Multiply or Divide?
A Best Practice for Factor Analysis

7 – 10 June 2011
Dr. Shu-Ping Hu
Alfred Smith CCEA

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
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
Given a set of n observations \((Z_1, Z_2, \ldots, Z_n)\), 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 (\(i = 1, \ldots, n\))

**Note:** \(HM(Z) = (SM(1/Z))^{-1}\)

\(SM \geq GM \geq HM\) 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
The weighted harmonic mean (WHM) is defined by

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

where $w_i$ is the weighting factor of the $i$th observation ($i = 1, \ldots, n$)

- **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 = GM²/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 $\frac{40}{2} = 20$ 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 \( x_1 = 1, \ x_2 = 2, \) \( \Rightarrow \) SM=1.5, GM = 1.4, HM = 1.3
- If \( x_1 = 0.1 \) & \( x_2 = \ldots = x_{10} = 10 \) \( \Rightarrow \) SM=9.0, GM = 6.3, HM = 0.9
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 \( Y = f(x)*\varepsilon \)

- **Log-Error**: \( \varepsilon \sim LN(0, \sigma^2) \)  \Rightarrow  Least squares in log space
  - Error = \( \log(Y) - \log f(X) \)
  - Minimize the sum of squared errors; process is done in log space

- **MUPE**: \( E(\varepsilon) = 1, V(\varepsilon) = \sigma^2 \)  \Rightarrow  Least squares in weighted space
  - Error = \( (Y-f(X))/f(X) \)
  - Minimize the sum of squared % errors iteratively (i.e., minimize \( \sum_i ((y_i-f(x_i))/f_{k-1}(x_i))^2 \), \( k \) is the iteration number)
  - Note: \( E((Y-f(X))/f(X)) = 0 \)
  - \( V((Y-f(X))/f(X)) = \sigma^2 \)

- **ZMPE**: \( E(\varepsilon) = 1, V(\varepsilon) = \sigma^2 \)  \Rightarrow  Least squares in weighted space
  - Error = \( (Y-f(X))/f(X) \)
  - Minimize the sum of squared (percentage) errors with a constraint:
    \[ \sum_i (y_i-f(x_i))/f(x_i) = 0 \]

Cost variation is proportional to the level of the function

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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com

Approved for Public Release

PR-70, 01 Apr 2011

9 of 25
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 \((SLOC_1, \text{hour}_1), (SLOC_2, \text{hour}_2), \ldots, (SLOC_n, \text{hour}_n)\), SM, WM, and HM of the productivity rate are given by

\[
\begin{align*}
\text{SM}(\text{SLOC/hour}) &= \text{SM}_{-\text{Rate}} = \frac{\sum^n_{i=1} (\text{SLOC}_i / \text{hour}_i)}{n} = \left( \frac{1}{n} \sum^n_{i=1} \frac{1}{(\text{hour}_i / \text{SLOC}_i)} \right) \\
\text{WM}(\text{SLOC/hour}) &= \text{WM}_{-\text{Rate}} = \frac{\sum^n_{i=1} w_i (\text{SLOC}_i / \text{hour}_i)}{\sum^n_{i=1} (w_i)} = \frac{\sum^n_{i=1} \text{hour}_i (\text{SLOC}_i / \text{hour}_i)}{\sum^n_{i=1} (\text{hour}_i)} = \frac{\sum^n_{i=1} (\text{SLOC}_i)}{\sum^n_{i=1} (\text{hour}_i)}
\end{align*}
\]

\[
\text{HM}(\text{SLOC/hour}) = \text{HM}_{-\text{Rate}} = \left( \frac{1}{n} \sum^n_{i=1} \frac{1}{\text{SLOC}_i / \text{hour}_i} \right)^{-1} = \left( \frac{1}{n} \sum^n_{i=1} (\text{hour}_i / \text{SLOC}_i) \right)^{-1}
\]

\[
\begin{align*}
(\text{HM}(\text{SLOC/hour}))^{-1} &= \text{SM}(\text{hour/SLOC}); (\text{SM}_{-\text{Rate}})^{-1} &= \text{HM}(\text{hour/SLOC})
\end{align*}
\]

\[
\begin{align*}
\text{HM}(X) &= (\text{SM}(1/X))^{-1} \\
\text{SM}(X) &= (\text{HM}(1/X))^{-1}
\end{align*}
\]

