Journal of Cost Analysis and Parametrics, 6:23–42, 2013
Copyright © 2013 Tecolote Research, Inc.
ISSN: 1941-658X print / 2160-4746 online
DOI: 10.1080/1941658X.2013.766550

www.iceaaonline.com

Accuracy Matters: Selecting a Lot-Based Cost Improvement Curve

SHU-PING HU and ALFRED SMITH

Tecolote Research, Inc., Santa Barbara, California

There are two commonly used cost improvement curve theories: unit cost theory and cumulative average cost theory. Ideally, analysts develop the cost improvement curve by analyzing unit cost data. However, it is common that instead of unit costs, analysts must develop the cost improvement curve from lot cost data. An essential step in this process is to estimate the theoretical lot midpoints for each lot, to proceed with the curve-fitting process. Lot midpoints are generally associated with unit cost theory, where the midpoint is always within the lot. The more general lot plot point term is used in the context of both the unit cost and cumulative average cost theories. Many research papers have been published on cost improvement curves, including several that discuss estimating the lot midpoint. A two-term formula has traditionally been used as a useful approximation to derive the lot total cost, as well as the lot midpoint under unit cost theory (see SCEA, 2002–2011; CEBoK, Module 7). There is, however, a more accurate six-term formula to better approximate the lot total cost and lot midpoint. This increase in accuracy may be substantial for high-cost items or an aggregated estimate, consisting of many cost improvement curve-related items. The more accurate formula can also impact cost uncertainty analysis results, especially when thousands of iterations are performed. This article describes how to derive and use lot plot points for both the unit cost and cumulative average cost theories. We describe how the analyst can use lot plot points to construct prediction intervals for cost uncertainty analysis. This approach is more efficient and appropriate than using the unit cost curve directly. In addition, this article will (1) detail an iterative, two-step regression method to implement the six-term formula, (2) describe the advantages of generating the lot plot points for cost improvement curves, (3) recommend an iterative (not direct) approach to fit a cost improvement curve under cumulative average theory, and (4) compare cost improvement curves derived using the two-step regression method with cost improvement curves generated by the simultaneous minimization process. Different error term assumptions and realistic examples are also discussed. In the example section, we show why the goodness-of-fit measures alone should not be used for selecting a best model, especially when either the fit spaces or the dependent variables are different.

Introduction

Background

This article was originally presented at the 2012 SCEA/ISPA Conference. The main objectives of this article are three-fold. First, we introduce a six-term formula as an alternative to the traditional two-term formula to compute the lot total cost (LTC) and lot midpoint (LMP) for unit theory. We then recommend using (1) an iterative approach to fit cumulative average cost improvement curves (CICs) and (2) the lot plot point (LPP) to construct prediction intervals (PIs) for cost uncertainty analysis. We also explain why the goodness-of-fit measures should *not* be the basis for selecting a best model.

Address correspondence to Alfred Smith, Tecolote Research, Inc., 5266 Hollister Ave., Ste. 301, Santa Barbara, CA 93111. E-mail: asmith@tecolote.com

Our goal is to provide analysts a better understanding of the methods used to derive both unit and cumulative average CICs from lot data and how to use the LPP effectively for cost estimating and cost uncertainty analysis.

Theories

Two theories are commonly used to fit CICs to lot data: unit cost (UC) theory and cumulative average cost (CAC) theory. In general, the CIC theory states that as the total quantity of units produced doubles, the "unit cost" goes down by a constant percentage. This unit cost may be either the average cost of a given number of units (CAC curve) or the cost of a specific unit (UC curve).

T. P. Wright first described the theory of the learning curve in 1936 (Wright, 1936). Wright's research led to the cumulative average formulation of learning curve theory, also known as the Northrop formulation. Unit learning curve theory is attributed to J. R. Crawford, who first formulated his theory in a booklet prepared for Lockheed Aircraft personnel in 1947 (Crawford, 1947). Unit theory is also known as the Boeing formulation or the Stanford formulation. Either formulation results in a hyperbolic function, which appears linear when plotted in logarithmic space (i.e., log-log grids).

To determine which theory to use for a data set, plot the data on log-log graph paper to see which plot most closely resembles a straight line. That is, for unit theory, plot the Lot Average Cost against the True Lot Midpoints (i.e., the theoretical lot midpoints) on a Cum Unit scale. For CAC theory, plot the Cum Average Cost against the cumulative units.

Note that the regression results from each of the cost theories cannot be used interchangeably. If you apply UC theory to develop a UC curve, you cannot use its T1 and slope to generate a CAC theory estimate and vice versa (Anderson, 2003). We also recommend that you do not mix learning theory CICs in the same cost model.

First, we explain in detail the CIC algorithms using the two-step regression method for both UC and CAC theories.

Regression Method to Generate UC and CAC Curves

In this section, we describe a conventional two-step method to develop CICs for both UC and CAC theories; we start with unit theory.

Unit Theory

Unit theory can be summarized as follows: when the total number of units produced is doubled, the unit cost decreases by some constant percentage. The constant percentage by which the cost decreases when the quantity is doubled is called the rate of improvement. In the improvement curve analysis, the "slope" is the difference between 100% (no improvement) and the percent of cost reduction. For example, if the unit cost reduces by 10% each time the quantity doubles, the improvement curve slope is 90% (i.e., 100% - 10%).

Mathematically, the individual cost of the *n*th unit under unit theory is given by:

$$UC_n = T_1\, n^b f(\underline{\mathbf{x}}_n)\varepsilon_n, \tag{1}$$

where T_1 is the first unit cost, $b = \frac{\ln(slope/100)}{\ln(2)}$, $f(\underline{\mathbf{x}}_n)$ is a multiplicative function of the independent variables $x_1, x_2, \ldots, x_m$, namely, $f(\underline{\mathbf{x}}_n) = x_{n1}^{c1}\, x_{n2}^{c2} \ldots x_{nm}^{cm}$ for the *n*th unit (e.g., Rate^c, weight^d, Area^e, etc., or a combination of them; if the additional predictors are not present, Equation (1) is a basic UC curve) and ε_n is the multiplicative error term assumed to follow a log-normal distribution with a mean of zero and variance σ^2 in log space, i.e., ε_n is distributed as LN(0, σ^2).

Note that the CIC slope is usually expressed in percent; if there is "cost improvement," then the exponent b should be less than zero. Also, Equation (1) is directly related to the equation listed on page 16 of Large, Hoffmayer, and Kontrovich (1974).

We use the following definitions to explain the iterative procedure for estimating the exponents (b, c_1, c_2, . . ., c_m) of the independent variables and the first unit cost, T_1:

Prior Total Quantity of lot $i = PQ_i$
Lot Total Quantity of lot $i = LQ_i$
Cum. Total Quantity of lot $i = Q_i$
Lot Total Cost of lot $i = LTC_i$
Lot Average Cost of lot $i = LAC_i = LTC_i/LQ_i$

When dealing with lot cost data, the approach is to find lot midpoints such that the unit cost at the lot midpoint equals the lot average cost. Therefore, the above unit theory equation can be rewritten for lots as:

$$LAC_i = T_1\, LMP_i^b f(\underline{\mathbf{x}}_i)\, \varepsilon_i \quad (= UC_{LMP_i}) \quad for\, i = 1, \ldots, k, \tag{2}$$

where LMP_i is the (true) Lot Midpoint (LMP) of lot i, k is the total number of lots to be included in the curve fit, and $f(\underline{\mathbf{x}}_i)$ is defined above.

