Statistical Foundations of Adaptive Cost-Estimating Relationships

Stephen A. Book, MCR LLC
Melvin A. Broder, The Aerospace Corporation
Daniel I. Feldman, MCR LLC

Abstract

Traditional development of cost-estimating relationships (CERs) has been based on "full" data sets consisting of all available cost and technical data associated with a particular class of products of interest, e.g., components, subsystems or entire systems of satellites, ground systems, etc. In this paper, we review an extension of the concept of "analogy estimating" to parametric estimating, namely the concept of "adaptive" CERs – CERs that are based on specific knowledge of individual data points that may be more relevant to a particular estimating problem than would the full data set. The goal of adaptive CER development is to be able to apply CERs that have smaller estimating error and narrower prediction bounds. Several examples of adaptive CERs were provided in a paper (Reference 2) presented by the first two authors to the May 2008 SSCAG Meeting in Noordwijk, Holland, and the July 2008 ISPA/SCEA Conference in Industry Hills CA.

This paper focuses on statistical foundations of the derivation of adaptive CERs, namely the method of weighted least-squares (WLS) regression. Ordinary least-squares (OLS) regression has been traditionally applied to historical-cost data in order to derive additive-error CERs valid over an entire data range, subject to the requirement that all data points are weighted equally and have residuals that are distributed according to a common normal distribution. The idea behind adaptive CERs, however, is that data points should be "deweighted" based on some function of their distance from the point at which an estimate is to be made, i.e., each historical data point should be assigned a "weight" that reflects its importance to the particular estimation that is to be made using the derived CER. This presentation describes technical details of the WLS derivation process, resulting quality metrics, and the roles it plays in adaptive-CER development.

Introduction

Weighted least-squares (WLS) regression is the statistical technique applied in Reference 1 to develop adaptive CERs. WLS regression is a straightforward extension of classical ordinary least-squares (OLS) regression, which is the 18th Century curve-fitting technique commonly taught in elementary statistics courses.

OLS regression "best" fits a straight line $y = a + bx$ to a set of ordered pairs (x_k, y_k), $1 \le k \le n$, of data points in two-dimensional Euclidean space. We will get to the OLS definition of "best" momentarily. Procedures based on OLS philosophy and mathematical principles can extend OLS regression to the case of curved lines, primarily logarithmic, as well as a multidimensional context. However, for our purposes of deriving adaptive CERs, the linear two-dimensional context suffices.

Suppose we have n data points such as those in Table 1, labeled (x_1,y_1), (x_2,y_2), …, (x_n,y_n), where, for $1 \le k \le n$, y_k is the actual cost associated with a program whose cost driver (perhaps weight, power, etc.) is x_k. Were we to use the OLS regression line $y = a + bx$ to predict the cost of the program in question, our cost estimate would have been $a + bx_k$, rather than the actual cost y_k. The equation $y = a + bx$ is therefore called a "cost-estimating relationship" (CER).

Program	Cost-Driver Value x	Unit Cost y
A	156.12	51,367.22
B	179.40	5,885.00
C	180.30	7,060.00
D	217.50	139,483.12
E	419.14	3,386.00
F	437.09	6,738.00
G	440.93	6,812.00
H	494.45	3,291.34
I	789.90	5,723.14
J	826.10	10,992.00
K	864.30	11,590.00
L	869.30	15,973.00
M	976.50	7,970.67
N	1,355.80	9,524.10
O	1,360.90	35,927.22
P	1,463.21	11,238.73
Q	2,332.10	92,059.97
R	3,017.73	74,649.00
S	3,253.00	42,915.23

Table 1. Example of Historical Cost Data (19 Data Points)

The error in our estimate of the cost of any program is the difference $d_k = y_k - (a+bx_k) = y_k - a - bx_k$ between the actual cost y_k and the CER-estimated cost $a + bx_k$. The principle of least squares asserts that, in order to calculate the "best"-fitting straight line, we ought to choose the coefficients a and b, which determine the CER, so that the sum of squared differences (i.e., estimating errors)

$$f(a,b) = \sum_{k=1}^{n} d_k^2 = \sum_{k=1}^{n} (y_k - a - bx_k)^2$$

is as small as possible. By considering this problem as a two-dimensional minimization problem, we can take the partial derivatives of $f(a,b)$ with respect to a and b, respectively, set both partial derivatives equal to 0, and solve the resulting simultaneous equations for the two unknowns a and b. This process results in the following OLS explicit expressions for the slope b and the intercept a of the linear CER $y = a + bx$:

$$b = \frac{n\sum_{k=1}^{n} x_k y_k - \left(\sum_{k=1}^{n} x_k\right)\left(\sum_{k=1}^{n} y_k\right)}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

$$a = \frac{\sum_{k=1}^{n} y_k}{n} - b\frac{\sum_{k=1}^{n} x_k}{n}.$$

The above discussion summarizes what can be referred to as "naïve" regression. It is naïve, because a number of unstated assumptions that critically affect the nature of the CER and how it can be correctly applied are being made, often without the knowledge or concurrence of the cost analyst. The most important of these assumptions is that all n data points are and ought to be treated equally by the mathematical computations. An immediate unfortunate corollary is that extreme outlying data points, those far away from the bulk of the data and/or the cost-driver value at which the analyst wants to make an estimate, exert excessive influence on the location of the regression line and all estimates made using it.

What is it about OLS that requires us to consider each data point of equal merit? The answer to this question goes back to the early part of the 18th Century when it was mathematically derived from reasonable assumptions that estimation errors are well-modeled by the normal distribution. In fact, use of the word "normal" was introduced in the context of "the normal law of error" by Karl Pearson (1857-1936), a British scientist who was one of the founders of modern statistical theory. (It is said that Pearson later regretted his use of the word "normal," coming to believe that its common usage biased less knowledgeable analysts against other statistical distributions, which they assumed to be "abnormal" in some sense.) The theory of regression assumes that the regression line is the truth and any departures from it, e.g., those in Figure 1 below, are errors. This means that the actual y values corresponding to any particular x value are normally distributed with mean equal to the number $a + bx$. Another way of looking at the OLS regression model is as $y_k = a + bx_k + \varepsilon_k$, where ε_k is a normally distributed random variable with mean 0 and standard deviation σ.

So far so good. The killer as far as CERs are concerned, though, is the OLS requirement that all normal distributions of y values (i.e., ε_k values), one for each x value, have the same standard deviation σ. It is this requirement that forces OLS to consider all data points to be of equal merit. The requirement of equal σ values as a general rule, though, is highly questionable in the case of CERs, especially when the wide range of parameters on which CERs may be based is considered. Take a look at Figure 1. It seems clear that, for some technical reason as yet uninvestigated, cost is much more variable for cost-driver values near 300 than for other cost-driver levels. Why this happens should be studied in detail from the engineering point of view, but nevertheless we have to take account of it when estimating costs.

Figure 1 illustrates the data of Table 1, along with the OLS regression line that best fits the points in the least-squares sense. The dashed vertical lines in Figure 1 represent the distances d_k whose sum of squared values is to be minimized.

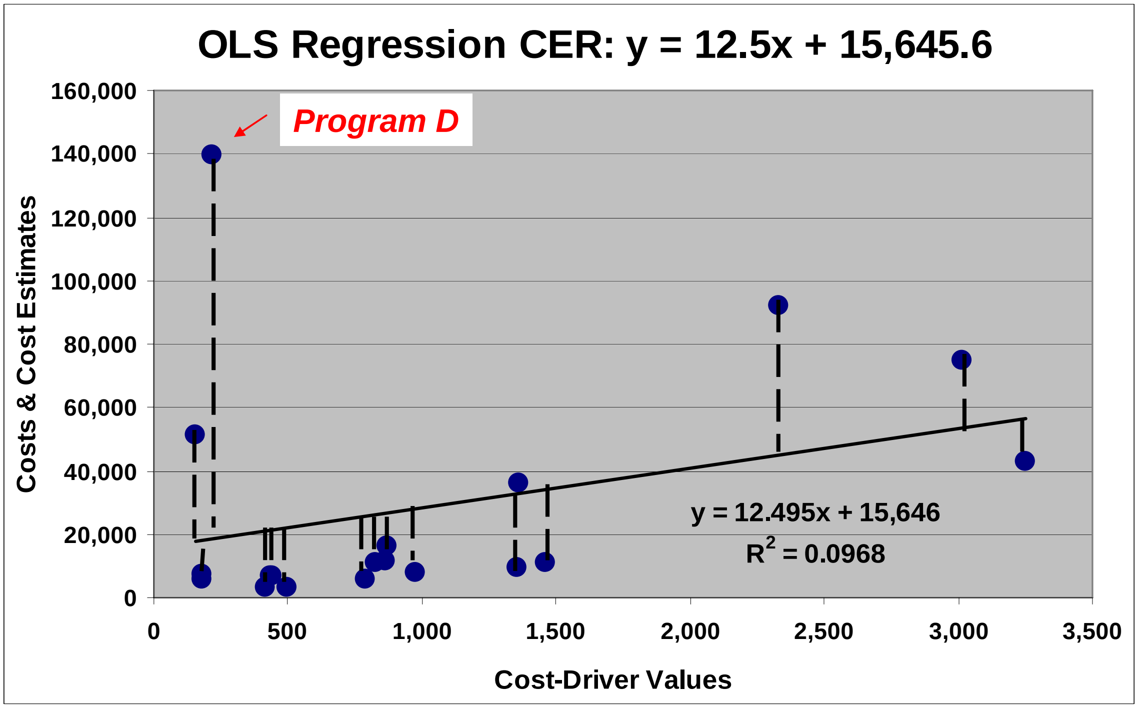

Figure 1. The Data Points of Table 1 and their OLS Regression Line

Consider the data point in Table 1 associated with Program D. From Figure 1, we see that this data point's d_k value will contribute the largest amount to the sum of squared estimating errors. In its attempt to minimize the sum of squared errors, the mathematics of OLS will take special pains to pull the regression line toward the Program D data point and thereby reduce the size of Program D's contribution to the total squared error. It is its very extremeness that gives the Program D data point its undue influence on the OLS regression line.

OLS CER Quality Metrics

Three quality metrics allow the cost analyst to assess the applicability of the CER to estimating problems involving the kinds of subsystems and/or components of which the supporting data base is comprised and the validity of estimates made using it. These three quality metrics are the following: (1) standard error of the estimate ***SEE***; (2) bias ***B***; and (3) R^2. We will discuss each of these in turn.

The standard error of the estimate ***SEE*** is an estimate of the σ value, which is the standard deviation of the normal distribution of $\varepsilon_k = y_k - a - bx_k$. Its expression is

$$SEE = \sqrt{\frac{\sum_{k=1}^{n}(y_k - a - bx_k)^2}{n-2}} = \sqrt{\frac{\sum_{k=1}^{n} y_k^2 - a\sum_{k=1}^{n} y_k - b\sum_{k=1}^{n} x_k y_k}{n-2}}\ .$$

In the OLS context, ***SEE*** is expressed in the same units as the costs and cost estimates, usually dollars. Because the coefficients of the OLS CER are calculated by minimizing the numerator under the square-root sign, the smaller the ***SEE*** turns out to be, the "better" the CER is. Choosing the denominator above as ***n-2*** makes ***SEE*** an "unbiased" estimator of σ. If the denominator were simply ***n***, ***SEE*** would be the "maximum-likelihood" estimator of σ, but not unbiased. "Unbiased" and "maximum likelihood" are statistical terms, for which we refer you to any advanced statistics text for further explanation.

The bias ***B*** of a CER is the average (sample mean) of the "residuals," namely the differences between the cost estimates and their respective actual costs, corresponding to all points in the supporting data base. In the OLS context, the bias always turns out to be zero, viz.

$$B = \frac{1}{n}\sum_{k=1}^{n}(a + bx_k - y_k) = \frac{1}{n}\sum_{k=1}^{n} a + \frac{1}{n} b\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} y_k$$

$$= \frac{1}{n} na + b\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) - \frac{1}{n}\sum_{k=1}^{n} y_k = a - \left(\frac{1}{n}\sum_{k=1}^{n} y_k - b\frac{1}{n}\sum_{k=1}^{n} x_k\right) = a - a = 0.$$

Finally, R^2, often called the coefficient of determination, is the square of the Pearson correlation between the cost estimates and their respective actual costs, corresponding to all points in the supporting data base. R^2 indicates the proportion of variation in the costs that is attributable to the OLS linear relationship between costs and cost drivers. It is usually expressed as a percentage between 0% and 100%. An R^2 of 80%, for example, means that 80% of the variation in the cost values seen in the data base is attributable to variations in the cost-driver values, while the remaining 20% of the variation is attributable to other factors not taken account of in the model, typically additional unidentified cost drivers.

