92.059E

Prepared by:
Proctor Engineering Group, Ltd.
San Rafael, CA 94901
(415) 451-2480

The Need for Statistical Analysis And Reporting Requirements: Some Suggestions for Regression Models

Prepared for
1995 Energy Program Evaluation
Conference, Chicago
August 22-25, 1995

Final Report

Prepared by:
Michael Blasnik

Creators of CheckMe!®

THE NEED FOR STATISTICAL ANALYSIS AND REPORTING REQUIREMENTS: SOME SUGGESTIONS FOR REGRESSION MODELS

Michael Blasnik, Proctor Engineering Group, Boston, MA

Abstract

Several regulatory bodies have developed DSM impact evaluation standards. The technical aspects of these standards have primarily focused on sample sizes, methodology choices, and internal measures of uncertainty. While these issues are important, greater challenges to evaluation reliability may come from external sources of uncertainty such as sample bias, reliance on statistical methods whose underlying assumptions are not met, and from choices made during the analysis which are not fully explained or justified. This paper describes some of these threats to reliability and provides examples which indicate the potential magnitude of their impact on results, particularly focusing on SAE-type regression models. Reporting and analysis requirements are proposed which may help in identifying and assessing these potential problems.

The general approach proposed is based on the idea that quality evaluations describe and test key assumptions to the extent possible, accompany the analysis with a well-reasoned narrative which explains the role and impact of analytical choices and statistical models, provide readers with sufficient information to assess the conclusions drawn, and include appropriate caveats.

The proposed requirements are assessed in terms of some of the arguments against evaluation standards (excessive cost of compliance, stifling of innovation). While it is hoped that the proposals will enable evaluation consumers to become better informed of the true uncertainties and analytical choices involved in impact evaluations, they should not be interpreted as providing "quality assurance". The recommendations are only intended to help uncover some of the more common problems, no assurances can be given that a model which appears "OK" actually provides a reliable estimate. As a professional courtesy, and in some cases to maintain confidentiality, examples from actual evaluations and published papers are generally used without citation.

Background

Impact evaluations of DSM programs have grown in importance as regulators in some jurisdictions have tied shareholder incentives to measured results. Regulators in several states have developed protocols for conducting impact evaluations in an effort to produce more reliable estimates of measured savings. In addition to addressing the basic evaluation issues of what, when, and how often, evaluation protocols have devoted a good deal of attention to the selection of precision requirements and evaluation methodologies.

Precision requirements have been the subject of ongoing debate in the DSM field. However, one critical issue missing from the discussion has been the reliability of the reported precision; i.e., Does the reported confidence interval accurately reflect the uncertainty in the savings estimate? It is often assumed that the reported standard errors and confidence intervals are accurate reflections of uncertainty. However, these measures of precision are themselves statistical estimates subject to potential bias and are predicated on the assumption that the impact estimate is unbiased. The accuracy of uncertainty estimates needs to be assessed if the debate over precision levels is to be truly useful.

Multiple regression models, such as SAE or CDA models, have been the primary approach promoted by the evaluation industry and by evaluation protocols for estimating kWh impacts of major programs. Regression approaches have been considered superior to simpler analyses because they are designed to control for confounding non-program effects and potentially biased comparison groups. The implicit assumption is that they will work as designed and provide more precise and less biased estimates of program impacts. Until recently, little attention has been given to assessing the reliability or stability of these models.

It is the author's belief, based on experiences working as an evaluator and as an evaluation reviewer on behalf of regulators, intervenors, and implementors, that much of the statistical analyses, and most of the regression models, presented in evaluation reports are subject to many potentially significant threats to their ability to provide reliable impact estimates. The proposals in this paper are based on the observation that these problems are often not adequately identified or addressed.