“Flip” twice to relate HM to SM, and vice versa.
SW Dev hours for a program are commonly estimated based on an estimate of SLOC divided by an average factor of SLOC per Hour (an estimate of productivity rate)

- **Divide**: Dev-Hours = \( \frac{\text{Est}(\text{SLOC})}{\text{SM}(\text{SLOC per hour})} \)
  
  \[ \frac{\text{Est}(\text{SLOC})}{\left(\frac{\sum_i \text{SLOC}_i/\text{Hour}_i}{n}\right)} = \text{Est}(\text{SLOC}) \cdot \left\{\frac{\sum_i 1/(\text{Hour}_i/\text{SLOC}_i)}{n}\right\}^{-1} \]

  = \text{Est}(\text{SLOC}) \cdot \text{HM}(\text{Hour}/\text{SLOC})

- **Multiply**: Dev-Hours = \( \text{Est}(\text{SLOC}) \cdot \text{SM}(\text{Hours per SLOC}) \)
  
  = \text{Est}(\text{SLOC}) \cdot \text{SM}(\text{Hour}/\text{SLOC})

- The “Divide” method generates a point estimate (PE) smaller than the “Multiply” method unless all ratios are the same

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Another method to estimate SW development hours is to divide an estimate of SLOC by a weighted productivity rate

\[ \text{Dev-Hours} = \frac{\text{Est}(\text{SLOC})}{(\text{Wtd SLOC per hour})} \]

\[ = \frac{\text{Est}(\text{SLOC})}{\left\{ \sum w_i \left( \frac{\text{SLOC}_i}{\text{Hour}_i} \right) \right\} / \left( \sum w_i \right)} \]

\[ = \frac{\text{Est}(\text{SLOC})}{\left\{ \sum \text{Hour}_i \left( \frac{\text{SLOC}_i}{\text{Hour}_i} \right) \right\} / \left( \sum \text{Hour}_i \right)} \]

\[ = \frac{\text{Est}(\text{SLOC})}{\left\{ \sum \text{SLOC}_i / \sum \text{Hour}_i \right\}} = \text{Est}(\text{SLOC}) \times \frac{\sum \text{Hour}_i}{\sum \text{SLOC}_i} \]

\[ = \frac{\text{Est}(\text{SLOC})}{\text{WM}(\text{SLOC}_i/\text{Hour}_i)} = \text{Est}(\text{SLOC}) \times \text{WM}(\text{Hour}_i/\text{SLOC}_i) \]

Using WM appears to be a better approach because it provides the same 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.

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Question #1: Which method is more appropriate to estimate SW development hours, “Divide” or “Multiply”? 
Divide:  \[ \text{DevHours} = \frac{\text{Est(SLOC)}}{\text{SM(SLOC/ Hour)}} = \text{Est(SLOC)} \times \text{HM (Hour/SLOC)} \]  
Multiply: DevHours = \text{Est(SLOC)} \times \text{SM(Hour/SLOC)} 

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:  \[ \text{DevHours} = \frac{\text{Est(SLOC)}}{\text{HM(SLOC/ Hour)}} \]  
Multiply:  \[ \text{DevHours} = \text{Est(SLOC)} \times \text{SM(Hour/SLOC)} \]  

Using HM to represent the productivity rate seems correct as it provides the same result when compared to the “Multiply” method.
Y = \beta X \varepsilon, \quad \varepsilon \sim \text{Distrn}(1, \sigma^2); \text{ e.g., } Y = \text{Hour}, X = \text{SLOC}

- By the definition above, the variance of Y is proportional to the square of its driver X: \text{Var}(Y) = \beta^2 \sigma^2 (X^2)

- \text{SM(Hour/SLOC)} is the solution for \beta using both the MUPE and ZMPE methods

\[ \hat{\beta}_{\text{MUPE}} = \hat{\beta}_{\text{ZMPE}} = \frac{\sum_{i=1}^{n} (Y_i / X_i)}{n} = \frac{\sum_{i=1}^{n} (\text{Hour}_i / \text{SLOC}_i)}{n} = \text{SM(Hour/SLOC)} \]

- 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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Analyzing a Factor CER by the WLS Method

\[ Y = \beta X + \delta, \quad \delta \sim \text{Distrn}(0, V \sigma^2), \, \text{e.g.,} \, Y = \text{Hour}, \, X = \text{SLOC} \]

