Eliminating the Guesswork: The Information-Theoretic Approach to Model Selection

Kathryn Parlin, West Hill Energy and Computing, Inc.
Larry Haugh, University of Vermont

Abstract

This paper briefly covers the theoretical underpinning of the information-theoretic method for model selection, explores the strengths and weaknesses of this approach in comparison to other model selection techniques and provides a step-by-step description of how to implement this approach. This process is then illustrated by applying these concepts to a billing analysis conducted for a low income retrofit program, based on a pooled, cross-sectional, time series data set. Although the example is related to impact evaluation, the potential applications of this methodology extend to many other types of evaluations, including, but not limited to, market characterizations, process evaluations and attribution studies.

Introduction

Model selection is an important, but often neglected, aspect of evaluation. Specific modeling decisions can have a major impact on the results of the analysis, and variations in methodology, ranging from structure of the model to the types and content of included variables, can be nearly as numerous as the individuals performing the analysis. Without a clear and objective standard for identifying the "best" model, the researcher is left in the position of making a decision based on his or her judgment and lacks a strong foundation to support the choice of models.

The information-theoretic approach provides a framework and theoretical justification for identifying the top-ranking model and counterbalancing the competing objectives of minimizing both the bias and error variance in the model. It can be used to rank models, incorporate model selection uncertainty and establish a framework for making inferences based on multiple models.

The goal of model selection is to find the most parsimonious model, i.e., the simplest model that adequately fits the data. Underfit models, i.e., those with too few variables or model terms, will tend to produce biased estimators, whereas overfit models will lead to a lack of precision of the estimators. Akaike's Information Criterion (AIC) is gaining popularity as a tool for model selection and has been shown to perform well in balancing between these competing objectives (Burnham and Anderson 2001:35-37, Kmeta 1980, McQuarrie and Tsai 1998).

This strategy of developing a parsimonious model using the AIC creates a powerful tool for developing robust and defensible estimates of the impacts from energy efficiency programs. One favorable aspect of the AIC is that it incorporates a penalty for adding variables, creating a situation in which the improvement in fit to the model must outweigh the negatives associated with expanding the variable list.

While the purpose of moving toward an objective standard for model selection is to avoid results-based decisions, no method completely substitutes for the judgment and experience of the researcher. Rigid adherence to any set of rules can easily run afoul of basic common sense, and all results should be assessed within the context of other research in the field and the knowledge of the analyst.

The next section provides a overview of model selection methods and short description of the theory behind the information-theoretic approach to model selection. This discussion is followed by a step-by-step explanation of the model selection process. This approach is then applied to estimating impacts for a low income retrofit programs using a billing analysis. The final two sections cover the results from the example analysis and some conclusions and thoughts for further model selection efforts.

Model Selection Theory

As any statistics student is likely to know, the earliest and possibly best known model selection criterion is the adjusted R^2. It is common knowledge that the (unadjusted) R^2 will continue to increase as variables are added to the model regardless of the relative improvement in fit, providing little insight into identifying the simplest model that fits the data. The adjusted R^2 and other methods attempt to address this drawback, with only partial success (McQuarrie and Tsai, 1998).[1] This concern is particularly acute for time series, cross-sectional models used for billing analysis because the R^2 values tend to be quite high and the differences between the R^2 statistics of the candidate models can be extremely small.

Over the last forty years, research into model selection has been energetic and fruitful, with numerous competing approaches to identify the "best" model. Bayesian indicators include the Bayesian Information Criterion (BIC) and the Schwartz Information Criterion (SIC). Akaike and others have worked on alternative methods that are better suited for predictive modeling.[2] Akaike's work has been further explored and expanded by other statisticians (McQuarrie and Tsai, 1998).[3]

Within this cluttered field, the AIC has been compared to numerous other methods for model selection, and it performs very well (Burnham and Anderson 2001:35-37, Kmeta 1980, McQuarrie and Tsai 1998). In contrast, the adjusted R^2 performs poorly, almost always resulting in an overfit model (McQuarrie and Tsai 1998).

In choosing among these model selection methods, the analyst should consider both the features of the underlying model and the ease of use. The model selection method has to be appropriate for the specific analysis under discussion and suit the known characteristics of the underlying model. There are two general strategies for approaching model selection. If the analyst believes that the "true" model is so complex that it is not possible to identify and measure all of the variables that contribute to the model, the goal is to find the model that best approximates the true model by using a model selection criterion that is asymptotically *efficient*. The AIC is an asymptotically efficient (efficient) selection criterion (McQuarrie and Tsai 1998).

