Presented at the 2009 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

A Distribution-Free Measure of the Significance of CER Regression Fit Parameters Established Using General Error Regression Methods

Timothy P. Anderson

MCR, LLC

Introduction

General Error Regression Methods (GERM) have earned a strong following in the cost estimating community as a means of establishing cost estimating relationships (CERs) using non-linear functional forms. An example of the type of functional form developed using GERM includes:

$$Y = a + bX^{c}W^{d}Q^{e}f^{Type}$$

where

Y is the independent variable;

X, *W*, *Q* and *Type* are the dependent variables; and

a, *b*, *c*, *d*, *e* and *f* are the regression fit parameters[1].

GERM has given rise to a wide variety of functional forms for CERs, but has so far lacked a means for evaluating the "significance[2]" of the individual regression fit parameters in a way that is analogous to the roles played by the *t*-statistic and associated *p*-value in ordinary least squares (OLS) regression. This research attempts to remedy that situation by developing and describing an analogous "significance" metric for GERM regression fit parameters that is independent of the nature of the underlying error distribution.

The significance metrics developed herein are comparable across CERs regardless of the functional form of the regression equation or the underlying error specification. Moreover, they are developed heuristically, they require no distributional assumptions, and they provide a collection of simple metrics by which to judge the "significance" of the individual regression fit parameters.

These metrics will be beneficial to anyone who uses GERM to develop CERs. They will enable cost modelers to judge whether or not cost-driver variables used in CERs that are derived using GERM have any real impact on the mean result of the CER, and whether or not they contribute to the overall reduction in the CER's variance. Those variables that, if excluded, would not

[1] The term "fit parameter" is used herein instead of the more commonly used term "coefficient." This is because the term "coefficient," mathematically speaking, refers to a multiplier only. Since generalized functional forms make use of multipliers, adders, and exponents, it would be incorrect to refer to all of these as "coefficients."

[2] In classical statistics, a result is deemed "statistically significant" if that result is not likely to have occurred by chance. This concept will be described more fully within this paper.

substantially impact the mean result or significantly change the CER variance, can be removed from the cost model with little or no consequence. This is often desirable because technical data collection can be difficult. Moreover, one less variable means one more degree of freedom, and that can be important when a small number of data points are available. So it is important to be able to eliminate from study those variables that are not significant cost drivers.

Background

The "significance" test for OLS

OLS regression and its application to linear relationships has been in use since 1795, when it was first developed by Carl Friedrich Gauss. OLS is used most frequently to establish an estimate of the linear relationship between a vector **Y** and a matrix **X** consisting of one or more vectors $\mathbf{X_i}$, whose actual relationship is assumed to be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon ,$$

where β_i are the actual parameters, and ε is a random error term with mean equal to zero and constant variance.

The typical result of OLS is an equation of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n ,$$

whose fit parameters b_i are estimates of the actual, though unknown, parameters β_i. The fit parameters b_i, comprising the vector **b**, are calculated by solving the following matrix equation:

$$\mathbf{b} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} .$$

Modern statistical applications typically perform tests of the "significance" of the fit parameters b_i under the assumption that the error term ε has a *normal* distribution. Statistical theory shows that the *Student's t* distribution, which is derived from the normal distribution, should be used to test significance hypotheses involving the fit parameters of a linear regression equation. In statistics, a result is deemed "statistically significant" if that result is not likely to have occurred by chance. Therefore, what is desired is a way to test whether or not an actual parameter is really just zero given that a non-zero estimate for it was produced.

In the case of OLS, each individual regression fit parameter is scored using the *t-statistic*. The *t-statistic* is the solution to the test statistic obtained by performing the following hypothesis test:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

where:

H_0 represents the null hypothesis that the *i*th regression parameter is equal to zero; and

H_a represents the alternate hypothesis that the *i*th regression parameter is not equal to zero.

The test statistic is:

$$t_{b_i} = \frac{b_i - \beta_i}{SE / \sqrt{\sum (x_i - \bar{X})^2}},$$

where:

b_i is the estimated value of the parameter calculated by performing OLS;

β_i is the true value of the parameter, assumed under the null hypothesis to be zero;

SE is the standard error of the regression, $SE = \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n - m}}$;

$\hat{y}_i$ is the estimated value of the i^{th} *Y* value;

y_i is the actual value of the i^{th} *Y* value;

n is the number of data points used in the regression;

m is the number of parameters estimated;

x_i is the i^{th} observation of vector *X*; and

$\bar{X}$ is the mean value of vector *X*.

The test statistic is then compared to a *Student's t* distribution with mean zero and standard deviation $s_{b_i} = SE / \sqrt{\sum (x_i - \bar{X})^2}$. If the test statistic falls within a critical region, determined in advance, of the tails of the *Student's t* distribution, then the null hypothesis H_0 is rejected, and the fit parameter b_i is said to be "statistically significant," meaning it is unlikely to have been arrived at by chance.

In practical terms, those fit parameters that are deemed to be significant imply that the corresponding cost-driving variables X_i have something important to say about the value of *Y*. Conversely, those fit parameters that are *not* significant are probably due merely to chance, and thus imply that the corresponding variables have little of importance to say about the value of *Y*, and therefore can be removed from the regression equation with little consequence. In cases such as this, the typical response is to remove an insignificant independent variable from consideration, establish a new regression equation without that variable, then re-examine the remaining variable fit parameters for significance, repeating this process until all cost drivers used in the regression model are deemed significant.

The General Error Regression Method (GERM)

For the purposes of this research, GERM refers to the regression method in which estimates of the parameters of generalized functional forms (e.g., non-linear) are derived through

constrained optimization. At least one alternative to GERM is a procedure commonly known as *iteratively re-weighted least squares* (IRLS). It is important to note that this research is not necessarily intended to support the IRLS method.

GERM was first popularized in the cost estimating community in 1997 by S. A. Book and P. H. Young[3], and then improved upon in 1998 by S. A. Book and N. Y. Lao[4]. GERM enables one to derive a regression model regardless of the functional form or the nature of the error distribution by seeking fit parameter values that minimize the sum of the squared errors, or the sum of the squared percent errors, relative to any generalized functional form. There are two common varieties of GERM models – those with additive errors, and those with multiplicative errors. The error specification determines the appropriate regression procedure.

