Statistics and Reality—Addressing the Inherent Flaws of Statistical Methods Used in Measurement and Verification

John Avina, CEM, CEA, CMVP, CxA

ABSTRACT

When we use regressions for Option C M&V, we use statistical indicators, such as R^2 and CV(RMSE) to determine whether the regression is valid and whether the regression should be used to determine energy and demand savings. Although introduced decades ago, fractional savings uncertainty (FSU), a more complex statistical indicator, is finally becoming known among M&V practitioners. All of these statistical indicators used to qualify Option C regressions are human creations and are not based in reality. This article explains how these indicators are inconsistent, unscientific, arbitrary, and often not well-suited for Option C M&V. The industry needs rules that are simple and that will work for all regressions. In this article, I have presented a set of rules that I believe are understandable to practitioners and will avoid the drawbacks of the current statistical indicator thresholds as recommended by ASHRAE and EVO.

STATISTICS AND REALITY

Coming from an energy engineering point of view, I find statistics to be difficult. The math is not too difficult and in a rudimentary sense, the concepts are not difficult to understand. But the more advanced concepts are confounding to an engineer whose way of thinking is based on what is observable in the material world.

Definitions of Terms Used

Abbreviation*	Definition
ASHRAE	American Society of Heating, Refrigeration, and Air Conditioning Engineers (www.ashrae.org)
CDD	Cooling degree-days
CV(RMSE)	Coefficient of variation, root-mean squared error (of the regression model)
ESCO	Energy services company
EVO	Efficiency Valuation Organization (www.evo-world.org) (EVO is an international group that publishes the IPMVP and other M&V guidance)
F	Fractional savings or savings fraction, defined as (Savings) / (Baseline energy use)
FSU	Fractional savings uncertainty
HDD	Heating degree-days
IPMVP	International Performance Measurement and Verification Protocol (available at www.evo-world.org)
M&V	Measurement and verification (which describes the methods used to verify and estimate energy savings from energy savings retrofits)
Option C	A whole-building M&V approach, which typically uses utility bills or metering information, and linear regressions to establish a relationship between usage and weather conditions
R^2	Correlation coefficient of the regression

* Many of these terms were taken from Stetz [1].

Both statistics and engineering use the language of math to describe reality. But the big difference between statistics and engineering is that engineers use mathematics to describe physical phenomena that can be observed and measured. We can compare our calculated results to these real phenomena, whereas statisticians use mathematics to describe concepts. For many of these concepts, there is no reality out there with which to compare our statistical calculations. In engineering, we do believe we are correct in our mathematic equations and calculations when they match physical reality. In statistics, there is often no such physical reality to compare to, only a conceptual reality.

Oh, we can weigh 100 apples and find an average weight, a high weight, a low weight, even a standard deviation. These statistical concepts are based in reality. But as we go further and further afield, statistics gets more and more conceptual and further and further from the real world.

Certified Measurement and Verification Professional

Get CMVP® Certified Today

A Certified Measurement & Verification Professional (CMVP®) is an individual who measures and verifies energy usage and energy requirements throughout a building or across multiple facilities. They develop metrics so that investment in energy, water, demand management, retrofit, and renewable energy projects can be evaluated, and prioritized. A CMVP® can often help a company realize substantial savings.

The certification program is of greatest value to those undertaking or assessing M&V energy projects, such as existing M&V professionals, energy engineers, energy managers, energy analysts, financial executives, manufacturing and facilities managers, and energy consultants. Obtaining AEE's CMVP® certification provides international credibility among energy management and M&V communities.

Boost **Your Career**

Get Certified Today...

aeecenter.org/cmvp

IN CONJUNCTION WITH

Efficiency Valuation Organization

Association of Energy Engineers | 3168 Mercer University Drive | Atlanta, Georgia 30341 | (770) 447-5083

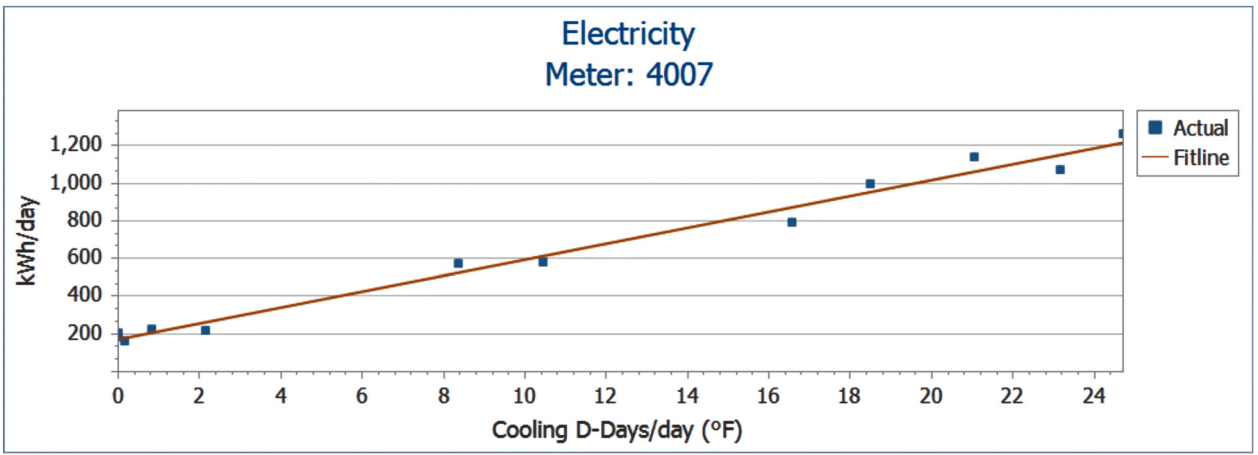
Find me a Newtonian physics equation and through experimentation engineers can tell you whether the equation accurately represents whatever entity it is supposed to represent. But in statistics, we are stuck with these conceptual constructs that exist only in our minds. For example, CV(RMSE) and R^2 are two statistical indicators that can be used to gauge the goodness of a linear regression to a set of points. Figure 1 presents such a fit, and the CV(RMSE) and R^2 are calculated.