Weighted Least Squares

Weighted least-squares (WLS) regression allows the cost analyst to take into account, not only the historical-cost data themselves, but also the data-collection or estimating context within which the data were gathered or the use to which any resulting CER will be put. Sometimes, the analyst will know that certain data points are less reliably known than others, so he or she can "deweight" the less reliable ones. Sometimes, the analyst will need a CER that estimates cost only within a certain cost-driver range, and then he or she can deweight data points outside that range. Once WLS theory is understood, further application contexts will almost certainly present themselves.

In addition to the actual values of cost driver and cost, each data point is assigned a weight, based on considerations discussed above, so that the set of data consist of triples (x_k,y_k,w_k), where the weight w_k represents the influence that the data point (x_k,y_k) is to have on the CER derived from the data set. In WLS regression, we weight each squared difference $d_k^2 = (y_k - (a+bx_k))^2 = (y_k - a - bx_k)^2$ by its weight w_k. We may express the principle of <u>weighted</u> least squares as choosing the numerical values of the coefficients a and b by minimizing the <u>weighted</u> sum of squared errors:

$$g(a,b) = \sum_{k=1}^{n} w_k d_k^2 = \sum_{k=1}^{n} w_k (y_k - a - bx_k)^2 .$$

What effect on the numerical values of a and b does the weighting procedure have? Well, suppose a particular value w_k is "small," indicating that we do not want the data point (x_k,y_k) to exert a major influence on the CER. Then, regardless of the choice of a and b, the term $w_k (y_k - a - bx_k)^2$ is not going to contribute too much to the sum of squared errors. Therefore, the mathematics does not have to move the regression line too close to the data point (x_k,y_k) in order to minimize the sum, because not much will be gained by making an already small summand a little smaller. On the other hand, suppose w_k is "large," indicating that we do want the corresponding data point (x_k,y_k) to exert a major influence on the CER. In this case, the term $w_k (y_k - a - bx_k)^2$ will be a major contributor to the sum of squared errors. In order to make the sum of squared errors as small as possible, a and b will have to be selected to push the resulting CER very close to the point (x_k,y_k).

Normalizing the Weights

Given an initial set of weights $\{w_1^*, w_2^*, \ldots, w_n^*\}$, we can define a new set of weights $\{w_1, w_2, \ldots, w_n\}$ that is equivalent to the initial set in the sense that the relative weights of all data points are the same as they were, but such that $\sum_{k=1}^{n} w_k = n$. The new weights are defined, for each $j = 1, 2, \ldots, n$, as $w_j = \dfrac{n w_j^*}{\sum_{k=1}^{n} w_j^*}$. Notice that, for all i and j values, the ratio $\dfrac{w_i}{w_j}$ is the same as the ratio $\dfrac{w_i^*}{w_j^*}$, i.e., the relative values of the new weights with respect to each are the same as the relative values of the original weights with respect to each other. In the sequel, we shall therefore consider all sets $\{w_1,w_2,\ldots,w_n\}$ of weights to be "normalized" in the sense that $\sum_{k=1}^{n} w_k = n$. Normalization plays a role in simplifying the expressions for the regression coefficients a and b, as is shown in the next section.

Derivation of WLS Regression Coefficients

To obtain the mathematical expression for ***a*** and ***b*** in the WLS context, we apply calculus to minimize the weighted sum of squared errors ***g(a,b)*** by first taking the partial derivatives with respect to ***a*** and ***b***:

$$\frac{\partial g}{\partial a} = \sum_{k=1}^{n} 2w_k\left(y_k - a - bx_k\right)(-1) = -2\left(\sum_{k=1}^{n} w_k y_k - a\sum_{k=1}^{n} w_k - b\sum_{k=1}^{n} w_k x_k\right)$$

and

$$\frac{\partial g}{\partial b} = \sum_{k=1}^{n} 2w_k\left(y_k - a - bx_k\right)(-x_k) = -2\left(\sum_{k=1}^{n} w_k x_k y_k - a\sum_{k=1}^{n} w_k x_k - b\sum_{k=1}^{n} w_k x_k^2\right)$$

Setting the two partial derivatives equal to ***0***, we obtain the following two simultaneous equations in the unknowns ***a*** and ***b***:

$$a\sum_{k=1}^{n} w_k + b\sum_{k=1}^{n} w_k x_k = \sum_{k=1}^{n} w_k y_k$$

$$a\sum_{k=1}^{n} w_k x_k + b\sum_{k=1}^{n} w_k x_k^2 = \sum_{k=1}^{n} w_k x_k y_k \ .$$

The solution to these equations is

$$b = \frac{\left(\sum_{k=1}^{n} w_k\right)\left(\sum_{k=1}^{n} w_k x_k y_k\right) - \left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)}{\left(\sum_{k=1}^{n} w_k\right)\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{n} w_k x_k\right)^2}$$

$$a = \frac{\left(\sum_{k=1}^{n} w_k y_k\right)}{\left(\sum_{k=1}^{n} w_k\right)} - b\frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{\left(\sum_{k=1}^{n} w_k\right)} .$$

Because the weights are normalized, the expressions for ***b*** and ***a*** can be reduced to, respectively,

$$b = \frac{n\left(\sum_{k=1}^{n} w_k x_k y_k\right) - \left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)}{n\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{n} w_k x_k\right)^2}$$

$$a = \frac{\left(\sum_{k=1}^{n} w_k y_k\right)}{n} - b\frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{n}$$

It is should be noted that when all w_k values are equal (i.e., all equal to 1 assuming normalization), the WLS expressions for a and b reduce to the OLS expressions. In addition, we refer to the expressions

$$\bar{x}_w = \frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{n} \quad \text{and} \quad \bar{y}_w = \frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{n}$$

as the "weighted means" of the x and y values, respectively. Note that the expression for a guarantees that the point $(\bar{x}_w, \bar{y}_w)$ falls exactly on the WLS regression line. Again, when each $w_k = 1$ or, more specifically, when all w_k values are equal, the expressions for the weighted means reduce to the expressions for the ordinary means (i.e., the averages) of x and y.

WLS CER Quality Metrics

The same three quality metrics used for OLS allow the cost analyst to assess the applicability of the WLS CER to estimating problems involving the kinds of subsystems and/or components of which the supporting data base is comprised and the validity of estimates made using it. These three quality metrics are again the following: (1) standard error of the estimate SEE_w; (2) bias B_w; and (3) R_w^2. However, as one would expect, the formulas for them are slightly different in the WLS situation.

Because there is nothing in the WLS setup that plays the OLS role of σ, we consider the standard error of the estimate SEE_w to measure the closeness of the estimated costs $a + bx_k$ to the actual costs y_k in the data base. Its expression is

$$SEE_w = \sqrt{\frac{\sum_{k=1}^{n} w_k (y_k - a - bx_k)^2}{\sum_{k=1}^{n} w_k - 2}} = \sqrt{\frac{\sum_{k=1}^{n} w_k y_k^2 - a\sum_{k=1}^{n} w_k y_k - b\sum_{k=1}^{n} w_k x_k y_k}{n-2}}.$$

In the WLS context, SEE_w is expressed in the same units as the costs and cost estimates, usually dollars. Because the coefficients of the WLS CER are calculated by minimizing the numerator under the square-root sign, the smaller SEE_w turns out to be, the "better" the CER is. Because the weights are normalized, the denominator reduces to $n-2$. If all weights are equal, SEE_w reduces to the unbiased form of the OLS SEE.

The bias B_w of a CER is the weighted mean of the "residuals," namely the differences between the cost estimates and their respective actual costs, corresponding to all points in the supporting data base. As noted earlier, in the OLS context, the bias always turns out to be zero, but this is not true in the WLS context.

$$
\begin{aligned}
B_w &= \frac{1}{n}\sum_{k=1}^{n}(a + bx_k - y_k) = \frac{1}{n}\sum_{k=1}^{n} a + \frac{1}{n} b\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} y_k \\
&= \frac{1}{n} na + b\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) - \frac{1}{n}\sum_{k=1}^{n} y_k = a + b\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) - \frac{1}{n}\sum_{k=1}^{n} y_k \\
&= a - \frac{\left(\sum_{k=1}^{n} y_k\right) - b\left(\sum_{k=1}^{n} x_k\right)}{n} \quad a = \frac{\left(\sum_{k=1}^{n} w_k y_k\right)}{n} - b\frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{n} - \frac{\left(\sum_{k=1}^{n} y_k\right) - b\left(\sum_{k=1}^{n} x_k\right)}{n} \\
&= \frac{\left(\sum_{k=1}^{n} w_k y_k\right)}{n} - \frac{\left(\sum_{k=1}^{n} y_k\right)}{n} - b\frac{\left(\sum_{k=1}^{n} w_k x_k\right)}{n} + \frac{\left(\sum_{k=1}^{n} x_k\right)}{n} \\
&= \frac{\left(\sum_{k=1}^{n} (w_k - 1)y_k\right)}{n} - b\frac{\left(\sum_{k=1}^{n} (w_k - 1)x_k\right)}{n}
\end{aligned}
$$

which reduces to 0 when all $w_k = 1$ or, more specifically, are all the same when normalized. However, the bias is, in general, not typically zero in the weighted least-squares situation.

Finally, R^2, just as in the OLS situation, measures the worth of the linear-regression equation as a model of the relationship underlying the data base. To derive the formula for R^2 in the WLS situation, let's start with some reasoning that applies in the OLS situation. Referring to the data points (x_1,y_1), (x_2,y_2), …, (x_n,y_n), we ask why the y values vary, i.e., why are they not all the same. There are two basic reasons that the y values vary: (1) the x values vary, and y is related to x through the hypothesized linear relationship, and (2) any other reason you can think of that does not involve the hypothesized linear relationship, e.g., nonlinearity, random errors in the data, additional cost drivers, that affects y. What R^2 does is to allocate the variation in y between these two sources. In particular R^2, usually expressed as a percentage, indicates the proportion of variation in y that is attributable to the linear relationship between x and y.

If the y values did not vary at all from the WLS regression line, they all would be equal to their weighted mean $\bar{y}_w = \left(\sum_{k=1}^{n} w_k y_k\right) / n$. If, on the other hand, we had no knowledge at all about the relationship between x and y, the best we could do to predict the value y at any given x would be to predict $y = \bar{y}_w$. This is equivalent to using the horizontal line $y = \bar{y}_w$ in place of the regression line $y = a + bx$. The sum of squared errors from the horizontal line $y = \bar{y}_w$ is called the "total variation" of y and is denoted $TV = \sum_{k=1}^{n} w_k (y_k - \bar{y}_w)^2$.

Suppose now that the only variation in y were due to the influence of the regression line $y = a + bx$. Then every y_k would be equal to its corresponding $a+bx_k$. The resulting total variation would then be

$$\sum_{k=1}^{n} w_k (y_k - \bar{y}_w)^2 = \sum_{k=1}^{n} w_k (a + bx_k - \bar{y}_w)^2$$

since each y_k and $a+bx_k$ would be one and the same. It would follow that the quantity $VR = \sum_{k=1}^{n} w_k (a + bx_k - \bar{y}_w)^2$, called the "variance due to regression" is the variation in y that can be attributed to the impact of the regression relationship.

