TECOLOTE RESEARCH, INC.
Bridging Engineering and Economics
Since 1973

Comparing Different Methods for Deriving Cost-Dependent Cost Estimating Relationships (CER)

Dr. Shu-Ping Hu

2009 ISPA/SCEA Professional Development and Training Workshop
St. Louis, MO
2 to 5 June 2009

TECOLOTE RESEARCH, INC.
1 S. Los Carneros Road, Suite 125
Goleta, CA 93117-5506
(805) 571-6366

TECOLOTE RESEARCH, INC.
5266 Hollister Ave., Suite 301
Santa Barbara, CA 93111-2089
(805) 964-6963

Comparing Different Methods for Deriving Cost-Dependent CERs

Dr. Shu-Ping Hu

ABSTRACT

Parametric cost estimating relationships (CER) are developed using historical data. Regression analysis is used to determine whether the independent variables can help explain the variation in the dependent variable. There are two different types of independent variables in CERs: technical parameters (e.g., weight, power, etc.) and cost-dependent parameters (e.g., first unit production cost, Prime Mission Product [PMP] cost, etc.). When CERs are generated, they are based upon actual data from completed projects whether or not the independent variables are cost-dependent or hardware design-driven.

Although CERs are developed in parallel using the "actual" data set, they may be used in series in cost uncertainty analysis, especially when the aggregated costs are used as independent variables. For example, let us consider the System Engineering and Program Management (SEPM) cost as a function of the PMP cost. The SEPM cost can be estimated only after the PMP cost is derived by the cost model (through another CER or a series of CERs). Since we use the "CER-estimated" (not actual) PMP to predict SEPM when analyzing cost uncertainties, the estimate and variance for SEPM as well as total cost will be inaccurate if the SEPM CER is built using the "actual" cost. In other words, this two-step process introduces error into the independent variable PMP, which is further compounded with the error of estimating SEPM cost and the total project cost. Reference 1 suggests using an alternative approach to avoid these errors: develop the SEPM CER using the "estimated" PMP cost instead of the "actual" PMP cost.

This paper will first examine whether this alternative method makes sense from a statistician's perspective. A mathematical proof is provided. We will then apply this alternative method to analyze the SEPM and integration, assembly, and test (IA&T) CERs in the Unmanned Space Vehicle Cost Model, Eighth Edition (USCM8) database. The USCM8 CERs were developed using the Minimum-Unbiased-Percentage Error (MUPE) method to model multiplicative errors. The MUPE method is also known as an iterative, weighted least squares (WLS) regression. The goal of this paper is to compare the cost-dependent CERs generated by the current (based on actual data) and the alternative methods to determine whether there are any significant differences. It will also compare their respective standard percent errors (SPE).

OUTLINE

The objectives of this paper are twofold. First we will examine the alternative method suggested by Reference 1 to see if we can offer any statistical validations for the new method under the linear models. We will then apply this alternative approach to the cost-dependent USCM8 CERs to examine whether there are any significant differences when compared to the current CERs generated by actual data. Several analysts have expressed their interest in the past in comparing results between the current and alternative methods using "real" data sets; they were especially interested in the published USCM8 CERs for this comparison.

We will discuss the topics below in the following sections:

- Statistical Derivations under WLS

- For One-Independent-Variable Models
- For CERs with Multiple Independent Variables

- Comparisons of Standard Percent Errors
  - USCM8 IA&T CER
  - USCM8 SEPM CER (for Communication Satellites)
- Limitations and Concerns about the Alternative Method

STATISTICAL DERIVATIONS

We will prove the alternative method is statistically sound for linear models using weighted least squares (WLS), which includes MUPE. We will first provide the math derivations using simple linear models.

One-Independent-Variable Models. Given: the cost variable Y is a function of an independent variable Z (e.g., PMP) but the value of Z is estimated from another explanatory variable X (e.g., weight). We want to prove that regressing Y on the "estimated" Z value is the same as regressing Y on X directly if the functional form is linear. (In essence, the cost variable Y is simply a function of the independent variable X.)