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

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} (1/X_i) (X_i Y_i)}{\sum_{i=1}^{n} (1/X_i) (X_i^2)} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} \text{Hour}_i}{\sum_{i=1}^{n} \text{SLOC}_i} = \frac{\sum_{i=1}^{n} \text{SLOC}_i (\text{Hour}_i / \text{SLOC}_i)}{\sum_{i=1}^{n} \text{SLOC}_i} = \hat{\beta}_{(wm)}
\]

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

- OLS is a special case when \( V = I \):

\[
\hat{\beta}_{(ols)} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} (\text{Hour}_i \times \text{SLOC}_i)}{\sum_{i=1}^{n} (\text{SLOC}_i^2)} = \frac{\sum_{i=1}^{n} \text{SLOC}_i^2 (\text{Hour}_i / \text{SLOC}_i)}{\sum_{i=1}^{n} \text{SLOC}_i^2}
\]

- See Reference 1 on page 25 for details

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Which method should we use?

- **Divide:** \( \frac{\text{Est}(\text{SLOC})}{\text{SM}(\text{SLOC}/\text{Hour})} = \text{Est}(\text{SLOC}) \times \text{HM}(\text{Hour}/\text{SLOC}) \)
- **Multiply:** \( \text{Est}(\text{SLOC}) \times \text{SM}(\text{Hour}/\text{SLOC}) = \frac{\text{Est}(\text{SLOC})}{\text{HM}(\text{SLOC}/\text{Hour})} \)

Which metric should we use to capture productivity rate?

- **SM(\text{SLOC}/\text{Hour})**
- **HM(\text{SLOC}/\text{Hour})** \( \Rightarrow \text{Est}(\text{SLOC}) / \text{HM}(\text{SLOC}/\text{Hour}) = \text{Est}(\text{SLOC}) \times \text{SM}(\text{Hour}/\text{SLOC}) \) (M)
- **SM(\text{Hour}/\text{SLOC})**
- **HM(\text{Hour}/\text{SLOC})** \( \Rightarrow \text{Est}(\text{SLOC}) \times \text{HM}(\text{Hour}/\text{SLOC}) = \frac{\text{Est}(\text{SLOC})}{\text{SM}(\text{SLOC}/\text{Hour})} \) (D)

Note: Only two measures (circled above) are left for discussion

- Using HM(\text{SLOC}/\text{Hour}) is the “Multiply” method
- Using HM(\text{Hour}/\text{SLOC}) is the “Divide” method

---

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
What Is Your Dependent Variable?

\[ Y = \beta X \varepsilon, \quad \varepsilon \sim \text{Distrn}(1, \sigma^2) \]

- If \( Y = \text{Hour} \) and \( X = \text{SLOC} \), 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 (\$/hour) is available
  - \( b_1 = (\Sigma_i (Y_i/X_i))/n = (\Sigma_i \text{Hour}_i/\text{SLOC}_i)/n = \text{SM(Hour/SLOC)} \)
  - \( \text{Dev} = \text{Dev-Hours} \times (\$/\text{Hour}) = b_1 \times \text{Est(SLOC)} \times (\$/\text{Hour}) \)

- If \( Y = \text{SLOC} \) and \( X = \text{Hour} \), use “SM(SLOC/\text{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 “\$/SLOC”)
  - \( b_2 = (\Sigma_i (Y_i/X_i))/n = (\Sigma_i \text{SLOC}_i/\text{Hour}_i)/n = \text{SM(SLOC/\text{Hour})} \)
  - \( \text{Dev} = \text{Est(SLOC)} \times (\$/\text{SLOC}) = b_2 \times \text{Est(Hour)} \times (\$/\text{SLOC}) \)

- 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 “\$/SLOC” not available

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Uncertainty Analysis

- **SM(Hour/SLOC)**
  - Dev$ = Dev-Hours * ($/Hour)
  - $\{SM(Hour/SLOC) \times Est(SLOC)\} \times ($/Hour)$
  - $\{Distribution \#1\} \times \{Distribution \#2 \text{ for SLOC Estimate if uncertain}\}$
  - Use Prediction Interval (PI) to analyze the uncertainties for Distribution \#1, as SM is the solution for a MUPE factor CER

- **WM(Hour/SLOC)**
  - Dev$ = \{(\sum_i Hour_i / \sum_i SLOC_i) \times Est(SLOC)\} \times ($/Hour)$
  - $\{Distribution \#1\} \times \{Distribution \#2 \text{ for SLOC Estimate if uncertain}\}$
  - 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 $\sum_i Hour_i$, another for $\sum_i SLOC_i$, and the other for $Est(SLOC)$
    - Note: $E(1/X) \neq 1/E(X)$; mean inputs may not develop a mean output