This implies the following:

$$LMP_i = \left(\frac{1}{LQ} \sum_{Q=PQ+1}^{PQ+LQ} (Q^b) \right)^{1/b}. \tag{3}$$

(Note that the multiplicative function $f(\underline{\mathbf{x}})$ does not appear in Equation (3) due to the canceling effect.)

Equation (3) is an exact solution for calculating the LMP when the slope is given. If the slope is not known, use an iterative approach to solve for both slope and LMP (see the descriptions given below). We first introduce a "two-term" formula to estimate the LMP.

Two-Term vs. Six-Term Formula to Estimate LMP. The direct computation of the lot total cost (Equation (1)) becomes very cumbersome if there are many units in the lot. Many analysts have thus used the traditional two-term formula to approximate the summation in Equation (3) to obtain the LMP. We will present both two-term and six-term formulas below.

The two-term formula can be derived from simple calculus over a range that begins a half a unit before the first unit and ends half a unit after the last unit. This simplification of "half a unit" is not precise, but does yield a reasonably useful result (note the subscript i is eliminated from the notations, PQ_i and LQ_i, to simplify the illustration):

$$\sum_{Q=PQ+1}^{PQ+LQ} Q^b \cong \int_{PQ+0.5}^{PQ+LQ+0.5} x^b dx = \frac{(PQ + LQ + 0.5)^{b+1} - (PQ + 0.5)^{b+1}}{b+1}, \tag{4}$$

$$LMP_i = \left(\frac{1}{LQ} \sum_{Q=PQ+1}^{PQ+LQ} Q^b \right)^{1/b} \cong \left(\frac{(PQ+LQ+0.5)^{b+1} - (PQ+0.5)^{b+1}}{LQ(b+1)} \right)^{1/b}. \tag{5}$$

This approximation is not very accurate, especially for small quantities; many analysts questioned the accuracy of the two-term formula and some have recommended other solutions (for details, see Coleman et al., 2010; Goldberg & Touw, 2000, 2005; Lee, 2005). Therefore, we employ a six-term formula to better approximate the lot total cost, as well as the LMP, for unit theory CICs:

$$\text{If } PQ > 0 \Rightarrow LTQ_i = \sum_{Q=PQ+1}^{PQ+LQ} Q^b \cong \frac{(PQ+LQ)^{b+1} - (PQ)^{b+1}}{b+1} + \frac{(PQ+LQ)^b - (PQ)^b}{2} + \frac{b}{12}\left((PQ+LQ)^{b-1} - (PQ)^{b-1}\right), \tag{6}$$

$$\text{If } PQ = 0 \Rightarrow LTQ_i = \sum_{Q=1}^{LQ} Q^b \cong \frac{(LQ)^{b+1}}{b+1} + \frac{(LQ)^b}{2} + \frac{b}{12}(LQ)^{b-1} - \frac{1}{b+1} + \frac{1}{2} - \frac{b}{12}. \tag{7}$$

The true lot midpoint is then given by:

$$LMP_i = \left(\frac{LTQ_i}{LQ} \right)^{1/b}. \tag{8}$$

(Here, "LTQ_i" stands for the lot total cost of lot i, assuming T_1 and $f(\underline{\mathbf{x}}_i)$ equal to one.)

The six-term formula (Equations (6) and (7)) represents a simple and accurate approximation to the summation in Equation (3). It is proven to be more accurate than the traditional two-term approximation formula (for a detailed proof, see Cho & Schmidt, 1984). Table 1 compares the exact lot total cost (using Equation (1)) to the lot total cost estimated using the two-term (Equation (4)) and six-term formulas (Equation (7)) when the prior quantity is zero ($PQ = 0$).

TABLE 1 Comparisons of 2-term and 6-term approximation with exact formula ($T_1 = 100$, slope $= 70\%$, $b = -.5146$)

No. of Units	Total from exact formula	Total from 2-term approximate formula	Total from 6-term approximate formula	% Error 2-term	% Error 6-term
1	100.00	100.00	100.00	0.00	0.00
2	170.00	174.25	170.19	2.50	0.11
3	226.82	231.28	227.02	1.97	0.09
4	275.82	280.38	276.03	1.65	0.08
10	493.18	497.90	493.39	0.96	0.04
30	930.50	935.27	930.71	0.51	0.02
100	1,779.07	1,783.85	1,779.28	0.27	0.01

As shown by Table 1, when we use a learning slope of 70%, the approximation using the six-term formula is about 20 times more accurate (when comparing % error) than the traditional two-term approximation formula. This improvement may have a substantial impact on cost uncertainty analysis results (e.g., the 80th percentile) when dealing with high-cost items or an aggregated estimate, consisting of many elements estimated using CICs derived from lot data (for the six-term formula, see Cho & Schmidt, 1984). Note that the percentage error for the two-term formula is smaller if the learning slope is shallower or when there is a prior quantity.

For a small lot quantity, we simply sum individual unit costs using Equation (1) to compute lot total cost (LTC). The six-term formula is only used to compute LTC and LMP when there are many units in the lot.

Two-Step Regression Method (a Conventional Approach). We now explain the regression method and we use the term "two-step" to describe the nature of this method because the solution is derived from a two-step process, not a single pass. As shown by Equation (3), the LMP_i is a function of b, so we cannot calculate it directly unless the exponent b is given. As a result, we apply an iterative approach. The calculation of LMPs is performed in two steps. First, we use the initial estimates of the LMPs to fit the CIC (Equation (2)). Then, we use new exponent b (resulting from the curve fit) to re-estimate the LMPs (see Equation (3)), and then we use these new LMPs to refit the curve, etc., until the solution for b converges. A detailed explanation is given below.

Both Equations (1) and (2) are log-linear models, so the actual curve-fitting is done in log space and Equation (2) can be equivalently stated as:

$$\begin{aligned}\ln(LAC_i) &= \ln(T_1) + b^* \ln(LMP_i) + \ln(f(\underline{\mathbf{x}}_i)) + \ln(\varepsilon_i)\\ &= A + b^* \ln(LMP_i) + c_1 \ln(x_{i1}) + c_2 \ln(x_{i2}) + \ldots + c_m \ln(x_{im}) + \ln(\varepsilon_i) \qquad (9)\\ &\quad \mathit{for}\, i = 1, \ldots, k.\end{aligned}$$

Listed below is the iterative solution, which determines the true lot midpoints, as well as A, b, and $c_1, c_2, \ldots, c_m$:

1. Estimate initial values of LMP_i for $i = 1, \ldots, k$ using arithmetic midpoint of each lot or an alternative formula: $LMP_i = (PQ + 1 + PQ + LQ + 2^*sqrt((PQ + 1)(PQ + LQ)))/4$.
2. Use these initial values to regress $\ln(LAC_i)$ against $\ln(LMP_i)$ and $\ln(x_{i1}), \ldots, \ln(x_{im})$ to obtain A, b, and $c_1, c_2, \ldots, c_m$.
3. Use the value of b obtained in step (2) to re-estimate true lot midpoints using the six-term equations (Equations (6) and (7)).
4. Use the re-estimated LMP_i (Equation (8)) to refit the curve (Equation (9)), and derive new values for A, b, and $c_1, c_2, \ldots, c_m$.
5. Repeat steps (3) and (4) above until the changes in successive values of b and $c_1, c_2, \ldots, c_m$ are sufficiently small.