In mathematical terms, the given conditions can be stated as

$$Y = a + b\,Z + \varepsilon \quad \text{where } E(\varepsilon) = 0 \text{ and } Var(\varepsilon) = V\sigma_1^2 \quad \text{(e.g., Z = PMP)} \qquad (1)$$

$$Z = c + d\,X + \delta \quad \text{where } E(\delta) = 0 \text{ and } Var(\delta) = V\sigma_2^2 \quad \text{(e.g., X = Weight)} \qquad (2)$$

Note that the variance matrix V is a diagonal matrix, with the non-negative value in the diagonals and zeros elsewhere, and ε/δ are the error terms. See Appendix A for more detailed definitions of the variance matrix, error terms, etc. and some introductory information on regression analysis. Note also the variance matrix V in Equation 1 and Equation 2 does not have to be the same. For simplicity, we use the same symbol V to illustrate the proof.

Normally, analysts would develop cost-dependent CERs using the actual cost from a historical database:

1. Regress Y on the variable Z (e.g., the **actual** cost of PMP) Method (1)

Reference 1 suggests using an **alternative** approach to avoid introducing unaccounted errors into the driver variable Z:

2. Regress Y on the **estimated** value of Z rather than the actual value of Z Method (2a)

Naturally, the **direct** and intuitive approach will be the following:

3. Regress Y on the variable X directly if possible Method (2b)

Our goal is to prove that **Methods 2a and 2b will deliver the same regression equation under WLS**. In other words, **the alternative and direct methods (Methods 2a and 2b) are equivalent under WLS**.

It follows from Equations 1 and 2 that the dependent variable Y can be modeled directly by the variable X as given below:

$$Y = (a + bc) + bd(X) + b\delta + \varepsilon = a' + b'\,X + r \qquad (3)$$

where:

$$r = b\delta + \varepsilon \quad (E(r) = 0,\ Var(r) = V\sigma^2)$$
$$\sigma^2 = \sigma_1^2 + b^2\sigma_2^2 \tag{4}$$

Two proofs are provided below for a one-independent variable case. The first one is for ordinary least squares (OLS) when all the observations have the same variance (V = I); the other is for weighted least squares (WLS) when the observations do not have the same variance (V ≠ I). (Note that the MUPE method is a WLS.)

<u>Proof of OLS (V = I)</u>

If V = I, then the OLS solution for Equation 2 is given by

$$\hat{Z}_j = \bar{Z} + \frac{SS_{XZ}}{SS_{XX}}(X_j - \bar{X}) \qquad \text{for } j = 1, \ldots, n \tag{5}$$

where:

$$SS_{XZ} = \sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}),$$
$$SS_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2,$$

$\hat{Z}_j$ is the predicted value of the jth observation, and

$\bar{X}$ *and* $\bar{Z}$ are the means of the X and Z variables, respectively.

If we regress Y against the predicted value of Z, which is denoted by $\hat{Z}$ (see Equation 5), then the regression equation should be given as follows:

$$\begin{aligned}
\hat{Y}_j &= \bar{Y} + \frac{SS_{Y\hat{Z}}}{SS_{\hat{Z}\hat{Z}}}(\hat{Z}_j - \bar{\hat{Z}}) = \bar{Y} + \frac{SS_{Y\hat{Z}}}{SS_{\hat{Z}\hat{Z}}}(\hat{Z}_j - \bar{Z}) = \bar{Y} + \frac{\sum_i (Y_i - \bar{Y})(\hat{Z}_i - \bar{\hat{Z}})}{\sum_i (\hat{Z}_i - \bar{\hat{Z}})^2}\frac{SS_{XZ}}{SS_{XX}}(X_j - \bar{X}) \\
&= \bar{Y} + \frac{\sum_i (Y_i - \bar{Y})\frac{SS_{XZ}}{SS_{XX}}(X_i - \bar{X})}{\sum_i \left(\frac{SS_{XZ}}{SS_{XX}}(X_i - \bar{X})\right)^2}\frac{SS_{XZ}}{SS_{XX}}(X_j - \bar{X}) \\
&= \bar{Y} + \frac{\frac{SS_{XZ}}{SS_{XX}}SS_{YX}}{(\frac{SS_{XZ}}{SS_{XX}})^2 SS_{XX}}\frac{SS_{XZ}}{SS_{XX}}(X_j - \bar{X}) \\
&= \bar{Y} + \frac{SS_{YX}}{SS_{XX}}(X_j - \bar{X}) \qquad \textit{for } j = 1, \ldots, n
\end{aligned} \tag{6}$$

