The Brattle Group

MEASUREMENT AND VERIFICATION PRINCIPLES FOR BEHAVIOR-BASED EFFICIENCY PROGRAMS

May 2011

Sanem Sergici, Ph. D.
Ahmad Faruqui, Ph.D.

Prepared for

OPOWER

ACKNOWLEDGEMENTS

The authors would like to acknowledge the research assistance of Lamine Akaba. We would also like to thank the members of the OPOWER team for their helpful suggestions and comments on earlier drafts of this document. Finally, we would like to thank Dr. Helen Connolly of the Luxembourg Income Study for her helpful suggestions on sample design.

Sanem Sergici
Ahmad Faruqui

OUTLINE

1. OBJECTIVE

An increasing number of utilities across the United States have deployed or are planning to deploy behavior-based energy efficiency programs. Just as these programs use a relatively new approach to driving efficiency at utility scale, new measurement approaches are also needed. There has been recognition at the regulatory and advocacy level that an experimental design approach – that is, the use of statistically equivalent treatment and control groups – renders rigorous results at high confidence when properly executed.[1] Because of this growing interest in behavioral efficiency solutions and a limited number of resources detailing their implementation, the need for a document that discusses best practices for designing and evaluating behavior-based programs has become increasingly evident.

This guideline document aims to fill in this gap and lays out the main principles of scientific research that yields a statistically valid program design and program impact metrics. The measurement and verification principles in this document may apply to a broad group of residential energy efficiency behavioral programs that promote efficient usage behavior, customer engagement, and individual energy management. More specifically, these programs may have one or more of the following features:

- Normative comparison of a customer's usage against comparable customers in the same geographical area;
- Targeted conservation and peak reduction tips based on an analysis of a customer's past usage and individual profile;
- Encourage participation in other utility programs based on previous usage patterns and individual consumer profile.

It is important to note that the objective of this document is *not* to develop a comprehensive measurement and verification protocols document that addresses all possible program design and evaluation decisions one can make. Nevertheless, it is our intent to identify the best practices in program design and impact evaluation and provide guidance to utilities in their efforts to design statistically valid programs which will yield reliable impact metrics.

2. PROGRAM DESIGN

2.1. Ensure Internal and External Validity

To be credible and useful to policy makers, programs need to have both internal and external validity. "Internal validity" means that a cause and effect relationship can be established between the various treatments being tested and the variables of interest such as peak demand

[1] Decision 10-04-029 (April 8, 2010) from the California Public Utilities Commission recognizes behavior-based programs that measure savings ex-post and utilize experimental design as eligible efficiency resources to count toward statewide efficiency goals. The National Action Plan for Energy Efficiency (NAPEE) also describes different approaches to applying experimental design to measures to efficiency measures in their discussion of large-scale data analysis (See: *Model Energy Efficiency Program Impact Guide*, p. 4-9, November 2007).

and overall energy consumption. The effect of all other variables needs to be purged. “External validity” means that the program results can be extrapolated to the population of interest. Both require careful design although it is generally easier to ensure internal validity than to ensure external validity.

To ensure internal validity, the “gold standard” of program design stipulates that every treatment that is being tested should also have a control associated with it so that a scientifically valid “but for” world can be constructed from which deviations can be successfully measured[2]. In other words, cause-effect relationships cannot be inferred with any precision and any conclusions derived from the program may be subject to the charge that they simply measure spurious correlation without this control. It is also likely that genuine cause-effect relationships (*e.g.*, conservation tips lead to lower usage by X percent) may not be measured accurately because other factors such as a changing economy or weather may obscure the true relationship. The best way to create a “but for” environment is to select a matching group of customers who can serve as a proxy for the behavior of the treatment group customers. In addition, to further anchor the measurements, it is best to have pre-treatment data on both the control and treatment groups as well as the treatment-period data on both groups of customers.

In the past, programs have been carried out without matching control groups and sometimes with no control groups at all. Others have been conducted with control groups but with no pre-treatment measurements. All such inadequacies impair the internal validity of the programs to varying degrees. Without a control group in the design, it is impossible to control for non-treatment variables that change between the pre-treatment and treatment periods (such as the economy, or general changes in attitudes toward energy use brought about by other exogenous factors). Without pre-treatment data, it is difficult to know if the treatment and control groups were comparable or not before the treatment was introduced. If systematic pre-treatment differences exist, they suggest that there may be a self-selection bias in the sample that needs to be dealt with. Having a randomized control group and sufficient amount of pre-treatment data for both treatment and control groups could address majority of these concerns.

A program must also have *external validity* so that its conclusions are transferable to the population at large. In the case of a behavior-based program, it will be useful to know if such programs will ultimately be offered on a universal basis, a default basis with opt-out provisions, or an opt-in basis. The sampling strategy for a program will vary across these three scenarios. For example, if universal deployment is contemplated, then both the control and treatment groups should be chosen randomly. On the other hand, if an opt-in deployment is envisioned, then opt-in sampling would be appropriate for both groups.

These are the general principles of program design to ensure internal and external validity of results. As with most things in the real world, they serve as guidelines and not mandates. Utilities will need to temper these principles in execution given their time and resource constraints.

[2] For a discussion of the gold standard of the program design, see Ahmad Faruqui, Ryan Hledik, and Sanem Sergici, “ Piloting the Smart Grid,” Electricity Journal, Aug./Sept. 2009.

2.2. Determine Sampling Frame and Program Design Approach

The first step before designing a program is to determine the objective of the program. The objective should clearly state the: (i) treatment(s) that will be applied; (ii) metrics that will be measured, and (iii) population about which inferences will be made.