The process usually converges within two to three iterations.

CAC Theory

Cumulative average cost theory can be stated as follows: When the total number of units produced is doubled, the cumulative average cost decreases by some constant percentage.

As with the unit improvement curve, the constant percentage by which the cost of the doubled quantities decreases is called the rate of improvement and the slope of the improvement curve is the difference between 100% and the rate of improvement. However, the rate of improvement and the slope are both measured using *cumulative averages* rather than unit values.

Under CAC theory, a log-linear equation is hypothesized to relate the *cumulative* number of units produced to the average cost of these units:

$$CAC_n = T_1 n^b f(\underline{\mathbf{x}}_n)\varepsilon_n, \tag{10}$$

where T_1 is the first unit cost, n is the unit number, CAC_n is the cumulative average cost through unit n from unit one assuming $f(\underline{\mathbf{x}}_n)$ is identical for all n units, and other definitions are as given before (see Equation (1)). Note that Equation (10) is almost identical to the equation form listed on page 16 of Large et al. (1974) after multiplying both sides of Equation (10) by the total unit n. In fact, this equation was first introduced by Levenson et al. (1971).

There are two approaches to generate the coefficients for Equation (10): the direct (traditional) and iterative approaches. The traditional, direct approach is simply fitting a log-linear model using the ordinary least squares (OLS) method in the log space when all the costs of the consecutive lots are available, i.e., no missing lot. The iterative method can be applied when there are voids in the data set. See the discussions of both methods below.

Direct (Traditional) Approach. Given the LTC for several consecutive lots, the CACs are obtained as follows:

$$\begin{aligned} CAC_{Q_1} &= \tfrac{LTC_1}{Q_1} \\ CAC_{Q_2} &= \tfrac{LTC_1+LTC_2}{Q_2} \\ &\vdots \\ CAC_{Q_k} &= \tfrac{LTC_1+LTC_2+\ldots+LTC_k}{Q_k} \end{aligned}, \tag{11}$$

where Q_i stands for the cumulative quantity through lot i ($i = 1,\ldots, k$) and k is the total number of lots. We use the term "CAC-Direct" to denote the CAC curve generated by the direct approach; we also use this term to indicate the method.

If the cumulative average costs for all consecutive lots are present, then the direct approach can be applied to the lot data with the last unit in the lot as the lot plot point (LPP). T_1, b, and other exponents (c_1, c_2, . . . c_m) can be obtained directly from the ordinary least squares (OLS) method by regressing CACs vs. cumulative quantities (Q_i's), along with other potential cost drivers (if any) in log space.

Downside of Using Direct Approach. This traditional, direct approach under CAC theory has a few drawbacks. It tends to smooth the cost data by summing and averaging the lot total costs from the very first lot. Potential outliers are not easily identified due to the "summing and averaging process." The smoothing also generates better goodness-of-fit statistics than the unit method, even though the ability to predict lot costs is not necessarily better (see an example on page 33). Further, if any of the LTCs are missing (a missing lot problem), it is not possible to calculate CACs as given in Equation (11). Therefore, we use an *iterative* approach using lot average cost instead.

Iterative Approach. Calculate the lot total cost directly by Equation (10), assuming $f(\underline{\mathbf{x}})$ is identical for units in any lot:

$$LTC_i = T_1[(PQ_i + LQ_i)^{b+1} - (PQ_i)^{b+1}]f(\underline{\mathbf{x}}_i)\varepsilon_i \quad \text{for } i = 1, \ldots, k. \tag{12}$$

Solve Equation (12) using non-linear regression. This equation form has been used to analyze the CICs for the Unmanned Space Vehicle Cost Model, 8th Edition (USCM8) database (see Hu, Fong, & Enser, 2006 for details). However, if we develop the effective LPP using CAC theory, we can solve this equation iteratively in the log space using OLS. In mathematical terms, we will relate the lot average cost (LAC) to the CAC through the effective LPP as follows:

$$\begin{aligned} LAC_i &= T_1\,((PQ_i + LQ_i)^{b+1} - (PQ_i)^{b+1})/LQ_i{}^*f(\underline{\mathbf{x}}_i)\varepsilon_i \\ &= T_1(LPP_i)^b f(\underline{\mathbf{x}}_i)\varepsilon_i \quad (= CAC_{LPP_i}) \end{aligned} \quad \text{for } i = 1, \ldots, k, \tag{13}$$

where

$$LPP_i = \left(\frac{(PQ_i + LQ_i)^{b+1} - (PQ_i)^{b+1}}{LQ_i}\right)^{1/b}, \tag{14}$$

and

$PQ_i =$ prior total quantity of lot i,
$LQ_i =$ lot quantity of lot i,
$k =$ total number of lots.

In other words, this approach finds the LPPs (on a log-linear curve) such that the cumulative average cost at the LPP is equal to the LAC (Kluge, 1975). Thus, the iterative solution for cumulative average CIC follows the same steps described in the unit theory section, with step (3) revised by using Equation (14) to derive the LPPs. Note that the actual cumulative lot average costs are *not* used in the iterative approach. Instead, the LAC or LTC is used as the dependent variable to generate CAC curves. Also, the LPP for the direct approach is the last unit of the lot (i.e., Q_i), while the LPP for the iterative approach lies outside the lot except for lot 1. To compare with CAC-Direct curves, we use the term "CAC-Iterative" to denote the CAC curve generated by the iterative approach; we also use this term to indicate the method.

Both unit and CAC curves are biased low when the curve fit is done in log space. Although a least squares optimization in log space produces an unbiased estimator in log space, the estimator is biased low when transformed back to unit space. Therefore, we should apply a correction factor to adjust the cost estimating relationship (CER) result to produce the mean in unit space. The commonly used correction factors are Goldberger's Factor, the Smearing Estimate, the PING Factor, etc. (Hu, 2005; Hu & Sjovold, 1989; Duan, 1983; Goldberger, 1968). Alternatively, we can use the Minimum-Unbiased-Percentage-Error (MUPE) method for modeling multiplicative errors directly *in unit space* to eliminate the bias. The MUPE method is an Iteratively Reweighted Least Squares (IRLS) regression technique (Hu, 2001; Hu & Sjovold, 1994; Seber & Wild, 1989; Weisberg, 1985). Both unit and CAC curves can be generated directly in unit space using the MUPE method to eliminate the bias. (The traditional OLS method is for additive-error models. Since the multiplicative error is assumed for CICs, the OLS method is not appropriate.) Note: the point estimate (PE) can be left undefined (it does not have to be the mean) when using a PI to model the CER uncertainty distribution for cost risk analysis.

Advantages of Using LMP for Unit Curves

In the next two sections, we explain why it is essential to use the LPP to build the PI to analyze the LTC for cost uncertainty analysis instead of summing individual unit cost distributions. Unit theory's LMP is also very useful when predicting the LTC for a future lot, which contains many production units.