R^2 VALUE AND CV(RMSE)

The R^2 value—also known as the coefficient of determination—indicates the proportion of the variance in the dependent variable that is a function of the independent variable. A good regression model will have a high R^2 value. The CV(RMSE) is defined as a measure of all other variation in the dependent variable. Often, the CV(RMSE) is called a measure of randomness or scatter. A good regression model will have little scatter, and thus a small CV(RMSE). You may think then, because the R^2 is expressed on a 0 to 1 scale, that the CV(RMSE) would also be on a 0 to 1 scale, and that the two values would add to 1. That is, if the R^2 is 0.81, then 81% of the variation of the dependent variable can be ascribed to variations in the independent variable. It should follow, that all other variation, or scatter, should be 19%, or 0.19, but the CV(RMSE) value is not 19% or 0.19, and could be any number between 0 to infinity. The CV(RMSE) isn't even measured on the same scale as R^2! Why is that?

So why don't the two measures add up to one when their lay verbal definitions imply they should? If we define CV(RMSE) as an indicator of all the variation in the dependent variable that is not related to the independent variable, then it seems that the two measures should be related mathematically to each other.

Most likely, the answer is that in trying to simplify these statistical concepts, statisticians have generalized them so that they can be vaguely understandable to the layperson (like myself). However, the generalization has leached out the accuracy of the explanations. If the generalizations were accurate, then the R^2 value and the CV(RMSE) would together sum to unity. And not only that, as the R^2 value drops, the CV(RMSE) should increase. But that doesn't always happen, as you will see below.



kWh=170.497×#Days + 42.428×CDD
R2 = 0.985
CVRMSE = 8.44%

Figure 1. Sample Data from Electric Meter Illustrating Fit of Linear Regression

Let's take a closer look at the R^2 value. Linear regressions that have steep fits have higher R^2 values than linear regressions with flatter slopes—even if the scatter is exactly the same. Figure 2 shows linear fit regressions for two sets of 12-month bill data: each data set has the same mean value, and the same CV(RMSE). The only difference between the data sets is that the orange data set has a steeper linear fit slope, and consequently has a higher R^2 value.

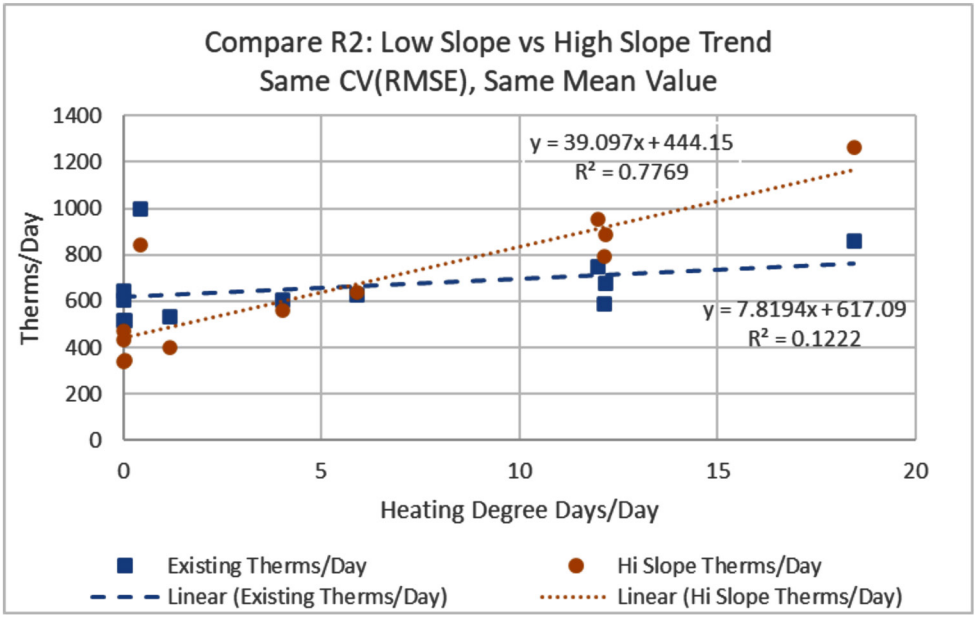


Figure 2. Linear Fit Regressions for Two Sets of Billing Data

The difference in R^2 values in the example is the natural result of the definition of R^2, which, as I said before, represents the proportion of the variance in the dependent variable that is a function of the independent variable. If the slope is flat, then the dependent variable is not varying by that much due to changes in the independent variable. And if there is the same random scatter in both data sets, then the proportion of the variance due to the independent variable in the flat slope is less, and the proportion due to the randomness is more. This definition of R^2 works.

So then, perhaps we shouldn't call the R^2 value an indicator of the "goodness of fit," because "goodness of fit" has nothing to do with

slope. In addition, because both linear fits in Figure 2 have the same random scatter, they should be considered equally valid models to predict energy usage. So, the R^2 value really has little to do with how well the model predicts the actual values either.

In addition, we should not be using the R^2 value to determine whether a fit is acceptable or not. It is not a fair measure, as areas with little variation in weather have a small likelihood of passing the R^2 criteria. Honolulu and San Francisco are two places that come to mind where a low R^2 value may not indicate a poor model. The point is that the R^2 value is an imperfect construct created by a pair of academics in the 1920s that is not based on a law of nature or mathematics. For decades, energy efficiency professionals have been taught the R^2 value as an indicator of "goodness of fit" as if it were a law of nature not to be questioned. It is not.* Agami Reddy and David Claridge made this clear in 2000[2], and the ASHRAE Guideline 14 committee took this information to heart when they did not include the R^2 value in the ASHRAE Guideline 14 in 2002 and again in 2014[3].

