another method that combines IV with an additional term. A simple version of the IV-Only method proceeds in two steps:

Step 1. Participation Model. Estimate a model of selection (or participation) into the program. Participation is modeled as a function of a set **z** of observable variables. A corresponding set of coefficients $\boldsymbol{\gamma}$ summarizes the relationship between the predictors **z** and the participation decision. This gives us a predicted probability model of the general form

$$\hat{P}_j = f(\mathbf{z}_j, \boldsymbol{\gamma}).$$

Step 2. Outcome Equation. Substitute each customer's predicted participation probability $\hat{P}_j$ for the actual participation dummy P_j in the primary regression equation. This gives the regression

$$\Delta E_j = \mathbf{X}_j\boldsymbol{\alpha} - b\hat{P}_j .$$

The terms included in the participation model's predictors **z** must satisfy three conditions:

Z1. The participation predictor set **z** must include all the predictor variables X_k that appear in the outcome equation (Step 2). If instead there's a consumption predictor X that also affects participation but is left out of the participation model, the contribution of X to participation will get picked up by the regression in the coefficient $\boldsymbol{\alpha}$ rather than in the participation coefficient b, leading to a biased estimate of net savings.[2]

Z2. The participation predictor set **z** must include one or more variables that are *not* among the direct predictors of consumption **X**, and are appropriately excluded from the outcome equation[3]. If instead the only variables available to predict participation are also direct drivers of consumption change, the regression will rely entirely on the functional form $f(\mathbf{z}_j, \boldsymbol{\gamma})$ to separate direct consumption effects from participation effects.

Z3. If there are additional (observable or unobservable) consumption drivers that are left out of the outcome equation, the participation predictors **z** must be unrelated to any of these omitted terms. If instead there is an omitted consumption driver X* that is related to the participation drivers **z**, the outcome regression equation will tend to pick up the effect of X* in the coefficient of $\hat{P}$. In this case, at least some of the bias we're attempting to correct will still be present.

For example, suppose the analysis data set includes house size from assessor's data, and neighborhood average education and income from Census data. We could include size and education as direct predictors of consumption in the **X** variables, then let the participation predictors **z** include size, education, and income. Then the variable **z** satisfies the condition Z1. However, we would only satisfy the condition Z2 if we believe income has no direct relation to the change in consumption (that's not already captured by the size and education terms in the primary equation). This assumption should at least be tested. Below, we describe how use of a RED allows us to satisfy condition Z2 unambiguously.

Now consider another determinant of participation that's not unobservable from any available data: the occurrence of particular life events that lead to increased attention to energy consumption. It's reasonable to think that the occurrence of such events in a particular year is unrelated to size, education, or income. Thus, condition Z3 might be satisfied.

To see why the instrumental variable approach works, consider again the simple regression

$$\Delta E_j = a - bP_j$$

[2] If a particular consumption driver X has no relationship to the participation decision, that variable could be omitted from the participation equation. However, to avoid bias it is better to retain all the X terms in the participation model unless they are found to have no effect on the result.

[3] This requirement is known as the *exclusion restriction*.

and suppose there is some variable that substantially determines the choice to participate in the program, but has no direct effect on subsequent energy consumption. This variable can be used to split the combined set of participating and nonparticipating customers into two groups, one with high participation probability P_{HI} and one with low probability P_{LO}. We have corresponding average consumption change for each group ΔE_{HI} and ΔE_{LO}. The simple regression equation will give the estimate

$$b = (\Delta E_{HI} - \Delta E_{LO})/(P_{HI} - P_{LO}).$$

Using our decomposition of the change in consumption

$$\Delta E_{HI} - \Delta E_{LO} = (noc_{HI} - (pnet*P)_{HI}) - ((noc_{LO} - (pnet*P)_{LO})$$
$$= (noc_{HI} - noc_{LO}) - ((pnet*P)_{HI}) - (pnet*P)_{LO}).$$

The term $(pnet*P)_{HI}$ is the average realized net savings among the high-probability group, times its participation probability, and similarly for $(pnet*P)_{LO}$. Hence the numerator becomes

$$\Delta E_{HI} - \Delta E_{LO} = (noc_{HI} - noc_{LO}) + net_{HI}P_{HI} - net_{Lo}P_{LO}.$$