We then compare TV and VR with the weighted sum of squared (SS) errors, where SS = $\sum_{k=1}^{n} w_k (y_k - a - bx_k)^2$. It can be proved by elementary, though tedious, calculations that $TV = SS+VR$. These calculations are reproduced in the Appendix. Simple algebra then ensures that $\frac{SS}{TV} + \frac{VR}{TV} = 1$. From this equation, it is evident that VR/TV is the proportion of the total variation in y that can be attributed to the impact of the linear-regression relationship. The proportion of variation in y due to all other effects is equal to SS/TV. The WLS coefficient of determination is then

$$R_w^2 = \frac{VR}{TV} = \frac{\sum_{k=1}^{n} w_k (a + bx_k - \bar{y}_w)^2}{\sum_{k=1}^{n} w_k (y_k - \bar{y}_w)^2} = \frac{\sum_{k=1}^{n} w_k (a + bx_k - a - b\bar{x}_w)^2}{\sum_{k=1}^{n} w_k (y_k^2 - 2y_k \bar{y}_w + \bar{y}_w^2)}$$

$$= \frac{b^2 \sum_{k=1}^{n} w_k (x_k - \bar{x}_w)^2}{\sum_{k=1}^{n} w_k y_k^2 - 2\bar{y}_w \sum_{k=1}^{n} w_k y_k + n\bar{y}_w^2} = \frac{b^2 \left(\sum_{k=1}^{n} w_k x_k^2 - 2\bar{x}_w \sum_{k=1}^{n} w_k x_k + n\bar{x}_w^2 \right)}{\sum_{k=1}^{n} w_k y_k^2 - 2\bar{y}_w \sum_{k=1}^{n} w_k y_k + n\bar{y}_w^2}$$

$$= \frac{b^2 \left(\sum_{k=1}^{n} w_k x_k^2 - \left(\sum_{k=1}^{n} w_k x_k \right)^2 / n \right)}{\sum_{k=1}^{n} w_k y_k^2 - \left(\sum_{k=1}^{n} w_k y_k \right)^2 / n}$$

$$= \frac{\left\{ n\left(\sum_{k=1}^{n} w_k x_k y_k \right) - \left(\sum_{k=1}^{n} w_k x_k \right)\left(\sum_{k=1}^{n} w_k y_k \right) \right\}^2}{\left\{ n\left(\sum_{k=1}^{n} w_k x_k^2 \right) - \left(\sum_{k=1}^{n} w_k x_k \right)^2 \right\}^2} \times \frac{\sum_{k=1}^{n} w_k x_k^2 - \left(\sum_{k=1}^{n} w_k x_k \right)^2 / n}{\sum_{k=1}^{n} w_k y_k^2 - \left(\sum_{k=1}^{n} w_k y_k \right)^2 / n}$$

$$R_w^2 = \frac{\left\{ n\left(\sum_{k=1}^{n} w_k x_k y_k\right) - \left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)\right\}^2}{\left\{ n\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{n} w_k x_k\right)^2\right\}\left\{ n\left(\sum_{k=1}^{n} w_k y_k^2\right) - \left(\sum_{k=1}^{n} w_k y_k\right)^2\right\}}$$

Adaptive CERs via Quadratic-Distance Weighting

An "adaptive" CER is an extension of the concept of analogy estimating to the CER context. The standard way doing analogy estimating is by finding one historical program that has several characteristics in common with the subsystems or components of a program that is being estimated, for example, the program's objective, hardware or software design proposed to carry it out, materials of which any hardware is constructed, use of similar legacy components, and Government or contractor approach to program development or production. The idea behind an adaptive CER is to build a data base consisting of as many programs as we can find that have subsystems or components of the same basic kind as in the program being estimated. Normally, we would use all the points of this data base to derive a CER that expresses the subsystem or component cost in terms of an appropriate cost-driver.

However, in any particular estimating context, we are interested only in one particular value of the cost driver or, at most, a relatively short interval of such values. We know from classical OLS theory (see below) that, if the value at which we are interested in estimating is relatively far away from the cost-driver values in the data base, the accuracy of our estimate is substantially reduced. Adaptive CERs look at the flip side of this situation: If a cost-driver value of a data point is relatively far away from the value at which we want to do our estimate, maybe we don't want to use that data point to calculate our CER or, at least, maybe we don't want to consider it of equal weight with data points whose cost-driver values are closer to where we want to estimate.

The mechanics of calculating adaptive CERs is therefore based on measurements of the distance between cost-driver values in the data base and the cost-driver value at which we want to conduct our estimate. Data points are treated differently, according to their distance from the estimating point. To carry out the process, we assign each point in the data base a "weight" that indicates how important that data point is to our estimating problem. Then we apply "weighted least-squares" (WLS) regression to derive the CER.

For purposes of illustration in this paper, we shall consider quadratic-distance weighting. This weighting method calls for weighting points according to the squared distance of its cost-driver value along the x-axis from a cost-driver value of interest. If x_0 is the cost-driver value of interest and x_k is the cost-driver value of the k^{th} data point, then $QD_k = (x_0 - x_k)^2$ is the squared distance between the two cost-driver values. Because the greater that distance is, the less we want its weight to be, we define the weight of the data point (x_k, y_k) to be the reciprocal of QD_k, namely $w_k = (x_0 - x_k)^{-2}$.

Why choose quadratic-distance weighting from among the infinite number of ways to define the weighting in terms of a cost driver's distance from x_0? We prefer the squared (quadratic) distance, because OLS calculations use the squares of residuals for best fit – this process forces the CER to pass through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean of the cost-driver values and $\bar{y}$ is the mean of the cost values in the data base. In the WLS case, the regression line based on minimizing the squares of residuals passes through the point $(\bar{x}_w, \bar{y}_w)$, where $\bar{x}_w = \left(\sum_{k=1}^{k} w_k x_k\right) \div \left(\sum_{k=1}^{k} w_k\right)$ is the weighted mean of the cost-driver values and $\bar{y}_w = \left(\sum_{k=1}^{k} w_k y_k\right) \div \left(\sum_{k=1}^{k} w_k\right)$ is the weighted mean of the cost values. However, other weighting schemes can be used if there is a compelling reason to do so.

Starting with the historical-cost data in Table 1, suppose we want to estimate the cost of a similar subsystem or component of interest whose cost-driver value is 800. We then weight each of the data points according to the quadratic distance of its cost-driver value from 800. The results are listed in Table 2. Note that the normalized weights sum to 19, which is the number of data points.

Program	Cost-Driver Value x	Unit Cost y	Initial Weight w	Normalized Weight w
A	156.12	51,367.22	0.00000241	0.003881827
B	179.40	5,885.00	0.00000260	0.004178521
C	180.30	7,060.00	0.00000260	0.004190667
D	217.50	139,483.12	0.00000295	0.004743012
E	419.14	3,386.00	0.00000689	0.011094695
F	437.09	6,738.00	0.00000759	0.012219353
G	440.93	6,812.00	0.00000776	0.012482106
H	494.45	3,291.34	0.00001071	0.017237787
I	789.90	5,723.14	0.00980296	15.77623429
J	826.10	10,992.00	0.00146798	2.362463352
K	864.30	11,590.00	0.00024187	0.389245992
L	869.30	15,973.00	0.00020823	0.335104011
M	976.50	7,970.67	0.00003210	0.05166027
N	1,355.80	9,524.10	0.00000324	0.005209656
O	1,360.90	35,927.22	0.00000318	0.005115348
P	1,463.21	11,238.73	0.00000227	0.003658845
Q	2,332.10	92,059.97	0.00000043	0.000685602
R	3,017.73	74,649.00	0.00000020	0.000327212
S	3,253.00	42,915.23	0.00000017	0.000267455
Sums	19,633.77	542,585.74	0.01180613	19.00000000

Table 2. Historical-Cost Data Weighted According to their Quadratic Distances from 800

The next step is to calculate the adaptive CER, i.e., the CER adapted to estimating at a cost-driver value of 800. We apply WLS methods to derive this CER, i.e., using the formulas for ***a*** and ***b*** derived earlier. The required preliminary computations appear in Table 3.

Cost-Driver Value of Interest =			800						
Program	Cost-Driver Value x	Unit Cost y	Normalized Weight w	wx	wy	wx^2	wy^2	wxy	WLS EST y
A	156.12	51,367.22	0.00388183	0.61	199.40	94.61	10,242,556.04	31,130.12	-5,734.14
B	179.40	5,885.00	0.00417852	0.75	24.59	134.48	144,715.65	4,411.55	-5,280.91
C	180.30	7,060.00	0.00419067	0.76	29.59	136.23	208,877.91	5,334.37	-5,263.39
D	217.50	139,483.12	0.00474301	1.03	661.57	224.37	92,277,865.87	143,891.50	-4,539.16
E	419.14	3,386.00	0.01109470	4.65	37.57	1,949.10	127,200.63	15,745.68	-613.53
F	437.09	6,738.00	0.01221935	5.34	82.33	2,334.48	554,766.51	35,987.37	-264.07
G	440.93	6,812.00	0.01248211	5.50	85.03	2,426.76	579,211.44	37,491.44	-189.31
H	494.45	3,291.34	0.01723779	8.52	56.74	4,214.31	186,735.55	28,052.83	852.64
I	789.90	5,723.14	15.77623429	12,461.65	90,289.60	9,843,455.33	516,740,007.13	71,319,753.08	6,604.62
J	826.10	10,992.00	2.36246335	1,951.63	25,968.20	1,612,242.35	285,442,423.23	21,452,327.68	7,309.38
K	864.30	11,590.00	0.38924599	336.43	4,511.36	290,772.40	52,286,674.49	3,899,169.35	8,053.07
L	869.30	15,973.00	0.33510401	291.31	5,352.62	253,232.23	85,497,341.14	4,653,029.40	8,150.42
M	976.50	7,970.67	0.05166027	50.45	411.77	49,260.77	3,282,058.62	402,090.44	10,237.44
N	1,355.80	9,524.10	0.00520966	7.06	49.62	9,576.36	472,559.94	67,271.11	17,621.85
O	1,360.90	35,927.22	0.00511535	6.96	183.78	9,473.87	6,602,713.33	250,106.54	17,721.14
P	1,463.21	11,238.73	0.00365884	5.35	41.12	7,833.53	462,145.19	60,168.32	19,712.96
Q	2,332.10	92,059.97	0.00068560	1.60	63.12	3,728.78	5,810,500.30	147,193.92	36,628.96
R	3,017.73	74,649.00	0.00032721	0.99	24.43	2,979.82	1,823,378.13	73,711.14	49,977.16
S	3,253.00	42,915.23	0.00026746	0.87	11.48	2,830.21	492,576.73	37,337.61	54,557.52
Sums	19,633.77	542,585.74	19.00000000	15,141.45	128,083.89	12,096,899.99	1,063,234,307.81	102,664,203.45	215,542.66
			Num b =	11,243,876.6334		Std Error =	3,147.8208		
			Den b =	577,541.5425		Num R^2 =	126,424,761,747,155.0000		
			b =	19.4685		Den R^2 =	2,192,330,157,360,000.0000		
			Wtd Mean x =	796.9185		R^2 =	5.7667%		
			Wtd Mean y =	6,741.2572					
			a =	-8,773.5633					

Table 3. WLS Computations Leading to Adaptive CER at a Cost-Driver Value of 800

Figure 2 compares the full-data-set CER with the CER adapted, via quadratic-distance weighting, to a cost-driver value of ***800***. It should be noticed that the standard error of the full-data-set CER is ***34,336.83***, while the standard error of the adaptive CER with points far from ***800*** deweighted considerably is only ***3,147.82***, a decrease in magnitude of over 90%.

Note also that the adaptive CER ***y = -8,773.56 + 19.4685x*** appears to estimate more accurately around ***x = 800***, while essentially ignoring data points whose ***x*** values are far removed from ***800***. This view is supported by the relative values of the standard errors of both CERs.

For additional illustration, we compare in Figure 3 the full-data-set CER with the CER adapted, via quadratic-distance weighting, to a cost-driver value of ***300***. It is still true, of course, that the standard error of the full-data-set CER is ***34,336.83***, while the standard error of the adaptive CER with points far from 300 deweighted considerably and those near 300 more heavily weighted is now ***55,556.56***. This large standard error undoubtedly occurs, because the actual data points vary quite a bit near the ***300*** cost-driver value. In Figure 4, we compare the full-data-set CER with the CER adapted, via quadratic-distance weighting, to a cost-driver value of ***3,000***. While the standard error of the full-data-set CER remains at ***34,336.83***, the standard error of the adaptive CER with points far from 3,000 deweighted is now ***2,838.37***.