On the other hand, we shouldn't be using the CV(RMSE) to determine whether a fit is acceptable or not either. Whereas the R^2 value is comparing points to the slope of the fit, the CV(RMSE) compares points to the average bill. In Figure 3, we have two sets of points; the only difference is that the intercept (3085.5) associated with the data set of orange dots is 500% higher than the intercept (617.09) of the data set signified by the blue squares.

The CV(RMSE) for the lower line (signified by squares) is 21%, while it is 5% for the higher line. You can see that the distance from the points to the fit line is the same. The R^2 is the same for both. The CV(RMSE) increases when the average bill drops.

So then, R^2 value is low for fits with low slopes, and the CV(RMSE) is high when the average bill is lower. Neither indicator determines the quality of the fit for all conditions. Really, neither indicator should be used to determine whether a fit is acceptable.

---

*I admit, I was one of those teachers, who, in trying to simplify the concept of R^2 called it "goodness of fit."

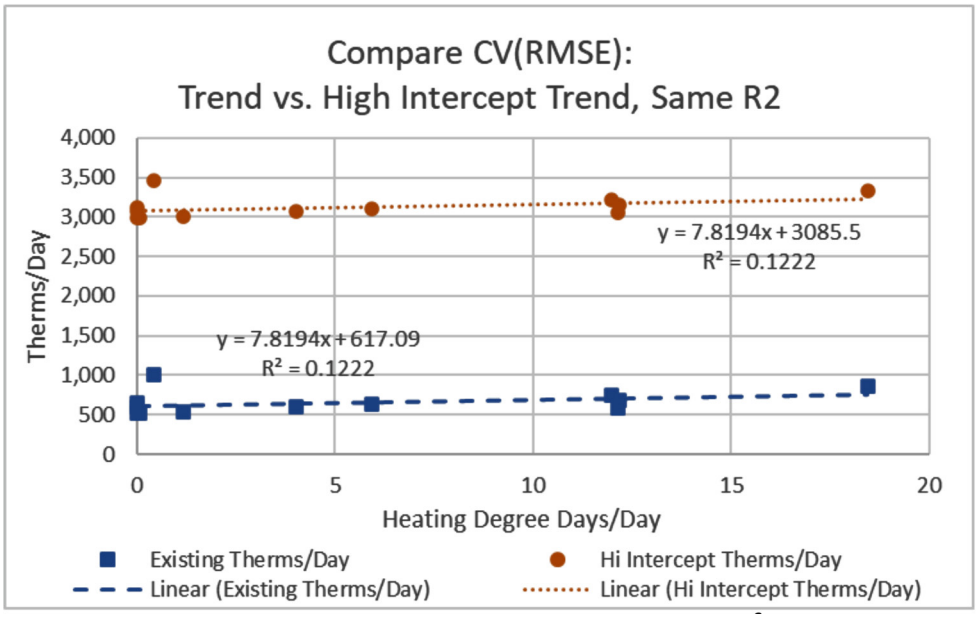


Figure 3. Comparison of Two Trend Lines with the Same R^2

A CLOSER LOOK AT THE DIFFERENCE BETWEEN THE R^2 AND CV(RMSE) EQUATIONS

Let's take a closer look at the data in Figure 2. The data are from 12 months of consecutive utility bills. Each data point represents a month's therms/day as a function of the heating degree days (HDD)/day.*

This difference in R^2 values in the two plots in the graph is due to the denominator of the R^2 term (see Equation 1), which represents how far the individual points are from the average point. If the points are clumped together to form a flat slope, then the points are close to the average point, and the denominator would be lower, the fraction is therefore higher, and finally, the R^2 value is lower.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \qquad \text{(Eq 1)}$$

*If we merely plotted therms vs. HDD, a 25-day bill would carry just as much weight as a 35-day bill. We divide by number of days to remove this bias.

where:
y_i represents the actual bill
$\hat{y}_i$ represents what the fit line estimates the bill to be
$\bar{y}_i$ represents the average bill in the base year
and can translated into lay terms as:

$$R^2 = 1 - \frac{\sum(difference\ between\ each\ bill\ and\ fitline)^2}{\sum(difference\ between\ each\ bill\ and\ average\ bill)^2} \quad \text{(Eq 2)}$$

Or written another way even simpler:

$$R^2 = 1 - \text{Scatter \%} \quad \text{(Eq 3)}$$

What is interesting (to the non-statistician like myself) is that the R^2 value is defined in the equation as everything except the scatter.

Let's take a look at the "Scatter %" term.

The numerator in Equation 2, $\sum$(difference from each bill to fit line)^2, is the "scatter" in absolute terms (in other words, it is not a percentage, but the number of kWh^2, therms^2, or the square of whatever unit we are dealing with). If the model were perfect, each month's bill would lie on the fit line, the numerator would be 0 (i.e., the "scatter" term would be 0), and the R^2 value would be 1.

The denominator makes the scatter in the numerator into a percentage of the total scatter, so that the R^2 value, which is 1 minus this percentage, will always be between 0 and 1. But why do they use the difference between each monthly bill and the yearly average in the denominator? Couldn't they have used the difference between each monthly bill and the model's prediction of that monthly bill? I know there is a reason behind the definition. I am just not aware of it.

