

# Regression for M&V: Reference Guide

May 2012



## **Regression for M&V: Reference Guide**

**Version 1.1**

**May 2012**

**Prepared for**

**Bonneville Power Administration**

**Prepared by**

**Research Into Action, Inc.**

**Quantum Energy Services & Technologies, Inc. (QuEST)**

**Stetz Consulting, LLC**

**Kolderup Consulting**

**Warren Energy Engineering, LLC**

**Left Fork Energy, Inc.**

**Schiller Consulting, Inc.**

**Contract Number 00044680**

# Table of Contents

- 1. Introduction.....1
  - 1.1. Purpose ..... 1
  - 1.2. Background ..... 1
- 2. Background .....3
  - 2.1. Description ..... 3
  - 2.2. Regression Applicability ..... 4
  - 2.3. Advantages of Regression ..... 5
  - 2.4. Disadvantages of Regression ..... 5
- 3. The Regression Process .....7
  - 3.1. Step 1 - Identify All Independent Variables ..... 7
  - 3.2. Step 2 - Collect Data ..... 8
  - 3.3. Step 3 - Synchronize the Data ..... 8
  - 3.4. Step 4 - Graph the Data ..... 8
  - 3.5. Step 5 - Select and Develop Model..... 8
  - 3.6. Step 6 - Validate Regression Model..... 9
  - 3.7. Requirements for Regression..... 9
- 4. Models..... 13
  - 4.1. One Parameter Model (Mean Model)..... 13
  - 4.2. Two Parameter Model (Simple Regression) ..... 13
  - 4.3. Simple Regression Change Point Models ..... 14
    - 4.3.1. Three-Parameter Change Point Model..... 15
    - 4.3.2. Four-Parameter Change Point Model..... 16
    - 4.3.3. Five-Parameter Change Point Model ..... 17
  - 4.4. Multiple Regression..... 17
    - 4.4.1. Categorical Variables .....18
    - 4.4.2. Multiple Regression Change Point Models..... 19
  - 4.5. Uncertainty and Confidence Intervals ..... 20
    - 4.5.1. Uncertainty ..... 20
    - 4.5.2. Confidence Level and Confidence Interval ..... 21
    - 4.5.3. Prediction Interval ..... 22
    - 4.5.4. Confidence Levels and Savings Estimates ..... 23

5. Validating Models ..... 25

    5.1. Statistical Tests and Measures for the Model ..... 25

        5.1.1. R-Squared (Coefficient of Determination) ..... 25

        5.1.2. Adjusted R-Squared ..... 25

        5.1.3. Degrees of Freedom ..... 26

        5.1.4. Root Mean Squared Error (Standard Error of the Estimate)..... 26

        5.1.5. Coefficient of Variation of the Root Mean Squared Error ..... 26

        5.1.6. Bias ..... 26

        5.1.7. F-Statistic ..... 28

    5.2. Statistical Tests and Measures for the Model’s Coefficients ..... 28

        5.2.1. Standard Error of the Coefficient (Intercept or Slope) ..... 28

        5.2.2. t-Statistic ..... 28

        5.2.3. p-value ..... 28

    5.3. Tables of Statistical Measures ..... 29

    5.4. Other Tests of Model Validity ..... 31

        5.4.1. Check for Autocollinearity..... 31

        5.4.2. Check for Multicollinearity ..... 32

    5.5. Analysis of Residuals ..... 33

6. Example ..... 35

    6.1. Use of Monthly Billing Data in a 2-Parameter Model to Evaluate Whether It  
    Will Make a Satisfactory Baseline ..... 35

    6.2. Background on Heating and Cooling Degree-Days (HDD and CDD)..... 41

7. Minimum Reporting Requirements..... 43

8. References and Resources..... 45

Appendix: Glossary of Statistical Terms ..... 47

    Sources: ..... 52

# 1. Introduction

## 1.1. Purpose

This document presents a *Regression for M&V: Reference Guide*<sup>1</sup> as a complement to the Measurement and Verification (M&V) protocols used by the Bonneville Power Administration (BPA). The *Regression Reference Guide* assists the engineer in conducting regression analysis to control for the effects of changing conditions (i.e., weather) on energy consumption.

This document is one of many produced by BPA to direct M&V activities. The *Measurement and Verification (M&V) Protocol Selection Guide and Example M&V Plan* provides the region with an overview of all of BPA's M&V protocols, application guides, and reference guides, and gives direction as to the appropriate document for a given energy efficiency project. The document *Glossary for M&V: Reference Guide* defines terms used in the collection of BPA M&V protocols and guides. In addition, an appendix to this *Regression Reference Guide* provides a glossary specific to this guide.

Chapter 8 of this guide provides full citations (and web locations, where applicable) of documents referenced.

## 1.2. Background

In 2009, BPA contracted with a team led by Research Into Action, Inc. to assist the organization in revising the M&V protocols it uses to assure energy savings for the custom projects it accepts from its customer utilities. The team has conducted two phases of research and protocol development under the contract, Number 00044680.

In the first phase, Research Into Action directed a team comprised of:

- Quantum Energy Services & Technologies, Inc. (QuEST), led by David Jump, Ph.D., PE and assisted by William E. Koran, PE;
- Left Fork Energy, Inc., the firm of Dakers Gowans, PE;
- Warren Energy Engineering, LLC, the firm of Kevin Warren, PE;
- Schiller Consulting, Inc., the firm of Steven Schiller, PE; and
- Stetz Consulting, LLC, the firm of Mark Stetz, PE.

In the second phase, Research Into Action directed a team comprised of:

- David Jump, Ph.D., PE, William E. Koran, PE, and David Zankowsky of QuEST;

---

<sup>1</sup> Hereinafter, *Regression Reference Guide*.

- Mark Stetz, PE, CMVP, of Stetz Consulting;
- Erik Kolderup, PE, LEED AP, of Kolderup Consulting; and
- Kevin Warren, PE, of Warren Energy Engineering.

The Research Into Action team was led by Jane S. Peters, Ph.D., and Marjorie McRae, Ph.D. Assisting Drs. Peters and McRae were Robert Scholl, Joe Van Clock, Mersiha Spahic, Anna Kim, Alexandra Dunn, Ph.D., and Kathleen Gygi, Ph.D.

For BPA, Todd Amundson, PE, directed the M&V protocol research and development activities. Mr. Amundson was working under the direction of Ryan Fedie, PE, and was assisted by BPA engineers. Mr. Amundson coordinated this work with protocol development work undertaken by the Regional Technical Forum. In addition, Mr. Amundson obtained feedback from regional stakeholders.

William Koran is the primary author of this *Regression for M&V: Reference Guide*; team members reviewed and provided guidance. We thank Andie Baker, Ph.D., Senior Conservation Evaluator for Tacoma Power, for her thoughtful comments on the draft document.

## 2. Background

### 2.1. Description

Regression is a statistical technique that estimates the dependence of a variable of interest (such as energy consumption) on one or more independent variables, such as ambient temperature. It can be used to estimate the effects on the dependent variable of a given independent variable while controlling for the influence of other variables at the same time. It is a powerful and flexible technique that can be used in a variety of ways when measuring and verifying the impact of energy efficiency projects.

These guidelines are intended to provide energy engineers and M&V practitioners with a basic understanding of the relevant statistical measures and assumptions necessary to use regression analysis properly. The guidelines should be followed whenever the technique is required. While this is not a comprehensive guide to regression, following the approaches described here should make most M&V regressions valid for their intended purpose. Please refer to a textbook for more comprehensive information.

Additional information on regression analysis is available from many sources. Resources that may be valuable references for energy efficiency M&V practitioners include the following:

- *IPMVP: International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings, Volume 1*
- *ASHRAE Guideline 14-2002 – Measurement of Energy and Demand Savings*
- *California Commissioning Collaborative’s Guidelines for Verifying Existing Building Commissioning Project Savings, Using Interval Data Energy Models: IPMVP Options B and C*

*IPMVP Appendix B, Uncertainty*, and *ASHRAE Guideline 14 Annex B, Determination of Savings Uncertainty*, and *Annex D, Regression Techniques*, have information very relevant to regression analysis. *Guideline 14* is scheduled to be updated in 2011.

The *Guidelines for Verifying Existing Building Commissioning Project Savings* is a relatively easy-to-read document that focuses on regression methods. Although written with a focus on commissioning of existing buildings, the methods described are applicable to a variety of projects.

In addition to these documents, a general reference for exploratory data analysis and statistical inference, the *NIST/SEMATECH Engineering Statistics Handbook*, is available online from the National Institute of Standards and Technology. The *Engineering Statistics Handbook* site includes a detailed table of contents for the web-based handbook, and also includes downloadable PDF files for off-line reading.

## 2.2. Regression Applicability

Regression estimation is applicable when the energy use affected by the efficiency measure is proportional to one or more independent variables. Note that the technique of energy indexing is a simple application of the regression guide that can be used when energy use is linearly proportional to one normalizing variable. There are other constraints upon using energy indexing in lieu of a more generalized approach. Please refer to BPA's *Verification by Energy Use Indexing Protocol* for further information on this technique.

In M&V, energy usage is typically (and optimally) the dependent variable, whether energy usage is measured monthly through bills or measured more frequently through meter monitoring. The regression model attempts to predict the value of the dependent variable based on the values of independent, or explanatory, variables such as weather data.

- ➔ **Dependent Variable** – the outcome or endogenous variable; the variable described by the model; for M&V, the dependent variable is typically energy use
- ➔ **Independent Variable** – an explanatory or exogenous variable; a variable whose variation explains variation in the outcome variable; for M&V, weather characteristics are often among the independent variables
- ➔ **Simple Regression** – a regression with a single independent variable
- ➔ **Multiple Regression** – a regression with two or more independent variables

One of the most common applications of regression in M&V is when the primary source of data is monthly utility consumption. The initial step is to establish the baseline dependence of building usage on weather conditions by modeling the period prior to the retrofit that is illustrative of pre-retrofit usage – the baseline period. Then, post-retrofit weather is applied to the baseline model in order to estimate the energy use of the building had the energy efficiency improvements not been made (the *counterfactual situation*). In M&V, this projection of the baseline energy use into the post period is typically called the *adjusted baseline*. Finally, the adjusted baseline (predicted counterfactual energy use) is compared to the actual post-retrofit energy use and the difference provides an estimate of energy savings.<sup>2</sup>

Regression techniques can be applied to data with a much smaller time interval than a monthly billing period, such as hourly or daily data. This is useful when a simple spot measurement is not adequate to establish the baseline. These smaller interval data are frequently applicable to *IPMVP Options A (Key Parameter Measurement)*, *B (All Parameter Measurement)*, and *C (Whole Facility)*, and can also be used to assist in model calibration for *IPMVP Option D (Calibrated Simulation)*.

---

<sup>2</sup> Note that this is the general approach followed by most M&V practitioners to estimate energy savings. Economists, who typically conduct impact evaluations, typically estimate a single model from both baseline and post-retrofit data, and use a dummy (*categorical*) variable applied to post-retrofit observations to estimate energy use savings. The resulting savings estimates are comparable, although not necessarily identical.

## 2.3. Advantages of Regression

Regression is a very flexible technique that can be used in conjunction with other M&V methods to help provide a deeper understanding of how and when energy is used. Regression can also be used to extrapolate short-term measurements to annual energy. The ideal case for regression is when the measurement period captures the full annual variation in the dependent and independent variables – that is, the full range of operation conditions. If the relationship between the independent and dependent variables is not expected to change over the range of operating conditions, then short-term measurements can be extrapolated to annual energy use, even if the measurement period does not capture the full annual variation.

A particular advantage of regression is that it not only facilitates an estimate of energy savings, but it also can provide an estimate of the uncertainty in savings calculations. Further, a baseline regression model can be used to estimate how much data is required in the post-retrofit period to keep savings uncertainty below a desired threshold.

Regression is conceptually simple, most M&V practitioners have at least a basic familiarity with it, and usage and weather data – the variables typically needed for a basic model – are usually readily available.

## 2.4. Disadvantages of Regression

Although simple in concept, proper use of regression requires a clear understanding of statistical methods and application guidance, which this document seeks to provide to the M&V practitioner. The information in this guide should cover the great majority of M&V projects, but situations can occur that require a more detailed understanding of statistical methods. While the basic technique is fairly straightforward, complications to the site or the data can easily require more advanced techniques and a more thorough understanding of regression methods.

Regression models require multiple observations on the dependent and independent (*explanatory*) variables. There are times, however, when explanatory variables are not readily available or we only have access to proxies. Explanatory variables that are not included in the regression model often introduce added error. If energy use is not a strong function of the independent variable(s) in the equation, or if there is large variability in energy use (“scatter” in the  $x$ - $y$  chart) relative to strength of the predictive relationship, regression analysis generates estimates that have high uncertainty.

It is important to note that regression is often performed without an estimate of the degree of uncertainty involved, so the validity of the resulting savings estimates is unknown. Currently, there is no robust means of estimating the uncertainty introduced when extrapolating short-term data to an annual savings estimate. The *ASHRAE Research Project RP-1404, Measuring, Modeling, Analysis and Reporting Protocols for Short-Term M&V of Whole Building Energy Performance*, due for completion in early 2012, will attempt to address this shortcoming for whole building methods.



## 3. The Regression Process

The regression process can be summarized in six steps:

1. Identify all independent variables to be included in the regression model
2. Collect data
3. Synchronize data into appropriate time intervals (if necessary)
4. Graph the data
5. Select and develop the regression model
6. Validate the model

### 3.1. Step 1 - Identify All Independent Variables

To properly identify all independent variables, you should consider the facility and how different factors play into its energy use. Then, you will compile a list of the variables that are likely to have an impact on the energy use of the facility or system being modeled. When variable values are not numeric or are not continuous, the data can be separated into several regression models, rather than including all variables within a single model.