Now if there's no relationship between the participation probability and naturally occurring change, the difference (nocHI – nocLO) is expected to be zero. If in addition the potential net savings is unrelated to the participation probability, the average net savings is the same for those who participate from the high-participation probability group as for those who participate from the low-probability group. Thus, the expected difference in consumption change for the two groups is

$$E(\Delta E_{HI} - \Delta E_{LO}) = E(noc_{HI} - noc_{LO}) - E(net_{HI}P_{HI} - net_{Lo}P_{LO})$$
$$= 0 + net_{AVG}(P_{HI} - P_{LO}).$$

Thus, the coefficient b is expected to yield the correct average net savings.

This relationship holds in the more general case, where the predicted probability isn't just two different values but varies across customers based on their variables z_j, and with additional explanatory variable Xj included in the primary regression. That is, there will be no bias due to different naturally occurring change, and the substitution of participation probability $\hat{P}$ for the participation dummy P (instrumental variables method) gives an unbiased estimate of net savings. This property of the the outcome equation providing an unbiased estimate of average net savings per participant holds provided the following both are true:

1. The conditions Z1-Z3 above on the predictor variables z are met.
2. The potential net savings is not related to the probability of participation. That is, (a) customers who have higher or lower potential net savings have the same probability of participating, and (b) customers who have higher or lower probability of participating have the same average potential net savings.

Limitations. The limitations of the IV-Only method are linked to these two provisos. First, it's necessary to find good participation predictors **z** that satisfy conditions Z1-Z3. If the participation model is not able to distinguish well between high and low probability of participation, the savings estimate will tend to have high variance compared to that from the biased direct regression that doesn't use the IV approach.

For example, in the illustration above, the DID estimator is comparing the change in consumption for two groups, one with high participation rate and one with low. If the participation model is weak, the two groups will have only a small difference in participation rate, and the effect of higher or lower participation will be harder to detect above the noise.

The second limitation of the IV method as a means to estimating net savings for the full program is that the method requires that the participation decision be unrelated to the net savings a customer will have if they join the program. As indicated in the Framework section, this is a reasonable assumption only if the program is applicable to everyone in the comparison group, and the free rider rate is negligible. Low-income programs using future participants as a comparison group for the current participants may satisfy these conditions.

5. Random Encouragement Design with IV-Only

One way to enhance the effectiveness of the IV approach is to use a Random Encouragement Design (RED). With this design, a random set of customers is selected to receive additional encouragement to join the program. The additional encouragement could be additional messaging and outreach, or a higher incentive level. Examples of RED for energy efficiency program evaluation are described in State and Local Energy Efficiency Action Network. 2012. Because of the random assignment, the encouragement indicator is a perfect instrument (provided that the encouragement substantially increases the probability that encouraged customers to participate). That is, the encouragement indicator is a predictor of the likelihood of participation, but is unrelated to naturally occurring change.

Even with the RED, it's still valuable to include other participation predictors besides encouragement. Thus, for example, we would include size, education, and income as direct predictors X of consumption change, and use these plus the encouragement variable for the participation predictor **z**, to meet conditions Z1 – Z3**.** Incorporating the RED with the study improves the participation model, and thereby reduces the variance of the savings estimate. The variance is still likely to be greater than for the biased estimate not using IV or RED.

Limitations. The limitation using the IV method with RED remains that the validity of the method as an estimator of total program net savings requires that participation be unrelated to the potential net savings a customer will have if they participate. In particular, the IV-Only method with RED will provide the average net savings for those who participate with encouragement but otherwise would not. If there is free ridership in the program, it's likely that the free rider rate is lower among those who require extra encouragement than among those who participate without extra encouragement. As a result, the net savings for the incrementally encouraged isn't a reliable estimate of the net savings for the base program without encouragement. This point is sometimes overlooked in interpreting RED results. If free ridership is likely to be negligible, this limitation might be less of an issue. Alternatively, the IV-Only result could be considered to be an upper bound for the base program net savings.