Equation 6 is exactly the OLS regression equation when regressing Y against X. Note that in OLS the average of the predicted values is equal to the average of the observations (i.e., the dependent variable):

$$\bar{\hat{Z}} = \bar{Z} \tag{7}$$

Proof of WLS (V ≠ I)

Let us define an n-by-n weighting matrix W as follows (same as Equation 31 in Appendix A):

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ . & . & \dots & 0 \\ 0 & 0 & \dots & w_n \end{pmatrix} = \begin{pmatrix} 1/v_1 & 0 & \dots & 0 \\ 0 & 1/v_2 & \dots & 0 \\ . & . & \dots & 0 \\ 0 & 0 & \dots & 1/v_n \end{pmatrix} \tag{8}$$

Here v_i is used to denote the ith diagonal element of the variance matrix V. If the variance matrix V is different from the identity matrix (I), the weighted least squares (WLS) solution for Equation 2 is given below (see Equation 34 in Appendix A for details):

$$\hat{Z}_i = \overline{Z}_w + \frac{SS_{wXZ}}{SS_{wXX}}(X_i - \overline{X}_w) \qquad \text{for } i = 1, \dots, n \tag{9}$$

where:

$$\overline{Z}_w = \sum_{i=1}^{n} w_i * Z_i \Big/ \sum_{i=1}^{n} w_i \tag{10}$$

$$\overline{X}_w = \sum_{i=1}^{n} w_i * X_i \Big/ \sum_{i=1}^{n} w_i \tag{11}$$

$$SS_{wXZ} = \sum_{i=1}^{n} w_i (X_i - \overline{X}_w)(Z_i - \overline{Z}_w) \tag{12}$$

$$SS_{wXX} = \sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2 \tag{13}$$

Similar to the case of OLS (Equation 7), the <u>weighted</u> average of the predicted values is equal to the <u>weighted</u> average of the observations for WLS:

$$\overline{\hat{Z}}_w = \sum_{i=1}^{n} w_i * \hat{Z}_i \Big/ \sum_{i=1}^{n} w_i = \overline{Z}_w \tag{14}$$

If we regress Y against $\hat{Z}$ (i.e., Equation 5) using WLS, the regression equation is given by

$$\begin{aligned} \hat{Y}_j &= \overline{Y}_w + \frac{SS_{wY\hat{Z}}}{SS_{w\hat{Z}\hat{Z}}}(\hat{Z}_j - \overline{\hat{Z}}_w) = \overline{Y}_w + \frac{SS_{wY\hat{Z}}}{SS_{w\hat{Z}\hat{Z}}}(\hat{Z}_j - \overline{Z}_w) \\ &= \overline{Y}_w + \frac{\sum_i w_i (Y_i - \overline{Y}_w)\frac{SS_{wXZ}}{SS_{wXX}}(X_i - \overline{X}_w)}{\sum_i w_i \left(\frac{SS_{wXZ}}{SS_{wXX}}(X_i - \overline{X}_w)\right)^2} \frac{SS_{wXZ}}{SS_{wXX}}(X_j - \overline{X}_w) \\ &= \overline{Y}_w + \frac{\frac{SS_{wXZ}}{SS_{wXX}} SS_{wYX}}{(\frac{SS_{wXZ}}{SS_{wXX}})^2 SS_{wXX}} \frac{SS_{wXZ}}{SS_{wXX}}(X_j - \overline{X}_w) = \overline{Y}_w + \frac{SS_{wYX}}{SS_{wXX}}(X_j - \overline{X}_w) \quad \textit{for } j = 1, \dots, n \end{aligned} \tag{15}$$

This is exactly the WLS solution when regressing Y against X (see Equation 34 in Appendix A).

Based upon Equation 15, we can conclude that Method 2a and Method 2b generate the **same** regression equation and the same fit statistics. This conclusion holds whether or not the error term is assumed to be additive or multiplicative such as MUPE. This is because the MUPE CER is generated using iterative weighted least squares. Note that the resultant weighting factors when regressing Y against X under MUPE should be the same as those when regressing Y against $\hat{Z}$ as $\hat{Z}$ is a linear function of X. These mathematical derivations can be further verified by the empirical examples using different error terms.

CERs with Multiple Independent Variables. Now let us extend the above conclusion to CERs with multiple drivers. In other words, we want to prove that Method 2a is the same as Method 2b when the design matrix **X** for the variable Z consists of multiple independent variables. Note that we use "**X**" to denote the design matrix in the derivations below. For simplicity, the discussion given below is for the centered models; the dependent variables and the design matrices are all centered.