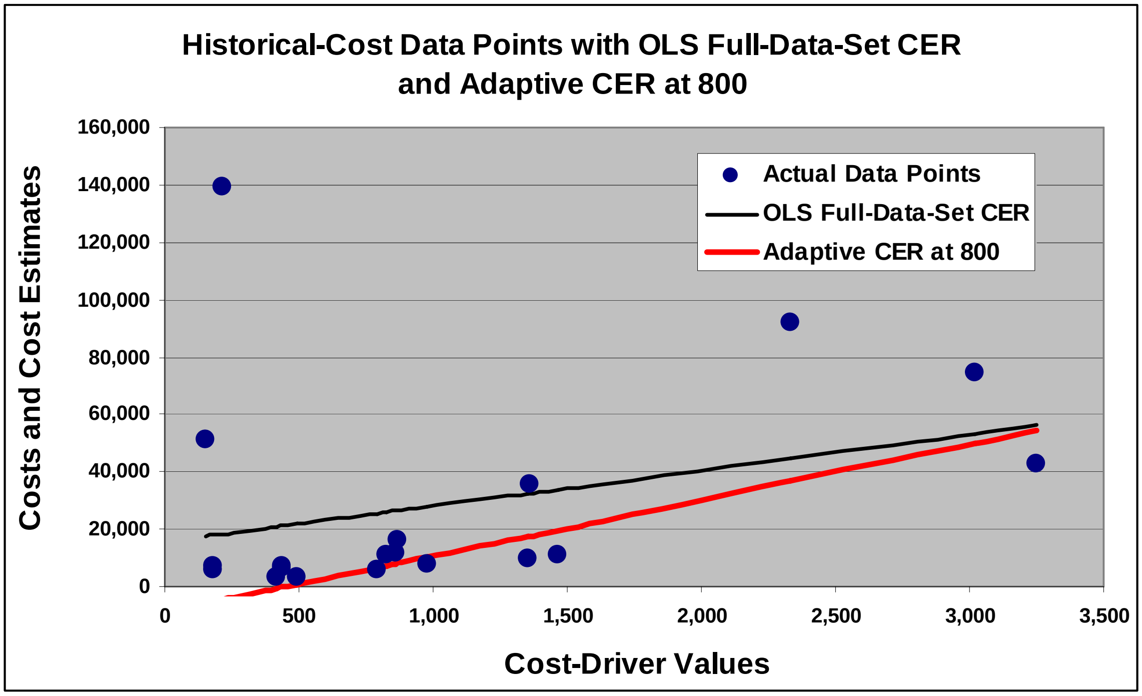

Figure 2. OLS Full-Data-Set CER Compared with Adaptive CER at a Cost-Driver Value of 800

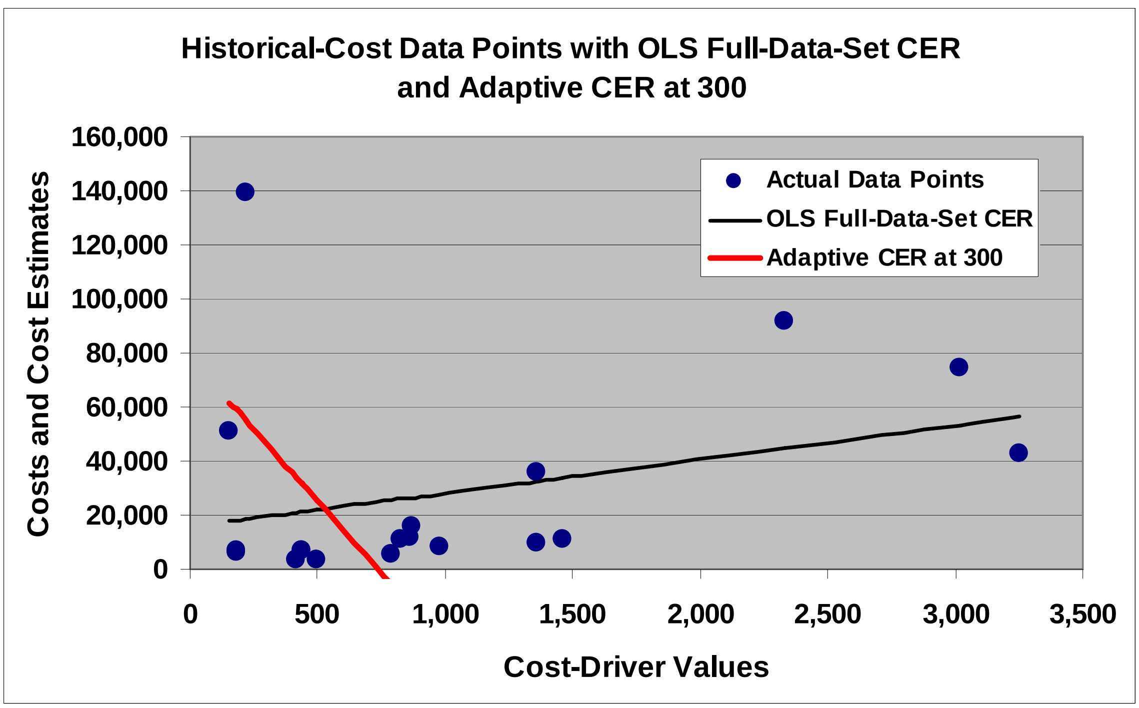

Figure 3. OLS Full-Data-Set CER Compared with Adaptive CER at a Cost-Driver Value of 300

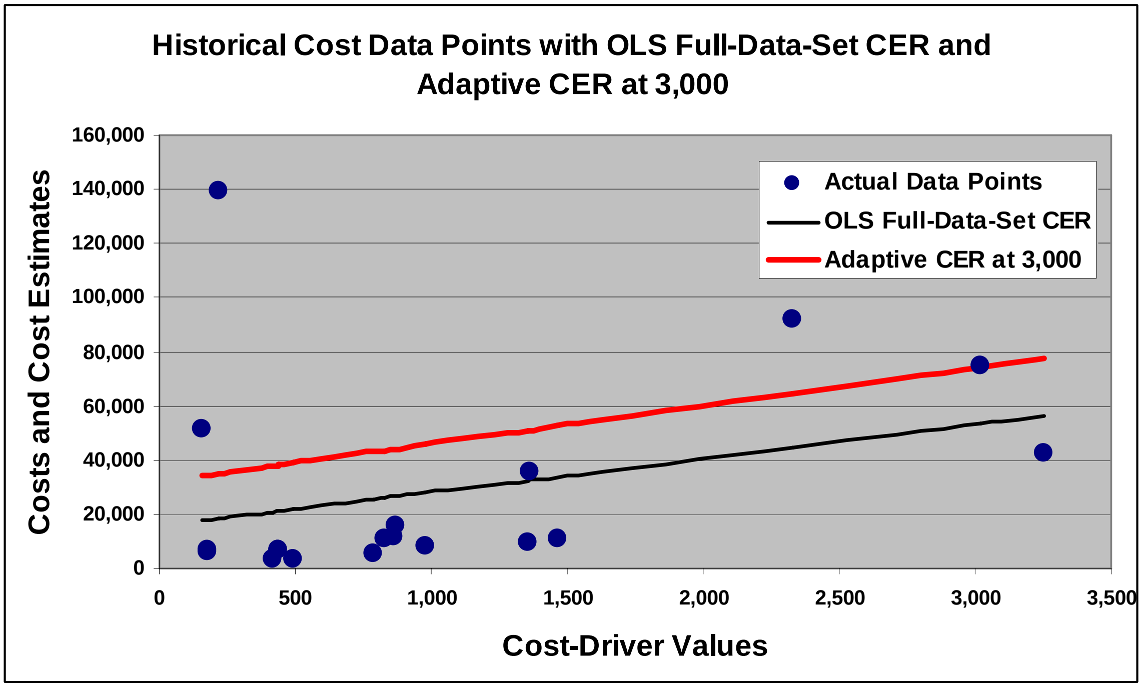

Figure 4. OLS Full-Data-Set CER Compared with Adaptive CER at a Cost-Driver Value of 3,000

The "Universal Adaptive CER"

The "universal adaptive CER" is formed by combining* the various individual adaptive CERs, of the sort derived above, over the range of cost drivers into one CER that applies over the entire range. This "universal adaptive CER" is, as P. Foussier (Reference 3, Chart 5) presciently noted, "highly nonlinear." For the data set we have been working with, we can consider the cost-driver range to go from 50 to 3,500, and we calculate a quadratic-distance-weighted CER and an estimated cost at each increment of 50 for each of those cost-driver values. Then we string all these estimates together and interpolate between successive ones to form the universal adaptive CER.

To complete the picture of estimating at each point along the cost-driver axis, we record and graph the standard error at each point as well. Table 4 contains the estimates and standard errors at 50 units apart along the cost-driver axis. The numbers in Table 4 form the basis for the graphs of the universal adaptive CER and the corresponding standard errors in Figure 5. For comparison purposes, the standard error of the OLS CER is a constant ***34,336.83*** across the data base. Notice how the standard error of the universal adaptive CER varies with the distance of the cost-driver value (x axis) from the nearest point in the data base. The numbers in red (between the 50-unit points) in Table 4 identify the actual data points underlying the analysis.

----

* The idea of combining estimates at various points of the cost-driver range into one all-inclusive CER was suggested to us by Paul Wetzel of OpsConsulting LLC.

Driver	EST Cost	Std Error
50.00	42,739.31	46,098.71
100.00	40,817.29	41,490.92
150.00	49,546.82	15,013.91
156.12	50,880.53	20,862.57
179.40	55,953.88	43,110.41
180.30	56,150.02	43,970.50
200.00	60,443.18	62,797.07
217.50	69,749.17	63,712.78
250.00	87,031.73	65,413.39
300.00	46,425.71	57,676.55
350.00	22,733.56	36,873.63
400.00	7,006.95	11,986.04
419.14	6,760.42	9,109.80
437.09	6,529.22	6,412.39
440.93	6,479.76	5,835.34
450.00	6,362.94	4,472.36
494.45	3,589.46	3,084.58
500.00	3,243.16	2,911.31
550.00	6,829.12	17,776.83
600.00	9,959.40	22,010.11
650.00	11,310.17	21,033.96
700.00	10,929.01	16,492.92
750.00	8,652.67	9,456.12
789.90	7,175.24	4,565.75
800.00	6,801.25	3,327.84
826.10	9,756.59	3,386.63
850.00	12,462.82	3,440.47
864.30	12,666.50	4,059.71
869.30	12,737.72	4,276.23
900.00	13,174.99	5,605.64
950.00	9,208.15	5,651.88
976.50	8,832.68	5,342.38
1,000.00	8,499.71	5,067.91
1,050.00	11,462.16	11,841.54
1,100.00	14,296.49	15,323.02
1,150.00	16,537.15	16,912.27
1,200.00	18,230.99	17,020.52
1,250.00	19,495.31	16,029.95
1,300.00	20,310.23	14,631.94
1,350.00	14,974.31	11,522.07
1,355.80	15,774.27	11,821.74
1,360.90	16,477.67	12,085.24
1,400.00	21,870.45	14,105.41
1,450.00	11,840.86	4,214.92
1,463.21	12,101.01	5,274.84

Driver	EST Cost	Std Error
1,500.00	12,825.54	8,226.72
1,550.00	16,621.72	13,974.93
1,600.00	20,492.26	17,569.25
1,650.00	24,526.56	20,350.34
1,700.00	28,831.03	22,668.31
1,750.00	33,415.50	24,632.61
1,800.00	38,247.16	26,275.33
1,850.00	43,285.50	27,589.48
1,900.00	48,497.85	28,534.71
1,950.00	53,862.57	29,032.00
2,000.00	59,364.10	28,954.26
2,050.00	64,981.01	28,118.23
2,100.00	70,666.52	26,286.86
2,150.00	76,319.27	23,197.58
2,200.00	81,744.09	18,634.07
2,250.00	86,609.89	12,543.91
2,300.00	90,430.47	5,163.31
2,332.10	91,836.14	3,730.10
2,350.00	92,619.98	2,930.89
2,400.00	92,676.25	10,907.76
2,450.00	90,463.37	17,895.26
2,500.00	86,410.39	23,227.16
2,550.00	81,412.53	26,603.62
2,600.00	76,466.46	28,091.64
2,650.00	72,322.92	27,995.50
2,700.00	69,366.76	26,697.11
2,750.00	66,431.86	24,540.98
2,800.00	67,242.40	21,772.29
2,850.00	67,904.22	18,495.58
2,900.00	69,545.45	14,613.82
2,950.00	71,913.26	9,720.21
3,000.00	74,219.40	3,000.69
3,017.73	74,164.83	4,164.89
3,050.00	74,065.53	6,283.82
3,100.00	67,141.02	15,848.64
3,150.00	54,415.99	17,689.83
3,200.00	45,424.15	9,943.35
3,250.00	42,927.10	501.90
3,253.00	42,978.65	868.74
3,300.00	43,786.36	6,615.99
3,350.00	45,762.39	11,482.72
3,400.00	47,971.96	14,864.14
3,450.00	50,126.95	17,319.87
3,500.00	52,149.51	19,185.52

Table 4. Universal Adaptive-CER-Based Estimates and Standard Errors at 50-Unit Increments Along the Cost-Driver Axis