6. The IV-EXN Method

A key condition we would like the comparison group to account for is the rate of natural adoption among the participants. However, as discussed in the Framework section, it's unlikely that the natural adoption rate among participants is the same as among nonparticipants. Put another way, it's unlikely that natural adopters participate in the program at the same rate as natural non-adopters. As a result, among those who do participate, the proportion of natural adopters will be greater than among those who don't. Thus, free ridership will be only partially netted out by the comparison group in a DID or equivalent analysis.

The IV-Only method doesn't fix this problem. The IV-Only method can eliminate any expected difference in naturally occurring change. Even so, the method won't give an unbiased estimate of program net savings if customers with different potential savings participate at different rates.[4] Thus, IV-Only is likely to be biased if free ridership is present.

Consider again a participation model that divides customers into a high and a low participation likelihood group. This is what a RED with no additional predictors would do—divide the customers into those who received supplemental encouragement and those who did not. As indicated above, the savings estimate is given by

$$b = (\Delta E_{HI} - \Delta E_{LO})/(P_{Hi} - P_{LO})$$

which has expectation

$$E(b) = E[(net_{HI}P_{HI} - net_{LO}P_{LO})/(P_{Hi} - P_{LO})].$$

When free ridership is present, we no longer assume that the average net savings for those who participate from the high-probability group is the same as that for the low-probability group.

[4] That is, an instrumental variables approach yields an unbiased estimate of the local average net savings among those customers whose participation status was determined by the instrument or encouragement. Net savings among this subset of customers could be different from the average net savings across *all* program participants.

In the RED context, the savings estimate b gives the average net savings per participant for those who participate with encouragement but otherwise not. We would typically expect that the free ridership will be higher among those who participate in the base program than among those who need extra encouragement to join. In that case, the savings provided by the RED with the standard IV analysis would tend to overstate net savings.

When the IV is applied without a RED, the interpretation of the savings coefficient b is less straightforward. However, it is still the case that the coefficient does not provide the average net savings for the program as a whole, unless it's assumed that net savings is the same regardless of participation probability.

An approach that can provide an unbiased estimate of net savings for any subset of customers in the study expands the primary regression equation to include an additional term. The additional term allows for the average net savings to vary systematically as a function of the participation probability. The expanded regression form is

$$\Delta E_j = \mathbf{X_j}\boldsymbol{\alpha} - b\hat{P}_j - c\,\hat{P}_j\,EXN_j$$

where EXN_j is determined from the fitted participation probability model, and is related to the expected potential net savings for customer j.

The simplest version of the expected net savings term is available when a normal distribution is assumed for each of the underlying variance components in the primary regression and in the participation model. In this case, the participation model form is a probit, which can be written as

$$\hat{\boldsymbol{P}}(\mathbf{z}_j\boldsymbol{\gamma}) = 1 - \Phi(\mathbf{z}_j\boldsymbol{\gamma}).$$

The term EXN_j for this case is the Inverse Mills Ratio,[5] calculated for each customer j as

$$IMR_j = \phi(\mathbf{z}_j\boldsymbol{\gamma})/(1 - \Phi(\mathbf{z}_j\boldsymbol{\gamma})).$$

The expanded regression equation becomes

$$\Delta E_j = \mathbf{X_j}\boldsymbol{\alpha} - b\hat{P}_j - c\,\hat{P}_j\,IMR_j$$

From the estimated model, the average savings per participant net_P is calculated from the estimated coefficients and the average value of the IMR for the participants, IMR_P:

$$net_P = b + c\,IMR_P.$$

Limitations. As for the IV-Only method, the IV-EXN method requires that the conditions Z1-Z3 for the participation model be satisfied. Also as for the IV-Only method, the variance of the estimated savings tends to increase when the participation dummy Pj is replaced by the predicted participation $\hat{P}_j$. This problem can be exacerbated with the IV-EXN method, because the terms $\hat{P}_j$ and $\hat{P}_j$ EXNj tend to be correlated.