Potential problems with the quality of statistical analyses have also been recognized in California and led to the recent development of quality assurance guidelines (which are currently under review). These guidelines address many of the key issues that may undermine the reliability of impact estimates. The approach taken by the guidelines is to require evaluators to describe how they have dealt with certain common analysis problems, particularly those related to regression models. The guidelines are not prescriptive in that they do not require specific methods for identifying

problems and do not indicate how problems should be resolved. The rationale for this approach is that it provides evaluators with flexibility to choose among a number of legitimate methodological choices. One disadvantage of this flexibility is the difficulties it may create for readers in both comparing different evaluations and in becoming familiar enough with the variety of techniques to properly assess them. An alternative approach, taken in this paper, is to create minimum requirements which evaluators would be free to expand upon or add to as they see fit.

General Approach to Evaluation

Because they are intended to evaluate the effectiveness of an operating program, DSM evaluations are seldom based on a true experimental design with random assignment of treatments and substantial control over experimental conditions. Instead, they are observational studies of a complex system of engineering and behavioral effects which tend to be based on "quasi-experimental" designs. The problems with observational studies are well known in the statistical literature. For example, noted statistician William Cochran has stated [1]that an investigator

> "may do well to adopt the attitude that, in general, estimates of the effect of a treatment or program from observational studies are likely to be biased. The number of variables affecting y on which control can be attempted is limited, and the controls may be only partially effective on these variables. One consequence of this vulnerability to bias is that the results of observational studies are open to dispute."

Most advanced impact evaluation techniques are intended to reduce bias by controlling for as many potentially confounding factors as possible. But, as Cochran points out, one can never be certain whether all or even most of the important sources of bias have been identified. Even for those sources properly identified, the effectiveness of the techniques which attempt to deal with them is uncertain. A quality evaluation recognizes the existence of this fundamental challenge and attempts to identify and address threats to reliability through a combination of careful data analysis (employing multiple approaches where feasible), well reasoned conclusions, and appropriate caveats.

In contrast, DSM impact evaluations often display a great deal of confidence in the approaches and results. For example, many SAE-based evaluations include a statement such as "SAE models are able to control *for* confounding factors that affect energy use" (emphasis added). Such statements are not confined to evaluation reports but also appear in DSM evaluation handbooks, one such example is "regression models can control most of the confounding factors that determine energy usage, so the evaluation researcher *can be certain* that the effects being measured are due to the DSM program and not to other, non-program, factors" (emphasis added). Caveats about potential biases or modeling problems are relatively rare.

For reasons explained by Cochran above, and elaborated upon in this paper, the optimism displayed in many impact evaluations may be unfounded. Instead, when performing or assessing an evaluation, it is generally wise to assume that all samples are biased and that the data fail to meet the underlying assumptions behind the statistical analyses performed. The burden of proof rests upon the evaluator to investigate identifiable threats to validity and to provide supporting evidence that the conclusions drawn are reasonable.

Regression Analysis: Some Potential Problems

Regression analysis has been termed the most used and most abused statistical tool [2]. When it works as intended, regression is a powerful tool for analyzing data and uncovering relationships. Yet, the reliability of a regression model is dependent upon many assumptions which are virtually never fully satisfied in practice, particularly for observational studies. Quality evaluations recognize the assumptions which the analysis methods rely upon and, to the extent possible, test the degree to which they are satisfied. Because of their sensitivity to certain violations of assumptions, regression models are particularly challenging to employ successfully without being misled by faulty analysis or interpretation. Meeting this challenge usually requires a healthy degree of skepticism combined with considerable expertise about not only data analysis, but about the subject being evaluated.

SAE Models

SAE-type regression models are intended to control for non-program factors which influence energy usage and therefore improve precision and/or reduce bias in savings estimates. The typical model specification attempts to predict post-program energy usage as a function of pre-program usage, engineering-based predicted savings, and a variety of survey responses to questions concerning changes in facility use, business activity and equipment level. Models may also include some demographic variables and often incorporate a variable derived from a logit participation model. The coefficient on the predicted savings is interpreted as the "realization rate", which is meant to represent the average proportion of predicted savings actually realized by the program.