Once the objectives have been clearly stated, a "sampling frame" must be developed. A sampling frame refers to a population from which a sample will be selected to participate in a program and expected to yield inferences about the population from which it originates. For instance, if a utility is interested in measuring the impact of conservation tips on reducing usage for high-usage homes, then the sampling frame consists of the population of high-usage customers.

After determining the sampling frame, the next step is to determine the "program design approach." Selection and implementation of a design approach have important consequences for internal and external validity of a program, therefore should be decided upon by considering how a given approach would affect a program's internal and external validity. Most behavioral-based efficiency programs are likely to be offered on a universal basis to a population or a sub-population when it is time to offer them as a full-scale program. In that case, the most suitable design approach is a "randomized controlled trial" (RCT) approach in which participants from a sampling frame are *randomly* allocated to treatment and control groups. By ensuring that the participants are selected from the sampling frame using an approach that best approximates the participant mix of a full-scale implementation, the design approach meets the external validity requirement. By randomly allocating participants to treatment and control groups and therefore avoiding potential selection biases, the recruitment approach meets the internal validity requirement (although some additional analysis may still be needed to make sure that the control and treatment groups are comparable even with the randomization).

2.3. Determine Impact Evaluation Method

After the sampling frame and program design approach are determined, the next step is to decide on the impact evaluation methodology. It is important to determine the impact evaluation methodology relatively early on during the program design process, as it has implications for data requirements as well as sample sizes required for statistically valid results. All potential methodologies are essentially based on some form of a mean comparison between treatment and control groups, and there are two main variations based on the frequency of measurements:

1. Studies with a single-measurement of the outcome: these studies employ an approach that is based on comparison of the means between treatment and control groups, measured only once after the introduction of a treatment.
2. Studies with repeated-measurements of the outcome: these studies are also based on comparison of the means between treatment and control groups, but measured at multiple times both before and after the introduction of a treatment.

In studies with repeated measurements taken at points preceding and following a treatment, it is possible to achieve a substantial increase in efficiency (variance reduction) due to the correlation between measurements at different time points as compared to studies with single measurement.

Therefore, in this principles document, we will make the very probable assumption that it is administratively feasible to take multiple measurements for control and treatment groups both before and after the treatment period. The longer a program is run, the more the treatments can be tested in it and the greater the confidence one can have in the results. However, it is rarely possible to run these programs for extended periods of time due to financial and administrative constraints. For behavior-based energy efficiency programs, it is important to allow the program to run at least one full year, so that seasonal effects can be properly captured in the data. In the same fashion, it is ideal to capture at least one full year of “pre-treatment” data so that it can be examined against “post-treatment” data and comparisons can be made on a similar seasonal basis. Our discussion of the data analysis methods in Section 3 of this document will also be based on these assumptions.

2.4. Determine the Sample Design

Once the impact evaluation approach and data requirements have been determined, statistical power analyses must be conducted to determine the treatment and control group sample sizes required to achieve a pre-determined statistical precision level.

The following factors determine the sample sizes required in a program:

- Significance level of the test (Type I error)
- Power of the test (1-Type II error)
- One-sided or two-sided hypothesis testing
- Ratio of treatment and control group sizes
- Number of the pre-treatment measurements planned in a repeated-measure study
- Number of the post-treatment measurements planned in a repeated-measure study
- Correlation between pre-treatment measurements in a repeated-measure study
- Correlation between post-treatment measurements in a repeated-measure study
- Correlation between pre-treatment and post-treatment measurements in a repeated-measure study

It is possible to obtain substantially different sample sizes, which would meet given statistical precision levels and detection limit requirements, based on the selected impact evaluation approach and frequency of measurements in both the pre- and post-treatment periods. Table 1 compares sample sizes that would detect 1% change in the average usages of the treatment customers with 90% statistical power and 95% statistical confidence level. The appendix presents the sample size formula used in these calculations.

Table 1: Comparison of Sample Sizes: Single-measurement vs. Repeated-measurement (Statistical Power=90%, Confidence Level=95%)

One-sided Hypothesis Testing

Assumptions	**Mean Usage (kWh/mo)=1,439** **St. Deviation of Usage (kWh/mo)=779.2** **Impact Detection Limit = 1%**					
	One-sided Test					
	C/T=1		**C/T=2**		**C/T=1/2**	
	Treatment	Control	Treatment	Control	Treatment	Control
Pre=0, Post=1	50,220	50,220	37,665	75,330	75,330	37,665
Pre=6, Post=6	7,701	7,701	5,776	11,552	11,552	5,776
Pre=12, Post=12	5,357	5,357	4,018	8,036	8,036	4,018

Two-sided Hypothesis Testing

Assumptions	**Mean Usage (kWh/mo)=1,439** **St. Deviation of Usage (kWh/mo)=779.2** **Impact Detection Limit = 1%**					
	Two-sided Test					
	C/T=1		**C/T=2**		**C/T=1/2**	
	Treatment	Control	Treatment	Control	Treatment	Control
Pre=0, Post=1	61,618	61,618	46,213	92,426	92,426	46,213
Pre=6, Post=6	9,448	9,448	7,086	14,172	14,172	7,086
Pre=12, Post=12	6,573	6,573	4,930	9,860	9,860	4,930

