

**MULTIPLICATIVE UNBIASED
REGRESSION TECHNIQUE (MURT)**

**62nd MORS Symposium
Colorado Springs, Colorado
June 7-9, 1994**

**Tecolote Research, Inc.
5266 Hollister Avenue, #301
Santa Barbara, CA 93111
(805) 964-6963**

ABSTRACT

MULTIPLICATIVE ERROR REGRESSION TECHNIQUE

by Dr. Shu Ping Hu
Arve R. Sjøvold
Tecolote Research, Inc.

A new Multiplicative Unbiased Regression Technique (MURT) has been developed to model multiplicative error in least-squares regressions. Multiplicative error is an appropriate assumption when modeling systems in which the dependent random variable ranges over more than an order of magnitude and errors in the dependent variable are believed to be proportional to the level of the variable. Previous methods to model multiplicative error have usually depended on log-transforms, either log-linear regressions or non-linear regressions of the log-transformed dependent variable. Unfortunately, log-transforms involve transformation bias such that the unit space equation is not unbiased. MURT involves an iterative, weighted least squares regression that is shown to provide unbiased percentage error regression results while modeling a multiplicative error. This represents a significant addition to the regression tool box for cost and systems analysts.

MULTIPLICATIVE ERROR REGRESSION TECHNIQUE

INTRODUCTION

Regression analysis is a powerful tool in scientific investigations. Once an hypothesis has been developed and appropriate data have been collected, the analyst can resort to regression analysis to test the validity of his hypothesis. The development of cost estimating relationships (CERs) in cost analysis represents one such body of scientific investigation. The CER to be developed attempts to establish the relationship between the cost of a certain class of equipment to parameters describing its performance or physical characteristics. Most often the costs of the equipment items in the data set will range over several orders of magnitude. It is this feature of the cost analysis problem that gives rise to the notion of a multiplicative error, by which we mean that errors in cost are in proportion to cost.

Simple linear regression involves the process most often referred to as “ordinary least squares,” or OLS. When using OLS, the assumption is usually made that the distribution of the error in the dependent variable (i.e., cost) is identical throughout the space. For example, a standard error of a distribution that is identical throughout the range of the data may be, say \$10, such that it would represent only 1% on a \$1000 item, but 100% on a \$10 item. Clearly, these quite different percentage errors do not represent well what we believe to be the case concerning costs. It is more appropriate in such a case to assume that the percentage errors of the item estimates will be identically distributed throughout the space. This is what a multiplicative error is supposed to represent.

When the absolute error (i.e., the dollar error) is identically distributed throughout, the error is said to be “additive.” The term OLS refers to the particular mathematical optimization technique used in regression analysis by which the sum of the squares of the estimating errors is minimized. It is the equivalent of visually fitting a curve through a plot of the data so as to balance the data points on both sides of the curve. The minimization is said to be “ordinary least squares” if the function is linear and the error is additive.

For most cost analyses, it will be necessary to assume a multiplicative error just due to the range of the cost data. In the past, several optimization techniques have been used to model

multiplicative error in studies. One is the use of a log-transform of the cost data. The transformation may be log-linear, whereby the dependent and independent variables are transformed to logarithms such that the transformed function to be fit is linear in log-space. When the transformed function is linear in log-space, OLS can be applied to achieve a multiplicative error. If the equation form is not log-linear, the log-transform can still be used to model a multiplicative error, but the optimization will be non-linear least squares. A second technique to model a multiplicative error is by weighted least squares. The sense of a multiplicative error is created by dividing the normal additive error by the value of the observed data at each point in the data set. The mathematical optimization is achieved by weighting each data point (as well as the residual) by the inverse of the observed value while performing ordinary least squares, or just least squares if the function is not linear in unit space (this requires a non-linear optimization calculation). Mathematically, the objective is described as

$$\text{Minimize } \sum [(y_i - \hat{y}_i)/y_i]^2 \quad (1)$$

where

y_i = the i^{th} observation

\hat{y}_i = the estimate of the i^{th} point

Both log-space and weighted least squares have some undesirable properties. Even though a least-squares optimization in log-space produces an unbiased estimator in log-space, on transformation back to unit space the estimator is no longer unbiased. Weighted least squares can also be shown to produce a biased estimator. In both cases, the magnitude of the bias is proportional to the variance about the estimator. However, in the case of the log-space least squares, the magnitude of the bias can be corrected with a simple factor if one is willing to assume that the errors are distributed normally in log-space (ordinarily a reasonable assumption.)