Because an SAE model is based on the change in energy usage (since pre-program usage is included as an explanatory variable), it may be seen as a way of adjusting a simple pre/post treatment/comparison savings estimate for differences captured in the variables representing non-

program effects. In fact, one can view the simple comparison approach as a regression model of change in usage as a function of a constant and a dummy variable indicating participation. SAE models attempt to improve this simple model by including more variables to explain changes in usage. If there are no systematic differences between the participant and comparison groups, then no adjustment to simple pre/post results is needed and the SAE model should produce essentially the same savings estimate as the simple comparison but with greater precision (because of usage variations "explained" by the variables in the model). However, if the comparison group differs from the participant group, then the regression model attempts to control for these differences and adjusts the savings estimates to account for this bias.

If an SAE model does not properly control for non-program effects, the savings estimate may be adjusted inappropriately. Given this possibility, a quality evaluation provides a narrative which explains what the model accomplished (or attempted to accomplish). This narrative would include a discussion of how and why the results differ from a simpler pre/post analysis and would describe any important confounding factors which were identified and how they affected the results. Without such a narrative, the reader is not given enough information to assess whether the model is believable.

Model Specification Issues

The most fundamental assumption made by regression modeling is that the model is "correct" -- it includes all of the variables which influence the dependent variable and the functional form of the relationship is properly specified. Of course, few evaluators would claim that their SAE model includes all factors influencing energy usage. However, some would point out that the model does not necessarily have to be correct for the impact estimate to be unbiased. This statement is true if all of the variables which are omitted from the model are unrelated to the variable representing the impact estimate (i.e., predicted savings in SAE models). There is no method available which can prove that this is the case for a particular model, although there are some tests for omitted variables which may disprove it. This fundamental threat to the reliability of regression coefficients is well known in the statistical literature [3], yet it has received little recognition, and has even been disputed, in the DSM impact evaluation field.

A simple example may help illustrate this problem. Using data from a residential conservation program, the author fit a regression model of pre-program energy usage in terms of house airtightness (measured in CFM50 by a blower door). The model indicated that each CFM50 increased gas usage by 0.12 ccf/yr. (+/-.02 @90% conf.). Engineering algorithms indicated that the impact should only be half as large, yet this value is far outside the confidence interval. The discrepancy is due to omitted variable bias -- there are factors correlated with airtightness that also affect energy usage. The airtightness of the building acts as a proxy for related omitted variables, most obviously the size of the house. When the model was re-estimated including the area of the house, the new coefficient on CFM50 was .06 (+/- .01) ccf/yr. This new value is consistent with expectations and is statistically significantly different from the initial model's coefficient. The model with the omitted variable produced a biased coefficient and the confidence interval provided no indication of a potential problem. In fact, the model indicated that the coefficient estimate was very precise, yet it was precisely wrong. This example is simple and perhaps obvious to many readers. Unfortunately, the problem that it illustrates is often not obvious and quite difficult to detect in practical applications of SAE models which are considerably more complex.

Potential problems with model specification threaten the reliability of all regression models. This statement should not be interpreted as suggesting that all SAE models give "bad" answers or that regression should be abandoned or that simple pre/post comparisons will give better answers. The intent is that such model results should be presented with this point in mind and that evaluations which rely solely upon a single regression model coefficient are at risk of providing misleading answers. Quality evaluations are cautious in drawing conclusions about regression coefficients, attempt to estimate savings using multiple approaches, and compare results to related studies.

In addition to potential omitted variable bias, SAE modelers need to be aware of a variety of other specification issues, including collinearity and model selection subjectivity.

Collinearity can cause problems in SAE models when "control" variables accidentally capture program effects. If an SAE model includes a variable (or set of variables) which is strongly related to participation, then such a variable may absorb some of the program impact and reduce the estimated realization rate. Regression models have difficulty fully distinguishing the separate impacts of correlated explanatory variables. In extreme situations (unlikely to occur in most SAE models, but common in CDA models), coefficients are poorly determined because two or more variables are highly correlated. A variety of approaches are available for identifying such extreme situations (e.g., variance inflation factors, condition indices) and several possible approaches may be pursued (e.g., dropping a variable, ridge regression).

In the context of SAE models, the problem is usually not as severe, but the impact on savings estimates may be substantial. One approach for detecting potential problems is to examine the correlation matrix on the estimated coefficients. Coefficients which are well correlated with the realization rate coefficient may deserve further scrutiny. The related coefficients may not cause a problem and, indeed, are considered quite valuable as they represent the confounding factors which SAE models are meant to control for. However,

their impact on savings estimates needs to be assessed and explained.