Given: $Y = b\,Z + \varepsilon$ where $E(\varepsilon) = 0$ and $Var(\varepsilon) = V\sigma_1^2$ (16)

$$Z = d\,X_1 + e\,X_2 + f\,X_3 + \ldots + \delta$$
$$= \mathbf{X}\,\beta + \delta \qquad \text{where } E(\delta) = 0 \text{ and } Var(\delta) = V\sigma_2^2 \tag{17}$$

⇨

$$Y = bd(X_1) + be(X_2) + bf(X_3) + \ldots + b\delta + \varepsilon$$
$$= b'\,X_1 + c'\,X_2 + d'\,X_3 + \ldots + r$$

where:

$r = b\delta + \varepsilon$ $(E(r) = 0 \text{ and } Var(r) = V\sigma^2)$

$\sigma^2 = \sigma_1^2 + b^2\,\sigma_2^2$

You can find the definition of the design matrix **X** in Appendix A; the variance matrix V is as defined above.

We want to use Equation 33 in Appendix A to prove the case of multiple drivers because it is easier and more straightforward to derive the least squares solution using the vector and matrix notations. Given Equation 17, it follows from Equation 33 in Appendix A that the predicted value of Z using the design matrix **X** is given by

$$\hat{Z} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WZ} \tag{18}$$

By the same rationale, if we use $\hat{Z}$ to predict Y, the weighted least squares solution of Y is given below:

$$\hat{Y} = \hat{\mathbf{Z}}(\hat{\mathbf{Z}}'\mathbf{W}\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\mathbf{WY} \tag{19}$$

Substituting Equation 18 into Equation 19, we can derive the CER-predicted Y as

$$\begin{aligned}\hat{Y} &= \hat{\mathbf{Z}}(\hat{\mathbf{Z}}'\mathbf{W}\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\mathbf{WY}\\ &= \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WZ}(\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WZ})^{-1}\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WY} \qquad (20)\\ &= \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WZ}(\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WZ})^{-1}\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WY}\end{aligned}$$

If we multiply both sides of Equation 20 by $(\mathbf{ZZ'W})^{-1}(\mathbf{ZZ'W})$, we will obtain the following equation:

$$\begin{aligned}\hat{Y} &= (\mathbf{ZZ'W})^{-1}\mathbf{ZZ'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WZ}(\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WZ})^{-1}\mathbf{Z'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WY} \\ &= (\mathbf{ZZ'W})^{-1}\mathbf{ZZ'WX}(\mathbf{X'WX})^{-1}\mathbf{X'WY} \qquad (21) \\ &= \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WY}\end{aligned}$$

Equation 21 is the WLS solution of $\hat{Y}$ using **X** as the design matrix (see Equation 33 in Appendix A). However, the inverse of (**ZZ'W**) may not exist; hence Equation 21 is not necessarily true. Therefore, Methods 2a and 2b may not generate the **same** regression equation when the design matrix **X** consists of multiple independent variables.

Degrees of Freedom (DF) when using Multiple Drivers. The alternative and direct methods (Methods 2a and 2b) do not have the same degrees of freedom when multiple drivers are used for estimating the cost-dependent variable Z (e.g., PMP). Although both the current and alternative methods (Methods 1 and 2a) may seem to have fewer DF than the direct method, that is not true in reality. In fact, the DF for both Methods 1 and 2a is artificially larger than it should be in this case.

COMPARISONS OF STANDARD PERCENT ERRORS

Which method will generate a smaller standard percent error (SPE)? Mathematically, we cannot conclude which method will generate a smaller SPE when comparing the commonly-used, current method with the alternative method (i.e., Method 1 vs. Method 2a). The goal of this paper is to compare the cost-dependent CERs generated by both methods (Methods 1 and 2a) using real examples. It will then compare their respective SPEs to determine whether there are any significant differences. By definition, the SPE measure is given by

$$SPE = \sqrt{\sum_{i=1}^{n}((y_i - \hat{y}_i)/\hat{y}_i)^2/(n-p)} \qquad (22)$$

where:

y_i is the ith observation,
$\hat{y}_i$ is the predicted value of the ith observation,
p is the total number of the estimated coefficients, and
n is the sample size.