Because of these shortcomings, there has been a need to discover a method for modeling multiplicative error with least squares optimization which produces unbiased estimators. This need was recognized in the cost analysis community in 1990 by Dr. S. A. Book and Mr. P. H. Young of the Aerospace Corporation.⁽¹⁾

Book/Young hypothesized that a multiplicative error can be best defined by dividing the ordinary unit space error by the predicted value of the cost function in contrast to weighted least

squares, where the unit space error is divided by the observed value of cost for each data point. Their hypothesis is succinctly stated by

$$y_i = f(\underline{x}_i, \underline{a}) * \epsilon(1, \sigma^2) \quad (2)$$

where

- y_i = value of the i th data point
- $f(\underline{x}_i, \underline{a})$ = the function to be fit
- \underline{x}_i = a vector of the independent variables @ each i
- \underline{a} = a vector of the coefficients to be calibrated by the regression
- $\epsilon(1, \sigma^2)$ = is a multiplicative error distribution with mean of 1 and variance σ^2

For mathematical calculation, the error term can be redefined as

$$\epsilon'_i = (y_i - f(\underline{x}_i, \underline{a})) / f(\underline{x}_i, \underline{a}), \quad (3)$$

which now has a mean of zero with variance σ^2 . The optimization objective is to find coefficient vector \underline{a} which will minimize the sum of ϵ'^2_i .

Book/Young developed an optimization method, referred to as GERM*, which performs a non-linear calculation to find the \underline{a} which minimizes the sum of squares.⁽¹⁾ It can be stated mathematically as

$$\text{Minimize } \Sigma [(y_i - f_i) / f_i]^2 \quad (4)$$

where

- y_i is as defined before, and
- f_i is short for $f(\underline{x}_i, \underline{a})$.

In effect, the optimization finds the vector set \underline{a} , which determines the value of the function to be used in the numerator while simultaneously providing the value used as the weighting parameter in

*GERM stands for General Error Regression Model, a term coined by Book/Young.⁽¹⁾

the denominator. The optimization differs from the weighted least squares noted above in that the predicted value of the function appears as the weighting parameter rather than the observed value for each data point. Intuitively, this should represent a correct weighting, given that the objective of the optimization is to find an unbiased predictor function.

However, because the GERM optimization involves the simultaneous determination of the function's coefficients for both the calculation of the residual and the weighting factor, the resultant function is biased high. This is seen intuitively by noting that the ϵ_i 's are minimized if the $f(\underline{x}_i, \underline{a})$ (to be used in the numerator and denominator) is made higher than if it were established independently. That fact has been confirmed by theoretical investigations on simple cases, such as a simple univariate or a simple factor equation, and was further confirmed numerically on real cost data drawn from the data base used in the development of the Unmanned Spacecraft Cost Model, 7th edition (USCM7).⁽²⁾

The first inclination on discovering the bias in the GERM optimization was to cast doubt on the entire method, both the error term formulation and the optimization calculations. However, further examination has shown that the error term formulation is a correct one, but that a different method of optimization has to be used. The most obvious approach is to uncouple the simultaneous aspect of the optimization calculation by employing an iterative calculation to arrive at $f(\underline{x}_i, \underline{a})$. This approach has now been investigated and has been shown to produce a percentage unbiased estimator. The optimization calculation can be stated mathematically as

$$\text{Minimize } \sum [(y_i - f_j(\underline{x}_i)) / f_{j-1}(\underline{x}_i)]^2, \quad (5)$$

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

This new optimization technique is termed **Multiplicative Unbiased Regression Technique (MURT)**. It allows the analyst to model a multiplicative error without introducing bias to the resultant regression equation, and as such represents a significant addition to the regression tool box that should be included in any general-purpose statistical package.