GERM with Additive Errors

A relationship between arrays **Y** and **X** that displays additive errors is similar to that shown in the scatterplot in Figure 1. Notice that the variance around the regression curve appears to stay relatively constant across the *X*-axis.

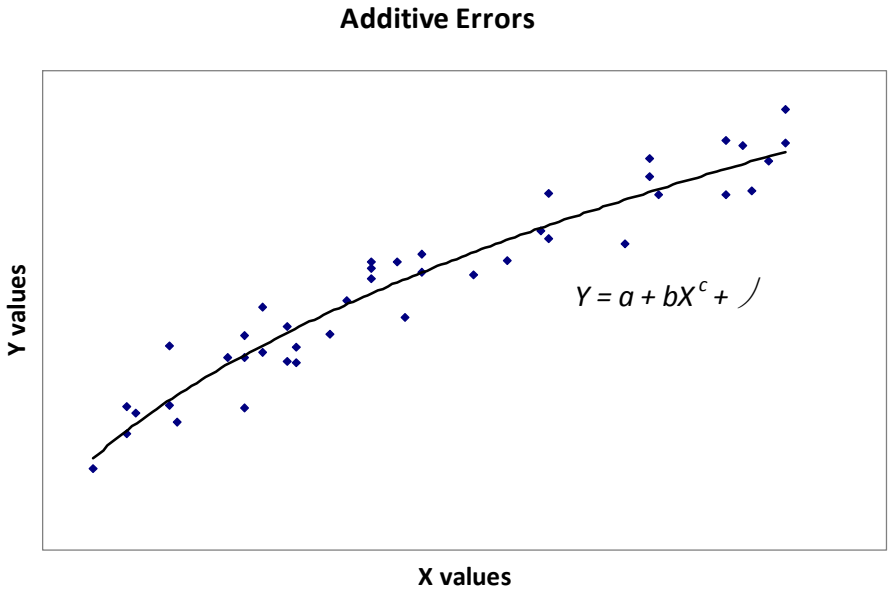

Figure 1 - Regression Model with Additive Errors

To solve for the estimates of the parameters of a general model, $f(X)$, the following steps are performed:

[3] Book, S.A., and Young, P.H., "General-Error Regression for Deriving Cost-Estimating Relationships," *The Journal of Cost Analysis*, Fall 1997, pages 1-28.

[4] Book, S.A., and Lao, N.Y., "Minimum-Percentage-Error Regression under Zero-Bias Constraints," *Proceedings of the Fourth Annual U.S. Army Conference on Applied Statistics*, 21-23 October 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.

Let **Y** = the array containing observations of the dependent variable Y_i;

X = the array or matrix containing observations of the independent variables X_i; and

f(X) = the estimated values of *Y*.

For each y_i, the actual value equals the estimated value plus a random error ε_i, with mean zero and constant variance,

$$y_i = f(x_i) + \varepsilon_i .$$

The error is the difference between the actual value and the estimated value,

$$\varepsilon_i = y_i - f(x_i).$$

The problem, then, is to choose the fit parameters of $f(X)$ so that the sum of squared errors (*SSE*),

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 ,$$

is as small as possible.

The solution can be found using optimization techniques, as shown in the next example.

Example:

A typical non-linear functional form for this type of data is given as

$$Y = a + bX^c + \varepsilon ,$$

where the error term ε has mean zero and constant variance.

To solve for the specific case of the estimates of the parameters *a*, *b*, and *c*, the following steps are performed:

For each y_i, the actual value equals the estimated value plus a random error,

$$y_i = a + bx_i^c + \varepsilon_i .$$

The error is the difference between the actual value and the estimated value,

$$\varepsilon_i = y_i - a - bx_i^c .$$

The problem, then, is to choose the fit parameters *a*, *b*, and *c*, using optimization techniques, so that the *SSE*,

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i^c\right)^2 ,$$

is as small as possible.

In order to ensure that the resulting function, $f(X)$ is sample unbiased[5], the optimization should be constrained such that the average bias is zero:

$$Average\ Bias = \frac{1}{n}\sum_{i=1}^{n}\left[a + bx_i^c - y_i\right] \times 100\% = 0\,.$$

In terms of the optimization routine, we seek *a*, *b* and *c* that provide an optimum solution to the following non-linear program:

Decision Variables: *a*, *b*, *c*

Objective Function: minimize $SSE = \sum_{i=1}^{n}\left(y_i - a - bx_i^c\right)^2$

Subject to: $Average\ Bias = \frac{1}{n}\sum_{i=1}^{n}\left[a + bx_i^c - y_i\right] \times 100\% = 0\,.$

Several methods exist for solving this type of non-linear programming problem. The specific method to be used in each particular case is left to the reader.

Key statistics that may be derived as a result of the GERM method for models with additive errors include the standard error (*SE*) of the estimate, and Pearson's r^2 – the squared correlation between the actual values and the estimated values. These are calculated as follows:

The *SE* in GERM with additive errors has the same interpretation as the *SE* in OLS. It is the root mean square of all errors made in estimating the points in the database, and is calculated directly from the data as follows:

$$SE = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2}\,.$$

The Pearson's r^2 is the squared correlation between the estimated values, $f(x_i)$, and the actual values, y_i. This metric serves as a rough analog to the *Coefficient of Determination* (R^2) that is calculated as part of OLS, but instead measures the relationship of the *estimates* to the corresponding *actuals*, and can be calculated regardless of the CER's functional form or method of derivation. Pearson's r^2 is calculated as follows:

$$\text{Pearson's } r^2 = \left[\frac{n\sum_{i=1}^{n} y_i f(x_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} f(x_i)}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}\sqrt{n\sum_{i=1}^{n} f(x_i)^2 - \left(\sum_{i=1}^{n} f(x_i)\right)^2}}\right].$$

[5] In this context, a regression equation is "sample unbiased" if the average of the residuals is zero. That is, the regression equation is biased neither upward nor downward relative to the data used to create the regression equation.

In a successful application of GERM with additive errors, *SE* is minimized, Pearson's r^2 is maximized, and bias is zero.

