Presented at the 2012 SCEA/ISPA Joint Annual Conference and Training Workshop - www.iceaaonline.com

Joining Effort and Duration in a Probabilistic Method for Predicting Software Cost and Schedule

MICHAEL A. ROSS

r2Estimating, LLC, Lakeside, Montana USA

Abstract

*This paper describes a **data-driven** method for estimating the cost and schedule of developing software items. This method correlates the estimates of cost and schedule such that constraining (reducing or increasing) the budget will impact the estimated schedule and constraining (relaxing or compressing) the schedule will impact the estimated cost. This method provides estimates of cost and schedule that are probabilistic (i.e., provide a range of possible outcomes with associated probabilities of attainment); a capability that is essential to analyzing the impact that **affordability and budget constraints** have on program cost and schedule and their associated risks. This method incorporates a software development Cost and Duration Estimating Relationship (CDER) system of two equations that can be easily calibrated to any historical data set that includes the size, effort, and duration of several completed software items. The paper includes a practical example of developing a software development CDER from an example data set. It also includes various forms of the equations that can be used to perform analyses such as joint confidence level (JCL), minimum time solution, minimum effort solution, compressed schedule solution, time as an independent variable, effort as an independent variable, cost as an independent variable, and average staff as an independent variable to name a few.*

Introduction

This paper describes the ***r2 Software Estimating Framework (r2SEF)***, a *data-driven* method for joining, correlating, and calibrating software Cost Estimating Relationships (CERs) and corresponding Schedule Estimating Relationships (SERs) to construct Cost and Duration Estimating Relationship (CDER) systems of two equations. These equations are probabilistic (i.e., provide a range of possible outcomes with associated probability of attainment) and can be calibrated to any historical data set that includes the size, effort, and duration of several completed Software Items (SIs). The paper then describes a simple method for determining joint and conditional probabilities (confidence levels) of estimated cost and schedule based on these relationships and a specific example historical data set. The paper concludes with equation forms that can be used to perform a variety of analyses such as Joint Confidence Level (JCL), minimum time solution, minimum effort solution, compressed schedule solution, time as an independent variable, effort (cost) as an independent variable, and average staff as an independent variable. All of these solutions can be presented as distributions (S-curves) or as single values with an associated probability of attainment (confidence level).

Software CDER Basics

Generalized Software CDER System of Equations

A Software CDER (Ross, 2008) is a system of two fundamental equations that have the following general forms[1]:

- ***Work Relation***

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\textbf{\textit{Resources Applied} } \leftrightarrow \textbf{\textit{Work Performed}}$$

$$\textbf{\textit{Effort}}^{\textit{Effort Nonlinearity}} \times \textbf{\textit{Duration}}^{\textit{Duration Nonlinearity}} = \textbf{\textit{Difficulty}} \times \textbf{\textit{Size}}^{\textit{Size Nonlinearity}} \quad (1)$$

$$\boldsymbol{E}^{\alpha_E}\boldsymbol{T}^{\alpha_T} = \boldsymbol{DS}^{\alpha_S}$$

where

E ≡ Random variable representing the range and distribution of possible outcomes of ***Effort*** (labor); typically measured in units of person-months or person-hours. Effort combined with a given labor rate yields cost.

T ≡ Random variable representing the range and distribution of possible outcomes of ***duration*** (Time); typically measured in units of calendar months. Duration combined with a start date or a finish date yields schedule.

D ≡ Random variable representing the range and distribution of possible values of ***Difficulty***; a unitless factor that scales Size units to Work units. The Difficulty factor quantifies the amount of work necessary to develop one unit of non-linear software size.[2] Each and every calibrated CDER has a unique range and distribution of possible difficulty values.

S ≡ Random variable representing the range and distribution of possible outcomes of software ***Size***; typically measured in some type of software sizing unit that directly relates to the amount of work that must be done to develop the software. Examples of such units are Effective Source Lines of Code (ESLOC) and Unadjusted Function Points (UFP).

$\alpha_E, \alpha_T, \alpha_S$ ≡ ***Effort nonlinearity*** (Greek alpha subscript E), ***duration*** (Time) ***nonlinearity*** (Greek alpha subscript T), and ***size nonlinearity*** (Greek alpha subscript S). Each and every calibrated CDER has a unique set of constant values for these exponents.

- ***Intensity Relation***

$$\textbf{\textit{Cost}} \leftrightarrow \textbf{\textit{Schedule}}$$

$$\textbf{\textit{Effort}} = \textbf{\textit{Intensity}} \times \textbf{\textit{Duration}}^{\textit{Effort}-\textit{Duration Tradeoff Nonlinearity}} \quad (2)$$

$$\boldsymbol{E} = \boldsymbol{IT}^{\alpha_{ET}}$$

where

I ≡ Random variable representing the range and distribution of possible outcomes of ***Intensity***; a unitless factor that correlates nonlinear duration units to units of effort irrespective of size and difficulty. The intensity factor quantifies the magnitude of effort with respect to the magnitude of duration. A project that expends a relatively large amount of effort over a relatively short period of time (i.e., requires a relatively large team of people) is said to have high intensity whereas a project that expends a relatively small amount of effort over a relatively long period of time (i.e., requires a relatively small team of people) is said to have low intensity. Compressing the schedule (reducing the duration) for developing an SI by increasing the team size in or-

©2012 ***r2Estimating, LLC***. All rights reserved.

der to meet some schedule goal or constraint increases its intensity. For most SI developments, increasing the project's intensity tends to increase the effort (cost) necessary to complete the project. This tendency implies that increasing a project's intensity tends to reduce its productivity (its size per unit of effort). Additionally, for most SI developments, increasing the intensity also tends to increase the total number of defects that the project team must find and fix (Ross, 2008).[3] Each and every calibrated CDER has a unique range and distribution of possible intensity values.

α_{ET} ≡ ***Effort-duration tradeoff nonlinearity*** (Greek alpha subscript ET). Each and every calibrated CDER has a unique constant value for its time-effort tradeoff nonlinearity.

Constructing a Calibrated Software CDER from Historical Data

How do we calibrate the generalized Equations (1) and (2) above to model the behavior of a specific set of developed SI historical data in order to estimate the development of future SIs? By calibration we mean determining unique constant values for the exponents α_E, α_T, α_S, and α_{ET} and unique ranges and distributions for the random variables $\boldsymbol{D}$ and $\boldsymbol{I}$. Obviously, any data-driven methodology must start with some data. In this case we require a list of completed SI actual size values $\mathbf{S}$ with lists of corresponding actual effort values $\mathbf{E}$ and actual duration (time) values $\mathbf{T}$; these three lists we collectively refer to as an historical data set. A primary objective when organizing historical data sets is to maximize the similarity between included SIs while at the same time maximizing the number of SIs (data points) included in the set. This objective, while often difficult to achieve, is nonetheless intended to both reduce the amount of variability and to increase the statistical significance of the relationships that are derived from the historical data set.

It is possible to construct a software CDER system of equations that is specifically calibrated to a particular historical data set of sizes, efforts, and durations of completed software items by performing the three power function regressions described below. For each regression, the paper first describes its associated assumptions, theory, and mathematics. These descriptions are followed by an example using the data set contained in the Appendix, Table 1.

First Regression: Effort as a Power Function of Size

Theory

The first regression deals with the notion of the cost ***inefficiency*** associated with developing a particular SI. We begin by assuming that the amount of effort E expended to develop an SI (and hence its cost) is proportional to some function its size S, that function being reasonably modeled as a power function of the form $\boldsymbol{f}(x) = x^a$ (Ross, 2008). This assumption can be described mathematically as:

$$E \propto \boldsymbol{f}_{\boldsymbol{Inefficiency}}(S) \rightarrow E = \omega\, \boldsymbol{f}_{\boldsymbol{Inefficiency}}(S) \rightarrow E = \omega S^{a_1} \qquad (3)^4$$

©2012 ***r2Estimating, LLC***. All rights reserved.

The scale factor ω (Greek omega) of size quantifies the cost inefficiency present in the software development process. Inefficiency within this context can be thought of as the inverse notion of productivity. For a given size, greater inefficiency results in more effort (and hence more cost) while lower inefficiency results in less effort (and hence less cost). The exponent of size a_1 quantifies the nonlinearity associated with cost inefficiency. An a_1 value of greater than 1 implies proportionally more effort (cost) as the size goes up (a diseconomy of scale) while an a_1 value of less than 1 implies proportionally less effort as the size goes up (an economy of scale).

The first step in the calibration process is to determine the data-set-specific constant value of the exponent a_1. This can be done by performing the power form of an appropriate regression function $\boldsymbol{r}_{<\textbf{regression method}>}(\mathbf{Y},\mathbf{X})$ where $\mathbf{Y}$ is a list of values representing the regressand (dependent variable; in this case effort $\mathbf{E}$) and where $\mathbf{X}$ is a corresponding list of values representing the regressor (independent variable; in this case size $\mathbf{S}$). Regression methods that are suitable for this purpose include *Log-transformed Ordinary Least Squares* (Log OLS) regression (USAF, 2007 pp. 52,53), *Minimum Unbiased Percentage Error* (MUPE) regression (USAF, 2007 pp. 56-58), and *Zero Percent Bias – Minimum Percentage Error* (ZMPE) regression (USAF, 2007 pp. 58,59). Each of these regression methods yields a vector $\langle a, \bar{b} \rangle$ where a represents the regressed power function's exponent and where $\bar{b}$ represents the mean value of the regressed power function's scale factor. The first regression can therefore be described as:

$$\left[\langle a_1, \bar{\omega} \rangle_{<regression\ method>} = \boldsymbol{r}_{<\textbf{regression method}>}(\mathbf{E},\mathbf{S})\right]_{<\textbf{historical data set name}>} \tag{4}$$

The second step in the calibration process is to compute a list of inefficiencies $\boldsymbol{\omega}$ that corresponds to the $\mathbf{S}$ and $\mathbf{E}$ lists we already have as part of the historical data set. We can rearrange the factors in Equation (3) to get $\omega = E/S^{a_1}$; therefore

$$\left[\boldsymbol{\omega} = \frac{\mathbf{E}}{\mathbf{S}^{a_1}}\right]_{<\textbf{historical data set name}>} \rightarrow \left|\boldsymbol{\omega}_{[i]} = \omega_i = E_i / S_i^{a_1}\right|_{i=1}^{N} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_N \end{bmatrix} = \begin{bmatrix} E_1/S_1^{a_1} \\ E_2/S_2^{a_1} \\ \vdots \\ E_N/S_N^{a_1} \end{bmatrix} \tag{5}$$

where

$\boldsymbol{\omega}$ ≡ SI-ordered list of Inefficiency values; one for each SI in the data set

$\mathbf{E}$ ≡ SI-ordered list of Effort values; one for each SI in the data set

$\mathbf{S}$ ≡ SI-ordered list of Size values; one for each SI in the data set

a_1 ≡ The exponent value yielded by performing the regression in Equation (4)

N ≡ Number of SIs in the data set

©2012 ***r2Estimating, LLC***. All rights reserved.

Example

In order to demonstrate a practical example of performing this first regression (first and second steps above) we use the effort list **E** and size list **S** from the data set contained in the Appendix, Table 1 which is a Military Ground C/C++ stratification of the Aerospace Corporation (2004) Software Cost and Productivity Model database. We must first choose a particular regression method and stick with it throughout the calibration process. The author, while favoring both MUPE and ZMPE, chooses to use ZMPE in this case because it eliminates bias as part of optimization criteria (not the case with Log OLS) and because of its ease of implementation in Microsoft Excel with the Solver add-in (the MUPE optimization process is more complex).[5] The regression can be stated mathematically as

$$\left[\langle a_1, \bar{\omega}\rangle_{ZMPE} = \mathbf{r}_{\boldsymbol{ZMPE}}(\mathbf{E},\mathbf{S})\right]_{\textbf{Aerospace 2004: Military Ground C/C++}} \tag{6}$$

and results in $a_1 = 1.0373$. The subsequently-computed inefficiency list $\boldsymbol{\omega}$ is contained in the Appendix, Table 2. A graphic representation of the regression is shown in Figure 1 below.

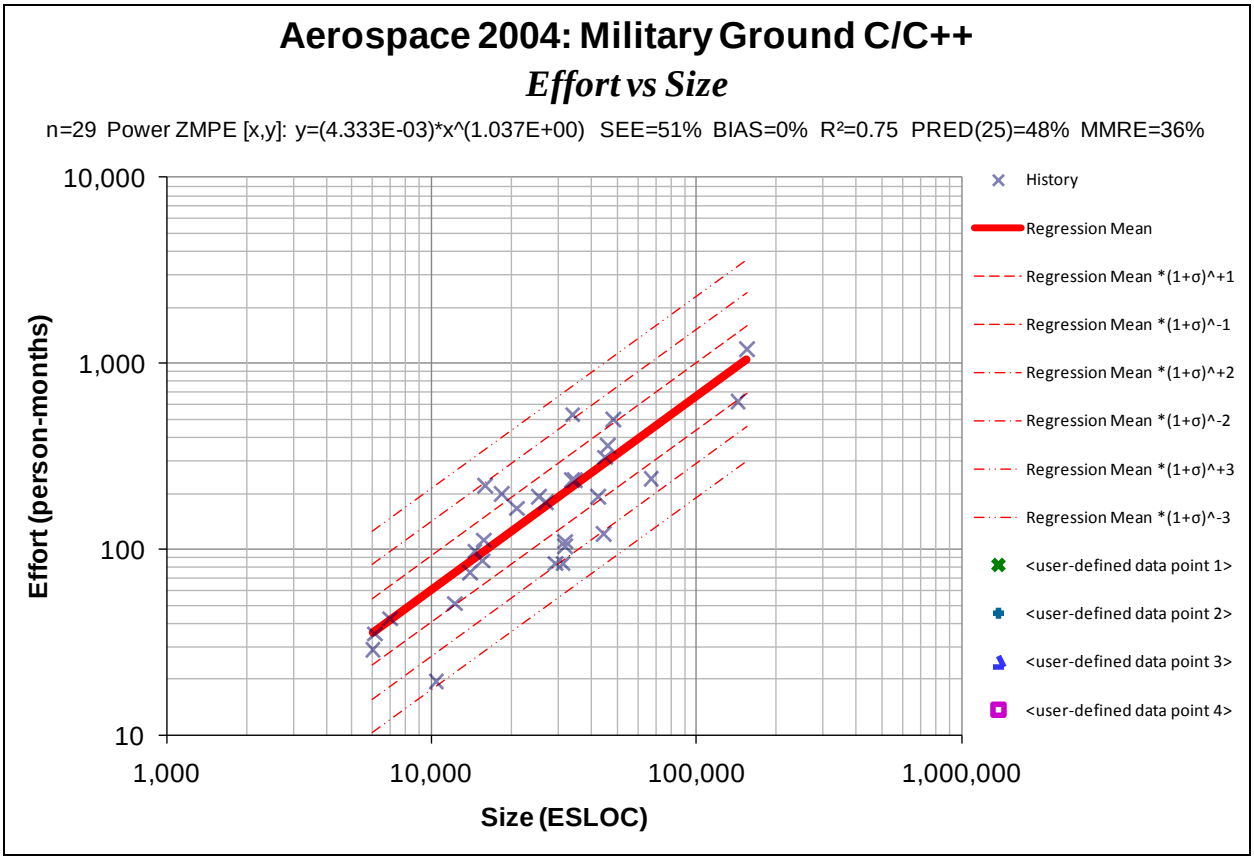

Figure 1: *Power regression of effort vs. size using the ZMPE regression method*

©2012 ***r2Estimating, LLC***. All rights reserved.