The following discussion presents the theoretical derivations used on simple cases to demonstrate the bias in the GERM optimization and the empirical results established for more complex

cases, as well as simulation examples that confirm the theoretical findings. The examples also demonstrate that the iterative approach does indeed produce an unbiased estimator. Furthermore, in the course of this investigation, a theoretical correction factor has been derived that can eliminate, approximately, the bias in the simultaneous optimization. This would only be of interest in an application where a large body of work was accomplished using the simultaneous optimization. There are other considerations in the adoption of an optimization technique involving the proper use of statistical measures, and these are also discussed.

PROBLEM STATEMENT AND DEFINITIONS

The objective of the research was to find a regression technique to model a multiplicative error term without bias. We have already described above what is meant by a multiplicative error term and its importance to cost analysis. Figure 1 graphically represents the notion of a multiplicative error by portraying error bars along the function with the span of the error bar proportional to the value of the function at any point. For cost, this is the equivalent of saying “errors in cost are in proportion to cost.”

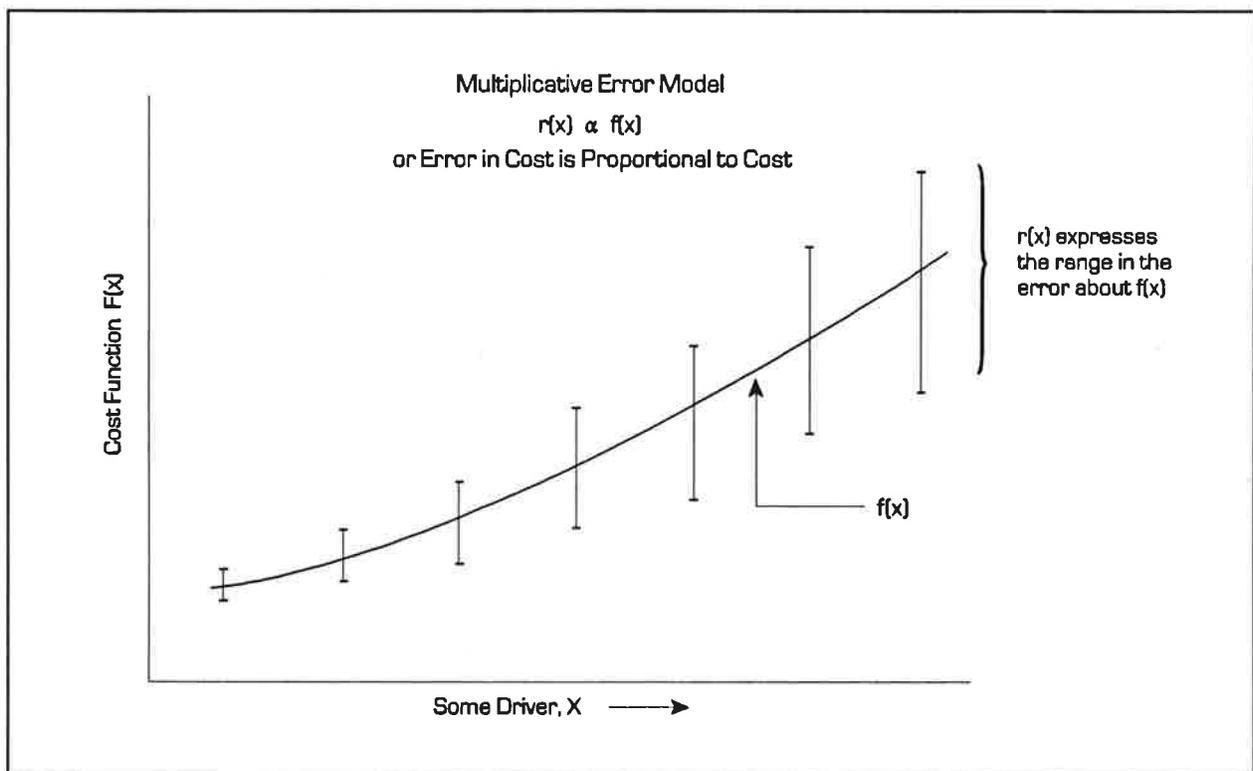


Figure 1

We will also use the term “bias” often in referring to the quality of the estimator, throughout the discussion, and so it is important to define what we mean by this term. In general, we define an estimator to be unbiased if it is an expected value estimator throughout the range of the estimating function. Figure 2 portrays this notion of bias. In the figure, the true function is shown with a solid line along with