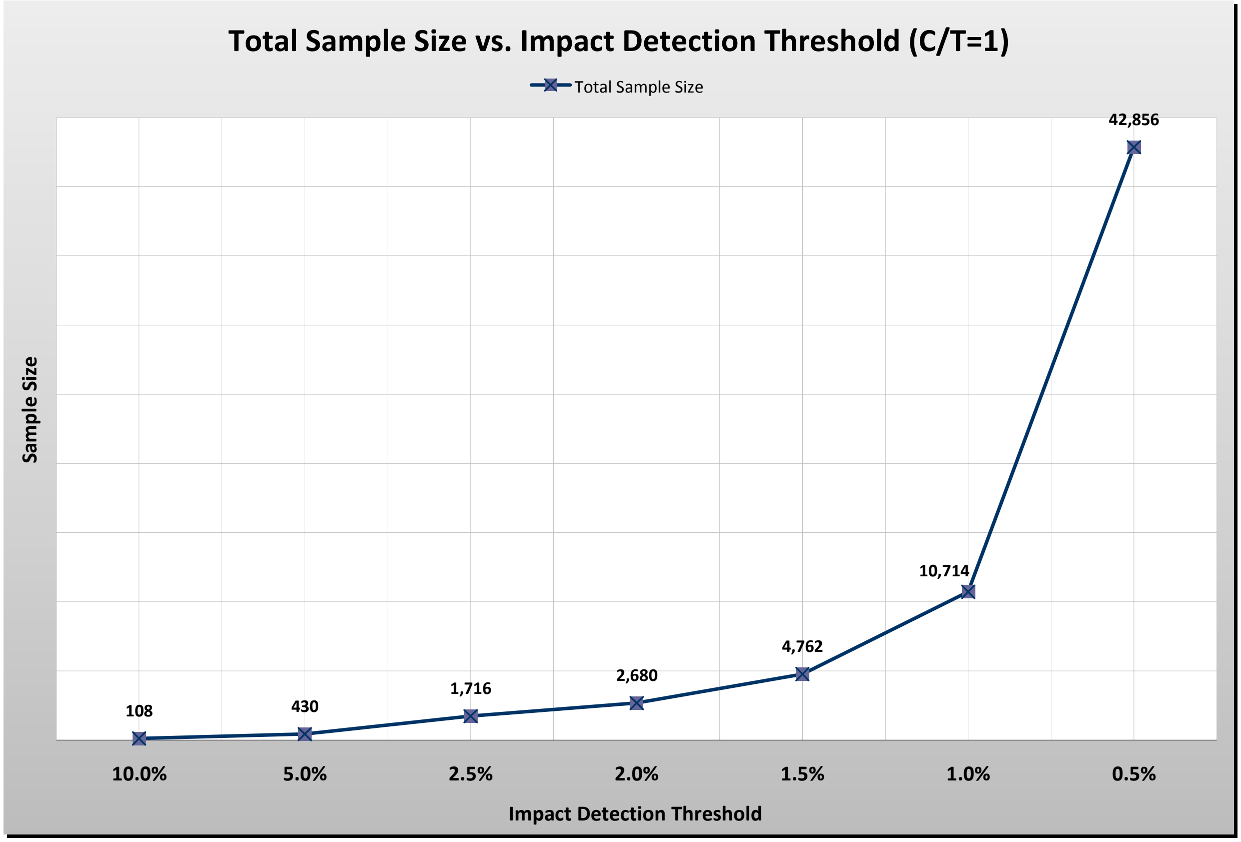
Notes:

1- Our calculations assume r_0=0.73, r_1=0.71, r_{01}=0.69

2- We use the "change" method to calculate the adjustment factors for the standard errors. See Appendix for a discussion of this method.

As Table 1 clearly indicates, a researcher who chooses to employ a single-measurement mean comparison analysis for the impact evaluation of the program would need to recruit 50,220 customers for each of the treatment and control groups assuming that she chooses to have equal sample sizes for both groups and tests a one-sided alternative hypothesis. On the other hand, this researcher would only need to recruit 5,357 customers for each of the treatment and control groups to meet the same criteria, if she chooses to employ a repeated-measurement mean comparison analysis and collects 12 months of pre-treatment and post-treatment usage data for each of the treatment and control customers in her sample. These sample sizes would allow the researcher to detect changes in the mean usage which are greater than or equal to 1% with 90% statistical power and 95% confidence. As the detection threshold becomes smaller, *i.e.*, 0.5%, the sample sizes that are required to detect these impacts with the same statistical power and confidence interval criteria become larger. Figure 1 demonstrates the trade off between the sample size requirement and the detection limit, assuming 12 pre-treatment and 12 post-treatment data points are available for the analysis.

Figure 1: Trade-off between Sample Size and Impact Detection Threshold



Note: These calculations employ the same underlying assumptions in Table 1 and report the total sample sizes (C+T) assuming 12 months of pre- and post-treatment data are available.

It is also important to note that the treatment and control groups do not need to be equal in size in a program. More observations (regardless of being control or treatment) increase the amount of available information which, in turn, decrease the standard deviation and improve the power of the test and significance level of the analysis. It is preferable to collect more treatment observations in a program, but the efficiency of the analyses increases from the increases in either of control and treatment group sample sizes. Therefore, if increasing the control group size is "cheaper" than increasing the treatment group size, it is acceptable to have more control customers than treatment customers in the design. For instance, in Table 1, "C/T=2" case represents a design in which the number of the control customers is twice that of the treatment customers. Alternatively, there may be situations in which increasing the control group size is more expensive than increasing the treatment group sample size. In Table 1, "C/T=1/2" case represents a design in which the number of the control customers is half of that of the treatment customers. In all cases, it is still possible to detect 1% change in mean usage levels with 90% statistical power and 95% confidence.