Second Regression: Effort as a Power Function of Duration

Theory

The second regression deals with the notion of the ***intensity*** associated with a given SI development. We continue the calibration process by assuming that the amount of effort expended to develop an SI (and hence its cost) is proportional to some function of its duration (its schedule), that function also being reasonably modeled as a power function of the form $\boldsymbol{f}(x) = x^a$ (Ross, 2008).

$$E \propto \boldsymbol{f}_{\boldsymbol{Intensity}}(T) \rightarrow E = I\ \boldsymbol{f}_{\boldsymbol{Intensity}}(T) \rightarrow E = IT^{a_2} \tag{7}$$

The scale factor of duration I quantifies the intensity present in the software development process, the concept of intensity having already been described earlier in this paper as quantifying the magnitude of effort with respect to the magnitude of duration.

The third and fourth steps in the calibration process are essentially the same as the first and second steps except this time they are applied to the effort list $\mathbf{E}$ and the duration list $\mathbf{T}$. The third step yields a value for the exponent a_2 and the fourth step yields a list of project intensities $\mathbf{I}$.

Example

The third step above can be applied to our example by using the effort list and the duration list from the same example data set (Appendix, Table 1) which yields $a_2 = 2.2020$. The fourth step results in the subsequently-computed intensity list $\mathbf{I}$ shown in the Appendix, Table 2. A graphic representation of the regression is shown in Figure 2 below.

©2012 ***r2Estimating, LLC***. All rights reserved.

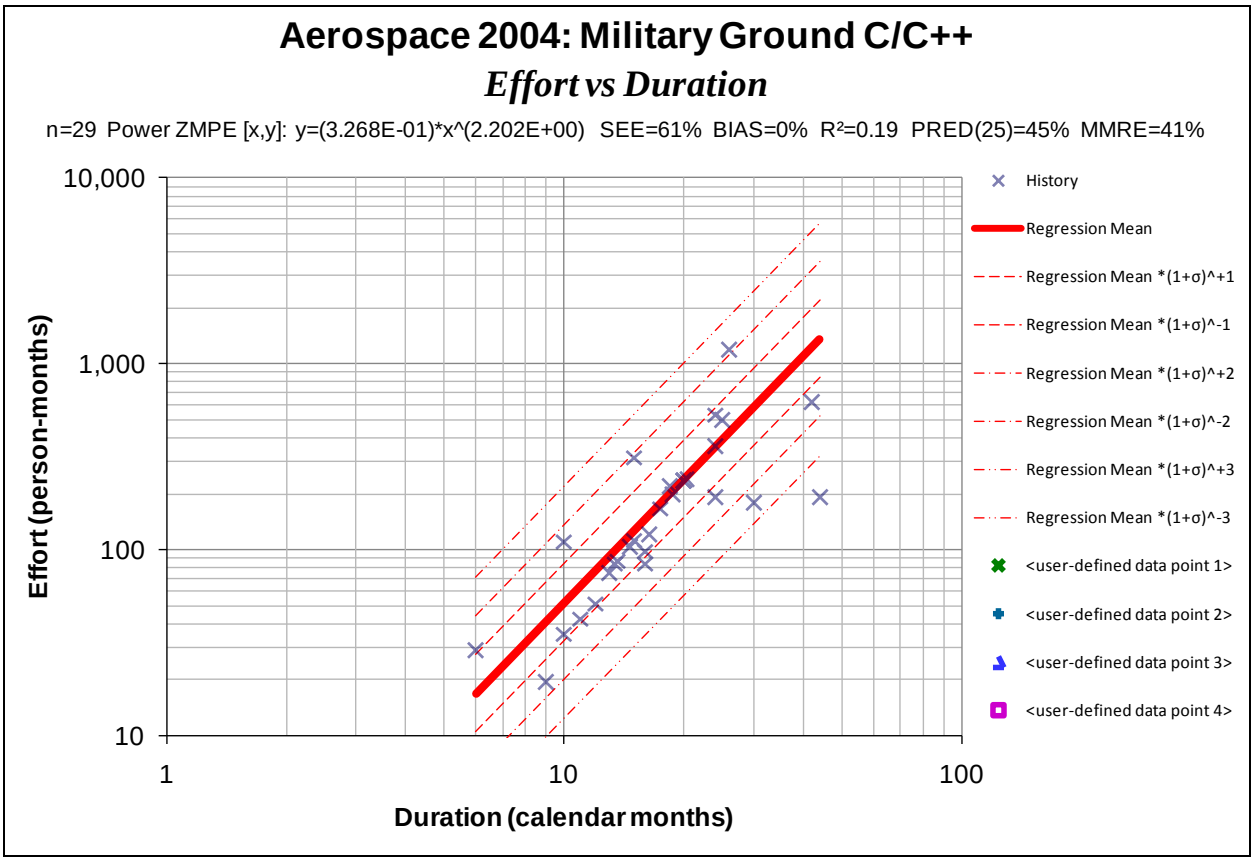

Figure 2: *Power regression of effort vs. duration using the ZMPE regression method*

Third Regression: Inefficiency as a Power Function of Intensity

Theory

The third regression deals with the notion of the ***difficulty*** associated with developing a particular SI with given inefficiency and intensity. We complete the calibration process by assuming that the amount of inefficiency present in the development of some SI is proportional to some function its intensity, that function again being reasonably modeled as a power function of the form $\boldsymbol{f}(x) = x^a$.

$$\omega \propto \boldsymbol{f}_{\boldsymbol{Difficulty}}(I) \rightarrow E = D\,\boldsymbol{f}_{\boldsymbol{Difficulty}}(I) \rightarrow \omega = DI^{a_3} \tag{8}$$

This proportionality assumption effectively correlates the notion of inefficiency with the notion of intensity and therefore ensures that we end up with a CDER where the estimated effort and estimated duration are correlated (positively or negatively) in a way that is consistent with the supporting historical data. Choosing a power function to model this correlation allows for the possibility that this correlation is nonlinear for a particular data set.

The scale factor D in Equation (8) quantifies the difficulty present in developing the given SI, the concept of difficulty having already been described earlier in this paper as being the

©2012 ***r2Estimating, LLC***. All rights reserved.

resources necessary (nonlinear effort and nonlinear duration) to develop one unit of non-linear software size.

The fifth and sixth steps in the calibration process are essentially the same as the first and second steps except this time they are applied to the previously-computed inefficiency list $\boldsymbol{\omega}$ and the previously-computed intensity list $\mathbf{I}$. The fifth step yields a value for the exponent a_3 and the sixth step yields a list of project difficulties $\mathbf{D}$.

Example

The fifth and sixth steps above can be applied to our example by using the inefficiency list and its corresponding intensity list (Appendix, Table 2), these lists having been computed in the previously-described calibration steps from the data set in the Appendix, Table 1. The resulting exponent is $a_3 = 0.1932$. The resulting difficulty list $\mathbf{D}$ is shown in the Appendix, Table 2. A graphic representation of the regression is shown in Figure 3 below.

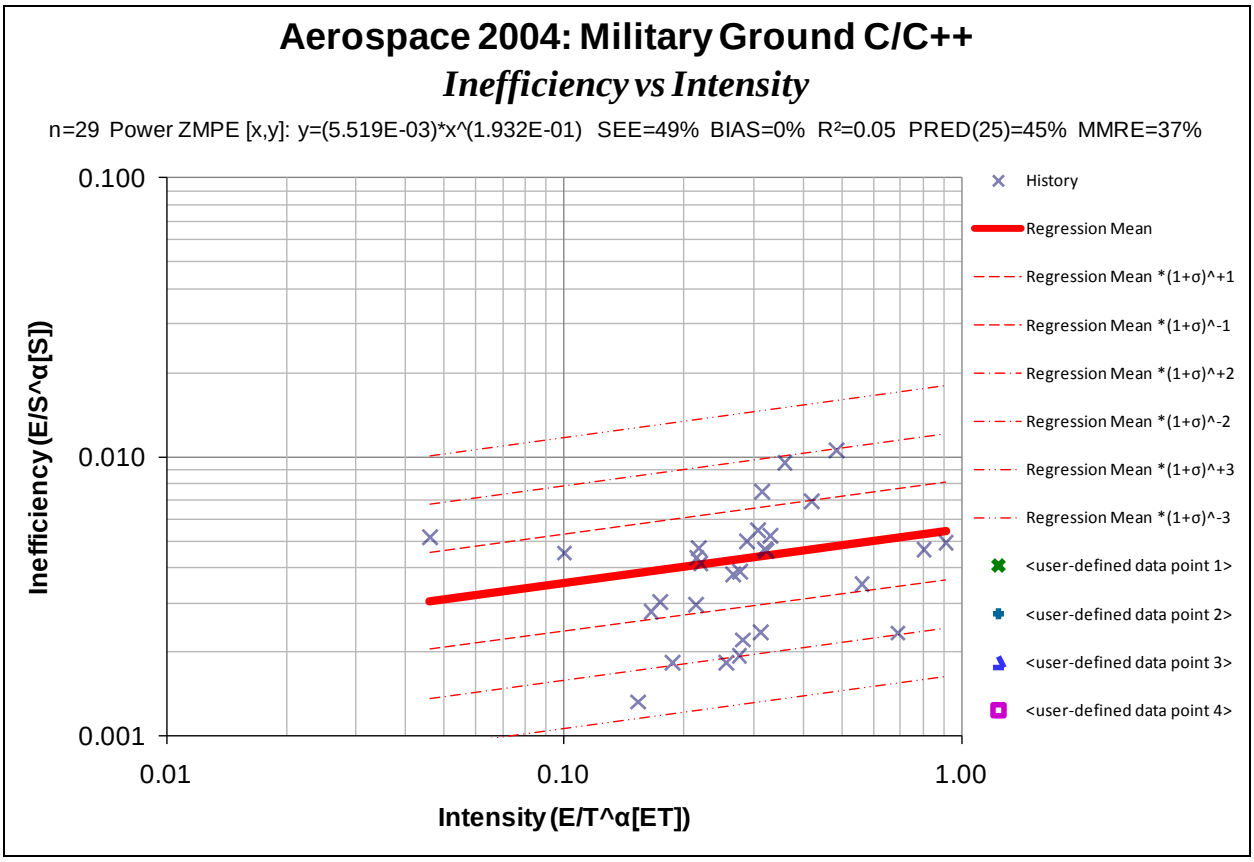

Figure 3: *Power regression of inefficiency vs. intensity using the ZMPE regression method*

©2012 ***r2Estimating, LLC***. All rights reserved.

Consolidating the Exponents

Theory

The seventh step of the calibration process produces a calibrated version of general Equation (1) and the eight step produces a calibrated version of general Equation (2).

With respect to the seventh step, we need calibrated values for the exponents α_E, α_T, and α_S. We use the following algebra to relate a_1, a_2, and a_3 from Steps 1, 3, and 5 above to α_E, α_T, and α_S in general Equation (1):

Solving for ω in Equation (3) yields

$$E = \omega S^{a_1} \rightarrow \omega = \frac{E}{S^{a_1}} \tag{9}$$

Solving for I in Equation (7) yields

$$E = IT^{a_2} \rightarrow I = \frac{E}{T^{a_2}} \tag{10}$$

Substituting ω and I in Equation (8) with the equivalent of ω in Equation (9) and the equivalent of I in Equation (10) yields

$$\frac{E}{S^{a_1}} = D\left(\frac{E}{T^{a_2}}\right)^{a_3} \rightarrow \frac{E}{S^{a_1}} = D\left(\frac{E^{a_3}}{T^{a_2 a_3}}\right) \rightarrow \frac{E}{E^{a_3}} T^{a_2 a_3} = DS^{a_1} \rightarrow E^{1-a_3} T^{a_2 a_3} = DS^{a_1} \tag{11}$$

Next we let

$$\alpha_E \equiv 1 - a_3 \text{ and } \alpha_T \equiv a_2 a_3, \text{ and } \alpha_S \equiv a_1 \tag{12}$$

and substitute each for their equivalents in Equation (11) to get

$$E^{\alpha_E} T^{\alpha_T} = DS^{\alpha_S} \tag{13}$$

Note that Equation (13) becomes a calibrated version of generalized Equation (1) when we perform the above-described substitutions with the calibrated values for a_1, a_2, and a_3 from the first, third, and fifth steps of the calibration process and use the calibrated list $\mathbf{D}$ from the sixth step of the regression process to get

$$\mathbf{E}^{\alpha_E} \mathbf{T}^{\alpha_T} = \mathbf{D}\mathbf{S}^{\alpha_S} \tag{14}$$

Regarding the eighth step, we let $\alpha_{ET} \equiv a_2$ and substitute α_{ET} for a_2 in Equation (7) to get

$$E = IT^{\alpha_{ET}} \tag{15}$$

Note that Equation (15) becomes a calibrated version of Equation (2) when, we perform the above-described substitution with the calibrated value for a_2 from the third step of the calibration process and the calibrated list $\mathbf{I}$ from the fourth step of the calibration process to get

$$\mathbf{E} = \mathbf{I}\mathbf{T}^{\alpha_{ET}} \tag{16}$$

©2012 ***r2Estimating, LLC***. All rights reserved.

Example

Applying the seventh step of the calibration process to our example we compute $\alpha_E \equiv 1 - a_3 = 1 - 0.1932 = 0.8068$, compute $\alpha_T \equiv a_2 a_3 = 2.2020 \times 0.1932 = 0.4255$, use $\alpha_S \equiv a_1 = 1.0373$ from the first step of the calibration process, and use the example difficulty list $\mathbf{D}$ shown in the Appendix, Table 2 to get

$$\left[\mathbf{E}^{0.8068}\mathbf{T}^{0.4255} = \mathbf{D}\mathbf{S}^{1.0373}\right]_{\dagger} \quad (17)^6$$

Applying the eighth step of the calibration process to our example we use $\alpha_{ET} \equiv a_2 = 2.2020$ from the third step of the calibration process and use the example intensity list $\mathbf{I}$ shown in the Appendix, Table 2 to get

$$\left[\mathbf{E} = \mathbf{I}\mathbf{T}^{2.2020}\right]_{\dagger} \quad (18)$$

Creating Calibrated List CDFs

Theory

Recall that part of the calibration process is to determine, for Equations (1) and (2), unique ranges and distributions for the random variables ***D*** and ***I*** from a given historical data set. Specifically, in order to support yet-to-be-described probabilistic cost and schedule analysis, we would like these random variables to be defined by their Cumulative Distribution Functions (CDFs). The fourth and sixth steps above have already produced lists $\mathbf{I}$ and $\mathbf{D}$ respectively that represent samples of their corresponding distributions. We could attempt to approximate the CDFs by fitting some known continuous CDF (e.g., normal, lognormal, beta, triangular, etc.) to each sample list. We choose, instead, to produce discrete CDFs from the sample lists by first sorting each list in ascending order and then determining the percentile rank of each sample within its corresponding list. The results are discrete CDF approximations of ***D*** and ***I*** which we label $\mathbf{D}'$ and $\mathbf{I}'$ respectively. $\mathbf{D}'$ and $\mathbf{I}'$ are lists of two-element vectors; each vector's first element is a sample value and each vector's second element is a sample's percentile rank.