Remember, CV(RMSE) is also defined as scatter. The definition of CV(RMSE) is:

$$CV(RMSE) = \frac{1}{\bar{y}}\left[\frac{\sum(y_i - \hat{y}_i)^2}{n-p}\right]^{0.5} \quad \text{(Eq 4)}$$

where
n is number of bills
p is number of independent variables + 1
y_i represents the actual bill
$\hat{y}_i$ represents what the fit line estimates the month's bill to be
$\bar{y}_i$ represents the average bill in the base year

Equation 4 can be rearranged:

$$CV(RMSE) = \frac{1}{\sqrt{n-p}} \frac{\sqrt{\sum(y_i - \hat{y}_i)^2}}{\bar{y}} \quad \text{(Eq 5)}$$

and translated into lay terms as:

$$CV(RMSE) = \frac{1}{\sqrt{n-p}} \frac{\sqrt{\sum(difference\ between\ each\ bill\ and\ fit\ line)^2}}{average\ bill} \quad \text{(Eq 6)}$$

In the format of equations 5 and 6, the CV(RMSE) looks similar to the R^2 value equation. Let's compare the two equations: Equation 6 above, and equation 2 repeated below:

$$R^2 = 1 - \frac{\sum(difference\ between\ each\ bill\ and\ fit\ line)^2}{\sum(difference\ between\ each\ bill\ and\ average\ bill)^2} \quad \text{(Eq 2)}$$

The numerators for R^2 and CV(RMSE) are nearly the same. In fact, they only differ by the square root of the summation term. It is the denominators that are different. The CV(RMSE) uses the average bill in the denominator, whereas the denominator of the R^2 value is based on the difference from each bill to the average bill. (You have to divide by something to get a percentage, and the designers of these two indicators unfortunately used different denominators.)

Essentially, we are using scatter in two ways. The R^2 value is the anti-scatter, R^2 = 1- Scatter, while the CV(RMSE) = Scatter.

And to confound it all, the problem is that the definitions of scatter are different. These different definitions of scatter explain why the R^2 value, and the CV(RMSE) do not add up to one. So why is the CV(RMSE) using the average bill, and the R^2 using the difference from each bill to the average bill? Is there a good reason for this difference? I doubt it. My guess is that the CV(RMSE) had a different independent evolution from the R^2 value. The CV(RMSE) evolved from the coefficient of variation, the CV, an old concept. As the CV was then extended from applying to a single number to a regression over a series of numbers, it became the CV(RMSE).

AN EXAMPLE GREAT FIT WITH A LOW R^2 VALUE

The meter data in Figure 4 are from a department store in Daly City, California. For those who have not visited Daly City, I can tell you it is nearly always foggy there, and the weather is uniformly dreary, every month of the year. The model for kWh vs. CDD shows a very flat fit, with a consequent low R^2 value of 0.06. But the CV(RMSE) is fantastic, at 3%. In spite of the poor R^2 value, this is a fantastic fit. So, what should we do? Is this an acceptable regression?

Old school ESCO thinking would be to throw the model out—don't use it. The IPMVP states that fits should have R^2 values above 0.75, and this model is clearly lower. Most ESCOs follow the IPMVP. ASHRAE Guideline 14 does not mention R^2 value and instead requires that fits have low CV(RMSE)s. According to the IPMVP recommendation of 0.75, then, this is an unacceptable fit. According to ASHRAE Guideline 14, it is stellar. So, what should we do with this model? The regression model appears very accurate at predicting the bills. Why not use it?

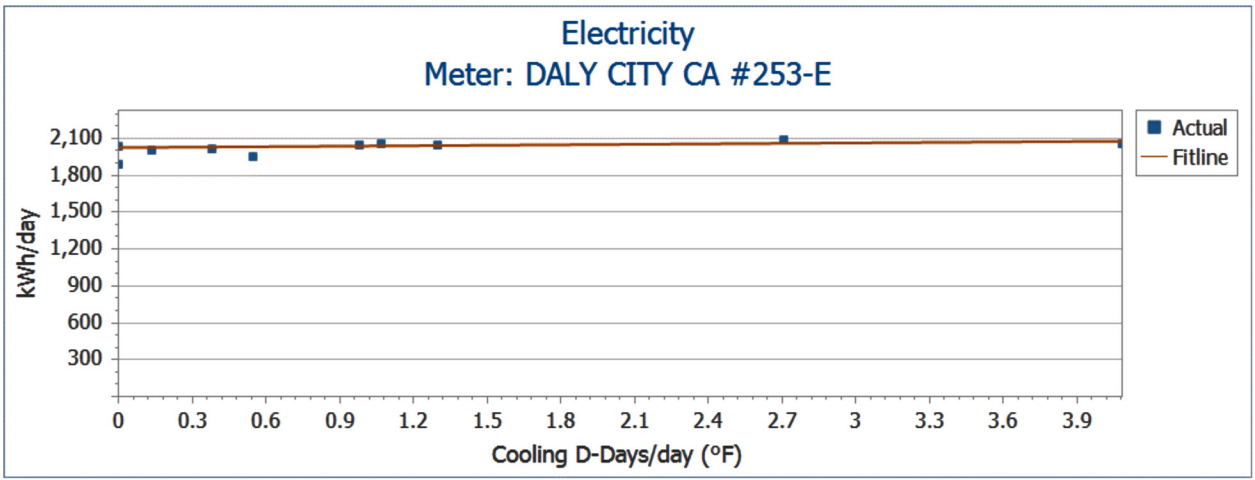
PERHAPS A SOLUTION TO THE CV(RMSE) AND R^2 INCONSISTENCY

Perhaps a solution to this jumble of the inconsistency of R^2 and CV(RMSE) would be to at least make them consistent. Why not have the CV(RMSE), or my proposed replacement of it, be associated with the R^2 value. The R^2 value compares deviation to the slope. Why not have the replacement for the CV(RMSE) also compare deviation to the slope. An easy equation would be:

$$\text{Proposed Measure of Scatter} = 1 - R^2 \qquad \text{(Eq 7)}$$

This way, if an R^2 value was 0.8, we are saying that 80% of the deviation of the dependent variable is associated with the independent variable. And therefore, the Scatter = 1 − 0.8 = 0.2, which means that 20% of the fluctuation of the independent variable we cannot account for. Doesn't that seem more elegant?

