AP17: *Multiply or Divide? A Best Practice for Factor Analysis*

Shu-Ping Hu
Tecolote Research, Inc.
5266 Hollister Ave. Ste 301, Santa Barbara, CA 93111
shu@tecolote.com
805-964-6963

Co-Author: Alfred Smith

Abstract: Factors are commonly used in engineering build-up equations to derive cost estimates. For example, software development hours are often estimated based on an estimate of source lines of code (SLOC) divided by a productivity rate: SLOC per hour. To develop the productivity rate, the analyst collects software size (SLOC) and development time (hours) metrics from a variety of programs. Ideally, these programs have similar characteristics to the one being estimated, for instance, similar complexity, language, resources (people and tools) and environment. Once convinced the collected data is relevant to the estimate, the analyst has an important decision to make: develop a factor (hours per SLOC) that is multiplied by the size estimate (SLOC), or develop a factor (SLOC per hour) that is divided into the SLOC estimate. Both methods seem logical, but which method is more appropriate?

This paper discusses the pros and cons and examines the properties of these two methods. In particular, the development hour point estimate and its associated uncertainty may be quite sensitive to the method selected. The distributions that best fit the two factors (hours per SLOC and SLOC per hour) and their impacts on results are also explored. Realistic examples will be discussed using both methods. The results provide the analyst with a better understanding of the math behind each method and the impact of this very important choice: multiply or divide. The selection should not be arbitrary - there should be a sound basis and a “best practice” to follow. This study provides simple-to-follow examples, mathematical insight and general guidelines to help the analyst decide the best practice for their data and objectives.

While our central example focuses on software productivity rates, the analysis in this paper is applicable to any factor relationship. The software productivity factor is used as an illustrative example.

REFERENCE:

Hu, S. “Simple Mean, Weighted Mean, or Geometric Mean?,” 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010.

TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics

Since 1973

Multiply or Divide?

A Best Practice for Factor Analysis

7 – 10 June 2011

Dr. Shu-Ping Hu

Alfred Smith CCEA

■ Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara

■ Albuquerque ■ Colorado Springs ■ Ft. Meade ■ Ft. Monmouth ■ Goddard Space Flight Center ■ Ogden ■ Patuxent River ■ Silver Spring ■ Washington Navy Yard

■ Cleveland ■ Dahlgren ■ Denver ■ Johnson Space Center ■ Montgomery ■ New Orleans ■ Oklahoma City ■ Tampa ■ Tacoma ■ Vandenberg AFB ■ Warner Robins ALC

Objectives

- **It is common to estimate hours as a simple factor of a technical parameter such as weight, aperture, power or source lines of code (SLOC), i.e., hours = a*TechParameter**
  - "Software development hours = **a** * SLOC" is used as an example
  - Concept is applicable to any factor cost estimating relationship (CER)
- **Our objective is to address how to best estimate "a"**
  - *Multiply* SLOC by Hour/SLOC or *Divide* SLOC by SLOC/Hour?
  - Simple, weighted, or harmonic mean?
  - Role of regression analysis
  - Base uncertainty on the prediction interval rather than just the range
- **Our goal is to provide analysts a better understanding of choices available and how to select the right approach**

Outline

- **Definitions and Examples:**
  - Means (Simple, Weighted, and Harmonic Means)
  - Regression Methods (MUPE, ZMPE, and Log-Error)
  - Harmonic Mean Examples
- **Estimating SW development Hours by a Factor of**
  - *SLOC per Hour*
  - *Hour per SLOC*
- **Analyzing Factor CERs Using Different Methods**
- **Analyzing Uncertainties**
- **Realistic Example**
- **Conclusions**

Definitions of Means

- **Given a set of n observations (Z_1, Z_2, …, Z_n), simple mean (SM), weighted mean (WM), geometric mean (GM), and harmonic mean (HM) are given by**

$$SM = \frac{\sum_{i=1}^{n} Z_i}{n}$$

$$WM = \frac{\sum_{i=1}^{n} w_i Z_i}{\sum_{i=1}^{n} w_i}$$

$$GM = \left(\prod_{i=1}^{n} Z_i\right)^{1/n}$$

$$HM = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Z_i}\right)^{-1} = \frac{n}{\sum_{i=1}^{n}\frac{1}{Z_i}}$$

where w_i is the weighting factor of the ith observation (i = 1,…, n)

- **Note: HM(Z) = $(SM(1/Z))^{-1}$**
- **SM ≥ GM ≥ HM if all observations (Z_i's) are non-negative**
  - These means are the same if all observations are the same
- **Whether SM is greater or less than WM depends upon the relative magnitudes of the weighting factors**

