DEALING WITH OUTLIERS IN IMPACT EVALUATIONS BASED ON BILLING DATA

Scott Pigg
Wisconsin Energy Conservation Corporation
Madison, Wisconsin

Michael Blasnik
GRASP
Philadelphia, Pennsylvania

Introduction

The goal of using utility billing data for impact evaluation is to discern the program energy savings from among all the other factors that influence energy usage. However, at a given facility, extraneous factors unrelated to the program can have a much larger impact on energy usage than what we are seeking to measure—the program itself. This can have a deleterious effect on the precision with which program impacts can be measured, and in some cases can strongly bias the results.

The purpose of this paper is to discuss, in a general way, the identification and treatment of unusual customer billing data. The goal of this process is to improve the precision with which we can measure program impacts without producing biased estimates of program savings. As such, it is focused on reducing the influence of random non-program factors that influence energy usage.

What is an Outlier?

We begin with the definition of an outlier. In its broadest sense, an outlier is an unusual data point. This definition is not useful, though, unless one can also address three questions: (1) "Unusual relative to what?"; (2) "How unusual?"; and (3) "Unusual for what reasons?"

Unusual Relative to What?

Outliers can occur in billing data both within and between facilities. Within an account, there may be months that show unusual usage relative to the pattern exhibited by other months. And facilities themselves may have unusually high or low usage, or exhibit unusual changes in usage, relative to other facilities. The presence of within-facility outliers does not necessarily imply that the facility will be a between-facilities outlier. Nor does being a between-facilities outlier necessarily mean that there are outliers in the monthly billing data; the facility may simply be unusual, and have internally consistent monthly billing data.

It is important to keep this distinction in mind. We typically perform two different operations with utility billing data: (1) we look within houses or facilities for patterns in the monthly billing data that explain the variation in month-to-month usage; and (2) we look across facilities for assessing the average usage or savings for populations of customers. Depending on the methods used, these steps may be performed independently (two-stage analysis) or they may be performed simultaneously (one-stage analysis).

In the context of multivariable regression, which is often employed to help control for the factors that affect usage and changes in usage, data points may well be outliers only in a multivariate sense: that is, the data point may be an outlier only when we examine it in relation to the combination of factors used in a regression model to explain variation. We discuss this situation in more detail later in this paper.

How Unusual?

The question of how different a data point must be in order to be classified as an outlier is a vexing one, because we must make decisions about the boundary of normal versus abnormal data based on the sample or study group at hand. One might draw these boundaries differently given a picture of the entire population of interest, a larger sample, or even simply a different sample.

The "how unusual" question is often addressed by comparing the observed distribution with a normal (Gaussian) distribution. Data points are called outliers if they lie more than x standard deviations away from the mean (where x is often 2.5 or 3). The problem with this classification scheme is that the mean and standard deviation must be calculated from the data at hand, and therefore are subject to influence by the very outliers one is seeking to identify.

The problem is one of obtaining robust measures of the center and the variability of the data—what statisticians call *location* and *scale*. The standard deviation is the

most common measure of scale, but it is also the most strongly influenced by outliers. We discuss robust alternatives to the mean later. Two more robust measures of scale are the interquartile range (IQR) and the median absolute deviation (MAD). The interquartile range is the distance between the 25th and 75th percentiles. As the name implies, the MAD is the median of the absolute deviations from the median.

Unusual for What Reasons?

Explaining exactly *why* a data point is an outlier is perhaps the most critical activity, yet it requires more detailed knowledge of facilities than is usually available. Nevertheless, it is important to recognize the distinctions among different sources of outliers.

Some data are outliers because the data are simply wrong. Meters can be misread, mistranscribed, or misdated (though these errors all tend to be self-correcting). In addition, the billing data may contain corrections, periods of no usage, or other customer billing system information that is mishandled by the unwary evaluator. Some of these errors can be detected and screened correctly by the analyst, while some cannot. Meter reading errors can sometimes be identified as roughly symmetric, opposed outliers in consecutive months.

Billing data represent usage for all end uses connected to the meter used in the analysis, and hence tend to contain a lot of noise that is not relevant to the program being assessed. Electricity billing data tend to be noisier than gas data because they represent more end uses (and more idiosyncratic end uses). Some unusual periods of usage are transient, and are caused by temporary factors such as the occurrence of a holiday. Other changes are due to changes in appliance holdings or structural modifications to a facility; these are more permanent.

The final category of outliers is that of firms or households that show unusually large changes in usage because they really have unusually large savings. These data points look unusual relative to other facilities, but they should not be eliminated because they represent legitimate program effects. Of course, much of the difficulty in handling outliers stems from an inability to unambiguously distinguish these program-induced outliers from outliers due to other factors.

The goal of using utility billing data for impact evaluation is to discern the program energy savings from among all the other factors that influence energy usage. However, at a given facility, extraneous factors unrelated to the program can have a much larger impact on energy usage than what we are seeking to measure—the program itself. This can have a deleterious effect on the precision with which program impacts can be measured, and in some cases can strongly bias the results.

The purpose of this paper is to discuss, in a general way, the identification and treatment of unusual customer billing data. The goal of this process is to improve the precision with which we can measure program impacts without producing biased estimates of program savings. As such, it is focused on reducing the influence of random non-program factors that influence energy usage.