The lot midpoints lie within their respective lots. For unit theory, we treat LMP as a representative for a given lot and, intuitively, this point should always remain in its own lot. A proof is given below. Let PQ and LQ denote the prior quantity and lot quantity for a lot, respectively; the LMP should be in the interval $[PQ+1, PQ+LQ]$. This is due to the fact that

$$\left(\frac{1}{LQ}LQ^{*}(PQ+LQ)^{b}\right) \leq (LMP_i)^b = \left(\frac{1}{LQ}\sum_{Q=PQ+1}^{PQ+LQ}(Q^b)\right) \leq \left(\frac{1}{LQ}LQ^{*}(PQ+1)^{b}\right)$$

$$\textit{if } b \leq 0.$$

The above inequality can be simplified as:

$$(PQ+LQ)^b \leq (LMP_i)^b \leq (PQ+1)^b \quad \textit{if } b \leq 0.$$

Since the exponent b is less than zero, we can derive the following inequality:

$$(PQ+1) \leq (LMP_i) \leq (PQ+LQ) \quad \textit{if } b \leq 0.$$

Estimation of LAC

Given a first unit cost and a CIC slope, we can predict the LAC for a lot with a prior quantity, PQ_0, and a lot quantity, LQ_0, using Equation (2) under unit theory CIC:

$$L\hat{A}C_0 = T_1\, LMP_0^b f(\underline{\mathbf{x}}_0), \tag{15}$$

where LAC_0 is the estimated lot average cost, LMP_0 is the theoretical lot midpoint, and $f(\underline{\mathbf{x}}_0)$ is the multiplicative function of the independent variables for this particular lot, respectively. Note that $b = \ln(\text{slope}/100)/\ln(2)$.

It follows from Equation (5) that the lot midpoint, LMP_0, can be approximated by the formula below if both PQ_0 and LQ_0 are fairly large:

$$LMP_0 \cong \left(\frac{(PQ_0+LQ_0+0.5)^{b+1}-(PQ_0+0.5)^{b+1}}{(b+1)LQ_0}\right)^{1/b}. \tag{16}$$

(As noted above, we applied the six-term formula to compute LMP_0 to achieve better accuracy.)

Once LMP_0 is derived, its lot average cost LAC_0 (as well as its lot total cost) can be easily estimated by Equation (15). This computation is faster and more straightforward than using UC curve (Equation (1)) directly, especially when there are thousands of units in the lot. The respective PI of this particular lot is then calculated in log space (just like a linear model) using LMP_0, along with other potential predictors (if any). Since the PI computation is done in log space, we need to exponentiate the PI back to unit space. See the PI discussion below.

Cost Uncertainty Analysis—PI for LTC (and LAC)

The proper measure of the quality of the estimate is the prediction interval. A PI provides a range of values around the PE at different probability levels to show the degree of confidence in the estimate based upon the sample evidence. The upper and lower bounds of the intervals form the branches of hyperbolas about the regression equation and illustrate the usefulness of the equation for predicting individual values from the independent variables.

A prediction interval can be thought of as a range defined by the PE plus or minus some number of adjusted standard errors (standard errors adjusted for prediction), depending upon the level of confidence. This adjusted standard error (Adj. SE) is a function of the standard error (SE) of the regression, the sample size, and the "distance" of the estimating point from the center of the database used to generate the CER.

Use of LMP for Cost Uncertainty Analysis

Although we can easily generate a PI for a particular unit cost, we cannot easily derive a PI for the total cost or average cost of a future production lot using a UC curve directly. In fact, it is difficult to generate a PI for the LTC based upon the unit cost equation (i.e., $UC_n = T_1{}^*n^{b*}f(x)$), because the LTC is the sum of all the units in the lot. Although the Monte Carlo simulation method has been suggested to generate the percentiles of the LTC for cost uncertainty analysis (Coleman et al., 2010), the process becomes very cumbersome if there are many production units in the lot. Also, it may not be appropriate to use UC curves directly to analyze cost uncertainty for a lot since the LMP (not the individual unit number) is used as a predictor when developing a UC curve on lot cost data. For consistency, we should also use the estimated LMP to generate the PI for a future lot using Equation (2), i.e., $LAC = T_1{}^*LMP^{b*}f(x)$. I believe this approach is cleaner, faster, and more realistic. For example, if a lot consists of 50 units, we should use its LMP and other cost drivers (if any) to construct a PI for its total cost rather than adding up 50 individual risk distributions by Monte Carlo simulation. The worst part: we still have to specify 50 PIs for the individual risk distributions when using the "simulation" method to sum unit costs to derive the percentiles for the LTC.

In a simple linear CER with an additive error term, where $Y = \beta_0 + \beta_1 X + \varepsilon$, $a(1 - \alpha)100\%$ PI for a future observation Y, when $X = \mathrm{x}_0$ is given by:

$$\begin{aligned} PI &= \hat{y}_0 \pm (t_{\alpha/2,n-2})SE\sqrt{1+\tfrac{1}{n}+\tfrac{(x_0-\bar{x})^2}{SS_{xx}}} \\ &= \hat{y}_0 \pm (t_{\alpha/2,n-2})SE\sqrt{1+\tfrac{1}{n}+\tfrac{((x_0-\bar{x})/S_x)^2}{n}}\,, \\ &= \hat{y}_0 \pm (t_{\alpha/2,\,n-2})^*(Adj.SE) \end{aligned} \tag{17}$$

where

$\hat{y}_0 =$ the estimated value from the CER when $X = x_0$,
$x_0 =$ the value of the independent variable used in calculating the estimate,
$n =$ the number of data points,
$t_{(\alpha/2,n\text{-}2)} =$ the upper $\alpha/2$ cut-off point of the student's t distribution with (n - 2) degrees of freedom (DF),
$SE =$ CER's standard error of estimate (also referred to as SEE),

Adj. SE = the adjusted standard error for PI,
$\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$; the mean of the independent variable in the data set,
$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$; the sum of squares of the independent variable about its mean,
$S_x = \sqrt{SS_{xx}/n}$; the uncorrected sample standard deviation of the independent variable,
ε = the error term with mean of 0 and variance σ^2 (assumed to follow a normal distribution).

For a CIC model ($y = f(x)^{*}\varepsilon$), the actual curve-fitting is done in log space when the CER error term is assumed to be log-normally distributed. Therefore, the (1 - α)100% PI for this particular lot when $LMP = LMP_0$ is given by:

$$\begin{aligned} PI &= Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,n-2})^{*}SE^{*}\sqrt{1+\frac{1}{n}+\frac{(\ln(LMP_0)-\overline{\ln(LMP)})^2}{\sum_{i=1}^{n}(\ln(LMP_i)-\overline{\ln(LMP)})^2}}\right), \\ &= Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,n-2})^{*}(Adj.SE)\right) \end{aligned} \tag{18}$$

where $\hat{y}_{\log}$ is the estimated value in log space when the lot midpoint is at LMP_0, $\overline{\ln(LMP)}$ is the average value of all LMPs evaluated in log space, SE is the standard error of estimate in log space, and “ln” stands for the natural logarithm function.