Weighted Harmonic Mean

- **The weighted harmonic mean (WHM) is defined by**

$$WHM = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Z_i}}$$

$$WHM = \left(\frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} \frac{w_i}{Z_i} \right)^{-1}$$

where w_i is the weighting factor of the ith observation (i = 1,…, n)

- **WMH ➔** $HM = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{Z_i} \right)^{-1}$ **when all w_i's are the same**

- **HM = GM^2/SM for two observations**
  - HM(a,b) = 2/(1/a+1/b) = 2ab/(a+b) = ab/[(a+b)/2]
  - If a = 1, b = 2, then SM = 1.5, GM = 1.414, HM = 1.333

Harmonic Mean Examples

- **Productivity:**
  - John can finish a data entry job in 60 hours
  - Mary can do it in 30 hours
  - The harmonic mean is 40 hours per person
  - It will take them 40/2 = 20 hours to finish the job if they work together
- **Speed:**
  - A car travels 120 miles at a speed of 60 miles/hour
  - It travels the next 120 miles at a speed of 30 miles/hour
  - Its average speed is the harmonic mean of 60 and 30, i.e., 40 miles/hour
    - This means that if you would like to cover the 240 miles in the same time at a single speed, you would travel at 40 miles/hour
- **SM is not applicable here and often mistakenly used in places where HM should be used**

Contrasting Simple and Harmonic Means

- **Every statistic that measures the central tendencies has its own pros and cons**
- **Simple Mean Properties:**
  - It can be highly influenced by extreme data points
  - It does not give greater weights to large data points
  - It treats all data points equally, as the weights are the same in the computation formula
- **HM strongly favors the smaller numbers in the data set; it does not give equal weight to each data point**
  - If $x_1 = 1$, $x_2 = 2$, ➔ SM=1.5, GM = 1.4, HM = 1.3
  - If $x_1 = 0.1$ & $x_2 = \ldots = x_{10} = 10$ ➔ SM=9.0, GM = 6.3, HM = 0.9

Developing the Factors by Regression Analysis

- **Simple and Weighted Means are also the solutions derived by weighted least squares (WLS) regression analysis for factor equations**
- **Multiplicative regression methods will be discussed next**

Multiplicative Error Models – Log-Error, MUPE, & ZMPE

Definition of error term for $Y = f(x)*\varepsilon$

Cost variation is proportional to the level of the function

- **Log-Error: $\varepsilon \sim LN(0, \sigma^2)$ ⇨ Least squares in log space**
  - Error = Log (Y) - Log f(X)
  - Minimize the sum of squared errors; process is done in log space
- **MUPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space**
  - Error = $(Y-f(X))/f(X)$
  - Minimize the sum of squared % errors iteratively (i.e., minimize $\Sigma_i \{(y_i-f(x_i))/f_{k-1}(x_i)\}^2$, k is the iteration number)

Note: $E((Y-f(X))/f(X)) = 0$

$V((Y-f(X))/f(X)) = \sigma^2$

- **ZMPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space**
  - Error = $(Y-f(X))/f(X)$
  - Minimize the sum of squared (percentage) errors with a constraint:

$$\Sigma_i (y_i-f(x_i))/f(x_i) = 0$$

We will focus on MUPE/ZMPE factor equations in this paper

SW Dev Hours As a Factor of SLOC/Hour & Hour/SLOC (1/4)

- **SW Dev hours for a program are commonly estimated using the *source lines of code* (SLOC) and productivity rate (*SLOC per hour*)**
- **Given n paired-data ($SLOC_1$,$hour_1$), ($SLOC_2$,$hour_2$), …, ($SLOC_n$,$hour_n$), SM, WM, and HM of the productivity rate are given by**

$$\mathrm{SM(SLOC/hour)} = \mathrm{SM_Rate} = \frac{\sum_{i=1}^{n}(\mathrm{SLOC_i}/\mathrm{hour_i})}{n} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\mathrm{hour_i}/\mathrm{SLOC_i})}\right)$$

$$\mathrm{WM(SLOC/hour)} = \mathrm{WM_Rate} = \frac{\sum_{i=1}^{n} w_i(\mathrm{SLOC_i}/\mathrm{hour_i})}{\sum_{i=1}^{n}(w_i)} = \frac{\sum_{i=1}^{n}\mathrm{hour_i}(\mathrm{SLOC_i}/\mathrm{hour_i})}{\sum_{i=1}^{n}(\mathrm{hour_i})} = \frac{\sum_{i=1}^{n}(\mathrm{SLOC_i})}{\sum_{i=1}^{n}(\mathrm{hour_i})}$$

$$\mathrm{HM(SLOC/hour)} = \mathrm{HM_Rate} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathrm{SLOC_i}/\mathrm{hour_i}}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n}(\mathrm{hour_i}/\mathrm{SLOC_i})\right)^{-1}$$