Developing separate models is just one approach to working with categorical variables, an approach favored by many M&V practitioners. One can also use *binary* variables to indicate the presence or absence of a given condition (that is, to create a category) and apply these binary variables to develop estimates of either the slope or the intercept, or both, when the given condition is satisfied. (See Section 4.4.1 for a discussion of the use of categorical variables.)

We advise caution when including many variables. A model should only use the variables that explain the relationship and not include additional, extraneous information. *ASHRAE Guideline 14, Appendix D*, provides additional information on regression estimation with two or more independent variables (*multiple regression*).

Some independent variables commonly used in energy regressions are:

- ➔ Ambient dry bulb temperature (actual or averaged over a time-period such as a day)
- ➔ Heating degree-days (HDD: See Section 6.2)
- ➔ Cooling degree-days (CDD: See Section 6.2)
- ➔ Plant output (number of widgets produced in some period)
- ➔ Number of occupants in a facility each hour

## 3.2. Step 2 - Collect Data

Prior to installation of the measure, identify and collect data for a monitoring period that is representative of the facility, operation, or equipment. This is the *baseline* period, sometimes referred to as the *tuning or pre* period. The baseline monitoring period should be long enough to represent the full range of operating conditions. For example, when using monthly data for a weather-sensitive measure, the baseline period typically includes 12 or 24 months of billing data, or several weeks of meter data. Using a partial year may overemphasize portions of the year and add variability to your model.

It is vital that the collected baseline data accurately represent the operation of the system before improvements were made. Anomalies in these data can have a large effect on the outcome of the study. Examine data outliers – data points that do not conform to the typical distribution – and seek an explanation for their occurrence. Atypical events that result in outliers include equipment failure, any situations resulting in abnormal closures of the facility, and a malfunctioning of the metering equipment. Truly anomalous data should be removed from the data set, as they do not describe the operations prior to the installation of the measure.<sup>3</sup>

## 3.3. Step 3 - Synchronize the Data

To accurately represent each independent variable, the intervals of observation must be consistent across all variables. For example, a regression model using monthly utility bills as the outcome variable requires that all other variables originally collected as hourly, daily, or weekly data be converted into monthly data points. In such a case, it is common practice to average points of daily data over the course of a month, yielding synchronized monthly data. There are problems with this approach because varying data lengths can cause net bias in the model. *Net bias* means that the total predicted energy use over the period being analyzed will differ from the actual energy use over that period.

## 3.4. Step 4 - Graph the Data

Create one or more scatter plots to begin to visualize the relationships between the dependent and independent variables. Most commonly, one graphs the independent variables on the X axis and the dependent variable on the Y axis. Figure 4-1 illustrates a scatter plot for the linear relationship between electrical demand and ambient temperature.

## 3.5. Step 5 - Select and Develop Model

To create a baseline equation, perform a regression analysis on the measured variables. The analysis is typically called an *ordinary least squares regression*, because the mathematics

---

<sup>3</sup> Again, the approach typically used by engineers and by economists diverges. Economists typically collect and clean both the baseline and the post-installation data as part of Step 2 and conduct the subsequent steps on the entire pre- and post-period.

generates a model that minimizes the sum of squared deviations between the actual and predicted values.

The equation calculated from the regression analysis represents the baseline relationship between the variables of interest. Figure 4-1 shows the data and the model estimated for the value of the outcome variable as a function of one independent variable – a *simple regression*.

Frequently, however, more than one independent variable influences the outcome variable. For example, the electricity used by a chiller system might be affected by variations in outside temperature, relative humidity, hours of facility use, and number of occupants. To accurately model cooling energy consumption, we need to include additional independent variables, creating a *multiple regression* model. Subsequent sections provide more detailed explanations, with examples of multiple regression analysis given in Section 4.4.

### 3.6. Step 6 - Validate Regression Model

Once you have created a baseline model, you can generate the following statistical measures or tests to help validate that your estimated model relationships provide a good description of the data. At a minimum, use the following three measures to determine if your baseline equation is appropriate:  $R^2$  (or *R-squared*), *Net Determination Bias*, and *t-statistic*. Subsequent sections provide additional detail as follows:

- ➔  $R^2$  – Section 5.1.1
- ➔ *Root Mean Squared Error (RMSE) or Standard Error of the Estimate* – Section 5.1.4
- ➔ *CV(RMSE) Coefficient of Variation of the Root Mean Squared Error* – Section 5.1.5
- ➔ *Net Determination Bias* – Section 5.1.6
- ➔ *F-statistic* – Section 5.1.7
- ➔ *t-Statistic* – Section 5.2.2
- ➔ *p-value* – Section 5.2.3

### 3.7. Requirements for Regression

There are four requirements for the appropriate application of linear regression. They are easy to remember with the acronym *LINE*:

1. **Linearity:** There must exist a linear relationship between variables. (The linearity can be between transformations of the variables, but a discussion of methods with transformed variables is beyond the scope of this document.)
2. **Independence:** Two or more regressor variables are independent if their conditional probability distributions are unrelated.
3. **Normality:** A continuous probability density function roughly characterizing a random variable that is the sum of a large number of independent random events; usually

represented by a smooth bell-shaped curve symmetric about the mean. In a normal distribution, the mean (average) of the residuals is zero.

4. **Equal Variance (or *Homoscedasticity*):** Under assumptions of homoscedasticity, different response variables will have the same variance in their errors, regardless of the values of the predictor variables. (Variance is a measure of the average distance between each of a set of data points and their mean value, and it is equal to the sum of the squares of the deviation from the mean value, or the square of the standard deviation.)

To provide accurate predictions, the sample of data used for a regression model should be representative of the overall population. For energy M&V, the baseline modeling period should cover most of the full range of operating conditions. Ideally, the sample observations should be random, but often they will not be. A typical situation is that an engineer will have data for only a subset of the ambient conditions encountered over a year. Depending on the season(s) during which you acquire data, it is likely that the various ambient conditions are not represented in equal proportion to conditions occurring over a full year. Hence, when ambient temperature is an independent variable, individual data points will be improperly treated as having equal weight in a regression, unless you make an effort to adjust, or weight, the data appropriately using *weighted least squares regression*.

A related difficulty that occurs with monthly data is that energy use differs month-to-month, not only because of the weather, but also because the number of days in the months may also differ. You may think that using (heating or cooling) degree-days addresses this issue, because the value of the independent variable would go up as the number of days in a month increases. However, the degree-day observations only affect the slope portion of the equation, yet the intercept of the equation might also be affected.

A common way to address this issue of varying days in a month is to standardize data into daily units, such that the independent variable is expressed as degree-days per day, and the dependent variable is expressed as energy-use-per-day. This is an improvement over the use of monthly data not expressed in daily units. Yet as part of the standardization effort, points that represent more days (such as an observations for 31-day months) should be made more important, or weighted more heavily in the regression, than points that represent fewer days (such as months with 30 days, which in turn should have more weight than observations for February). You can use a weighted least squares regression to appropriately represent the data gathered. This method gives a data point representing a longer period of time proportionally more weight than a point representing less time. The need of using a weighted least squares regression model can be evaluated by checking the model's bias error, described in *Chapter 5, Validating Models*. Weighted least squares regression is outside the scope of this document, but we can generalize the issue by stating that if the baseline modeling period is not very similar to the performance period, then the model may be incorrect. *ASHRAE Guideline 14, Annex D* addresses weighted least squares regression.

It is important to note that linear regression assumes that the x values are known exactly, with no measurement errors. However, in practice, we often ignore this requirement when the variability of the independent variable is small relative to the variability in the dependent variable. Also, the uncertainty calculated in the regression often accounts for the variability in the independent variable measured in the baseline period. The model does not, however, account for the variability measured in the post period and can introduce uncertainty in savings estimates that is

not accounted for by the methods described here. But, with sufficient data, this increased uncertainty should become minor.

One other note regarding the common approach to regression employing the ordinary least squares method to determine regression lines: The squaring used to get the mean squared error weights outliers more than methods based on simple differences, assigning relatively greater importance to large errors than to small ones. Therefore, if the data has outliers, they should be understood and removed if not representing operating conditions, or an alternative regression approach that reduces the impact of outliers (such as one based on the *mean absolute error*) should be considered.



## 4. Models

This chapter describes the various types of linear regression models that are commonly used for M&V. In certain circumstances, other model functional forms, such as second-order or higher polynomial functions, can be valuable. The M&V practitioner should always graph the data in a scatter chart (Step 4 in the process) to verify the type of curve that best fits the data.

The *ASHRAE Inverse Model Toolkit*, a product that came out of research project RP-1050, provides Fortran code for automating the creation of the various model types described below. Spreadsheets and statistical software can create simple linear regressions, polynomial, logistic, and other types of models.

### 4.1. One Parameter Model (Mean Model)

Single parameter (1P), or *mean models*, estimate the mean of the dependent variable and are the simplest models described in this guide. They are not really regression models, but are included here for completeness. A mean model would describe energy use that is not related to other independent variables, such as that of a light that runs continuously.

### 4.2. Two Parameter Model (Simple Regression)

Two parameter (2P) models are the simple linear regression models with which most M&V practitioners are familiar through the use of popular spreadsheet software. They are appropriate for modeling building energy use that varies linearly with a single independent variable, such as ambient temperature. In most commercial buildings, metered whole-building energy use varies linearly with ambient temperature above 75° F due to changes in cooling energy use.

A linear least squares regression with only two parameters is often called a *simple* regression. The equation below is the standard form of a simple regression, illustrated in Figure 4-1 with actual building data.

■ **Simple Regression:**  $Y = \beta_1 + \beta_2 X_1$

where:  $Y$  = the value of the dependent variable

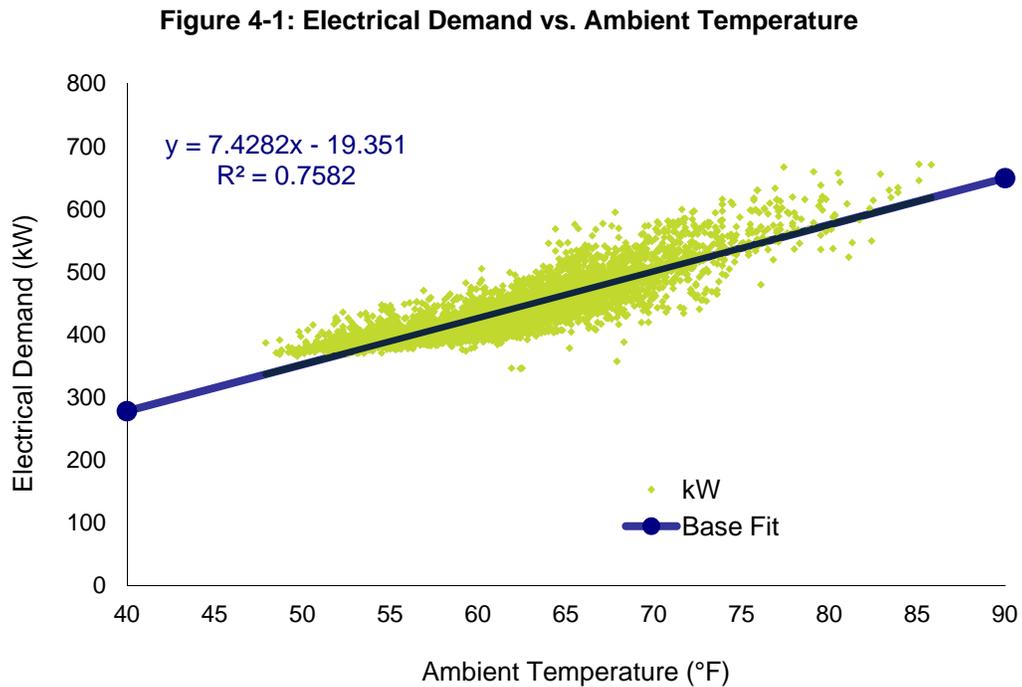
$\beta_1$  = the parameter that defines the *y-intercept* (the value of  $y$  when  $x$  equals zero)

$\beta_2$  = the parameter that describes the linear dependence on the independent variable (*slope*)

$X_1$  = the value of the independent variable

(Note that statisticians typically describe this model as  $Y = \beta_0 + \beta_1 X_1$ . In this text, we use the former notation, as it is consistent with the common engineering terminology *two parameter model*.)

The following graph is an example of a simple regression.



