Simple Mean, Weighted Mean, or Geometric Mean?

Dr. Shu-Ping Hu

2010 ISPA/SCEA Professional Development and Training Workshop
San Diego, CA
8 to 11 June 2010

TECOLOTE RESEARCH, INC.
420 S. Fairview Avenue, Suite 201
Goleta, CA 93117-3626
(805) 571-6366

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

Simple Mean, Weighted Mean, or Geometric Mean?

Dr. Shu-Ping Hu
Tecolote Research, Inc.

ABSTRACT

There are three commonly used methods to calculate an average (i.e., mean): the simple average, weighted average, and geometric average. Analysts often use averages as estimating guidance when predicting nonrecurring (NR) costs using ratios of NR costs to first unit (T1) production costs. For example, when the "average" of the NR to T1 ratios is determined, the NR cost can be estimated as a "factor" of its T1 cost. Here, the simple average is simply the arithmetic mean of these ratios while the weighted average is derived by weighting each ratio by its T1 cost. Consequently, deciding which average (i.e., factor) is the most appropriate metric for estimating the nonrecurring cost is frequently debated.

There are some academic concerns about the relevance of using simple and weighted averages. Some analysts insist that simple averages are derived by the "wrong math," as averaging the averages breaks the fundamental rules of math. These analysts believe weighted averages are based upon the "correct math" and should be applied accordingly. Other analysts argue that both simple and weighted averages have their shortcomings and the geometric average is the appropriate approach.

This paper discusses these concerns and examines the properties of these methods. In addition, this study analyzes all three different methods of calculating an average using weighted least squares (WLS) in regression analysis. As a result, analysts will have a better understanding of the math behind each method, and this will help them select an appropriate and suitable method to meet their requirements for calculating an average of the NR to T1 ratios (or any ratios of interest). Realistic examples will be discussed using all three methods.

Note that there are many kinds of factor relationships, such as cost-to-cost, cost-to-weight, and weight-to-power factors. The analysis in this paper is applicable to any factor relationship and we simply use the NR to T1 factor as an illustrative example.

OUTLINE

The objectives of this paper are three-fold. First we will address the criticisms and concerns about using simple, weighted, and geometric averages. We will then present examples using Simpson's Paradox to explain why these concerns are not relevant. We will also analyze simple, weighted, and geometric averages derived by different fitting methods using regression analysis. As a result, analysts may have a better understanding of the characteristics of these measures and how to apply them appropriately in their applications. A realistic example will be discussed to illustrate various NR to T1 factors using different regression methods.

We will discuss the topics below in the following sections:

- Definitions of Simple, Weighted, and Geometric Means
- Concerns about Using Simple, Weighted, and Geometric Means (SM, WM, and GM)

- Simpson's Paradox
- NR Cost As a Factor of T1 Cost Example
  - Math Formulas of SM, WM, and GM for NR to T1 ratios
  - Brief Introduction of Several Popular Regression Methods
- Analyzing a Factor CER Using Different Methods
  - Simple Mean for Factor CER by MUPE and ZMPE Methods
  - Weighted Mean for Factor CER by WLS Method (Including OLS Solution)
  - Geometric Mean for Factor CER by Log-Error Method
- A Realistic Example
- Conclusions

We will define the acronyms used above in the following sections.

DEFINITIONS OF SIMPLE, WEIGHTED, AND GEOMETRIC MEANS

Let us first define simple, weighted, and geometric means before discussing the concerns about using them.

Definitions. Given a set of n observations (Z_1, Z_2, …, Z_n), the simple mean (SM), weighted mean (WM), and geometric mean (GM) are defined by the following formulas, respectively:

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

$$WM = \frac{\sum_{i=1}^{n} w_i Z_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

$$GM = \left(\prod_{i=1}^{n} Z_i\right)^{1/n} \tag{3}$$

where n is the sample size and w_i is the weighting factor of the i^{th} observation (i = 1,…, n).

Simple Mean ≥ Geometric Mean. It can be shown that the arithmetic mean (i.e., simple mean) is greater than or equal to the geometric mean if all the observations are non-negative:

$$\frac{\sum_{i=1}^{n} Z_i}{n} \geq \left(\prod_{i=1}^{n} Z_i\right)^{1/n} \tag{4}$$

SM ≥ WM or SM ≤ WM? Can we conclude whether the simple mean is greater than or less than the weighted mean? The answer is **no** because it depends upon the relative magnitudes of the weights when applying different weighting factors. This can help us explain *Simpson's Reversal of Inequalities*, or *Simpson's Paradox*.

CONCERNS ABOUT USING SIMPLE, WEIGHTED, AND GEOMETRIC MEANS

There are some growing concerns about the relevance of using simple averages as factors to predict the NR cost based upon the T1 cost (see Reference 1). Some analysts insist that simple averages are derived by the "wrong math," as averaging the averages breaks the fundamental rules of math. These analysts believe weighted averages are based upon the "correct math" and should be applied accordingly. Some other analysts argue that both simple and weighted averages have their shortcomings and the geometric average is the appropriate approach. Below are a few criticisms about the use of simple mean that have been raised recently:

- Simple mean favors small programs.
- Simple mean is incapable of predicting the error of the estimate.
- Simple mean contains no risk due to modeling error.
- Simple mean is calculated based upon the wrong math; statistical summing is not done arithmetically and we should never average percents, ratios, or averages to create cost estimating relationships (CER), factors, or models.