- **$(\mathrm{HM(SLOC/hour)})^{-1} = \mathrm{SM(hour/SLOC)}$; $(\mathrm{SM_Rate})^{-1} = \mathrm{HM(hour/SLOC)}$**

$\mathrm{HM}(X) = (\mathrm{SM}(1/X))^{-1}$

$\mathrm{SM}(X) = (\mathrm{HM}(1/X))^{-1}$

"Flip" twice to relate HM to SM, and vice versa

SW Dev Hours As a Factor of SLOC/Hour & Hour/SLOC (2/4)

- **SW Dev hours for a program are commonly estimated based on an estimate of SLOC <u>divided</u> by an average factor of *SLOC per Hour*** (an estimate of productivity rate)

  Average Productivity Rate

  - **Divide**: Dev-Hours = Est(SLOC) **/** SM(SLOC per hour)

    $= \text{Est(SLOC)} / \{(\Sigma_i \text{SLOC}_i/\text{Hour}_i)/n\} = \text{Est(SLOC)} * \{(\Sigma_i 1/(\text{Hour}_i/\text{SLOC}_i))/n\}^{-1}$

    = Est(SLOC) * **HM(Hour/SLOC)**

- **SW Dev hours can also be predicted by an estimate of SLOC <u>times</u> an average factor of *Hour per SLOC***
  - **Multiply**: Dev-Hours = Est(SLOC) * SM(Hours per SLOC)

    = Est(SLOC) * **SM(Hour/SLOC)**

- **The "Divide" method generates a point estimate (PE) <u>smaller</u> than the "Multiply" method unless all ratios are the same**

SW Dev Hours As a Factor of SLOC/Hour & Hour/SLOC (3/4)

- **Another method to estimate SW development hours is to divide an estimate of SLOC by a <u>weighted</u> productivity rate**
  - Dev-Hours = Est(SLOC) / (Wtd SLOC per hour)

    $= \text{Est(SLOC)} / \{(\Sigma_i w_i * (\text{SLOC}_i / \text{Hour}_i))/(\Sigma_i w_i)\}$

    $= \text{Est(SLOC)} / \{(\Sigma_i \text{Hour}_i * (\text{SLOC}_i / \text{Hour}_i))/(\Sigma_i \text{Hour}_i)\}$

    $= \text{Est(SLOC)} / \{\Sigma_i \text{SLOC}_i / \Sigma_i \text{Hour}_i\} = \text{Est(SLOC)} * (\Sigma_i \text{Hour}_i)/(\Sigma_i \text{SLOC}_i)$

    $= \text{Est(SLOC)} / \text{WM}(\mathbf{SLOC_i/Hour_i}) = \text{Est(SLOC)} * \text{WM}(\mathbf{Hour_i/SLOC_i})$

Weighted Productivity Rate (weighted by hours)

- **Using WM appears to be a better approach because it provides the <u>same</u> result regardless of the methods**
- **Be wary of the "common" uncertainty approach for WM**
  - Three distributions are often applied to model the uncertainty for Dev-hours

SW Dev Hours As a Factor of SLOC/Hour & Hour/SLOC (4/4)

- **Question #1: Which method is more appropriate to estimate SW development hours, "Divide" or "Multiply"?**

  Divide: DevHours = Est(SLOC)**/SM**(SLOC/Hour) = Est(SLOC)***HM** (Hour/SLOC)(1)

  Multiply: DevHours = Est(SLOC)***SM**(Hour/SLOC) (2)

- **Question #2: Which factor is more appropriate to capture the productivity rate (*SLOC per Hour*), SM or HM?**

- **Equation 3 = Equation 4:** the **same** solution can be derived when dividing the HM of the productivity rate into the SLOC estimate

  Divide: DevHours = Est(SLOC) **/ HM(SLOC/Hour)** (3)

  Multiply: DevHours = Est(SLOC) * **SM(Hour/SLOC)** (4)

- **Using HM to represent the productivity rate seems correct as it provides the <u>same</u> result when compared to the "Multiply" method**

Analyzing a Factor CER by the MUPE/ZMPE Method

$Y = \beta * X * \varepsilon$, $\varepsilon \sim$ Distrn(1, σ^2); e.g., Y = Hour, X = SLOC