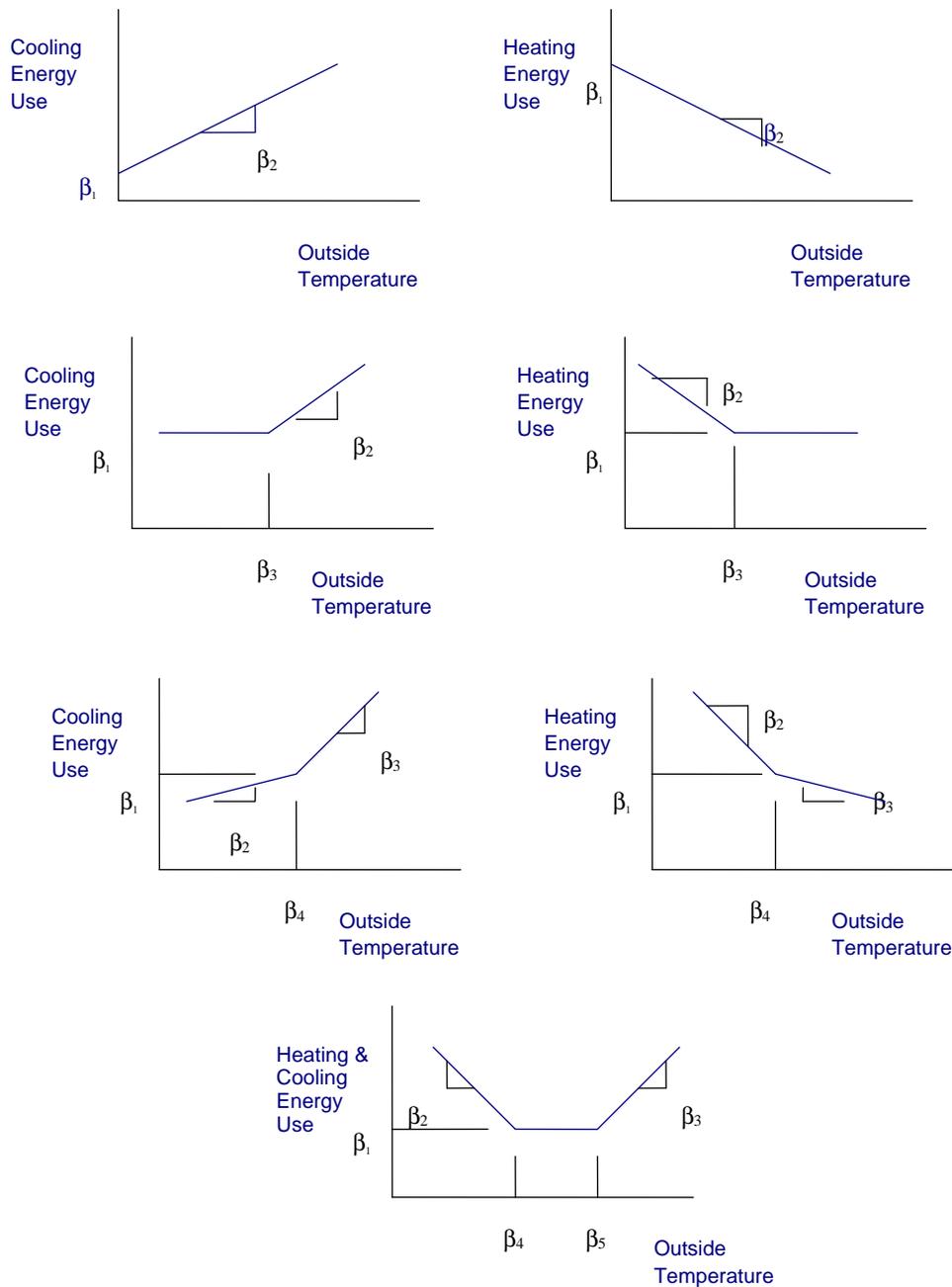
### 4.3. Simple Regression Change Point Models

Some systems are dependent on a variable, but only above or below a certain value. For example, cooling energy use may be proportional to ambient temperature, yet only above a certain threshold. When ambient temperature decreases to below the threshold, the cooling energy use does not continue to decrease, because the fan energy remains constant. In commercial buildings with economizer cooling, this threshold is often 55° F. Similar behavior is often seen in building gas usage, because the heating energy is proportional to ambient temperature during the space heating season and the energy associated with hot water use is constant across all seasons.

In cases like these, simple regression can be improved by using a *change-point* linear regression. Change point models often have a better fit than a simple regression, especially when modeling energy usage for a facility. Because of the physical characteristics of buildings, the data points have a natural 2-line angled pattern to them. Sometimes it is even appropriate to use multiple change points.

The following diagrams (Figure 4-2) illustrate the major models used for temperature-dependent loads. The top row illustrates 2-parameter heating and cooling models; the second row illustrates 3-parameter models; the third row illustrates 4-parameter models; and the bottom row illustrates a 5-parameter combined heating and cooling model.

**Figure 4-2: Figure from ASHRAE Research Project 1050-RP, *Development of a Toolkit for Calculating Linear, Change-point Linear and Multiple-Linear Inverse Building Energy Analysis Models***



### 4.3.1. Three-Parameter Change Point Model

Three-parameter (3P) models are appropriate for energy use that increases or decreases with changes in an independent variable over either the upper or lower part of the range of the independent variable, and remains constant over the remaining part of the independent variable's

range, such as previously described for heating and cooling energy use that varies with temperature only below or above a threshold. (The second row of Figure 4-2 illustrates the three-parameter model.)

■ **Three-Parameter (3P) Cooling Change-Point Model:**  $Y_c = \beta_1 + \beta_2(X_1 - \beta_3)^+$

■ **Three-Parameter (3P) Heating Change-Point Model:**  $Y_h = \beta_1 + \beta_2(X_1 - \beta_3)^-$

where:  $\beta_1$  = the intercept

$\beta_2$  = the parameter defining temperature dependency (slope)

$\beta_3$  = the change-point

$(...)^+$  = indicates that the values of the parenthetic term are set to zero when they are negative

$(...)^-$  = indicates that the values of the parenthetic term are set to zero when they are positive

Another way to think about the mathematics described by the  $(...)^+$  and  $(...)^-$  notations is to consider that the model is run with dummy variables to indicate the  $(...)^+$  and  $(...)^-$  conditions. The dummy variables enter as multipliers on the terms  $(X_1 - \beta_3)$ , which has the result of setting the terms to 0 when they do not meet the criteria; thus the slope,  $\beta_2$ , is only pertinent for the non-zero condition.

### 4.3.2. Four-Parameter Change Point Model

Similar to three-parameter models, four-parameter models incorporate a change point, but do so by incorporating an additional non-zero slope that best fits the relationship over that range of data. Thus, you can use a four-parameter model to better model heating and cooling electricity use with outdoor air temperature as your independent variable for such applications as variable-air-volume systems, certain types of controls, or buildings with both electric heating and cooling. For example, above a certain minimum temperature, there may be two slopes associated with cooling – the portion of the temperature range that includes economizer cooling, and the portion with minimum outside air and compressor cooling only.

Note that the slopes of the two sides of the model can be either of the same or opposite sign, depending upon what is being modeled; in many applications, the slopes have the same sign.

The equation is:

■ **Four-Parameter (4P) Change-Point Model:**  $Y = \beta_1 + \beta_2(X_1 - \beta_4)^- + \beta_3(X_1 - \beta_4)^+$

where:  $\beta_1$  = the constant term

$\beta_2$  = the left slope (heating)

$\beta_3$  = the right slope (cooling)

$\beta_4$  = the change point

$(...)^+$  = indicates that the values of the parenthetic term are set to zero when they are negative

$(...)^-$  = indicates that the values of the parenthetic term are set to zero when they are positive

### 4.3.3. Five-Parameter Change Point Model

Five-parameter models using outdoor air temperature are appropriate in many of the same situations as the four-parameter models. Five-parameter models incorporate slopes of opposite signs on the left and right side, with a constant value in the middle. While these models are used in other regions of the country, they are rarely used to model West Coast climates and systems. The models typically are more appropriate for daily data than for hourly data.

The equation is:

■ **Five-Parameter (5P) Change-Point Model:**  $Y = \beta_1 + \beta_2 (X_1 - \beta_4)^- + \beta_3 (X_1 - \beta_5)^+$

where:  $\beta_1$  = the constant term

$\beta_2$  = the left slope (heating)

$\beta_3$  = the right slope (cooling)

$\beta_4$  = the left change point

$\beta_5$  = the right change point

$(...)^+$  = indicates that the values of the parenthetic term are set to zero when they are negative

$(...)^-$  = indicates that the values of the parenthetic term are set to zero when they are positive

## 4.4. Multiple Regression

The models discussed thus far have all used a single independent variable. Of course, for many building systems, energy use is dependent on more than one variable. In such cases, single variable models will typically result in low  $R^2$  values. When using only one independent variable, the equation has only limited ability to predict the dependent variable, because it does not account for other important factors that should be present in the model.

In such cases, including other variables that are known to influence energy usage will provide a more accurate model. Commonly used variables whose variation is related with variation in energy use include: hours of occupancy in buildings, number of employees on given day, meals served at a restaurant, amount of conditioned floor space, equipment or appliances in use, and water usage. Including two or more independent variables produces a multiple regression model.

Simple regression can be visualized as fitting a line. Multiple regression models with two independent variables fit a plane, and a three variable model fits a 3-dimensional space. The general format of the model is.

$$\blacksquare Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_i X_{i-1}$$

where:  $i$  = the number of predictors

Note that in common statistics terminology, *multiple regression* typically refers to regression models with two or more independent variables and a single dependent variable. In *multivariate regression*, by contrast, there are multiple dependent variables and any number of predictors. The *ASHRAE Inverse Model Toolkit* refers to multiple regression models and change-point models with multiple independent variables as *multiple-variable* or *multi-variable* models.

With multiple regression, additional independent variables will always increase the model's fit. However, this does not necessarily mean that the model is improved, since a model can be over-specified so that the additional independent variables are not statistically significant, or the additional variables are correlated with other independent variables already included in the model. (Refer to *Chapter 5, Validating Models*.)

#### 4.4.1. Categorical Variables

Energy use modeling can account for change of states (broadly, the influence of categorical variables, defined and discussed in this section) by estimating separate models for each state, estimating a single model with categorical variables, and estimating change-point models (a specific type form of a model with categorical variables, described in the next section). Most energy models for M&V will have only one continuous independent variable, but may also incorporate categorical variables.

Variables can be divided into two general types: *continuous* and *categorical*. Continuous variables are numeric and can have any value within the range encountered in the data. Continuous variables are either interval or ratio numbers (where a value of 10 is twice the magnitude as a value of 5). Continuous variables are measured things, such as energy use or ambient temperature. Categorical variables include things like daytype (weekday or weekend, or day of week), occupancy (occupied or unoccupied), and equipment status (on or off). As examples, *occupancy status* is a categorical variable, while *number of occupants* is a continuous variable.

For use in a regression analysis, any categorical variable must be expressed in a binary form, such as taking the value of 1 for Monday and taking the value of 0 for all other days. This is because all the variables in a regression model must be linearly related to the dependent variable. A conceptual category such as day-of-week therefore cannot be included in a regression if it takes values such 1 for Monday, 2 for Tuesday, on up through 7 for Sunday; Tuesday does not have twice the impact on the dependent variable than Monday, nor does Wednesday have three times the impact.

As mentioned at the end of the prior section, one needs to take care in adding additional variables – such as multiple binary variables to describe a composite concept (i.e., day-of-week) – because the model can become overspecified, and the parameter estimates inaccurate and imprecise.

Thus, when needing to create a set of binary variables to capture a composite categorical concept, the M&V practitioner should consider the most concise way to express the underlying relationships between these categories and the dependent variable. Continuing with the day of week example, it may be that activity ramps up during the week; appropriate categories might be Monday/Tuesday, Wednesday/Thursday/Friday, and Saturday/Sunday, where *Mon\_Tues* has the value of 1 if the day is a Monday or Tuesday and 0 otherwise, and similarly for the other variables.

Finally, when working with binary variables describing composite categories, the modeler includes one less binary variable in the equation than the total number of categories in the set. Continuing with the example, when the variables *Mon\_Tues* and *Wed\_Thu\_Fri* both have the value of 0, the day must be a Saturday or Sunday; it would be redundant (that is collinear) to add the variable *Sat\_Sun*.

A common issue in multiple regression for M&V is that categorical variables are included as part of a multiple regression in an improper fashion. Specifically, the categorical variable is often, yet incorrectly, included simply as an additional variable in the regression, which yields a model with different intercepts, depending on the categorical state. When the binary condition is not met (for binary variable  $X_1$ ,  $X_1=0$ ), the model intercept is  $\beta_1$ . When the binary condition is met ( $X_1=1$ ), the model intercept is  $\beta_1 + \beta_2$ . Instead, the true relationship may be that the slope of  $X_2$  changes depending on the categorical state. In that case, the appropriate model includes both  $X_2$  and an interactive term ( $X_1 * X_2$ ). When the binary condition is not met, the value of ( $X_1 * X_2$ ) is 0 and the slope of  $X_2$  is  $\beta_2$ ; when the binary condition is met, the value of ( $X_1 * X_2$ ) is  $X_2$  and the slope of  $X_2$  is  $\beta_2 + \beta_3$ .

According to *ASHRAE RP-1050* (see Section **Error! Reference source not found.**), a common weakness of regression models using categorical variables is that the practitioner creates models with the same slope for all categories. The M&V practitioner needs to carefully consider whether the categorical variable is expected to effect the model's intercept term, a slope term, or both.

An appropriate statistical approach to apply with categorical variables is the *General Linear Model* (GLM). Multiple regression is typically used where the independent variables are continuous, but a general linear model can accommodate both categorical and continuous predictor variables. In avoiding the common pitfall of all categories having the same slope, it is important to use the proper GLM method. (Please refer to a statistics text for further discussion of general linear models.)

Instead of using a multiple regression of the format in *ASHRAE RP-1050*, you can create separate models for each category or combination of categories, and then combine these individual models into a complete model. The basic process is similar to using *IF* statements to determine, for each data point, the category of the categorical independent variable, and then using the intercept and slope that are appropriate for that category.

#### 4.4.2. Multiple Regression Change Point Models

Combining a multiple regression model with a change point model can dramatically improve fits. The methodology to combine these models is similar to applying a change point to a simple

regression. But be cautious not to use more parameters (the  $\beta$ s) than there are variables (the  $X$ s), as this will result in an infinite number of possible solutions.

The following three equations show the formulation of the generic model for three, four, and five-parameter change point models with multiple independent variables. In these models, the *n-parameter* adjective describes the form of the model, but there are actually more parameters because of the added independent variables. Similar models with fewer variables and fewer parameters could also be constructed. The model forms shown below use six parameters, the maximum allowed by the *ASHRAE Inverse Model Toolkit*.