The second approach is to assume that the underlying model can be specified with measured variables. Under this assumption, the true model is among the set of candidate models and the purpose of the selection process is to identify the correct model. Selection methods that work under these conditions are known as *consistent*. The BIC and SIC are examples of consistent selection criteria (McQuarrie and Tsai 1998).

Energy use is quite complex and defining all of the contributing factors is elusive as best. Often, attitudes and habits affect energy use as much or more than the installation of efficiency measures. The same can be said for market forces that affect purchases of energy-related equipment. Consequently, efficient model selection criteria, such as the AIC, are more appropriate for the types of analysis typically conducted for impact evaluation or market characterization.

Ease of use is also an important consideration. If writing the code to calculate the chosen selection criteria requires a substantial investment of time and energy, it is of limited use to many analysts. The AIC has the advantage of being easily computed from the output from standard statistical packages, either from the log maximum likelihood in mixed models or from the residual sum of squares in linear regression.

Information-Theoretic Approach

[1] Examples are Mallow's Cp and Akaike's FPE.

[2] PRESS and Akaike's FPE are used for prediction.

[3] Takeuchi developed a more generalized form of the AIC, referred to as Takeuchi's Information Criterion (TIC).

The information-theoretic approach is designed to allow a group of candidate models to be compared and ranked by use of Akaike's Information Criterion (AIC). The model with the lowest value of the AIC is the one that best fits the data set, i.e., the model that minimizes the information loss. Only logistics and common sense limit the number of models that can be compared.

The AIC is calculated from the log likelihood function with an added penalty reflecting the number of parameters in the model, as shown below:

$$AIC = -2\log(\mathrm{L}(\hat{\theta}|y)) + 2K\,, \qquad (1)$$

where $\log(\mathrm{L}(\hat{\theta}|y))$ is the value of the log likelihood function at its maximum point for the vector of parameters designated by θ, given the data *y*, and K is the number of estimable parameters, including the intercept and the residual variance.[4] If the candidate models are fit by least squares regression and the outcomes are not transformed, the maximum likelihood estimate (MLE) of the residual variance can be calculated directly from the residual sum of squares (RSS/n) (Burnham and Anderson 2002).[5]

The AIC's of all models in the set of candidates can be rescaled to simplify the comparison and ranking process:

$$\Delta_i = \mathrm{AIC}_i - \min(\mathrm{AIC})\,, \qquad (2)$$

where index *i* indicates the number of the model and min(AIC) is the smallest AIC value. The relative values of Δ_i indicate the level of support for the given model. A rule of thumb is that models varying by only 1 or 2 from the best model have strong support; models with Δ_i's between 3 and 7 show less support and a value of 10 or more indicates little to no support (Burnham and Anderson 2002). However, these ground rules presume that all of the basic assumptions of linear regression are met.

The model weights reflect the probability that a given model is the best one among the set of candidate models. These weights are calculated as shown in equation (3),

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{i=1}^{R}\exp(-\frac{1}{2}\Delta_i)}\,, \qquad (3)$$

where R is the total number of models under consideration and *i* is the index for each model.

Model uncertainty can also be incorporated into the sample variances of the estimators, as shown in equation (4),

$$\mathrm{var}(\hat{\bar{\beta}}) = \sum_{i=1}^{R} w_i\left(\mathrm{var}(\hat{\beta}|g_i) + (\hat{\beta}_i - \hat{\bar{\beta}})^2\right) \qquad (4)$$

where β is the parameter of interest and $\hat{\bar{\beta}}$ is the model averaged estimator. This approach allows the researcher to reflect the uncertainty inherent in the identification of the candidate models and develop estimates based on a selected set of the candidates. Clearly, large differences in the values of the estimators of the top candidate models will result in wide variances.

There are some limitations to applying the information-theoretic approach. The candidate models must have the same number of observations and a similar structure. Models in which the dependent variable

[4] Maximum likelihood methods allow for the estimation of the parameters of interest, given a set of data and an assumed model. A brief introduction to maximum likely theory is provided in the Burham and Anderson text.

[5] The maximum likelihood estimator (MLE) is the value of the parameter for which the log likelihood function is at its maximum.

is transformed or that assume a lognormal distribution of errors (for example) cannot be compared with untransformed models (Burnham and Anderson, 2002).