- **By the definition above, the variance of Y is proportional to the square of its driver X: $Var(Y) = \beta^2\sigma^2(X^2)$**
- **SM(Hour/SLOC) is the solution for β using both the MUPE and ZMPE methods**

$$\hat{\beta}_{(MUPE)} = \hat{\beta}_{(ZMPE)} = \frac{\sum_{i=1}^{n}(Y_i/X_i)}{n} = \frac{\sum_{i=1}^{n}(Hour_i/SLOC_i)}{n} = SM(Hour/SLOC)$$

  - MUPE is an iterative, weighted least squares (WLS) method
  - SM treats all programs equally
  - There are many kinds of factor CERs; the software productivity factor is used as an illustrative example
- See Reference 1 on page 25 for details

Analyzing a Factor CER by the WLS Method

$Y = \beta * X + \delta$, $\delta \sim$ Distrn$(0, V\sigma^2)$, e.g., Y = Hour, X = SLOC

- **If the variance of Y is proportional to its driver X, e.g., Var(Y) = $\sigma^2(X)$, then WM is the solution for β by the WLS method when the weighting factor (w) is set to be 1/X:**

$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} (1/X_i)(X_i Y_i)}{\sum_{i=1}^{n} (1/X_i)(X_i^2)} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} Hour_i}{\sum_{i=1}^{n} SLOC_i} = \frac{\sum_{i=1}^{n} SLOC_i (Hour_i / SLOC_i)}{\sum_{i=1}^{n} SLOC_i} = \hat{\beta}_{(wm)}$$

  - WM favors programs large in size, as each Hour to SLOC ratio is weighted by its SLOC

- **OLS is a special case when V = I:**

$$\hat{\beta}_{(ols)} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} (Hour_i * SLOC_i)}{\sum_{i=1}^{n} (SLOC_i^2)} = \frac{\sum_{i=1}^{n} SLOC_i^2 (Hour_i / SLOC_i)}{\sum_{i=1}^{n} SLOC_i^2}$$

- See Reference 1 on page 25 for details

Available Choices

- **Which method should we use?**
  - Divide: Est(SLOC) / SM(SLOC/Hour) = Est(SLOC)*HM(Hour/SLOC)
  - Multiply: Est(SLOC) * SM(Hour/SLOC) = Est(SLOC)/HM(SLOC/Hour)
- **Which metric should we use to capture productivity rate?**
  - SM(SLOC/Hour)
  - HM(SLOC/Hour) ➔ Est(SLOC)/HM(SLOC/Hour) = Est(SLOC)*SM(Hour/SLOC) (M)
  - SM(Hour/SLOC)
  - HM(Hour/SLOC) ➔ Est(SLOC)*HM(Hour/SLOC) = Est(SLOC)/SM(SLOC/Hour) (D)
- **Note: Only two measures (circled above) are left for discussion**
  - Using HM(SLOC/Hour) is the “Multiply” method
  - Using HM(Hour/SLOC) is the “Divide” method

What Is Your Dependent Variable?

$Y = \beta * X * \varepsilon, \qquad \varepsilon \sim Distrn(1, \sigma^2)$

- **If Y = Hour and X = SLOC, use "SM(Hour/SLOC)" as a common factor to estimate SW Dev hours based upon an estimate of SLOC if an estimate of labor cost per hour ($ per hour) is available**
  - $b_1 = (\Sigma_i(Y_i/X_i))/n = (\Sigma_i \text{Hour}_i/\text{SLOC}_i)/n =$ **SM(Hour/SLOC)**
  - Dev$ = Dev-Hours * ($/Hour) = $\mathbf{b_1}$ * Est(SLOC) * (**$/Hour**)

an estimate of SLOC

- **If Y = SLOC and X = Hour, use "SM(SLOC/Hour)" as a common factor to estimate total SLOC of a SW program based upon an estimate of Dev hours (if you have a metric to estimate "$ per SLOC")**
  - $b_2 = (\Sigma_i(Y_i/X_i))/n = (\Sigma_i \text{SLOC}_i/\text{Hour}_i)/n =$ **SM(SLOC/Hour)**
  - Dev$ = Est(SLOC) * ($/SLOC) = $\mathbf{b_2}$ * Est(Hour) * (**$/SLOC**)
- **Factor b_1 is applicable if we know the labor cost per hour, while Factor b_2 is applicable if the labor cost per SLOC is available → Recommend using b_1 not b_2 since "$ per SLOC" not available**

Uncertainty Analysis

- **SM(Hour/SLOC)**
  - Dev$ = Dev-Hours * ($/Hour)