Three-parameter multi-variable regression models (3P-MVR) with four independent variables are written:

$$\blacksquare Y_c = \beta_1 + \beta_2 (X_1 - \beta_3)^+ + \beta_4 X_2 + \beta_5 X_3 + \beta_6 X_4$$

$$\blacksquare Y_h = \beta_1 + \beta_2 (X_1 - \beta_3)^- + \beta_4 X_2 + \beta_5 X_3 + \beta_6 X_4$$

Four-parameter multi-variable regression models (4P-MVR) with three independent variables are written:

$$\blacksquare Y = \beta_1 + \beta_2 (X_1 - \beta_4)^- + \beta_3 (X_1 - \beta_4)^+ + \beta_5 X_2 + \beta_6 X_3$$

Five-parameter multi-variable regression models (5P-MVR) with two independent variables are written:

$$\blacksquare Y = \beta_1 + \beta_2 (X_1 - \beta_4)^- + \beta_3 (X_1 - \beta_5)^+ + \beta_6 X_2$$

Note that the additional parameters used in the multiple regression (e.g.,  $\beta_5$  and  $\beta_6$  for the four-parameter model), are multiplied by their corresponding independent variable, unadjusted for any change point. This method may be appropriate for some continuous independent variables, but it is typically inadequate for categorical variables (see Section **Error! Reference source not found.**).

## 4.5. Uncertainty and Confidence Intervals

### 4.5.1. Uncertainty

Uncertainty in regression analysis can come from multiple sources, *including measurement uncertainty* and *regression uncertainty*. Of these, regression uncertainty is typically of greater importance in the estimation of energy use.

*Measurement uncertainty* has two principal components: *measurement bias* and *measurement precision*. *Bias* relates to issues of calibration and accuracy; *precision* relates to the magnitude of random variation that occurs when multiple measurements are made. For example, when the dependent variable in a regression is energy use, and energy use data are recorded from a utility-grade meter, the uncertainty derived from the accuracy of the measurement (measurement bias) is stipulated to be zero.

*Regression uncertainty* can occur even with perfect measurement (such as when some explanatory variables are omitted from the model) and because of unpredictable behaviors of

people affecting energy use. Uncertainty in regression typically refers to the uncertainty in the output from a regression, that is, uncertainty in the predicted *y-value*. Uncertainty in the regression coefficients is typically referred to in a more explicit manner as the *uncertainty of the slope*.

#### 4.5.2. Confidence Level and Confidence Interval

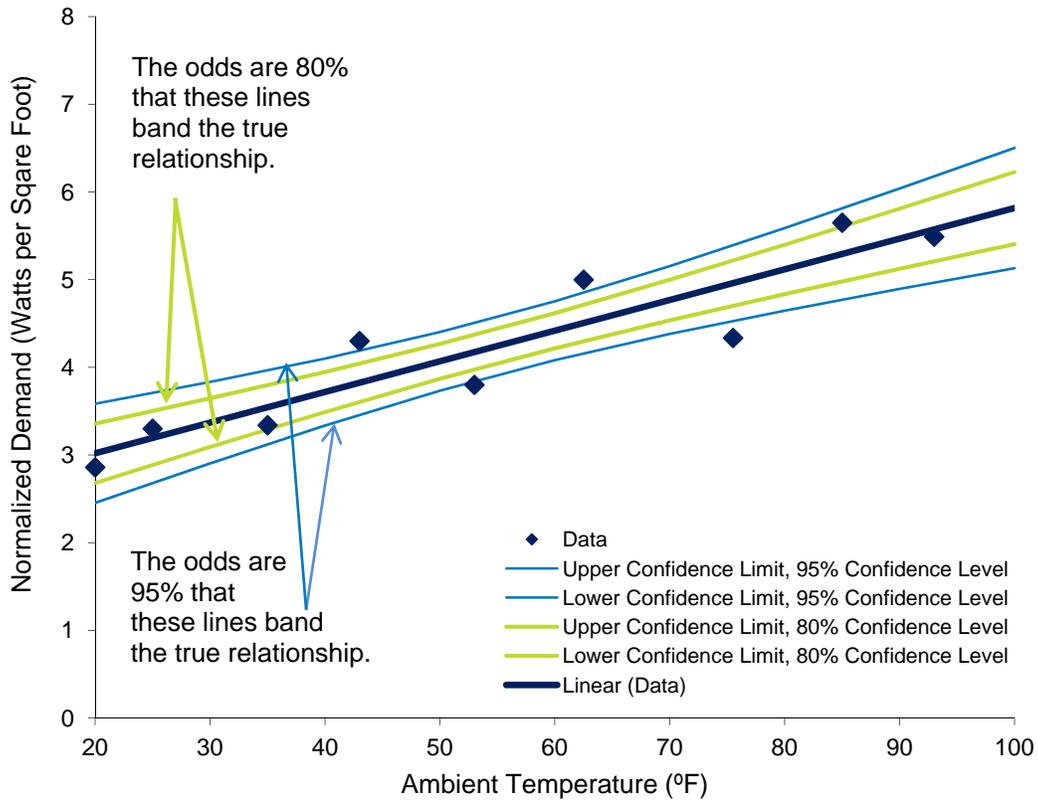
Uncertainty is associated with a given confidence level or probability – for example, “We are 90% confident that the range 433 and 511 kWh bands the true value,” or, as it is more commonly but less accurately expressed, “We are 90% confident that the true value lies between 433 and 511 kWh.” Confidence level is an input number; for a given sample and regression, the higher the confidence level specified, the larger the estimated range that is likely to contain the true value that proportion of the time.

A 95% confidence level implies that there is a 95% chance that the *confidence interval* resulting from a sample contains the true parameter. Confidence intervals define the range – an uncertainty band – that is expected to band the true regression, with a certain probability. The width of the confidence interval provides some idea of uncertainty about the estimated parameters. For example, the results of a regression analysis of savings may be reported as “500 kWh  $\pm 5\%$  at the 95% confidence level.” This means that there is a 95% chance that the confidence interval of 475 to 525 kWh contains the true value of savings. A statement of “500 kWh  $\pm 5\%$  at the 68% confidence level” means that there is only a 68% chance that the true savings value is between these calculated limits, and a 32% chance that it is outside them.

The practitioner should note that the true value does not fluctuate; rather, because of regression uncertainty (and, perhaps, measurement uncertainty), there cannot be complete certainty that the true savings value lies within these limits. Confidence limits are the bounds of the confidence interval.

Figure 4-3 provides a graphical representation of confidence intervals. The bounded confidence intervals in this figure demonstrate that higher chances an interval contains the true regression line require wider intervals than lower chances (that is, the wider the confidence interval, the more likely it is to contain the true value). The lines in this figure represent upper and lower confidence limits.

**Figure 4-3: Confidence Intervals for a Regression**

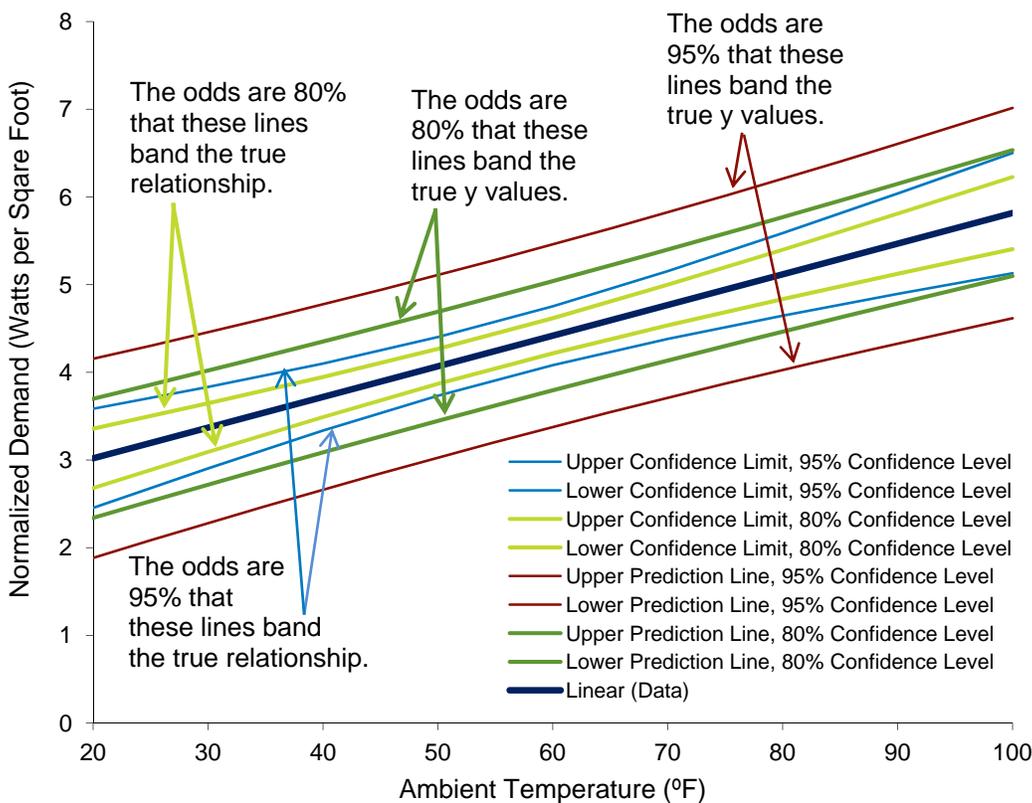


### 4.5.3. Prediction Interval

A prediction interval is an estimate based on earlier observations of the interval in which future data points will fall, with a certain probability. More simply, we can predict the distribution of future points by using the fitted slope and intercept values derived from our regression model. Prediction intervals are similar to confidence intervals, but rather than estimating the distribution of a true parameter, prediction intervals are used to predict the distribution of future samples by indicating the uncertainty in the value of future points. Prediction intervals are wider than confidence intervals since, under the identical conditions, it is more difficult to predict the value of a future point than it is to predict the distribution of the population parameter.

Figure 4-4 illustrates prediction intervals, adding them to Figure 4-3, above.

**Figure 4-4: Prediction Intervals for a Regression**



#### 4.5.4. Confidence Levels and Savings Estimates

Savings estimated from regression analyses should describe the range of values corresponding to a given confidence level. If a single savings estimate, rather than a range, is required, the savings estimate should be the lower value estimated for the required confidence.

The less scatter, or variability, in the data, the narrower the confidence intervals; greater scatter results in wider confidence intervals. However, regardless of the degree of scatter, the confidence interval will be wider when requiring a higher probability that it contains the true regression line or the true value of savings than when requiring a lower probability. For example, the interval estimated for a 99% confidence interval will be wider than it will be for a 95% confidence interval.

For a single value of savings, requiring a greater probability that an interval contains the true value results in a wider uncertainty band and thus a lower estimated minimum savings. If a lower probability is acceptable, the uncertainty band will be narrower and the estimated minimum savings will be higher. To summarize, the minimum savings estimated is higher with a lower confidence level and is lower with a higher confidence level.



## 5. Validating Models

### 5.1. Statistical Tests and Measures for the Model

After developing the regression model, you must assess its *goodness of fit*. There are many ways of testing regression models. The following is an engineering layperson's description of some of the statistical measures and methods used for validating models. Interim measures needed for the statistical tests, such as *root mean squared error*, are also described in this section.

#### 5.1.1. R-Squared (Coefficient of Determination)

The *coefficient of determination* ( $R^2$ ) is the measure of how well future outcomes are likely to be predicted by the model. It illustrates how well the independent variables explain variation in the dependent variable.  $R^2$  values range from 0 (indicating none of the variation in the dependent variable is associated with variation in any of the independent variables) to 1 (indicating all of the variation in the dependent variable is associated with variation in the independent variables, a "perfect fit" of the regression line to the data). The rule-of-thumb for an acceptable model using monthly billing data is an  $R^2 > 0.75$ .

If the  $R^2$  is low, you may wish to return to Step 5 in the regressions process (see Chapter 3) and select additional independent variables that make sense to add to your model; then use the adjusted  $R^2$  (see Section 5.1.2) as a goodness-of-fit test for a multiple regression.

The  $R^2$  value can be thought of as a goodness-of-fit test; but a high  $R^2$  value is not enough to say the selected model fits the data well, nor that a low  $R^2$  indicates a poor model. Fit criteria in addition to  $R^2$  should be assessed. For CR-RMSE (see Section 5.1.5), a low value (often interpreted as 10% or 15%) is desirable. For example, a model with a low  $R^2$  is acceptable when there is a clear relationship between the dependent and independent variables, as evidenced by the following: The scatter of the observed  $y$ -values around the regression line is low, yet large in relationship to the total scatter of  $y$ -values from the mean of  $y$ , and total  $y$  scatter is much smaller than the total scatter of  $x$ -values from its mean (this results in a low slope estimate). In a situation where the total scatter of  $y$  and  $x$  compared to their means is more comparable, a low  $R^2$  can be acceptable when the estimated coefficient of  $x$  is significant, in spite of the unexplained variation; however, there will be relatively high uncertainty in the resulting savings estimates.

The calculations for estimating uncertainty are described in Section 4.5.

#### 5.1.2. Adjusted R-Squared

In multiple regression models, the addition of an independent variable will always result in an increase in the model's  $R^2$ , which means the basic  $R^2$  value is not an appropriate indicator of model fit. Instead, one should judge model fit using adjusted  $R^2$ , a value produced by adjusting  $R^2$ , dividing  $R^2$  by the associated degrees of freedom (discussed next). The value of the adjusted  $R^2$  only increases from one model specification to another if the additional independent variable(s) improve the model more than by random chance.

### 5.1.3. Degrees of Freedom

*Degrees of freedom* is a common input for statistical calculations. Degrees of freedom is the number of values in a calculation that are free to vary and is calculated by subtracting the number of parameters in the model from the total number of data points.

### 5.1.4. Root Mean Squared Error (Standard Error of the Estimate)

*Root mean squared error* (RMSE) is an indicator of the scatter, or random variability, in the data, and hence is an average of how much an actual  $y$ -value differs from the predicted  $y$ -value. It is the standard deviation of errors of prediction about the regression line. *Standard error of the estimate* (SE) is always adjusted by the number of parameters in the model. Keep in mind, however, that some sources include the adjustment for the number of parameters in their definition of RMSE; others do not. In this document, SE and RMSE are synonymous, and include the adjustment for the number of parameters in the model. Standard error of the estimate is sometimes called *standard error of prediction*.