GERM with Multiplicative Errors

A relationship between arrays **Y** and **X** that displays multiplicative errors is similar to that shown in the scatterplot in Figure 2. Notice that the variance around the regression curve appears to increase as the values on the *X*-axis increase.

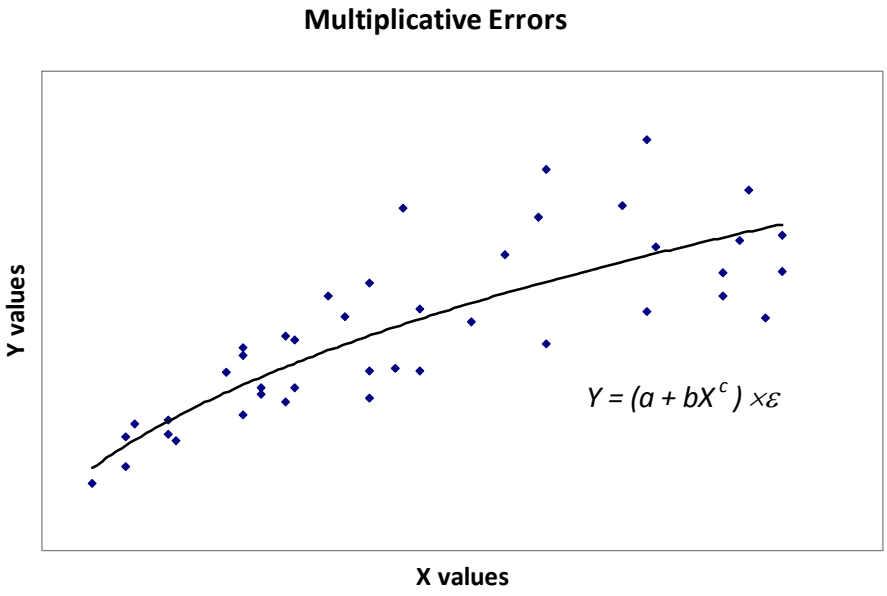

Figure 2 - Regression Model with Multiplicative Errors

To solve for the estimates of the parameters of a general model, $f(X)$, the following steps are performed:

Let **Y** = the array containing observations of the dependent variable Y_i;

X = the array or matrix containing observations of the independent variables X_i; and

f(X) = the estimated values of *Y*.

For each y_i, the actual value equals the estimated value multiplied by a random error ε_i, with mean 1.0 and constant variance,

$$y_i = f(x_i) \cdot \varepsilon_i .$$

The error is the ratio of the actual value to the estimated value,

$$\varepsilon_i = \frac{y_i}{f(x_i)} = \frac{\text{Actual}}{\text{Estimate}} .$$

The problem, then, is to choose the fit parameters of $f(X)$ so that the summation,

$$\sum_{i=1}^{n}(\varepsilon_i - 1)^2 = \sum_{i=1}^{n}\left[\frac{y_i - f(x_i)}{f(x_i)}\right]^2,$$

is as small as possible.

The solution can be found using optimization techniques as shown in the next example.

Example:

A typical non-linear functional form for this type of data is given as

$$Y = \left(a + bX^c\right)\cdot\varepsilon,$$

where the error term ε has mean equal to 1.0 and constant variance.

To solve for the specific case of the estimates of the parameters *a*, *b*, and *c*, the following steps are performed:

For each y_i, the actual value equals the estimated value multiplied by a random error,

$$y_i = \left(a + bx_i^c\right)\cdot\varepsilon_i.$$

The error is the ratio of the actual value to the estimated value,

$$\varepsilon_i = \frac{y_i}{a + bx_i^c}.$$

We desire ε_i to be as close to 1.0 as possible. The problem, then, is to choose the fit parameters *a*, *b*, and *c*, using optimization techniques, so that the summation,

$$\sum_{i=1}^{n}(\varepsilon_i - 1)^2 = \sum_{i=1}^{n}\left(\frac{y_i - a - bx_i^c}{a + bx_i^c}\right)^2,$$

is as small as possible.

In order to ensure that the resulting function, $f(X)$, is sample unbiased, the optimization should be constrained such that

$$Average\ Percent\ Bias = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{a + bx_i^c - y_i}{a + bx_i^c}\right]\times 100\% = 0.$$

In terms of the optimization routine, we seek *a*, *b* and *c* that provide an optimum solution to the following non-linear program:

Decision Variables: *a*, *b*, *c*

Objective Function: minimize $\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i^c}{a + bx_i^c}\right)^2$

Subject to: $\frac{1}{n}\sum_{i=1}^{n}\left[\frac{a + bx_i^c - y_i}{a + bx_i^c}\right]\times 100\% = 0.$

Again, several methods exist for solving this type of non-linear programming problem. The specific method to be used in each particular case is left to the reader.

Key statistics that may be derived as a result of the GERM method for models with multiplicative errors include the standard percent error (*SPE*) of the estimate, and Pearson's r^2 – the squared correlation between the estimated values and the estimated values. These are calculated as follows:

The *SPE* in GERM with multiplicative errors is the root mean square of all percent errors made in estimating the points in the database, and is calculated directly from the data as follows:

$$SPE = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[\frac{f(x_i)-y_i}{f(x_i)}\right]^2}\ .$$

The Pearson's r^2 is the squared correlation between the estimated values, $f(X_i)$, and the actual values, Y_i. This metric serves as a rough analog to the *Coefficient of Determination* (R^2) that is calculated as part of OLS, but instead measures the relationship of the *estimates* to the corresponding *actuals*, and can be calculated regardless of the CER's functional form or method of derivation. Pearson's r^2 is calculated in the same way as done in the case of additive errors, as follows:

$$\text{Pearson's } r^2 = \left[\frac{n\sum_{i=1}^{n} y_i f(x_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} f(x_i)}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}\sqrt{n\sum_{i=1}^{n} f(x_i)^2 - \left(\sum_{i=1}^{n} f(x_i)\right)^2}}\right].$$

In a successful application of GERM with multiplicative errors, *SPE* is minimized, Pearson's r^2 is maximized, and percent bias is zero.