    = {SM(Hour/SLOC) * Est(SLOC)} * (**$/Hour**)

    = {Distribution #1} * {Distribution #2 for SLOC Estimate if uncertain}
  - Use Prediction Interval (PI) to analyze the uncertainties for Distribution #1, as SM is the solution for a MUPE factor CER
- **WM(Hour/SLOC)** Use WM under certain constraints
  - Dev$ = {(Σ_i Hour$_i$ / Σ_i SLOC$_i$) * Est(SLOC)} * (**$/Hour**)

    = {Distribution #1} * {Distribution #2 for SLOC Estimate if uncertain}
  - Use PI to assess Distribution #1, since WM is a solution under WLS
  - Do **not** recommend using **three** distributions to capture the uncertainty for SW Dev hours:
    - ~~One for Σ_i Hour$_i$, another for Σ_i SLOC$_i$, and the other for Est(SLOC)~~ Excessive
    - Note: E(1/X) ≠ 1/E(X); **mean inputs may not develop a mean output**

Uncertainty Analysis for Weighted Linear CER: f(x) = a + bx

- **WLS PI when X = x_o (an estimating point):**

$$f(x_0) \pm t_{(\alpha/2,\, n-2)} * SE * \sqrt{\frac{1}{w_o} + \frac{1}{\sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{SS_{wxx}}}$$

Adjusted SE

where:

- $\bar{x}_w = \sum_{i=1}^{n} w_i (x_i) \Big/ \sum_{i=1}^{n} w_i$, $\quad SS_{wx} = \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2$
- x_o is the value of the independent variable used in calculating the estimate
- w_i is the weighting factor for the ith data point ($w_i = 1/(f(x_i))^2$ for MUPE)
- w_o is the weighting factor for y when x = x_o ($w_o = (1/f(x_o))^2$ for MUPE)
- SE is CER's standard error of estimate & "n-2" is degrees of freedom (DF)
- t(α/2, n-2) is the upper α/2 cut-off point for a t distribution with "**n-2**" DF

- **If the data set is unavailable, we can use a heuristic approach to approximate the PIs based upon the formula above**

Uncertainty Analysis for Weighted Factor CER: f(x) = bx (1/2)

- **PI for Weighted Factor CER when X = x_o:**

$$f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{SS_{wxx}}} = f(x_0) \pm t_{(\alpha/2,\, n-1)} * AdjustedSE$$

where:

- $SS_{wx} = \sum_{i=1}^{n} w_i (x_i)^2$
- x_o the value of the independent variable used in calculating the estimate
- w_i is the weighting factor for the ith data point ($w_i = 1/(bx_i)^2$ for MUPE)
- w_o is the weighting factor for y when x = x_o ($w_o = (1/bx_o)^2$ for MUPE)
- SE is CER's standard error of estimate and "n-1" is DF
- t(α/2, n-1) is the upper α/2 cut-off point for a t distribution with "**n-1**" DF

- **If the data set is unavailable, we can use a heuristic approach to approximate the PIs based upon the formula above**

Uncertainty Analysis for Weighted Factor CER: f(x) = bx (2/2)

- **PI for Weighted Factor CER when X = x_o:**

$$PI = f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{SS_{wxx}}} = f(x_0) \pm t_{(\alpha/2,\, n-1)} * AdjustedSE$$

$$= \begin{cases} f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{(1+\frac{1}{n}) b^2 x_0^2} & \text{for SM (a MUPE/ZMPE solution)} \\ f(x_0) \pm t_{(\alpha/2,\, n-1)} * SE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{\sum_{i=1}^{n} X_i}} & \text{for WM (weighted by 1/X)} \end{cases}$$

where:

- $SS_{wx} = \sum_{i=1}^{n} w_i (x_i)^2$
- w_i is the weighting factor for the ith data point ($w_i = 1/(bx_i)^2$ for MUPE)
- w_o is the weighting factor for y when $x = x_o$ ($w_o = 1/(bx_o)^2$ for MUPE)
- SE is CER's standard error of estimate and "n-1" is DF
- $t(\alpha/2, n-1)$ is the upper $\alpha/2$ cut-off point for a t distribution with "**n-1**" DF

A Realistic Example - the CER Approach (1/2)