the proportional error bands centered on the line. (The bell-shaped curve is meant simply to show a normal distribution, although this is not necessary to demonstrate the notion of bias.) The figure also shows eight sample points that one might expect to draw as a random sample from the distributions about this function. The lower dashed line represents what best-fit regression line might represent this sample. Even though this sample line does not fall exactly on the true line, it would be considered an expected value estimator based on the sample drawn, and therefore is unbiased. The upper dashed line is clearly not consistent with the sample, and we would say that it is biased. We demonstrate unbiased by showing that an estimator provides an expected value estimate throughout its range. Further in the discussion, we will show some specific properties that can be used to demonstrate whether or not an estimator is unbiased.

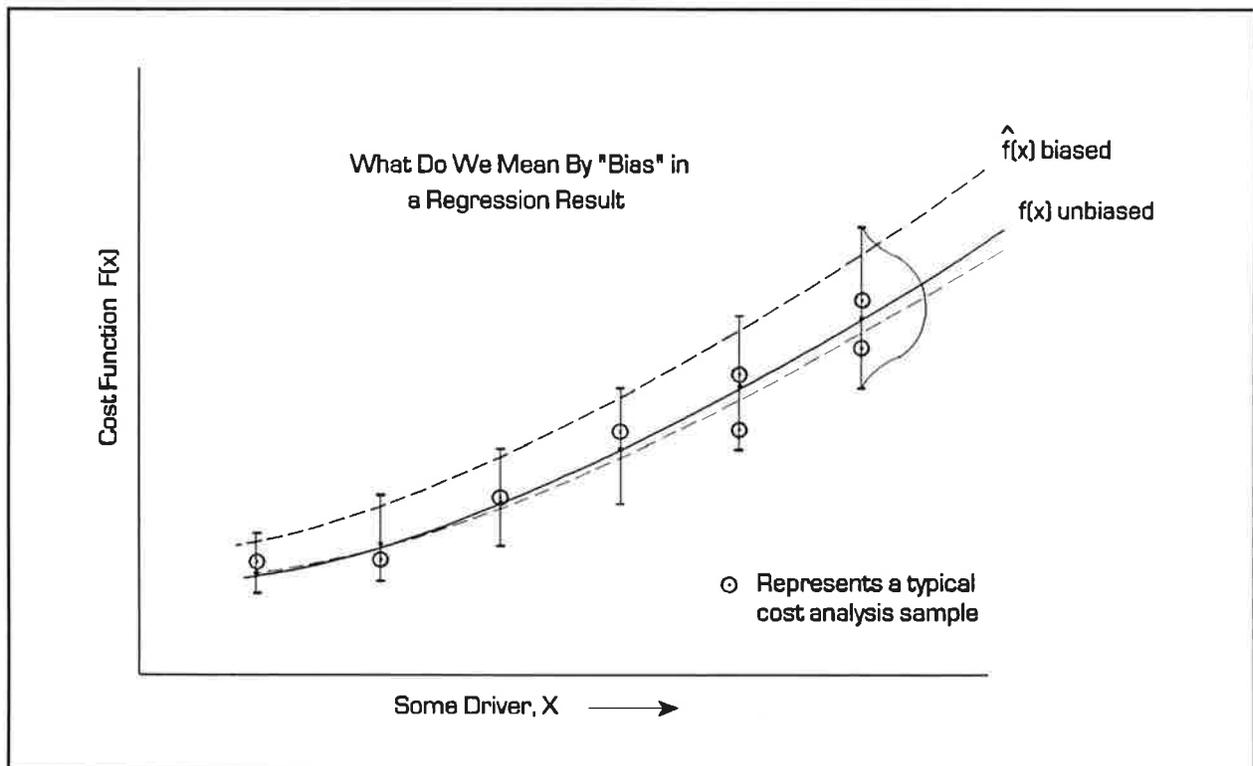


Figure 2

CANDIDATE MODEL FORMS TO EXPRESS A MULTIPLICATIVE ERROR

There are at least three distinct error formulations that have been posited and used to express a multiplicative error term for regression analysis. One is the “log-error,” which expresses error residuals by

$$\text{ERROR}_i = \log y_i - \log f(\underline{x}_i) \quad (6)$$

Two additional forms to express a multiplicative error are developed through the use of weighted residuals. One form weights residuals by the value of the observation shown by

$$\text{ERROR}_i = [y_i - f(\underline{x}_i)]/y_i \quad (7)$$

The second weights residuals by the value of the predicted function as shown by

$$\text{ERROR}_i = [y_i - f(\underline{x}_i)]/f(\underline{x}_i) \quad (8)$$

This latter form is the one postulated by Book/Young of Aerospace and, by definition, should produce an unbiased estimator.