What to Do with Outliers

Once an observation has been classified as an outlier, the analyst can either correct, delete, explain, ignore, or downweight it. The proper approach depends on the cause of the outlier (if identifiable), its impact on program results, and often on the nature and objectives of the analysis.

The first step after identifying an outlier is to look for its cause. Double-checking the original data may lead to the discovery of an error. If the error can be corrected, then the point can be retained; otherwise, the error should be deleted and reported in the analysis of sample attrition.

If no error is found, then the data can be investigated more thoroughly to (one hopes) explain its cause. Other available data from the tracking system or from surveys can be examined for unusual values or patterns that may explain the outlier (such as vacation absences or equipment breakdowns). A phone call or site visit may be worthwhile when feasible, particularly for large facilities that dominate the results. If these efforts lead to an explanation, then a decision must be made regarding the proper course of action. If the outlier was caused by something that can be attributed to the program or that also occurs in the comparison group, then one might lean toward retaining the outlier. This approach can lead to problems if the treatment or comparison samples do not properly represent the true population frequency of whatever unusual circumstances caused the outlier. The safest approach may be to report the results both with and without such outliers, and then provide a rationale for which our answer is considered best.

A lucky analyst may find that the cause of the outlier has been captured in other data available for the entire population (e.g., through survey or tracking data). The analyst may then be able to "explain" the unusual observation statistically through stratification or regression approaches, making it no longer an outlier and improving precision. Frequently, no explanation is found for outliers and the analyst is left with a more difficult decision. The primary options have been to either retain the outlier and accept lower precision (and potential bias if it should have been deleted) or delete the outlier and provide some

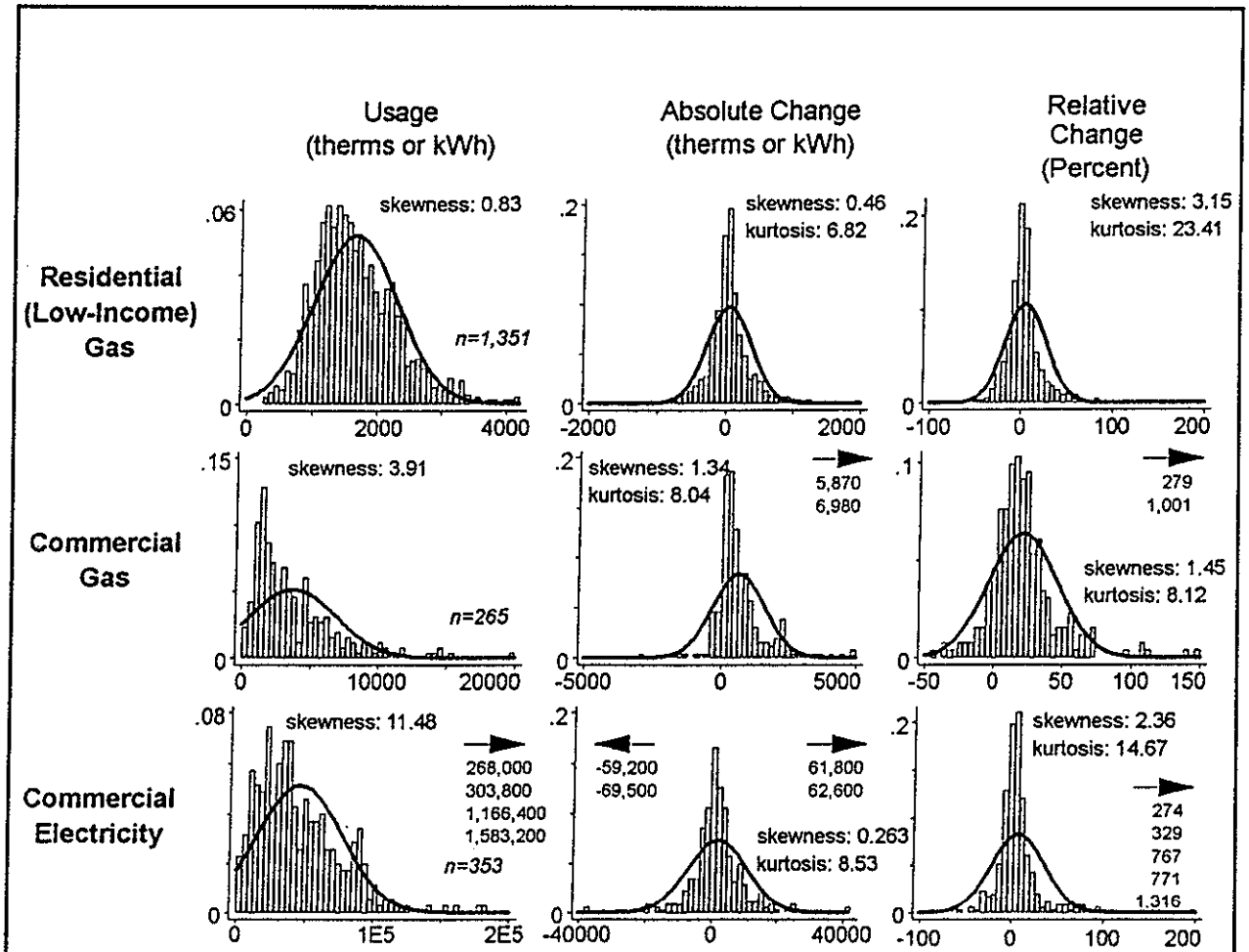
justification. Another approach is to retain the outlier but reduce its influence through the use of robust statistical procedures, which we discuss in more detail below.