Example

The ninth and tenth steps of the calibration process are to produce lists $\mathbf{D}'$ and $\mathbf{I}'$ from lists $\mathbf{D}$ and $\mathbf{I}$ (Appendix, Table 2) respectively using the previously-described percentile rank process. The resulting difficulty CDF $\mathbf{D}'$ and the resulting intensity CDF $\mathbf{I}'$ are shown in the Appendix, Table 3.

Summarized Steps of the Calibration Process

(1) Regress $\left[\mathbf{E} = \bar{\omega}\mathbf{S}^{a_1}\right]_{\text{<dataset name>}}$ to get the value of a_1.

(2) Compute $\boldsymbol{\omega}$ as $\left[\left|\boldsymbol{\omega}_{[i]} = \omega_i = E_i \big/ S_i^{a_1}\right|_{i=1}^{N}\right]_{\text{<dataset name>}}$ using the value of a_1 from (1).

©2012 ***r2Estimating, LLC***. All rights reserved.

(3) Regress $\left[\mathbf{E} = \bar{I}\mathbf{T}^{a_2}\right]_{\text{<data set name>}}$ to get the value of a_2.

(4) Compute $\mathbf{I}$ as $\left[\left|\mathbf{I}_{[i]} = I_i = E_i \big/ T_i^{a_2}\right|_{i=1}^{N}\right]_{\text{<data set name>}}$ using the value of a_2 from (3).

(5) Regress $\left[\boldsymbol{\omega} = \bar{D}\mathbf{I}^{a_3}\right]_{\text{<data set name>}}$ using the list $\boldsymbol{\omega}$ from (2) and the list $\mathbf{I}$ from (4) to get the value of a_3.

(6) Compute $\mathbf{D}$ as $\left[\left|\mathbf{D}_{[i]} = D_i = \omega_i \big/ I_i^{a_3}\right|_{i=1}^{N}\right]_{\text{<data set name>}}$ using the list $\boldsymbol{\omega}$ from (2), the list $\mathbf{I}$ from (4), and the value of a_3 from (5).

(7) Compute $\left[\alpha_E = 1 - a_3\right]_{\text{<data set name>}}$ using the value of a_3 from (6).

(8) Compute $\left[\alpha_T = a_2 a_3\right]_{\text{<data set name>}}$ using the value of a_2 from (3) and a_3 from (6).

(9) Compute $\left[\boldsymbol{I} \cong \left|\mathbf{I}'_{[i]} = \left\langle \boldsymbol{sort}_{\uparrow}\left(\mathbf{I}\right)_{[i]}, \boldsymbol{PercentileRank}\left(\boldsymbol{sort}_{\uparrow}\left(\mathbf{I}\right)\right)_{[i]} \right\rangle\right|_{i=1}^{N}\right]_{\text{<data set name>}}$

(10) Compute $\left[\boldsymbol{D} \cong \left|\mathbf{D}'_{[i]} = \left\langle \boldsymbol{sort}_{\uparrow}\left(\mathbf{D}\right)_{[i]}, \boldsymbol{PercentileRank}\left(\boldsymbol{sort}_{\uparrow}\left(\mathbf{D}\right)\right)_{[i]} \right\rangle\right|_{i=1}^{N}\right]_{\text{<data set name>}}$

Performing the process summarized above on the example set of stratified historical SI data (Aerospace Corporation, 2004) shown in the Appendix, Table 1 yields the calibrated resources-work relation illustrated Figure 4 below.

©2012 ***r2Estimating, LLC***. All rights reserved.

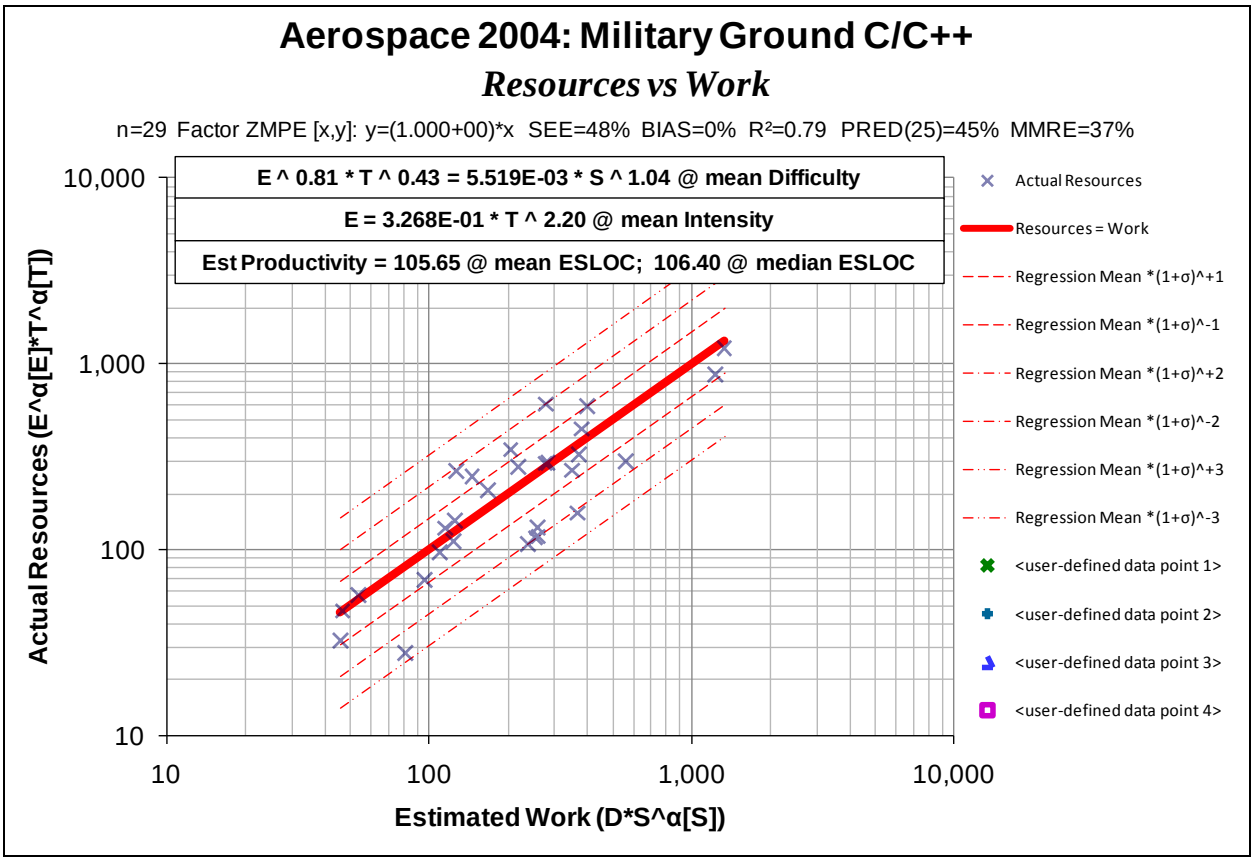

Figure 4: *Example resources-work relation – nonlinear effort and nonlinear duration as a function of mean difficulty and nonlinear size*

Incorporating Probability in CDER Mathematics

Single-point CDERs

Up until this point, we have been treating the independent variables representing intensity I, size S, and difficulty D as either certain; i.e., single-point values or as lists of samples. If we treat all these variables as single-point values, we have the single-point CDER system of equations

$$\left(E/\mathcal{E}_{af}\right)^{\alpha_E}\left(T/\mathcal{T}_{af}\right)^{\alpha_T} = DS^{\alpha_S} \tag{19}$$

and

$$E = I\left(T/\mathcal{T}_{af}\right)^{\alpha_{ET}}\mathcal{E}_{af} \tag{20}$$

where

$\mathcal{E}_{af}$ ≡ Effort Adjustment Factor; the factor that converts effort associated with the portion of the Software Development Life Cycle (SDLC) covered by the data set to effort associated with the entire SDLC (ATP through acceptance of

©2012 ***r2Estimating, LLC***. All rights reserved.

the system; referred to as all-all effort because it covers all activities and all disciplines); for the example data set used in this paper $\mathcal{E}_{af} = 1.44$.

$\mathcal{T}_{af}$ ≡ Time (duration) Adjustment Factor; the factor that converts duration associated with the portion of the SDLC covered by the data set to duration associated with the entire SDLC (referred to as all-all duration because it covers all activities and all disciplines); for the example data set used in this paper $\mathcal{T}_{af} = 1.37$.

E ≡ All-all effort.

T ≡ All-all duration.

If we rearrange the factors in Equation (19) to solve for effort E, our system of equations becomes

$$E = \left(DS^{\alpha_S} \right)^{1/\alpha_E} \left(T / \mathcal{T}_{af} \right)^{(-1)\alpha_T/\alpha_E} \mathcal{E}_{af} \tag{21}$$

and

$$E = I \left(T / \mathcal{T}_{af} \right)^{\alpha_{ET}} \mathcal{E}_{af} \tag{22}$$

Since both Equations (21) and (22) yield effort E as some function of duration T, they can be easily graphed in the same two-dimensional space. If we calibrate Equations (21) and (22) to the data set (Aerospace Corporation, 2004) that we have been using as an example, assume mean difficulty and mean intensity which are 0.005519 and 0.3268 respectively as shown in Figure 4, and assume an SI with a mean effective software size of 127,571 ESLOC[7], we get the specific CDER system of equations

$$E = \left(\left(0.005519 \right) \left(127{,}571 \right)^{1.0373} \right)^{1/0.8068} \left(T / 1.37 \right)^{(-1)0.4255/0.8068} 1.44 \tag{23}$$

and

$$E = 0.3268 \left(T / 1.37 \right)^{2.2020} 1.44 \tag{24}$$

Figure 5 below is a graph of Equations (21) and (22). The intersection of the two curves represents the CDER single-point solution; i.e., the value of effort E and the value of duration T that, together, satisfy both equations of the CDER.

©2012 ***r2Estimating, LLC***. All rights reserved.

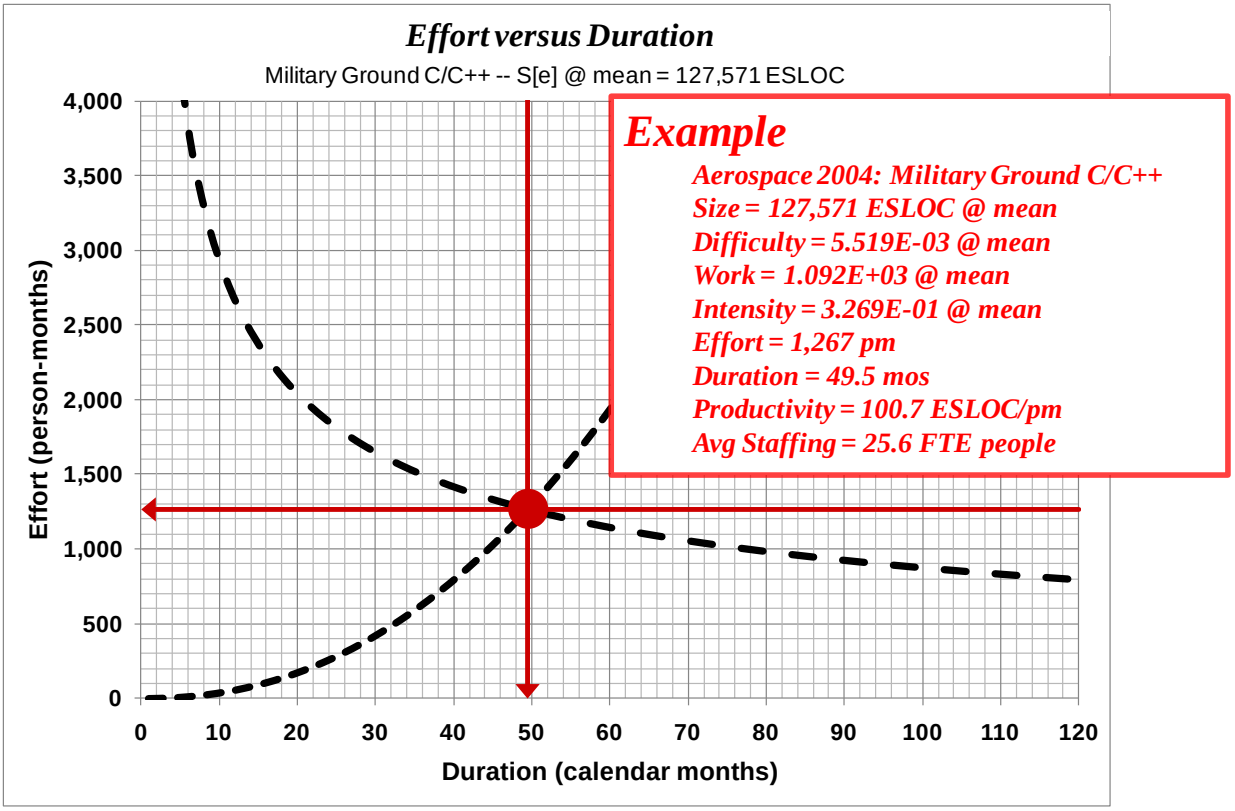

Figure 5 *Single-point CDER – work and intensity functions using single-point values*

The example CDER solution can also be found algebraically. Since the CDER solution implies both Equations (21) and (22) must be simultaneously true, we can first substitute E in Equation (21) with the equivalent of E in Equation (22) and solve for T to get

$$\begin{aligned} & I\left(T/\mathcal{T}_{af}\right)^{\alpha_{ET}}\mathcal{E}_{af} = \left(DS^{\alpha_S}\right)^{1/\alpha_E}\left(T/\mathcal{T}_{af}\right)^{(-1)\alpha_T/\alpha_E}\mathcal{E}_{af} \\ & \rightarrow T = I^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)}\left(DS^{\alpha_S}\right)^{1/(\alpha_{ET}\alpha_E+\alpha_T)}\mathcal{T}_{af} \end{aligned} \tag{25}$$

We can then substitute T in Equation (22) with the equivalent of T in Equation (25) and solve for E to get

$$\begin{aligned} & E = I\left(\left(I^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)}\left(DS^{\alpha_S}\right)^{1/(\alpha_{ET}\alpha_E+\alpha_T)}\mathcal{T}_{af}\right)\Big/\mathcal{T}_{af}\right)^{\alpha_{ET}}\mathcal{E}_{af} \\ & \rightarrow E = I^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)}\left(DS^{\alpha_S}\right)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)}\mathcal{E}_{af} \end{aligned} \tag{26}$$

If we calibrate Equations (25) and (26) to the Military Ground C/C++ data set that we have been using as an example, assume mean intensity and difficulty, and assume an SI with a mean effective software size of 127,571 ESLOC, we get

©2012 ***r2Estimating, LLC***. All rights reserved.

$$T = 0.3268^{(-1)0.8068/((2.2020)(0.8068)+0.4255)} \left((0.005519)(127,571)^{1.0373}\right)^{1/((2.2020)(0.8068)+0.4255)} 1.37 \quad (27)$$

$$\therefore T = 49.5 \text{ calendar months}$$

and

$$E = 0.3268^{0.4255/((2.2020)(0.8068)+0.4255)} \left((0.005519)(127,571)^{1.0373}\right)^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44 \quad (28)$$

$$\therefore E = 1,267 \text{ person-months}$$

Note that the solved values for duration T and effort E in Equations (27) and (28) are the same as those shown for the CDER solution in Figure 5 above (each arrow points to the solution value corresponding to its respective axis).