The regression results and the corresponding statistical properties attending these different error forms depend a great deal on the mathematical optimization techniques chosen. Least-squares regression with the log-error form may be either linear or non-linear in log-space. It can only be linear in log-space if the functional form is log-linear of the general form

$$f = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \dots x_p^{\beta_p} \epsilon, \quad (9)$$

where

x_1, x_2, \dots, x_p are the independent variables

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the coefficients to be determined

ϵ is a multiplicative error term

This equation on transformation to log-space becomes linear with an additive error term so it can be addressed with OLS. The same equation form can be addressed as a non-linear regression using the

definition of the error term given in equation (1) but without transforming the independent variables. Furthermore, any equation form postulated with a multiplicative error term can be addressed with non-linear regression applied to the error term defined by equation (1) above when f_i is not log-linear. We shall refer to regressions applied to definitions of the error term given by equation (1) as “log-error” regressions, with the choice of whether it is log-linear or non-linear up to the user.

Weighted least squares, wherein the ordinary error residual is divided by the value of the observation at each point, has been used to express a multiplicative error. However, analysis of this error form with a simple univariate system of data shows that any such estimator will be biased low, and that the magnitude of the bias will be in some proportion to the variance.

The third error form, hypothesized by Book/Young, is weighted least squares wherein the residual is divided by the value of the predictor function: that is, the estimator at each point. This is a more intuitive form, but in its straightforward form requires a random variable, the predictor function, to appear in both the numerator and denominator, of the objective function. This is mathematically stated as

$$\text{MIN } \Sigma [(y_i - f(\underline{x}_i))/f(\underline{x}_i)]^2 \quad (10)$$

Following the notation of Book and Young, we will denote the process of solving this minimiation problem using a traditional non-linear optimization procedure as GERM-SIMUL (simultaneously solves for the fit parameters as well as the weighting factors).

In searching for an unbiased solution, our preliminary investigations of these three different error models convinced us that the GERM error model was intuitively correct but that the “simultaneous optimization” solution method introduced bias. We reasoned that an iterative solution approach would produce the desired result. The iterative optimization problem is stated mathematically as

$$\text{MIN } \Sigma [(y_i - f_j(\underline{x}_i))/f_{j-1}(\underline{x}_i)]^2 \quad (11)$$

where $f_{j-1}(\underline{x}_i)$ is the value at each [i] of the predictor function of the previous iteration. By using an iterative method, the denominator is a fixed value in each iteration of the optimization. All that remained was to show that the iterations would converge and that the resulting predictor would be

unbiased. In the ensuing discussions, this optimization is termed GERM-ITER since it preserves the error model originally hypothesized by Book/Young, but is combined with an iterative solution method.

Our preliminary investigations, both theoretical and empirical, have confirmed our hypothesis. The four possible solution approaches for multiplicative error (log-error, weighted least squares, GERM-SIMUL, and GERM-ITER) produce results generally depicted in Figure 3. The figure presents a mass of data, shown by the squash-shaped boundary, that is supposed to represent a system with multiplicative error. There is a solid curve running through the centroid of the system of data which depicts truth and, therefore, the desired unbiased function. The four possible optimizations produce the four curves that are shown approximately parallel to the true function. GERM-SIMUL will always be high and the weighted-by-observation least squares will always be low. Log-error will generally always be low, but its position depends on the probability density function assumed for the error residuals (normally for log-linear regressions a log-normal distribution is implicitly assumed, and this has been shown theoretically to always be biased low). GERM-ITER falls approximately on the true function, deviating only in accord with the sampling statistics.