An additional limitation of the IV-EXN method is that it depends on the assumed functional form of the participation prediction model. Simulation tests described in DNV GL (2017) indicate that when the functional form is the probit probability model with corresponding IMR term, the method appears to be robust to certain kinds of departures from normality.

7. IV-EXN with RED

As with the IV-Only method, using the IV-EXN method together with RED can mitigate the limitations related to the participation model. An effective RED can improve the precision of the IV-EXN savings estimate and ensure that the exclusion restriction (requirement Z2) is satisfied.

Limitations. IV-EXN limitations related to the distributional dependence of the method remain.

[5] The EXN term is the expected difference between the net savings for customer j and the average potential net savings for the population, given that customer j did participate, and given the customer's predicted participation probability. The normal distribution is particularly easy to work with, leading to the probit participation model and IMR as the EXN term.

Summary and Conclusions

Table 1 summarizes the issues successfully addressed by each of the methods discussed. Without an RCT, a comparison group helps to control for general increases or decreases unrelated to the program (naturally occurring change). A general comparison group leaves bias due to the trends being different for different types of customers. This bias can be mitigated by using matched comparison groups or including customer characteristics as terms in the regression. However, these methods can still leave substantial bias if the tendency to join the program is related to other circumstances that tend to increase or decrease consumption. The IV-Only method can mitigate this bias due to unobserved factors that affect both participation and naturally occurring savings, but not if those factors are related also to potential net savings. The IV-EXN method can eliminate bias due to a relationship between potential net savings and the likelihood of participating, but depends on the Normal (or other specific) distribution of the residual errors—the unexplained portion in the models.

Table 1. Billing Analysis Error Sources Addressed by Different Methods.

Potential Error Source or Issue	Error Type	Method to Address the Error Source							
		RCT	General Comparison Group	Matched Comparison Group	Regression w/ Additional Variables	IV-Only	IV-Only with RED	IV-EXN	IV-EXN with RED
General tendency to increasing or decreasing consumption apart from the program	Bias	X	X	X	X	X	X	X	X
General consumption trend has different magnitude change for different household characteristics	Bias	X		X	X	X	X	X	X
Observable characteristics don't account for key factors that determine both tendency to join the program in a given year and naturally occurring change	Bias	X				X	X	X	X
Hard to get participation predictors with good predictive power	Variance	N/A					O	X	X
Natural adopters are more likely to join the program than natural non-adopters	Bias	X						X	X
Potentially poor precision of IV-EXN estimate	Variance	N/A							O
Potentially incorrect distribution assumption for IV-EXN	Bias	N/A							

X: Method addresses the issue
O: Method partially addresses the issue
N/A: Issue doesn't apply when the method is used.

Both of the IV-based methods have the disadvantage that they increase the variance of the estimated savings. The variance inflation may be less of an issue today than in the past, as rich data analysis including tens of thousands of customers in both the participant and comparison group is now practical. Still, getting reliable estimates from either IV method requires good predictors of participation that are not also closely correlated with direct drivers of consumption apart from program effects. Using an RED, if the encouragement is effective, can improve the power of the participation model and the precision of the estimates.

Table 1 describes issues in terms of naturally occurring savings and potential net savings. Table 2 describes what self-selection correction is likely to be needed based on program characteristics. Measure applicability—specifically, program measures not applicable to the full comparison group-- is one key factor that leads to differences in naturally occurring savings between participants and the comparison group. A second key factor that can lead to such a difference is participation being triggered by life events that increase consumption. Neither of these factors is likely to be observable from available data.

Free ridership will also typically lead to naturally occurring change being different for participants than for the comparison group. More importantly, free ridership will tend to mean that potential net savings is different for participants than the comparison group, making the IV-Only method still biased.

Table 2 gives general guidelines. There are always nuances to the choice of methods for particular contexts. In addition, while one of the bias reduction methods may be indicated, it will still be necessary to determine empirically if the variance increase creates a worse problem than is solved.