The problem with this simplification, as I mentioned already, is that R^2 values are low for flatter slopes and high for steeper slopes. Although



12 bills spanning Mar 02, 2004 to Mar 01, 2005
Weather location: SAN FRANCISCO, CA, USA
CDD balance point = 64F°
kWh = 2,024.012×#Days + 12.247×CDD
CVRMSE = 3.04%
R2 = 0.060
SE = 1,879

Figure 4. Sample Linear Fit Regression with a Good CV(RMSE) but Low R^2

this follows the definition of R^2, as "to what degree do changes in the independent variable correspond to changes in the dependent variable," my definition of scatter still holds. For flat slopes, we cannot account for most of the variation; therefore, for the Daly City meter above, the 0.06 R^2 value, would yield a scatter of 94%. Just like we cannot use R^2, or CV(RMSE) to determine whether a fit is acceptable to use in a model, we would not be able to use my proposed scatter term either. It is a good fit, and my scatter is 94%, which appears, on the surface, way too high.

But at least R^2 and Scatter terms would be consistent. We would be able to say: "6% of the variation in kWh/day is associated with changes in CDD/day and 94% of the variation in kWh/day, well, we don't know where that comes from." Isn't that more elegant than saying, the fit has an R^2 value of 0.06 and a CV(RMSE) of 3.04%?

CV(RMSE) AND SAVINGS FRACTION

The IPMVP Core Concepts module states that the savings fraction should be more than two times the CV(RMSE). The savings fraction, F, is the percentage of energy you expect to save from the total bill. In other words, if you expect to save 40% of energy usage, then the CV(RMSE) should be less than 20%. Two questions come to mind. Why two times and not three times, or one and half times? Where did that come from?

The second question is, why does this matter in the first place? If you want to see savings over a month or a few months, sure, this makes sense. You want the savings to be larger than the CV(RMSE). Then you can be somewhat certain that the difference in usage is due to your retrofits and not due to the noise the CV(RMSE) is indicating. But over the course of a year, it all evens out. Because the model has a net mean bias error of 0, for an entire year of data the model is not biased to show more or less savings than there really is. For example, suppose you have a model with a good fit. The CV(RMSE) is 18%, and your energy projects are supposed to save 20% of the total meter's usage. Does it matter that the CV(RMSE) is not below 10%? If you are looking at 1 month of data, it might matter. It may be hard to tell the noise from the savings. But looking at an entire year of post-retrofit data, I don't think it matters at all.

FRACTIONAL SAVINGS UNCERTAINTY

But let's look at uncertainty. Many in the M&V community have been writing papers about uncertainty lately, all trying to forge some measure that can accurately capture the amount of uncertainty we have in our Option C energy savings calculations.* One indicator of uncertainty that has been discussed often is the fractional savings uncertainty (FSU).

Let's examine the concept behind FSU. Suppose you expect to save 5% of the total electricity usage of a building. To determine baseline energy usage, you modeled the electricity usage versus cooling degree days, and used a linear regression model that had a CV(RMSE) of 50%. For any given month, you would never be able to tell whether any month's decrease in energy usage was because of the substantial scatter, as evidenced by the high CV(RMSE), or because of the retrofit that you implemented. One of the points behind FSU is that the smaller the scatter in the data around your model and the larger your savings fraction, the more certain you can be that the savings are a result of the retrofit. FSU also addresses the fact that the more post-retrofit data you have, the less the uncertainty, because there are more points and more savings.

The equation for FSU is:

FSU = (Uncertainty in the Baseline Model)/
(Expected Energy Savings) (Eq 8)

The greater the expected energy savings, the lower the FSU. The higher the uncertainty in the baseline model, the greater the FSU. At first glance, this is entirely reasonable. But it is a little more complicated than that. There is an adjustment for the number of post-retrofit measurements you are planning on taking.

FSU = [(Uncertainty in the Baseline Model) *
(Adjustment for # of Post Retrofit Measurements)]/
(Expected Energy Savings) (Eq 9)

*Never mind that uncertainty rarely comes up in performance contracting discussions (negotiations). There is no need for calculating uncertainty. So why even go there? If practitioners do not calculate uncertainty for Option A, B and D M&V, then why would they do it for Option C? See Avina [4].

More post-retrofit measurements drops the FSU value. There will be less uncertainty associated with a 10-year M&V job than for a 1-year M&V job, that is, assuming that the building energy usage pattern does not change in the performance year. There will be less uncertainty with daily data than with monthly data for an equal time period. This all makes intuitive sense to me.

But how do you put this concept into a mathematical equation? You can give the problem to 100 statisticians, and you will get 100 different equations to represent this concept. And each of the equations will unfairly bias the FSU towards one or another input. One equation may put more significance in the number of post-retrofit measurements, while another in the savings fraction. How would we know which is right? As it stands now, if you use daily or hourly post-retrofit data, instead of monthly or weekly, your FSU is significantly lower. In fact, a model with a poor fit but daily data scores a much better FSU than a model with a good fit and monthly data.

Suppose we are looking at a meter both with daily and monthly data. Suppose:

- Confidence interval = 95%
- M = number of months of post-retrofit tracking is 5 years or 60 months
- P = 2, representing there is one independent variable
- CV(RMSE) = 10%
- F = 25%. This is the amount we expect to save.
- n = n' is the number of pre-retrofit points in the regression. For daily it is 365, for monthly, 12.
- m = number of post-retrofit points. For daily it is 5 years or 1824 points. For monthly it is 60 points.