Simple judgement-based screening procedures are sometimes used to eliminate a few very wild data points. For example, all facilities that show more than a 75% change in usage may be dropped. If prior engineering predictions of savings are available, the data may be screened on the difference between the predicted and observed change in usage. The analyst must be careful that such procedures do not eliminate important but idiosyncratic facilities from the data set.

Regardless of the approach used, the responsible analyst should report outlier identification methods, the rationale behind any judgments made, and the ultimate impact on the results. Additionally, the same procedures and approaches must be used for both the treatment and comparison groups.

Typical Variation in Usage and Changes in Usage

Figure 1 shows examples of typical distributions of annual billing usage and the year-to-year change in usage for residential and commercial customers. The graphs represent usage and changes in usage that occurred in the two years *prior* to participation in energy conservation programs in Wisconsin. The figure highlights three quantities that are critical to measuring program impacts: usage



Notes:

Line represents Gaussian distribution with same mean and std. dev. as data over range of figure.

Skewness for usage calculated using all data points; skewness and kurtosis for change data is over range of figure.

Figure 1. Examples of Pre-participation Usage and Changes in Usage for Program Participants

level; absolute (*i.e.*, therm or kWh) change in usage from year to year; and relative (percentage) change in usage. For obtaining estimates of therm or kWh impacts, the absolute change in usage is the primary quantity of interest, but understanding usage level distributions and relative change distributions are also important.

We have found that usage typically is lognormally distributed—more strongly so for commercial customers than for residential customers. The distribution of usage, of course, depends on the kinds of customers that participate in a program. Some programs are limited to certain size classes of customers, while others may span a wide range of usage.

The change in usage from year to year (both absolute and relative) is typically leptokurtic; that is, more "peaky" than a Gaussian distribution with the same mean and standard deviation. Another way to put this is to say that the tails of the distribution tend to be longer than would be encountered by a Gaussian distribution that fits the middle of the data. Kurtosis is a measure of the peakedness of a distribution; a Gaussian distribution has a kurtosis of 3. We have found that the year-to-year change in annualized energy usage typically has a kurtosis of 6 to 30. Kurtosis is very sensitive to the presence of outliers, and may be much higher if there are some extreme observations. One possible explanation for this shape is that such distributions represent a mixture of many facilities that show minor year-to-year variations in usage, and a few that experience large changes due to some type of structural change.

A strongly lognormal usage distribution can also create a long-tailed change distribution. If large and small facilities have a similar propensity for *percentage* change in usage, the *absolute* changes in usage for the large facilities will tend to wind up in the tails of the change distribution. This is an important point to keep in mind when dealing with outliers, since it implies that large facilities may look like outliers in a savings distribution, even though their savings may not be abnormal in relation to their usage. Deleting or downweighting these facilities may bias the results in favor of smaller facilities, as we discuss in more detail later.

Concepts of Robustness

A robust estimator is one that is not unduly influenced by the presence of outliers. Statisticians distinguish between two types of robustness: robustness of *validity* and robustness of *efficiency*. Robustness of validity refers to the ability of an estimator and its calculated confidence interval to maintain the desired confidence level in the face of outliers. In other words, if the desired confidence level for an analysis is 90%, the confidence interval for an estimator that is robust with respect to validity will still enclose the true population value in 90 out of 100 random samples, even with outliers present.

Robustness of efficiency refers to the ability of an estimator to maintain its efficiency in the presence of outliers. To give a vague definition, the efficiency of an estimator is its ability to measure the desired quantity with as much precision as possible for a given sample size. Robustness of efficiency thus refers to the ability of an estimator to maintain the width of its confidence interval in the face of outliers.

It has long been recognized that the most common measure of central tendency—the mean—is robust with respect to validity, but not with respect to efficiency (for reasonable sized samples). In other words, the confidence interval for the mean tends to maintain its validity in the presence of outliers, but it does so by getting wider. In practical terms, this means that estimates of program savings are less precise than they otherwise would be if outliers were not present.

This fact has led to a search for alternative estimators that can maintain the validity *and* efficiency of their confidence intervals in the presence of outliers. An entire branch of statistical research is devoted to robustness. We will examine just a few of these estimators here—all from the perspective of trying to find an alternative estimator to the mean for assessing average energy savings from energy-efficiency programs.

Some Robust Alternatives to the Mean

We discuss here three robust alternatives to the mean: the trimmed mean, the bi-weight mean, and the median. Hoaglin *et al.* (Ref. 4) and Gross (Ref. 3) give more detail on the performance of these and other robust estimators of location and scale.

The trimmed mean. The trimmed mean is calculated by simply discarding a fraction of the data points from each end of the ranked data, and proceeding to calculate the mean as before. Typically, either 5% or 10% of the data is trimmed from each end. The idea is that, by removing the most extreme data, the estimate of the mean can be based on the majority of the data that is less extreme. Staudte and Sheather (Ref. 6) provide a method for calculating the standard error of the trimmed mean.