In fact, none of the above criticisms are true. Refuting each in turn:

- Simple mean does not favor small programs; it treats all the data points equally (percentage-wise).
- We can evaluate the error of the simple mean as an estimate in regression analysis under the distribution assumption.
- There is definitely uncertainty associated with simple mean.
- Simple mean is the solution of the Minimum-Unbiased-Percentage-Error (MUPE) and Zero-Percentage Bias (ZMPE) methods for factor equations. It is not wrong mathematically—in fact, it may be a logical and sound approach for many applications. A well-known and often cited fact in risk simulation analysis is that the mean of the sum is the sum of the individual means.

We will return to the points above in the following sections.

The weighted mean, however, does favor programs that are large in size. This statistic can be easily influenced by outliers in the data set such as extremely large data points, which may be a shortcoming of using this method. On the other hand, geometric mean seems to minimize the influence of extreme data points, which is a benefit of using this measure. However, the value of geometric mean is biased low in unit space and hence it should be adjusted to reflect the mean in unit space. The commonly used correction factors are Goldberger's Factor, the Smearing Estimate, the PING Factor, etc. (see References 4, 10, 14, and 16 for details).

SIMPSON'S PARADOX

What Is Simpson's Paradox? Simpson's Paradox is used to denote an intriguing phenomenon where an apparent relationship in different groups seems to be reversed when the groups are combined. By the same token, it also refers to the situation where an association between two variables can consistently be inverted in each subgroup of a group when the group is partitioned. This is also commonly expressed as "an association between two variables can disappear or reverse direction when a third variable is considered." The third variable is

commonly referred to as a *hidden* or *confounding* variable. We now use the mathematical notations to describe this counter-intuitive result:

- $a/b < A/B$
- $c/d < C/D$
- $e/f < E/F$
- …
- $(a+c+e+\ldots)/(c+d+f+\ldots) > (A+B+E+\ldots)/(C+D+F+\ldots)$

where a, A, b, B, c, C, d, D, e, E, f, F, etc. are all positive numbers. The switching of inequalities above is called *Simpson's Reversal of Inequalities*, *Simpson's Paradox*, or *Reversal Paradox*. Table 1 below illustrates this reversal of inequalities. Here, Treatment #2 seems to have a better success rate in Group 1 as well as Group 2. However, Treatment #1 has a better success rate when Groups 1 and 2 are combined.

Table 1: Reversal of Success Rates between Treatments #1 and #2

	Treatment #1		Treatment #2		Results:
	Success	Total	Success	Total	
Group 1	a	b	A	B	a/b < A/B
Group 2	c	d	C	D	c/d < C/D
Groups 1 & 2	a+c	b+d	A+C	B+D	(a+c)/(b+d) > (A+C)/(B+D)

Simpson's Paradox is a popular topic in statistics and is often encountered in social-science and medical-science statistics. Many statisticians use it to caution people against hastily making causal interpretations on the mere association between two or more variables, especially when dealing with frequency data. This is because the hidden variable can influence the results, but it may not be evident in the data set collected.

Berkeley Gender Bias Case. A very famous gender bias case study illustrates Simpson's Paradox. In the Fall of 1973, there were 12,763 applicants for graduate admission to the University of California at Berkeley (UCB). Table 2 below lists the numbers of total applicants and admitted students by gender. Based upon Table 2, there appeared to be a clear (yet misleading) bias against female applicants because the male admittance rate was almost 10% higher than the female admittance rate, which may be unlikely due to chance.

Table 2: List of UCB Graduate School Admission Data in 1973

	Admitted	Rejected	Total	%Accepted
Male	3738	4704	8442	44.3%
Female	1494	2827	4321	34.6%
M & F	5232	7531	12763	41.0%

Hypothetically, if the gender bias was not a factor in the admission process, what could be the possible causes for the discrepancy shown in Table 2? Were female applicants less qualified than the male applicants? According to the UCB application data in 1973, however, there was no significant difference between the qualifications of the male and female applicants in terms of college GPAs or standardized test scores, e.g., GRE scores.

Now let us take another variable into account. Just like every university or college, the decisions on graduate admissions at UCB are made independently by each department (not at the university level). When examining the individual departments, female applicants had similar

admittance rates as male applicants—**no** department was significantly biased against female applicants. On the contrary, many departments had a small bias in favor of women. Why would women have a lower overall admittance rate when the individual departments were not found to favor men? We will use Table 3 below to explain why female applicants had a lower overall admittance rate.

Table 3 represents the admission data by various departments, gender, and results (accepted or denied). Although there were 101 different graduate schools at UCB in 1973, for simplicity, we only list the six largest departments to illustrate the reversal of inequalities. Note that these six largest departments (A through F in Table 3) accounted for 4526 of the applicants, which is more than 35% of the total applicants.

Table 3: List of Admission Data for Six Largest UCB Graduate Schools in 1973

	Male			Female		
Department	Admitted	Total	%Admitted	Admitted	Total	%Admitted
A	512	825	62%	89	108	82%
B	353	560	63%	17	25	68%
C	120	325	37%	202	593	34%
D	138	417	33%	131	375	35%
E	53	191	28%	94	393	24%
F	22	373	6%	24	341	7%
Combined (WM)	1198	2691	44.5%	557	1835	30.4%
Simple Mean			38.1%			41.7%

(Table 3 is adapted from the admission data provided in Reference 9.)