Another technique which can help identify which variables most affect the savings estimate and may also provide a better understanding of the model, is to re-fit the model in steps. One can start by fitting the simplest model and then examine how the savings estimate changes as additional terms are added. For example, a model of change in usage with just a constant and a participation dummy variable can provide a baseline equivalent to a simple pre/post analysis. The participation variable can then be changed to predicted savings with pre-program usage added, then the other variables can be added in order of perceived importance. This exercise can help identify which variables affect impact estimates the most and may help the evaluator to describe what the SAE model actually accomplished. If the savings estimate is stable under a variety of specifications, then the SAE model is not adjusting for sample biases, it is merely attempting to improve precision. If the savings estimate varies dramatically when a particular variable is included, then an explanation can be sought.

For example, if including a variable which is intended to reflect changes in business activity shifts the impact estimate upward, an examination of the data may reveal that the non-participant sample was more likely to be downsizing and therefore their consumption declined at a greater rate than would have happened to participants without the program. While the accuracy of these explanations cannot be tested, the fact that the evaluator can create a sensible narrative which describes how and why the model affected the impact estimates and why the final model is reasonable can provide crucial supporting evidence for their conclusions. If such a "story" can't be created, then the evaluator needs to look closer at the model and perhaps consult with program implementors for potential theories. (Note: fitting a model in steps can be quite sensitive to the order in which the variables are entered, although in the author's experience it is often quite useful for SAE models.)

In addition to technical problems such as collinearity, SAE modelers need to be aware of potential biases which may be introduced in the model building process. In the course of performing an evaluation, the search for the "best" model is typically a key part of the analysis. The process of fitting and comparing different models is considered by many an "art", which renders it subjective and open to potential manipulation. Experienced evaluators know that, by choice of model specification, they can usually have a meaningful effect on the impact estimates. Decisions concerning data screening and sample selection can exert similar influence on virtually all analysis methods. Because these threats to unbiased results usually can not be eliminated, they need to be addressed through reporting. Quality evaluations document key decisions which are likely to affect impact estimates (data screening, sample selection, and model specification choices) and provide a rationale for the particular choices made. The impact of such decisions on the final results is provided and compared to other reasonable choices that could have been made. The range of values for the realization rate under the different model specifications tested is often a useful part of this reporting.

Heteroscedasticity

Regression models, and particularly the estimated standard errors, rely upon the assumption that the residuals are independently and identically distributed. Two common violations of this assumption are serial correlation and heteroscedasticity. For the typical SAE model based on annual (not monthly) pre and post program consumption, serial correlation is not a significant issue. However, heteroscedasticity, which refers to non-constant error variance, is a common problem in SAE models especially those applied to the commercial and industrial sector.

Because high use buildings tend to have more variable energy usage than low use buildings (in absolute kWh), SAE models which include buildings of widely varying usage levels experience heteroscedasticity. The primary effect attributed to heteroscedasticity is that it biases the estimated standard errors. However, it can also exacerbate other model problems leading to substantial changes in the estimated realization rate and reduced model accuracy. Heteroscedasticity often reveals itself through large influential outliers because modest usage variations for large facilities appear as tremendous changes in usage compared to the variations seen in the majority of (smaller) buildings in the sample. These high use facilities with large influence may substantially reduce the accuracy of an SAE model, while the standard errors indicate that the model is quite accurate. An example using synthetic data may help demonstrate the potential importance of heteroscedasticity and its relation to outlier problems in SAE models.