The bi-weight mean. The bi-weight (or bi-square) mean has the same concept as the trimmed mean, but it refines the process in two ways. First, it weights observations in inverse proportion to their distance from the center (median) of the data: the further away from the middle of the data, the less weight a data point is given. Data points beyond a certain distance from the center are given zero weight.

Second, the bi-weight tailors the weighting level according to a robust measure of scale. The bi-weight we use here is based on the MAD. Instead of arbitrarily deciding that 10% or 20% of the data should be thrown away as outliers, the bi-weight considers the demarkation point between the "bulk of the data" and outliers individually for each batch of data. The formula for the bi-weight used here is:

$$\text{bi-weight}(y) = \frac{\sum w_i y_i}{\sum w_i}$$

where,

$$w_i = \left[1 - \left(\frac{y_i - \text{median}\{y\}}{cS}\right)^2\right]^2$$

$$\text{if } \left(\frac{y_i - \text{median}\{y\}}{Cs}\right)^2 < 1,$$

otherwise, $w_i = 0$.

Here S = MAD(y) and c = 9. The formula for the standard error of the bi-weight mean is complicated, and not shown here. It can be found in Ref. 5, p. 208. In contrast to the trimmed mean, which can be considered a "boxcar" function, the bi-weight has a smooth weighting distribution, as Figure 2 shows.

The median. The median, of course, simply measures the middle observation (or average of the two middle observations). Because the median is completely determined by one or, at most, two data points, outliers have very little influence on the median. But, for the same reason, the median tends to be an inefficient estimator of location relative to the mean. We calculate the standard error of the median as the interquartile range divided by the square root of sample size, though other approaches also exist.

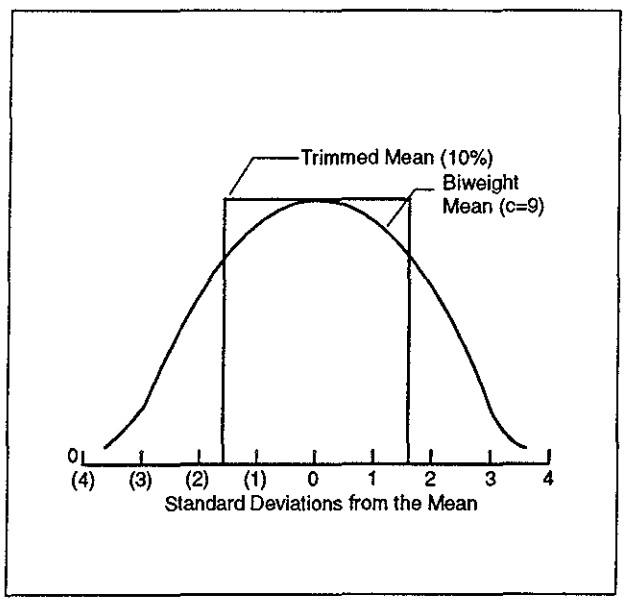


Figure 2. Weighting Functions for a Gaussian Distribution with a Standard Deviation of 1

It is important to recognize that all of the above estimators can be considered unbiased estimators of the population *mean* only if they are applied to symmetric distributions; that is, distributions that are not skewed to one direction or the other.

Synthetic Billing Data as a Test of Efficiency in the Presence of Outliers

To test the above estimators in a setting typical of energy conservation program evaluation, we devised a Monte Carlo simulation of monthly gas billing data for residential customers. We chose residential gas consumption because it is simpler to model than electricity usage or energy usage in other sectors. However, we believe the results are broadly applicable to other sectors, with some caveats, which we discuss.

Our model generates billing data for households with a defined distribution of usage, savings due to program intervention, and billing data anomalies relating to factors we describe. We used the model to generate 100 samples of 150 households each, or 15,000 households in all. Since we built the true program impacts into the model, we can directly compare sample means and the other estimators of centrality against the population true value to gauge the robustness of validity and efficiency of these estimators.

Throughout, our notation for probability distributions is:

dist. type(mean, std. dev., [minimum], [maximum])

where *minimum* and *maximum* are optional parameters that truncate the distribution to within a certain range. For example, *normal(1,0.5,0,5.0)* means a normal distribution with a mean of 1, a standard deviation of 0.5 (prior to truncation), and limited to the range between 0 and 5.0. In addition, when we specify the mean and standard deviation of a lognormal distribution, we are referring to the mean and standard deviation of the actual distribution, not the mean and standard deviation of the logged values.

The Base Model

Initial gas usage. The model begins by generating initial usage levels for each household, assuming that each house perfectly fits the PRISM model (Ref. 2): a constant base-usage per day plus a heating component determined

by heating degree days (HDDs) at some reference temperature. Specifically, the distribution of initial usage was:

heating use: lognormal(0.15,0.075) therms/degree day

base use: lognormal(1,0.25) therms/day

ref. temp.: normal(60,3,50,70) F

We chose lognormal distributions for heating and base use because usage is necessarily bounded by zero on the low end and it typically displays skewness toward high usage households, as mentioned previously. These three parameters were modelled independently of one another. When combined (and modelled using weather data for Madison, Wisconsin), they gave a distribution of annual gas usage with a mean of about 1,310 therms and a range of about 360 to 5,000 therms, with 90% of the houses falling between about 720 and 2,260 therms. This was consistent with typical gas usage for residences in Wisconsin.