As shown by Table 3, female applicants were more likely to apply to competitive departments with higher rejection rates (such as the E and F departments), while men tended to apply to less competitive departments with higher acceptance rates (such as the A and B departments). This is why the overall acceptance rate for female applicants is lower than the male applicants. Note that the overall admission rate at the college level (in Table 2) is calculated as a weighted average by weighting each department's admission rate by the number of its applicants. Hence, it is biased towards the departments with large number of applicants.

Based upon Table 3, using simple mean—the average of each department's acceptance rates—is appropriate and no evidence was found to support the conjecture that male applicants applying to the UCB graduate schools in 1973 were more likely than female applicants to be admitted. This gender bias study is probably the most famous example of Simpson's Paradox and the data contained in Table 3 is frequently used. See References 9 and 15 for details.

If the decisions were made at the university level rather than the department level, then using the weighted mean for the overall admission rate would be appropriate.

P/E Ratio in Stock Portfolio. This is a hypothetical example provided by MCR. An investor has a stock portfolio, which consists of three stocks, A, B, and C. He wants to know whether the overall price to earning (P/E) ratio for his portfolio has met his investment target. See Table 4 below for each stock's number of shares, unit price, and P/E ratio.

Table 4: List of Unit Price, Dollars Invested, and P/E Ratios for Stocks A, B, C

Stock	Share	Unit Price	$Invested	% Inv (w)	P/E	w*P/E

A	1	5	5	2%	3.00	0.07
B	1	10	10	5%	7.00	0.33
C	10	20	200	93%	11.00	10.23
Sum:			215	100%	21.00	10.63
SM				= 21 /3 =		7.00
WM				= 10.63 / 1		10.63
GM				= (3*7*11)^(1/3) =		6.14

For a given portfolio, the decision to purchase or sell any stocks is generally made by one investor, not by several independent investors, and the evaluation should be considered for the entire portfolio. Consequently, the use of the weighted mean would be an appropriate choice for this example, and each P/E ratio should be weighted by its respective dollar amount (i.e., percent of investment).

Note that the weighted average method favors the data point large in size; the ordinary least square method (OLS) does so even more (see the detailed discussion in the following section). This is why the overall P/E ratio (10.63) is almost the same as the P/E ratio for Stock C, because 95% of the investment is in this stock. The purchases of Stocks A and B are not meaningful; the proceeds of selling these two stocks may not even cover the fees.

If A, B, and C are independent data points in a data set (true for most cases) and the P/E ratio is synonymous with the NR to T1 ratio, then using WM would not be a relevant choice as data points A and B have almost no effect in this measure.

NR COST AS A FACTOR OF T1 COST

The NR cost to T1 cost factor is an example of cost-to-cost factors. A common estimating technique is to estimate the NR cost of a program based on a factor of its first unit (T1) production cost. (T1 is often viewed as a good surrogate for the complexity factor of a program.) The factor of the NR to T1 ratios is generally derived from the average of the NR to T1 ratios across a number of similar programs. However, there are various ways to derive the average factor. For example, the factor based upon the simple average is simply the arithmetic mean of these ratios, while the factor by the weighted average is derived by weighting each ratio by its T1 cost. Therefore, for the weighted average, the larger the T1, the more influence the data point has on the resultant factor. As for the geometric average, this factor is computed as the n^{th} root of the product of these ratios. Mathematically, these average factors denoted by SM, WM, and GM, respectively, are given below:

$$SM = \frac{\sum_{i=1}^{n}(NR_i / \mathrm{T1}_i)}{n} \tag{5}$$

$$WM = \frac{\sum_{i=1}^{n} w_i (NR_i / \mathrm{T1}_i)}{\sum_{i=1}^{n}(w_i)} = \frac{\sum_{i=1}^{n} \mathrm{T1}_i (NR_i / \mathrm{T1}_i)}{\sum_{i=1}^{n}(\mathrm{T1}_i)} = \frac{\sum_{i=1}^{n}(NR_i)}{\sum_{i=1}^{n}(\mathrm{T1}_i)} \tag{6}$$

$$GM = (\prod_{i=1}^{n} NR_i / \mathrm{T1}_i)^{1/n} \tag{7}$$

Before analyzing their properties in details, we will briefly introduce three different methodologies for fitting multiplicative error models; we will first define a multiplicative error model.

Multiplicative Model. The multiplicative error model is generally stated as follows:

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta})\,\varepsilon_i \qquad \text{for i = 1, …, n} \tag{8}$$

where:

n	= sample size
Y_i	= observed cost of the ith data point, i = 1 to n
$f(\mathbf{x}_i,\boldsymbol{\beta})$	= the value of the hypothesized equation at the ith data point
$\boldsymbol{\beta}$	= vector of coefficients to be estimated by the regression equation
$\boldsymbol{x}_i$	= vector of cost driver variables at the ith data point
ε_i	= error term (assumed to be independent of the cost drivers)

Unlike the additive error model (i.e., $Y = f(\boldsymbol{x}, \beta) + \varepsilon$), the standard error of the dependent variable in Equation 8 is proportional to the level of the hypothetical equation rather than some fixed amount across the entire data range.