There is no analogous "significance" test for GERM

Until now, there has been no "significance" test for fit parameters that are derived using GERM that is analogous to the *t*-statistic and associated *p*-value found in OLS. One reason for this is that GERM-derived CERs are typically non-linear, and there is no reason to assume that their errors are normally distributed. In fact, it is usually agreed that the error distributions are anything but normal. It is commonly assumed in practice that the error distributions of these CERs are lognormal, but that is primarily for convenience – not necessarily a result of the mathematics. Therefore, CER developers usually determine, in advance, which independent variables they desire in their models, then they judge the goodness of the models based solely on *SE* or *SPE* and Pearson's r^2. If a CER can be developed with the desired independent variables, low *SE* or *SPE*, and high Pearson's r^2, then the model is considered a success.

But clearly, it would be beneficial to be able to judge whether or not one or more of the "desired" independent variables actually have something important to say about the dependent variable. Therefore, the goal of this research is to "invent" a collection of metrics that are analogous to the *t*-statistic, that are comparable between CERs regardless of the

functional form or the underlying error specification, that require no distributional assumptions, and that provide a set of simple metrics by which to judge the "significance" of the individual regression fit parameters.

Introducing the *SIG* Test

The *SIG* test, so named because it measures the "significance" of fit parameters independently of the underlying error distribution, is proposed herein and meets the goals of the research – works on generalized functional forms, needs no distributional assumptions, and provides a set of simple numerical scores against which each fit parameter can be measured.

It is different from the *t*-statistic. The *t*-statistic is an analytical result based on the assumption that the underlying error distribution is normal. The *SIG* test, on the other hand, is a heuristic result, and does not care about the nature of the underlying error distribution, other than that it has a mean and a standard deviation. However, in many ways the *SIG* test is similar to the *t*-statistic in that it tests the hypothesis that the true parameter is negligible.

Consider a CER of the form

$$Y = a + bX^{c}W^{d}Q^{e}f^{Type}\ .$$

The basic idea is that if any of the fit parameters *a*, *b*, *c*, *d*, *e*, or *f* does not substantially impact the calculated value of the CER, or reduce the model's variance, whatever the nature of the error distribution might be, then it might as well be removed from the model. It immediately becomes apparent that some fit parameters are more important than others. In the CER form shown above, fit parameters *c*, *d*, *e* and *f* relate directly to hypothesized cost drivers, so those are more important than fit parameters *a* and *b*, which relate to the overall model. However, the method described herein will test the "significance" of all six fit parameters – whether tied to a hypothesized cost driver or not.

Hypothesis: "Insignificant" Fit Parameters Will Have Little or No Impact on CER Mean or Variance if Nullified

It is postulated that "insignificant" fit parameters would have little to no impact on the mean or variance of the CER if nullified. On the other hand, "significant" fit parameters should substantially alter the CER's mean or variance if nullified. Take note of Figure 3, in which two Estimates vs. Actuals plots are shown – one based on a CER containing multiple independent variables, and the other on a re-optimized CER with one of the independent variables removed.

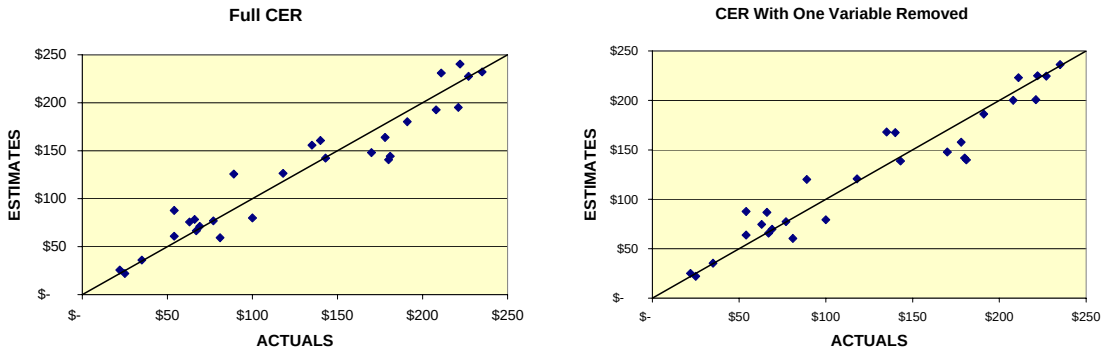

Figure 3 – Two Similar Actuals vs. Estimates Plots

Notice that the two plots appear nearly identical. This is an indication that the CER performs about the same with or without that independent variable; thus, one could conclude that the missing variable is "insignificant." This phenomenon was actually observed during a recent CER development effort, leading to the ideas developed in this paper.

On the other hand, consider the two plots shown in Figure 4. Again, the plot on the left is based on a CER containing multiple independent variables, and the one on the right on a re-optimized CER with one of the independent variables removed.

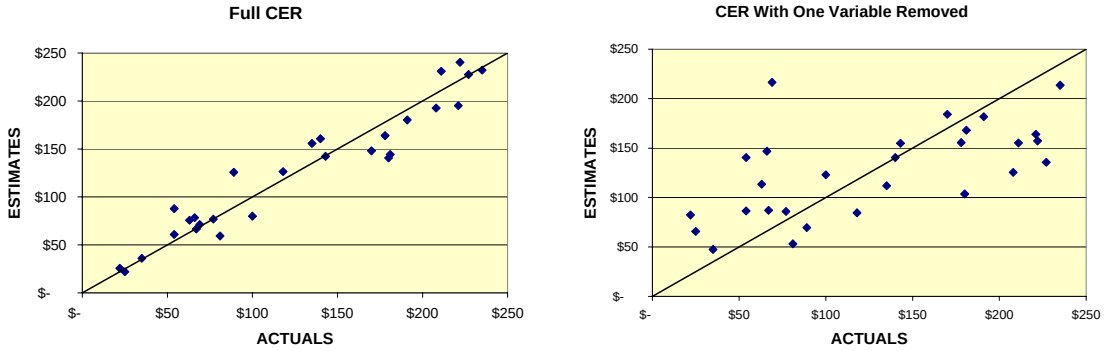

Figure 4 – Two Different Actuals vs. Estimates Plots

In this example, the two plots are quite different. The CER with one variable removed does not perform nearly as well as the full CER. This is an indication that the missing variable is indeed "significant" and should not be removed from the CER.