Monte Carlo simulations allow one to create a synthetic world where the true answers are known and all sources of variability are specified. Repeated replications of this known world allow one to assess the performance of different statistical estimators under the given assumptions. The author performed a Monte Carlo analysis of a simple commercial DSM program. The mean values and variability in usage, predicted savings, realization rates, and post-program usage were specified as follows: pre-program usage was log-normally distributed (where log(usage-20,000) has mean 4.8 and std. dev of 0.42), predicted savings averaged 15% of this usage (with 5% std. dev), the average true realization rate was 75% (with 15% std. dev.), and post program usage averaged pre-program usage minus true savings with an added random variation of 10% of usage. These values provide a relatively well-behaved data set with fairly tight distributions and no sources of bias. The log-normal usage distribution leads to a ratio of about 100:1 for largest to smallest usage rate. The 10% random variation in post-program usage is the source of the heteroscedasticity

since it makes the standard deviation proportional to usage. This assumption is believed to be more realistic than the constant kWh value assumed by ordinary regression.

The Monte Carlo analysis involved generating the values of all variables using these specifications for each of 1000 buildings with half the buildings randomly declared non-participants. The resulting data set was analyzed using a simple SAE model of post program usage with pre-program usage and predicted savings as the explanatory variables. A simple pre/post analysis was also performed. This entire 1000 building data generation and analysis process was replicated 1000 times and the resulting 1000 realization rate estimates and confidence intervals were compiled.

The analysis revealed that the 90% confidence interval from the SAE model included the true realization rate only 35% of the time! The true uncertainty in the SAE realization rate was three and a half times greater than reported. In contrast, the 90% confidence interval from the simple pre/post analysis included the true value 94% of the time. In addition to providing a conservative confidence interval, the simple pre/post analysis proved to be more than twice as accurate at estimating the realization rate than the SAE model (as measured by the median discrepancy between the estimate and the true value which was .09 for the SAE model and .04 for the simple pre/post).

Overall, the SAE model claimed to be about twice as precise as the simple pre/post but was only about half as precise. The failure of the SAE model to properly cover its confidence interval is an expected result of heteroscedasticity. The relatively poor accuracy of the SAE model is also due to heteroscedasticity as the greater absolute usage variations in high use buildings lead to relatively wide fluctuations in the estimated realization rate because of their large influence on the model fit. This problem also manifested itself through apparent outliers. An average of 20 observations per replication (2% of the sample) had studentized residuals greater than three in absolute value, while one should only expect about 2 such observations in a sample of 1000 (see next section). Additional simulations performed using different specifications (including different usage distributions) found varying but similar results for all but the homoscedastic error case (where the SAE model properly covered its confidence interval and was slightly more accurate than the simple pre/post).

In addition to the Monte Carlo findings, heteroscedasticity can lead to other problems under more complex situations. Heteroscedasticity can be viewed as improper weighting of the observations. If realization rates are thought to vary across facility or measure types and the model is attempting to estimate the "average" rate, then one result of this improper weighting may be incorrect "averaging" of these realization rates. Fixing heteroscedasticity involves downweighting observations with higher variability. Such fixes may be at odds with efforts to properly weight samples for representativeness. For example, if much of a program's predicted impacts occur in very large facilities then one would want these facilities to have greater weight in the impact estimate. But if usage rates are more variable in large facilities, then correcting for heteroscedasticity would involve downweighting these facilities.

Given the issues described above, principles of sound data analysis dictate testing for heteroscedasticity in all regression models, particularly those involving samples with a large range of usage rates. There are a variety of tests available (e.g., Breusch-Pagan, White, Cook-Weisberg). When applied to C&I SAE models which use ordinary least squares, the tests virtually always indicate a problem. When heteroscedasticity is found, the estimates are suspect and the standard errors are invalid. The situation may be improved by respecifying the model, using stratification or weighted least squares, or calculating standard errors that aren't dependent on homoscedasticity. The rationale for the selected approach needs to be stated and the impact reported.

Outliers and Influence

Ordinary regression models are notorious for their sensitivity to outliers. The topic of detecting and dealing with outliers and influential data points is broad and somewhat controversial (see [4] for more detailed treatment). Until recently, the issue received little attention in DSM evaluations, but has become more widely recognized as investigations have found some SAE models that are tremendously influenced by just a few buildings.