If there are multiple drivers in the CER, we will compute PI using matrix operations. See the (1 - α)100% PI formula below for a log-linear CER at a given driver vector $\underline{\mathbf{x_0}}$:

$$\begin{aligned} PI &= Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,n-p})^{*}SE^{*}\sqrt{1+\ln(\underline{\mathbf{x}}_0)(\mathbf{X}'\mathbf{X})^{-1}\ln(\underline{\mathbf{x}}_0)'}\right) \\ &= Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,n-p})^{*}(Adj.SE)\right) \end{aligned}, \tag{19}$$

where

p = the total number of estimated parameters, including the intercept,
$t_{(\alpha/2,n\text{-}p)}$ = the upper $\alpha/2$ cut-off point of the student’s t distribution with (n - p) DF,
SE = CER’s standard error of estimate (evaluated in log space),
$\ln(\underline{\mathbf{x}}_0) = (1, \ln(x_{10}), \ldots, \ln(x_{k0}))$, a row vector of given driver values in log space and 1 is for the intercept (Note: $p = k + 1$),
X = the design matrix of the independent variables in log space.
(The apostrophe superscript denotes the transpose of a vector or a matrix.)

Since the computation of PI for learning curves is done in log space, the resultant PI in unit space will be asymmetrical.

DF Factor. As shown by Equation (17), if the error term follows approximately a normal distribution, we should use the student’s t distributions provided in risk analysis tools (such as Crystal Ball, @Risk, or ACEIT) to model the CER uncertainty. We can enter the *Adj. SE* into the scale field and specify the DF in the degrees of freedom field when modeling risk using a Student’s t distribution. Alternatively, we can enter the low/high bounds to specify the distribution. Note that the upper cut-off point ($t_{\alpha/2,df}$) is derived from a Student’s t distribution that has the same degrees of freedom as the CER. However, if we use the *Adj. SE* to model the CER uncertainty by a different distribution, a DF adjustment factor should be applied to account for small samples. For example, we should multiply the *Adj. SE* measure by the DF factor to account for the broader tails of the t distribution for small samples if we use normal instead of t distribution for cost uncertainty analysis:

$$\text{DF factor} = \sqrt{\frac{df}{df-2}}, \tag{20}$$

where "*df*" stands for the degrees of freedom of the *t* distribution. In fact, Equation (20) is the standard deviation of the student's *t* distribution with a scale parameter one and "*df*" degrees of freedom.

For CICs, we should, in fact, apply the log-t distributions directly to model uncertainties, since the CER errors are commonly assumed to follow the log-normal distribution. (Just as with a Student's *t* distribution, we can enter the *Adj. SE* into the scale field and specify the DF in the degrees of freedom field when selecting a log-t distribution. Alternatively, we can enter the low/high bounds to specify the distribution.) If the log-t distributions are not available in the risk tools, we can specify the Student's *t* distributions in log space, but we should ensure the PI is transformed back to unit space as given in Equations (18) and (19).

The DF factor can be ignored when the sample size is fairly large (e.g., $df > 50$) or the upper/lower bounds of PI (rather than the *Adj. SE*) are specified to model the risk distribution.

Distance Factor. The last term in these PI equations is a "distance" adjustment factor, which should be applied to account for the location of the estimating point. It assesses the "distance" of the estimating point from the centroid of the predictors. The *Adj. SE* (as well as PI) gets larger when the estimating point moves farther away from the center of the database. This is especially true when the CER is used beyond the range of the data used in developing the CER. Hence, using the CER's standard error alone for risk assessment may significantly underestimate the risk associated with the PE unless the PE is very close to the center of the database and the sample size is fairly large. For a single variable model, the range of PI is the smallest when the estimating point is exactly the mean of the independent variable and the last term is reduced to sqrt($1 + 1/n$) when it happens.

Both the *DF* and *distance* adjustments should be considered when constructing PIs. Otherwise, the range of PI (based upon SE alone) will be smaller than it should be.

Advantages of Using LPP for CAC-Iterative Curves

As mentioned above in the regression method section, the LPP for a CAC-Direct curve is the last unit of the lot, while the LPP for a CAC-Iterative curve lies outside the lot except for lot 1. Before describing the benefits of using LPPs for CAC-Iterative curves, we will first discuss this characteristic, beginning with the estimation of LAC.

Estimation of LAC

Unlike the unit cost curve, we can easily compute the LTC and LAC for any given lot using CAC theory (see the first line of Equation (21)). We do not have to derive the lot plot point for this computation; in fact, it is easier to use the CAC curve directly. With a given first unit cost and a CIC slope, we can predict the lot average cost for a lot with a prior quantity PQ_0 and a lot quantity LQ_0 using Equation (10) based upon CAC theory:

$$\begin{aligned} LAC_0 = LTC_0/LQ_0 &= T_1 \tfrac{(PQ_0+LQ_0)^{b+1}-(PQ_0)^{b+1}}{LQ_0} f(\underline{\mathbf{x}}_0) \\ &= T_1\, LPP_0^b f(\underline{\mathbf{x}}_0) \quad (= CAC_{LMP_0}) \end{aligned}, \tag{21}$$

where LAC_0 is the estimated lot average cost, LTC_0 is the estimated lot total cost, LPP_0 is the lot plot point, and $f(\underline{\mathbf{x}}_0)$ is the multiplicative function of the predictors for this particular lot, respectively. It follows from Equation (21) that the lot plot point, LPP_0, is calculated by:

$$LPP_0 = \left(\frac{(PQ_0 + LQ_0)^{b+1} - (PQ_0)^{b+1}}{LQ_0} \right)^{1/b}. \qquad (22)$$

For CAC-Iterative Curves, LPPs lie outside the bounds of their respective lots except for lot 1. The effective LPP represented by Equation (14), as well as Equation (22), is *not* a lot midpoint at all. It is just a plot point for the curve-fitting process. This point will always lie to the right of the last unit of the lot except for lot 1. In the first lot, when the prior quantity is zero, the plot point is at the last unit of the lot, and the lot average cost is the cumulative average cost (as it should be). In all later lots, Equation (14) (as well as Equation (22)) must yield an effective LPP that is outside the bounds of the lot! Although many analysts have found these "astronomical" LPPs confusing and hard to explain, the LPPs are in fact quite useful for developing CAC-Iterative curves and generating the statistical measures for cost uncertainty analysis.

Advantages of Using the Iterative Approach for the CAC Curve

Listed below are the benefits of using the iterative approach to generate the CAC curve:

- The iterative approach can easily handle missing, non-consecutive, and concurrent lots.
- With the LPP as a driver variable, it is no longer necessary to apply non-linear regression to derive a solution. Equation (12) (as well as Equation (13)) becomes a log-linear model when the (composite) independent variable, LPP, is used in the CER. This would be a big advantage in the early days to avoid using non-linear regression.
- The traditional goodness-of-fit measures for the log-linear CAC curves can be applied to judge the quality of the fit in log space, and the outliers can be easily identified for further scrutiny.
- The statistical measures generated by the iterative approach are much more reliable and realistic than those produced by the direct approach.
- The slopes derived by these two methods (Unit and CAC Iterative) are generally very close to each other when the production units are above 30 and the noise in the curve is small. Furthermore, the T_1 based upon CAC theory is approximately $1/(1+b)$ times the T_1 derived from UC theory.
- If the prior quantity of a future lot is present (i.e., $PQ > 0$), it is also difficult to generate a PI for the LTC (or LAC) for CAC-Iterative curve (see Equation (12) or (21)). However, PIs can be easily built in log space using the LPP, along with other potential cost drivers (if any). The process of generating PIs is given above in the unit curve section.