Another important factor is the selection of a hypothesis testing rule. In a one-sided statistical test, the values for which we can reject the "null hypothesis of zero impact" lie on one side of the probability distribution. For instance, if a researcher is interested in knowing whether a treatment in her research led to a statistically significant reduction in usage, then she would want to use a one-sided test. However, if she is interested in determining whether the impact of treatment is not zero (regardless of a decrease or an increase), then she would want to use a two-sided test. The choice between a one-sided and a two-sided test is determined by the purpose of the program at hand and the questions a researcher wants to answer.

2.5. Determine the Data Requirements

Data requirements of a study will be determined based on the selected impact evaluation approach and practical considerations of a program. The following list summarizes the data requirements for the impact evaluation approach recommended in this guideline document:

1. Monthly kWh usage data for each of the treatment and control customers for at least 3 months and preferably 12 months prior to the treatment period.
2. Monthly kWh usage data for each of the treatment and control customers for at least 12 months during the treatment period.
3. Meter reading date for each of the customers if the billing is based on billing cycles.
4. Tariff designation.
5. Effective treatment start date.
6. For customers leaving the program, the date they left.
7. Socio-demographic and appliance data (if available, this data can be used to further assess whether the treatment and control groups are balanced in their observable characteristics)
8. Weather data based on weather station(s) that are in the closest geographical proximity to the program customers

For monthly kWh usage data collection, it is important to make sure to identify the billing cycle for each of the customers if the bills are prepared on a bill cycle rather than on a calendar month basis. Capturing this data will allow the researchers to align the weather variables, *i.e.*, cooling degree days and heating degree days more closely with the usage variables.

3. MEASUREMENT AND VERIFICATION

3.1. Recommended Impact Evaluation Approach

It is possible to estimate precise load impacts using a single-measurement study design by measuring treatment and control group usages once in the post-treatment period, provided that *sufficiently large sample sizes are available*. These estimations employ a "test of differences" of the mean usages of control and treatment groups. If the difference in the mean values is found to be statistically significant, then the treatment is found to yield an observable effect in the usages of the treatment customers.

If it is feasible to measure treatment and control group usages once in the pre-treatment period and once in the post-treatment period, the load impacts can be measured more precisely compared to the single measurement case with the *added benefit of requiring smaller sample sizes*. These estimations employ a "difference-in-differences" approach which is based on netting out the mean difference between treatment and control groups in the pre-treatment period from the mean differences between treatment and control groups in the post-treatment period. If the difference in differences of the mean values is statistically significant, then the treatment is found to yield an observable effect in the usages of the treatment customers.

Finally, if it is feasible to obtain multiple measurements of treatment and control group customers both in the pre-treatment and treatment periods, then the *precision of the impacts can be improved even more with much smaller sample sizes*. These estimations employ a "panel data or cross-sectional time-series" estimation technique which is based on following and comparing the same individuals over time as well as comparing different individuals at a given point in time through regression models. Panel regressions also allow for the testing of a broad range of hypotheses in addition to the estimation of the load impacts provided that the program is run and measurements are taken over a sufficiently long time period. For example, do the treatment impacts persist over time?[3] Do the treatment impacts vary seasonally?

For the purposes of this document, we will only discuss the panel data regression analysis technique. The motivation behind this choice is several-fold:

- Most behavior-based programs are designed to run at least one year. This implies that repeated measurements on the treatment and control group data will be readily available.
- Behavior-based programs require billing data at a minimum and therefore they are possible to administer using a utility's legacy metering and billing systems. This also implies that several months' worth of pre-treatment data will be available for both treatment and control group customers.
- Since several repeated measures of pre-treatment and treatment data are likely to be available for both treatment and control group customers, the panel data estimation would yield the most precise impact estimate at much lower sample sizes for both treatment and control groups.
- It is possible to explicitly control for the weather variables within the panel data regression framework to remove weather impacts on customers' usage behaviors and therefore reveal the true impact of the treatment.
- It is possible to account for the impacts of time-invariant unobservable variables on the usage levels through a procedure known as "fixed-effects" estimation which is embedded in the panel data regression approach. These variables, if remain uncontrolled, lead to biased estimates of the load impacts.

3.2. Compare Treatment and Control Groups in the Pre-Treatment Period

As discussed in Section 2.3, most behavior-based efficiency programs are likely to be offered on a universal basis to a population or a sub-population when it is time to offer them as full-scale programs. Moreover, existing experience with these programs show that a very small number of the customers opt-out of the program after their initial enrollment.[4] In that case, the most

[3] OPOWER programs assume a single-year measure life. However, there may be some permanent behavior changes that can be empirically explored.

[4] Previous research shows that the opt-outs are in the range of 1 to 2 percent. See, for example, Hunt Allcott. *Social Norms and Energy Conservation*. MIT and NYU. October 10, 2009, and Ian Ayres et al.

suitable design approach is a "randomized controlled trial" approach in which participants from a sampling frame are *randomly* allocated to treatment and control groups. The expectation as a result of the random allocation is that the treatment and control groups would be comparable to each other in terms of their usage, socio-demographic, and appliance characteristics. However, despite the randomization, it is still a good practice to assess the comparability of the treatment and control groups in the pre-treatment period before the treatment customers started to participate in the behavior-based energy efficiency programs. We recommend the following analyses for this assessment:

1. Compare monthly average daily usages between treatment and control groups by month. Conduct mean comparison tests to determine whether the difference between the treatment and control group usages is statistically significant.

2. Plot average daily usages for treatment and control groups for each pre-treatment month and visually inspect whether the average daily usages follow the same pattern across a given month for both groups.