What changed when the variable in question was removed? Clearly, the CER's variance increased, and it is possible that its mean shifted also (although a shift in the mean is not immediately apparent from the plots shown). This would indicate that "significance" can be measured relative to the change in variance and possibly a change in the mean value of the CER as well.

A Note on Fit Parameter "Nullification"

How, exactly, does one "nullify" a fit parameter? There is an element of art to this, and it depends on the functional form of the CER. Recall that what we desire is a method to eliminate independent variables one by one from the CER. Moreover, in keeping with the idea of having

an analog to the *t*-statistic, we also desire to test the value of each fit parameter in the regression equation. Consider again, the typical GERM functional form shown earlier:

$$Y = a + bX^{c}W^{d}Q^{e}f^{Type}.$$

The independent variables are *X*, *W*, *Q*, and *Type*, and the fit parameters are *a*, *b*, *c*, *d*, *e*, and *f*. The method used to nullify each fit parameter depends on its position in the CER. Fit parameter *a* is an additive constant and is not directly related to any of the independent variables. Therefore, nulling this fit parameter is as simple as setting it equal to zero. Fit parameter *b* is a multiplier. Setting it to zero would effectively render all independent variables zero, so the best way to nullify *b* is to set it equal to 1. Fit parameters *c*, *d*, and *e* are exponents. Each of them can be nullified by setting them equal to zero, which has the effect of setting their associated independent variables to 1. Fit parameter *f*, like fit parameter *b*, is another multiplier. Setting *f* to zero would effectively render all independent variable zero, so the best approach is to set it equal to 1. Different functional forms will require a common sense approach like this in order to determine how best to nullify the fit parameters.

Proposition: Fit Parameters Can Be Scored Based on Their Impact on CER Mean and Variance

It is proposed that individual CER fit parameters can be "scored" based on their impact on the CER's mean and variance if nullified. The scoring metric should be such that it will produce a small number if nullification of the fit parameter causes an inconsequential change to the CER's mean and/or variance, and should provide a large number otherwise.

Central to this idea is that the CER has a probability distribution (although its shape may be unknown) with a quantifiable mean and variance. GERM CERs easily fit this criterion. They are produced in such a way that they are sample unbiased – so, while yet to be proven, it is generally thought that the value of the CER, when evaluated with given independent variables, represents the mean of a cost distribution. And, they also have a variance, usually represented by the *SE* or *SPE*.

Figure 5 shows an overlay of the cost probability distributions of two multivariable CERs evaluated at the mean values of their independent variables[6]. One of the CERs was based on a full model, and the other was based on the same functional form, but with one of the variables removed.

[6] These distributions are assumed to be lognormal, but the theory should apply to any distributional shape.

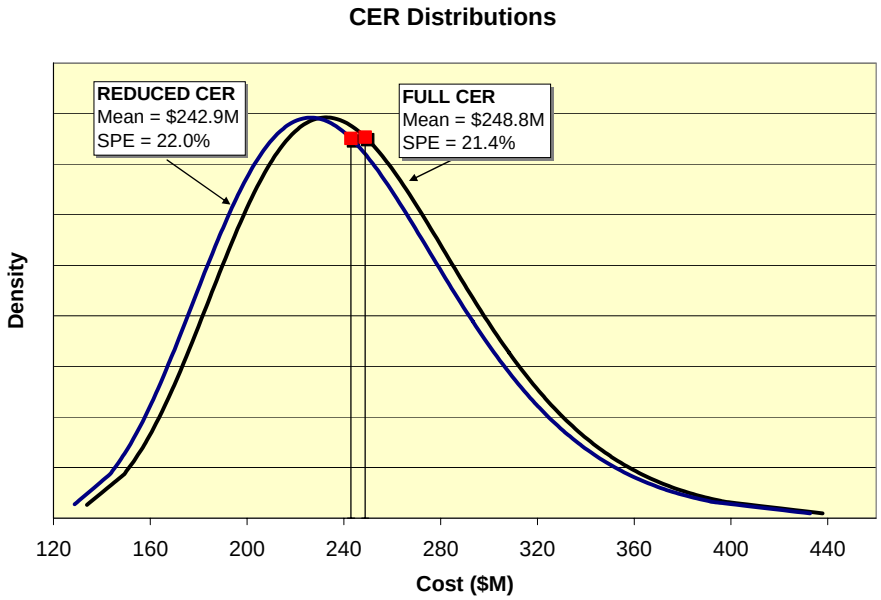

Figure 5 – Two Similar Cost Probability Distributions

The full CER has a mean of $248.8M and an *SPE* of 21.4% when evaluated at the mean values of its independent variables. However, the reduced CER – same functional form, but with one variable removed from the model – has a nearly identical distribution, with a mean[7] of $242.9M and an *SPE* of 22.0%. The effect of removing the independent variable was to decrease the mean by just over 2%, and increase the *SPE* by just under 3%. One could easily argue that removing the independent variable from the model had almost no effect on the resulting CER distribution, and therefore, that independent variable – or its fit parameter – is "insignificant." Either CER would produce nearly the same result with nearly the same error distribution.

Contrast this with the overlay shown in Figure 6. Again there are two multivariable CER distributions – one based on a full model, and the other based on the same functional form, but with one of the variables removed.

[7] Here, the mean of the cost distribution is considered to be the evaluated result of the CER given the mean values of the cost drivers that were used in the creation of the CER. However, there is no requirement to use the mean values of the cost driver data points in this calculation. Under the assumption that the CER has a probability distribution, any set of values will serve as valid input variables. Once the input variables are provided and the CER evaluated, the result of the CER will represent the mean of a cost distribution. One caveat, however, is that the "mean" is actually a "pseudo-mean," since the CER is based on a sample, so the true mean of the cost distribution is unknown.