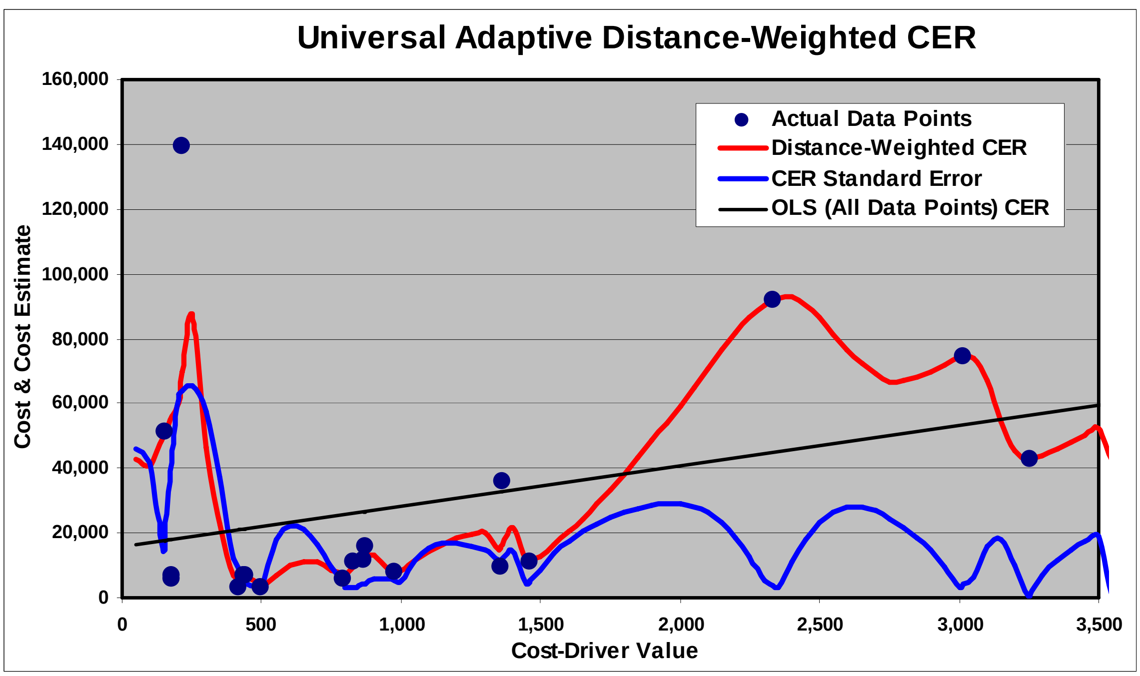

Figure 5. Universal Adaptive-CER-Based Estimates and Standard Errors Graphed at 50-Unit Increments along the Cost-Driver Axis

Prediction Bounds

Estimating the cost of developing or producing a new subsystem or component is essentially trying to predict the future, which means that any such estimate contains uncertainty. A portion of this uncertainty is described by the "standard error of the estimate" of a cost-estimating relationship (CER), which is basically the standard deviation of errors made (the "residuals") in using that CER to estimate the (known) costs of the subsystems or components comprising the supporting historical data base. The standard error of the estimate depends primarily on the extent to which those (known) costs fit the CER that purports to model them. However, additional uncertainty arises from the location of the particular cost-driver value (x) within or without the range of cost-driver values for programs comprising the historical cost data base. For example, if x were located near the center of the range of its historical values, the CER would provide a more precise measure of the element's cost than if x were located far from the center of the range. The total uncertainty in the estimate can then be expressed in terms of prediction bounds that involve both sources of uncertainty.

The first kind of uncertainty, represented by only one number characteristic of the CER, is fairly easy to measure for any CER shape or error model. The second kind, which involves both the CER itself and the value of the cost-driving parameter, however, is more complicated, and the way to calculate it is completely understood only in the case of classical OLS linear regression. As a result, an explicit formula exists for "prediction intervals" that bound cost estimates based on CERs that have been derived by applying OLS to historical cost data. In fact, the formula for the $(1-\alpha)^{th}$ percent upper and lower prediction bounds on the true cost y, based on the estimate $ESTy$ from the CER is the following:

$$ESTy \pm t_{\alpha/2,n-2} * SEE\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

where $t_{\alpha/2,n-2}$ is the $(1-\alpha)^{th}$ percentage point of the t distribution, $\bar{x}$ is the mean of the cost-driver values in the data base, x is the cost-driver value at which the estimate is being made, and SEE is the standard error of the estimate. Table 5 displays the sequence of 80% upper and lower prediction bounds for the OLS CER based on our data set. Figure 6 graphs the prediction bounds, along with the actual data points and the OLS CER.

Program	Cost-Driver Value x	Unit Cost y	80% Upper Bound	OLS EST y	80% Lower Bound
A	156.12	51,367.22	65,673.53	17,596.30	-30,480.93
B	179.40	5,885.00	65,907.23	17,887.18	-30,132.88
C	180.30	7,060.00	65,916.29	17,898.42	-30,119.45
D	217.50	139,483.12	66,292.88	18,363.23	-29,566.43
E	419.14	3,386.00	68,400.42	20,882.67	-26,635.08
F	437.09	6,738.00	68,593.51	21,106.95	-26,379.62
G	440.93	6,812.00	68,634.94	21,154.93	-26,325.09
H	494.45	3,291.34	69,216.65	21,823.65	-25,569.35
I	789.90	5,723.14	72,574.56	25,515.22	-21,544.12
J	826.10	10,992.00	73,003.23	25,967.53	-21,068.17
K	864.30	11,590.00	73,459.69	26,444.83	-20,570.03
L	869.30	15,973.00	73,519.75	26,507.30	-20,505.14
M	976.50	7,970.67	74,824.83	27,846.74	-19,131.35
N	1,355.80	9,524.10	79,710.04	32,586.00	-14,538.05
O	1,360.90	35,927.22	79,778.56	32,649.72	-14,479.12
P	1,463.21	11,238.73	81,168.85	33,928.06	-13,312.74
Q	2,332.10	92,059.97	94,145.23	44,784.62	-4,576.00
R	3,017.73	74,649.00	105,728.61	53,351.39	974.17
S	3,253.00	42,915.23	109,940.12	56,291.03	2,641.94

Table 5. 80% Upper and Lower OLS Prediction Bounds

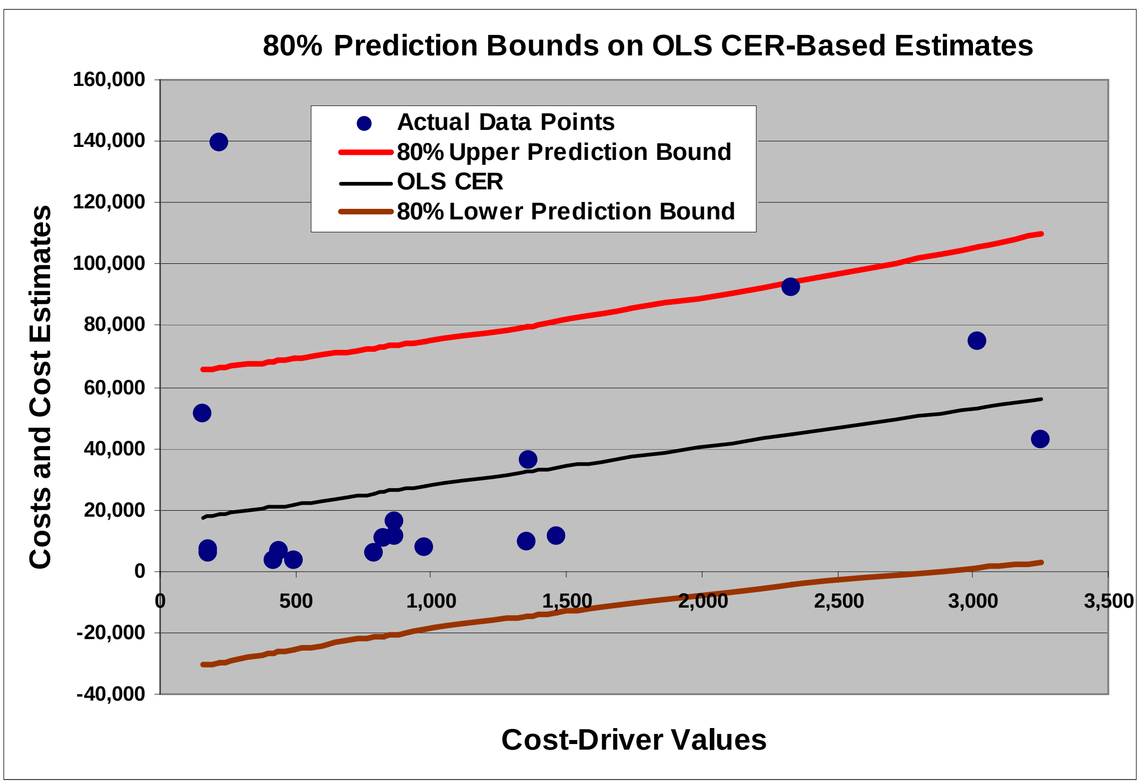

Figure 6. 80% OLS Prediction Bounds with Actual Data Points and OLS CER

When the weights are normalized, the expressions for the $(1-\alpha)^{th}$ percent upper and lower prediction bounds on the true cost y at the cost-driver value x_p, based on estimates ***ESTy*** from WLS-based adaptive CERs are the following:

$$ESTy \pm t_{\alpha/2,n-2} * SEE_w \sqrt{\frac{1}{w_p} + \frac{1}{n} + \frac{n(x_p - \bar{x})^2}{n\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{k} w_k x_k\right)^2}}$$

One way to obtain a usable value, if needed, for w_p when x_p is not in the data base from which the adaptive CERs are derived is to interpolate between the weights of the nearest data-base points. That is what is effectively done in the graphs based on Tables 6, 7, and 8 below.

In Table 6, 7, and 8 we compile the 80% upper and lower prediction bounds on adaptive CERs at the cost-driver values, respectively, of 800, 300, and 3,000. Figures 7, 8, and 9 display the graphs of these respective prediction bounds. Notice how the prediction bounds narrow in the region very near the cost-driver value of interest.

Program	Cost-Driver Value x	Unit Cost y	80% Upper Bound	WLS EST y	80% Lower Bound
A	156.12	51,367.22	67,335.731428	-5,734.14	-78,804.008697
B	179.40	5,885.00	65,146.948025	-5,280.91	-75,708.771200
C	180.30	7,060.00	65,062.330513	-5,263.39	-75,589.110360
D	217.50	139,483.12	61,564.835765	-4,539.16	-70,643.158038
E	419.14	3,386.00	42,608.518817	-613.53	-43,835.578046
F	437.09	6,738.00	40,921.251654	-264.07	-41,449.391167
G	440.93	6,812.00	40,560.306422	-189.31	-40,938.927733
H	494.45	3,291.34	35,529.986321	852.64	-33,824.697703
I	789.90	5,723.14	8,126.533982	6,604.62	5,082.700610
J	826.10	10,992.00	10,459.318778	7,309.38	4,159.436356
K	864.30	11,590.00	15,439.587849	8,053.07	666.561891
L	869.30	15,973.00	16,099.371097	8,150.42	201.463800
M	976.50	7,970.67	30,313.734118	10,237.44	-9,838.849438
N	1,355.80	9,524.10	80,730.945765	17,621.85	-45,487.245014
O	1,360.90	35,927.22	81,409.009710	17,721.14	-45,966.730098
P	1,463.21	11,238.73	95,011.748000	19,712.96	-55,585.820690
Q	2,332.10	92,059.97	210,542.762967	36,628.96	-137,284.838305
R	3,017.73	74,649.00	301,708.981386	49,977.16	-201,754.659776
S	3,253.00	42,915.23	332,992.265384	54,557.52	-223,877.228359
Sums	19,633.77	542,585.74		215,542.66	

Table 6. 80% Upper and Lower Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 800

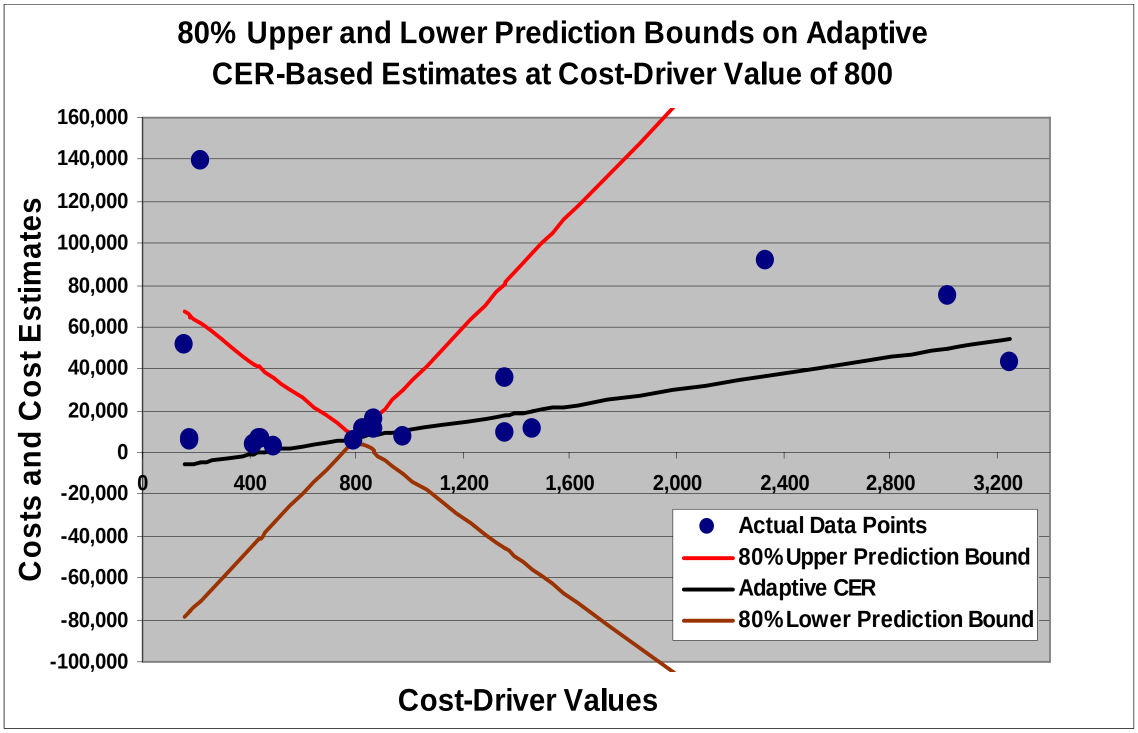

Figure 7. 80% Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 800 with Actual Data Points and Adaptive CER

What is characteristic about the prediction bounds whose graphs appear in Figures 7, 9, and 11 is their excessive widening as the cost-driver value moves away from its base value (800 in Figure 7, 300 in Figure 9, and 3,000 in Figure 11. The point to remember about adaptive CERs is that it is our intention to apply them only in the vicinity of the base cost-driver value, where the prediction bounds are at their narrowest. Therefore, their width in other estimating regions is essentially irrelevant. By the way, the upper and lower prediction bounds do not touch, as Figures 8, 10, and 12 below show. In addition, because these are prediction bounds on cost estimates, which as a practical matter cannot be negative, the region of applicability is further constrained beyond cost-driver values at which the lower prediction bounds go negative.