3. Compare the distributions of socio-demographic and appliance characteristics between the treatment and control groups to the extent that data is available. Determine whether these characteristics are statistically similar between the two groups.

If these analyses imply that the treatment and control groups are similar to each other in most of these dimensions, the control group is further verified to be a reliable “but for” group for the treatment group. To the extent that there are some dissimilarities between the two groups in terms of their usages and largely time-invariant (at least during the study time frame) socio-demographic and appliance characteristics, these differences can be accounted for in the impact evaluation framework as we will discuss in Section 3.3.2. Alternatively, the randomization procedure can be repeated until balance is achieved[5].

3.3. Specify the Impact Evaluation Framework

The primary objective of impact evaluation is to obtain the most accurate impacts that can be attributed to a treatment tested in a program. To ensure that the results are free from errors and can stand the scrutiny of internal and external stakeholders, the impact evaluation should adhere to the academic standards of applied econometrics. Therefore, an impact evaluation approach adopted for a program must follow some generally accepted rules and conventions.

As discussed previously, there are three main approaches that can be employed for impact evaluation of the behavior-based energy efficiency programs: (i) Difference of Means or

Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage. http://ssrn.com/abstract=1434950.

[5] In addition to pure randomization discussed in this document, there are other randomization methods namely stratified randomization, pair-wise matching randomization, and re-randomization methods. See, for example, Miriam Bruhn and David McKenzie. *In Pursuit of Balance: Randomization in Practice in Development Field Experiments*. American Economic Journal: Applied Economics 2009, 1:4, 200-232. They show that with large sample sizes, the method of randomization matters much less for the degree of balance of treatment and control groups.

Analysis of Variance (ANOVA); (ii) Difference-in-Differences of Mean Values; (iii) Panel Data Regression Analysis. Our recommended approach is Panel Data Regression Analysis technique because it is possible to increase the efficiency and precision of the estimates using repeated measures on each program participant and to account for time-invariant unobservable variables that would otherwise lead to biased estimates, using this technique. Several other reasons for this recommendation were provided in Section 3.1.

3.3.1. Overview of Panel Data Regression Analysis

In simple terms, a regression model relies on a dataset and statistical analysis to develop a mathematical relationship between a variable of interest (dependent variable) and other variables (independent or explanatory variables) that influence the variable of interest.[6] A regression analysis may utilize a cross-sectional dataset, a time-series dataset, or a hybrid of the two, a panel dataset.

A *cross-sectional dataset* consists of different individuals measured for certain characteristics at a given point in time. Consequently, a cross-sectional regression analysis defines a mathematical relationship between dependent and independent variables by utilizing data on different individuals measured at one point in time.

On the other hand, a *time-series dataset* consists of one individual measured for certain characteristics at different points in time. Driven by the nature of this dataset, a time-series regression analysis develops a relationship between dependent and independent variables by utilizing the data on one individual measured at different points in time.

Finally, a *panel dataset (*also known as cross-sectional time-series or longitudinal data) consists of different individuals measured for certain characteristics at different points in time. A panel data regression utilizes the variation in the data across individuals, as well as across time, to fit a relationship between dependent and independent variables.[7] Naturally, observing many individuals over time is more advantageous compared to having either cross-sectional data alone or time-series data alone. A panel dataset makes it possible to study different research hypotheses related to impacts that only emerge over time. For instance, in the context of a behavior-based energy efficiency program, if a researcher is interested in understanding the persistence of a program's impacts, she will need to employ a panel dataset to explore this question. However, the most important benefit of a panel dataset is that it allows a researcher to account for time-invariant unobservable characteristics of individuals that could otherwise introduce bias to the estimation results. These biases could be certain socio-demographic and appliance characteristics such as education level of head of household, income level, central air conditioning ownership, and so on. If a researcher does not observe, or have reliable data on, these characteristics, it is not possible to employ these variables as independent variables even

[6] California Public Utilities Commission, Energy Division. *Attachment A- Load Impact Estimation for Demand Response: Protocols and Regulatory Guidance,* April 2008. pp. 60

[7] Jeffrey M. Wooldridge. *Introductory Econometrics- A Modern Approach.* Fourth Edition, South-Western Cengage Learning, 2009. This book is a good reference for regression analysis in general and introductory panel data regression technique.

though they have the potential to explain the variation in the dependent variable. As we will discuss in Section 3.3.3, omission of these variables from the regression model leads to an "omitted variable" problem which may result in biased parameter estimates.

There are two widely used panel data estimators that account for the unobservable factors that may vary across individuals, but are constant over the course of the study: (i) fixed-effects estimator, (ii) random effects estimator.[8]

One of the key assumptions for a regression model to produce unbiased estimates, the error term, u, must have an expected value of zero given any value of the model's independent variables ($E(u \mid X) = 0$, zero conditional mean assumption). This implies that the error term must not be related to any of the independent variables in the model. However, when an independent variable is omitted from the regression, it is automatically included in the error term. If this omitted variable is related to one of the model's independent variables, then the error term becomes related to one of the independent variables violating the zero conditional mean assumption and leading to biased parameter estimates.