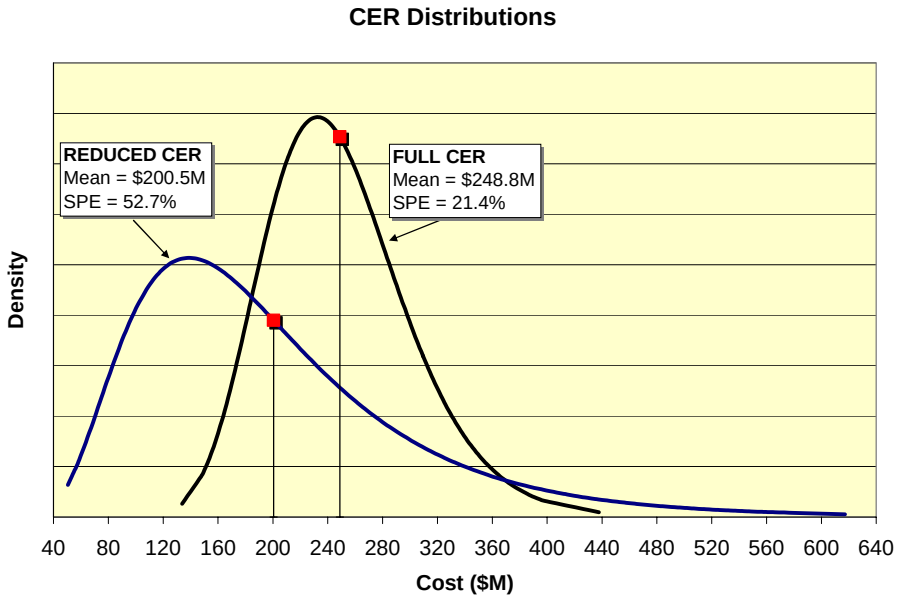

Figure 6 – Two Different Cost Probability Distributions

Again, the full CER has a mean of $248.8M and an *SPE* of 21.4% when evaluated at the mean values of its independent variables. However, the reduced CER – same functional form, but with one variable removed from the model – has a substantially different distribution, with a mean of $200.5M and an *SPE* of 52.7%. In this case, the effect of removing the independent variable was to decrease the mean by nearly 20%, and increase the *SPE* by over 100%. Here, one would argue that removing the independent variable from the model had a very large effect on the resulting CER distribution, and therefore, that independent variable – or its fit parameter – is "significant." The CERs would produce very different results with the same set of independent variable values.

"Significance" can be Measured by the Change in the CER Mean and Variance

In the previous example, it was shown that a CER's cost probability distribution can be impacted by the removal of an independent variable. Moreover, it was shown that the impact to the distribution can be measured by the change in the CER's mean – when evaluated at the means of the independent variables – as well as by the change in the CER's variance under the same circumstances. Therefore, we propose that a set of metrics for a given CER that indicates the degree of change when a fit parameter is nullified, or equivalently, when an independent variable is removed from the model.

The Significance Relative to the Mean, SIG_{Mean}

The first significance metric is the significance of the fit parameter relative to the mean of the CER distribution. For purposes of this discussion, the mean used herein refers to the mean of

the CER distribution when evaluated at the mean value of each of the independent variables as shown below[8].

$$f_{\bar{Y}}(X) = f\left(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N\right).$$

SIG_{Mean} is defined as the percentage difference between the mean of the *full* CER and the mean of the *reduced* CER:

$$SIG_{Mean} = \frac{f_{\bar{Y}}(X)_{\text{Reduced}} - f_{\bar{Y}}(X)_{\text{Full}}}{f_{\bar{Y}}(X)_{\text{Full}}}.$$

For example, consider the following CER:

$$f(X) = -79.65 + 31.35\left(X_1\right)^{0.3664}\left(X_2\right)^{0.1094}\left(X_3\right)^{0.0576}\left(1.44\right)^{X_4}$$

where X_1, X_2, and X_3 are continuous variables, and X_4 is a binary (0, 1) variable. Assuming there exist underlying data for these four variables, then calculating the means of the independent variables is trivial:

$$\bar{X}_i = \frac{1}{N}\sum_{j=1}^{N} x_{i,j}.$$

It follows, then, that the mean of the *full* CER can be calculated as:

$$f_{\bar{Y}}(X)_{\text{Full}} = -79.65 + 31.35\left(\bar{X}_1\right)^{0.3664}\left(\bar{X}_2\right)^{0.1094}\left(\bar{X}_3\right)^{0.0576}\left(1.44\right)^{\bar{X}_4}.$$

Suppose the values are $\bar{X}_1 = 108$, $\bar{X}_2 = 12$, $\bar{X}_3 = 9.9$, and $\bar{X}_4 = 0.6$. Then:

$$f_{\bar{Y}}(X)_{\text{Full}} = -79.65 + 31.35(108)^{0.3664}(12)^{0.1094}(9.9)^{0.0576}(1.44)^{0.6} = 245.22.$$

Now suppose the exponent associated with X_1 is nullified, by setting it to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f_{\bar{Y}}(X)_{\text{Reduced}} = 151.96 + 9.03\left(\bar{X}_1\right)^{0}\left(\bar{X}_2\right)^{-0.3739}\left(\bar{X}_3\right)^{0.5594}\left(3.60\right)^{\bar{X}_4},$$

or

$$f_{\bar{Y}}(X)_{\text{Reduced}} = 151.96 + 9.03\left(\bar{X}_2\right)^{-0.3739}\left(\bar{X}_3\right)^{0.5594}\left(3.60\right)^{\bar{X}_4}.$$

Re-entering the values for $\bar{X}_2$, $\bar{X}_3$ and $\bar{X}_4$, we have:

$$f_{\bar{Y}}(X)_{\text{Reduced}} = 151.96 + 9.03(12)^{-0.3739}(9.9)^{0.5594}(3.60)^{0.6} = 179.69.$$

[8] This is not intended to imply that in general the mean of the CER is a function of the mean of the independent variables! That most certainly is not true unless the CER is a linear function. On the contrary, the CER has infinitely many means depending on the choice of values used for the independent variables. However, for consistency, we compare the CER means in this methodology using the *same* set of independent variables throughout – specifically, the means of those independent variables. But note that this is a matter of convenience only. Any other set of independent variable values could be chosen, so long as they are chosen consistently.

In this case, the difference between the *full* and *reduced* CERs is quite large. The value calculated for SIG_{Mean} is:

$$SIG_{Mean} = \frac{179.69 - 245.22}{245.22} = -0.267 = -26.7\% .$$

In other words, the mean of the *reduced* CER is nearly 27% lower than the mean of the *full* CER. This would indicate that the fit parameter for the exponent associated with X_1 is "significant."