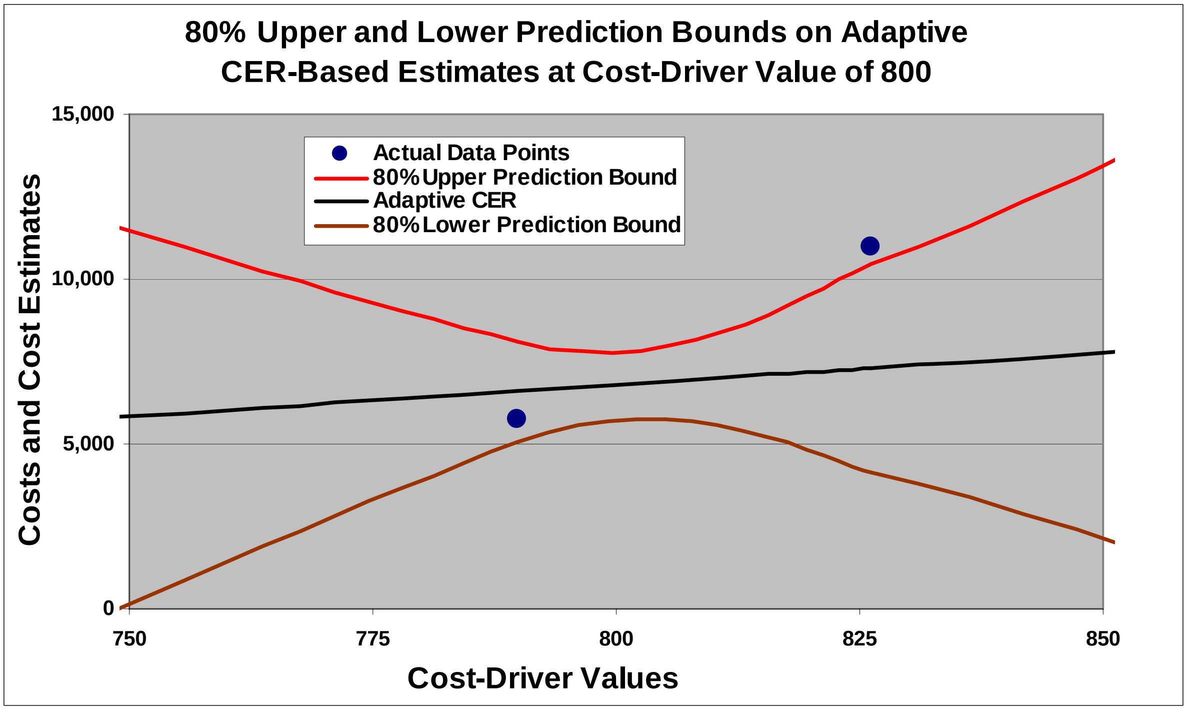

Figure 8. Gap between Upper and Lower Prediction Bounds in the Vicinity of the Cost-Driver Value 800

Program	Cost-Driver Value x	Unit Cost y	80% Upper Bound	WLS EST y	80% Lower Bound
A	156.12	51,367.22	65,389.279544	61,698.97	58,008.663971
B	179.40	5,885.00	62,372.227016	59,227.74	56,083.244080
C	180.30	7,060.00	62,255.776784	59,132.20	56,008.619347
D	217.50	139,483.12	57,462.441876	55,183.32	52,904.189048
E	419.14	3,386.00	36,867.788626	33,778.67	30,689.557986
F	437.09	6,738.00	35,381.736102	31,873.23	28,364.726492
G	440.93	6,812.00	35,064.501531	31,465.60	27,866.707881
H	494.45	3,291.34	30,658.711130	25,784.31	20,909.907048
I	789.90	5,723.14	6,491.040727	-5,578.52	-17,648.087346
J	826.10	10,992.00	3,534.857637	-9,421.25	-22,377.363947
K	864.30	11,590.00	415.759782	-13,476.29	-27,368.336816
L	869.30	15,973.00	7.527753	-14,007.05	-28,021.632368
M	976.50	7,970.67	-8,743.802865	-25,386.63	-42,029.453100
N	1,355.80	9,524.10	-39,698.603983	-65,650.37	-91,602.134324
O	1,360.90	35,927.22	-40,114.762116	-66,191.75	-92,268.734323
P	1,463.21	11,238.73	-48,463.042557	-77,052.24	-105,641.431258
Q	2,332.10	92,059.97	-119,355.526647	-169,287.31	-219,219.087245
R	3,017.73	74,649.00	-175,292.373781	-242,068.82	-308,845.271266
S	3,253.00	42,915.23	-194,486.501042	-267,043.38	-339,600.262830
Sums	19,633.77	542,585.74		-597,019.57	

Table 7. 80% Upper and Lower Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 300

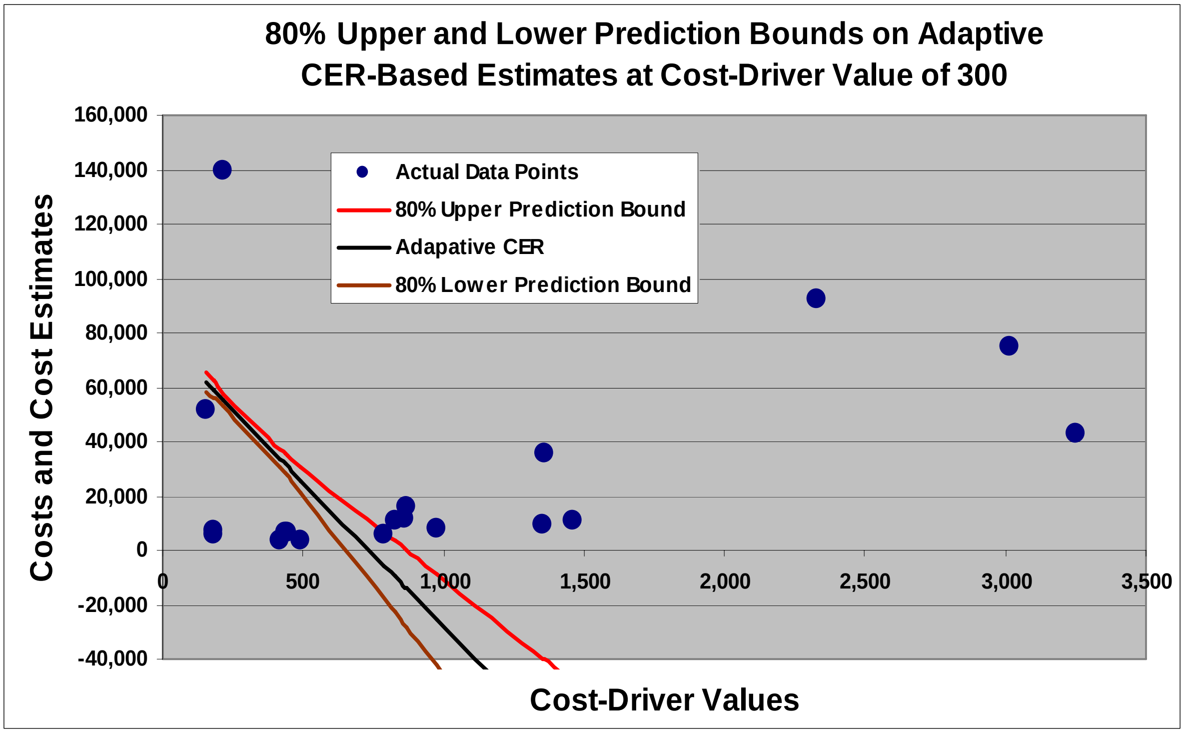

Figure 9. 80% Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 300 with Actual Data Points and Adaptive CER

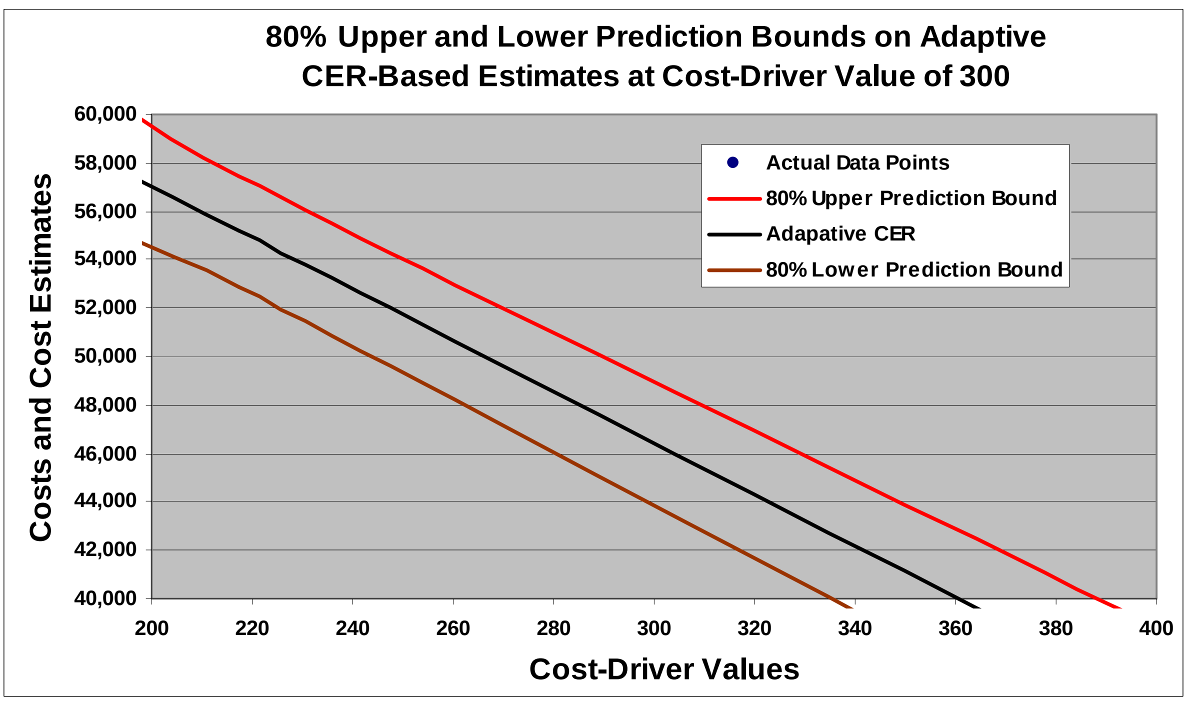

Figure 10. Gap between Upper and Lower Prediction Bounds in the Vicinity of the Cost-Driver Value 300

Program	Cost-Driver Value x	Unit Cost y	80% Upper Bound	WLS EST y	80% Lower Bound
A	156.12	51,367.22	202,434.005312	34,104.71	-134,224.591913
B	179.40	5,885.00	201,384.901034	34,433.09	-132,518.729992
C	180.30	7,060.00	201,344.342887	34,445.78	-132,452.781730
D	217.50	139,483.12	199,667.940092	34,970.51	-129,726.920845
E	419.14	3,386.00	190,581.137616	37,814.77	-114,951.604146
F	437.09	6,738.00	189,772.232090	38,067.96	-113,636.306880
G	440.93	6,812.00	189,599.184936	38,122.13	-113,354.928569
H	494.45	3,291.34	187,187.341936	38,877.06	-109,433.220060
I	789.90	5,723.14	173,873.151720	43,044.57	-87,784.019292
J	826.10	10,992.00	172,241.840172	43,555.19	-85,131.460894
K	864.30	11,590.00	170,520.403443	44,094.02	-82,332.354836
L	869.30	15,973.00	170,295.084698	44,164.55	-81,965.979897
M	976.50	7,970.67	165,464.262738	45,676.67	-74,110.913120
N	1,355.80	9,524.10	148,371.862469	51,026.94	-46,317.989913
O	1,360.90	35,927.22	148,142.044389	51,098.87	-45,944.294515
P	1,463.21	11,238.73	143,531.737673	52,542.02	-38,447.695941
Q	2,332.10	92,059.97	104,382.272484	64,798.25	25,214.232669
R	3,017.73	74,649.00	75,911.693364	74,469.49	73,027.283557
S	3,253.00	42,915.23	92,744.870060	77,788.12	62,831.365052
Sums	19,633.77	542,585.74		883,094.70	