Program energy savings. Assuming a rather comprehensive program, we modelled savings for the program as a combination of heating savings and base use (*i.e.*, water heating) savings. Both types of savings were allowed to vary from house to house, and were modelled to be independent of each other.

In addition, each type of savings had a relative component and an absolute component. The relative component prescribed a level of percentage savings, while the absolute component specified a level of therm savings. We modelled savings in this way to mimic the fact that some measures (such as heating-system efficiency improvements or low-flow showerheads) affect gas usage on a percentage basis, while others (such as infiltration reduction or water-heater blankets) operate more in an absolute fashion. The percent savings for both heating and base usage were defined as a normal (15,7.5,0,50) distribution. The absolute savings for heating usage (in therms per degree day) followed a normal (0.005,0.0025,0,0.75•x) distribution, where x represents total heating usage *after* the percentage component of savings was accounted for. In other words, total heating savings were limited to 75% of pre-participation usage. The absolute savings component for base gas usage were defined similarly as a normal (0.05,0.025,0,0.75 • x) distribution. On average, the relative component of savings represented three-quarters of the total savings, and the absolute component represented the other quarter.

When these savings distributions were combined with the initial gas usage distributions, they resulted in an expected value of 20.6% savings in each of heating and base gas usage. In therms, the average savings were 194 therms and 75 therms on heating and base usage, respectively, for a total of 269 therms. This 269 therm signal was the population average savings that we were looking for in our subsequent analyses.

Monthly billing periods. Usage for every house was started on a randomly selected day in May 1988. Monthly billing periods were then generated by randomly selecting the number of days in each billing period according to a normal (30,2) distribution. Twenty-five months of billing periods were generated in this manner. The first 12 months represented pre-participation periods, the middle month (which was not used) represented the month of treatment, and the last 12 months represented post-participation periods. Using each house's initial parameters and the number of days and heating degree days (at the selected reference temperature) for each billing period, we calculated the gas usage for 25 months, with use after the 13th month being reduced according to the determined savings for each house.

Perturbation Factors

Billing data generated using the base model described above are unrealistic because they are perfectly correlated to heating degree days (HDDs). Therefore, we built into the model several perturbation parameters that added a random component to monthly usage.

"Normal" monthly noise. We multiplied the usage for each month by a normal (1,0.10,0,5) factor to represent the "normal" month-to-month variation in usage. This factor reduced the r^2 between usage and degree days from 1.0 to an average of about 0.96, but produced only a few households with an r^2 of less than 0.90.

Monthly anomalies. We also gave each month a 5% chance of being anomalous. An anomalous month was perturbed by the same process as the monthly noise above, but with five times the standard deviation, generating an occasional month of billing data that did not fit the normal pattern for the household.

Structural change. We considered one of the more important types of changes that could occur to household gas use to be those that are permanent, or at least more long-lasting than a simple monthly anomaly. Examples of these kinds of changes include changing the thermostat setpoint, changes in base-usage because of changes in occupancy level, and structural changes to the house itself that affect gas used for heating. We modelled these structural changes at two levels: minor changes that were pervasive in our population, and major changes for a small fraction of households.

We specified that every household had a minor change to each of heating slope, base-use, and reference temperature exactly once during the 24-month period of analysis. The period in which this change occurred was randomly and independently selected for each of the three parame-

ters, but the change carried forward into all subsequent periods. The heating and base-usage changes were modelled as a multiplicative lognormal (1,0.1) factor. The reference temperature change was modelled as a normal (1.0,0.2,0.75,1.25) factor multiplied onto the existing reference temperature.

We used a lognormal distribution for heating and base use to account for the fact that usage can increase without bound, but can decrease by no more than 100%. The multiplicative factor specifies that some households will increase usage and some will decrease usage, but on average there is no net change in the population. The lognormal distribution is skewed, so that a few large increases in usage are balanced by many smaller decreases.

In addition to the minor structural changes for every household, we specified that the model give a 5% probability of a major structural change to each of heating and base gas usage for each household at some point during the 24-month analysis period. Major structural changes were modelled as multiplicative lognormal (1,1.25) factors on heating and base use. The few households that experience one of these structural changes have a much higher probabilities of having significant change in gas usage. But, as before, the average effect is for no net change in usage across the population of households.

Using the above model, we simulated usage data for 100 samples of 150 households each (a total of 15,000 households). We then weather-normalized the pre- and post-participation synthetic billing data using PRISM, and calculated the average savings using the estimators described previously. Figure 3 shows the distribution of pre-participation usage and savings over the entire data set of 15,000. The shapes of these distributions are similar to those for the actual residential gas data shown in Figure 1, though the scale and location are different owing to different assumptions about base conditions. The distribution of NAC savings has a kurtosis of 111, but this drops to 22 if the most extreme four observations are dropped from each end.