### 5.1.5. Coefficient of Variation of the Root Mean Squared Error

*Coefficient of variation of the root mean squared error* – CV(RMSE) – is the RMSE normalized by the average  $y$ -value. Normalizing the RMSE makes this a nondimensional that describes how well the model fits the data. It is not affected by the degree of dependence between the independent and dependent variables, making it more informative than R-squared for situations where the dependence is relatively low.

### 5.1.6. Bias

*Energy* models should always be checked for bias: Does the model re-create the baseline energy use? *Demand* models, on the other hand, generally do not require a bias check, since demand is not summed over time. Also, demand models will generally not require different points to have different weights and so that potential for bias error (from not using a weighted regression when one is warranted) is not a concern. Therefore, since regression itself minimizes the error for each point, there will typically be no need to check bias for a demand model. M&V practitioners should take care to understand any unique situations that may require checking for bias in a demand model.

Two indices are defined in *ASHRAE Guideline 14* for checking energy model bias. These two indices are *net determination bias error* (or *mean bias error*) and *normalized mean bias error*. Be forewarned that the Guideline is somewhat confusing, since these two indices are nearly the same and one of the indices is called two different things at different places in the document.

Net determination bias is simply the percentage error in the energy use predicted by the model compared to the actual energy use. The sum of the differences between actual and predicted energy use should be zero. If the net determination bias = 0, then there is no bias. *ASHRAE Guideline 14-2002* accepts an energy model if the net determination bias error is less than 0.005%.

Often, bias may be minor, but it still will affect savings estimates. If the savings are large relative to the bias, bias may not be important, but in many cases, bias could be influential.

- **Net Determination Bias Error (NBE):** 
$$NBE = 100 * \frac{\sum_i (E_i - \hat{E}_i)}{\sum_i E_i}$$
- **Normalized Mean Bias Error (NMBE):** 
$$NMBE = 100 * \frac{\sum_i (E_i - \hat{E}_i)}{(n - p) * \bar{E}}$$

Note that the two indices are identical if, in NMBE,  $p=0$ . Therefore, the only difference between the two bias error calculations is an adjustment for the number of parameters in the model.

Since there is no averaging occurring, it seems that *mean bias error* is a misnomer. The *net determination bias error* is simply the percentage error in total energy use predicted by the model over the relevant (baseline) time period. In the equation for *normalized mean bias error*, there is an average term in the denominator, but the result is still simply a percent error, just one that is adjusted for the number of parameters in the model.

Regression models *by themselves* will not typically have any bias if created properly. However, as stated above, there can be bias when using regression models, either because multiple categories need to be considered, or because an unweighted regression was used when data points should not have equal weights.

Checking for model bias is an important part of model validation, but there does not seem to be any value in using *both* of these very similar bias calculations. Keep it simple and just use *net determination bias error*, which provides a net percentage error in the model.

To clarify some of the confusion between guidelines, we have listed the terms and uses for various guidelines below.

- ➔ **Normalized Mean Bias Error** – is called *net mean bias error* in the *Guidelines for Verifying Existing Building Commissioning Project Savings*.
- ➔ **Net Determination Bias Error** – is called by this same term in the *Guidelines for Verifying Existing Building Commissioning Project Savings*.
- ➔ **Mean Bias Error** – is referenced by *ASHRAE Guideline 14* in 6.3.3.4.2.2 *Statistical Comparison Techniques*, but the verbal definition of this term is the same as the equation for *net determination bias error*.
- ➔ **Net Determination Bias** – is a term not found in the statistical literature. References on the Internet point exclusively to *ASHRAE Guideline 14*. Consider *net determination bias* as simply a percentage error.

### 5.1.7. F-Statistic

The *F-statistic* is similar to the *t-statistic* (described subsequently), but is for the entire model rather than for individual variables. When testing a model, the larger the value of *F*, the better.

In the Excel Regression tool output, *Significance F* is the whole-model equivalent of *p-value* for an individual variable. For a simple regression (no change points) with a single independent variable, the *Significance F* value is the same as the *p-value* for the independent variable. It is the probability that the model does *not* explain most of the variation in the dependent variable. Therefore, low values for *Excel's Significance F* are desirable.

## 5.2. Statistical Tests and Measures for the Model's Coefficients

### 5.2.1. Standard Error of the Coefficient (Intercept or Slope)

The *standard error of the coefficient* is similar to the *standard error of the estimate*, but is calculated for a single coefficient rather than the complete model. The standard error is an estimate of the standard deviation of the coefficient. For simple linear regression, it is calculated separately for the slope and intercept: there is a *standard error of the intercept* and *standard error of the slope*. These are necessary to get the *t-statistic* for each.

### 5.2.2. *t*-Statistic

The *t-statistic* is the coefficient ( $\beta_i$ ) divided by its standard error. Within regression, the *t-statistic* is a measure of the significance for each coefficient (and, therefore, of each independent variable) in the model. The larger the *t-statistic*, the more significant the coefficient is to estimating the dependent variable. The coefficient's *t-statistic* is compared with the critical *t-statistic* associated with the required confidence level and degrees of freedom. For a 95% confidence level and a large number for degrees of freedom (associated with a lot of data), the comparison *t-statistic* is 1.96. Measure the *t-statistic* for every independent variable used, and if the *t-statistic* is lower than the critical value (such as 1.96) for any variable, reconsider your model. Go back to Step 5 (see Chapter **Error! Reference source not found.**) and consider if a different model specification is more appropriate. Note that the more variables used in a regression, the lower will be the significance of each variable.

### 5.2.3. *p*-value

The *p-value* is the probability that a coefficient or dependent variable is not related to the independent variable. Small *p-values*, then, indicate that the independent variable or coefficient is a significant (important) predictor of the dependent variable in your model. The *p-value* is an alternate way of evaluating the *t-statistic* for the significance of a regression coefficient. Rather than requiring an input confidence level as you would to compare the *t-statistic*, the *p-value* provides probability as an output.

### 5.3. Tables of Statistical Measures

Table 5-1 through Table 5-4, below, present the definitions of the relevant statistical measures, their equation formulas, and their calculation in *Microsoft Excel*.

**Table 5-1: Definitions of Regression Model Statistics**

Regression Model Statistic	Equation or Definition
<b>n</b>	Number of points
<b>p</b>	Number of parameters
<b>df</b>	Degrees of freedom, =n-p
<b>Y<sub>avg</sub></b>	= $\sum(Y)/n$
<b>X<sub>avg</sub></b>	= $\sum(X)/n$
<b>SSQ<sub>total</sub></b>	= $\sum((Y-Y_{avg})^2)$
<b>SSQ<sub>reg</sub></b>	= $\sum((Y_{Calc}-Y_{avg})^2)$
<b>SSQ<sub>res</sub> (or SSE)</b>	= $\sum((Y-Y_{calc})^2)$
<b>SSQ<sub>x</sub></b>	= $\sum((X-X_{avg})^2)$
<b>F</b>	= $SSQ_{Reg}/(SSQ_{res}/(n-p))$
<b>RMSE</b>	= $\sqrt{SSQ_{res}/(n-p)}$
<b>Standard Error of Estimate</b>	= $\sqrt{(1/(n-p)*(SSQ_{total})-(\sum((X-X_{avg})*(Y-Y_{avg}))^2)/(\sum((X-X_{avg})^2)))}$
<b>CV-RMSE</b>	= $RMSE/Y_{avg}$
<b>R-Squared</b>	= $SSQ_{reg}/SSQ_{total}$
<b>R-Squared</b>	= $1 - SSQ_{res}/SSQ_{total}$
<b>Adjusted R-Squared</b>	= $1-((1-R^2)*((n-1)/(n-p-1)))$
<b>Net Determination Bias</b>	= $\sum(Y-Y_{Calc})/\sum(Y)$
<b>Confidence Half-Interval</b>	= $t\text{-statistic}*SE*\sqrt{1/n+(X-X_{avg})^2/SSQ_x}$
<b>Prediction Half-Interval</b>	= $t\text{-statistic}*SE*\sqrt{1+1/n+(X-X_{avg})^2/SSQ_x}$

**Table 5-2: Microsoft Excel Functions for Regression Model Statistics**

Regression Model Statistic	Microsoft Excel Function	Excel LINEST (Where Applicable)
<b>n</b>	= <code>COUNT(XVals)</code>	
<b>p</b>	2	
<b>df</b>	= <code>n-p</code>	= <code>INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 4,2)</code>
<b>Y<sub>avg</sub></b>	= <code>AVERAGE(Yvals)</code>	
<b>X<sub>avg</sub></b>	= <code>AVERAGE(XVals)</code>	
<b>SSQ<sub>total</sub></b>	= <code>DEVSQ(Yvals)</code>	
<b>SSQ<sub>reg</sub></b>	= <code>DEVSQ(YvalsCalc)</code>	= <code>INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 5,1)</code>

Continued

Regression Model Statistic	Microsoft Excel Function	Excel LINEST (Where Applicable)
<b>SSQ<sub>res</sub> (or SSE)</b>	= SUM((Yvals-YvalsCalc)^2)	= INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 5,2)
<b>SSQ<sub>x</sub></b>	= DEVSQ(XVals)	
<b>F</b>	= DEVSQ(YvalsCalc)/(SUM((Yvals-YvalsCalc)^2)/(n-p))	= INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 4,1)
<b>RMSE</b>	= SQRT(SUM((Yvals-YvalsCalc)^2)/(n-p))	
<b>Standard Error of Estimate</b>	= STEYX(Yvals,XVals)	= INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 3,2)
<b>CV-RMSE</b>	= SQRT(SUM((Yvals-YvalsCalc)^2)/(n-p))/AVERAGE(Yvals)	
<b>R-Squared</b>	= RSQ(Yvals,XVals)	= INDEX(LINEST(Yvals,XVals, TRUE,TRUE), 3,1)
<b>R-Squared</b>	= RSQ(Yvals,XVals)	
<b>Adjusted R-Squared</b>	= 1-((1-RSQ(Yvals,XVals))*((n-1)/(n-p-1)))	
<b>Net Determination Bias</b>	= SUM(Yvals-YvalsCalc)/SUM(Yvals)	
<b>Confidence Half-Interval</b>	Evaluated at each x	
<b>Prediction Half-Interval</b>	Evaluated at each x	

**Table 5-3: Definitions of Coefficient Statistics**

Coefficient Statistic	Equation or Definition
<b>Confidence Level</b>	Input required probability that the coefficient is not zero
<b>t-Statistic, Critical</b>	From table
<b>Intercept</b>	= Yavg-Slope*Xavg
<b>Slope</b>	= $\sum((X-X_{avg})*(Y-Y_{avg}))/(\sum(X-X_{avg})^2)$
<b>Standard Error of Intercept</b>	= $\sqrt{(SSQ_{res}/(n-p)*(1/n+X_{avg}^2/\sum((XVals-\sum(XVals)/n)^2))}$
<b>Standard Error of Slope</b>	= $\sqrt{(SSQ_{res}/(n-p))/(SSQ_x)}$
<b>t-Statistic for Intercept</b>	= intercept/(Standard Error of intercept)
<b>t-Statistic for Slope</b>	= slope/(Standard Error of slope)
<b>p-Value for Intercept</b>	—
<b>p-Value for Slope</b>	—

**Table 5-4: Microsoft Excel Functions for Coefficient Statistics**

Coefficient Statistic	Microsoft Excel Function	Excel LINEST (Where Applicable)
<b>Confidence Level</b>	0.95	
<b>t-Statistic, Critical</b>	= <i>TINV</i> (1-ConfLvl,n-p)	
<b>Intercept</b>	= <i>INTERCEPT</i> (Yvals,XVals)	
<b>Slope</b>	= <i>SLOPE</i> (Yvals,XVals)	= <i>INDEX</i> ( <i>LINEST</i> (Yvals,XVals, TRUE,TRUE), 1,2)
<b>Standard Error of Intercept</b>	= <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/n+Xavg^2/ <i>DEVSQ</i> (XVals))	= <i>INDEX</i> ( <i>LINEST</i> (Yvals,XVals, TRUE,TRUE), 1,1)
<b>Standard Error of Slope</b>	= <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/ <i>DEVSQ</i> (XVals))	= <i>INDEX</i> ( <i>LINEST</i> (Yvals,XVals, TRUE,TRUE), 2,2)
<b>t-Statistic for Intercept</b>	= ( <i>INTERCEPT</i> (Yvals,XVals))/( <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/n+Xavg^2/ <i>DEVSQ</i> (XVals)))	= <i>INDEX</i> ( <i>LINEST</i> (Yvals,XVals, TRUE,TRUE), 2,1)
<b>t-Statistic for Slope</b>	= ( <i>SLOPE</i> (Yvals,XVals))/( <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/ <i>DEVSQ</i> (XVals)))	
<b>p-Value for Intercept</b>	= <i>TDIST</i> ( <i>ABS</i> ( <i>INTERCEPT</i> (Yvals,XVals))/( <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/n+Xavg^2/ <i>DEVSQ</i> (XVals))),n-p,2)	
<b>p-Value for Slope</b>	= <i>TDIST</i> ( <i>ABS</i> ( <i>SLOPE</i> (Yvals,XVals))/( <i>STEYX</i> (Yvals,XVals)* <i>SQRT</i> (1/ <i>DEVSQ</i> (XVals))),n-p,2)	

## 5.4. Other Tests of Model Validity

### 5.4.1. Check for Autocollinearity

*Autocorrelation*, sometimes called *serial correlation*, is the correlation of values in a time series with prior and future values. When autocorrelation exists, the model violates the requirement that the *y*-values be independent of each other. Autocorrelation can be common in energy models, especially with data taken at short time intervals. For example, hourly meter data will generally exhibit autocorrelation.