Log-Error Method. If the multiplicative error term (ε_i) is further assumed to follow a log-normal distribution, then the error can be measured by the following:

$$e_i = \ln(\varepsilon_i) = \ln(Y_i) - \ln(f(\mathbf{x}_i, \boldsymbol{\beta})) \tag{9}$$

where "ln" stands for nature logarithmic function. The objective is then to minimize the sum of squared e_is (i.e., $(\Sigma(\ln(\varepsilon_i))^2$). If the transformed function is linear in log space, then OLS can be applied in log space to derive a solution for $\boldsymbol{\beta}$. If not, we need to apply the non-linear regression technique to derive a solution.

Although a least squares optimization in log space produces an unbiased estimator in log space, the estimator is biased low when transformed back to unit space (see References 4, 11, and 16). However, the magnitude of the bias can be corrected with a simple factor if the errors are distributed normally in log space (see References 4 and 11). Because of this shortcoming, the MUPE method is recommended for modeling multiplicative errors directly *in unit space* to eliminate the bias.

MUPE Method. The general specification for a MUPE model is the same as given above (Equation 8), except that the error term is assumed to have a mean of 1 and variance σ^2. Based upon this assumption of a multiplicative model, a generalized error term is defined by:

$$e_i = \frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})}{f(\mathbf{x}_i, \boldsymbol{\beta})} \tag{10}$$

where e_i now has a mean of 0 and variance σ^2.

The difference between this percentage error (Equation 10) and the traditional percentage error is in the denominator where MUPE uses predicted cost, not actual cost, as the baseline. The optimization objective is to find the coefficient vector $\boldsymbol{\beta}$ that minimizes the sum of squared e_is:

$$\text{Minimize} \quad \sum_{i=1}^{n}\left(\frac{y_i - f(\mathbf{x}_i,\boldsymbol{\beta})}{f(\mathbf{x}_i,\boldsymbol{\beta})}\right)^2 = \sum_{i=1}^{n} e_i^{\,2} \tag{11}$$

The solution for Equation 11 is biased high if it is solved directly in a single pass (see Reference 10). To eliminate this bias, the MUPE method solves for the function ($f(\mathbf{x},\boldsymbol{\beta})$) in the numerator separately from the function in the denominator through an iterative process.

$$\text{Minimize} \quad \sum_{i=1}^{n}\left(\frac{y_i - f(\mathbf{x}_i,\boldsymbol{\beta}_k)}{f(\mathbf{x}_i,\boldsymbol{\beta}_{k-1})}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_i - f_k(\mathbf{x}_i)}{f_{k-1}(\mathbf{x}_i)}\right)^2 \tag{12}$$

where k is the iteration number and the other terms are as defined previously.

The weighting factor of each residual in the current iteration is equal to the reciprocal of the predicted value from the previous iteration. Since the denominator in Equation (12) is kept fixed throughout the iteration process, the MUPE technique turns out to be a weighted least squares (WLS) with an additive error. The final solution is derived when the change in the estimated coefficients ($\boldsymbol{\beta}$ vector) between the current iteration and the previous iteration is within the analyst-specified tolerance limit. This optimization technique (Equation 12) is commonly referred to as Iteratively Reweighted Least Squares (IRLS; see References 12 and 13). The corresponding standard error of estimate for the MUPE CER is commonly termed multiplicative error or standard percent error (SPE):

$$SPE = \sqrt{\sum_{i=1}^{n}((y_i - \hat{y}_i)/\hat{y}_i)^2/(n-p)} \tag{13}$$

Note that $\hat{y}_i$ is the predicted value in unit space for the ith data point and p is the total number of estimated coefficients. The MUPE CER provides consistent estimates of the parameters and has zero proportional error for all points in the data set. See Reference 7 or 10 for detailed descriptions of the MUPE method.

ZMPE Method. There is another alternative method to reduce the positive proportional error when minimizing Equation 11 directly. Mathematically, it is stated as follows:

$$\text{Minimize} \quad \sum_{i=1}^{n}\left(\frac{y_i - f(\mathbf{x}_i,\boldsymbol{\beta})}{f(\mathbf{x}_i,\boldsymbol{\beta})}\right)^2$$

$$\text{Subject to} \quad \sum_{i=1}^{n}\frac{y_i - f(\mathbf{x}_i,\boldsymbol{\beta})}{f(\mathbf{x}_i,\boldsymbol{\beta})} = 0 \tag{14}$$

This alternative method (Equation 14) is a "constrained" minimization process. It is commonly referred to as the Zero-Percentage Bias method under MPE, i.e., the ZPB/MPE or ZMPE method by Book and Lao, 1999 (see Reference 8). Both MUPE and ZMPE CERs have zero proportional error for all the points in the data set.

ANALYZING A FACTOR CER USING DIFFERENT METHODS

We now analyze a simple factor CER (e.g., NR = β*T1) using various methods, beginning with the MUPE and ZMPE methods. ("NR = β*T1" is used for illustration purposes.)

Simple Mean for Factor CER by MUPE and ZMPE Methods. Given the following factor equation with a multiplicative error term:

$$Y_i = \beta * X_i * \varepsilon_i \qquad \text{for } i = 1, \ldots, n \tag{15}$$

where:

β = the factor (to be estimated by the regression equation)
n = sample size
Y_i = the observed NR cost of the i^{th} data point, $i = 1$ to n
X_i = the T1 cost of the i^{th} data point
ε_i = the error term (with a mean of 1 and a variance σ^2)

Unlike the ordinary least squares (OLS), every data point is treated equally (percentage-wise) regardless of the size of the program under the assumption of the multiplicative error term.