Table 2. Program Characteristics Indicating a Need for IV-Only or IV-EXN Self-Selection Correction

Conditions:			**Self-selection correction needed:**
Is random assignment used in the program delivery?	**Is there free ridership?**	**Do participants and comparison group customers have the same naturally occurring change in consumption? (Are the measures applicable to the full comparison group? Is there minimal association between increased consumption and a decision to join?)**	
Randomized control trials (RCT)			No correction is needed
Random encouragement design (RED)	No		IV-Only
	Yes		IV-EXN
No random assignment	No	Yes	No correction is needed
		No	IV-Only
	Yes		IV-EXN

Making it work

As the tables indicate, selecting an appropriate method requires understanding of how the program is operating in the market, and an initial assessment of what drives participation. Implementing either self-selection correction further requires data on customer characteristics related to the decision to participate or not.

Obtaining variables that predict participation well--but do not directly affect energy consumption trajectories--is a challenge. While many variables may be available for program participants, corresponding detail is rarely available for nonparticipants. If surveys are used to collect data to support participation models, we trade survey nonresponse bias for the initial self-selection bias. Without such surveys, the participation models are limited to indicators available from utility customer information systems, together with Census area average demographics, or imputed values from commercially available data bases. These variables may be only weak indicators of key participation factors including natural inclination to adopt efficiency measures, the applicability of particular measures, or unrelated consumption changes.

Key Lessons

Following are key lessons from this work.

- Even with explanatory variables included (or matching), billing analysis regression without self-selection correction terms is likely to be biased for net savings unless (a) free ridership is negligible and (b) there is no relationship between participation and other factors that tend to increase or decrease consumption.

- Adding the IV term by itself will correct for the correlation between participation and other factors that change consumption, but bias is still likely if there is free ridership. In particular:
    - RED with standard analysis provides net savings due to incremental encouragement, but is likely to overstate net savings of the base program if there is free ridership.
- Adding the IV-EXN terms can provide an unbiased estimate of net savings even when there is free ridership, and appears to be somewhat robust to departures from the normal distribution assumption.
- Use of the IV-Only or IV-EXN methods requires good predictors of participation, otherwise the savings estimates will have high variance.
- Use of an RED can improve the viability of both the IV-Only and IV-EXN methods.

Next Steps

The next steps in developing this work will be to assess the performance of the IV-EXN method in practice. This work will include applications to existing data sets as well as additional simulations using parameters based on particular real-world examples, with alternative models of the underlying participation drivers. Correction form for other distributional assumptions will also be explored. Key questions to be considered include

- How can we get variables that account well for measure applicability and other participation drivers?
- With realistic simulations, what are the bias and variance using no correction, IV-only, or IV-EXN?
- How should an RED be designed to support the IV-EXN approach?

References

Applied Energy Group. 2014. SCE's Home Energy Report Program Savings Assessment: Ex-Post Evaluation Results, Program Year 2013. Prepared for Southern Californial Edison. http://www.calmac.org/publications/SCE_2013_HER_Evaluation_Final_Report_v10-24-14.pdf

Churchwell, C. A. et al. 2013. "Default Critical Peak Pricing for Non-residential Customers: Do Demand Reductions Persist? Are the Reductions Reliable?" *Getting It Done! Evaluation Today, Better Programs Tomorrow*, Chicago, IL: International Energy Program Evaluation Conference, August 2013.

DNV GL, 2017. A White Paper: Mitigating Self-Selection Bias in Billing Analysis for Impact Evaluation, Prepared for Pacific Gas and Electric Company, San Francisco (to appear)

DNV GL, 2015. Impact Evaluation of PSE Web-Enabled Thermostat Program. https://conduitnw.org/_layouts/Conduit/FileHandler.ashx?rid=2965

National Renewable Energy Laboratory. 2013. "Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol," *The Uniform Methods Project: Methods for Determining Energy Efficient Savings for Specific Measures.* Prepared by K. Agnew and M. Goldberg, DNV GL. https://energy.gov/sites/prod/files/2013/11/f5/53827-8.pdf

State and Local Energy Efficiency Action Network. 2012. Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. http://behavioranalytics.lbl.gov.

XENERGY, 1996. *Net Savings Estimation: An Analysis of Regression and Discrete Choice Approaches.* Prepared for CADMAC by M. Goldberg, DNV GL, and Kenneth Train, NERA Economic Consulting.