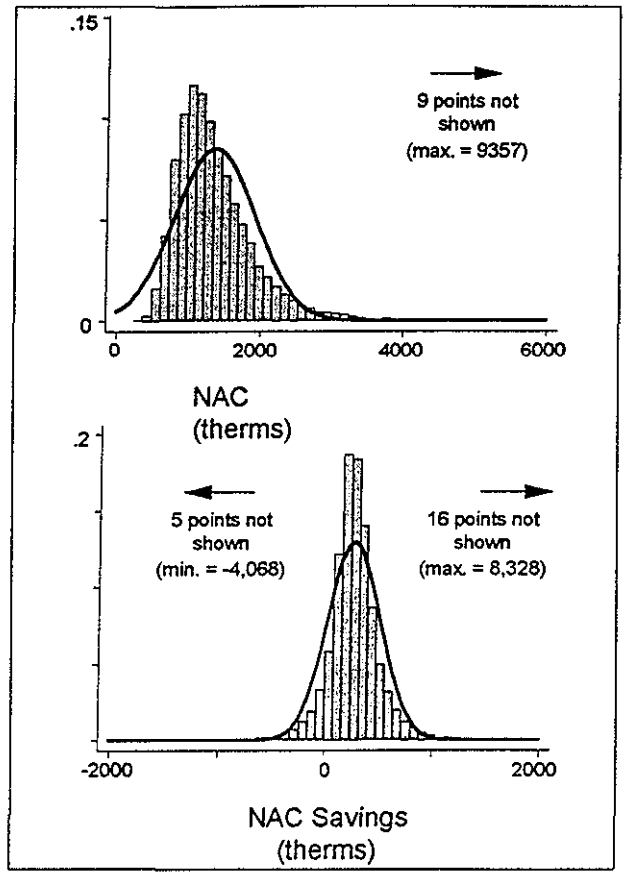


Figure 3. Distribution of NAC and NAC Savings for 15,000 Houses Generated by Monte Carlo Model

Results

The results of the runs are shown graphically as boxplots and one-way scatterplots in Figure 4, and as summary statistics in Table 1. Each stripe in the one-way scatterplot represents the results for one sample of 150 households. The boxplots show the first and third quartiles (ends of the boxes) and the medians (lines inside boxes) across samples.

The most extreme sample estimates came from the mean, which, as expected, was susceptible to influences by outliers. The trimmed mean, bi-weight, and median, were all somewhat less variable than the mean because they were less susceptible to influence by outliers. The average standard error across the 100 samples was 16.2 therms for the bi-weight, compared to 22.5 therms for the mean. On average, the bi-weight therefore was about 28% more efficient than the mean for this scenario, and would be the preferred estimator. In addition, two other measures of efficiency are given in Table 1: (1) the percent of the estimates that were within 10% of the population expected value of 269 therms; and (2) the root-mean-square (RMS) error from the true savings of 269 therms. In all cases, the bi-weight performed best, followed closely by the trimmed mean, with the median somewhat lagging, and the mean clearly in last place.

Of course, it could be argued that *all* of the estimators were precise in this scenario. If we add in a hypothetical comparison group that is similarly distributed, the 90% confidence interval on net percent savings would be about 4% using the mean ($100 \bullet 1.65 \bullet \sqrt{2} \bullet 22.5 \div 1310$), and about 3% using the bi-weight. Since the savings are about 20%, the gain in precision is not much relative to

Table 1. Summary Statistics for Various Estimators of Savings

	Estimator Performance Across 100 Samples of n=150 each[a]					Full Sample (n=15,000)	
Estimator	Mean Estimate	Standard Deviation of Estimates	% Within 10% of True Value	RMS Error from True Value	Average Precision (Standard Error)	Estimate	Standard Error
Mean	272.7	26.3	74	26.7	22.5	273	2.34
Trimmed mean	268.7	17.6	89	17.5	16.9	269	1.71
Bi-weight mean	265.5	16.2	90	16.5	16.2	263	1.56
Median	260.5	17.4	85	19.3	18.6	260	1.89

[a]The population expected value of savings is 269 therms.