Fixed-effects (FE) estimation assumes that the unobservable factor (the error term) is related to one or more of the model's independent variables. Therefore, it removes the unobserved effect from the error term prior to model estimation using a data transformation process. During this process, other independent variables that are constant over time are also removed. This drawback of the FE estimation implies that it is not possible to estimate the impact of variables that remain constant over time, such as ownership of a single-family house. However, it is still possible to estimate the impact of the ownership of a single-family house in the post-treatment period, by interacting the single-family home variable with a post-treatment period indicator variable (which is time-variant)

Random-effects (RE) estimation is a reasonable alternative when a researcher is able to explicitly control for all potential independent variables and has a good reason to think that any unobservable variable that may be pooled in the error term is not correlated with any of the model's independent variables. If this assumption holds, then removing it from the error terms, as in the case with FE estimation, would result in inefficient estimates. Therefore, RE estimator is a more efficient estimator compared to that of FE when the unobserved effect is uncorrelated with independent variables. Moreover, RE estimator has the advantage of allowing for the estimation of variables that remain constant over time. However, it is important to note that if the assumption about the unobservable effect does not hold, then the RE estimator would yield biased parameter estimates.

Most of the time, the primary reason for using panel data is to account for the unobservable time-invariant effects, which are thought to be correlated with the independent variables using, an FE estimator. If this assumption does not hold however, the parameter estimates would be less efficient compared to those that can be estimated using an RE estimator. Fortunately, there is a statistical procedure called the "Hausman test" which is used to assess whether the RE or FE is a

[8] Jeffrey M. Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. First Edition, Massachusetts Institute of Technology , 2001.

more suitable estimator for a given panel regression model.[9] The Hausman test is based on estimating a model using both FE and RE and then formally testing for differences in the parameter estimates. Rejection of Hausman test implies that the RE assumption is not valid; therefore, the researcher should employ the FE routine to obtain unbiased parameter estimates.

3.3.2. An Example Model Specification

In this section, we present a general model specification that can be used to estimate the impact of behavior-based energy efficiency programs. *This specification is just one of many alternative specifications that can be employed by researchers and it is included in this guideline document to demonstrate the concepts we have previously introduced in this document.*[10]

$$\begin{aligned}
&\ln\_kWh_{it} = \alpha_0 + \alpha_1 Treatment_i + \alpha_2 Post_t + \alpha_3 TreatmentxPost_{it} \\
&+\beta_1 HDD_t + \beta_2 HDDxTreatment_{it} + \beta_3 HDDxPost_{it} + \beta_4 HDDxTreatmentxPost_{it} \\
&+\delta_1 CDD_t + \delta_2 CDDxTreatment_{it} + \delta_3 CDDxPost_{it} + \delta_4 CDDxTreatmentxPost_{it} \\
&+v_i + u_{it}
\end{aligned} \tag{1}$$

Where:

$\ln\_kWh_{it}$: Natural logarithm of monthly average kWh/day for customer i and month t.

$Treatment_i$: Dummy variable that takes the value of 1 if customer i is a treatment customer.

$Post_t$: Dummy variable that takes the value of 1 if month t is in the treatment period.

$TreatmentxPost_{it}$: Dummy variable that takes the value of 1 if customer i is measured in the treatment period month t.

HDD_t : Heating degree days per day for month t

$HDDxTreatment_{it}$: Interaction of HDD_t with $Treatment_i$

$HDDxPost_{it}$: Interaction of HDD_t with $Post_t$

[9] Some econometric packages, such as STATA, have a routine to calculate the Hausman test.

[10] It is possible that the interactions of the CDD and HDD variables with the *Post*, *Treatment*, and *TreatmentxPost* variables will turn out to be insignificant when the model is estimated. In that case, it may be preferable to estimate a more simplified version of this model where the CDD and HDD variables are included in the model without interactions.

$HDDxTreatmentxPost_{it}$: Interaction of HDD_t with $TreatmentxPost_{it}$

CDD_t : Cooling degree days per day for month t

$CDDxTreatment_{it}$: Interaction of CDD_t with $Treatment_i$

$CDDxPost_{it}$: Interaction of CDD_t with $Post_t$

$CDDxTreatmentxPost_{it}$: Interaction of CDD_t with $TreatmentxPost_{it}$

v_i : Time invariant fixed effect term for customer i.

u_{it} : Independent and identically distributed random error term for customer i at month t.

This equation is estimated using data on both treatment and control customers before and during the treatment period. This type of database allows one to isolate the true impact of a program by controlling for any potential biases due to (i) differences between control and treatment groups in the pre-treatment period (ii) any changes in the consumption behavior of the treatment customers between the pre-treatment and treatment periods that are not related to the treatment *per se*. These potential confounding factors are controlled for by introducing dummy variables pertaining to the customer type ($Treatment_i$) and the analysis period ($Post_t$).

It is important to properly control for the impact of weather conditions on the usage behavior of the customers and isolate the impact of the program treatments. If there is a usage reduction in the treatment period, a researcher must ensure that the reduction due to possibly milder weather conditions (therefore less CAC or electric heating load) in the treatment period is properly identified and not attributed to the behavior-based energy efficiency program that is being tested. In order to compare pre-treatment and post-treatment usages on a seasonal basis, it is recommended that at least one full year of usage data for both periods is collected.

Due to the nature of the FE estimator, it is not possible to measure the impact of time-invariant socio-demographic and appliance characteristics on customer's energy usage in this model since they will be removed along with the time-invariant un-observables. However, it is possible to determine how the treatment impact varies with these time-invariant customer characteristics through interactions with $Post_t$ and $TreatmentxPost_{it}$ variables. For instance, if a researcher is interested in learning the incremental treatment impact of single-family housing, three additional variables can be included in the model above: SF_i, $SFxPost_{it}$, and $SFxTreatmentxPost_{it}$. The model will only estimate the parameters for the last two variables as FE routine will remove away the SF_i variable.