Uncertainty about Intensity, Size, Difficulty, Effort, and Duration

Unfortunately, until project completion, we have exact values for neither intensity, nor size, nor the difficulty of the SI being estimated; these values are ***uncertain***; i.e., they have ranges of possible outcomes. We, therefore, choose to represent intensity, size, and difficulty as random variables $\boldsymbol{I}$, $\boldsymbol{S}$, and $\boldsymbol{D}$, and our estimation process must therefore include modeling the distribution function for each. From the $\boldsymbol{S}$ and $\boldsymbol{D}$ distributions we define the size$\times$difficulty product (the work distribution $\boldsymbol{\Psi}$) as

$$\boldsymbol{\Psi} \equiv \boldsymbol{D}\boldsymbol{S}^{\alpha_S} \quad (29)$$

The choice of specific distributions for $\boldsymbol{I}$, $\boldsymbol{S}$, and $\boldsymbol{D}$ is a subject worthy of debate and a future paper. The author currently takes the position that when the data from a statistically-significant number of past projects exists, it is best to create a discrete mapping between the metric's range values and their corresponding cumulative probabilities[8] (i.e., a Cumulative Distribution Function (CDF) list) rather than assume some mathematically defined distribution such as normal, lognormal, or triangular. For the example scenario used in this paper $\boldsymbol{I}$, $\boldsymbol{S}$, and $\boldsymbol{D}$ are modeled this way, the intensity and difficulty range values coming directly from the example regression process described above and the size range values coming from a fictitious estimate of growth-adjusted Effective Source Lines of Code (ESLOC)[9] with CDF as shown in the Appendix, Table 4. Arithmetic operations involving random variables represented as CDF lists are performed row-wise on versions of these CDF lists that have been randomly shuffled, the resulting list then being percentile ranked to yield the desired CDF list result. We have used this random variable arithmetic process

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\left[\boldsymbol{\Psi} \cong \left| \boldsymbol{\Psi}'_{[i]} = \left\langle \begin{array}{l} \boldsymbol{sort}_{\uparrow}\left(\boldsymbol{shuffle}\left(\mathbf{D}\right)_{[i]} \times \boldsymbol{shuffle}\left(\mathbf{S}\right)_{[i]}{}^{\alpha_S}\right)_{[i]}, \\ \boldsymbol{PercentileRank}\left(\boldsymbol{sort}_{\uparrow}\left(\boldsymbol{shuffle}\left(\mathbf{D}\right)_{[i]} \times \boldsymbol{shuffle}\left(\mathbf{S}\right)_{[i]}{}^{\alpha_S}\right)_{[i]}\right) \end{array} \right\rangle \right|_{i=1}^{N}\right]_{\text{<data set name>}} \tag{30}$$

to compute the CDF list $\left[\boldsymbol{\Psi} = \mathbf{DS}^{\alpha_S}\right]_{\uparrow}$; the result is contained in the Appendix, Table 4.

Work and Intensity Confidence Level (Attainment Probability) as Fields

Recall that the relationship between our two dependent variables effort and duration is based on the expected size×difficulty product (expected work $\boldsymbol{\Psi}$) and the intensity $\boldsymbol{I}$; both of which we now treat as a random variables.

Our work relationship curve can now be described as a probability field (see Figure 6 below).

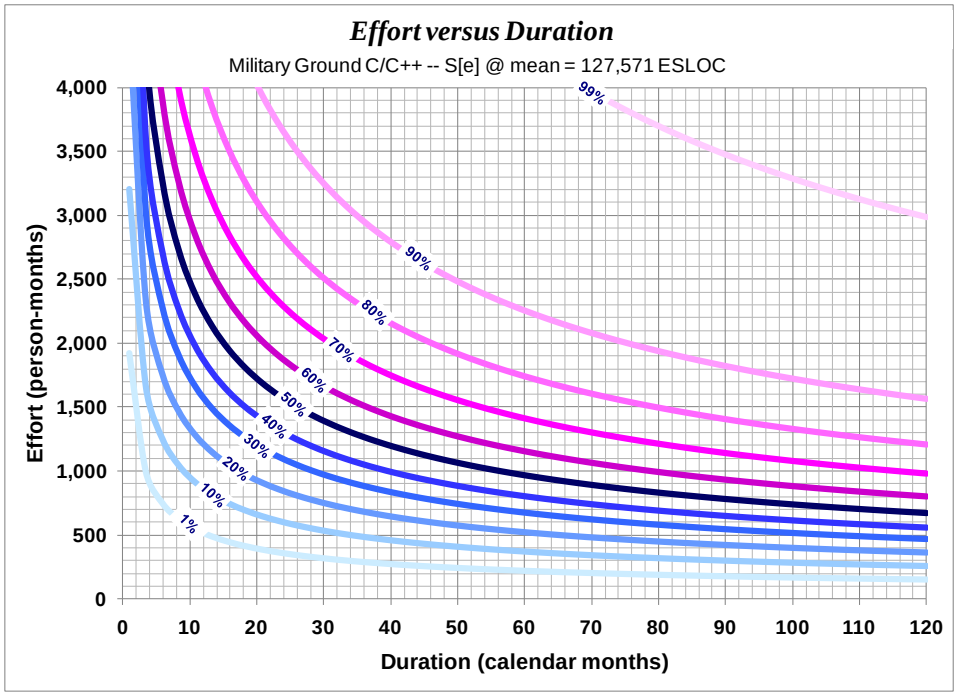

Figure 6: *Work relation as a probability field*

We can do the same thing with the intensity relationship as is shown in Figure 7 below.

©2012 ***r2Estimating, LLC***. All rights reserved.

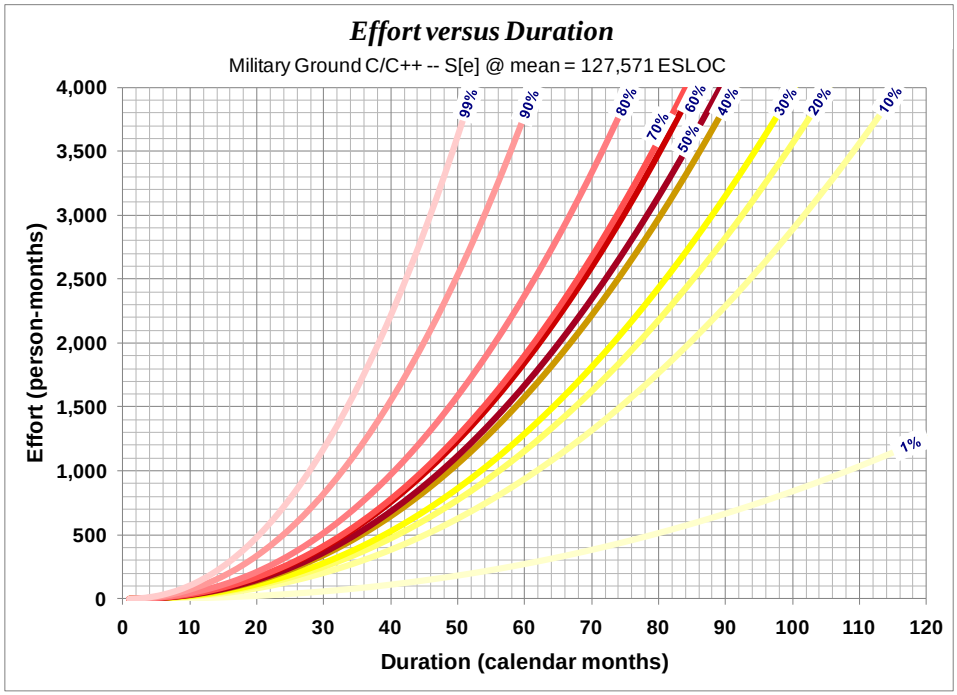

Figure 7: *Intensity relation as a probability field*

We can overlay Figure 6 and Figure 7, as is shown in Figure 8 below, to see how these two relations interact and to get a feel for the estimating scenario's solution space (range of possible outcomes with associated confidence levels (attainment probabilities).

©2012 ***r2Estimating, LLC***. All rights reserved.

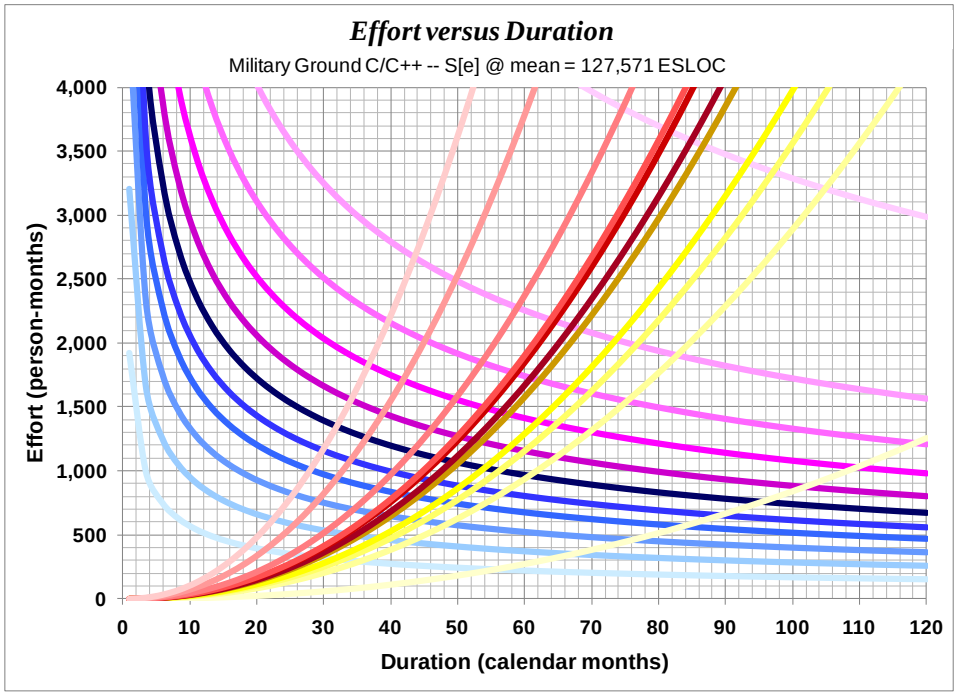

Figure 8: *Work and intensity probability fields*

Figure 6, Figure 7, and Figure 8 above show specific members (decade probability field lines) of the represented field. The mathematical expressions for describing any work or intensity field line are:

1) Work

$$\boldsymbol{E}\left(p\,|\,T,\boldsymbol{\Psi}\right)=\left(\left(\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}\left(p\right)\right)\left(T/\mathcal{T}_{af}\right)^{-\alpha_T}\right)^{1/\alpha_E} \quad \text{and}$$
$$\boldsymbol{T}\left(p\,|\,E,\boldsymbol{\Psi}\right)=\left(\left(\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}\left(p\right)\right)\left(E/\mathcal{E}_{af}\right)^{-\alpha_E}\right)^{1/\alpha_T} \tag{31}$$

2) Intensity

$$\boldsymbol{E}\left(p\,|\,T,\boldsymbol{I}\right)=\left(\boldsymbol{F}_{\boldsymbol{I}}^{-1}\left(p\right)\right)\left(T/\mathcal{T}_{af}\right)^{\alpha_{ET}} \quad \text{and}$$
$$\boldsymbol{T}\left(p\,|\,E,\boldsymbol{I}\right)=\left(\frac{\left(E/\mathcal{E}_{af}\right)}{\left(\boldsymbol{F}_{\boldsymbol{I}}^{-1}\left(p\right)\right)}\right)^{1/\alpha_{ET}} \tag{32}$$

where

$\boldsymbol{\Psi}$	≡	Work random variable
$\boldsymbol{I}$	≡	Intensity random variable
p	≡	Probability of attainment (confidence level)

©2012 ***r2Estimating, LLC***. All rights reserved.

$F_{\Psi}^{-1}(p)$ ≡ Inverse CDF (quantile function) of work random variable Ψ at probability p

$F_{I}^{-1}(p)$ ≡ Inverse CDF (quantile function) of intensity random variable I at probability p

Random Variable forms of CDER Equations

These are a collection of probabilistic equations that form the basis of a data-driven parametric joint cost and schedule estimating model (Ross, 2007a).

Bivariate Estimating Form

$$E^{\alpha_E}\Psi T^{\alpha_T} = \tag{33}$$

Work and Intensity Equations Solved for Effort

$$E = \Psi\left(T^{-1}\right)^{\alpha_T/\alpha_E} \quad 1/\alpha_E \tag{34}$$

and

$$E = IT^{\alpha_{ET}} \tag{35}$$

Work and Intensity Equations Solved for Duration

$$T = \Psi\left(E^{-1}\right)^{\alpha_E/\alpha_T} \quad 1/\alpha_T \tag{36}$$

and

$$T = \left(I^{-1}\right)^{1/\alpha_{ET}} E^{1/\alpha_{ET}} \tag{37}$$

Intensity-Correlated CER and SER Equations

These intensity-correlated CER and SER equations facilitate finding the appropriate distributions of effort and duration at some known intensity.

Cost (Effort) Estimating Relationship (CER)

$$E \; \Psi I^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \quad \alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T) \tag{38}$$

Schedule (Duration) Estimating Relationship (SER)

$$T \; \Psi\left(I^{-1}\right)^{\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \quad 1/(\alpha_{ET}\alpha_E+\alpha_T) \tag{39}$$

Intensity-Correlated CER and SER Equations for Intensity

We provide solved-for-intensity forms of the CER and SER equations to facilitate finding intensity associated with some particular solution or distribution of solutions. These equations make it

©2012 ***r2Estimating, LLC***. All rights reserved.

possible to find the duration that corresponds (correlates) to a particular effort solution and vice versa.