We will examine two recurring cost-dependent CERs in USCM8: the IA&T first unit cost (T1) CER and program-level (SEPM) total recurring cost CER.

USCM8 IA&T T1 CER. In the USCM8 online publication, the IA&T T1 CER (in FY00$K) is given by

Y = **0.124** *(Space Vehicle First Unit Cost in FY00$K) (23)

where Space Vehicle (SV) First Unit Cost = Spacecraft + Communications total first unit costs.

According to the alternative method, the IA&T T1 CER is generated using the sum of the estimated subsystem T1 costs as the new driver. The updated CER, based upon the "estimated" driver variable, is given by

Y = **0.134** * (Estimated Space Vehicle First Unit Cost in FY00$K) (24)

Note that the estimated space vehicle first unit cost is derived by summing all the subsystem-level T1 CERs as listed below.

Estimated SV First Unit Cost

= Structure&Thermal + ADCS + EPS + TTC + COMM + Propulsion Estimated T1 Costs

$$= 7.556*\text{Structure\&Thermal\_Wt} + 379.872*\text{ADCS\_Wt}^{(0.593)} + 10.811*\text{EPS\_Wt} + 441.546*\text{TTC\_Wt}^{(.4914)}*1.13^{\text{GEO}} + 77.3*\text{COMM\_Wt} + 3931 \quad (25)$$

(An average cost CER is used for the propulsion subsystem; the factor 77.3 for the communication subsystem T1 CER is found in the USCM8 estimating guidance section.)

Comparing Equation 24 to Equation 23 yields two major findings:

- The updated factor by the alternative method is 0.134, which is about eight percent higher than the old factor (0.124).
- The updated SPE is 32.4%, which is in fact five percent less than the old SPE (34%).

USCM8 SEPM CER. According to the USCM8 online publication, the SEPM total recurring cost in FY00$K (for communication satellites) is given by

Y = **0.234** *(Space Vehicle Total Recurring Cost in FY00$K) (26)

where Space Vehicle (SV) Total Recurring Cost = Spacecraft + Communications + IA&T total recurring costs.

By the same rationale given above, Equation 26 should be updated using the sum of the estimated subsystem recurring costs (i.e., the CER-predicted space vehicle total recurring cost) as the new driver. The updated CER, based upon the "estimated" driver, is given by

Y = **0.355** * (Estimated Space Vehicle Total Recurring Cost in FY00$K) (27)

(Note that the alternative method would generate a factor of 0.395 for the SEPM CER if we choose the overall factor 63.016, based on 25 programs, instead of 77.3, which was based on a smaller subset, for the communication subsystem T1 CER.)

The following results are noted when comparing Equation 27 with Equation 26:

- The updated factor by the alternative method is 0.355, which is 52% greater than the old factor (0.234).
- The updated SPE by the alternative method is **63%**, which is more than four times greater than the old SPE (12%).

The above results are more unsatisfactory than those calculated for updating the IA&T CERs. Note also that this alternative SEPM CER (Equation 27) is based upon all the "estimated" subsystem recurring costs while the IA&T portion (in the driver) is derived from Equation 24,

which is also based upon the "estimated" subsystem costs. This can be a little cumbersome and confusing.

Y = 0.355 *(Estimated Space Vehicle Total Recurring Cost in FY00$K) (28)

= 0.355 *(ST+ADCS+EPS+TTC +PROP+COMM+**IA&T**)←all estimated subsystem rec costs

= 0.355(ST+ADCS+EPS+TTC+PROP+COMM+.134(ST+ADCS+EPS+TTC+PROP+COMM))

(Note: ST stands for the Structure and Thermal subsystems combined.) To use these alternative equations repeatedly and correctly, we should ensure that (1) the IA&T CER is an alternative CER and all its drivers are the estimated costs, not the actual costs and (2) the IA&T CER is developed before the SEPM CER. Therefore, CERs may not be developed in parallel when deriving cost-dependent CERs. Further, since the cost driver consists of all estimated subsystem recurring costs, we should use the equation below to estimate the total recurring cost for each subsystem based upon the 95% cost improvement curve (CIC) slope:

$$\text{Subsystem Recurring Cost} = (\text{Subsystem T1 Estimate}) * ((PQ+LQ)^{(0.926)} - PQ^{(0.926)}) \quad (29)$$

where LQ is the total space vehicle quantity and PQ is the prior quantity for each satellite.