Having discussed the variables included in the model, we can now define the average treatment impact (ATC) which is the sum of all terms multiplying the interaction term $TreatmentxPost_{it}$:

$$\hat{ATC} = \hat{\alpha}_3 + \hat{\beta}_4\, HDD_t + \hat{\delta}_4\, CDD_t \qquad (2)$$

Where HDD_t and CDD_t are the average values of the *actual* weather terms in the treatment period.

As the dependent variable of our model is in logarithms, the average treatment impact calculated from (2) will be in percentage terms. If a researcher chooses to define the dependent without the logarithmic conversion, then the appropriate calculations would need to be undertaken to convert the impacts into percentages.

The average treatment impact estimated from the regression above is only a point estimate and does not mean much without its estimated standard error and confidence interval. Confidence interval can be easily calculated as follows:

$$\hat{ATC} \pm c * se(\hat{ATC}) \qquad (3)$$

For a 95% confidence interval, c is the 97.5th percentile in a t_{df} distribution and $df = n - k - 1$

Lower and upper bounds of the confidence interval are given by:

$$LB = \hat{ATC} - c * se(\hat{ATC}) \text{ and } UB = \hat{ATC} + c * se(\hat{ATC}) \qquad (4)$$

A 95% confidence interval for $\hat{ATC}$ implies that if random samples are drawn repeatedly, with LB and UB computed each time, the unknown population value for ATC would lie in the interval (LB, UB) for 95 percent of the samples.[11]

3.3.3. Issues in Regression Analysis

It is often said that regression analysis is as much art as it is science. *A priori*, it is not possible to come up with a prescriptive list or blueprint that can lead a researcher to the best model specification and the most accurate estimates of the regression parameters. For that reason, it is important that impact evaluations of programs are undertaken by experienced professionals who have developed a modeling intuition over the years and excelled at the art of regression analysis.

Although it is not possible to be prescriptive, it is still possible to identify major issues one should be aware of in regression analysis. Below, we introduce these issues briefly and discuss their implications for the regression analysis:[12]

[11] For a detailed discussion of confidence interval, see Wooldridge (2009).

[12] TecMarket Works. *The California Impact Evaluation Framework*. June 24. pp. 113-117. See for detailed discussion of regression issues.

1) Model Misspecification: A model misspecification refers to incorrect specification of a relationship between dependent and independent variables. It can take different forms:

   a) Omitted variables: this is the most common form of model specification error. When a variable is omitted from a regression model, it is pooled in the error term. If the omitted variable is correlated with at least one of the independent variables, then the zero conditional mean assumption is violated leading to biased estimates for all parameters.
   b) Inclusion of irrelevant variables (over-specification of the model): this issue emerges when one or more of the independent variables in a regression model have no effect on the dependent variable. Inclusion of irrelevant variables into the regression model does not lead to biased parameter estimates, however, it leads to larger variances for the estimated parameters.
   c) Incorrect functional form: if the functional form used to describe a relationship between the dependent and independent variables does not reflect the true relationship, it will lead to biased estimates as well as a meaningless estimate of the relationship. Omission of interaction terms, quadratic terms, or defining variables in levels rather than logarithmic terms when the model calls for these terms, are some examples of functional form misspecification. It is often difficult to tell whether a model has a misspecification error or not. Although there are some tests to find out whether a model has an omitted variable problem, there is not a prescriptive way of testing for model misspecification[13]. Most of the time, the best guidance a researcher can rely on is the economic theory, relevant literature, and experience.

2) Measurement Error: When the actual data on dependent or independent variables differ from data on the reported or measured variables, the estimated regression model includes a measurement error. If a variable (dependent or independent) includes a measurement error, and if this error is uncorrelated with the explanatory variables, the parameter estimates will still be unbiased. The variance will be, however, larger compared to a case without a measurement error, a situation which can only be addressed by collecting better data. When a variable (dependent or independent) includes a measurement error which is correlated with one or more independent variables in the model, the problem is more serious. The parameter estimates from the model will be biased and inconsistent.

3) Heteroscedasticity: in order to obtain correct inferences based on hypothesis testing, the error term of a regression model must have the same variance for any given value of an independent variable. If this assumption is violated, then the regression estimation suffers from heteroscedasticity. If a model has the heteroscedasticity problem, the parameter estimates would still remain unbiased; however, standard errors would be inaccurate. A relatively simple way to detect heteroscedasticity is to plot a model's error term against the independent variable(s) it is thought to be correlated with and assess whether there is a detectable correlation between the two series. As a good practice, it is recommended to use "heteroscedasticity-consistent standard errors" (or Huber-White standard errors) in the OLS estimations. If the heteroscedasticity is present, the errors will be fixed, otherwise the errors will remain unchanged. For more serious forms of heteroscedasticity, the researcher may need to resort to the weighted least squares estimation.

[13] Peter Kennedy. *A Guide to Econometrics*. Fifth Edition, The MIT Press, 2003.

4) Autocorrelation: if the errors from two different time periods are correlated, then we say that the errors suffer from autocorrelation or serial correlation. Autocorrelation is a very common problem in time series and panel-data regression analyses due to the time nature of the data. When the errors are serially correlated, the estimated parameters remain unbiased, but the standard errors are biased. This implies that statistical inference based on the estimated standard errors would be inaccurate. There are standard tests to detect the existence of autocorrelation such as Durbin-Watson and Breusch-Godfrey tests. It is recommended to rely on "autocorrelation-robust standard errors" through clustering of the error terms to correct for the unknown form of autocorrelation in the error terms[14]. If the form of autocorrelation is known, it is possible to obtain more efficient estimators by using a feasible GLS estimator such as Cochrane-Orcutt or Prais-Winsten estimators.