Now consider a different example. Using the same *full* CER, nullify the exponent associated with X_3 by setting it equal to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f_{\bar{Y}}(X)_{\text{Reduced}} = -63.10 + 29.23\left(\bar{X}_1\right)^{0.3962}\left(\bar{X}_2\right)^{0.0924}(1.485)^{\bar{X}_4} .$$

Re-entering the values for $\bar{X}_1, \bar{X}_2$ and $\bar{X}_4$, we have:

$$f_{\bar{Y}}(X)_{\text{Reduced}} = -63.10 + 29.23(108)^{0.3962}(12)^{0.0924}(1.485)^{0.6} = 234.91 .$$

In this case, the difference between the *full* and *reduced* CERs is much smaller. The value calculated for SIG_{Mean} is:

$$SIG_{Mean} = \frac{234.91 - 245.22}{245.22} = -0.042 = -4.2\% .$$

In other words, the mean of the *reduced* CER is only about 4% lower than the mean of the *full* CER. This would indicate that the fit parameter for the exponent associated with X_3 is relatively "insignificant."

The Significance Relative to the Standard Error, SIG_{SE}

The second significance metric is the significance of the fit parameter relative to the standard error (*SE*) of the CER distribution. Note that this particular discussion relates to GERM CERs that have *additive* errors. A similar treatment for GERM CERs with *multiplicative* errors follows in the next section. For purposes of this discussion, the *SE* of the CER distribution is computed as part of the GERM procedure as shown below:

$$SE = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2} .$$

SIG_{SE} is defined as the percentage difference between the *SE* of the *full* CER and the *SE* of the *reduced* CER:

$$SIG_{SE} = \frac{SE_{\text{Reduced}} - SE_{\text{Full}}}{SE_{\text{Full}}} .$$

For example, consider the following CER, derived using GERM with *additive* errors:

$$f(X)_{\text{Full}} = 11.41 + 5.38\left(X_1\right)^{0.6115}\left(X_2\right)^{0.1487}\left(X_3\right)^{0.0793}(1.71)^{X_4} .$$

The standard error of the *full* CER is calculated, after optimization, as:

$$SE_{\text{Full}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[y_i - f(x_i)_{\text{Full}}\right]^2} = \$40.84\,.$$

Now suppose the exponent associated with X_1 is nullified, by setting it to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f(X)_{\text{Reduced}} = 165.47 + 2.68\times10^{-6}\left(X_1\right)^{0}\left(X_2\right)^{1.1959}\left(X_3\right)^{4.3383}\left(18.37\right)^{X_4},$$

or

$$f(X)_{\text{Reduced}} = 165.47 + 2.68\times10^{-6}\left(X_2\right)^{1.1959}\left(X_3\right)^{4.3383}\left(18.37\right)^{X_4}.$$

The standard error of the *reduced* CER is calculated, after optimization, as:

$$SE_{\text{Reduced}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[y_i - f(x_i)_{\text{Reduced}}\right]^2} = \$95.06\,.$$

In this case, the difference between the *full* and *reduced* CERs is quite large. The value calculated for SIG_{SE} is:

$$SIG_{SE} = \frac{\$95.06 - \$40.84}{\$40.84} = 1.328 = 132.8\%\,.$$

In other words, the *SE* of the *reduced* CER is well over double the *SE* of the *full* CER. This would indicate that the fit parameter for the exponent associated with X_1 is "significant."

Now, using the same *full* CER but testing a different fit parameter, nullify the exponent associated with X_3 by setting it equal to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f(X)_{\text{Reduced}} = 15.35 + 5.52\left(X_1\right)^{0.6395}\left(X_2\right)^{0.1412}\left(1.71\right)^{X_4}.$$

The standard error of the *reduced* CER is calculated, after optimization, as:

$$SE_{\text{Reduced}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[y_i - f(x_i)_{\text{Reduced}}\right]^2} = \$44.04\,.$$

In this case, the difference between the *full* and *reduced* CERs is much smaller. The value calculated for SIG_{SE} is:

$$SIG_{SE} = \frac{\$44.04 - \$40.84}{\$40.84} = 0.078 = 7.8\%\,.$$

In other words, the *SE* of the *reduced* CER is only about 8% larger than the *SE* of the *full* CER. This would indicate that the fit parameter for the exponent associated with X_3 is relatively "insignificant."

The Significance Relative to the Standard Percent Error, SIG_{SPE}

The third significance metric is the significance of the fit parameter relative to the standard percent error (*SPE*) of the CER distribution. Note that this particular discussion relates to GERM

CERs that have *multiplicative* errors. For purposes of this discussion, the *SPE* of the CER distribution is computed as part of the GERM procedure as shown below:

$$SPE = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[\frac{f(x_i)-y_i}{f(x_i)}\right]^2}\,.$$

SIG_{SPE} is defined as the percentage difference between the *SPE* of the *full* CER and the *SPE* of the *reduced* CER:

$$SIG_{SPE} = \frac{SPE_{\text{Reduced}} - SPE_{\text{Full}}}{SPE_{\text{Full}}}\,.$$

For example, consider the following CER, derived using GERM with *multiplicative* errors:

$$f(X)_{\text{Full}} = -79.65 + 31.35(X_1)^{0.3664}(X_2)^{0.1094}(X_3)^{0.0576}(1.44)^{X_4}\,.$$

The standard percent error of the *full* CER is calculated, after optimization, as:

$$SPE_{\text{Full}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[\frac{f(x_i)_{\text{Full}}-y_i}{f(x_i)_{\text{Full}}}\right]^2} = 21.4\%\,.$$

Now suppose the exponent associated with X_1 is nullified, by it to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f(X)_{\text{Reduced}} = 151.96 + 9.0322(X_1)^{0}(X_2)^{-0.3739}(X_3)^{0.5594}(3.60)^{X_4}\,,$$

or

$$f(X)_{\text{Reduced}} = 151.96 + 9.0322(X_2)^{-0.3739}(X_3)^{0.5594}(3.60)^{X_4}\,.$$

The standard percent error of the *reduced* CER is calculated, after optimization, as:

$$SPE_{\text{Reduced}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[\frac{f(x_i)_{\text{Reduced}}-y_i}{f(x_i)_{\text{Reduced}}}\right]^2} = 52.7\%\,.$$

In this case, the difference between the *full* and *reduced* CERs is quite large. The value calculated for SIG_{SPE} is:

$$SIG_{SPE} = \frac{52.7\% - 21.4\%}{21.4\%} = 1.463 = 146.3\%\,.$$

In other words, the *SPE* of the *reduced* CER is about two and a half times the *SPE* of the *full* CER. This would indicate that the fit parameter for the exponent associated with X_1 is "significant."