In one large scale impact evaluation, the realization rate varied from 0.27 to 0.87 depending on the inclusion of fewer than ten buildings out of a sample of more than a thousand. While the concerned parties can debate the merits of keeping, removing or downweighting these buildings, none of these options is necessarily the "correct" answer. Assuming that the outliers have not been identified as some type of data error, the real issue is whether the SAE model is viable as specified. Such tremendously influential outliers usually indicate a problem with the model specification. An examination of that model reveals a questionable specification that also undoubtedly suffers from heteroscedasticity (which may be responsible for some of these outliers). Only after such a model is re-estimated with a better specification should the issue of removing or keeping outliers be addressed.

If the model is still sensitive to a few observations, then one needs to assess whether these buildings' influence on savings is appropriate (based on participant population characteristics). If a building was expected to save a significant fraction of the total savings of a program, then a large effect on the impact estimate may be acceptable. If a building with only 1% of the predicted program impact changes the savings estimate by 30%, then the observation may have too much influence.

There are many techniques for assessing outliers and influence and different analysts may have different favored approaches. However, it may be wise to require some minimal analysis and reporting requirements otherwise the resulting array of approaches employed may make it difficult to interpret or compare studies to each other. For standard SAE models, three approaches are proposed as a useful minimum requirement.

First, studentized residuals should be calculated to identify outliers. Observations with studentized residuals greater than 3 in absolute value are usually worth further investigation (since their values should be distributed approximately as a t-statistic, there should only be about 2 such observations in a sample of a thousand). The evaluator should identify and describe the observations with the five largest values above this cut off. The impact on savings estimates from removing all such observations should be presented. If more than about 1% of the observations are in this category (as in the Monte Carlo analysis in the previous section), then the model is probably suffering from uncorrected heteroscedasticity or other problems.

Second, df-betas should be calculated to identify observations with large influence. Df-betas are valuable diagnostic statistics for assessing SAE models because they directly measure how much each observation affects the realization rate. Observations which affect the realization rate by more than a few percent may be worth investigating (the cut-off number should depend on the size of the sample and the relative predicted impact of the observation, but three to six percent may be a reasonable starting point). Again, the observations with the five largest df-betas should be identified, and the impact of removing all observations beyond the cut-off reported.

Third, robust regression (e.g., bi-weighted least squares or least absolute values) should be used to re-estimate the model in order to assess the overall quality of the fit to the data. If a robust fit gives very different impact estimates, it implies that the model does not fit the bulk of the data very well and needs further investigation. Discrepancies need to be explained. The Monte Carlo simulations described in the previous section also included use of least absolute values (LAV) regression to estimate the SAE model. Due to its greater resistance to outliers, the LAV model proved to be almost twice as accurate as the standard SAE estimate and comparable to the simple pre/post. In addition, the LAV model properly covered its confidence interval (when standard errors were estimated through bootstrapping, but not when estimated analytically). These results do not necessarily mean that the LAV model should be used instead of ordinary least squares, but do tend to endorse the principle of using LAV estimates as a cross-check on the standard results.

The particular cut-offs for studentized residuals and df-betas cited above are suggestions based on theory and experience. Reasonable arguments could be made for using different values. However, some values need to be selected in order to provide consistency between evaluations and to help the evaluation field develop a better sense for what values may be typical and how large a problem outliers and influence points may be. Evaluators should feel free to supplement any minimum requirements with additional preferred tests or approaches. Regardless of the particular approach taken, quality evaluations provide information on the potential extent of outlier and influence problems and show the impacts of any analytical choices on the results.

Sample Representativeness and Participation Models

The issue of sample representativeness is critical in impact evaluations. Biases can arise from a participant sample which doesn't represent the participant population, or a non-participant sample which doesn't represent the participant population (since their usual role is to indicate what would have happened to participant usage if the program hadn't existed). Biases tend to occur in the initial sample selection process, through data screening, and from survey non-response. If both samples are perfectly representative, then a simple pre/post treatment/comparison evaluation design provides unbiased results. One of the prime justifications for SAE models is that unbiased samples are very hard to find and a regression model is one way to try to control for these problems.