It can be shown within two iterations that the MUPE solution for the above factor CER is a simple average of the NR to T1 ratios, i.e., Equation 5. It is even more straightforward to derive the factor "β" of Equation 15 using the ZMPE method due to the following constraint:

$$\sum_{i=1}^{n} \frac{Y_i - \beta X_i}{\beta X_i} = \left(\sum_{i=1}^{n} \frac{Y_i}{\beta X_i} \right) - n = 0$$

Therefore, the ZMPE solution for this factor CER is also a simple average of these ratios, i.e.,

$$\hat{\beta}_{(ZMPE)} = \hat{\beta}_{(MUPE)} = \frac{\sum_{i=1}^{n} (Y_i / X_i)}{n} = \frac{\sum_{i=1}^{n} (NR_i / T1_i)}{n} = \hat{\beta}_{(sm)} \tag{16}$$

Solve Equation 15 Using WLS. Since the MUPE technique is also a WLS, we can derive the same solution using the WLS method under an additive error model:

$$Y_i = \beta * X_i + \delta_i \qquad \text{for } i = 1, \ldots, n \tag{17}$$

Here, the mean of the new error term δ_i is zero for all observations. Since Equations 15 and 17 represent the same model, the variance of Y_i is equal to the variance of the new additive error term δ_i, and both are proportional to the square of its corresponding X_i:

$$V(Y_i) = V(\delta_i) = \sigma^2 \beta^2 X_i^2 \quad \text{for } i = 1, \ldots, n \tag{18}$$

Hence, we can choose the weighting factor for the i^{th} observation to be proportional to the reciprocal of its variance under WLS, e.g.,

$$w_i = 1 / X_i^2 \quad (e.g., w_i = 1 / T1_i^2)$$

Given the above weighting factor, the WLS solution of the factor CER (Equation 17) is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} (1/X_i^2)(X_i Y_i)}{\sum_{i=1}^{n} (1/X_i^2)(X_i^2)} = \frac{\sum_{i=1}^{n} (Y_i / X_i)}{n} = \frac{\sum_{i=1}^{n} (NR_i / T1_i)}{n} = \hat{\beta}_{(sm)} \tag{19}$$

It is exactly the same solution derived by both the MUPE and ZMPE methods, which is the simple average of these NR to T1 ratios.

Based upon the WLS method, the variance of $\hat{\beta}_{(sm)}$ is given below:

$$V(\hat{\beta}_{(sm)}) = \frac{\sigma^2\beta^2}{\sum_{i=1}^{n} w_i X_i^2} = \frac{\sigma'^2}{\sum_{i=1}^{n} (1/X_i^2) X_i^2} = \frac{\sigma'^2}{n} \tag{20}$$

(σ*β is denoted by σ'.)

Note that the variance of this factor $\hat{\beta}_{(sm)}$ does not depend upon the driver variable. Some analysts use the standard error of the factor (i.e., square root of Equation 20) to quantify the error of the estimate, but prediction interval should be the proper measure for cost uncertainty analysis.

Weighted Mean for Factor CER by WLS Method. When examining the derivation of the WLS solution, we can deduce the weighing factor for deriving the weighted mean as the common factor for the factor CER. If the variance of the dependent variable Y is proportional to its respective X value (i.e., $V(Y) = s^2X$), then its weighting factor can be given by

$$w_i = 1/X_i \quad (e.g., w_i = 1/T1_i)$$

(Note: We use s^2 to denote the constant term in the variance of Y.) With this weighting factor, the WLS solution for the factor β in the factor CER (e.g., NR = β∗T1) is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} (1/X_i)(X_i Y_i)}{\sum_{i=1}^{n} (1/X_i)(X_i^2)} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} NR_i}{\sum_{i=1}^{n} T1_i} = \hat{\beta}_{(wm)} \tag{21}$$

The above solution (Equation 21) is exactly the weighted mean of the NR to T1 ratios; each ratio is weighted by its respective T1 cost (see Equation 6).

Similarly, the variance of $\hat{\beta}_{(wm)}$ is given below according to the WLS analysis:

$$V(\hat{\beta}_{(wm)}) = \frac{s^2}{\sum_{i=1}^{n} w_i X_i^2} = \frac{s^2}{\sum_{i=1}^{n} (1/X_i) X_i^2} = \frac{s^2}{\sum_{i=1}^{n} X_i} \tag{22}$$

Although Equation 22 can be used to analyze the uncertainty of the coefficient in the factor CER, the prediction interval should be the proper measure for cost uncertainty analysis.

Obviously, when all the weighting factors are assumed to be the same, we would derive the OLS solution for a simple factor CER:

$$\hat{\beta}_{(ols)} = \frac{\sum_{i=1}^{n} w_i (X_i Y_i)}{\sum_{i=1}^{n} w_i (X_i^2)} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} (NR_i * T1_i)}{\sum_{i=1}^{n} (T1_i^2)} = \frac{\sum_{i=1}^{n} T1_i^2 (NR_i/T1_i)}{\sum_{i=1}^{n} T1_i^2} \tag{23}$$

And the variance of $\hat{\beta}_{(ols)}$ is given below by WLS:

$$V(\hat{\beta}_{(ols)}) = \frac{\sigma^2}{\sum_{i=1}^{n} w_i X_i^2} = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2} \tag{24}$$

Geometric Mean for Factor CER by Log-Error Method. Given the same factor CER with a multiplicative error term:

$$Y_i = \beta * X_i * \varepsilon_i \qquad \text{for } i = 1, \ldots, n \tag{25}$$

Now the error term ε_i is assumed to follow a lognormal distribution with a mean of 0 and a variance σ^2 in log space, i.e., $\varepsilon_i \sim LN(0, \sigma^2)$ for i = 1, …, n.