Now, using the same *full* CER but testing a different fit parameter, nullify the exponent associated with X_3 by setting it equal to zero. After re-optimizing, we are left with the following *reduced* CER:

$$f(X)_{\text{Reduced}} = -63.10 + 29.23(X_1)^{0.3962}(X_2)^{0.0924}(1.48)^{X_4}\,.$$

The standard percent error of the *reduced* CER is calculated, after optimization, as:

$$SPE_{\text{Reduced}} = \sqrt{\frac{1}{n-m}\sum_{i=1}^{n}\left[\frac{f(x_i)_{\text{Reduced}} - y_i}{f(x_i)_{\text{Reduced}}}\right]^2} = 22.5\%\,.$$

In this case, the difference between the *full* and *reduced* CERs is much smaller. The value calculated for SIG_{SPE} is:

$$SIG_{SPE} = \frac{22.5\% - 21.4\%}{21.4\%} = 0.051 = 5.1\%\,.$$

In other words, the *SPE* of the *reduced* CER is only about 5% larger than the *SPE* of the *full* CER. This would indicate that the exponent associated with X_3 is relatively "insignificant."

The Total Significance, SIG_{Total}

The last significance metric to be presented here represents the *total* significance of the fit parameter. The idea here is to combine the significance of the *mean* and the significance of the *SE* or *SPE* into one metric. The simplest approach is to add the absolute value of SIG_{Mean} to SIG_{SE} (or SIG_{SPE}). This will give an overall metric that combines the percentage shift in the mean with the percentage change in the variance. The equation is as follows:

$$SIG_{Total} = SIG_{SE} + \left|SIG_{Mean}\right| \text{ (for CERs with additive errors), and}$$

$$SIG_{Total} = SIG_{SPE} + \left|SIG_{Mean}\right| \text{ (for CERs with multiplicative errors).}$$

Note that there should be no need to take the absolute value of SIG_{SE} or SIG_{SPE}. These two metrics should always be non-negative. If either SIG_{SE} or SIG_{SPE} were negative, this would mean that the variance of the reduced CER was *less* than the variance of the full CER. However, this should not happen if the optimization is done correctly. If a lower variance solution of the full CER were available by setting one of the fit parameters to its null value, then the optimization should have found that solution. In other words, forcing any of the fit parameters to a value other than the value produced by the optimization of the full CER *should* result in a CER with larger variance.

SIG_{Mean}, on the other hand, can easily take on either positive or negative values, depending on the direction of the shift. Hence, we use the absolute value of SIG_{Mean} in the calculation of SIG_{Total}.

As an example, consider a fit parameter from an *additive* error CER with SIG_{SE} = 0.045 and SIG_{Mean} = -0.035. Calculating SIG_{Total} is trivial:

$$SIG_{Total} = SIG_{SE} + \left|SIG_{Mean}\right| = 0.045 + \left|-0.035\right| = 0.080 = 8\%\,.$$

In this example, the combined change due to nullifying that fit parameter is only 8%. Therefore, that fit parameter could arguably be considered as relatively "insignificant." Similarly, replace SIG_{SE} with SIG_{SPE} for a multiplicative error CER.

When is a Fit Parameter Significant? Insignificant?

An obvious question is "at what value of *SIG* is a fit parameter declared significant or insignificant?" This is still an open question, subject to interpretation, and more research is needed in order to answer it. For now, the author suggests as default values individual *SIG* values less than 5% and SIG_{Total} less than 10% as "insignificant." But, the real answer will depend on individual circumstances.

As Figure 7 shows, the "significance" metrics described herein cover a continuum of values, and therefore are relative. Clearly, however, significance values close to zero signify "insignificant" fit parameters, and values that deviate substantially from zero are "significant."

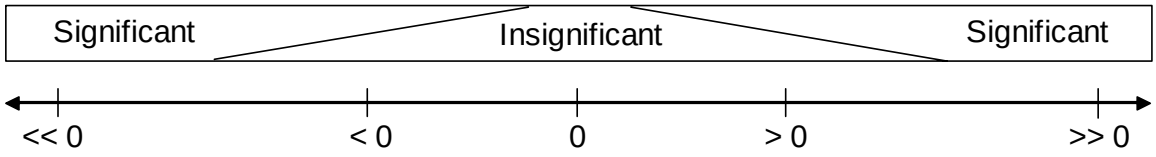

Figure 7 – "Significance" as a Continuum

Conclusions and Recommendations

This research has demonstrated at least one way of evaluating the "significance" of individual regression fit parameters related to CERs that are established using GERM. More research into this area is recommended. The metrics described in this study are comparable across CERs regardless of the functional form of the regression equation or the underlying error specification, are developed heuristically, require no distributional assumptions, and provide a collection of simple metrics by which to judge the "significance" of the individual regression fit parameters.

These metrics will be beneficial to anyone who uses GERM to develop CERS. They will enable cost modelers to judge whether or not independent variables used in CERs that are derived using GERM have any real impact on the mean result of the CER, and whether or not they contribute to the overall reduction in the CER's variance. Those independent variables that, if excluded, would not substantially impact the mean result or significantly change the CER variance, can be removed from the cost model with little or no consequence. This is often desirable because technical data collection can be difficult. Moreover, one less variable means one more degree of freedom, and that can be important when a small number of data points are available. So it is important to be able to eliminate from study those variables that are not significant cost drivers.

Areas for further study include:

1. The discovery of a "universal" value that determines whether a fit parameter is either "significant" or "insignificant."
2. Development of methods to simplify the calculation of each fit parameter's *SIG* value. E.g., is there an analytical way to determine this?

3. Determination as to whether more relevant or descriptive metrics can be derived.