LIMITATIONS AND CONCERNS ABOUT THE ALTERNATIVE METHOD

Given two significantly different CERs (Equations 26 and 27), which one should we use to predict the space vehicle program-level recurring cost? Further, if we were to choose Equation 27 (derived by the alternative approach) to estimate costs, should we also use its respective SPE for cost uncertainty analysis?

Based upon Equation 15, we can claim that there is a theoretical basis for using the alternative approach (i.e., Method 2a) to develop cost-dependent CERs. However, this mathematical proof does **not** hold when (1) there are **voids** in the data set, (2) the CERs are not linear, or (3) the cost-driver (e.g., PMP) is estimated by several CERs or multiple drivers.

Upon further analyses, there is a potential problem with using inconsistent data sets when developing cost-dependent CERs, especially when the cost-dependent CERs are used repeatedly, such as in Equation 28. Note that there were seven programs used to build this SEPM recurring cost CER (Programs B and C were combined):

A, B, C, D, E, F, and G

Here are the concerns (in order of decreasing impact):

- Program G was **not** used in any of the subsystem-level recurring cost CERs because of the programmatic issues.
- Program D was only used in the IA&T T1 CER (as well as the SEPM recurring cost CER) but not in any of the subsystem-level recurring cost CERs due to issues of the data.
- Program E was only used in the COMM, ADCS, and IA&T recurring cost CERs; it was not included in the Structure, Thermal, EPS, or TT&C subsystem T1 CER.
- Program F was not used in the IA&T T1 CER.

Consequently, using the CER-predicted cost instead of the actual cost can generate **very poor** space vehicle (SV) cost estimates for Programs D and G because these two programs were not part of the subsystem-level CERs. In fact, both were identified as **outliers** based upon engineering logic and had been excluded from the curve-fitting process. For instance, the SV total recurring cost for Program G was about 290% of the CER-predicted cost and Program D's actual SV total recurring cost was about 170% of the CER-predicted cost. These two poorly estimated SV recurring costs were then used to derive the SEPM factor for Equation 27. As a result, the "estimated" costs of these two programs caused the regression model to generate a SEPM factor higher than it should be and a much worse SPE under the alternative approach.

It may seem logical to remove Programs D and G from fitting this SEPM total recurring cost CER because they were not used in any of the subsystem-level CERs. However, it is hard to remove them because their IA&T and SEPM costs are legitimate. Also, the degrees of freedom (DF) will be only three if we eliminate Programs D and G. Note that "X" in Table 1 below indicates the program was used in fitting the CER.

Table 1: List of CERs used above and SEPM Data Points

WBS	USCM8 CERs	A	B/C	D	E	F	G
Space Vehicle less SEPM (**SV**)							
Spacecraft (**SC**)							
Structure & Thermal (ST)	7.566 * ST_Wt	X	X			X	
ADCS	365.332 * ADCS_Wt ^ 0.593	X	X		X	X	
EPS	10.811 * EPS_Wt	X	X			X	
TT&C	441.546 * TT&C_Wt ^0.491 * 1.130 ^ GEO	X	X			X	
Propulsion	3,931	X	X			X	
COMM Payload	77.3 * COMM_Wt	X	X		X	X	
IA&T	0.124 * (SC_T1+COMM_T1)	X	X	X	X		
Program Level (SEPM)	0.234 * SV_REC	X	X	X		X	X

CONCLUSIONS

Statistical validation of using the alternative method to derive cost-dependent CERs. There has been a concern about using cost-dependent CERs (derived by the actual cost) in cost uncertainty analysis. For example, if the PMP cost is estimated by the hardware design parameters and the SEPM cost is a function of the PMP cost, the SEPM cost can be estimated only after the PMP cost is estimated by the hardware-based CER. Since we use the estimated PMP cost (rather than the actual cost) in the SEPM equation when conducting cost uncertainty analysis, we should also use the estimated PMP cost to develop the SEPM CER for consistency. Otherwise, the SEPM estimate and the variance for SEPM will be inaccurate. This appears to be a legitimate concern, so Reference 1 suggests an alternative approach of using the "estimated" cost rather than the actual cost to derive cost-dependent CERs.