The impact of autocorrelation is that the effective number of data points is fewer than the actual number, since the information in each observation is not completely new. A consequence of this is that the variability looks lower than it actually is, making some predictors look significant when they are not. In the equations for the statistical tests, the effective number of data points needs to be substituted for *n*, the actual number of data points.

To calculate autocollinearity, *R-squared* is first calculated for the correlation between the residuals and the residuals for the prior time period. The autocorrelation coefficient  $\rho$  is then the square root of this value.<sup>4</sup> The effective number of data points is then given by:

$$\blacksquare \quad n = n*(1-\rho)/(1+\rho)$$

*Annex D of ASHRAE Guideline 14* suggests that autocorrelation can be ignored for values of  $\rho$  less than 0.5.

#### 5.4.2. Check for Multicollinearity

With multiple regression, models should be checked to avoid multicollinearity. *Multicollinearity* is a strong relationship between two or more of the *independent* variables. Broad discussion of multicollinearity is beyond the scope of this document. The key point is that allowing multicollinearity in a model can create a number of problems and lead to incorrect inferences from the model. Multicollinearity between two independent variables means that standard errors for coefficients are over-emphasized, and therefore larger. The coefficient estimates may change erratically in response to small changes in the model or the data. Even the signs of coefficients can be incorrect!

Multicollinearity may not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors. Significant relationships between independent variables make it difficult to determine which of the correlated independent variables are most significant – that is, which ones most explain variations in the dependent variable.

To avoid multicollinearity, use as few independent variables as possible to obtain a reasonable model, and have a good understanding of the variables you are using. Creating a good multivariate model begins with a strong understanding of what drives energy use. You can avoid multicollinearity by creating a model that you think best describes your dependent variable and then checking the coefficient of correlation among all the independent variables. Scatter plots of the independent variables together can assist in visually seeing whether one independent variable is correlated with another. When you are assured the correlation is low among the independent variables, check via scatter plots to see that the relationships between each independent variable and the dependent variable are viable and linear. This can give you a sense of the impact that each independent variable has on the dependent variable.

Understanding the theoretical impact that an independent variable has on the dependent variable can help you to avoid using two independent variables that are correlated. Finally, after running the whole multivariate model and checking scatter plots, if you are still concerned about multicollinearity, you can add independent variables one at a time. This is commonly known as *step-wise regression*. Evaluate the *t*-statistic or *p*-value for each variable as it is added, to make sure it is significant.

---

<sup>4</sup> Note, the English spelling of the Greek letter  $\rho$  is *rho*, not to be confused with “p.”

## 5.5. Analysis of Residuals

Analysis of residuals can provide a relatively easy way to confirm the assumptions required for a valid linear regression are met (see Section 3.7) and to help validate the model's predictions.

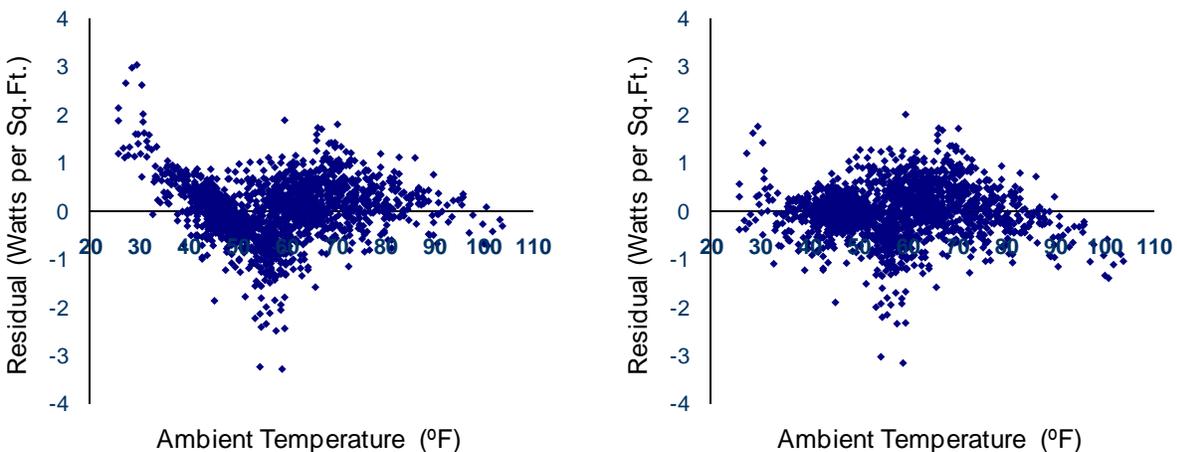
Plotting the predicted  $y$ -values against the actual  $y$ -values provides a quick way to validate the model's predictions; the slope should be close to 1.

A test for equality of variance (*homoscedasticity*) is to plot residuals against the independent variable(s). There should be no discernable pattern. The scatter in the residuals should be the same regardless of the value of the independent variable (visually blob-like). Note that building energy models may often fail this test, because the scatter in energy use typically varies with ambient temperature.

The assumption of normal distribution of residuals can be checked with a histogram of the residuals. The assumption for independence of residuals can be checked with a *lag plot*. A lag plot charts the value of a residual against the residual from one or more time periods earlier. To verify that the relationship is not changing over time, the residuals can be plotted against time.

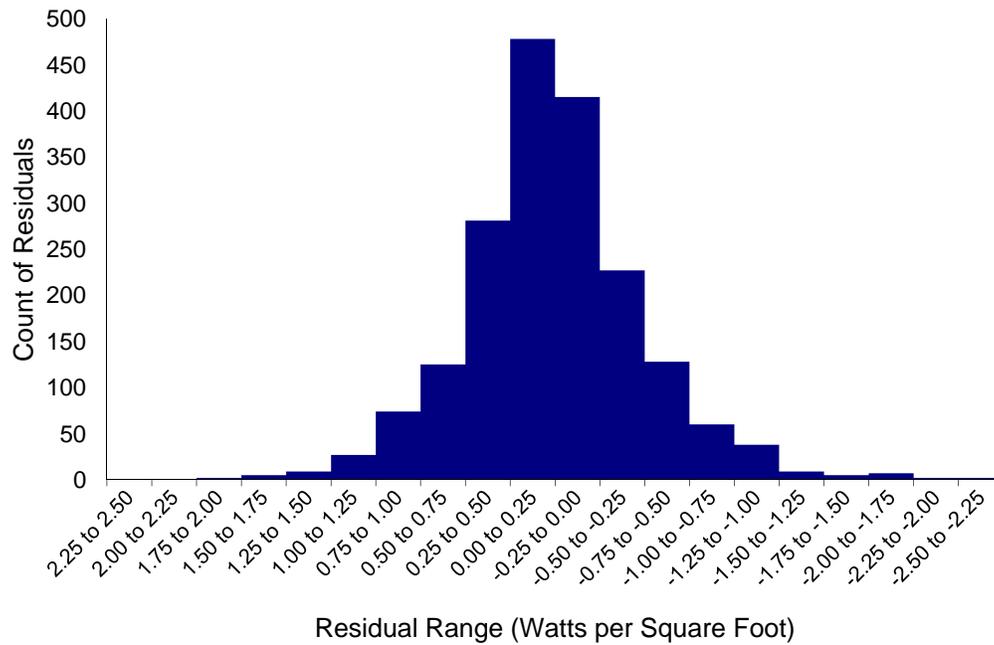
The two charts shown in Figure 5-1 demonstrate how a residual chart can be used for a test of equal variance. The first chart uses a linear 2P model and it is apparent that there is a pattern to the residuals. The second chart uses a 4P model and there is little pattern to the residuals. A practical rule-of-thumb for interpreting these plots is that if by covering up one or two points on a plot, it changes whether you see a pattern or not, those points are probably creating the impression of a pattern where there is not one.

**Figure 5-1: Residuals vs. Independent Variable**



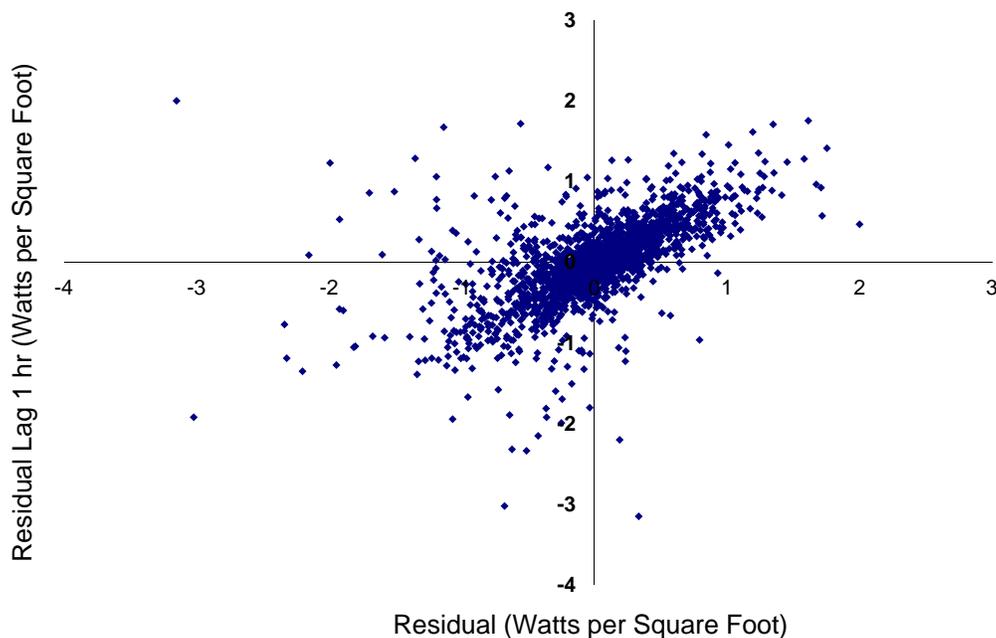
Next, the chart in Figure 5-2, for a 4P model, shows how a histogram can be used to qualitatively assess whether the distribution of residuals approaches a normal distribution.

**Figure 5-2: Histogram of Residuals**



The last chart for analysis of residuals (Figure 5-3) is a lag plot for a model using hourly data. It charts the residuals against the residuals from the prior hour. The strong relationship shown indicates that this model suffers from autocorrelation. For this model, the autocorrelation should be accounted for, or the uncertainty in the model will be underestimated.

**Figure 5-3: Residual Lag Plot**



## 6. Example

### 6.1. Use of Monthly Billing Data in a 2-Parameter Model to Evaluate Whether It Will Make a Satisfactory Baseline

Regression is commonly used to analyze monthly utility data. It is best applied to a package of measures whose total savings is a relatively high percentage of the building's baseline energy use. It is important to remember that the energy use of buildings is typically dependent on weather. More specifically, it can be dependent on the demand for cooling and heating. This is because energy usage is usually higher when it is either very cold (heaters) or very hot (AC units), since the temperature is far from the balance point.

In cases where only cooling or only heating is present or relevant, a simple 2-parameter (straight-line) regression is often satisfactory.

Consider the case of schools in the Northwest, especially on the west side of the Cascade Mountains. Many schools do not have cooling, and although cooling is not generally needed during the school year, heating is. Therefore, a model of energy use versus *average ambient temperature* or *heating degree-days* (HDD) may be appropriate.

Usually, degree-days are better than average-ambient-temperatures. An average temperature may indicate little need for heating or cooling if it is near the balance point for the building. However, a moderate average temperature can be made up of a series of cool temperatures and a series of warm temperatures. During the times of cool temperatures, heating is needed. Therefore, depending on climate, a better fit will typically be found by using degree-days. On the west side of the Cascades in the Northwest, winter temperatures may be relatively constant over a day, and almost always below a school building's balance point, so the greatest difference between degree-days and average temperature will be found in the spring and fall months.

The following analysis estimates the baseline for the electricity use of a group of modular classrooms heated by heat pumps. The planned measure is a web-enabled programmable thermostat. Prior similar projects have shown savings exceeding 45%.

The available data are the monthly electricity energy use, kWh, and ambient temperature during the billing period. There are 24 months of data to be used for the baseline. The data to be used for the regression will be normalized to *average kWh per day* in each billing period and *average heating degree-days* per day in each billing period. The base temperature for heating degree-days in this example is 65° F. (See Section 6.2 for a discussion of heating degree-days.)

The relevant equation is for a common 2-parameter *ordinary least squares regression*:

$$\blacksquare Y = \beta_1 + \beta_2 X_1$$

where:  $Y$  = electricity use per day in the billing period

$\beta_1$  = *y-intercept* – electricity use (kWh-per-day) for a day with zero heating degree-days

$\beta_2$  = slope – how much the energy use increases for a day as the temperature decreases below 65° F (kWh-per-day per heating degree-day)

$X_1$  = average heating degree-days per day in the billing period

Table 6-1 provides the data for the project.

**Table 6-1: Example Data for Classroom Heat Pump Project**

End of Billing Period	Billing Period Duration in Days	Billed Usage kWh	HDD in Billing Period
09/24/2007	30	6,080	113
10/24/2007	30	7,330	311
11/21/2007	28	7,470	463
12/19/2007	28	10,000	669
01/23/2008	35	11,480	877
02/25/2008	33	11,420	782
03/26/2008	30	9,970	560
04/24/2008	29	7,840	561
05/20/2008	26	6,800	265
06/21/2008	32	5,980	268
07/23/2008	32	4,310	73
08/22/2008	30	3,330	57

The consumption and heating degree-days are standardized by the number of days in the billing period (Table 6-2).