Looking at the two models, we can calculate FSU using the Sun and Baltazar model, shown in Table 1.

Notice, that for models with identical CV(RMSE), the difference in FSU uncertainty is very large. This difference might persuade someone to use daily data, rather than monthly data.

The monthly model gives a 26.5% FSU. What CV(RMSE) in the daily model would yield that same 26.5% FSU? See Table 2.

Table 1. Summary from the Sun and Baltazar Model

					PRE			POST				
Interval	CVRMSE	p	Confidence Level	F	n	n'_base	n_base	M	m_post	Critical t-Statistic	aM^2 + bM + c	FSU
Monthly	10.0%	2	95%	25%	12	12	12	60	60	2.228	2.132	26.5%
Daily	10.0%	2	95%	25%	365	365	365	60	1824	1.967	2.260	4.2%

Table 2.

					PRE			POST				
Interval	CVRMSE	p	Confidence Level	F	n	n'_base	n_base	M	m_post	Critical t-Statistic	aM^2 + bM + c	FSU
Monthly	10.0%	2	95%	25%	12	12	12	60	60	2.228139	2.13214	26.5%
Daily	63.5%	2	95%	25%	365	365	365	60	1824	1.966521	2.25986	26.5%

To get the same FSU that the monthly model with a CV(RMSE) = 10% gives, I would need a daily model with a 63.5% CV(RMSE), which, quite frankly, is a horrible model. This is what I mean when I say that someone determined that number of post-retrofit points is much more important to FSU than how good the actual regression model is. Another statistician would have come up with a different equation entirely, emphasizing the duration, the number of points and the CV(RMSE) differently.

Like the R^2 value and the CV(RMSE), the FSU is an arbitrary measure. In fact, if we look at the recent history of this indicator, we can see that it is been through several iterations. In 2002 ASHRAE Guideline 14 presented savings uncertainty as:

$$\frac{\Delta E_{save,m}}{E_{save,m}} = t \times \frac{1.26 \cdot CV \left[\frac{n}{n'}\left(1+\frac{2}{n'}\right)\frac{1}{m}\right]^{1/2}}{F} \qquad \text{(Eq 10)*}$$

where

t = students t-statistic
CV is the CV(RMSE)
n = number of points in the baseline period
m = number of points in the post period
p = number of model parameters, i.e., if there is one independent variable, p = 2
n' = the effective number of points after accounting for autocorrelation, which, in this simple analysis, I assume does not happen. So, n' = n.
F = the expected savings fraction.

In 2012, Sun and Baltazar improved the model by replacing the 1.26 factor with a polynomial using M, the number of months of post-retrofit points:

$$\frac{\Delta E_{save,m}}{E_{save,m}} = t \cdot \frac{(aM^2 + bM + c) \cdot CV \left[\frac{n}{n'}\left(1+\frac{2}{n'}\right)\frac{1}{m}\right]^{1/2}}{F} \qquad \text{(Eq 11)}$$

where a, b and c are given as constants.†

*It is not important to look deeply into the equations for the sake of this argument. I only want you to recognize that they keep changing. If you want a good history, along with a deep analysis, SBW Consulting [5] is an excellent reference.

†For monthly data: a = -0.00022, b = 0.03306, c = 0.94054. For Daily Data: a = -0.00024, b = 0.03535, c = 1.00286.

And recently Josh Rushton, took it further:

$$se\left(\hat{y} + \sum \varepsilon_i\right) = \sqrt{n \times s^2 + n \times s^2} = \sqrt{2n} \times s \quad \text{(Eq 12)}$$

These changes in the FSU equation occurred over 15 years. In 100 years, do we think any of these models will be remembered? I believe not. They will be replaced over and over again, with many different models, or perhaps, scuttled completely.

Let's take a closer look at the ASHRAE Guideline 14 and the Sun and Baltazar models. I used the following static inputs:

- Confidence interval = 95%
- M = number of months of post-retrofit tracking (used only in the Sun and Baltazar model), and this typically would be 12 months per year, for however many years of post-retrofit data.
- P = 2, representing there is one independent variable
- CV(RMSE) = 10%
- F = 25%. This is the amount we expect to save.
- n = n' is the number of pre-retrofit points in the regression. For daily it is 365, for monthly, 12.
- m = number of post-retrofit points. For daily it is 365 points per year, while for monthly, it is 12 points per year.

The ASHRAE Guideline 14 model in Figure 5 shows a lowering of FSU the longer the performance period is.

The Sun and Baltazar model shown in Figure 6, using the same inputs, is problematic. At 14.6 years, 175 months, FSU becomes negative, and this is because the squared value in the first polynomial term grows dramatically as the number of months of post-retrofit data, M, increases.* Because it's coefficient, a, is a negative number, you get negative FSU, which doesn't make any sense.

So, there it is, the improvement, perhaps is not an improvement after 175 months of post-retrofit data. According to this equation, it doesn't matter what the CV(RMSE), F or the confidence level is. After 14 years

*At 175 months using the monthly coefficients: $(175)^2 = 30{,}625$. Multiply that result by the negative constant -0.00022, and get -6.7375, which is greater than the other two terms combined, b x M = 0.03306 x 175 = 5.7855, and c = 0.94954.