Note that both unit and (iterative) CAC theories are, in fact, the "lot average cost" theories because they are applied to the lot average costs. In other words, the dependent variable used in the regression analysis is the lot average cost (or lot total cost), not the individual unit cost or the cumulative average cost unless we are dealing with single unit lot data.

Pitfalls of Using CAC-Direct Curves

In this section, we illustrate three CICs using a realistic learning curve example from ICEAA CEBoK® Module 7 (FD05, Learning Curve Analysis). We use this example to demonstrate that (1) we should not solely rely on the fit measures to select a best CER and (2) we cannot use a CAC-Direct curve for cost uncertainty analysis.

We are given the lot data in Table 2 for a generic missile program and we want to determine the cost of a lot produced in 1991 consisting of 1,430 units.

Three cost improvement curves are derived based upon the unit and CAC theories (see Table 3).

As shown by Table 3, Equation (I) is developed using unit theory; Equations (II) and (III) are both developed using CAC theory. However, Equation (II) is derived from the iterative approach while Equation (III) is developed using the direct approach. In other words, the dependent variable in "CAC-Iterative" is the lot average cost, while the dependent variable in "CAC-Direct" is the cumulative average cost.

Note that both "UC" and "CAC-Iterative" equations are very similar to each other. They have about the same fit and predictive measures (see Table 4); their CIC slopes are also the same. The only difference between these two curves is the first unit cost; namely, the T_1 based upon unit theory is about $(1+b)$ times the T_1 based upon CAC theory:

$$Unit\_T_1 = CAC\_T_1{}^*(1+b),$$

TABLE 2 Learning curves: Imperfect example from CEBoK, Module 7

Year	Lot quantity	First unit	Last unit	Lot total cost ($K)	Lot average cost ($K)	Cum Avg. cost ($K)
1981	1,106	1	1,106	60,275.2	54.5	54.5
1982	1,585	1,107	2,691	75,169.2	47.4	50.3
1983	2,447	2,692	5,138	87,791.7	35.9	43.4
1984	1,517	5,139	6,655	47,072.1	31.0	40.6
1985	1,983	6,656	8,638	62,745.4	31.6	38.6
1986	1,574	8,639	10,212	41,318.7	26.3	36.7
1987	2,643	10,213	12,855	74,676.3	28.3	34.9
1988	887	12,856	13,742	29,998.2	33.8	34.9
1989	1,871	13,743	15,613	48,931.4	26.2	33.8
1990	2,194	15,614	17,807	54,109.0	24.7	32.7
1991	1,430	17,808	19,237	?	?	?

TABLE 3 Three cost improvement curves—UC, CAC-Iterative, and CAC-Direct

	Equation	% CIC slope	SE	% Adj. R^2
I. UC	192.8 * Unit_Num ^ (−0.2043)	86.8	0.104	84.2
II. CAC-Iterative	241.1 * Cum_Qty ^ (−0.2038)	86.8	0.104	84.2
III. CAC-Direct	223.8 * Cum_Qty ^ (−0.1950)	87.4	0.027	97.6

Note: 241.1 * $(1+b) = 192$.

TABLE 4 Percentage error comparison table for "lot total" costs

Year	Actual lot cost	Pred unit theory	Pred CAC iterative	Pred CAC direct	% Error unit	% Error CAC iterative	% Error CAC direct
1981	60,275	63,890	63,916	63,111	6.0	6.0	4.7
1982	75,169	65,862	65,819	66,000	−12.4	−12.4	−12.2
1983	87,792	87,409	87,378	88,195	−0.4	−0.5	0.5
1984	47,072	49,667	49,658	50,313	5.5	5.5	6.9
1985	62,745	61,568	61,564	62,519	−1.9	−1.9	−0.4
1986	41,319	46,807	46,808	47,623	13.3	13.3	15.3
1987	74,676	75,440	75,449	76,898	1.0	1.0	3.0
1988	29,998	24,580	24,584	25,089	−18.1	−18.0	−16.4
1989	48,931	50,820	50,831	51,918	3.9	3.9	6.1
1990	54,109	58,036	58,052	59,362	7.3	7.3	9.7
1991	?	37,035	37,047	37,917			
MAD of % errors					7.0	7.0	7.5
RMS of % errors					8.9	8.9	9.3

where $b = \ln(\text{slope}/100)/\ln(2)$. Although "CAC-Direct" has the best goodness-of-fit measures among the three CICs, its ability to predict lot costs is not necessarily better than the other two curves. See Table 4 for details.

Common Practice

Analysts have noticed that (1) data smoothing is a problem for CAC-Direct curves and (2) it is unfair to compare the goodness-of-fit measures between UC and CAC-Direct curves (Cullis et al., 2008). However, the fit-statistics of the regression analysis (e.g., standard error) are still compared to determine which CIC theory best fits the data set. In fact, the fit measures should be used mainly to determine whether the fitted model and the regressed coefficients are significant.

Caution. We should not select a CER solely based upon the goodness-of-fit measures, especially when the dependent variables used in the equations are not the same. Even though the SE measure for "CAC-Direct" is almost four times better than both UC and CAC-Iterative, its ability to predict lot costs is, in fact, inferior to both of its competitors. The bottom line: we cannot compare the goodness-of-fit measures between CERs when (1) the fit spaces are different or (2) the dependent variables in the CERs are different.

Estimating PI for a Future Lot Cost

The 10%/90% PIs for the 1991 production lot by the three CICs are given in Table 5.

These prediction intervals are derived from Equation (18), using the values from the lot plot point column as LMP_0 in the PI equation. As shown by Table 5, the 10th and 90th percentiles under the UC and CAC-Iterative curves are about 15% below and 17% above the estimated values, respectively. However, the 10th and 90th percentiles of the CAC-Direct curve are just about 5% off the estimated value. This extremely tight PI is caused by the low SE of the CAC-Direct curve by the smoothing effect. In fact, we should not use

TABLE 5 Prediction interval results for 1991 production lot using equations I to III

PI result	10% Lower bound	Year 1991 estimate	90% Upper bound	10% Low multiplier	90% High multiplier	Lot plot point
UC Theory	31,590.6	37,034.9	43,417.5	85.3%	117.2%	18,517.0
CAC-Iterative	31,595.4	37,046.8	43,438.7	85.3%	117.3%	56,653.7
CAC-Direct	36,109.6	37,929.6	39,841.3	95.2%	105.0%	56,318.2

this SE (0.027) at all because it is generated by the cumulative average costs, rather than the lot costs. If we use the SE from the direct approach to develop the PI for a future lot, the results will be misleading and incorrect. (This is a common mistake when building the PI using a CAC-Direct curve.) Further, we should consider the following question.

Question. Given a CAC-Direct curve (e.g., Equation (III)), can we generate a PI for a lot when its prior quantity is greater than zero? The answer to this question is probably "no"—we cannot build a PI from a CAC-Direct curve. This is because we do not know the distribution for this lot, as it is the difference between two correlated log-normal distributions. In mathematical terms, let X denote the distribution of the lot total cost from year 1981 to year 1990 and let Y denote the distribution of the lot total cost from 1981 to 1991. Both X and Y are commonly assumed to follow log-normal distributions under the traditional CAC theory. However, the distribution of "$Y – X$" (for the cost of 1991) is not easily identified.