Model Selection Process

The model selection process involves four steps: defining the candidate models, diagnostics to assess goodness of fit, running the models, comparing and evaluating the results. Each of these items is discussed in more detail below.

Defining the models

The first step is to establish a set of defensible candidate models appropriate to the immediate researchable question and the available data. Experience has shown that this step is the most difficult, requires a substantial time investment and can be the limiting factor in the overall success of the endeavor. The "garbage in, garbage out" rule applies to model selection as well as computers. While in theory it is possible to have a thousand or more candidate models, in reality it makes sense to limit the candidate models to a more reasonable number.

It may also make sense to divide the model selection process into two stages, as is done in the example discussed below. The first stage may be a broad brush, with candidate models that are likely to represent wide variations in model fit (such as weather-dependent effects or various error structures), with the second stage more of a fine tuning to compare combinations of variables that have a smaller impact on fit. The top ranking model in the first stage would then be used for all of the models compared in the second stage.

The advantage of the two-stage approach is that there are fewer models to run and compare at each stage. For example, if there are twenty models in the first stage and ten in the second, conducting the model selection all together as one stage would require fitting 200 models, but using a two-stage approach would necessitate thirty models. However, this strategy only works if the top-ranked model in the first stage represents a substantial improvement in fit over the alternatives and the second-stage models have a lower level of impact on the model fit.

Diagnostics

As is appropriate with any modeling project, the next step is to run the global model, i.e., the one with the most parameters, and calculate the diagnostic statistics to check for violations of assumptions. Common issues with billing data include heteroskedasticity and autocorrelation. It is also wise to check for multicollinearity among the variables. These diagnostics can help to identify serious issues with the data and allow the researcher to consider possible mitigating strategies, if needed. This process may lead to an expansion of the candidate models to incorporate a variety of error structures or other factors that may have been missed in the initial consideration of viable models.

Fitting the models

Once the list of candidate models has been completed, the researcher can begin to fit the models and compare the AIC for each model. Common statistics software packages often include the AIC in the output for the mixed models procedures. However, these values may be calculated for a different purpose and do not necessarily correctly count the number of parameters in the model. The number of parameters must include the intercept and the residual variance, in addition to the regression coefficients.[6]

Assessing the results

Once the AIC has been calculated for each model, the models can be ranked by AIC in descending order and the weights calculated for each model. These weights reflect the probability that a model minimizes the loss of information in relationship to the other candidate models, and can be used to determine whether the results from multiple models should be incorporated into the estimates through model averaging. If the top model is substantially better than any of the others or the estimates are very close for all of the top models, then model averaging is clearly unnecessary. On the other hand, if two models are close contenders and the estimates vary between the models, then model averaging is a reasonable method to incorporate information from both of the top models. The same approach can be used to decide whether to incorporate model selection uncertainty into the confidence intervals calculated for the parameters. Wide variations among the estimates and relatively equivalent weighting could result in a substantial increase in the uncertainty associated with the parameter.

As with any project, the final results should be compared to other research. Results outside of a reasonable range may indicate underlying issues with the model.

An Example

This process was applied to a billing analysis for a residential retrofit program targeted to low income Vermonters living in single family homes, served between 2001 and 2004. The purpose was to determine the annual energy savings for major measures and by end use for smaller measures. The main component of the program is an energy audit, in which an auditor visits the home, installs energy conservation devices, conducts an analysis of the home and recommends other measures to be installed at a later date. All efficiency devices are installed at no cost to the participant.

The specific interventions included efficient lighting products and water heating conservation measures (installed at the time of the audit), efficient refrigerators, fossil fuel water or space heating systems to replace electric water or space heating, waterbed insulating pads and efficient ventilation fans. Lighting retrofits included the replacement of incandescent lamps with compact fluorescent lamps (CFL's) and the installation of fluorescent hardwired fixtures. The package of hot water conservation measures may include any combination of low flow showerheads and aerators, pipe insulation, tanks wraps and thermostat turn-down.

The data structure is pooled, cross-sectional, time series (CSTS), interrupted at the time of the intervention (installation of an energy efficient device). The data come from two primary sources: program tracking and utility billing records. The program tracking data set for all homes treated during the period forms the "cross sectional" component of the data. The billing data comprise the time series component and

[6] Calculating the AIC from the output from ordinary or generalized least squares is discussed above in the theory section. The regression output typically uses n - (p+1) in the denominator, whereas n should be used for calculating the AIC. If the sample size is small enough to make a difference, this correction should be made.

come from the utility tracking systems. To properly model weather-dependent effects, climate data are also incorporated into the analysis. This information was obtained from the National Oceanographic and Atmospheric Administration (NOAA). For the Vermont program, the data set has 835 homes, 57 time periods, six to eleven explanatory variables, 21,128 total observations and a minimum of seventeen monthly observations per home.