**Table 6-2: Data Standardized by Days in the Billing Period**

End of Billing Period	Billing Period Duration in Days	Billed Usage kWh	HDD in Billing Period	Average kWh per Day in Billing Period	Average HDD per Day in Billing Period
09/24/2007	30	6,080	113	202.7	3.7
10/24/2007	30	7,330	311	244.3	10.4
11/21/2007	28	7,470	463	266.8	16.5
12/19/2007	28	10,000	669	357.1	23.9
01/23/2008	35	11,480	877	328.0	25.1
02/25/2008	33	11,420	782	346.1	23.7
03/26/2008	30	9,970	560	332.3	18.7
04/24/2008	29	7,840	561	270.3	19.4
05/20/2008	26	6,800	265	261.5	10.2
06/21/2008	32	5,980	268	186.9	8.4
07/23/2008	32	4,310	73	134.7	2.3
08/22/2008	30	3,330	57	111.0	1.9

Table 6-3 provides the *Microsoft Excel* formulas for the regression. Note that  $p$  in the term  $(n-p)$  refers to the number of parameters, which is two for this simple linear regression.

**Table 6-3: Microsoft Excel Formulas for the Regression**

Output	Formula
R-squared	= $RSQ(Yvals, XVals)$
Number of Baseline Points, n	= $COUNT(YVals)$
CV-RMSE	= $SQRT(SUM((Yvals - YvalsCalc)^2) / (n - p)) / AVERAGE(Yvals)$
Intercept at HDD=0	= $INTERCEPT(Yvals, XVals)$
Slope	= $SLOPE(Yvals, XVals)$

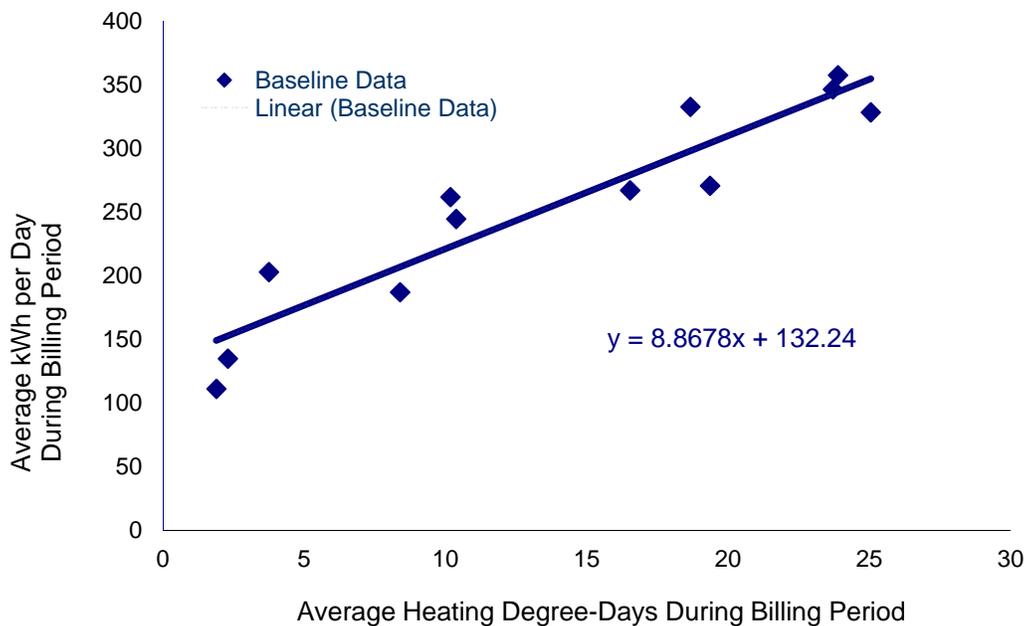
Table 6-4 provides the *Excel* output:

**Table 6-4: Microsoft Excel Output for Example Model**

Output	Data
R-squared	= 0.879
Number of Baseline Points, n	= 12
CV-RMSE	= 11.7%
Intercept at HDD=0	= 132.24
Slope	= 8.8678

Figure 6-1 shows the data graphed, with the regression equation and line included.

**Figure 6-1: Baseline Electricity Use vs. Heating Degree-Days**



Next, the uncertainty needs to be calculated. The input confidence level used to calculate the *t*-statistic will be 90%. The *t*-statistic will be used to get the confidence intervals, evaluated at each value of *X*. To calculate the *t*-statistic, some intermediate calculations need to be made, as shown in Table 6-5. In this table, *p* is the probability that the dependent variable is not significantly related to the independent variable.

**Table 6-5: Microsoft Excel Formulas for the Fit Statistics**

Output	Formula
Standard Error	= STEYX(Yvals,XVals)
Standard Error – Percent of Average	= STEYX(Yvals,XVals) / AvgY
Critical <i>t</i> -Statistic	= TINV(1-ConfLvl,n-p)
Sum of Squares of Differences: <i>X-avg(X)</i>	= DEVSQ(XVals)
Standard Deviation of the Residuals	= STDEV(Residuals)
<i>t</i> -Statistic	= CONFIDENCE(1-ConfLvl,STDEV(Residuals),n)
<i>p</i> -Value	= TDIST(ABS( <i>t</i> _statistic),n-p,2)

Table 6-6 provides the *Excel* outputs for the goodness-of-fit statistics.

**Table 6-6: Microsoft Excel Output for Example Fit Statistics**

Output	Data
Standard Error	= 29.69
Standard Error – Percent of Average	= 11.7%
Number of Baseline Points, <i>n</i>	= 12
Critical <i>t</i> -Statistic	= 12
Sum of Squares of Differences: <i>X-avg(X)</i>	= 1.81
Standard Deviation of the Residuals	= 818
<i>t</i> -Statistic	= 28.3
<i>p</i> -Value	= 13.44

Below is the equation for calculating the confidence intervals for the regression:

$$\blacksquare \Delta Y_{confidence} = \pm (t\text{-statistic}) * STEYX(Yvals,XVals) * SQRT(1/n + (X-xAvg)^2 / DEVSQ(XVals))$$

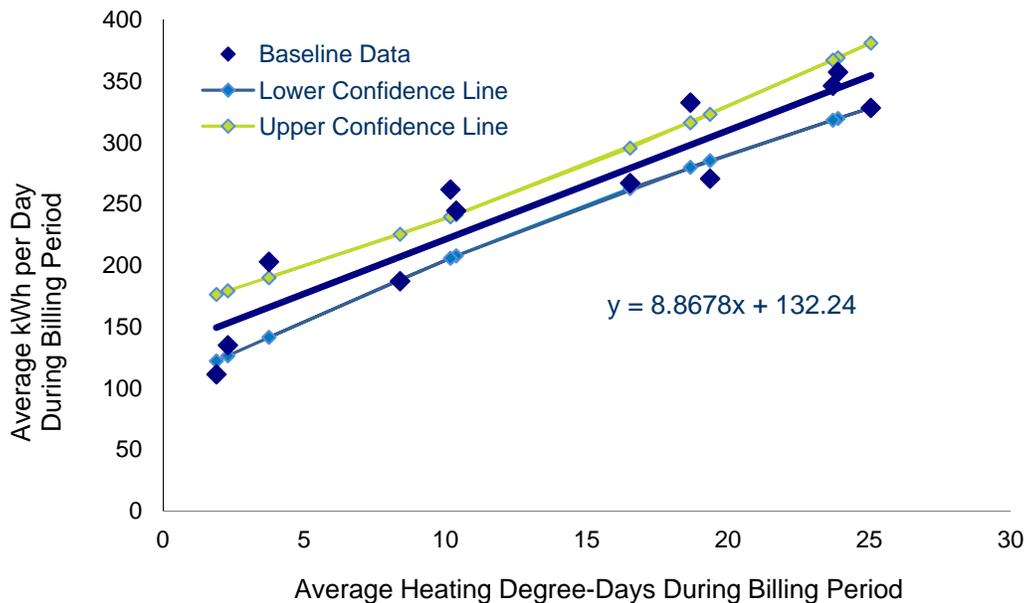
Table 6-7 provides the spreadsheet output, including the estimates for the confidence intervals of the regression. *Min Modeled* is the modeled value minus the confidence half-interval. *Max Modeled* is the modeled value plus the confidence half-interval.

**Table 6-7: Example Model Estimates**

Average HDD per Day in Billing Period	Average kWh per Day in Billing Period	Modeled kWh per Day	Residual	90% Confidence Half-Interval	Minimum Modeled kWh per Day	Maximum Modeled kWh per Day
3.7	202.7	165.5	37.2	24.3	141.2	189.8
10.4	244.3	224.2	20.1	16.7	207.5	241.0
16.5	266.8	278.8	-12.0	16.4	262.4	295.3
23.9	357.1	344.1	13.1	24.7	319.4	368.8
25.1	328.0	354.4	-26.4	26.5	327.9	380.8
23.7	346.1	342.5	3.6	24.5	318.0	366.9
18.7	332.3	297.7	34.6	18.2	279.6	315.9
19.4	270.3	303.9	-33.6	18.9	285.1	322.8
10.2	261.5	222.4	39.1	16.9	205.6	239.3
8.4	186.9	206.6	-19.7	18.4	188.2	225.1
2.3	134.7	152.6	-17.9	26.5	126.1	179.0
1.9	111.0	149.0	-38.0	27.1	121.9	176.1
<b>Total</b>	<b>3,041.8</b>	<b>3,041.8</b>	<b>0.0</b>	<b>258.9</b>	<b>2,782.8</b>	<b>3,300.7</b>

Figure 6-2 provides the scatter chart again, including the lines of 90% confidence intervals.

**Figure 6-2: Baseline Electricity Use vs. Heating Degree-Days with Confidence Intervals**



Note that the regression appears to reproduce the baseline totals. However, these values are for the average kWh-per-day, not for the total energy use over the year. Yet each point does not

represent the same number of days; consequently, the best approach would have been to use a weighted regression. Because a weighted regression was not used, the model’s bias should be checked.

To complete the model and check the bias, the modeled values for kWh-per-day are multiplied by the number of days in the billing period. The actual kWh values are reproduced in Table 6-8 for comparison with the modeled values.

**Table 6-8: Example Model Estimates with Actual Observations**

Average HDD per Day in Billing Period	Average kWh per Day in Billing Period	Modeled kWh per Day	Residual	90% Confidence Half-Interval	Minimum Modeled kWh per Day	Maximum Modeled kWh per Day	Actual kWh	Modeled kWh
3.7	202.7	165.5	37.2	24.3	141.2	189.8	6,080	4,964
10.4	244.3	224.2	20.1	16.7	207.5	241.0	7,330	6,727
16.5	266.8	278.8	-12.0	16.4	262.4	295.3	7,470	7,807
23.9	357.1	344.1	13.1	24.7	319.4	368.8	10,000	9,634
25.1	328.0	354.4	-26.4	26.5	327.9	380.8	11,480	12,403
23.7	346.1	342.5	3.6	24.5	318.0	366.9	11,420	11,301
18.7	332.3	297.7	34.6	18.2	279.6	315.9	9,970	8,932
19.4	270.3	303.9	-33.6	18.9	285.1	322.8	7,840	8,814
10.2	261.5	222.4	39.1	16.9	205.6	239.3	6,800	5,783
8.4	186.9	206.6	-19.7	18.4	188.2	225.1	5,980	6,612
2.3	134.7	152.6	-17.9	26.5	126.1	179.0	4,310	4,883
1.9	111.0	149.0	-38.0	27.1	121.9	176.1	3,330	4,469
<b>Total</b>	<b>3,041.8</b>	<b>3,041.8</b>	<b>0.0</b>	<b>258.9</b>	<b>2,782.8</b>	<b>3,300.7</b>	<b>92,010</b>	<b>92,331</b>

So what is the bias in the model?

■ **Net Determination Bias Error (NBE):** 
$$NBE = 100 * \frac{\sum_i (E_i - \hat{E}_i)}{\sum_i E_i}$$

$$NBE = (92,331 - 92,010) / 92,331$$

$$NBE = 0.3\%$$

The model predicts 0.3% higher energy use than the actual data.

ASHRAE Guideline 14 does not accept a model with bias >0.005%, so this model would be rejected. However, the uncertainty in the model is much, much greater than the bias, and the savings are expected to be much, much greater than the uncertainty. Thus, this model is actually acceptable:

■ **Modeled Uncertainty** =  $\pm (92,331 - 84,436) / 92,331 = \pm 8.6\%$ .

The expected energy savings for this measure is at least 45%. Since the uncertainty is low relative to the expected savings, this baseline model would be acceptable for projecting energy use under post-implementation conditions and could be used in the calculation of energy savings.

## 6.2. Background on Heating and Cooling Degree-Days (HDD and CDD)

Heating degree-days are a measure of how much cold weather there is in a specific period. The average daily temperature is determined for each day. The average temperature is then compared to a *base* temperature (often 65° F). If the average temperature (when only daily data are available, typically the average of the daily high and the daily low) is 55° F for a day, and the base is 65° F, then that day contributes 10 HDD to the period. The HDD for each day in the period (typically a calendar month or a utility billing period) are summed to create a single data point for the month. If the temperature difference for a day is negative, it is recorded as 0.

$$\blacksquare \text{HDD}_n = \sum_i^n (T_{base} - T_i)^+$$

Note that while HDD and CDD are often reported with a base or balance point of 65° F, results can often be improved by experimenting with different base temperatures. The base temperature should generally be the average temperature at which the building does not require any heating or cooling – the balance point temperature. For most commercial buildings, this temperature will typically be between 55° and 60° F, depending on building size, operating schedule, and other parameters. If regression models are created separately for occupied and unoccupied periods, the balance point temperature will be different for each: for the occupied period, it may be near 55° F, and for the unoccupied period it may be near 65° F.



## 7. Minimum Reporting Requirements

This document is a reference guide, a companion to the M&V protocols. Below are the minimum reporting requirements for the use of regressions within protocols. The overall M&V approach should be described according to the minimum reporting requirements of the protocol used. Please see the protocols for minimum reporting requirements.