This IR&D paper offers a statistical validation of using the alternative approach (Method 2a) to derive cost-dependent CERs for straightforward linear models (no WBS involvement). It shows that the alternative method (Method 2a) is the same as the direct method (Method 2b) for

linear models under WLS. (OLS is a special case here.) In fact, the proof to validate Method 2a is true not only for cost-dependent CERs, but also for any CERs whose drivers can be estimated by another linear model (see the Statistical Derivation Section for details). In addition, this statistical validation holds for simple linear models whether or not the error term is assumed to be additive or multiplicative such as MUPE. Therefore, we conclude that the alternative method is valid for straightforward linear models under WLS. (However, this proof does **not** hold if there are voids in the data set or the equation form is not linear.)

Comparisons of Standard Percent Errors. Mathematically, we cannot conclude which method will generate a smaller standard percent error when comparing the commonly-used method (Method 1) with the alternative method (Method 2a). The goal of this paper is to compare the cost-dependent CERs generated by both methods (Methods 1 and 2a) using the USCM8 examples to see whether there are any significant differences. We developed two cost-dependent CERs using the alternative approach for USCM8: the IA&T T1 CER and SEPM total recurring cost CER. The SPE for the USCM8 IA&T T1 CER using the alternative method is smaller than the published CER using the current method (32% vs. 34%). However, the SPE result is quite the opposite for the SEPM total recurring cost CER: 73% for the alternative method vs. 12.6% for the current method. We do not have a conclusion as to which method will deliver a smaller standard percent error.

Beware of the pitfalls of applying the alternative method to real examples. We noticed a potential problem with using inconsistent data sets when deriving the cost-dependent CERs in USCM8 under the alternative method. Two programs (denoted by Programs D and G) were identified to be outliers (due to data issues) and had been excluded from the curve-fitting process for deriving the subsystem-level CERs, but they were included in the SEPM total recurring cost CER. Consequently, including these two irrelevant "estimated" costs caused the regression model to generate a SEPM factor much higher than it should be and a very poor SPE under the alternative approach. Clearly, the actual and predicted space vehicle recurring costs for Programs D and G are not from the same population. We do not recommend applying the alternative method to complicated cases, especially to inconsistent or nested CERs such as Equation 28 (the USCM8 SEPM T1 CER). This is because the alternative method is tedious and is subject to errors in these cases as demonstrated in the sections above.

Avoid using cost as an independent variable in a CER. An independent variable based upon cost, whether or not it is actual or predicted, is not an ordinary independent variable. A cost driver is always subject to errors. In other words, a cost independent variable cannot be observed without error, which violates the basic assumption of regression analysis. If the cost is further estimated by CERs, analogies, or even expert opinions, then we may become more uncertain about the uncertainties associated with the cost driver, which is certainly less desirable. In conclusion, the rationale of using the alternative method for deriving cost-dependent CERs may be logical, but the actual implementation may not be practical. It can be tedious, confusing, and prone to error, especially when using the alternative method repeatedly for certain CERs. On the other hand, the current method (based upon the actual cost) is also undesirable because (1) it can be problematic for uncertainty analysis as discussed in Reference 1, (2) there are always more uncertainties inherited in the cost-dependent drivers than the technical drivers, and (3) the degrees of freedom for cost-dependent CERs can be over-estimated, which may affect cost uncertainty analysis. The bottom line: develop hardware design-based CERs whenever possible and avoid using cost-dependent CERs.

REFERENCES

1. Covert, R. P. and Anderson, T. P., "Regression of Cost Dependent CERs," the Aerospace Corporation, Space Systems Cost Analysis Group (SSCAG) meeting, February 2002.

2. Nguyen, P., Lozzi, N., et al., "Unmanned Space Vehicle Cost Model, Eighth Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Angeles AFB, CA, October 2001.

APPENDIX A

It is more convenient to write the regression model and the corresponding objective function using the vector and matrix notations than to use the actual equation directly. This way, not only can the least squares solutions be derived easily, but their statistical properties can be obtained and discussed much more directly. Therefore, we will briefly introduce least squares regression analysis using matrix notations.