Cost (Effort) Estimating Relationship (CER)

$$\Psi\left(\quad^{-1}\right)^{\alpha_{ET}/\alpha_T}\boldsymbol{E}^{\quad(\alpha_{ET}\alpha_E+\alpha_T)/\alpha_T} \tag{40}$$

Schedule (Duration) Estimating Relationship (SER)

$$\Psi\quad^{1/\alpha_T}\boldsymbol{T}\left(\quad^{-1}\right)^{(\alpha_{ET}\alpha_E+\alpha_T)/\alpha_E} \tag{41}$$

Joint Confidence Level (JCL)

Theory

Joint Confidence Level (JCL), also known as joint probability, is simply the probability or likelihood that two or more events will occur simultaneously. Suppose we have two events, actual cost being less than or equal to predicted cost and actual schedule being less than or equal to predicted schedule. Since it is desirable that both these events turn out to be true, we might like to know, in addition to the individual probabilities of occurrence, the probability that both will occur. Expressed mathematically

$$JCL \equiv \boldsymbol{P}\left(A, B\right) \text{ or } \boldsymbol{P}\left(A \wedge B\right) \tag{42}$$

where A and B represent the occurrence of the two events.

We can represent actual cost and schedule as random variables $\boldsymbol{E}$ and $\boldsymbol{T}$ respectively since their outcomes are uncertain; i.e., there is some range of possible outcomes. We treat predicted (estimated) cost and schedule as given specific values $\hat{E}$ and $\hat{T}$ respectively. We rewrite the JCL Equation (42) with these variables as

$$\boldsymbol{P}\left(\left(\boldsymbol{E} \le \hat{E}\right) \wedge \left(\boldsymbol{T} \le \hat{T}\right)\right) \tag{43}$$

where

$\wedge \equiv$ Boolean (logical) AND operator

Note that the two events in Equation (43) are each represented as a Boolean expression, the expressions being separated by a Boolean operator. This results in an overall expression that can evaluate to one of only two possible outcomes for a given pair of $\boldsymbol{E}$ and $\boldsymbol{T}$ draws, TRUE or FALSE. Therefore, the result of $\left(\boldsymbol{E} \le \hat{E}\right) \wedge \left(\boldsymbol{T} \le \hat{T}\right)$ is a random variable we will call $\boldsymbol{J}$ that can be modeled as a discrete distribution of TRUE and FALSE (1 and 0) values. A list approximation of j can be described as

$$\boldsymbol{J} \equiv \left(\boldsymbol{E} \le \hat{E}\right) \wedge \left(\boldsymbol{T} \le \hat{T}\right) \rightarrow \boldsymbol{J} \cong \mathbf{J}_i = \left(\mathbf{E}_i \le \hat{E}\right) \wedge \left(\mathbf{T}_i \le \hat{T}\right)\Big|_{i=1}^{N} \tag{44}$$

We can define a CDF $\boldsymbol{F}_{\boldsymbol{J}}$ on the random variable $\boldsymbol{J}$ such that

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\text{(CDF of } \boldsymbol{J}\text{)} \qquad \boldsymbol{F}_{\boldsymbol{J}}(x) \cong \boldsymbol{F}_{\mathbf{J}}(x) = \begin{cases} 0 & x = 0 \text{ (FALSE)} \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \neg \mathbf{J}_i & 0 < x < 1 \text{ (undefined transition range)} \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \left(\neg \mathbf{J}_i + \mathbf{J}_i\right) & x = 1 \text{ (TRUE)} \end{cases} \tag{45}$$

$$x \in \{0,1\}$$

where

$\neg \equiv$ Boolean (logical) NOT operator

JCL can now be defined as

$$\text{JCL} \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{J}_i \quad \text{(\% of the } \mathbf{J} \text{ elements that evaluate to TRUE)}$$

$$\frac{1}{N}\sum_{i=1}^{N} \left(\neg \mathbf{J}_i + \mathbf{J}_i\right) \equiv 100\% \quad \text{(a } \mathbf{J} \text{ element must evaluate to either TRUE or FALSE)} \tag{46}$$

$$\therefore \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{J}_i\right) = 100\% - \frac{1}{N}\sum_{i=1}^{N} \left(\neg \mathbf{J}_i\right) \text{ and by substitution JCL} = 100\% - \frac{1}{N}\sum_{i=1}^{N} \left(\neg \mathbf{J}_i\right)$$

Example

Suppose we are asked, *"What is the probability that the system can be delivered within the budgeted cost and schedule?"* We are told that the system was budgeted at the effort and duration positions derived from mean work and mean intensity. We therefore would like to know the JCL (joint cost and schedule probability) of staying within the cost and schedule budget. Figure 9, Figure 10, and Figure 11 below provide a graphic illustration of the effort and duration JCL using our example calibrated CDER and example size distribution.

©2012 ***r2Estimating, LLC***. All rights reserved.

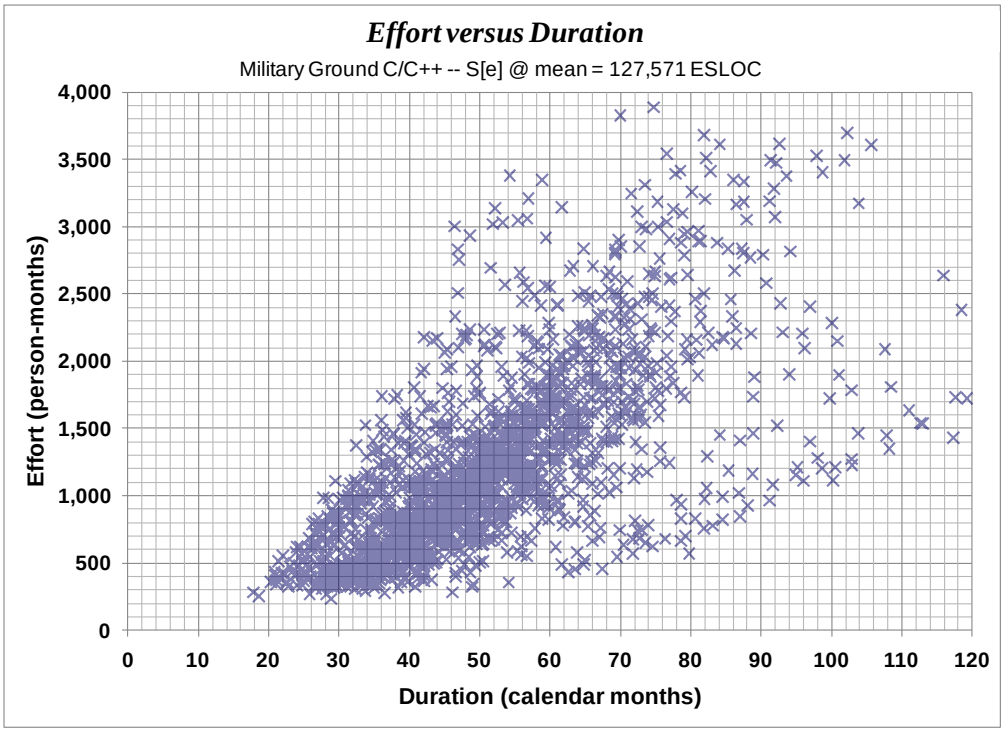

Figure 9: *Scatter Diagram of Monte Carlo draws using example calibrated CDER and example size distribution*

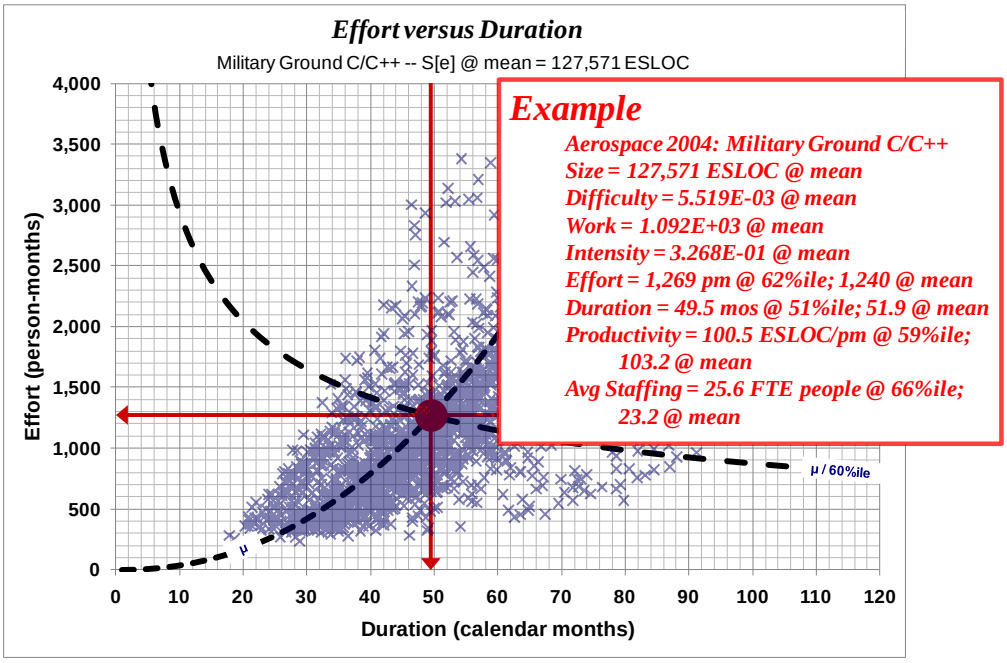

Figure 10: *Scatter Diagram of Monte Carlo Draws with mean work and intensity curves; red lines indicate the budget positions*

©2012 ***r2Estimating, LLC***. All rights reserved.

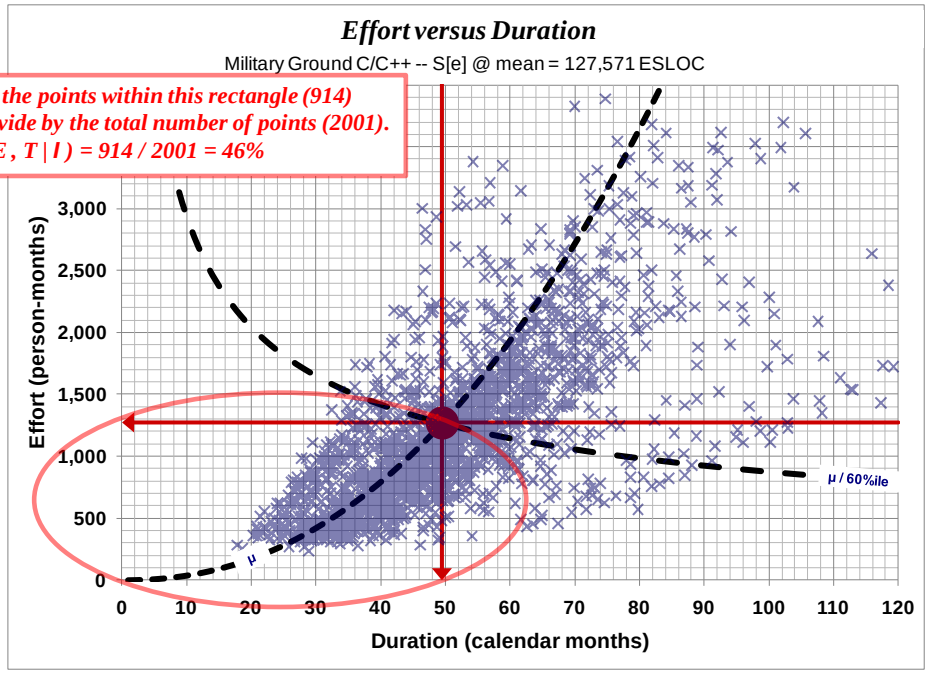

Figure 11: *Illustration of area where Monte Carlo draws satisfy both constraints*

Since it is not practical to actually count points on a scatter diagram, the JCL Boolean random variable CDF shown in Figure 12 and the mathematics shown in Equation (47) applied to lists of draws from the relevant random variables can be used to yield the same solution.

©2012 ***r2Estimating, LLC***. All rights reserved.

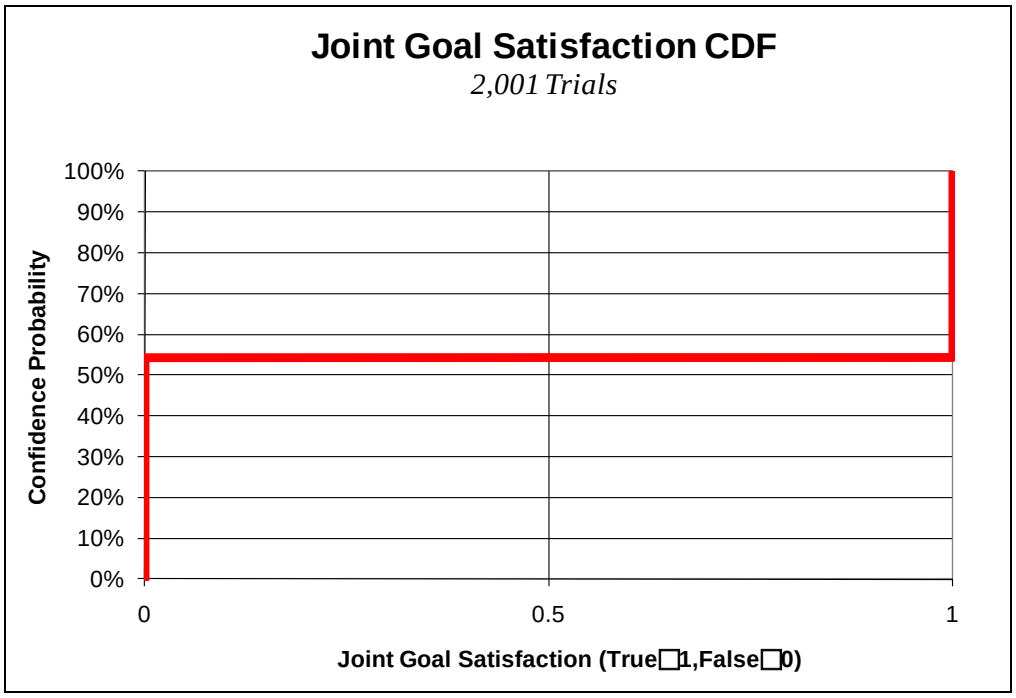

Figure 12: *CDF of JCL Boolean random variable*

$$\begin{aligned} \boldsymbol{F}_{\mathbf{J}}\left(0 < x < 1\right) &= \frac{1}{N}\sum_{i=1}^{N} \neg\mathbf{J}_i = 1097/2001 = 54\% \\ \text{JCL} &= 100\% - \frac{1}{N}\sum_{i=1}^{N}\left(\neg\mathbf{J}_i\right) = 100\% - 54\% = 46\% \end{aligned} \tag{47}$$

Conditional Confidence Level (CCL)

Conditional Confidence Level (CCL), also known as conditional probability, is simply the probability or likelihood of some event given the occurrence of some other event. Suppose we have two events, actual cost being less than or equal to predicted cost and an assumption that actual schedule will equal the predicted schedule. Conversely we could have two events, actual schedule being less than or equal to predicted schedule and an assumption that actual cost will equal the predicted cost. Expressed mathematically

$$\text{CCL (Conditional Probability)} \equiv \boldsymbol{P}\left(A \mid B = 1 \text{ (TRUE)}\right) \tag{48}$$

(read the probability that A will be true given that B is true)

or

$$\text{CCL (Conditional Probability)} \equiv \boldsymbol{P}\left(B \mid A = 1 \text{ (TRUE)}\right) \tag{49}$$

(read the probability that B will be true given that A is true)

where A and B represent the occurrence of the two events.