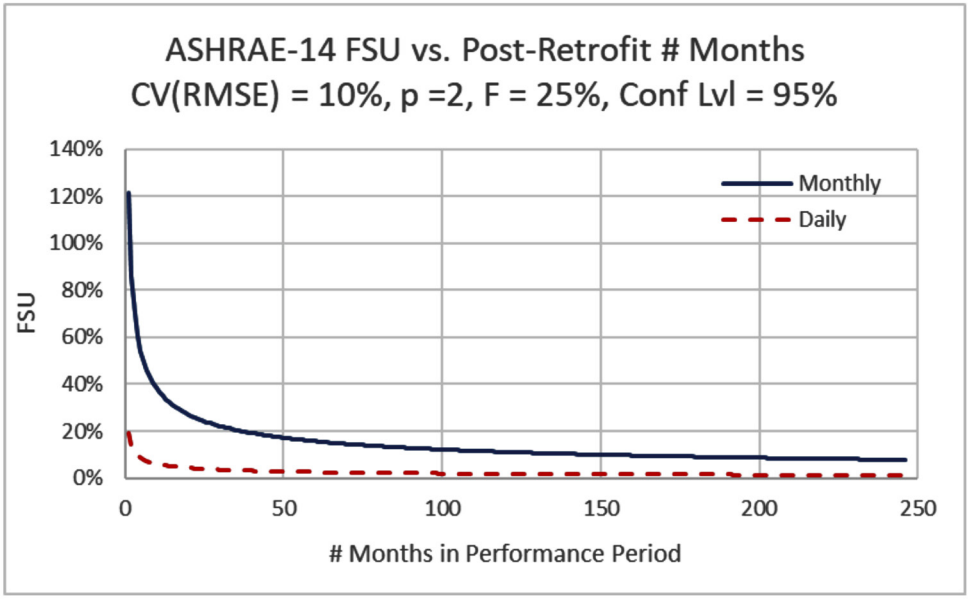


Figure 5. ASHRAE Guideline 14 Model Results

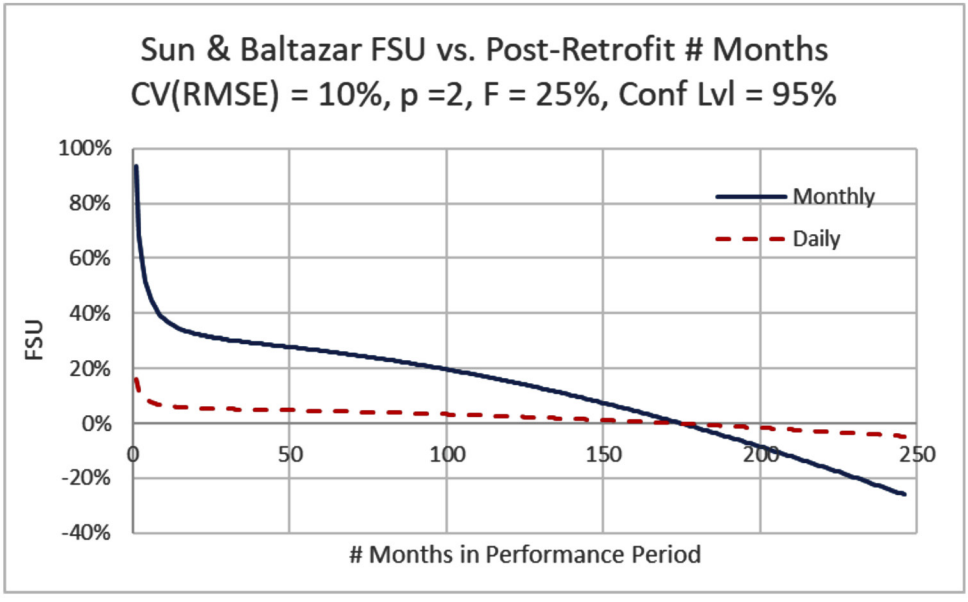


Figure 6. Sun and Baltazar Model Results

of data, FSU approaches 0. All projects and their models are acceptable. I don't know how to interpret negative FSU though. Perhaps it implies that there is no uncertainty, but instead supercertainty.

Josh Rushton's version, I am ashamed to say, involved such complex matrix algebra, that I avoided it entirely. It may be wonderful, but I couldn't tell you. For more information about the Rushton model, see SBW Consulting [5].

In sum, the equations the industry uses to represent FSU today or at any future point will continue to be arbitrary, that is a product of someone's imagination.* Therefore, coming up with some standard equation and some maximum acceptable FSU value to define all successful projects is also arbitrary, not based in reality, and really just a figment of someone's imagination, that we have all accepted as truth. Or as Gertrude Stein once wrote: "There is no 'there' there."

THE MODELS ARE ABSTRACTIONS AND UNSCIENTIFIC

Engineers are not statisticians. We have enough complexity in our own profession that we usually prefer to leave statistics to the statisticians. Many of us engineers, otherwise very intelligent people, take for granted things that may not be true. We take as a postulate that R^2/CV(RMSE)/FSU is the best method to gauge the applicability of linear regression models, when unfortunately, these methods may not actually be the best.

None of these models (R^2, CV(RMSE) or FSU) can be compared to reality. That is the problem. They are models of a concept, not models of reality, and as a result, they cannot be proven to be true or accurate.† What can we measure to prove the regression model is adequate? Nothing. The R^2 and CV(RMSE) values are flawed and only work sometimes. The entire concept of using these indicators to prove the soundness of a regression model is unscientific. They are agreed upon figments of our collective imagination composed for us by authorities such as ASHRAE

*Much like the difference between science, which is supposed to be based on observations, and metaphysics, which is created by the mind.

†A funny response by a well-respected statistician-engineer: "Sure they can. Well, it depends upon your definition of "prove" but I personally am extremely 'confident' of their reality."

and EVO. We take their recommendations as gospel.*

Rather than accept this collective delusion, let's recognize FSU, CV(RMSE) and R^2 values for what they really are. They are invented unscientific concepts that can never be tested against reality.