The respective objective function is given below when the fit is done in log space:

$$F = \sum_{i=1}^{n} (\ln(Y_i) - \ln(\beta X_i))^2 = \sum_{i=1}^{n} (\ln(Y_i) - \ln(\beta) - \ln(X_i))^2 \tag{26}$$

If we take the partial derivative of F with respect to β and set it to zero, we can derive the following equation:

$$\sum_{i=1}^{n} (\ln(Y_i) - \ln\beta - \ln(X_i))/\beta = 0 \tag{27}$$

Therefore, the solution for β is then given by

$$\begin{aligned} \hat{\beta}_{(\text{log-err})} &= \exp(\sum_{i=1}^{n} (\ln(Y_i) - \ln(X_i))/n) = \exp\left(\ln(\prod_{i=1}^{n} Y_i/X_i)/n\right) \\ &= \exp\left(\ln(\prod_{i=1}^{n} Y_i/X_i)^{1/n}\right) = (\prod_{i=1}^{n} Y_i/X_i)^{1/n} \\ &= \hat{\beta}_{(gm)} \end{aligned} \tag{28}$$

As shown by Equation 28, the geometric mean of the NR to T1 ratios is the solution derived by the log-error method, the so-called log-linear model. As shown in References 4, 11 and 16, this solution is biased low so we have to apply a correction factor to adjust for the mean in unit space.

In summary, Table 5 lists the solutions for the factor equation using various methods and their respective weighting factors under the WLS method. The last column in Table 5 reports the weighting factors for the NR to T1 ratios; these weighting factors are different from the weighting factors for deriving the factor in the factor CER.

Table 5: List of Solutions and Weighting factors by Various Methods

Method	**Solution for NR = βT1**	WF for Factor **CER**	WF for NR/T1 **Ratio**
MUPE/ZMPE	$\hat{\beta}_{(sm)} = \sum_i (NR_i/T1_i)/n$	$1/(T1)^2$	1
WM by WLS	$\hat{\beta}_{(wm)} = \sum_i (NR_i)/\sum_i (T1_i)$	1/(T1)	T1
OLS	$\hat{\beta}_{(ols)} = \sum_i (NR_i * T1_i)/\sum_i (T1_i^2)$	1	$T1^2$
Log Error	$\hat{\beta}_{(gm)} = (\prod_{i=1}^{n} (NR_i/T1_i))^{1/n}$	(fit done in log space)	

As shown by the last column in Table 5, the simple mean (derived by both the MUPE and ZMPE methods) treats all the programs equally; it does not favor the small size programs. As for the weighted mean (derived by the WLS method), it favors programs large in size, as each ratio is weighted by its T1 cost. The OLS factor also favors programs large in size because each ratio is weighted by the square of its T1 cost. (OLS is a special case of WLS when all weighting factors are the same.) Since the geometric mean is derived in log space through transformation, it does not have weighting factors listed in Table 5. Note that Table 5 is applicable to any factor CER and we use the NR to T1 factor CER as an illustrative example.

A REALISTIC EXAMPLE

We will use the following realistic example to point out some common mistakes that may occur when calculating simple, weighted, and geometric means, and the pitfalls of using small samples to develop CERs.

Factor CER. Here is a realistic example relating the NR cost to its T1 cost by a common factor, i.e., NR = β*T1. This example contains 12 satellite programs, denoted by Program A through Program L. The NR and T1 costs of these programs were extracted from the Unmanned Space Vehicle Cost Model, Eighth Edition (USCM8) database at a particular suite level. Due to the proprietary nature of the database, all the costs and program names were modified and scrambled. See Table 6 below for a listing of the "fictitious" data set, the NR to T1 ratios, and the three means (SM, WM, and GM).

Table 6: NR and T1 Cost Data Set

Programs	NR Cost $K	T1 Cost $K	NR/T1 Ratio
A	60.53	24.18	2.50
B	320.08	148.94	2.15
C	7.41	292.01	0.03
D	2.09	325.78	0.01
E	139.01	70.95	1.96
F	-	~~195.41~~	
G	101.37	35.90	2.82
H	0.98	56.14	0.02
I	222.61	65.05	3.42
J	142.75	65.41	2.18
K	115.89	80.08	1.45
L	11.51	203.97	0.06
Sum:	1,124.21	1,368.39	16.59
SM	= 16.59 / 11	=	1.51
WM	= 1124.21 / 1368.39	=	0.82
GM	= (2.5*2.15*…*0.06)^(1/11)	=	0.41

(Note: the T1 cost of Program F is not included in the computation of means because this program does not have any NR cost.)

Given three significantly different means (SM, WM, and GM), which one should we use to predict the NR cost as a factor of its T1 cost? As shown in Table 6, the factor based upon the simple mean (i.e., 1.51) is the largest, while the factor based upon the geometric mean (0.41) is the smallest. Some analysts have strong concerns about using simple mean as it is calculated as the average of the individual ratios. They argue that we should never average percents, ratios, or averages to create CERs, factors, or models. Besides, the simple mean (1.51) in this example is almost twice as large as the weighted mean (0.82). Consequently, the point estimate based upon the weighted mean will be significantly less than the simple mean when using the factor CER. Based upon this information, many analysts would then suggest using the weighted mean to estimate the NR cost.