©2012 ***r2Estimating, LLC***. All rights reserved.

As with JCL, we can represent actual cost and schedule as random variables $\boldsymbol{E}$ and $\boldsymbol{T}$ respectively since their outcomes are uncertain; i.e., there is some range of possible outcomes. We treat predicted (estimated) cost and schedule as given specific values $\hat{E}$ and $\hat{T}$ respectively. We rewrite the CCL Equations (48) and (49) with these variables as

$$\boldsymbol{P}\left(\boldsymbol{E} \le \hat{E} \mid \hat{T}\right) \tag{50}$$

or

$$\boldsymbol{P}\left(\boldsymbol{T} \le \hat{T} \mid \hat{E}\right) \tag{51}$$

A common requirement of cost analysts is the ability to perform what-if or excursion analysis from some baseline estimating scenario. Performing these kinds of analyses generally implies the use of some form of conditional probability; i.e., solving for one or more variables conditional on some other variable(s) being assumed to take on some specific value(s). Our defined system of correlated estimating relationships offers the opportunity to perform these kinds of analyses by taking advantage of the fact that the CER and SER are correlated by intensity. If we treat intensity as having a specific value rather than as a random variable, we can use intensity as a gradient function across the work probability field. Figure 13, Figure 14, and Figure 15 illustrate how this process works. Equations (52), (53), (54), and (55) provide the intensity-correlated CER and SER equations where intensity is a single value. The following subsections provide equations for calculating intensity as a single value based on several common estimating scenarios; each of which is followed by a specific example using the Military Ground C/C++ stratification of the Aerospace Corporation (2004) Software Cost and Productivity Model database (indicated by the dagger symbol †).

©2012 ***r2Estimating, LLC***. All rights reserved.

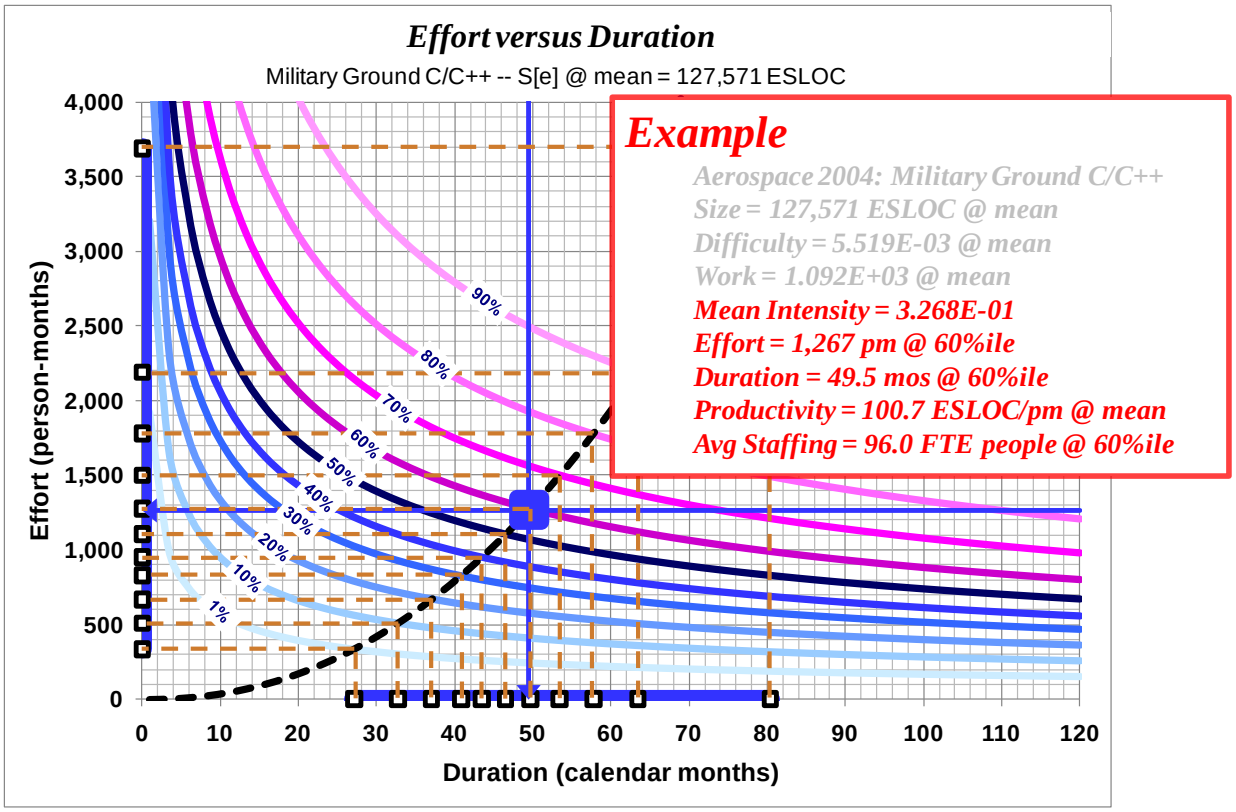

Figure 13 *Random variable (probabilistic) CDER – fix intensity according to some given constraint to create a gradient (a projection curve) across the work probability field, then project the work probability field horizontally and vertically to create a CDF representation on each of the effort and duration axes*

©2012 ***r2Estimating, LLC***. All rights reserved.

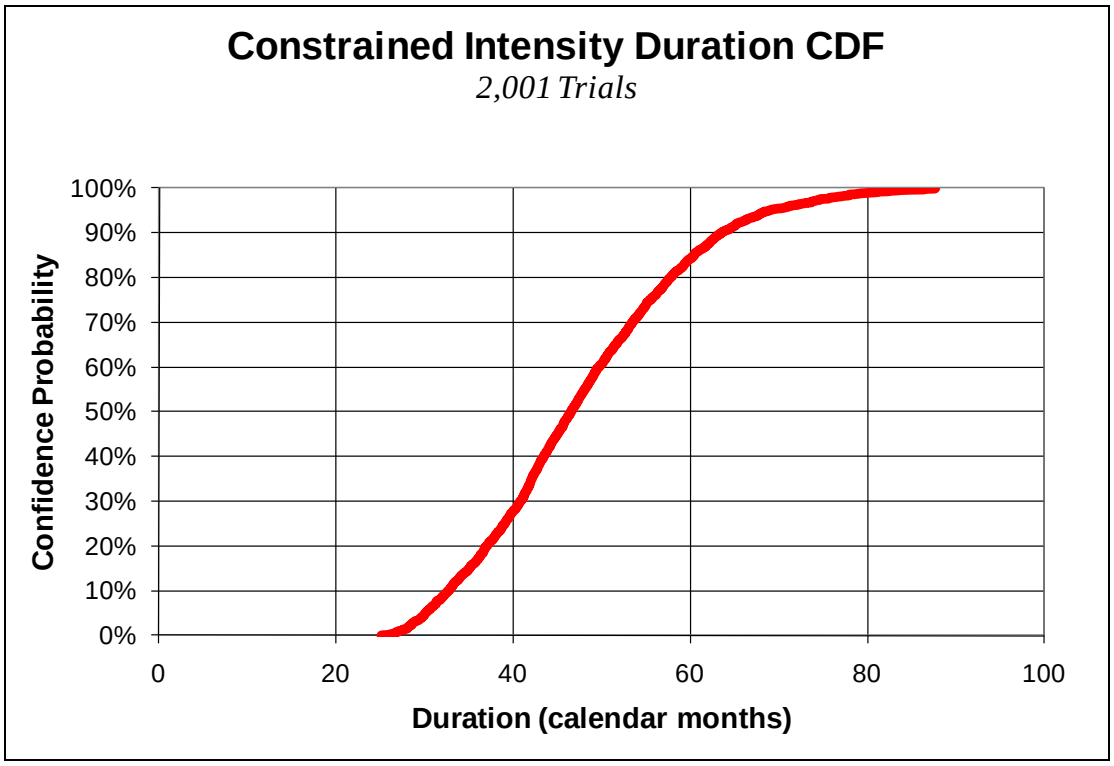

Figure 14 *All-all duration CDF – traditional form*

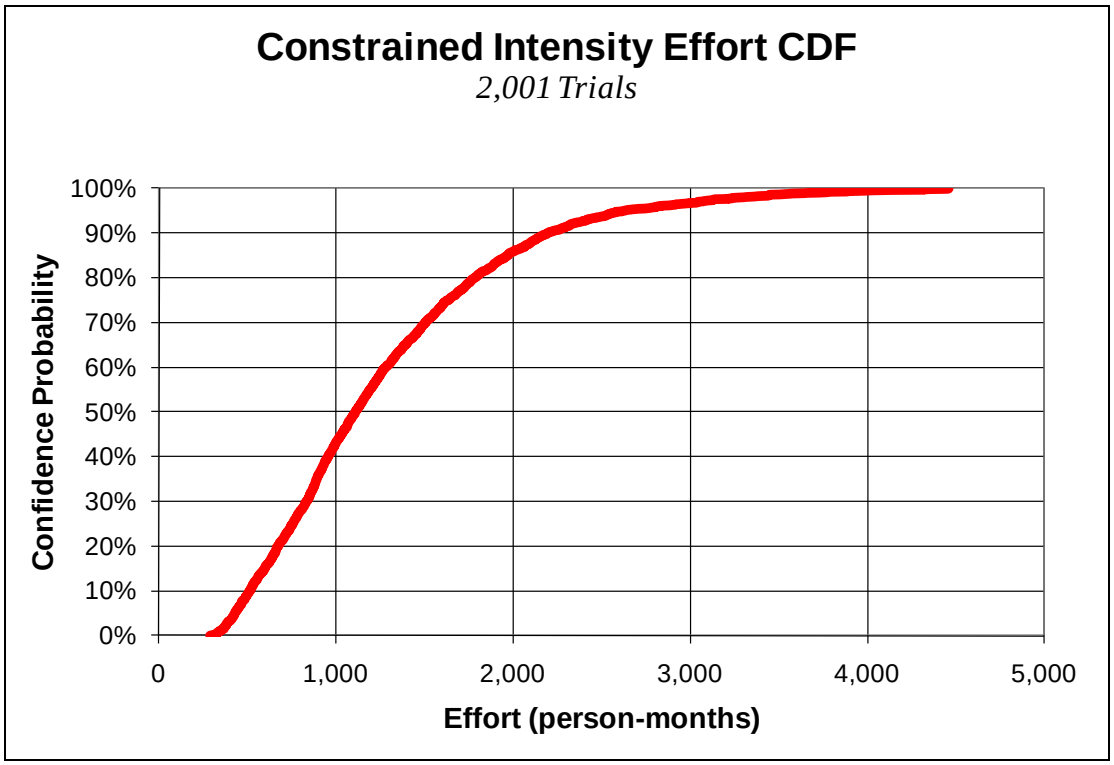

Figure 15 *All-all effort CDF – traditional form*

©2012 ***r2Estimating, LLC***. All rights reserved.

CER: Effort as a Distribution

$$\left[\boldsymbol{E\Psi}\, I^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \quad \alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)\, \mathcal{E}_{af}\right]_{\text{<dataset name>}} \tag{52}$$

CER: Effort at a Specific Confidence Level

$$\left[E = I^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)}\, \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)}\, \mathcal{E}_{af}\right]_{\text{<dataset name>}} \tag{53}$$

SER: Duration as a Distribution

$$\left[\boldsymbol{T\Psi}\,(1/I)^{\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \quad 1/(\alpha_{ET}\alpha_E+\alpha_T)\right]_{\text{<dataset name>}} \tag{54}$$

SER: Duration at a Specific Confidence Level

$$\left[T = (1/I)^{\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)}\, \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{1/(\alpha_{ET}\alpha_E+\alpha_T)}\right]_{\text{<dataset name>}} \tag{55}$$

Mean Intensity

General

$$\left[\bar{I} = \frac{1}{N}\sum \mathbf{I}\right]_{\text{<dataset name>}} \tag{56}$$

Note that $\mathbf{I}$ for † can be found in the Appendix, Table 2.

Example

Provide CDFs (S-curves) for effort and duration of our example 127,571 ESLOC SI assuming mean intensity.

$$\begin{aligned}
&\left[\bar{I} = 0.3268\right]_{\dagger}\\
&\left[\mathbf{E\Psi}\, \bar{I}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \quad \alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)\, \mathcal{E}_{af}\right]_{\dagger}\\
&\left[\mathbf{E\Psi}\, 0.3268^{0.4255/((2.2020)(0.8068)+0.4255)} \quad 2.2020/((2.2020)(0.8068)+0.4255)\, 1.44\right]_{\dagger}\\
&\left[\mathbf{T\Psi}\, \bar{I}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \quad 1/(\alpha_{ET}\alpha_E+\alpha_T)\, \mathcal{T}_{af}\right]_{\dagger}\\
&\left[\mathbf{T\Psi}\,(1/0.3268)^{0.8068/((2.2020)(0.8068)+0.4255)} \quad 1/((2.2020)(0.8068)+0.4255)\, \mathcal{T}_{af}\right]_{\dagger}
\end{aligned} \tag{57}$$

Figure 14 and Figure 15 happen to be illustrations of this estimating scenario.

If we further assume mean work then our solution becomes

©2012 ***r2Estimating, LLC***. All rights reserved.