To create a more manageable framework for model selection, the process was divided into two stages:

- an overall, macro-level assessment of the error structures and treatment of seasonal effects, and
- refining the composition of the intervention effects.

This approach dramatically reduced the total number of models to be compared without sacrificing the primary objective of model selection, i.e., to identify the best candidate model.

In the first stage, the intervention effects are held constant across all the models while various error structures and weather-dependent effects are compared. The difference in fit among the possible models is large and a single model tends to be clearly identified as the "best" model among the candidates. In contrast, the second stage compares models reflecting variations in the definitions of the intervention variables, which are relatively small in magnitude and do not affect the selection of the model with the best overall structure and treatment of seasonal effects.

The Models

A major concern arising from the pooling of CSTS data is that the resulting data set can exhibit substantial variation across the cross-sectional units and/or time periods, making it difficult to estimate the treatment effects. The possible structures for the model include the following:

- The intercept and slope regression coefficients are constant over all cross-sectional units and time periods.
- The intercept varies by cross-sectional unit and/or time period, but the slope coefficients remain constant.
- Both the intercepts and slope coefficients vary by cross-sectional unit and/or time period.

These concepts can be applied to the billing analysis for impact evaluation. The homes are the cross-sectional units and the read periods for the electric bills are the time periods. Average energy usage varies widely from one house to the next due to appliance holdings, lifestyle, size of the home, and many other factors. There may also be time effects, in that specific events (such as a cold snap) or seasonal variations (such as the length of daylight) may trigger a wide spread reaction across all homes. If not otherwise addressed, these variations are likely to overwhelm any efforts to tease out the savings from energy conservation measures. Expected savings, reflected in the coefficients of the treatment variables, may also vary from one house to the next. For example, installing a low flow showerhead in a home with a single occupant is likely to save much less than the same installation in a household of four. Thus, it is entirely feasible to consider whether both the intercepts and treatment effects could vary by house.

The second option, with varying intercepts and constant slope coefficients, is the most reasonable for the Vermont data set. Given the wide range of appliance holdings, number of occupants, lifestyle choices and house characteristics, the house-specific intercepts are virtually assured to vary from one home to the next. Although the case for the slope coefficients is not as clear, the hypothesis of equal slope coefficients was tested by comparing the restricted and unrestricted models (Dielman, 1989), leading to the conclusion that the slope coefficients did not vary significantly by home.

Models with varying intercepts and constant slope coefficients are referred to as "error component" models in this paper, stemming from the conceptualization of the variations between cross-sectional units as

a component of the error (Maddala, 1971). The general form of the error components model for a billing analysis is given in Equation 5.

$$C_{it} = \alpha_i + \tau_t + \sum_{j=1}^{p} x_{ijt}\beta_j + \sum_{k=1}^{q} z_{ikt}\gamma_k + \varepsilon_{it} \quad , \qquad (5)$$

where

C_{it} is the monthly consumption for the household i in period t, expressed in annualized kWh per day,

α_i is the "customer-specific" intercept for household *i*, accounting for unexplained differences in use between households associated with the number of occupants, appliance holdings and lifestyle,

τ_t is the "time-specific" effect (or error) for period *t*, reflecting the unexplained differences in use between time periods,

x_{ijt} are the predictor variables reflecting the installation of energy efficiency measure j for household i in period t,

β_j are the slope coefficients that quantify the average influence of modeled efficiency measure j on monthly consumption,

p is the total number of variables related to energy efficiency measures included in the model,

z_{it} are the predictor variables reflecting non-program related effect k (such as weather impacts) for household i in period t,

γ_k represents the slope coefficients that quantify the average influence of modeled non-program related effect (such as weather) *k* on monthly consumption,

q is the total number of non-program related effects included in the model, and

ε_{it} is the error term that accounts for the difference between the model estimate and actual consumption for household *i* in period *t*.

Types of Error Component Models. There are two common strategies for modeling CSTS data where the intercepts are likely to vary across cross-sectional units or time periods, or both, i.e., fixed effects and random effects models. The fixed effects model retains the assumption that the program effects are the same for all homes, but allows for a fixed term for each cross-sectional unit, representing the variation unique to that home, conditional on the data set, commonly referred to as the "customer-specific" intercept. The same type of strategy can be used to account for variations across time periods.