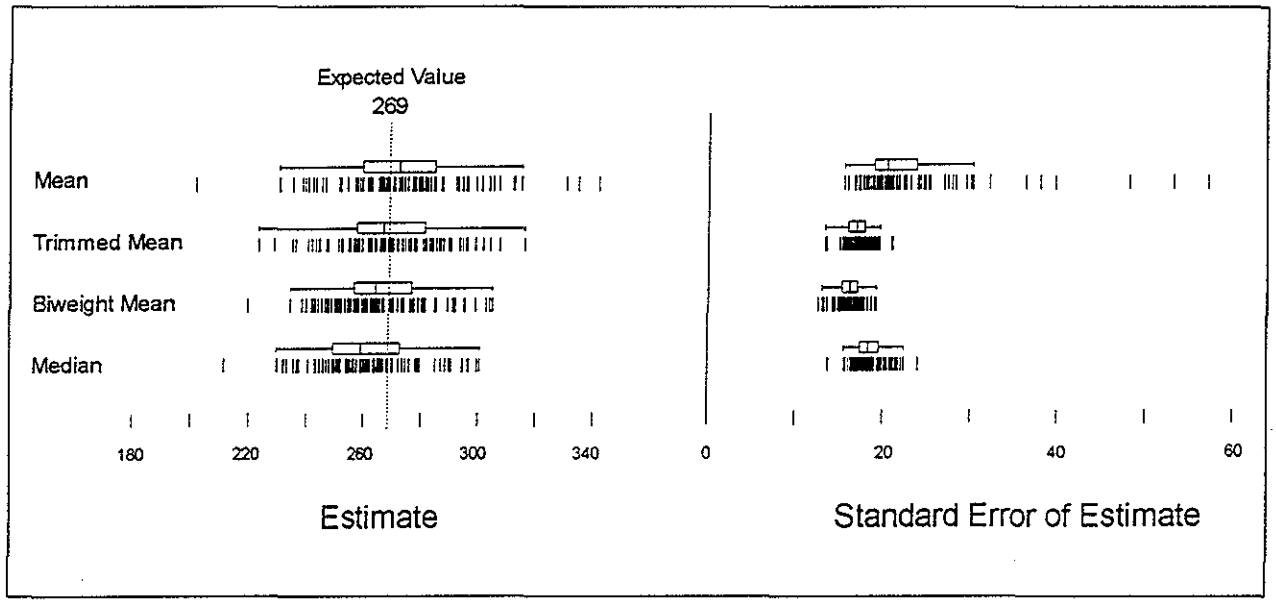


Figure 4. Distributions of Estimators and Standard Errors for 100 Samples of 150 Houses Each

the large measured savings. The rather trivial gain in precision seen here could be more dramatic and practically significant, though, in other situations that (1) have a smaller percentage impact on usage and (2) have more and larger non-program noise than we modelled here.

We should note that, overall, we found that the confidence intervals for the estimators all had about the same validity rate. Remarkably, three of the four estimators returned exactly the theoretical confidence level of 90% over the 100 samples. In other words, 90 out of the 100 samples had confidence intervals that enclosed the true population savings of 269 therms. The fourth, the median, differed only slightly, with a 91% validity rate.

Are the estimators biased? The results of our runs show an apparent tendency toward lower savings estimates for the trimmed mean, bi-weight, and the median. When we looked at these estimators applied over the entire 15,000 household data set, we found that they did indeed exhibit a slight downward bias, as Table 1 shows. In fact, for this large sample size, the confidence intervals for all but the mean did not enclose the true level of savings. The reason for this small bias was probably the fact that usage and, to some extent, savings were both skewed towards high values. Usage was skewed because we defined it to be a lognormal distribution; savings were skewed because a portion of the savings was relative to usage, which was skewed.

This is an important point because, in general, these estimators can only be considered unbiased estimators of the population mean if the underlying distribution is symmetric. If it is not, robust estimators tend to neglect data

from the skewed tail of the distribution, and hence come up with estimates that are on the low side, albeit trivially, as in the case here. Note that at the more typical sample size of 150 households, the confidence intervals for all of the estimators showed the correct coverage rate, and did so with greater efficiency than the mean.

In other populations, though, the bias could be more substantial. If savings are strongly skewed, or if usage is strongly skewed and savings tend to be a percentage of usage, then the bias from using the robust estimators could be unacceptable even for moderate sample sizes. To illustrate this point with an extreme example, consider a population of commercial customers who receive rebates for efficiency improvements in HVAC equipment. Ninety-nine percent of the participants are small customers, but one in 100 is much larger, and has savings that tend to be 100 times that of the smaller customers. If we blindly apply any of the three robust estimators to a sample of 99 small participants and one very large participant, the robust estimators will identify the large participant as an outlier, and return estimates that are good estimates of the average savings for small customers, which we will call x, but are very poor (and biased) estimators of the population savings, which are on the order of $2x$. The mean would give the correct estimate of savings, but would also provide very poor precision, reflecting the extreme heterogeneity of the population.

This situation could be handled by stratifying the population on usage (if usage is known for the population), and estimating savings for each stratum separately (although this would leave the problem of a sample size of one in the large-facility stratum). In less obvious situations, though, indiscriminate use of robust estimators could lead to biased estimates of impacts. The analyst must be careful to look for sources of such bias, particularly if usage is strongly skewed or predicted impacts are skewed. Stratifying the population and the study sample into more homogenous subgroups may help reduce the kind of bias shown above.

Using Prior Predictions of Savings

It has become common for energy conservation programs to employ tracking systems that contain engineering predictions of the impacts from measures installed in each facility. This information can be valuable in two ways. First, by focusing the analysis on the relationship between predicted savings and measured changes in usage, precision can be increased. Second, the existence of prior predictions of savings helps in the detection and treatment of outliers. A 50% drop in usage after participation in a program might seem to be a legitimate program effect, unless we know that the measures installed in the facility were predicted to save only 0.5% of pre-participation usage.

Prior predictions of savings can be incorporated into outlier screening criteria to help lessen the danger of excluding facilities that show unusual but legitimate program effects. This might involve trimming the facilities that are outliers with respect to differences between predicted and observed change in usage. As discussed above, the potential bias introduced by doing so should be assessed. Another strategy is to regress the observed change in usage on predicted savings using a regression procedure that is more robust than ordinary least squares. In this case, we are seeking a robust estimate of the realization rate of predicted savings. Refs. 4, 5, and 6 describe using the bi-weight and other robust estimators in the context of regression.

Often, billing data are combined with survey information and multivariable regression is performed, of which predicted savings constitute only one explanatory variable. Here the intent is to not only uncover the realization rate of predicted savings, but also to control for other non-program effects in an effort to increase the precision of the estimate and reduce bias between the treatment and comparison groups. Identifying and handling outliers in this setting is the subject of our last section.