These are the essential reporting requirements for regressions within an M&V plan and verification report:

- ➔ **Data:** variables, interval of observation – such as monthly, number of observations, or length of measurement period
- ➔ **Model:** the proposed or final model and alternative models proposed or tested (the verification report should include estimated model parameters)
- ➔ **Model Statistics:** statistics for assessing goodness of fit (proposed and, in the verification report, calculated statistics for final model)
- ➔ **Discussion:** why the final model was selected or weaknesses of the alternative models tested



## 8. References and Resources

- ASHRAE. 2002. *ASHRAE Guideline 14-2002 – Measurement of Energy and Demand Savings*. Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.  
Purchase at: [http://www.techstreet.com/cgi-bin/detail?product\\_id=1645226](http://www.techstreet.com/cgi-bin/detail?product_id=1645226).
- ASHRAE. 2004. *ASHRAE 1050-RP – Inverse Modeling Toolkit*.  
See: Kissock, J., J. Haberl, and D. Claridge. *Inverse Modeling Toolkit: Numerical Algorithms*.
- California Commissioning Collaborative. 2008. *Guidelines for Verifying Existing Building Commissioning Project Savings, Using Interval Data Energy Models: IPMVP Options B and C*. 2008. California Commissioning Collaborative.  
Available at: [http://resources.cacx.org/library/holdings/VoS%20Guide%20111308\\_final.pdf](http://resources.cacx.org/library/holdings/VoS%20Guide%20111308_final.pdf).
- IPMVP. 2010. *International Performance Measurement and Verification Protocol Volume 1: Concepts and Options for Determining Energy and Water Savings*. EVO 10000 – 1:2010. Washington, D.C.: Efficiency Valuation Organization.  
Available at: [http://www.evo-world.org/index.php?option=com\\_form&form\\_id=38](http://www.evo-world.org/index.php?option=com_form&form_id=38).
- Haberl, J., A. Sressthaputra, D. Claridge, and J. Kissock. 2003. *Inverse Model Toolkit: Application and Testing*. KC-03-02-2 (RP-1050). Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.  
Purchase at: [http://www.techstreet.com/cgi-bin/detail?product\\_id=1717581](http://www.techstreet.com/cgi-bin/detail?product_id=1717581).
- Hardy, Melissa A. 1993. *Regression with Dummy Variables*. Newbury Park, Calif.: Sage.
- Kissock, J., J. Haberl, and D. Claridge. 2004. *RP-1050 – Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models*. Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.  
Purchase at: [http://www.techstreet.com/cgi-bin/detail?product\\_id=1717813](http://www.techstreet.com/cgi-bin/detail?product_id=1717813).
- Kissock, J., J. Haberl, and D. Claridge. 2004. *Inverse Modeling Toolkit: Numerical Algorithms*. KC-03-02-1 (RP-1050). Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.  
Purchase at: [http://www.techstreet.com/cgi-bin/detail?product\\_id=1717580](http://www.techstreet.com/cgi-bin/detail?product_id=1717580).
- NIST/SEMATECH. 2011. *Engineering Statistics Handbook (NIST/SEMATECH e-Handbook of Statistical Methods)*. Gaithersburg, Md.: National Institute of Standards and Technology.  
Available at: <http://www.itl.nist.gov/div898/handbook/index.htm>.
- Stevens, James. 2002. *Applied Multivariate Statistics for the Social Sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.



## Appendix: Glossary of Statistical Terms

This Glossary provides definitions for the statistical terms used in this *Regression Reference Guide*. Additional M&V terms are defined in the companion document *Glossary for M&V: Reference Guide*.

**Accuracy:** An indication of how close the measured value is to the true value of the quantity in question. Accuracy is not the same as precision.

**Adjusted R-square ( $\bar{R}^2$ ):** A modification of  $R^2$  that adjusts for the number of independent variables (explanatory terms) in a model. The adjusted  $R^2$  only increases if the additional independent variables improve the model more than by random chance. It is calculated by taking  $R^2$  and dividing it by the associated degrees of freedom. Or as described below:

$$\bar{R}^2 = 1 - \frac{MSE}{MST}$$

**Autocollinearity:** The serial correlation over time of predictor values in a time series model. To calculate autocollinearity, R-squared is first calculated for the correlation between the residuals and the residuals for the prior time period. The autocorrelation coefficient  $\rho$  is then the square root of this value.<sup>5</sup> Autocollinearity is calculated as:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

**Categorical Variables:** Variables that have discrete values and are not continuous. Categorical variables include things like daytype (weekday or weekend, or day of week), occupancy (occupied or unoccupied), and equipment status (on or off). For example, occupancy (occupied or unoccupied) is a categorical variable, while number of occupants is a continuous variable.

**Coefficient of Variation (CV):** An indication of how much variability or randomness there is with any given data set. It quantifies variation within the population relative to the average and is dimensionless. The larger it is, the more variation there is in the population relative to the average. It is calculated as the ratio of the standard deviation to the average:

$$CV = \frac{\sigma}{x}$$

---

<sup>5</sup> Note, the English spelling of the Greek letter  $\rho$  is *rho*, not to be confused with “p.”

**Coefficient of Variation of the Root-Mean Squared Error [CV(RMSE)]:** A measure that describes how much variation or randomness there is between the data and the model, calculated by dividing the root-mean squared error (RMSE) by the average y-value. It is calculated as:

$$CV(RMSE) = \frac{1}{y} \left[ \frac{\sum (y_i - \hat{y})^2}{(n-p)} \right]^{1/2}$$

**Confidence Interval:** A range of uncertainty expected to contain the true value within a specified probability. The probability is referred to as the *confidence level*.

**Confidence Level:** A population parameter used to indicate the reliability of a statistical estimate. The confidence interval expresses the assurance (probability) that given correct model selection, the true value of interest resides within the proportion expressed by the confidence interval.

**Continuous Variables:** Variables that are numeric and can have any value within the range of encountered data (i.e., measurable things such as energy usage or ambient temperature).

**Dependent Variable:** The variable that changes in relationship to alterations of the independent variable. In energy efficiency, energy usage is typically treated as the dependent variable, responsive to the manipulation of conditions (independent variables).

**Homoscedasticity:** (Also known as *Homogeneity of Variance*.) Within linear regression, this means that the variance of the dependent values around the regression line is constant for all values of the independent variable.

**Independent Variable:** Also termed an *explanatory* or *exogenous variable*; a factor that is expected to have a measurable impact on the dependent, or outcome variable (e.g., energy use of a system or facility).

**Mean:** The most widely used measure of the central tendency of a series of observations. The Mean ( $\bar{Y}$ ) is determined by summing the individual observations ( $Y_i$ ) and dividing by the total number of observations ( $n$ ), as follows:

$$\bar{Y} = \frac{1}{n} \sum Y_i$$

**Mean Bias Error (MBE):** (Also known as the *Normalized Mean Bias Error* or the *Net Determination Bias Test*.) The Mean Bias Error is an indication of overall bias in a regression model. Positive MBE indicates that regression estimates tend to overstate the actual values. It is calculated as:

$$NMBE = \frac{1}{y} \sum (y_i - \hat{y}_i) / (n-p)$$

**Mean Model:** (Also known as a *Single Parameter Model*.) A model that estimates the mean of the dependent variable.

**Multicollinearity:** A statistical occurrence where two or more predictor variables in a multiple regression model are highly correlated (there are exact linear relationships between two or more explanatory variables). Allowing multicollinearity in a model can lead to incorrect inferences from the model.

**Net Bias:** Where there exists net bias, modeled or predicted energy usage will differ from actual energy usage for the period examined.

**Net Determination Bias Error:** The percentage error in the energy use predicted by the model compared to the actual energy use. See *Normalized Mean Bias Error*.

$$NBE = 100 * \frac{\sum_i (E_i - \hat{E}_i)}{\sum_i E_i}$$

**Net Determination Bias Test:** The savings resulting from applying the baseline period's independent variable data to the algorithms for savings determination. The data so applied must reflect all exclusions or adjustments to actual measured data as documented for the baseline model. See *Mean Bias Error*.

**Normal Distribution:** A continuous and symmetric population distribution in which the frequency of occurrence decreases exponentially as values deviate from the mean (or central) value. In a regression equation, the distribution of errors (residuals) at a given value of  $x$  is a normal distribution and the mean of residuals is zero. It is also referred to as a *Gaussian* or *bell curve*.

**Normalized Mean Bias Error (MBE):** (Also known as the *Mean Bias Error*.) Similar to *Net Determination Bias*, but adjusted for the number of parameters in the model. The Normalized Mean Bias Error is an indication of overall bias in a regression model. Positive MBE indicates that regression estimates tend to overstate the actual values. It is calculated as:

$$NMBE = 100 * \frac{\sum_i (E_i - \hat{E}_i)}{(n - p) * \bar{E}}$$

**Outliers:** Data points that do not conform to the typical distribution. Graphically, an outlier appears to deviate markedly from other members of the same sample.

**Overspecified Model:** A model with added independent variables that are not statistically significant or are possibly correlated with other independent variables.

**$p$ -value:** The probability that a coefficient or dependent variable is not related to the independent variable. Small  $p$ -values, then, indicate that the independent variable or coefficient is a significant (important) predictor of the dependent variable in a regression model. The  $p$ -value is an alternate way of evaluating the  $t$ -statistic for the significance of a regression coefficient, and is expressed as a probability.

**Precision:** The indication of the closeness of agreement among repeated measurements; a measure of the repeatability of a process. Any precision statement about a measured value must include a confidence level. A precision of 10% at 90% confidence means that we are 90% certain the measured values are drawn from samples that represent the population and that the “true” value is within  $\pm 10\%$  of the measured value. Because precision does not account for bias or instrumentation error, it is an indicator of predicted accuracy only given the proper design of a study or experiment.

**R-Squared ( $R^2$ ):** (Also known as the *Coefficient of Determination*.)  $R^2$  is the measure of how well future outcomes are likely to be predicted by the model. It illustrates how well the independent variables explain variation in the dependent variable.  $R^2$  values range from 0 (indicating none of the variation in the dependent variable is associated with variation in any of the independent variables) to 1 (indicating all of the variation in the dependent variable is associated with variation in the independent variables, a “perfect fit” of the regression line to the data). It is calculated as:

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}}$$

**Regression Analysis:** A mathematical technique that extracts parameters from a set of data to describe the correlation relationship of measured independent variables and dependent variables.

**Regression Model:** A mathematical model based on statistical analysis where the dependent variable is regressed on the independent variables which are said to determine its value. In so doing, the relationship between the variables is estimated from the data used. A simple linear regression is calculated as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where } i = 1, \dots, n$$

**Reliability:** When used in energy evaluation, refers to the likelihood that the observations can be replicated.

**Residual:** The difference between the predicted and actual value of the dependent variable. In other words, whether a point is above or below the regression line is a matter of chance and is not influenced by whether another point is above or below the line. Estimated by subtracting the data from the sample mean:

$$\hat{\varepsilon} = X_i - \bar{X}$$

**Root Mean Squared Error (RMSE):** (Also known as the *Standard Error of the Estimate*.) An indicator of the scatter, or random variability, in the data, and hence is an average of how much an actual  $y$ -value differs from the predicted  $y$ -value. It is the standard deviation of errors of prediction about the regression line. The RMSE is calculated as:

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)}$$

**Standard Deviation (s):** The square root of the variance, which brings the variability measure back to the units of the data. (With variance units in kWh<sup>2</sup>, the standard deviation units are kWh.) The sample standard deviation (*s*) is calculated as:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$$

**Standard Error (SE):** An estimate of the standard deviation of the coefficient. For simple linear regression, it is calculated separately for the slope and intercept: there is a *standard error of the intercept* and *standard error of the slope*. SE is calculated as:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

**Standard Error of the Coefficient:** Similar to the *standard error of the estimate*, but calculated for a single coefficient rather than the complete model.

**Standard Error of the Estimate:** (Also known as the *Root Mean Squared Error*.) When a model is used to predict a value for given independent variable(s), the reliability of the prediction is measured by the standard error of the estimate. The Root Mean Squared Error (RMSE) is calculated as:

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)}$$

**t-statistic:** A measure of the probability that the value (or difference between two values) is statistically valid. The calculated *t*-statistic can be compared to critical *t*-values from a *t*-table. The *t*-statistic is inversely related to the *p*-value; a high *t*-statistic (*t*>2) indicates a low probability that random chance has introduced an erroneous result. Within regression, the *t*-statistic is a measure of the significance for each coefficient (and, therefore, of each independent variable) in the model. The larger the *t*-statistic, the more significant the coefficient is to estimating the dependent variable. The *t*-statistic is calculated as:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

**Uncertainty:** The range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall within some stated degree of confidence. Uncertainty in regression analysis can come from multiple sources, including *measurement uncertainty* and *regression uncertainty*.

**Variance (S<sup>2</sup>):** A measure of the average distance between each of a set of data points and their mean value, and it is equal to the sum of the squares of the deviation from the mean value, or the square of the standard deviation. Variance is computed as follows:

$$S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n - 1}$$

**Weighted Regression:** A form of regression used when individual data points are weighted so as to represent more data than other points. An example is billing-period analysis, where billing periods may have different numbers of days and billing periods with more days are adjusted upward in weight relative to periods with fewer days. (Also, a form of regression used when data do not have equal weight in a model because error is not expected to be constant across all observations.)

## Sources:

- ASHRAE. 2002. *ASHRAE Guideline 14-2002 – Measurement of Energy and Demand Savings*. Atlanta, Ga.: American Society of Heating, Ventilating, and Air Conditioning Engineers.
- Brase, Charles Henry, and Corrinne Pellillo Brase. 2009. *Understandable Statistics: Concepts and Methods* (9th Ed.). New York, N.Y.: Houghton Mifflin.
- Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol Volume 1: Concepts and Options for Determining Energy and Water Savings*. (EVO 10000 – 1:2010), September 2010. Washington, D.C.: Efficiency Valuation Organization.
- Stevens, James. 2002. *Applied Multivariate Statistics for the Social Sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- TecMarket Works Team. *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. April 2006. San Francisco, Calif.: California Public Utilities Commission.
- Thompson, Steven K. 2002. *Sampling*. (2nd Ed.). New York, N.Y.: John Wiley & Sons, Inc.