Upon careful examination, we noticed that several follow-on programs, such as Programs C, D, and H, are used to compute the NR to T1 ratios. Clearly, this data set is not homogeneous because the follow-on programs do not have meaningful NR costs. Furthermore, Program L has data/programmatic issues (not a full design effort) and hence should be excluded from the computation.

When a method is applied to a mixed bag of programs, the analysis results will be misleading and inaccurate. Using a weighted mean as the common factor is not a remedy for this problem—in fact, it may be more misleading. As shown by Equation 6, every simple ratio is weighted by its respective T1 cost when calculating the weighted mean; the larger the T1, the more influence the data point has on the resultant factor. Therefore, Programs C, D, and L would have a lot more influence on the factor than the remaining programs. Hence, the weighted mean (0.82) is negatively skewed due to the presence of these follow-on programs.

$$WM = \Sigma\{(NR_i/T1_i)*T1_i\}/\Sigma(T1_i)$$

$$= (2.5*24.18+2.15*148.94+.03*292.01+\ldots+0.06*203.97)/(24.18+148.94+\ldots+203.97)$$

$$= 1124.21 / 1368.39 = 0.82$$

When all the follow-on programs and Program L are removed from the data set, the simple mean (average of the ratios), weighted mean (ratio of the averages), and geometric mean are very close to one another. See the updated numbers in Table 7 below for details. The bottom line: a homogeneous data set is essential for data analysis. If significant disparities are found in these means, this may be an indication that errors may occur in the computation or the data set is simply not homogeneous.

Table 7: NR and T1 Cost W/O Follow-on Programs

Programs	NR Cost $K	T1 Cost $K	NR/T1 Ratio
A	60.53	24.18	2.50
B	320.08	148.94	2.15
E	139.01	70.95	1.96
G	101.37	35.90	2.82
I	222.61	65.05	3.42
J	142.75	65.41	2.18
K	115.89	80.08	1.45
Sum:	1,102.23	490.51	16.49
SM	= 16.49 / 7	=	2.36
WM	= 1102.23 / 490.51	=	2.25
GM	= (2.5*2.15*1.96*…)^(1/7)	=	2.28

Power-Form CER. If we want to use a power-form CER to better explain the variation in cost, we can generate the following MUPE CER for this data set:

$$NR = 4.75 * T1 \wedge 0.8283 \quad (29)$$

This CER seems even better than the factor CER because (1) the exponent on T1 is reasonable by engineer's logic and (2) its standard percent error (SPE) is 28% and all the fit and predictive

measures are fairly tight. However, use this CER with caution as its degrees of freedom (DF) is very small. (See Equation 13 for the definition of SPE.)

Be Wary of Small DF. A CER developed upon a very small sample (e.g., less than 10) is vulnerable because it can be changed significantly when new data points become available or an update is made to an existing data point. In fact, the decimal points in the costs of Program G were found to be off by one place. The NR and T1 costs for Program G should be 1013.7, and 359, respectively. With these two updated costs for Program G, the power-form CER is now given by

NR = 2.12 * T1 ^ 1.02 (SPE = 30%, DF = 5) (30)

This updated CER is very different from the previous power-form CER due to this update. Since the exponent on T1 is so close to one, the exponent is fixed at one to save DF for small samples and the resultant CER is given by

NR = 2.36 * T1 (SPE = 27%, DF = 6) (31)

If we use 1/T1 as the weighting factor, the corresponding CER is given by

NR = 2.48 * T1 (SPE = 27%, DF = 6) (32)

This factor, 2.48, is also the result when calculating the weighted mean of the NR to T1 ratios. To save DF and treat all programs equally, we would suggest using the simple mean (2.36) as given in Equation 31 to estimate the NR cost.

CONCLUSIONS

Simpson's Paradox refers to the reversal of an association between two variables due to the impact of a third variable. This Paradox occurs when the association between two variables is actually due to the fact that each is strongly related to the third variable. There are many real-life examples on the Internet of Simpson's Paradox. These examples indicate that weighted means are the appropriate measures for some cases when the data points are correlated or under certain constraints, while simple means should be used for others. These examples can help to disprove the myth that using simple means or weighted means is always wrong mathematically. Note that SM is more relevant than WM for cost analysis because data points are generally independent.

Alternatively, we can use WLS to ease the concerns of using simple and weighted averages. By WLS, we can derive various solutions (SM, WM, etc.) for NR vs. T1 factor CERs. So it is proven that simple, as well as weighted, averages can be a common factor in factor CERs. Neither is wrong mathematically. For ease of reference, Table 5 is repeated below showing various factors regarding NR vs. T1 factor CERs.