Outliers and Multiple Regression

The subject of outliers in multiple regression is a broad, complex, and often controversial topic, which is the focus of several entire textbooks (for example, see Ref. 1) and which we could not hope to adequately cover here. Yet the widespread use of regression in impact evaluations using billing data demands at least a cursory review of the topic.

One of the fundamental problems with ordinary least-squares regression is its lack of resistance to outliers. Entire regression models can be made or broken based on a single outlier. Coefficients can be badly biased, while standard errors and r^2 can either be inflated or deflated by outliers. In other cases, outliers may just cause lost precision and no appreciable bias. The difficulty arises in identifying outliers, discerning their impact, and deciding what to do about them. This process is made even more complex by the attraction of the regression fit to unusual observations, causing the definition of outliers to become hazy.

Regression outliers can be simple univariate outliers, such as one wild value on one variable, or they can be multidimensional outliers that involve an unusual combination of values for two or more variables, although the value for each variable may not be unusual at all. In a simple regression involving just one explanatory variable, outliers are usually obvious in a basic scatter plot. Multi-

dimensional outliers are much more difficult to detect due to our limited ability to plot data in more than three dimensions. The effects of multi-dimensional outliers may also be exacerbated when collinearity between explanatory variables means that individual coefficients are poorly determined.

Examination of residuals from the regression fit is one of the primary and most valuable approaches for identifying outliers and unusual observations. But in many cases, the standard techniques do not work as well as one might like because the regression fit itself is so contaminated by the outliers that they no longer seem unusual. An unusual observation may not be an outlier in the residuals because it has attracted the regression fit; conversely, an outlier in the residuals may not be a particularly influential data point and therefore have little affect on the fit. This phenomenon has led to the search for not just "outliers" (in the residuals) but also for influential observations. Statistical measures of the impact that a given observation has on the fitted values or the value of a particular coefficient are available as a standard component of most statistics software (*e.g.*, influence, leverage, Cook's D, Dfbetas). These statistics are often quite valuable in identifying unusual and influential data points.

Many of the regression diagnostics are less discerning when there is a clump of outliers, since each single outlier has little influence on the results if the other outliers remain in the fit. One approach that can work under a variety of circumstances is to fit a "robust" regression that is resistant to outliers. The residuals from this fit will often provide easy identification of outliers because their influence is reduced. Comparison of these results to a standard regression can often help indicate whether the standard fit really represents the bulk of the data or just the influence of a few outliers.

As this brief discussion may indicate, the analysis of outliers and influence in multiple regression requires a careful and sophisticated analysis and therefore is often not done. One-stage multiple regression approaches to impact evaluation are particularly difficult to diagnose (in terms of outliers and influence, and probably more importantly in terms of model specification—a topic outside the scope of this paper) because of both the size and complexity of the combined time-series cross-sectional data set. The authors have not seen any published analysis of influence or outliers in studies of this type.

Two-stage approaches, which first summarize the billing data for each case (through PRISM modelling or other analysis) and then analyze savings across cases through simple summary statistics or multiple regression models, tend to be more amenable to identification of outliers and general statistical diagnostics. The first-stage analysis can be used to help identify within-facility data anomalies (*e.g.*, meter read errors or unusual usage patterns not consistent with expectations such as electric space heating in a baseload-only program). The results of the first stage can then be used to identify unusual between-facility variations such as unusual usage or savings relative to other facilities. The second-stage analysis can then be performed after cleaning up many of the outliers and with and without already-identified extreme cases included.

Conclusions

Identifying and treating outliers in customer billing data can lead to better estimates of program impacts. This process begins with an understanding of the factors that cause outliers, and how they affect the distribution of energy savings. Robust alternatives to mean savings are more resistant to the effect of outliers, but the analyst must be careful that these do not exclude or downweight facilities that are outliers *because* of the program or bias the results in favor of smaller facilities. Outliers and influence points deserve especially close scrutiny in multiple regression, where they can be difficult to detect, but can strongly affect results.

Because outliers are almost always a concern in billing-data-based impact evaluation, it is important that any methods used to ameliorate their effects be reported and discussed. It is perhaps best to report results for both traditional and robust estimators of savings, with an explanation of why the results might differ and which results are considered superior. This is especially true as billing-data-based impact evaluation is used increasingly to support utility cost-recovery and financial-incentive payments for energy-efficiency efforts. All parties involved in these processes need to be assured that the reported evaluation results are robust, and not unduly influenced by a few outliers that may well be due to factors unrelated to the program.

References

(1) Belsley, David A., Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics*. New York: John Wiley & Sons, Inc., 1980.

(2) Fels, M. F. "PRISM: An Introduction," *Energy and Buildings*, Vol. 9, No. 1, pp. 5-18.

(3) Gross, Alan M. "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association*, Vol. 71, No. 354, June 1976, pp. 409-416.

(4) Hoaglin, David C., Frederick Mosteller, and John W. Tukey. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons, Inc., 1983.

(5) Mosteller, Frederick, and John W. Tukey. *Data Analysis and Regression.* New York: Addison-Wesley, 1977.

(6) Staudte, Robert G., and Simon J. Sheather. *Robust Estimation and Testing.* New York: John Wiley & Sons, Inc., 1990.