Let us consider a regression model where a dependent variable Y can be estimated from k independent variables; namely, $X_1, X_2, \ldots, X_k$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad \text{for } i = 1,\ldots,n$$

The model can be rewritten in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{30}$$

where:

Y is the n by 1 vector of observations (i.e., the dependent variable),
X is the n by (k+1) design matrix, which consists of the independent variables,
$\boldsymbol{\beta}$ is the (k+1) by 1 vector of unknown coefficients, i.e., $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$,
$\boldsymbol{\varepsilon}$ is the n-by-1 vector of error terms with a variance matrix **V**, i.e., $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{V}\sigma^2$,
V is an n-by-n diagonal matrix with the non-negative value v_i in the diagonals (for i = 1, …, n) and zeros elsewhere, and
n is the sample size.

(Note that the variance matrix **V** is not necessarily an identity matrix **I**, so the MUPE CER is included in the analysis.)

To be more specific, the design matrix is given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & x_{2k} \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{j} & \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_k \end{pmatrix}$$

where

$\mathbf{x}_i$ = $(x_{1i}, x_{2i}, \ldots, x_{ni})^t$, a column vector for the ith independent variable
$\mathbf{j}$ = $(1, 1, \ldots, 1)^t$, a 1 by n column vector of all 1's

(Note that the superscript t stands for the transpose of a vector or a matrix.)

If we apply the concept of using a centered designed matrix, the design matrix X can also be expressed as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & . & x_{1k} - \bar{x}_k \\ 1 & x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & . & x_{2k} - \bar{x}_k \\ . & . & . & . & . \\ 1 & x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_{21} & . & x_{nk} - \bar{x}_k \end{pmatrix} = \begin{pmatrix} \mathbf{j} & \mathbf{x}_1 - \bar{\mathbf{x}}_1 & \mathbf{x}_2 - \bar{\mathbf{x}}_2 & \ldots & \mathbf{x}_k - \bar{\mathbf{x}}_k \end{pmatrix}$$

(The **j** vector is not necessary if the y variable is also centered.) Our goal is to find a set of constants ($\beta_0, \beta_1, \ldots, \beta_k$), which minimizes the objective function, i.e., the sum of squared errors

$$F = \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})\right)^2 = (\mathbf{Y} - \mathbf{X\beta})'\mathbf{W}(\mathbf{Y} - \mathbf{X\beta})$$

where **W** is an n by n weighting matrix as given below:

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ . & . & \ldots & 0 \\ 0 & 0 & \ldots & w_n \end{pmatrix} = \begin{pmatrix} 1/v_1 & 0 & \ldots & 0 \\ 0 & 1/v_2 & \ldots & 0 \\ . & . & \ldots & 0 \\ 0 & 0 & \ldots & 1/v_n \end{pmatrix} = \mathbf{V}^{-1} \tag{31}$$

By taking the partial derivative of F with respect to **β** and setting it to zero, the ordinary least squares (OLS) solution of **β** is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'WX})^{-1}\mathbf{X'WY} \tag{32}$$

Consequently, its respective predicted value is given by

$$\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WY} \tag{33}$$

If the error term (ε_i) is further assumed to follow a normal distribution, with a mean of 0 and variance $v_i\sigma^2$ for i = 1, …, n, i.e., $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V}\sigma^2)$, then statistical inferences can be made based upon regression analysis.

For a one-independent variable case, where $Y = \beta_0 + \beta_1 X + \varepsilon$, the CER-predicted value (i.e., Equation 33) can be expressed as

$$\hat{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'WY}$$

$$= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix} \left(\begin{pmatrix} 1 & 1 & . & 1 \\ x_1 & x_2 & . & x_n \end{pmatrix} \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ . & . & \ldots & 0 \\ 0 & 0 & \ldots & w_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & . & 1 \\ x_1 & x_2 & . & x_n \end{pmatrix} \begin{pmatrix} w_1 y_1 \\ w_2 y_2 \\ . \\ w_n y_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix} = \frac{1}{SS_{wXX}} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \sum w_i x_i^2 / \sum w_i & -\bar{x}_w \\ -\bar{x}_w & 1 \end{pmatrix} \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix}$$

$$= \frac{1}{SS_{wXX}} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \bar{y}_w \sum w_i x_i^2 - \bar{x}_w \sum w_i x_i y_i \\ \sum w_i x_i y_i - \bar{x}_w \bar{y}_w \sum w_i \end{pmatrix} = \left((\bar{y}_w + \frac{SS_{wXY}}{SS_{wXX}}(x_i - \bar{x}_w)) \right)_{nx1} \tag{34}$$