Recommendation. Data smoothing is problematic—it generates an artificially tight SE, which cannot be compared with the SE in unit theory curve. It cannot be used in cost uncertainty analysis either. Hence, we suggest using the iterative (not direct) approach to fit cumulative average CICs.

As noted above, we should use the LMP to generate the PI for a future production lot for cost uncertainty analysis because we usually observe lot cost data rather than individual unit cost data. Hypothetically, what if the all the observations are single unit data points, say from unit 1 to unit 10, and we want to use this information to analyze the total cost for the next production lot, say from unit 11 to unit 20? In this situation, should we generate a PI using the LMP to analyze the uncertainty for the total cost or should we use Monte Carlo simulation to sum ten risk distributions to determine the percentiles of the LTC? The answer is debatable. If a future lot consists of many units, say 100, then we should definitely use the LMP to construct a PI for the LTC. However, if this lot only consists of a handful of units, we should probably use Monte Carlo simulation to add up these risk distributions using unit cost curve. But where do we draw the line? (Note that we should still build a PI for each unit cost distribution in the lot when using the simulation method to sum the unit costs—it is a labor-intensive job!) We can certainly compare the difference between these two methods using unit cost data—this will be a future study item.

Validating CICs Generated by (Two-Step) Regression Method

We wanted to test the CIC features offered by CO$TAT using a third party tool. (CO$TAT is our in-house statistical package tailored for cost analysts.) However, we could not find a single tool to test all the CIC features, so we used the Air Force Cost Analysis Agency's (AFCAA) LEARN Program and the Solver tool (an MS Excel add-in program).

AFCAA's LEARN Program

We used AFCAA's LEARN program as a comparison baseline in the early 1990s to test the CICs generated by the traditional two-step regression method. We tested several examples using UC, CAC-Direct, and Learning with Rate curves in the LEARN program. (The rate slope could be either fixed or determined by the regression model.) The two-step CIC results were found to be very close to those generated by the LEARN program.

Solver (a Simultaneous Minimization Process)

However, the LEARN program cannot generate CAC-Iterative curves; it does not offer the MUPE option either when developing CICs. So we used Solver to validate the CIC features. We tested several lot cost data sets, including many of AFCAA's real examples, but did not find any substantial differences between the conventional two-step regression method and the Solver approach when developing CICs. The differences of the fitted coefficients between these two methods are all within 0.5%, and most of them are well within 0.1%. In summary, we used Solver to validate UC and CAC-Iterative curves using the conventional regression method:

- Compared CO$TAT's CIC results generated by the conventional two-step method (i.e., the curve fit is done in log space) against Solver's solutions.
- A rate term was also considered in the validation process.
- The MUPE method was explored.

These results will be discussed in a different paper.

Nonlinear least squares (NLS) regression has also been suggested as a way to fit the unit theory curves directly in unit space (Goldberg & Touw, 2005; Lee, 2005). The CER errors are assumed to be additive when using NLS to minimize the sum of squared differences between the actual and predicted in unit space. However, multiplicative (not additive) error terms are commonly used in the cost analysis field because experience tells us that the error of an individual observation (e.g., cost) is generally proportional to the magnitude of the observation, not a constant. (See the error term specification in Equations (1) and (2).) Therefore, we either take the log-transformation or use the MUPE method (a weighted least squares) to analyze multiplicative error models.

Conclusions

Both UC- and CAC-Iterative curves are, in fact, derived from the "lot average cost" theory. Instead of analyzing unit costs directly using unit cost data, analysts are often provided with lot cost data. Both unit and (iterative) CAC theories are in fact the "lot average cost" theory, as the *dependent* variable used in the regression analysis is the lot average cost (or lot total cost). The dependent variable is neither the individual unit cost nor the cumulative average cost unless we are dealing with single unit lot data.

The six-term formula is proven to be more accurate than the traditional two-term formula to approximate the lot total cost for unit cost curves. This six-term formula represents a simple and more accurate approximation than the traditional two-term formula to derive the lot total cost, as well as the lot midpoints, for unit theory CICs. Although the error of the two-term formula is generally quite small, the accuracy improves significantly using the six-term formula (see Table 1 above). The improvement can be substantial for estimating the costs of expensive lots or a Work Breakdown Structure with many different elements. This also helps during the curve-fitting process, as we need to closely estimate a very important independent variable—the lot midpoint.

Use the iterative (not the direct) approach to derive CAC curves. We cannot place too much credence on a CER generated by the direct approach as it smoothes the cost data by summing and averaging the lot total costs from the very first lot. Consequently, every observation (except for the first one) depends upon all previous observations, which violates the basic assumptions of OLS regression analysis. Because of the smoothing effect, the direct approach generates *artificially* better goodness-of-fit measures than those done by the iterative approach. (The fit measures produced by the iterative approach are more reliable and realistic.) More importantly, a CAC-Direct curve does not necessarily predict lot costs better. Hence, using the adjusted SE from the traditional cumulative average curve for cost uncertainty analysis will underestimate the uncertainty. In fact, we cannot accurately construct the PI for a production lot when its prior quantity is greater than zero using a CAC-Direct curve. Further, the direct approach cannot handle missing, nonconsecutive, and concurrent lots. The use of the traditional CAC curves is limited.

Do not select a CER solely based upon the goodness-of-fit measures, especially when the dependent variables used in the equations are *not* the same. Statistics of the regression analysis (e.g., standard error of estimate, adjusted R^2, etc.) are usually compared to determine which theory (UC or CAC) best fits the data set. However, we should not compare the fit measures across different models when either the fit spaces or the dependent variables are different. Use the goodness-of-fit measures to determine whether the fitted model is significant.

The conventional CIC is biased low as the actual curve fit is done in log space; it should be adjusted to reflect the mean in unit space. The commonly used correction factors are Goldberger's Factor, the Smearing Estimate, the PING Factor, etc. Alternatively, both UC and CAC curves can be derived directly in unit space using the MUPE method to eliminate the bias.

Both the LMP and LPP are useful statistics for developing CICs and generating PIs for cost uncertainty analysis. Also, use log-t distribution to model uncertainty distributions for the traditional CIC curves. The LMP (in UC curves) and LPP (in CAC-Iterative curves) are derived during the curve-fitting process. The LMPs in unit curves are very useful for (1) developing unit theory curves, (2) estimating the lot cost for a future lot, and (3) generating PIs for cost uncertainty analysis. The effective LPP for CAC-Iterative curves represented by Equation (14) is not a midpoint at all. It is just a plot point for the curve-fitting process. This point is always outside the bounds of the lot except for lot 1. Although the LPP (generated by the CAC-Iterative curve) is not intuitively appealing, it is, in fact, quite useful for cost uncertainty analysis under CAC theory.

Consider applying both DF and distance adjustment factors to a CER's standard error for cost uncertainty analysis. The proper measure for analyzing cost uncertainty is the PI if CERs are used as cost estimating methodologies. The DF and distance adjustments should be considered when constructing PIs. Otherwise, the range of PI (based upon a CER's SE alone) will be smaller than it should be. The DF adjustment accounts for the broader tails of Student's *t* (or Log-t) distribution for small samples while the distance adjustment factor accounts for the location of the estimating point. (Note that PI gets larger when the estimating point moves farther away from the center of the database.)