SAE models, like all evaluation methods, still depend on the representativeness of the participant sample. While a model may be able to capture some confounding factors affecting changes in energy usage, the sample must still represent the population in terms of specific technologies and applications of measures and their impacts. SAE models also depend on the non-participant sample to represent how the participant's usage would have changed if subject to the same factors included in the model. Potential sources of bias include systematic differences between the groups which aren't captured in the model and/or differences which are captured in the model but affect the participants and non-participants differently. Because SAE models require data from surveys, they may add to sample problems due to survey non-response bias.

In a more sophisticated attempt to deal with comparison group representativeness, and particularly self-selection bias, many evaluations combine a logit participation model with an SAE model. These models are subject to many of their own problems, including poor predictive ability. Participation models also make certain assumptions about sample representativeness and in many cases may be trying to correct for sample differences which are really due to differential non-response biases between the samples. In addition, there has been considerable debate about what exactly the nested logit/SAE approach is really trying to accomplish [5]. When the modeling approach works properly, it may be estimating the wrong thing -- what the savings would have been if everyone were forced to

participate, as opposed to removing the effect of what the participants would have saved if they hadn't participated.

Sample problems are frequently downplayed in DSM evaluations. For example, a recent residential evaluation found that the surveyed non-participant sample used 30% more energy in the pre-program year than either the participant population or sample. The text noted that the usage is "moderately higher" and then stated "These differences are controlled for in both the participation decision and energy impact model". In the same report, evaluating another customer sector, a table of summary statistics reveals that the surveyed participant sample used 20% less electricity in the pre-program period than the full participant population or the non-participant sample. The text accompanying the table stated "The three groups are roughly similar in terms of initial consumption". However, the differences were highly statistically significant and clearly of practical significance. In addition, a simple pre/post savings calculation indicates that the participant sample had apparent net savings 40% lower than the participant population. A logit participation model indicated that pre-program energy usage is the only significant determinant of participation. This "finding" was then incorporated into an SAE model, attempting to adjust the impact estimates for differences between the participants and non-participants that may only exist due to non-response bias, not actual population differences. Ironically, the net result of the two stage modeling process was a savings estimate indistinguishable (within 5 kWh/yr.) from the simple pre/post comparison of the analysis samples. It is not clear what relation this savings estimate bears to the actual program impact, although one could make a reasonable argument that it is 40% too low given the sample bias.

Because of the underlying assumptions about sample representativeness which all evaluation methods rely upon to some extent, a detailed assessment of sample representativeness should be an integral component of all evaluations. This assessment needs to go beyond simple means and t-tests (which are commonly misinterpreted as proving that the two groups are the same, and are subject to type II error). Comparisons should be made between all relevant groups (populations, initial samples, and final analysis samples) on all available variables (e.g., sector, building size, occupancy, major end uses, energy usage, measure types, predicted savings, and all variables used in statistical models). The comparisons should include an analysis of the similarity of the distributions (not just means) of the variables particularly including the "tails" (regression models tend to give the greatest weight to extreme observations, making the representativeness of such values in the sample critical). Graphical approaches (e.g. histograms or quantile-quantile plots) and/or simple reporting of percentiles (e.g., min, 1st, 5th, 10th, 25th, 50th, etc.) could be used. For variables in regression models which frequently take on zero values (e.g., dummy variables), the proportion of zeros and the distribution of the remaining values (if they vary) should be reported.

While these proposed requirements may seem onerous in comparison to typical practices, compliance should be relatively easy since almost any statistics package can produce these results easily. The sheer quantity of information may be overwhelming in some cases and will require a clear presentation format and a useful narrative. In addition to assessing representativeness, the resulting information should also provide greater insight into participant characteristics.

SAE Model Interpretations & Reporting

The typical presentation of an SAE model in an evaluation report includes: a brief description of the strengths of SAE models; a very brief, and often ambiguous, list of variable definitions; a table showing estimated coefficients, t-values, model r-squared, and perhaps total sample size; an assessment of model performance based on r-squared, "statistically significant" coefficients, and coefficients with the "right" sign, and ; a statement of what the realization rate means along with a 90% confidence interval on it.