Table 8. 80% Upper and Lower Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 3,000

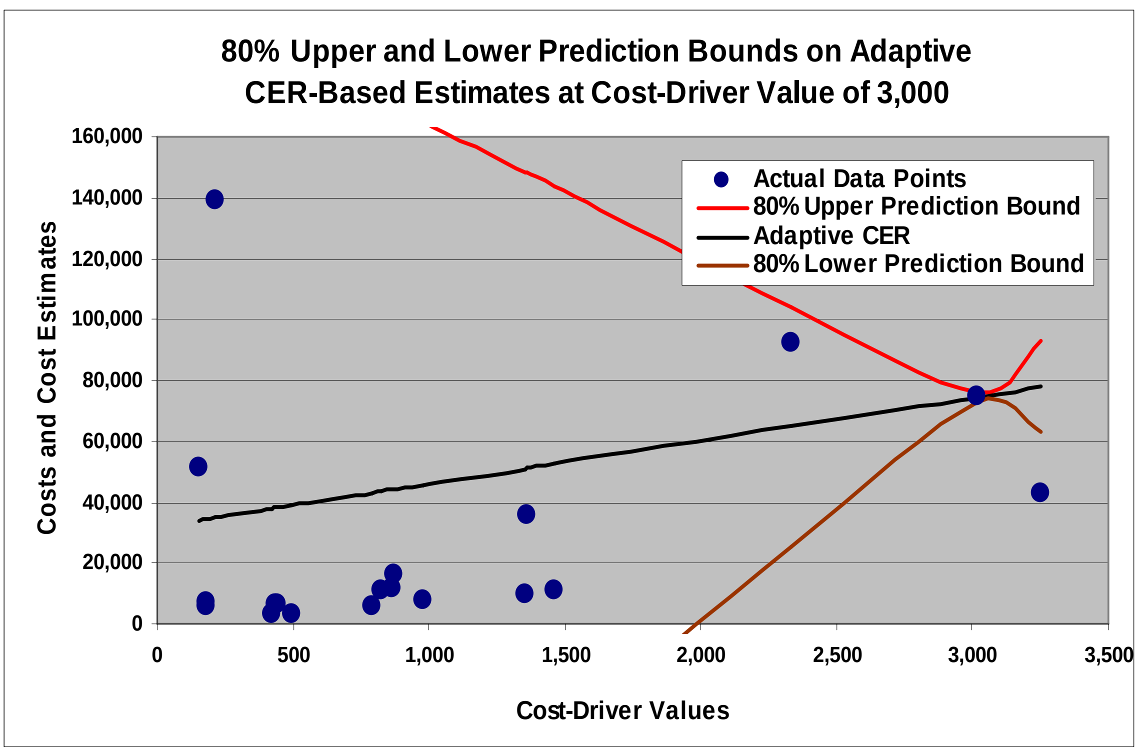

Figure 11. 80% Prediction Bounds for Adaptive-CER-Based Estimates at Cost-Driver Value 3,000 with Actual Data Points and Adaptive CER

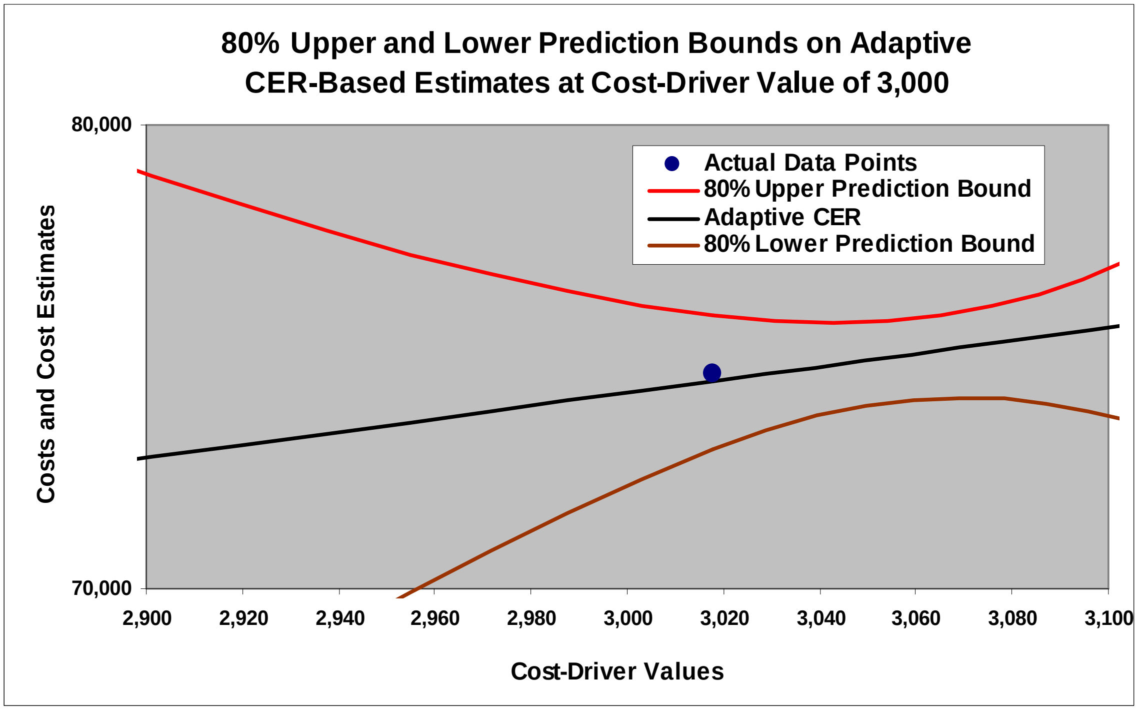

Figure 12. Gap between Upper and Lower Prediction Bounds in the Vicinity of the Cost-Driver Value 3,000

Prediction Bounds for the Universal Adaptive CER

The universal adaptive CER described in Table 4 and Figure 5 is formed by combining the various individual adaptive CERs, over the range of cost drivers into one CER that applies over the entire range. In the example we have been working with, adaptive CERs corresponding to 50-unit cost-driver increments are merged to form one continuous CER across the entire cost-driver range. The resulting universal adaptive CER is illustrated in Figure 5. Insofar as prediction bounds are concerned, we want to make use of the fact that prediction bounds on each individual adaptive CER are very narrow in the vicinity of the cost-driver value on which the adaptive CER is based, but they widen considerably as the cost-driver value moves away from that point. This effect can be seen very clearly in Figures 7, 9, and 11. The universal adaptive CER takes advantage of this situation by providing estimates that have the narrowest possible prediction bounds for all cost-driver values.

Table 11 below contains the numerical data on 80% upper and lower prediction bounds on estimate made using the universal adaptive CER. The prediction bounds themselves, along with the data points and the CER, appear in Figure 10. Note that the prediction bounds are much narrower in the adaptive context than in the standard least-squares-fit context.

Driver	Cost	80% Upper Bound	EST Cost	80% Lower Bound
50.00		62,922.60536	42,739.31	22,556.01954
100.00		58,907.24807	40,817.29	22,727.33210
150.00		56,054.74733	49,546.82	43,038.89123
156.12	51,367.22	59,905.78998	50,880.53	41,855.27867
179.40	5,885.00	74,603.67051	55,953.88	37,304.09301
180.30	7,060.00	75,171.88754	56,150.02	37,128.14511
200.00		87,612.53844	60,443.18	33,273.82964
217.50	139,483.12	97,311.65891	69,749.17	42,186.69003
250.00		115,347.90219	87,031.73	58,715.55405
300.00		71,377.71021	46,425.71	21,473.71561
350.00		38,704.87919	22,733.56	6,762.24433
400.00		12,204.28249	7,006.95	1,809.62688
419.14	3,386.00	10,701.37240	6,760.42	2,819.47622
437.09	6,738.00	9,303.25537	6,529.22	3,755.18780
440.93	6,812.00	9,004.15958	6,479.76	3,955.36231
450.00		8,300.59270	6,362.94	4,425.27919
494.45	3,291.34	4,923.86478	3,589.46	2,255.05196
500.00		4,503.97498	3,243.16	1,982.35231
550.00		14,529.66385	6,829.12	-871.42873
600.00		19,484.26578	9,959.40	434.52824
650.00		20,409.64947	11,310.17	2,210.70010
700.00		18,067.87906	10,929.01	3,790.13759
750.00		12,749.77204	8,652.67	4,555.56975
789.90	5,723.14	9,150.40455	7,175.24	5,200.06839
800.00		8,241.00254	6,801.25	5,361.49607
826.10	10,992.00	11,221.66628	9,756.59	8,291.51518
850.00		13,951.60979	12,462.82	10,974.03604
864.30	11,590.00	14,422.75320	12,666.50	10,910.25030
869.30	15,973.00	14,587.63569	12,737.72	10,887.80057
900.00		15,604.93947	13,174.99	10,745.03389
950.00		11,653.20568	9,208.15	6,763.08930
976.50	7,970.67	11,143.81693	8,832.68	6,521.53760
1,000.00		10,696.45067	8,499.71	6,302.97553
1,050.00		16,599.85083	11,462.16	6,324.47866
1,100.00		20,939.42063	14,296.49	7,653.56932
1,150.00		23,860.71099	16,537.15	9,213.58728
1,200.00		25,595.54200	18,230.99	10,866.44026
1,250.00		26,430.13239	19,495.31	12,560.49388
1,300.00		26,643.54266	20,310.23	13,976.92318
1,350.00		19,965.36192	14,974.31	9,983.26614
1,355.80	9,524.10	20,888.41315	15,774.27	10,660.11771
1,360.90	35,927.22	21,705.81036	16,477.67	11,249.53159
1,400.00		27,979.59574	21,870.45	15,761.29785
1,450.00		13,664.60151	11,840.86	10,017.11120
1,463.21	11,238.73	14,382.93075	12,101.01	9,819.08646

Driver	Cost	80% Upper Bound	EST Cost	80% Lower Bound
1,500.00		16,394.47396	12,825.54	9,256.59722
1,550.00		22,698.80390	16,621.72	10,544.64424
1,600.00		28,144.34523	20,492.26	12,840.17489
1,650.00		33,397.70393	24,526.56	15,655.41028
1,700.00		38,715.42463	28,831.03	18,946.63797
1,750.00		44,154.10037	33,415.50	22,676.89870
1,800.00		49,694.94147	38,247.16	26,799.37082
1,850.00		55,295.14895	43,285.50	31,275.84370
1,900.00		60,905.74386	48,497.85	36,089.94751
1,950.00		66,472.33508	53,862.57	41,252.81236
2,000.00		71,925.95418	59,364.10	46,802.23932
2,050.00		77,167.60488	64,981.01	52,794.40890
2,100.00		82,049.60587	70,666.52	59,283.42427
2,150.00		86,358.35570	76,319.27	66,280.18315
2,200.00		89,805.61668	81,744.09	73,682.56956
2,250.00		92,036.69169	86,609.89	81,183.08854
2,300.00		92,664.98297	90,430.47	88,195.96100
2,332.10	92,059.97	93,449.79687	91,836.14	90,222.47807
2,350.00		93,889.12293	92,619.98	91,350.84125
2,400.00		97,402.84769	92,676.25	87,949.65291
2,450.00		98,222.52317	90,463.37	82,704.22441
2,500.00		96,484.73846	86,410.39	76,336.03984
2,550.00		92,951.10708	81,412.53	69,873.94518
2,600.00		88,646.33546	76,466.46	64,286.59020
2,650.00		84,454.68294	72,322.92	60,191.16611
2,700.00		80,929.09901	69,366.76	57,804.41474
2,750.00		77,054.82704	66,431.86	55,808.89003
2,800.00		76,663.27434	67,242.40	57,821.52197
2,850.00		75,905.67799	67,904.22	59,902.76737
2,900.00		75,867.66447	69,545.45	63,223.22554
2,950.00		76,119.31586	71,913.26	67,707.21079
3,000.00		75,518.29497	74,219.40	72,920.49668
3,017.73	74,649.00	75,966.58830	74,164.83	72,363.08019
3,050.00		76,786.35756	74,065.53	71,344.69813
3,100.00		74,002.10190	67,141.02	60,279.92945
3,150.00		62,069.86209	54,415.99	46,762.11593
3,200.00		49,725.94543	45,424.15	41,122.36282
3,250.00		43,144.36743	42,927.10	42,709.82647
3,253.00	42,915.23	43,354.47526	42,978.65	42,602.82964
3,300.00		46,653.41208	43,786.36	40,919.29842
3,350.00		50,744.79550	45,762.39	40,779.98882
3,400.00		54,430.68793	47,971.96	41,513.24151
3,450.00		57,664.17277	50,126.95	42,589.72220
3,500.00		60,512.13570	52,149.51	43,786.89143