- This example contains 8 software programs (Programs A – H). The Dev hours and SLOC numbers of these programs are listed.
- HM is significantly smaller than SM
- A power-form CER is first developed for these eight programs:
  - Dev-Hours = 0.672 * SLOC ^ 1.02
    (Standard Percent Error (SPE) = 28.3%)
  - Degrees of freedom (DF) = 6
- A factor CER is fitted, as the exponent is very close to one
  - Dev-Hours = 0.83 * SLOC
    (SPE = **26**.3%; MAD of %Error = **21**%)
- Be wary of small DF for a CER
  - The CER can be changed significantly when new data points become available
- Suggest using Dev-Hours = 0.83 * SLOC to save DF, i.e., using **SM**

Programs	Hour	SLOC	Hour/SLOC
A	7,808	8,351	0.93
B	45,864	51,961	0.88
C	13,678	24,904	0.55
D	38,012	38,345	0.99
E	30,713	29,368	1.05
F	17,130	22,027	0.78
G	13,307	29,396	0.45
H	38,286	39,644	0.97
Sum:	204,798	243,996	6.60
WM(Hour/SLOC):		= 204798 / 243996	0.84
SM(Hour/SLOC):		= 6.6 / 8	0.83
HM(Hour/SLOC):		= 1 / (10.52/ 8)	0.76
GM(Hour/SLOC):			0.80

- Using HM (i.e., divide method) underestimates actual by **8.5**%
- SPE for HM = **30**%
- MAD% Error = **26**%

A Realistic Example - the CER Approach (2/2)

- Use Dev-Hours = 0.83 * SLOC for future prediction when SLOC = a specific PE
- An 80% PI (i.e., lower bound:10th percentile; upper bound: 90th percentile) when SLOC = 54000 is given by the following formula:

$$PI = f(x_0) \pm t_{(\alpha/2,\, n-1)} * SPE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{SS_{wxx}}}$$

$$= bx_0 \pm t_{(\alpha/2,\, n-1)} * SPE * \sqrt{(1+\frac{1}{n})\, bx_0}$$

$$= 0.825 * (54000) * \left(1 \pm TINV(0.2,7) * 0.263 * \sqrt{1+1/8}\right)$$

$$= 44550 * (1 \pm 0.3941) = (26994,\ 62108)$$

- Consider using a distribution finder tool to define the uncertainty distribution for the error term for large samples (e.g., n = 25)

Programs	Hour	SLOC	Hour/SLOC
A	7,808	8,351	0.93
B	45,864	51,961	0.88
C	13,678	24,904	0.55
D	38,012	38,345	0.99
E	30,713	29,368	1.05
F	17,130	22,027	0.78
G	13,307	29,396	0.45
H	38,286	39,644	0.97
Sum:	204,798	243,996	6.60
WM(Hour/SLOC):		= 204798 / 243996	0.84
SM(Hour/SLOC):		= 6.6 / 8	0.83
HM(Hour/SLOC):		= 1 / (10.52/ 8)	0.76
GM(Hour/SLOC):			0.80

WM ≥ SM ≥ GM ≥ HM

Conclusions

- **Use "Multiply": SM(Hour/SLOC), not SM(SLOC/Hour), should be used when estimating SW hours and SW development cost, as labor rate is calculated by each hour, not by each SLOC**
- **SM(Hour/SLOC) is more relevant than WM(Hour/SLOC) for cost analysis because data points are generally independent**
  - WM can be used under certain requirements, goals, or constraints (see Ref. 1 for details)
- **HM favors smaller programs; it is biased low**
  - "Divide" delivers same solution as "Multiply" when using HM(SLOC/Hour) for rate
- **WLS offers an alternative proof that SM, as well as WM, can be a common factor in factor CERs**
  - SM is the solution for MUPE and ZMPE factor CERs
- **Based upon WLS, PI should be applied to analyze uncertainties when either SM or WM is used as a common estimating factor**
  - Use Distribution Finder (a modeling tool) to hypothesize an appropriate distribution for the error term if necessary (a future topic)

TECOLOTE RESEARCH, INC.

References

1. Hu, S., "Simple Mean, Weighted Mean, or Geometric Mean?," 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010
2. Smith, A., "Build Your Own Distribution Finder," 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010
3. Hulkower, Neal D., "Numeracy for Cost Analysts; Doing the Right Math, Getting the Math Right," ISPA/SCEA workshop, San Pedro, CA, 15 September 2009
4. Hu, S., "Prediction Interval Analysis for Nonlinear Equations," 2006 Annual SCEA International Conference, Tysons Corner, VA, 13-16 June 2006
5. Hu, S., "The Impact of Using Log-Error CERs Outside the Data Range and PING Factor," 5th Joint Annual ISPA/SCEA Conference, Broomfield, CO, 14-17 June 2005
6. Hu, S., "The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development," 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001
7. Book, S. A. and N. Y. Lao, "Minimum-Percentage-Error Regression under Zero-Bias Constraints," Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 Oct 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56
8. Hu, S. and A. R. Sjovold, "Multiplicative Error Regression Techniques," 62nd MORS Symposium, Colorado Springs, Colorado, 7-9 June 1994
9. Hu, S. and A. R. Sjovold, "Error Corrections for Unbiased Log-Linear Least Square Estimates," TR-006/2, March 1989
10. Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons, 1989, pages 37, 46, 86-88
11. Weisberg, S., Applied Linear Regression, 2nd Edition," New York: John Wiley & Sons, 1985, pages 87-88
12. Goldberger, A. S., "The Interpretation and Estimation of Cobb-Douglas Functions," Econometrica, Vol. 35, July-Oct 1968, pages 464-472