The reporting on the model is often brief and provides few analyses or discussions such as described in this paper. Many evaluations don't provide even basic summary statistics on the variables in the model. In addition to what isn't reported, much of what is reported is unsupported in the data provided or indicative of a statistical misinterpretation. A typical example, from a peer-reviewed evaluation paper, presented the usual SAE model output table. The narrative with the model stated that the model was good because the r-squared was high, most variables were statistically significant, and all but one variable had the right sign. There are several problems with this narrative.

First, r-squared is a poor indicator of model performance, particularly for SAE models, because the dependent variable is post-program usage, not savings, and pre and post program usage are highly correlated. Therefore, the r-squared will typically be very high (>.9) even if the only explanatory variable is pre-program usage. The value of r-squared will be dominated by this underlying correlation, regardless of the quality of the model. In practice, few SAE models provide a substantial increase in this already high r-squared.

Second, if most variables in a model have t-statistics greater than 2 it doesn't mean that the model specification is correct, or that the t-statistics are correct, or that these statistically significant factors have any practical significance in understanding usage variations (particularly true when dummy variables take on almost all zero values). Many evaluators mistake statistical significance for accuracy. For example, in another recent paper the authors' stated that the "analysis produced very accurate (i.e., statistically significant) results". The problems with internal measures of uncertainty such as t-statistics was summarized quite well by

famed quality expert and statistician W.E. Deming who noted, "Statistical 'significance' by itself is not a rational basis for action." [6]. There are many threats to the validity of standard errors from SAE-type regression models (only some of which have been described in this paper) and therefore it may not be wise to rely upon them to assess uncertainty.

The third problem with the example narrative is the claim that all but one variable has the anticipated sign. A brief examination of the coefficients indicates that of the 11 "control" variables in the model which actually have an anticipated sign and supposedly had the right sign, four clearly have incorrect signs (often two variables indicating opposite responses to a question both had the same sign, e.g. floor space increased and floor space decreased were both associated with increased usage).

Unfortunately, the problems with model assessment and interpretation in this example are not uncommon. It is unusual to find a discussion which explains what the model accomplished, why it makes sense, which variables affected the results, what other specifications were tried, why the presented model was selected, whether there were any problems with outliers, the extent to which assumptions were violated, analytical choices made and their impacts on the results, etc.. Quality evaluations include these analyses and provide this level of detail because they recognize the many potential threats to validity.

Barriers to Quality Evaluations

Many DSM impact evaluations do not comply with the majority of proposals and recommendations in this paper. There are a variety of political pressures which may play a role in producing evaluations which minimize or ignore potential threats to validity. Some DSM evaluators have stated, at least in private, that evaluation problems are downplayed because their clients want firm answers, not excuses and caveats. The utilities, in turn, contend that regulators and intervenors will attack any weakness that is shown and so a balanced and open evaluation is an invitation to contentious shareholder incentive proceedings.

Given this political climate, regulators need to adopt minimum analysis and reporting requirements to change the status quo. The added cost of compliance should be modest for evaluators who follow principles of sound data analysis, because all of the proposed analyses need to be performed anyway. The only added costs are in presentation of results and more detailed explanations of the analysis process.

Conclusions

While regression models have much to offer for DSM impact evaluation, they are often used without the supporting analysis and discussion which is required to make them a reliable source for impact estimates. This paper described some of the common threats to validity caused by model specification, outliers, heteroscedasticity, and sample representativeness and the potential to be misled by standard model output was illustrated with real and synthetic examples. Minimum analysis and reporting requirements are recommended to help identify potential problems and improve the quality of impact evaluations. Similar requirements may be needed on other aspects of regression models and on other evaluation techniques not covered in this paper.

References

1. Cochran, W. G. (1983), *Planning and Analysis of Observational Studies*, New York, Wiley.
2. Box, G.E.P. (1966), "Use and Abuse of Regression", *Technometrics*, 8, pp. 625-629.
3. Mosteller, F. and J. Tukey (1977), *Data Analysis and Regression*, Reading, MA, Addison-Wesley.
4. Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York, Wiley.
5. Keating,. K., at al (1993), "Using Discrete Choice Models to Determine Free Ridership", Evaluation Exchange, 3(1), pp.9-15
6. Deming, W. Edwards (1943), *Statistical Adjustment of Data*, New York, Dover Publications.