This model is a popular choice for modeling energy savings (Medgal et. al. 1995, Solberg et. al. 2003, TecMarket Works et. al. 2004). The rationale is that each participant acts as their own control, netting out the unknown impacts of lifestyle, house size, occupancy patterns and appliance holdings. It allows us to use program and billing data for all (or most) participants without incurring the additional costs (and related reduction in sample size) associated with conducting surveys to obtain detailed information on occupancy patterns and other non-program related factors that affect energy use.

The random effects model has the same structure as the fixed effects model, with one exception, i.e., the α_i and τ_t in Equation 5 are assumed to vary randomly. This approach has also been used for impact evaluation (Sumi and Oblander 1993). By allowing the house and time intercepts to vary randomly, there is no substantial increase in the number of parameters. However, it is necessary to make an assumption about the distribution of the intercepts, typically that the intercepts are normally distributed with a mean of zero and a constant but unknown variance.

The fixed and random effects models should produce the same slope estimators, unless there is correlation between the intercepts and the explanatory variables (Dielman 1989, Mundlak 1978, Hausman 1978). In the presence of such correlation, the slope coefficients estimated by the random effects model will

be biased, but the fixed effects model will produce the best linear unbiased estimator (Dielman 1989:76). Hausman (1978) proposed a test statistic to determine whether the explanatory variables are correlated with the intercepts.

In the context of billing analysis, this type of correlation relates to the connection between the initial level of use and the potential savings from energy conservation measures. Researchers have shown that homes with lower initial use also tend to have lower savings (Blasnik, 2002). The Hausman test using the Vermont data set indicated that correlation between the intercepts and explanatory variables cannot be ruled out.

Fixed vs. Random Effects. There are a number of considerations to take into account when evaluating the applicability of the fixed and random effects models. The relative merits of the two approaches are summarized in Table 1 below.

Table 1: Fixed v Random Effects Models

Fixed Effects	Random Effects
Conditional on data set	Assume distribution of intercepts
Increases number of model parameters	Fewer parameters
Unbiased coefficients	Coefficients biased if correlation between explanatory variables and intercepts
Better for smaller samples	Need larger sample size (T-K>10)[7]

Weather-Dependent Use. The modeling of the seasonal and weather-dependent effects is also a critical component of estimating energy savings and requires careful consideration. For many homes, there are few if any weather-dependent electric end uses. Electric space heat is rare in Vermont, and air conditioning is not common in older homes. During winter months, furnaces will use some electricity for the burner and distribution system, whereas boilers generally draw only a minimal amount of electricity. Electric use tends to increase during the winter for non-weather-dependent reasons as well, such as higher lighting use during the shorter days in winter, and this seasonal increase may appear to be similar to weather-dependent use.

Separating out weather-dependent effects from other variations in use is critical to developing realistic estimates of energy savings. The lack of detailed, house-specific information often necessitates identifying weather-dependent usage patterns from the consumption data. Modeling weather-dependent use generally takes one of two forms:

- adding a dummy variable interacting with heating or cooling degree days to reflect weather-dependent impacts from groups of homes with the same characteristics and
- estimating weather-dependent effects for each home.

The second method can greatly increase the number of variables in the model if the sample includes a large number of homes, but it also accounts for the wide fluctuations in the consumption patterns from one home to the next.

Diagnostic Issues. Pooled CSTS data creates additional sources of variability. The underlying assumption behind pooling is that the cross-sectional units are homogenous. In real applications, this is rarely the case. Energy use in homes varies widely, as does the impact of the conservation treatments.

In CSTS data sets, variation among the cross-sectional units may contribute to heteroskedasticity and the series of observations within each house may well be autocorrelated. Multicollinearity among the

[7] T is the number of observations (time periods) available per house and K is the number of parameters in the model.

explanatory variables can also contribute to the uncertainty in the estimated intervention effects, sometimes resulting in estimators of the opposite sign from what would be expected (Sayrs, 1989).

Goodness of fit testing was conducted with the global model for Phase I, i.e., the model with the most parameters. The results of this analysis show some potentially serious departures from the standard OLS assumptions. The data set shows signs of autocorrelation, heterskedasticity and heavy tails. Multicollinearity was not a problem.

These issues indicate that the modeling should take into account methods to mitigate the effects of these departures from the standard assumptions.