[14] These errors are called panel cluster standard errors and they are robust to both heteroscedasticity in the cross-section dimension as well as unknown forms of serial correlation in the time series dimension.

APPENDIX

Calculating Sample Size in Studies with Single Measurement and Repeated Measurement[15]

The following is a list of terms and parameters that are used in the sample size formulas:

n1 & ***n2*** = the sample size in population 1 and population 2

samples = indicator for one or two-sample test (values = 1 or 2)

m1 & ***m2*** = the means of population 1 and population 2

sd1 & ***sd2*** = the standard deviations of population 1 and population 2

alpha = the significance level of the test

power = 1 $-\beta$ is the power of the test

ratio = the ratio of sample sizes for two-sample tests: ratio = n2/n1

base = the number of baseline measurements planned in a repeated- measure study

follow = the number of follow-up measurements planned in a repeated- measure study

r0 = the correlation between baseline measurements in a repeated-measure study

r1 = the correlation between follow-up measurements in a repeated-measure study

r01 = the correlation between baseline and follow-up measurements in a repeated-measure study

sides = indicator for one-sided or two-sided test (values = 1 or 2)

method = post, change, or ancova analysis method to be used with repeated measures

asd1 & a***sd2*** = the standard deviations of population 1 and population 2 adjusted for the relative efficiency gained from multiple observations

sdadj = the adjustment factor based on the relative efficiency gained from multiple observations, which is applied to the standard deviation

1- Studies with Single Measurement of the Outcome

For simple studies, with only one outcome measurement, the basic method to calculate the sample size of the treatment group (***n1***) is:

One-sample test (treatment group (***m1***) v. hypothesized mean (***m2***)):

$$\boldsymbol{n1} = \frac{\boldsymbol{sd1}^2 * (z_{1-\boldsymbol{alpha/sides}} + z_{\boldsymbol{power}})^2}{(\boldsymbol{m1}-\boldsymbol{m2})^2}$$

[15] See Hilbe, J. M. *Sample size determination for means and proportions*. Stata Technical Bulletin, 1993, 11:17-20, and Seed P.T. Sample Size Calculations for Clinical Trials with Repeated Measures Data. Stata Technical Bulletin, 1997, 40:16-18. Also see the discussion of the sampsi command in the Stata Technical Manual for more information.

Two-sample test (treatment (***m1***) v. control (***m2***)):

$$\boldsymbol{n1} = \frac{(\boldsymbol{sd1}^2 + \boldsymbol{sd2}^2/\boldsymbol{ratio})*(z_{1\text{-}\boldsymbol{alpha/sides}} + z_{\boldsymbol{power}})^2}{(\boldsymbol{m1}\text{-}\boldsymbol{m2})^2}$$

2- Studies with Repeated Measurement of the Outcome

In studies with repeated measurements taken at baseline and/or follow-up, there can be an increase in efficiency (as compared to a single outcome measurement) due to the correlation between measurements at different time points. The method to calculate the sample size of the treatment group (***n1***) is the same as with the single measurement studies, adjusting for the change in efficiency of the standard error:

One-sample test (treatment group (***m1***) v. hypothesized mean (***m2***)):

$$\boldsymbol{n1} = \frac{\boldsymbol{asd1}^2 * (z_{1\text{-}\boldsymbol{alpha/sides}} + z_{\boldsymbol{power}})^2}{(\boldsymbol{m1}\text{-}\boldsymbol{m2})^2}$$

Two-sample test (treatment (***m1***) v. control (***m2***)):

$$\boldsymbol{n1} = \frac{(\boldsymbol{asd1}^2 + \boldsymbol{asd2}^2/\boldsymbol{ratio})*(z_{1\text{-}\boldsymbol{alpha/sides}} + z_{\boldsymbol{power}})^2}{(\boldsymbol{m1}\text{-}\boldsymbol{m2})^2}$$

where:

asd1 = sd1 * sdadj
asd2 = sd2 * sdadj

and the relative efficiency is calculated as:

$$\textbf{\textit{Relative efficiency}} = \frac{1}{\boldsymbol{sdadj}^2}$$

The adjustment to the standard deviation (***sdadj***) is calculated according to the method used.

a) <u>Post-treatment Method (*post*)</u>

The post-treatment method (***post***) utilizes information from multiple follow-up observations, ignoring baseline measurements. Using the ***post*** method with a single follow-up observation is equivalent to using the unadjusted single-outcome calculation.

$$\boldsymbol{sdadj} = \sqrt{\frac{1 + (follow - 1) \times r1}{follow}}$$

b) <u>Change Method (*change*)</u>

The change in means method (***change***) utilizes information from both multiple baseline and multiple follow-up observations. Using the ***change*** method with a single baseline and a single follow-up observation is equivalent to using the unadjusted single-outcome calculation.

$$\textbf{\textit{sdadj}} = \sqrt{\frac{1+(follow-1)\times r1}{follow}+\frac{1+(base-1)\times r0}{base}-2r01}$$

c) <u>ANCOVA Method (*ancova*)</u>

The ANCOVA method (***ancova***) utilizes information from both multiple baseline and multiple follow-up observations. It adjusts the standard deviation to account for covariance among observations and corrects for the average at mean at baseline. Using the ***ancova*** method with a 0 correlation between baseline and follow-up measurements is equivalent to using the unadjusted single-outcome calculation.

$$\textbf{\textit{sdadj}} = \sqrt{\frac{1+(follow-1)\times r1}{follow}-\frac{r01^2\times base}{1+(base-1)\times r0}}$$