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Uncertainty Analysis for Weighted Linear CER: $f(x) = a + bx$

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

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

where:

- $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i(x_i)}{\sum_{i=1}^{n} w_i}$, $SS_{wx} = \sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2$

- $x_o$ is the value of the independent variable used in calculating the estimate
- $w_i$ is the weighting factor for the $i$th 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)
- $t(\alpha/2, n-2)$ is the upper $\alpha/2$ 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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Uncertainty Analysis for Weighted Factor CER: \( f(x) = bx^{1/2} \)

- **PI for Weighted Factor CER when X = \( x_0 \):**

\[
 f(x_0) \pm t(\alpha/2, n-1) \times SE \times \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{SS_{wxx}}} = f(x_0) \pm t(\alpha/2, n-1) \times AdjustedSE
\]

where:

- \( SS_{wxx} = \sum_{i=1}^{n} w_i (x_i)^2 \)
- \( x_0 \) the value of the independent variable used in calculating the estimate
- \( w_i \) is the weighting factor for the \( i \)th data point (\( w_i = 1/(bx_i)^2 \) for MUPE)
- \( w_0 \) is the weighting factor for y when \( x = x_0 \) (\( w_0 = (1/bx_0)^2 \) 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

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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
Uncertainty Analysis for Weighted Factor CER: $f(x) = bx$ (2/2)

**PI for Weighted Factor CER when $X = x_0$:**

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

$$= f(x_0) \pm t(\frac{\alpha}{2}, n-1) \cdot AdjustedSE$$

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

where:

- $SS_{wxx} = \sum_{i=1}^{n} w_i(x_i)^2$
- $w_i$ is the weighting factor for the $i$th data point ($w_i = 1/(bx_i)^2$ for MUPE)
- $w_o$ is the weighting factor for y when $x = x_0$ ($w_o = 1/(bx_o)^2$ 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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
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

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
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)

<table>
<thead>
<tr>
<th>Programs</th>
<th>Hour</th>
<th>SLOC</th>
<th>Hour/SLOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7,808</td>
<td>8,351</td>
<td>0.93</td>
</tr>
<tr>
<td>B</td>
<td>45,864</td>
<td>51,961</td>
<td>0.88</td>
</tr>
<tr>
<td>C</td>
<td>13,678</td>
<td>24,904</td>
<td>0.55</td>
</tr>
<tr>
<td>D</td>
<td>38,012</td>
<td>38,345</td>
<td>0.99</td>
</tr>
<tr>
<td>E</td>
<td>30,713</td>
<td>29,368</td>
<td>1.05</td>
</tr>
<tr>
<td>F</td>
<td>17,130</td>
<td>22,027</td>
<td>0.78</td>
</tr>
<tr>
<td>G</td>
<td>13,307</td>
<td>29,396</td>
<td>0.45</td>
</tr>
<tr>
<td>H</td>
<td>38,286</td>
<td>39,644</td>
<td>0.97</td>
</tr>
<tr>
<td>Sum:</td>
<td>204,798</td>
<td>243,996</td>
<td>6.60</td>
</tr>
</tbody>
</table>

WM(Hour/SLOC): \[= \frac{204,798}{243,996} = 0.84\]
SM(Hour/SLOC): \[= \frac{6.6}{8} = 0.83\]
HM(Hour/SLOC): \[= \frac{1}{(10.52/8)} = 0.76\]
GM(Hour/SLOC): \[= 0.80\]

WM ≥ SM ≥ GM ≥ HM

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
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).

Presented at the 2011 ISPA/SCEA Joint Annual Conference and Training Workshop - www.iceaaonline.com
References

1. Hu, S., “Simple Mean, Weighted Mean, or Geometric Mean?,” 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010
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
Backup Slides
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
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 $\sum_i \{(y_i-f(x_i))/f_{k-1}(x_i)\}^2$ iteratively

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

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

$y = ax^b \times \varepsilon$

Note: This equation is linear in log space.

Cost variation is proportional to the scale of the project

$y = (a + bx) \times \varepsilon$

Note: This requires non-linear regression.
Standard Percent Error (SPE) or Multiplicative Error:

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

(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**