General

$$\left[\bar{D} = \frac{1}{N}\sum \mathbf{D}\right]_{\text{<dataset name>}}$$
$$\left[\bar{\Psi} = \frac{1}{N}\sum \boldsymbol{\Psi} = \bar{D}\bar{S}^{\alpha_S}\right]_{\text{<dataset name>}} \quad (58)$$

Example

Note that $\mathbf{D}$ for † can be found in the Appendix, Table 2.

$$\left[\bar{I} = 0.3268\right]_{\dagger};\ \left[\bar{D} = 0.005519\right]_{\dagger}$$

$$\left[\bar{\Psi} = \bar{D}\bar{S}^{\alpha_S} = (0.005519)(127{,}571)^{1.0373} = 1{,}094\right]_{\dagger}$$

$$\left[T_{\bar{I}@\bar{\Psi}} = \bar{I}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)}\bar{\Psi}^{1/(\alpha_{ET}\alpha_E+\alpha_T)}\tau_{af}\right]_{\dagger}$$

$$\left[T_{\bar{I}@\bar{\Psi}} = (1/0.3268)^{0.8068/((2.2020)(0.8068)+0.4255)}\,1{,}094^{1/((2.2020)(0.8068)+0.4255)}\,1.37\right]_{\dagger}$$

$$\therefore T_{\bar{I}@\bar{\Psi}} = 49.5 \text{ calendar months}$$

$$\left[E_{\bar{I}@\bar{\Psi}} = \bar{I}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)}\bar{\Psi}^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)}\varepsilon_{af}\right]_{\dagger}$$

$$\left[E_{\bar{I}@\bar{\Psi}} = 0.3268^{0.4255/((2.2020)(0.8068)+0.4255)}\,1{,}094^{2.2020/((2.2020)(0.8068)+0.4255)}\,1.44\right]_{\dagger}$$

$$\therefore E_{\bar{I}@\bar{\Psi}} = 1{,}269 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{\bar{I}@\bar{\Psi}} = 25.6$ FTE people

$P_{\bar{I}@\bar{\Psi}} = 100.5$ ESLOC per person-month (59)

Minimum Time

General

$$\left[I_{max} = \boldsymbol{max}(\mathbf{I})\right]_{\text{<dataset name>}} \quad (60)$$

Example

Provide the 60% confidence level duration and effort associated with the minimum time solution of our example 127,571 ESLOC SI estimate. Note that $\mathbf{I}$ for † can be found in the Appendix, Table 2 and that $\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)$ for † can be found in the Appendix, Table 4.

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\left[I_{max} = 0.9126\right]_{\dagger}$$

$$\left[T_{min@60\%\Psi} = I_{max}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}\left(60\%\right)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{T}_{adj}\right]_{\dagger}$$

$$\left[T_{min@60\%\Psi} = \left(1/0.9126\right)^{0.8068/((2.2020)(0.8068)+0.4255)} 1{,}100^{1/((2.2020)(0.8068)+0.4255)} 1.37\right]_{\dagger}$$

$$\therefore T_{min@60\%\Psi} = 34.1 \text{ calendar months}$$

$$\left[E_{Tmin@60\%\Psi} = I_{max}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}\left(60\%\right)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{adj}\right]_{\dagger} \quad (61)$$

$$\left[E_{Tmin@60\%\Psi} = 0.9126^{0.4255/((2.2020)(0.8068)+0.4255)} 1{,}100^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44\right]_{\dagger}$$

$$\therefore E_{Tmin@60\%\Psi} = 1{,}557 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{Tmin@60\%\Psi} = 45.7$ FTE people

$P_{Tmin@60\%\Psi} = 82.0$ ESLOC per person-month

Minimum Effort

General

$$\left[I_{min} = \boldsymbol{min}\left(\mathbf{I}\right)\right]_{\text{<dataset name>}} \quad (62)$$

Example

Provide the 60% confidence level duration and effort associated with the minimum effort solution of our example 127,571 ESLOC SI estimate. Note that $\mathbf{I}$ for $\dagger$ can be found in the Appendix, Table 2 and that $\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}\left(60\%\right)$ can be found in the Appendix, Table 4.

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\left[I_{min} = 0.04594\right]_{\dagger}$$

$$\left[T_{Emin@60\%\Psi} = I_{min}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F_{\Psi}^{-1}}\left(60\%\right)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{T}_{af}\right]_{\dagger}$$

$$\left[T_{Emin@60\%\Psi} = \left(1/0.04594\right)^{0.8068/((2.2020)(0.8068)+0.4255)} 1{,}100^{1/((2.2020)(0.8068)+0.4255)} 1.37\right]_{\dagger}$$

$$\therefore T_{Emin@60\%\Psi} = 101.9 \text{ calendar months}$$

$$\left[E_{min@60\%\Psi} = I_{min}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F_{\Psi}^{-1}}\left(60\%\right)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{af}\right]_{\dagger} \quad (63)$$

$$\left[E_{min@60\%\Psi} = 0.04594^{0.4255/((2.2020)(0.8068)+0.4255)} 1{,}100^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44\right]_{\dagger}$$

$$\therefore E_{min@60\%\Psi} = 874 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{Emin@60\%\Psi} = 8.6$ FTE people

$P_{Emin@60\%\Psi} = 146.0$ ESLOC per person-month

Schedule Compression

General

$$\left[\%T = \left(\frac{\bar{I}}{I_{\%T}}\right)^{\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)}\right]_{\text{<dataset name>}}$$

$$\therefore \left[I_{\%T} = \bar{I}\left(\frac{1}{\%T}\right)^{(\alpha_{ET}\alpha_E+\alpha_T)/\alpha_E}\right]_{\text{<dataset name>}} \quad (64)$$

Example

Provide the 60% confidence level duration and effort associated with the 75% schedule-compressed solution of our example 127,571 ESLOC SI estimate. The 75% schedule-compressed solution constrains duration to 75% of the nominal duration (the duration that results from mean intensity). Both Boehm (1981 p. 472) and Putnam (1992 pp. 114-115) have stated that 75% represents the maximum amount of reasonably-achievable schedule compression. Note that $\boldsymbol{F_{\Psi}^{-1}}\left(60\%\right)$ for $\dagger$ can be found in the Appendix, Table 4.

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\left[I_{75\%C} = 0.7166 \right]_{\dagger}$$

$$\left[T_{75\%C@60\%\Psi} = I_{75\%C}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F_{\Psi}^{-1}}(60\%)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{T}_{af} \right]_{\dagger}$$

$$\left[T_{75\%C@60\%\Psi} = (1/0.7166)^{0.8068/((2.2020)(0.8068)+0.4255)} 1{,}100^{1/((2.2020)(0.8068)+0.4255)} 1.37 \right]_{\dagger}$$

$$\therefore T_{75\%C@60\%\Psi} = 37.2 \text{ calendar months}$$

$$\left[E_{75\%C@60\%\Psi} = I_{75\%C}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F_{\Psi}^{-1}}(60\%)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{af} \right]_{\dagger} \quad (65)$$

$$\left[E_{75\%C@60\%\Psi} = 0.7166^{0.4255/((2.2020)(0.8068)+0.4255)} 1{,}100^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44 \right]_{\dagger}$$

$$\therefore E_{75\%C@60\%\Psi} = 1{,}486 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{75\%C@60\%\Psi} = 39.9$ FTE people

$P_{75\%C@60\%\Psi} = 85.9$ ESLOC per person-month

Time as an Independent Variable

General

$$\left[I_{duration\_constraint@p\%} = \left(\frac{\boldsymbol{F_{\Psi}^{-1}}(p)}{T_{duration\_constraint@p\%}^{\alpha_{ET}\alpha_E+\alpha_T}} \right)^{1/\alpha_E} \right]_{\text{<dataset name>}} \quad (66)$$

Example

Provide the 60% confidence level effort associated with the solution of our example 127,571 ESLOC SI estimate that satisfies a duration constraint of 56 calendar months with 60% confidence. Note that $\boldsymbol{F_{\Psi}^{-1}}(60\%)$ for $\dagger$ can be found in the Appendix, Table 4.

©2012 ***r2Estimating, LLC***. All rights reserved.

Given: $T_{@60\%\Psi} = 56.0$ calendar months

$$\left[I_{56mos@60\%\Psi} = \left(\frac{\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)}{T_{@60\%\Psi}^{\alpha_{ET}\alpha_E+\alpha_T}} \right)^{1/\alpha_E} \right]_{\dagger}$$

$$\left[I_{56mos@60\%\Psi} = \left(\frac{1{,}100}{56.0^{((2.2020)(0.8068)+0.4255)}} \right)^{1/0.8068} \right]_{\dagger}$$

$$\therefore I_{56mos@60\%\Psi} = 0.2353$$

$$\left[E_{56mos@60\%\Psi@60\%\Psi} = I_{56mos@60\%\Psi}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{adj} \right]_{\dagger} \quad (67)$$

$$\left[E_{56mos@60\%\Psi@60\%\Psi} = 0.2353^{0.4255/((2.2020)(0.8068)+0.4255)} 1{,}100^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44 \right]_{\dagger}$$

$$\therefore E_{56mos@60\%\Psi@60\%\Psi} = 1{,}198 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{56mos@60\%\Psi@60\%\Psi} = 21.4$ FTE people

$P_{56mos@60\%\Psi@60\%\Psi} = 106.5$ ESLOC per person-month

Effort as an Independent Variable

General

$$\left[I_{effort\_constraint@p\%} = \left(\frac{E_{effort\_constraint@p\%}^{\alpha_{ET}\alpha_E+\alpha_T}}{\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{\alpha_{ET}}} \right)^{1/\alpha_T} \right]_{\text{<dataset name>}} \quad (68)$$

Example

Provide the 60% confidence level duration associated with the solution of our example 127,571 ESLOC SI estimate that satisfies an effort constraint of 1,198 person-months with 60% confidence. Note that $\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)$ for $\dagger$ can be found in the Appendix, Table 4. Note also that this example is using the effort solution from the time as an independent variable example above; therefore, the two solutions should be identical (which indeed they are).

©2012 ***r2Estimating, LLC***. All rights reserved.

Given: $E_{@60\%\Psi} = 1{,}198$ person-months

$$\left[I_{1198pm@60\%\Psi} = \left(\frac{\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)}{E_{@60\%\Psi}^{\alpha_{ET}\alpha_E+\alpha_T}} \right)^{1/\alpha_E} \right]_{\dagger}$$

$$\left[I_{1198pm@60\%\Psi} = \left(\frac{1{,}100}{1{,}198^{((2.2020)(0.8068)+0.4255)}} \right)^{1/0.8068} \right]_{\dagger}$$

$$\therefore I_{1198pm@60\%\Psi} = 0.2353$$

$$\left[T_{1198pm@60\%\Psi@60\%\Psi} = I_{1198pm@60\%\Psi}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{T}_{af} \right]_{\dagger} \quad (69)$$

$$\left[T_{1198pm@60\%\Psi@60\%\Psi} = (1/0.2353)^{0.8068/((2.2020)(0.8068)+0.4255)} 1{,}100^{1/((2.2020)(0.8068)+0.4255)} 1.37 \right]_{\dagger}$$

$$\therefore T_{1198pm@60\%\Psi@60\%\Psi} = 56.0 \text{ calendar months}$$

Solution implies: $\bar{\Omega}_{1198pm@60\%\Psi@60\%\Psi} = 21.4$ FTE people

$P_{1198pm@60\%\Psi@60\%\Psi} = 106.5$ ESLOC per person-month

Average Staff Level as an Independent Variable

General

$$\bar{\Omega} \equiv \frac{E}{T}$$

$$E = I_{staffing\_constraint@p\%\Psi}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{af}$$

$$T = \left(\frac{1}{I_{staffing\_constraint@p\%\Psi}} \right)^{\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{D}_{af} \quad (70)$$

$$\therefore I_{staffing\_constraint@p\%\Psi} = \frac{\left(\frac{\mathcal{D}_{af}}{\mathcal{E}_{af}} \bar{\Omega} \right)^{\alpha_{ET}\alpha_E+\alpha_T}}{\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(p)^{\alpha_{ET}-1}}$$

Example

Provide the 60% confidence level duration and effort associated with the solution of our example 127,571 ESLOC SI estimate that satisfies an average staffing constraint of 16 people with 60% confidence. Note that $\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)$ for $\dagger$ can be found in the Appendix, Table 4.

©2012 ***r2Estimating, LLC***. All rights reserved.

$$\left[I_{16\,\Omega@60\%\Psi} = \frac{\left(\frac{\mathcal{D}_{af}}{\mathcal{E}_{af}} \bar{\Omega} \right)^{(\alpha_{ET}\alpha_E+\alpha_T)/(\alpha_E+\alpha_T)}}{\boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)^{(\alpha_{ET}-1)/(\alpha_E+\alpha_T)}} \right]_\dagger$$

$$\left[I_{16\,\Omega@60\%\Psi} = \frac{\left(\frac{1.37}{1.44} 16.0 \right)^{((2.2020)(0.8068)+0.4255)/(0.8068+0.4255)}}{1{,}100^{(2.2020-1)/(0.8068+0.4255)}} \right]_\dagger$$

$$\left[I_{16\,\Omega@60\%\Psi} = 0.1400 \right]_\dagger$$

$$\left[T_{16\,\Omega@60\%\Psi@60\%\Psi} = I_{16\,\Omega@60\%\Psi}^{(-1)\alpha_E/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)^{1/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{T}_{af} \right]_\dagger$$

$$\left[T_{16\,\Omega@60\%\Psi@60\%\Psi} = (1/0.1400)^{0.8068/((2.2020)(0.8068)+0.4255)} 1{,}100^{1/((2.2020)(0.8068)+0.4255)} 1.37 \right]_\dagger$$

$$\therefore T_{16\,\Omega@60\%\Psi@60\%\Psi} = 67.7 \text{ calendar months}$$

$$\left[E_{16\,\Omega@60\%\Psi@60\%\Psi} = I_{16\,\Omega@60\%\Psi}^{\alpha_T/(\alpha_{ET}\alpha_E+\alpha_T)} \boldsymbol{F}_{\boldsymbol{\Psi}}^{-1}(60\%)^{\alpha_{ET}/(\alpha_{ET}\alpha_E+\alpha_T)} \mathcal{E}_{af} \right]_\dagger$$

$$\left[E_{16\,\Omega@60\%\Psi@60\%\Psi} = 0.1400^{0.4255/((2.2020)(0.8068)+0.4255)} 1{,}100^{2.2020/((2.2020)(0.8068)+0.4255)} 1.44 \right]_\dagger$$

$$\therefore E_{16\,\Omega@60\%\Psi@60\%\Psi} = 1{,}084 \text{ person-months}$$

Solution implies: $\bar{\Omega}_{16\,\Omega@60\%\Psi@60\%\Psi} = 16.0$ FTE people

$P_{16\,\Omega@60\%\Psi@60\%\Psi} = 117.7$ ESLOC per person-month (71)

Other Possible Solution Scenarios

- Cost as an independent variable
  - How long will it take to develop a certain amount of code with some given confidence (probability of attainment)?
- Size as a dependent variable
  - How much code can I develop within a certain budget and schedule and with some given confidence?
- Productivity as an independent variable
  - Given a certain schedule, how much money can I save if I can increase productivity by 5 ESLOC/pm?
- Confidence analysis

©2012 ***r2Estimating, LLC***. All rights reserved.

- What is the impact of budgeting at the mean versus budgeting at the 60th percentile?

Conclusion

The r2 Software Estimating Framework (r2SEF) is a *data-driven* software estimating methodology that features a generalized Cost and Duration Estimating Relationship (CDER) system of two equations. These two equations are probabilistic and can be calibrated to any historical data set that includes the size, effort, and duration of several completed Software Items (SIs). The methodology is completely open *(i.e., it is not a black box)* as evidenced by the fact that the paper contains sufficient detail with examples to permit reasonably easy implementation and deployment.