Questions. A number of questions arise in the process of trying to determine the "best" model for billing analysis designed to estimate energy savings.

- Do the benefits, i.e., improved fit, of the fixed effects model outweigh the loss of degrees of freedom?
- Is the variation over the time periods sufficiently important to warrant their inclusion in the model?
- Does explicitly modeling the within-house autocorrelation improve the model fit?
- Is estimating seasonal effects by home worth the large increase in the number of parameters?
- Does addressing heteroskedasticity have a beneficial impact on the model fit?

Candidate Models. The range of models included in Phase I addressed a variety of structural issues and seasonal effects. The random and fixed effects models were compared, although the initial diagnostics suggest that the fixed effects model may be more appropriate. House effects and time effects were considered separately, so that a model could include fixed house effects and random time effects. A first-order autoregressive error was explicitly modeled to address within-house autocorrelation and weighted least squares was added to mitigate the effects of heteroskedasticity (Judge 1980). Phase I included 36 models, as shown in Table 2 below.

Table 2: Phase I Candidate Models

	Fixed House Effect	Random House Effect
Time (3)	Fixed Random No time effect	Fixed Random No time effect
Weather-dependent Effects (2)	By house Aggregate variables	By house Aggregate variables
Within-house Error Structures (3)	Independent Weighted least squares First-order autoregressive	Independent Weighted least squares First-order autoregressive
Total number of Models	18	18

The second stage of modeling involved refining the intervention effects. The hot water conservation measures were bundled into a group, although each home typically received a subset of the five possible measures (low flow showerheads and aerators, tank wrap, pipe insulation and temperature turn down). The initial modeling approach was to use a single dummy variable reflecting the installation of any part of the DHW conservation package. The second strategy was to define a scaled variable that captured the relative weights of the specific conservation measures installed. These weights were developed based on the prescriptive savings values used for the measures. A third set of models were defined by dividing the homes into low and high use, according to whether initial electric use was above or below the median, and modeling the interaction between the low/high use and dummy or scaled measure variables for the hot water conservation measures.

The options for modeling efficient lighting included specifying a single variable for the installation of any type of fixture, separate variables for external and internal fixtures, a single variable holding the number of lamps installed and setting up a separate variables for homes with six or more lamp replacements as compared to those with fewer, on the hypothesis that a high number of lamp replacements may translate to lower savings per lamp.

Results

The results from Phase I of this analysis support the conclusion of other evaluators that the fixed effects model is the best strategy for this type of modeling. In this part of the analysis, the top ranked model stands well apart from the next best option (by a difference of over 1,000 in AIC), suggesting that choosing fixed effects has a large impact on the fit of the model.

Table 3: Phase I Ranking

Rank	Model Description	# of Para-meters	AIC	Δ_i
1	FE/WLS/Weather by House/Fixed Time	1,907	341,700	0
2	FE/WLS/Weather by House/Time Period Out	1,851	342,851	1,151
3	FE/WLS/Weather by House/ Time Random	1,852	345,410	3,710
4	FE/AR/Weather by House/Fixed	1,908	352,526	10,826
5	FE/AR/Weather by House/Random	1,853	353,221	11,521
6	FE/AR/Weather by House/Time Period Out	1,852	353,441	11,741
7	FE/WLS/Weather Aggregated/Fixed Time	903	356,739	15,039
8	RE/WLS/Weather by House/Random Time	1,018	356,964	15,264

FE=fixed effects, RE=random effects; AR=first-order autoregressive, I=independent, WLS=weighted least squares; Δ_i =the difference between the AIC of model i and the model with the lowest AIC.

The findings from this part of the analysis can be summarized as follows:

- While the random effects model tremendously reduces the degrees of freedom, the improvement in fit from the fixed effects model easily overcomes the impact of the increase in parameters.
- Estimating the weather-related effects with separate slopes for each home has a large and unmistakably positive contribution to improving the fit.
- Incorporating weighted least squares results in a major reduction in the AIC. To a lesser degree, explicitly modeling the autocorrelated error structure also has a beneficial impact on the model fit.
- The impact of including the time period as a fixed or random effect is smaller, but still important.

The results further indicate that conducting the initial modeling with weighted least squares, independent errors or autoregressive errors would result in the same decision regarding the fixed effects model and weather-dependent and time variables.

The model selection process identifies the best model, but does not explain the implications of choosing one model over another. Table 4 below summarizes the estimates of the main variables and related

standard errors for models with various error structures. In comparison to a model assuming independent errors and no heteroskedasticity, one would expect that correcting for autoregressive errors would result in estimates with higher standard errors and models with weighted least square would produce estimates with lower standard errors, but the estimates themselves would remain fairly constant in either case.