Do not use Monte Carlo simulation to generate cost uncertainty results for a lot total cost using a UC curve. For example, if a lot consists of 50 units, we should use its LMP and other cost drivers (if any) to construct a PI for its total cost rather than adding up 50 individual risk distributions by Monte Carlo simulation. (Note: we should still build a

PI for each of these 50 individual risk distributions when using the "simulation" method to sum the unit costs in the lot—it is a labor-intensive job!)

CICs produced by the conventional two-step regression method have been validated using a simultaneous minimization process, i.e., Solver. We used Solver to test several lot cost data sets, including many of the AFCAA's real examples, but did not find any substantial differences between the conventional two-step regression method and the Solver approach when developing CICs. The differences of the fitted coefficients between these two methods are all within 0.5%, and most of them are well within 0.1%. Besides Solver, we have also used AFCAA's LEARN program for validation, although the LEARN program generates the CICs through the conventional two-step regression method.

Future Study Items

The following topics should be addressed in the future:

1. Compare the CIC results between the iterative and direct approaches under CAC theory.
2. Compare the CIC results between the traditional two-step regression method and the MUPE method.
3. Repeat (1) and (2) with inclusion of a Rate term.
4. Examine the difference between the Monte Carlo simulation method and the LMP method for analyzing the cost uncertainty for a future production lot, which follows a unit cost curve. For example, if a UC curve is developed by unit cost data and if a future lot consists of 10 units, should we use its LMP to construct a PI to analyze its total cost or should we add up 10 individual uncertainty distributions by Monte Carlo simulation?

References

Anderson, T. (2003, January 28–31). The Trouble With Learning Curves. *36th Annual DoD Cost Analysis Symposium*, Williamsburg, VA.

Cho, C., Schmidt, B. K. (1984). Estimating Production Cost with Learning Curves: A Closed Form Approximation. *Journal of the National Estimating Society*, Spring, 8–11.

Coleman, R. L., Braxton, P. J., Druker, E. R., Cullis, B. L. (2010, February 16–19). Estimating the Cost and Risk Impact of Learning Curve Differences. *43rd Annual DoD Cost Analysis Symposium*, Williamsburg, VA.

Crawford, J. R. (1947). *Learning Curve, Ship Curve, Ratios, Related Data*. Burbank, California: Lockheed Aircraft Corporation.

Cullis, B. L., Coleman, R. L., Braxton, P. J., McQueston, J. T. (2008, June 24–27). CUMAV or Unit? Is Cum Average vs. Unit Theory a Fair Fight? *2008 SCEA/ISPA Joint Annual Conference*, Industry Hills, CA.

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association*, 78(383), 605–610.

Goldberg, M. S., Touw, A. E. (2000, February 1–4). Statistical Considerations in Estimating Learning Curves and Multiplicative CERs. *33rd Annual DoD Cost Analysis Symposium*, Williamsburg, VA.

Goldberg, M. S., Touw, A. E. (2005, February 15–18). Statistical Methods for Learning Curves and Cost Analysis. *38th Annual DoD Cost Analysis Symposium*, Williamsburg, VA.

Goldberger, A. S. (1968). The Interpretation and Estimation of Cobb-Douglas Functions. *Econometrica*, 35, 464–472.

Hu, S. (2001, June 12–15). The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development. *3rd Joint Annual ISPA/SCEA International Conference*, Vienna, VA.

Hu, S. (2005, June 14–17). The Impact of Using Log-Error CERs Outside the Data Range and PING Factor. *5th Joint Annual ISPA/SCEA Conference*, Broomfield, CO.

Hu, S., Fong, F., Enser, B. (2006, June 13–16). Cost Improvement Curve Analysis of the USCM8 Database Using Quantity As an Independent Variable (QAIV). *2006 Annual SCEA International Conference*, Tysons Corner, VA.

Hu, S., Sjovold, A. R. (1989, March). *Error Corrections for Unbiased Log-Linear Least Square Estimates* (TR-006/2). Santa Barbara, CA: Tecolote Research, Inc.

Hu, S., Sjovold, A. R. (1994, June 7–9). Multiplicative Error Regression Techniques. *62nd MORS Symposium*, Colorado Springs, CO.

Kluge, A. J. (1975, July). *Iterative Procedure for Fitting Cost Improvement Curves (COSTNOR)* (TM 28). Santa Monica, California: Tecolote Research, Inc.

Large, J. P., Hoffmayer, K., Kontrovich, F. (1974, December). *Production Rate and Production Cost* (R-1609-PA&E). Santa Monica, California: RAND Corporation.

Lee, D. A. (2005, February 15–18). Mathematical Methods for Cost Estimating and Analysis. *38th Annual DoD Cost Analysis Symposium*, Williamsburg, VA.

Levenson, G. S., Boren, H. E., Tihansky, D. P., Timson, F. (1971, December). Cost Estimating Relationships for Aircraft Airframes (R-761-PR). RAND Corporation.

SCEA. (2002–2011). FD05: Learning Curve Analysis. Cost Estimating Body of Knowledge (CEBoK), Module 7. Vienna, Virginia: SCEA Professional Development & Training Workshop.

Seber, G. A. F., Wild, C. J. (1989). *Nonlinear Regression*. New York: John Wiley & Sons, pp. 37, 46, 86–88.

Weisberg, S. (1985). *Applied Linear Regression* (2nd Edition). New York: John Wiley & Sons, pp. 87–88.

Wright, T. P. (1936). Factors Affecting the Cost of Airplanes. *Journal of Aeronautical Sciences*, 3(6), 122–128.

About the Authors

Shu-Ping Hu is a Chief Statistician at Tecolote Research, Incorporated. Shu-Ping joined Tecolote in 1984 and serves as a company expert in all statistical matters. She earned her Ph.D. in Mathematics, with an emphasis in Statistics, at the University of California, Santa Barbara.

She has published many technical papers, covering such topics as developing the PING Factor to adjust the log-linear CER to reflect the mean and suggesting an adjusted R-square measure for the Minimum-Unbiased-Percentage Error (MUPE) and Minimum-Percentage Error Regression under Zero-Percentage Bias (ZMPE) CERs.

Dr. Hu has 20 years of experience supporting Unmanned Space Vehicle Cost Model (USCM) CER development and the related database. She also has 25 years of experience designing, developing, and validating statistical, learning, and regression algorithms in CO$TAT. In addition, Dr. Hu developed many of the distribution and correlation algorithms implemented in the ACE RI$K simulation tool. For over 20 years, she has been a regular presenter of the most advanced cost analysis techniques at major cost conferences.

Alfred Smith earned a Bachelor Mechanical Engineering degree from the Canadian Royal Military College and a Master of Science with Distinction in naval architecture from the University College, London, England.

He served 21 years in the Canadian Navy in a variety of positions such as submarine Navigator and Operations Officer and then as a naval architect. In addition, he has over 20 years experience leading, executing, or contributing to life cycle cost model development and cost uncertainty analysis for a wide variety of DoD, Coast Guard, NASA, and international projects.

Alfred has been employed by Tecolote Research, Inc. since 1995 and became its General Manager for Software Products/Services Group in 2000. His team develops, distributes and supports of a variety of web and desktop products supporting the cost community.

Alfred has delivered numerous papers on cost risk analysis topics and was the lead writer of the AFCAA Cost Risk and Uncertainty Handbook. He is certified by ICEAA as a Certified Cost Estimator/Analyst (CCEA$^{®}$).