TECOLOTE RESEARCH, INC.
Bridging Engineering and Economics
Since 1973

Backup Slides

Summaries

- **Multiply or Divide?**
  - **Multiply!** Use SM(Hour/SLOC), not SM(SLOC/Hour)
    - Dev$ = LaborRate$ * Hours, not LaborRate$ * SLOC; **Hour/SLOC** more relevant than **SLOC/Hour**
- **How do we estimate Hour/SLOC: SM, WM, GM, or HM?**
  - **SM!**
    - Calculate the Hour/SLOC ratio for each historical project and take the Simple Mean of these ratios. This is appropriate when comparing projects across different organizations.
    - If the SAME development team worked for all historical projects, then WM is meaningful. WM is appropriate when comparing projects within the same organization (when attrition rate is low).
    - Besides, we just cannot find a situation in cost estimating (other than weighted inflation factors – see SCEA CEBoK) where harmonic mean is appropriate
- **How do we estimate the Hour/SLOC Uncertainty?**
  - Use the equation on slide 21 to develop the prediction interval for entire CER
  - Do not apply uncertainty to the numerator and denominator separately
  - Do apply appropriate uncertainty to SLOC
  - Instead of using a t distribution to define the prediction interval, consider using a distribution finder tool to derive a distribution more closely related to the error term