For refrigerator replacements and hot water fuel switching, which have substantial savings associated with them, the estimates are reasonably consistent. However, this is not the case for the lighting and the hot water conservation measures for low use homes, which tend to have smaller savings. The wide variations in estimates for lighting suggest that these savings may be too small to separate from the noise in the billing analysis, at least without a substantially larger sample size. The hot water conservation results indicate that homes with electric hot water may be sufficiently different from the other homes in the model to create instability in these estimates for low use homes.

Table 4: Comparison of Fixed Effects Models with Different Error Structures*

	Autoregressive Errors		Independent Errors		Weighted Least Squares	
	Savings (kWh)	Std Error	Savings (kWh)	Std Error	Savings (kWh)	Std Error
CFL Lamps	33	9.34	42	5.98	9	3.98
CFL Fixtures	44	20.17	60	12.58	42	6.21
Refrigerator Replacement	758	71.03	751	44.61	681	20.02
DHW Conservation High Use	335	99.43	333	101.22	402	74.58
DHW Conservation Low Use	-115	107.83	-96	124.18	100	47.62
DHW Fuel Switch High Use	4,182	165.43	4,189	102.06	3,783	94.90
DHW Fuel Switch Low Use	1,894	686.97	1,864	411.57	1,967	95.29

* Weather-dependent use was modeled by house for all three scenarios.

The AIC's of the Phase II models were much closer to each other than found in Phase I, as can be seen in Table 5. The primary differences between the top three models are the modeling of the lighting variables (lamps and fixtures). While the rule of thumb is that a difference in AIC of 10 or more is very strong support for the top model, the actual results from the top model were counterintuitive and turned out to be less than useful.

Table 5: Phase II Results

Rank	Model Description	Δ_i	w_i
1	2 lamp variables, interior/exterior fixtures, DHW hi/low dummy	-	0.999
2	2 lamp variables, combined fixture variable, DHW hi/low dummy	14	0.001
3	1 lamp variable, interior/exterior fixtures, DHW hi/low dummy	24	0.000
4	2 lamp variables, interior/exterior fixtures, DHW hi/low scaled	31	0.000
9	*2 lamp variables, interior/exterior fixtures, DHW combined scaled*	*241*	*0.000*

The top-ranked model has the most variables and included four variables for lighting measures, two for lamps (one for homes with less than six and another for homes with more than six) and two for fixtures (interior and exterior). The savings for lamps in homes with fewer than six bulbs installed and for exterior fixtures were negative, but the coefficients were not significant. However, significant and positive savings were found for lamps installed in homes with more than six products and for interior fixtures. In short, these results suggest that participants who received fewer than six lamps achieved no savings and homes with six or more saved 38 kWh per lamp, a finding that is difficult to reconcile with information from other studies.

The modeling option consistently ranked higher by a wide margin in Phase II was the low and high use differentiation of homes with hot water conservation measures. The ninth-ranked model was the first one with a combined DHW variable, and its AIC is 241 higher than the top-ranked model. Scaling the variable to try to account for the level of expected savings did not outperform the simple dummy variable, as can be seen by comparing Models 1 and 4, which are identical except for the DHW conservation variable.

Other than supporting the differentiation of DHW savings by high and low users, the fine tuning strategy attempted in Phase II was not particularly effective. The Phase I outcomes suggest that the lighting estimates were highly variable, and consequently it was not surprising that attempting to add more lighting variables in Phase II failed to yield a useful end product. Given the available sample size, the statistical model could not overcome the overwhelming combination of the small size of the expected savings and the wide variations in monthly electric use.

For this analysis, all of the candidate models were defined prior to beginning the analysis. In retrospect, it would have been worthwhile to review the results of Phase I and make modifications to the Phase II models at that point.

Conclusions

Model selection is a critical aspect of impact evaluation. Without a clear and objective standard for identifying the "best" model, the researcher is left in the position of making a decision based on his or her frame of reference and lacks a strong foundation to support the choice of models to other stakeholders. The information-theoretic approach using the AIC as a summary statistic provides the framework and theoretical justification for identifying the top-ranking model and counterbalancing the competing objectives of minimizing both the bias and variance in the model.