Table 9. Universal Adaptive-CER-Based Estimates and 80% Prediction Bounds at 50-Unit Increments Along the Cost-Driver Axis

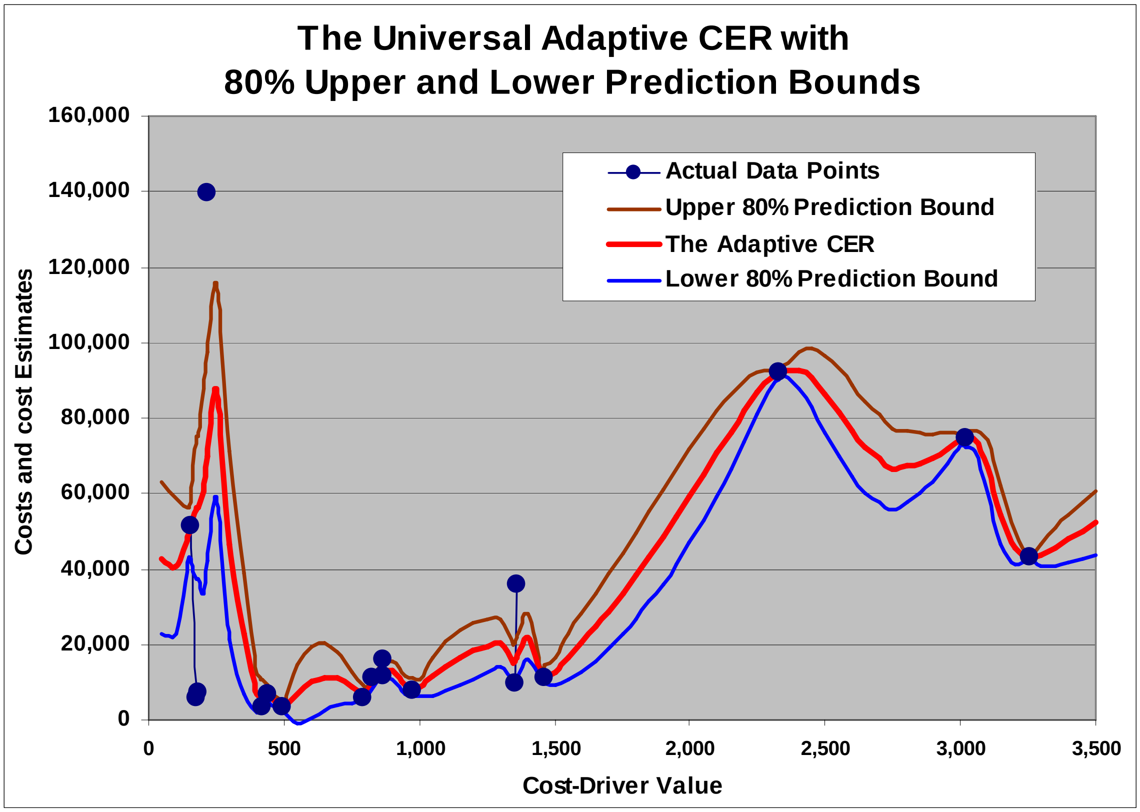

Figure 13. Universal Adaptive-CER-Based Estimates and 80% Prediction Bounds Graphed at 50-Unit Increments along the Cost-Driver Axis

As is characteristic of adaptive CERs, we see that the prediction bounds are much narrower in Figure 10 than they are in the OLS regression situation illustrated in Figure 6. Again, this narrowing is due to the fact that estimating using an adaptive CER near a cost-driver value is carried out using only data points near that cost-driver value. However, when there is significant variation in data points near a cost-driver value, the prediction bounds widen in that region. For an example, see what happens in the cost-driver region of 200-300 in Figure 13 above. The prediction bounds for OLS CERs, on the other hand, must be wide enough to provide the desired amount of confidence, e.g., 80%, throughout the entire cost-driver range.

Acknowledgements

The authors are grateful to the members of MCR's quality-review team: Dr. Neal D. Hulkower, Dr. Jerry D. Hofmann, Raymond P. Covert, Nathan J. Menton, Timothy P. Anderson, and Jan B. Sterbutzel. The combined efforts of the team resulted in a significant improvement in the clarity of explanation of specific details of the conclusions drawn, as well as the elimination of various typos and other errors.

Appendix

Algebraic Analysis of the Total Variation

$$TV = \sum_{k=1}^{n} w_k (y_k - \bar{y}_w)^2 = \sum_{k=1}^{n} w_k \left[(y_k - a - bx_k) + (a + bx_k - \bar{y}_w) \right]^2$$

$$= \sum_{k=1}^{n} w_k \left[(y_k - a - bx_k)^2 + 2(y_k - a - bx_k)(a + bx_k - \bar{y}_w) + (a + bx_k - \bar{y}_w)^2 \right]$$

$$= \sum_{k=1}^{n} w_k (y_k - a - bx_k)^2 + \sum_{k=1}^{n} w_k (a + bx_k - \bar{y}_w)^2 + 2\sum_{k=1}^{n} w_k (y_k - a - bx_k)(a + bx_k - \bar{y}_w)$$

$$= SS + VB + 2\sum_{k=1}^{n} w_k (y_k - a - bx_k)(a + bx_k - \bar{y}_w)$$

We now show that the third summand in the above equation is always zero, no matter what the data, so that $TV = SS + VB$ for every set of data points. The expression for a that results from solving for the WLS regression equation implies that

$$a = \frac{\left(\sum_{k=1}^{n} w_k y_k \right)}{\left(\sum_{k=1}^{n} w_k \right)} - b \frac{\left(\sum_{k=1}^{n} w_k x_k \right)}{\left(\sum_{k=1}^{n} w_k \right)} = \bar{y}_w - b\bar{x}_w ,$$

where $\bar{y}_w$ and $\bar{x}_w$ are the weighted means of the y and x values in the data set, respectively. Therefore $a + bx_k - \bar{y}_w = a + bx_k - (a + b\bar{x}_w) = b(x_k - \bar{x}_w)$, from which it follows that

$$2\sum_{k=1}^{n} w_k (y_k - a - bx_k)(a + bx_k - \bar{y}_w) = 2\sum_{k=1}^{n} w_k (y_k - a - bx_k) b(x_k - \bar{x}_w)$$

$$= 2b\sum_{k=1}^{n} w_k (x_k y_k - ax_k - bx_k^2 - \bar{x}_w y_k + a\bar{x}_w + b\bar{x}_w x_k)$$

$$= 2b\left[\sum_{k=1}^{n} w_k x_k y_k - a\sum_{k=1}^{n} w_k x_k - b\sum_{k=1}^{n} w_k x_k^2 - \bar{x}_w \sum_{k=1}^{n} w_k y_k + a\bar{x}_w \sum_{k=1}^{n} w_k + b\bar{x}_w \sum_{k=1}^{n} w_k x_k \right]$$

In view of the fact that $\sum_{k=1}^{n} w_k x_k = \bar{x}_w \sum_{k=1}^{n} w_k$, the two terms above that contain "a" can be canceled out. What remains is, except for the "$2b$" factor:

$$\sum_{k=1}^{n} w_k x_k y_k - b\sum_{k=1}^{n} w_k x_k^2 - \bar{x}_w \sum_{k=1}^{n} w_k y_k - b\bar{x}_w \sum_{k=1}^{n} w_k x_k$$

$$= \sum_{k=1}^{n} w_k x_k y_k - \frac{\left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)}{\sum_{k=1}^{n} w_k} - b\sum_{k=1}^{n} w_k x_k^2 - b\frac{\left(\sum_{k=1}^{n} w_k x_k\right)^2}{\sum_{k=1}^{n} w_k}$$

$$= \frac{\sum_{k=1}^{n} w_k \sum_{k=1}^{n} w_k x_k y_k - \left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)}{\sum_{k=1}^{n} w_k} - b\left[\frac{\sum_{k=1}^{n} w_k \sum_{k=1}^{n} w_k x_k^2 - \left(\sum_{k=1}^{n} w_k x_k\right)^2}{\sum_{k=1}^{n} w_k}\right]$$

$$= \frac{b\left[\left(\sum_{k=1}^{n} w_k\right)\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{n} w_k x_k\right)^2\right]}{\sum_{k=1}^{n} w_k} - b\left[\frac{\sum_{k=1}^{n} w_k \sum_{k=1}^{n} w_k x_k^2 - \left(\sum_{k=1}^{n} w_k x_k\right)^2}{\sum_{k=1}^{n} w_k}\right] = 0$$

Because $b = \dfrac{\left(\sum_{k=1}^{n} w_k\right)\left(\sum_{k=1}^{n} w_k x_k y_k\right) - \left(\sum_{k=1}^{n} w_k x_k\right)\left(\sum_{k=1}^{n} w_k y_k\right)}{\left(\sum_{k=1}^{n} w_k\right)\left(\sum_{k=1}^{n} w_k x_k^2\right) - \left(\sum_{k=1}^{n} w_k x_k\right)^2}$

References

1. Book, S.A., “Deriving Cost-Estimating Relationships Using Weighted Least-Squares Regression,” IAA/ISPA/AIAA Space System Cost Methodologies and Applications Symposium, San Diego CA, 10-11 May 1990, 19 pages.

2. Book, S.A.and Broder, M.A., “Adaptive Cost-Estimating Relationships,” Space Systems Cost Analysis Group (SSCAG) Meeting in Noordwijk, Holland, 15-16 May 2008, and ISPA/SCEA Joint Annual Conference & Training Workshop, Industry Hills CA, 24-27 June 2008, 80 Charts.

3. Foussier, P., “Space System Study,” Space Systems Cost Analysis Group (SSCAG), Montreal QB, 16 May 2007, 41 Charts.

Author Biographies

Dr. Stephen A. Book is Chief Technical Officer of MCR, LLC. In that capacity, he is responsible for ensuring technical excellence of MCR's products, services, and processes by encouraging process improvement, maintaining quality control, and training employees and customers in cost and schedule analysis and associated program-control disciplines. During his career, he has provided technical support in the cost, schedule, and earned-value areas to several Air Force, NASA, and Intelligence organizations and continues to do so. Dr. Book joined MCR in January 2001 after 21 years with The Aerospace Corporation, where he held the title "Distinguished Engineer" during 1996-2000 and served as Director, Resource and Requirements Analysis Department, during 1989-1995. Dr. Book was the last editor of ISPA's *Journal of Parametrics* prior to its merger with SCEA's *Journal of Cost Analysis and Management*, and is now co-editor of the combined journal. He was the 2005 recipient of ISPA's Freiman Award for Lifetime Achievement. He earned his Ph.D. in mathematics, with concentration in probability and statistics, at the University of Oregon.

Melvin A. Broder is a Senior Project Leader at The Aerospace Corporation. In that capacity he has developed cost models for the Concept Design Center, building and expanding tools for the cost seat and devising new processes. Prior to joining Aerospace he worked in cost estimating at Boeing's Satellite Systems where he was responsible for front end of the business cost tools and models for Boeing's commercial product line including support of the BSS Design Center's Integrated Engineering Laboratory. Mr. Broder has also been a Project Manager for cost tools and processes in the System Engineering Laboratory at Raytheon Systems Company. His responsibilities include the creation and maintenance of tools to support the Sensors and Electronic Systems in design to cost (DTC) and cost-as-an-independent-variable (CAIV) activities. Prior to working the aerospace industry he was an Instructor of Economics at La Verne College, teaching both upper and lower division course work in Micro and Macro Economics, Comparative Economics, Money and Banking, and Econometrics. He earned an M.S. in Economics from the University of Southern California.

Daniel I. Feldman is a Junior Cost Analyst at MCR, LLC. Since joining MCR in early September 2005, he has worked on developing new techniques in utilizing CER-based estimates, along with doing rocket modeling and trade-study analysis. Mr. Feldman earned his B.S. in mathematics in June 2005, with concentration in statistics, at the University of California, Irvine. He is currently working on his M.S. in applied statistics at California State University, Long Beach.