Method	**Solution for NR = βT1**	WF for Factor **CER**	WF for NR/T1 **Ratio**
MUPE/ZMPE	$\hat{\beta}_{(sm)} = \sum_i (NR_i / \mathrm{T1}_i)/n$	$1/(\mathrm{T1})^2$	1
WM by WLS	$\hat{\beta}_{(wm)} = \sum_i (NR_i)/\sum_i (\mathrm{T1}_i)$	1/(T1)	T1
OLS	$\hat{\beta}_{(ols)} = \sum_i (NR_i * \mathrm{T1}_i)/\sum_i (\mathrm{T1}_i^2)$	1	$\mathrm{T1}^2$
Log Error	$\hat{\beta}_{(gm)} = (\prod_{i=1}^{n} (NR_i / \mathrm{T1}_i))^{1/n}$	(fit done in log space)	

As shown by the last column in the table, the simple mean (derived by the MUPE and ZMPE methods) treats all the programs equally; it does not favor the small size programs. However, the weighted mean favors the programs large in size as each ratio is weighted by its T1 cost. The OLS factor also favors the programs large in size as each ratio is weighted by the square of its T1 cost. Since the geometric mean is derived in log space through transformation, it does not have weighting factors listed in Table 5. As mentioned above, GM is always biased low and it should be adjusted to reflect the mean in unit space. Note that Table 5 is applicable to any factor CER and we simply use the NR to T1 factor CER as an illustrative example.

Uncertainty can be specified for all three means (SW, WM, and GM). According to WLS, there are estimating errors for simple and weighted means, as well as the mean derived by the OLS method. There is also standard error associated with the geometric mean when the curve fit is done in log space. Although some analysts use *the standard error of the mean* to quantify the uncertainty when estimating the NR cost as a factor of the T1 cost, the prediction interval is the proper measure for cost uncertainty analysis.

A homogeneous data set is essential for data analysis. If a method is applied to a mixed bag of observations, the analysis results will be misleading and inaccurate. As shown by the realistic example in the paper, using a weighted mean as the common factor is not a remedy for this problem—in fact, it may be more misleading. If significant disparities are found in these means, it may be an indication that errors may occur in the computation or the data set is simply not homogeneous.

Apply appropriate methods for specific assumptions and goals. Mathematical and statistical models are the tools to help us meet the requirements and achieve the estimating goals. We should select an appropriate analysis approach to solve a particular problem based upon our specific assumptions and targets. For the NR vs. T1 factor equation, if the estimating goal is to find most similar programs and use analogy, then we should do so and should not use simple, weighted, or geometric mean of all programs in the data set for estimating the NR cost from the factor CER. (A method should be chosen based upon specific goal and requirements.)

Develop parametric CERs whenever possible. If we have enough data points available (e.g., 10 or more), we should consider developing a NR equation using T1 and/or other explanatory variables. (We do not favor using simple or weighted factors over general-purpose CERs.) However, beware of the pitfalls of using CERs with small DF (say five or less). We cannot place too much credence on a CER with just few degrees of freedom left because a CER built upon small DF is unstable and can be changed significantly when new programs become available.

Avoid using cost as an independent variable in a CER. An independent variable based upon cost is not an ordinary independent variable. A cost driver is always subject to errors. In other words, a cost independent variable cannot be observed without error, which violates the basic assumption of regression analysis. If the cost is further estimated by CERs, analogies, or even expert opinions, then we may become more uncertain about the uncertainties associated with the cost driver, which is certainly less desirable. In conclusion, although the T1 cost may be a good surrogate for the NR cost, we should use cost-dependent CERs with caution because (1) it can be problematic for uncertainty analysis as discussed in References 2 and 5, (2) there are always more uncertainties inherited in the cost-dependent drivers than the technical drivers, and (3) the degrees of freedom for cost-dependent CERs can be over-estimated, which may affect

cost uncertainty analysis. The bottom line: develop hardware design-based CERs whenever possible and avoid using cost-dependent CERs.

REFERENCES

1. Hulkower, Neal D., "Numeracy for Cost Analysts; Doing the Right Math, Getting the Math Right," ISPA/SCEA workshop, San Pedro, CA, 15 September 2009.
2. Hu, S. "Comparing Methods for Deriving Cost Dependent CERs," 2009 ISPA/SCEA Joint Annual Conference, St. Louis, Missouri, 2-5 June 2009.
3. Hu, S. and A. Smith, "Why ZMPE When You Can MUPE," 6th Joint Annual ISPA/SCEA International Conference, New Orleans, LA, 12-15 June 2007.
4. 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.
5. Covert, R. P. and Anderson, T. P., "Regression of Cost Dependent CERs," the Aerospace Corporation, Space Systems Cost Analysis Group (SSCAG) meeting, February 2002.
6. Nguyen, P., N. Lozzi, et al., "Unmanned Space Vehicle Cost Model, Eighth Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, October 2001.
7. 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.
8. 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.
9. Freedman, D., R. Pisani, and R. Purves, "Statistics, 3rd Edition," New York: WW Norton & Company, 1998.
10. Hu, S. and A. R. Sjovold, "Multiplicative Error Regression Techniques," 62nd MORS Symposium, Colorado Springs, Colorado, 7-9 June 1994.
11. Hu, S. and A. R. Sjovold, "Error Corrections for Unbiased Log-Linear Least Square Estimates," TR-006/2, March 1989.
12. Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons, 1989, pages 37, 46, 86-88.
13. Weisberg, S., Applied Linear Regression, 2nd Edition," New York: John Wiley & Sons, 1985, pages 87-88.
14. Duan, N., "Smearing Estimate: A Nonparametric Retransformation Method," Journal of the American Statistical Association, Vol. 78, Sep 1983, No. 383, pp. 605-610.
15. Bickel, P. J., E. A. Hammel, and J.W. O'Connell, "Sex Bias in Graduate Admissions: Data from Berkeley," Science, 187 (1975), pages 398-404.
16. Goldberger, A. S., "The Interpretation and Estimation of Cobb-Douglas Functions," Econometrica, Vol. 35, July-Oct 1968, pages 464-472.