This approach was successfully applied to a low income retrofit program in Vermont. These results suggest some effective modeling options and provide a starting point for future evaluation efforts of this type. The model selection process leads to the conclusion that the fixed effects model is the best choice for the data set and incorporating weighted least squares to address heteroskedasticity is an important component to developing a strong and defensible model. House-by-house estimates of weather-dependent use also have a dramatic improvement on model fit.

The implementation of Phase II, i.e., the fine tuning of the measure-level variables, further illustrates a potential pitfall of rigid adherence to the established process. Since all of the candidate models in this analysis were defined prior to commencing the analysis, the results from Phase I were not thoroughly considered before proceeding with the second phase, and consequently the top-ranked model from Phase II included additional refining of the lighting variables. A more reasonable interpretation of the Phase I output is that the savings from lighting measures are too small to be estimated by a model of this type with the sample size available for this analysis and no further parsing of the lighting variables would be worthwhile.

The Phase II component of the analysis has one useful outcome, in that is provided support for the strategy of dividing the participants into low and high users for the purposes of estimating savings from hot water conservation measures. This strategy substantially improved the model fit and allowed for the estimation of savings from the DHW conservation package for high use homes.

The information-theoretic approach to model selection shows substantial promise for moving toward a more objective and defensible method to identifying the best model out of a field of candidate models. While the example given here is heavily oriented toward error structures for pooled, CSTS billing data, the process works equally well for defining the best set of variables in any type of regression analysis. Thus, the method can be applied to a wide range of evaluation tasks, ranging from impact evaluation to market characterization and process issues.

References

Blasnik, Michael. 2004. Ohio Electric Partnership program impact evaluation, final report prepared for the Ohio Office of Energy Efficiency.

Burnham, Kenneth and David Anderson. 2002. *Model selection and multimodel inference*. New York: Springer-Verlag.

Dielman, Terry E. 1989. Pooled cross-sectional and time series data analysis. New York: Marcel Dekker.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica*, 46(6): 1251-1271.

Judge, George G., William E. Griffiths, R. Carter Hill and Tsoung-Chao Lee. 1980. *The theory and practice of econometrics*. New York: John Wiley and Sons.

Kmenta, Jan, and James B. Ramsey. 1980. *Evaluation of econometric models*. New York: Academic Press.

Maddala, G.S. 1971. The use of variance components models in pooling cross section and time series data. *Econometrica*, 39(2): 341-358.

McQuarrie, Allan D. R., Tsai, C.L. 1998. *Regression and time series model selection*. Singapore: World Scientific.

Medgal, Lori, E. Paquette, and J. Greer. 1995. The importance of using analysis of covariance, diagnostics, and corrections with billing analysis for large C&I customers. In *Energy program evaluation: uses, methods, and results; Proceedings of the 1995 international energy program evaluation conference held in Chicago, IL, August 22-25, 1995.* Madison, Wisconsin: OmniPress.

Mundlak, Yair. 1978. On the pooling of time series and cross section data. *Econometrica,* 46(1): 69-85.

Parlin, K., Al Bartsch, and Scott Pigg. 2005. Facing uncertainty from within and without: An impact evaluation of the California low income energy efficiency program. In *Proceedings of the 2005 international energy program evaluation conference, held in New York City August 17-19, 2005.* Madison, Wisconsin: OmniPress.

Sayrs, Lois W. 1989. *Pooled time series analysis*. Newbury Park, California: Sage Publications.

Solberg, Eric, A.M. Gill, and A.Y. Ahmed. 2003. Measurement of DSM program savings: Comparing estimates from treatment-effects and fixed-effects models. In *Proceedings of the 2003 international energy program evaluation conference, held in Seattle August 20-22, 2003***.** Madison, Wisconsin: OmniPress.

Sumi, David and Oblander, Paul. 1993. A comparison of model specifications in a billing data analysis of impacts from a commercial and industrial rebate program. In *Proceedings of the 1993 international energy program evaluation conference held in Chicago August 25-27, 1993***.** Madison, Wisconsin: OmniPress.

TecMarket Works, et. al. 2004. *The California evaluation framework, project number K2033910*. Prepared for the California Public Utilities Commission and the Project Advisory Group.

West Hill Energy and Computing, Ridge and Associates, Energy Center of Wisconsin, Wirtshafter Associates and Business and Economic Analysis and Research. 2005. Impact evaluation of the 2002 California low income energy efficiency program, final report prepared for Southern California Edison Company, Pacific Gas and Electric Company, San Diego Gas and Electric Company and Southern California Gas Company.