COMPLEXITY

If you accept that, then the next step is to recognize that adding more complexity, such as FSU, to the problem is not necessarily an improvement. More complexity just means a more complex arbitrary standard. We still cannot compare the results and prove scientifically that one statistical indicator is better than another. Yet industry guidelines have been tending towards more complexity. Complex statistical indicators may be understandable to statisticians, but they are leaving more and more practitioners behind. If the new methods are not easily comprehensible, then practitioners will continue to ignore the new advice and continue to use old "tried and true" guidelines, such as $R^2 > 0.75$ and CV(RMSE) < 25%, which really are not sufficient as they are now.

A BETTER WAY TO VALIDATE LINEAR REGRESSIONS

ASHRAE Guideline 14 recommended that linear regressions having CV(RMSE) values less than 25% are acceptable†‡. Because the CV(RMSE) is not always useful, we need a better method to determine whether our linear regression models to be used in performance con-

*This is akin to the early Church fathers (politicians in practice), who, in the 4th century Council of Nicaea selected the Christian canon and the gospels that everyone in the Roman Empire had to believe in. Over 40 gospels and letters were thrown out, and they edited those that they kept. They determined what was the truth, and those who believed otherwise were stigmatized, subject to book burnings, deprived of their property, fired from their bishoprics, exiled, and murdered by the state. (Don't worry, I would never claim that ASHRAE or EVO would ever go to such extremes.)

†Actually, ASHRAE Guideline 14-2014 says "the baseline model shall have a maximum CV(RMSE) of 20% for energy use and 30% for demand quantities when less than 12 months' worth of post-retrofit data are available for computing savings. These requirements are 25% and 35%, respectively, when 12 to 60 months of data will be used in computing savings. When more than 60 months of data will be available, these requirements are 30% and 40%, respectively."

‡ASHRAE Guideline 14 also requires that the savings uncertainty (FSU) be less than 50% of the annual savings at 68% confidence.

tracting are acceptable or not.

I have already pointed out the deficiencies of the R^2 and CV(RMSE) in certain conditions. Still, we need some way to identify a good from a bad regression. I am suggesting the following rules for validating the best fit of a linear regression line:

If the $R^2 > 0.75$ or the CV(RMSE) < 25%, the regression is valid.

If you don't immediately see how this is different, let me explain.

- If previously you only accepted fits with R^2 values above 0.75, now, you can accept with lower R^2 values. Often these are the fits with more horizontal slopes.
- If previously you only accepted regressions with CV(RMSE)s below 25%, now you can accept fits above 25% as well. These might be the fits with low average values, such as natural gas meters which have near zero usage during summer months.

Some of these models may not comply with ASHRAE Guideline 14's requirement of CV(RMSE) < 25%, but that is the point here. We are looking for a better way to validate regressions.

This is only a step, but a first step to making more sense of validating regressions. But we are left with a question: why use thresholds of 0.75, and 25%? Wouldn't 0.7 work as well? Probably. Wouldn't 30% work as well? Maybe. But this is a topic for another article.

CONCLUSION

It took me more than a decade to realize that being smart does not mean having memorized whatever guidelines and standards are out there. Most people content themselves with this level of mastery. We must remember normal people write these documents. There are few Einstein's out there. Most of them are just like us. It is important to understand and really think about the concepts in these documents.

I hope that I have convinced you of the arbitrary, unscientific nature of the statistical indicators we have been using to validate regressions used in M&V. More complex formulations, such as FSU, will not necessarily remedy the problem. As the statistical indicators and thresholds,

we have been using are of little use in some circumstances, we should recognize that we should not be bound by prior published recommendations and can seek a better way to validate regressions.

The industry needs rules that are easily understandable and that will work in more cases than existing ASHRAE Guideline 14 or EVO recommendations. I have presented a set of rules that I believe are understandable and easily implementable for M&V practitioners.

References

[1] Stetz, Mark. "Why R^2 Doesn't Matter." M&V Focus. Efficiency Value Organization. October 2019. Available at https://evo-world.org/en/news-media/m-v-focus/868-m-v-focus-issue-5/1164-why-r2-doesn-t-matter.

[2] Reddy, T.A. and D.E. Claridge. "Uncertainty of "Measured." Energy Savings from Statistical Baseline Models." HVAC&R Research. Vol 6, No. 1. January 2000.

[3] ASHRAE. ASHRAE Guideline 14-2014, Measurement of Energy, Demand, and Water Savings. American Society of Heating, Refrigerating, and Air Conditioning Engineers. Atlanta, GA.

[4] Avina, J. "Why Do We Calculate Uncertainty?" *International Journal of Energy Management*, Vol 3, No 3. Pages 26-31. Association of Energy Engineers. Atlanta, GA.

[5] SBW Consulting. Uncertainty Approaches and Analyses for Regression Models and ECAM. Bonneville Power Administration. 2017.

AUTHOR BIOGRAPHY

John Avina, CEM, CEA, CMVP, CxA, has worked in energy analysis and utility bill tracking for over 25 years. During his tenure at Thermal Energy Applications Research Center, Johnson Controls, SRC Systems, Silicon Energy and Abraxas Energy Consulting, Mr. Avina has managed the measurement and verification (M&V) for a large performance contractor, managed software development for energy analysis and M&V applications, created M&V software that is used by hundreds of energy professionals, taught over 250 energy management classes, created hundreds of building models and utility bill tracking databases, modeled hundreds of utility rates, and has personally performed energy audits and RCx on over 25 million square feet. Mr. Avina currently chairs the Certified Energy Auditor Exam Committee for the Association of Energy Engineers. Mr. Avina has an MS in Mechanical Engineering from the University of Wisconsin-Madison. John may be contacted via email at john.avina@abraxasenergy.com.