Appendix

Notation Convention

Use of Fonts

Numeric Constant - - Times New Roman : 54.0905

Simple Variable - - Times New Roman Italic : x

Function - - Times New Roman Bold Italic : $\boldsymbol{f}(x)$

Vector, Matrix, List, Array - - Times New Roman Bold : $\mathbf{X}$

Random Variable - - Arial Bold Italic : $\boldsymbol{X}$

Use of Symbols

$\equiv$	The left operand is defined as (assumed equivalent to) the right operand
$\cong$	The left operand is approximated by the right operand
$\square$	The left operand estimates the right operand
$\in$	The left operand is an element (member) of the right operand
$\propto$	The left operand is proportional to the right operand
$\wedge$	Logical AND operator
$\neg$	Logical NOT operator
$\dagger$	Indicates specific calibration to Aerospace 2004 : Military Ground C / C + + data set

©2012 ***r2Estimating, LLC***. All rights reserved.

Variable Dictionary

Letting $X \in \left\{ \begin{matrix} \text{E (effort), T (duration), S (software size), D (difficulty), I (intensity),} \\ \omega \text{ (inefficiency), } \Psi \text{ (work), } \Omega \text{ (staffing), and P (productivity)} \end{matrix} \right\}$

$\boldsymbol{X}$ ≡ Random variable of X; range and distribution of possible outcomes

$\mathbf{X}$ ≡ List of outcomes of X; $\boldsymbol{X} \cong \mathbf{X}$

X_i ≡ The i^{th} element (member) of list $\mathbf{X}$; $X_i \in \mathbf{X}$

$\mathbf{X}_{[i]}$ ≡ The list $\mathbf{X}$ indexed by i

$\bar{X}$ ≡ The expected (mean) value of $\boldsymbol{X}$ or $\mathbf{X}$

X ≡ Some specific value of X

$\alpha_E, \alpha_T, \alpha_S, \alpha_{ET}$ ≡ Effort, duration, size, and effort-duration tradeoff nonlinearities (exponents)

a_1, a_2, a_3 ≡ First, second, and third regression exponents

$\mathcal{Eaf}$ ≡ Effort adjustment factor; data set life cycle to all-all life cycle

$\mathcal{Taf}$ ≡ Duration adjustment factor; data set life cycle to all-all life cycle

Tables

Table 1: *Example data set – Military Ground C/C++* **(Aerospace Corporation, 2004)**

Observations	*Database Serial Number*	*Operating Environment*	*Application Domain*	*Size*	*Effort*	*Time*
Observation 1	299	Military Ground	Command/Control	6,000	29.0	6.00
Observation 2	102	Military Ground	Test	6,113	35.1	10.00
Observation 3	74	Military Ground	Support	6,974	42.3	11.00
Observation 4	407	Military Ground	Support	10,400	19.4	9.00
Observation 5	303	Military Ground	Test	12,200	51.0	12.00
Observation 6	431	Military Ground	Support	13,929	75.4	13.00
Observation 7	309	Military Ground	Support	14,551	97.6	16.00
Observation 8	210	Military Ground	Command/Control	15,608	86.6	13.60
Observation 9	254	Military Ground	Signal Processing	15,776	112.3	15.00
Observation 10	220	Military Ground	Database	15,978	218.9	18.40
Observation 11	71	Military Ground	OS/Exec	18,330	198.9	18.70
Observation 12	176	Military Ground	Database	20,965	165.8	17.40
Observation 13	129	Military Ground	Signal Processing	25,389	191.0	44.00
Observation 14	358	Military Ground	Test	27,008	179.0	30.00
Observation 15	387	Military Ground	Mission Planning	29,326	83.6	13.40
Observation 16	362	Military Ground	Test	31,300	84.0	16.00
Observation 17	242	Military Ground	Signal Processing	31,870	103.4	14.60
Observation 18	118	Military Ground	Signal Processing	32,000	110.0	10.00
Observation 19	107	Military Ground	Support	33,919	234.6	20.00
Observation 20	70	Military Ground	Test	34,000	529.0	24.00
Observation 21	373	Military Ground	Support	34,752	235.7	20.00
Observation 22	280	Military Ground	Command/Control	42,500	191.0	24.00
Observation 23	273	Military Ground	Signal Processing	44,447	121.1	16.40

©2012 ***r2Estimating, LLC***. All rights reserved.

Observations	Database Serial Number	Operating Environment	Application Domain	Size	Effort	Time
Observation 24	121	Military Ground	Command/Control	45,000	313.0	15.00
Observation 25	381	Military Ground	Support	46,208	360.8	24.00
Observation 26	138	Military Ground	Signal Processing	48,290	501.1	25.00
Observation 27	216	Military Ground	Mission Planning	67,280	239.4	20.40
Observation 28	155	Military Ground	Signal Processing	143,026	621.0	42.00
Observation 29	345	Military Ground	Support	154,378	1,191.3	26.00

Table 2: *Example data set scale factor lists – inefficiency* $\boldsymbol{\omega}$, *intensity* $\mathbf{I}$, *and difficulty* $\mathbf{D}$

Observations	Database Serial Number	Inefficiency	Intensity	Difficulty
Observation 1	299	3.493E-03	5.610E-01	3.906E-03
Observation 2	102	4.147E-03	2.205E-01	5.554E-03
Observation 3	74	4.359E-03	2.154E-01	5.865E-03
Observation 4	407	1.321E-03	1.537E-01	1.897E-03
Observation 5	303	2.942E-03	2.144E-01	3.962E-03
Observation 6	431	3.791E-03	2.658E-01	4.898E-03
Observation 7	309	4.690E-03	2.178E-01	6.297E-03
Observation 8	210	3.869E-03	2.764E-01	4.961E-03
Observation 9	254	4.962E-03	2.888E-01	6.308E-03
Observation 10	220	9.546E-03	3.590E-01	1.164E-02
Observation 11	71	7.522E-03	3.148E-01	9.405E-03
Observation 12	176	5.455E-03	3.076E-01	6.851E-03
Observation 13	129	5.152E-03	4.594E-02	9.343E-03
Observation 14	358	4.529E-03	1.001E-01	7.065E-03
Observation 15	387	1.942E-03	2.756E-01	2.491E-03
Observation 16	362	1.824E-03	1.874E-01	2.520E-03
Observation 17	242	2.203E-03	2.823E-01	2.813E-03
Observation 18	118	2.334E-03	6.909E-01	2.507E-03
Observation 19	107	4.686E-03	3.203E-01	5.839E-03
Observation 20	70	1.054E-02	4.834E-01	1.213E-02
Observation 21	373	4.591E-03	3.218E-01	5.715E-03
Observation 22	280	3.019E-03	1.745E-01	4.231E-03
Observation 23	273	1.827E-03	2.559E-01	2.378E-03
Observation 24	121	4.663E-03	8.050E-01	4.862E-03
Observation 25	381	5.229E-03	3.297E-01	6.480E-03
Observation 26	138	6.938E-03	4.185E-01	8.210E-03
Observation 27	216	2.350E-03	3.129E-01	2.941E-03
Observation 28	155	2.788E-03	1.655E-01	3.947E-03
Observation 29	345	4.941E-03	9.126E-01	5.029E-03

Table 3: *Example data set CDF lists – difficulty* $\mathbf{D}'$, *intensity* $\mathbf{I}'$, *and inverse intensity* $\mathbf{I}'^{-1}$

Percentile	Difficulty	Intensity	Inverse Intensity
1.72%	1.897E-03	4.594E-02	1.096E+00
5.17%	2.378E-03	1.001E-01	1.242E+00

©2012 ***r2Estimating, LLC***. All rights reserved.

8.62%	2.491E-03	1.537E-01	1.447E+00
12.07%	2.507E-03	1.655E-01	1.783E+00
15.52%	2.520E-03	1.745E-01	2.069E+00
18.97%	2.813E-03	1.874E-01	2.389E+00
22.41%	2.941E-03	2.144E-01	2.785E+00
25.86%	3.906E-03	2.154E-01	3.033E+00
29.31%	3.947E-03	2.178E-01	3.108E+00
32.76%	3.962E-03	2.205E-01	3.123E+00
36.21%	4.231E-03	2.559E-01	3.176E+00
39.66%	4.862E-03	2.658E-01	3.196E+00
43.10%	4.898E-03	2.756E-01	3.251E+00
46.55%	4.961E-03	2.764E-01	3.462E+00
50.00%	5.029E-03	2.823E-01	3.543E+00
53.45%	5.554E-03	2.888E-01	3.618E+00
56.90%	5.715E-03	3.076E-01	3.628E+00
60.34%	5.839E-03	3.129E-01	3.763E+00
63.79%	5.865E-03	3.148E-01	3.908E+00
67.24%	6.297E-03	3.203E-01	4.536E+00
70.69%	6.308E-03	3.218E-01	4.592E+00
74.14%	6.480E-03	3.297E-01	4.643E+00
77.59%	6.851E-03	3.590E-01	4.664E+00
81.03%	7.065E-03	4.185E-01	5.335E+00
84.48%	8.210E-03	4.834E-01	5.730E+00
87.93%	9.343E-03	5.610E-01	6.043E+00
91.38%	9.405E-03	6.909E-01	6.508E+00
94.83%	1.164E-02	8.050E-01	9.994E+00
98.28%	1.213E-02	9.126E-01	2.177E+01

Table 4: *Example effective size* $\mathbf{S}'$ *and work* $\mathbf{\Psi}'$ *CDF lists*

Percentile	*Effective Size (ESLOC)*	*Work*
5.00%	8.486E+04	3.652E+02
10.00%	8.939E+04	4.391E+02
15.00%	9.296E+04	5.113E+02
20.00%	9.606E+04	5.766E+02
25.00%	9.947E+04	6.483E+02
30.00%	1.033E+05	7.135E+02
35.00%	1.070E+05	7.645E+02
40.00%	1.109E+05	8.202E+02
45.00%	1.152E+05	8.861E+02
50.00%	1.194E+05	9.530E+02
55.00%	1.237E+05	1.024E+03
60.00%	1.294E+05	1.100E+03
65.00%	1.357E+05	1.195E+03
70.00%	1.422E+05	1.294E+03
75.00%	1.492E+05	1.404E+03
80.00%	1.582E+05	1.534E+03
85.00%	1.689E+05	1.694E+03
90.00%	1.819E+05	1.888E+03
95.00%	2.007E+05	2.262E+03

©2012 ***r2Estimating, LLC***. All rights reserved.

Notes

[1] See the Appendix for details about notation and the use of various typefaces.
[2] Difficulty is roughly analogous to the inverse of Putnam's (1980) productivity parameter and to the inverse of Jensen's (1983a) effective technology constant.
[3] Intensity is roughly analogous to Putnam's (1980) manpower buildup parameter and to Jensen's (1983a) software complexity.
[4] We use the symbol $\propto$ to indicate that the preceding expression is "proportional to" the succeeding expression. See the Appendix for details about notation.
[5] This paper does not go into the details about how to perform or implement regression processes. All three methods to which the paper refers are well understood and well documented. Log OLS can be performed using Microsoft Excel and its Linest function on log-transformed lists, MUPE can be performed using Tecolote's CO$TAT tool, and ZMPE can be performed using Microsoft Excel and its Solver add-in.
[6] We use the dagger symbol † to indicate the example historical data set Aerospace 2004: Military Ground C/C++. See the Appendix for details about notation.
[7] Mean position in an arbitrarily-selected example Effective Source Lines of Code (ESLOC) distribution that is based on a Delivered Source Lines of Code (DSLOC) estimate that has been growth-adjusted according to Ross (2011b) assuming 0% estimate maturity.
[8] Tecolote's ACEIT tool refers to this as a *Custom CDF*.
[9] The result of a growth-adjusted estimate of Delivered Source Lines of Code (DSLOC) (Ross, 2011b) from which an Effective Source Lines of Code (ESLOC) distribution has been calculated by applying appropriate rework.

References

Aerospace Corporation. 2004. Software Cost and Productivity Model. *Aerospace Report No. ATR-2004(8311)-1.* El Segundo, CA, USA : s.n., 2004.

Boehm, Barry W. 1981. *Software Engineering Economics.* Englewood Cliffs : Prentice-Hall, Inc., 1981. ISBN 0-13-822122-7.

Jensen, Randall W. 1983a. An Improved Macrolevel Software Development Resource Estimation Model. *Proceedings of the Fifth International Society of Parametric Analysts Conference.* St. Louis, Missouri, USA : The International Society of Parametric Analysts, April 1983a. pp. 88-92.

Putnam, Lawrence H. 1980. *Software Cost Estimating and Life-Cycle Control: Getting the Software Numbers.* New York City : IEEE Computer Society, 1980. IEEE Catalog No. EHO 165-1, Library of Congress No. 80-83083.

Putnam, Sr., Lawrence H. 1990. Measures for Excellence. 1990.

Ross, Michael A. 2011b. A Probabilistic Method for Predicting Software Code Growth. [ed.] Stephen A. Book and Edward White III. *Journal of Cost Analysis and Parametrics.* s.l. : Society of Cost Estimating and Analysis - International Society of Parametric Analysts, July-December 2011b. Vol. 4, 2, pp. 127-147. ISSN 1941-658X.

—. 2007a. Illustrating Probability in Software Cost and Schedule Estimating: Know the Odds Before Placing Your Bet. *Proceedings, AIAA SPACE 2007 Conference & Exhibition.* Long Beach, CA, USA : American Institute of Aeronautics and Astronautics, September 2007a. AIAA 2007-6022.

—. 2008. Next Generation Software Estimating Framework: 25 Years and Thousands of Projects

©2012 ***r2Estimating, LLC***. All rights reserved.

Later. [ed.] Stephen A. Book and Edward White III. *Journal of Cost Analysis and Parametrics.* s.l. : Society of Cost Estimating and Analysis - International Society of Parametric Analysts, Fall 2008. Vol. 1, 2, pp. 7-30. ISSN 1941-658X.

USAF. 2007. *Cost Risk and Uncertainty Analysis Handbook.* Hanscom AFB : Tecolote Research, Inc. for U.S. Air Force, 2007.

About the Author

Michael A. Ross has over 35 years of experience in software engineering as a developer, manager, process expert, consultant, instructor, and award-winning international speaker. Mr. Ross is currently President and CEO of r2Estimating, LLC (makers of the r2Estimator software estimating tool) and is employed as a Technical Expert by Tecolote Research, Inc. Mr. Ross's recent previous experience includes three years as Chief Scientist of Galorath Inc. (makers of the SEER suite of estimation tools) and seven years with Quantitative Software Management, Inc. (makers of the SLIM suite of software estimating tools) where he was a senior consultant and Vice President of Education Services. His prior experience includes 17 years with Honeywell Air Transport Systems (formerly Sperry Flight Systems) and two years with BAE Systems (formerly Tracor Aerospace) during which time he developed and/or managed the development of real-time embedded software for various military and commercial avionics systems. Mr. Ross is a Life Member of ISPA, is currently on the Board of Directors of its Southern California chapter, and regularly presents papers at ISPA/SCEA annual conferences (four of which have been recognized with Best Paper Awards). Mr. Ross did his undergraduate work at the United States Air Force Academy and Arizona State University, receiving a Bachelor of Science in Computer Engineering from ASU.

©2012 ***r2Estimating, LLC***. All rights reserved.