MUPE Regression Method

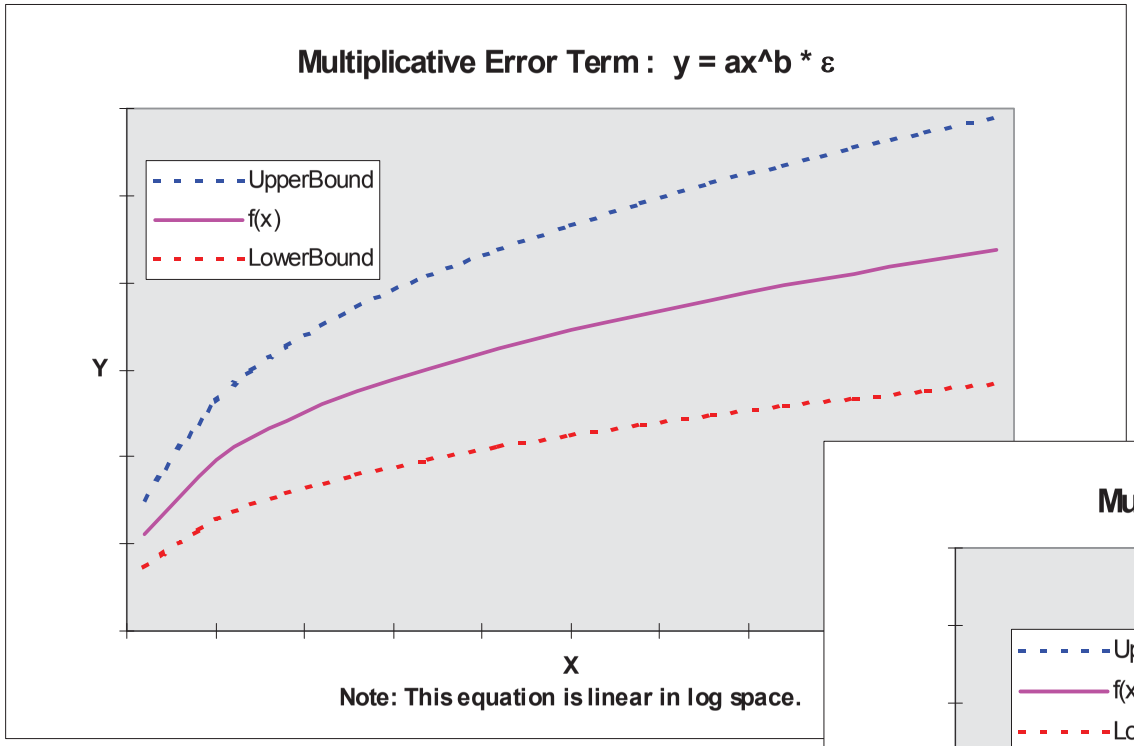
$Y = f(X)*\varepsilon$

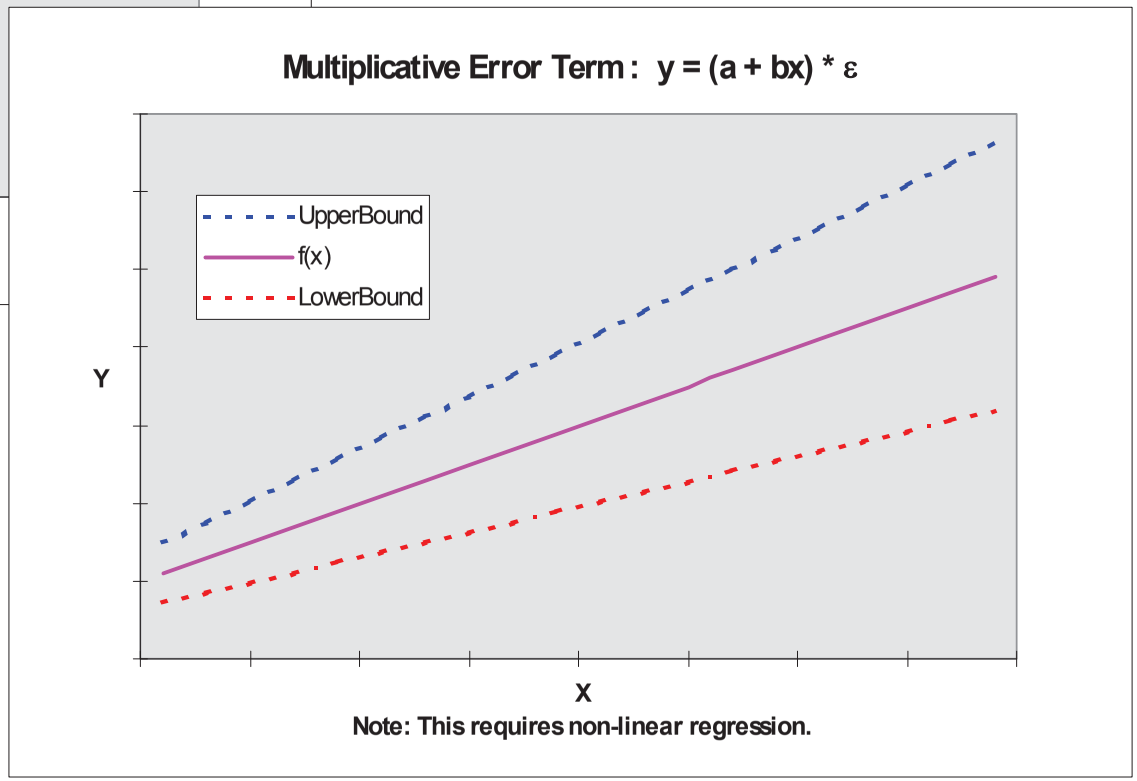
- **Minimum unbiased percentage error (MUPE) regression method is used to model CERs with multiplicative errors**
  - MUPE is an Iteratively Reweighted Least Squares (IRLS) regression technique
  - Multiplicative error assumption is appropriate when
    - Errors in the dependent variable are believed to be proportional to the value of the variable
    - Dependent variable ranges over more than one order of magnitude
- **MUPE assumptions: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in weighted space**
  - Error = $(Y-f(X))/f(X)$
  - Minimize $\Sigma_i \{(y_i-f(x_i))/f_{k-1}(x_i)\}^2$ iteratively where k is the iteration number
  - MUPE is an iterative, weighted least squares (WLS)

Note: $E((Y-f(X))/f(X)) = 0$

$V((Y-f(X))/f(X)) = \sigma^2$

Multiplicative Error Term





Cost variation is proportional to the scale of the project

Definition – SPE

- **Standard Percent Error (SPE) or Multiplicative Error:**

$$\mathrm{SPE} = \mathrm{SEE} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2}$$

(n = sample size and p = total number of estimated coefficients)

Note: SPE is CER's standard error of estimate (SEE)

- **SPE is used to measure the model's overall error of estimation; it is the one-sigma spread of the MUPE CER**
- **SPE is based upon the objective function; the smaller the value of SPE, the tighter the fit becomes**