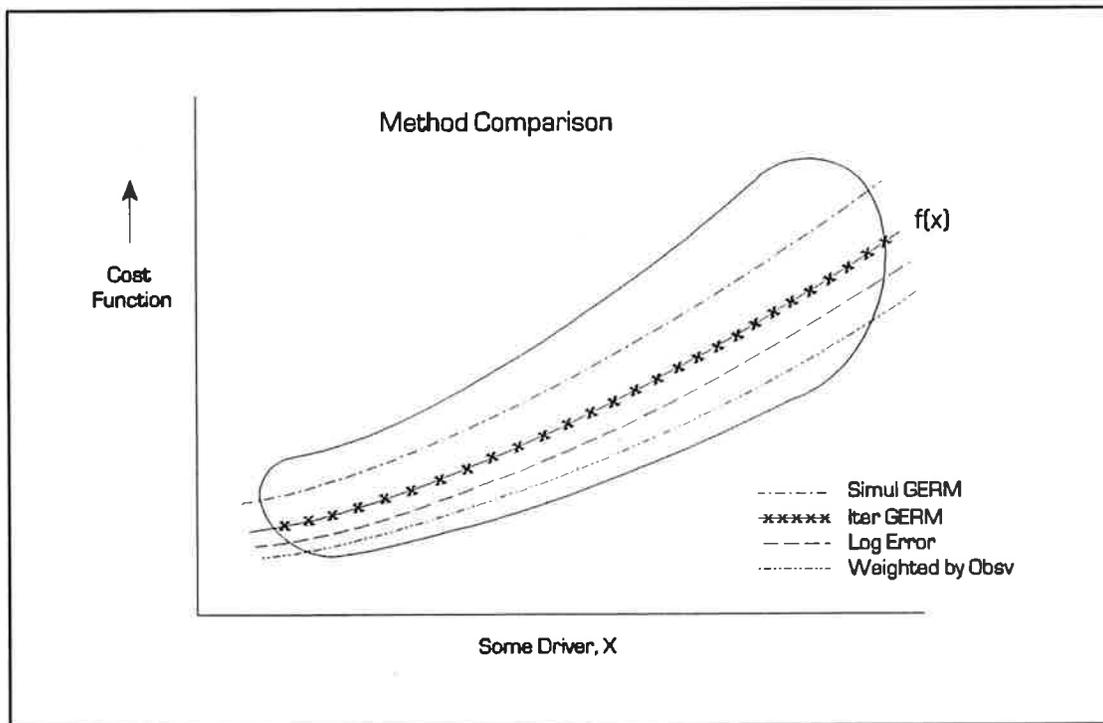


Figure 3

In the next section we will present the theoretical and empirical findings that show the biases and quantify their magnitudes. The focus will be primarily on the simultaneous weighted and iterative weighted least squares optimizations, since these are new forms that have been posited for multiplicative error.

THEORETICAL ANALYSES OF THE SIMPLE CASE

For the simple case, we assume a univariate distribution, by which we mean a sample of, n , observations, y_i , and there is no independent or driver variables. In such a case, we seek simply a single, fixed value as a predictor, P , for y_i ,

$$P = C, \text{ } C \text{ being a constant.}$$

Traditional OLS

With traditional ordinary least squares (OLS) the optimization would be stated as

$$\text{MIN } \sum (y_i - P)^2$$

and the result would be:

$$P = \bar{y}$$

This is the traditional result that is taught when no significant drivers for a problem can be found, “the mean is the ‘best’ estimator.” It is also a property of \bar{y} that it is unbiased (i.e., it is the expected value and the algebraic sum of the error residuals is zero).

Weighted-Simultaneous (GERM-SIMUL)

If we define the error term as a multiplicative error, wherein the ordinary residual is weighted by the value of the predictor, the optimization is

$$\text{MIN } \sum [(y_i - P)/P]^2$$

For this simple case the optimization can be solved in closed form, and the result is

$$P = \bar{y} + \sigma^2/\bar{y}$$

where

σ^2 is used here to represent the uncorrected variance about \bar{y} , $\Sigma (y_i - \bar{y})^2/n$.

Clearly, the term σ^2/\bar{y} will always be positive so that, P , is always biased above \bar{y} .

If we divide, P , through by \bar{y} , we obtain

$$P/\bar{y} = 1 + (\sigma/\bar{y})^2$$

which more clearly shows the bias in the multiplicative space, and $(\sigma/\bar{y})^2$ can be defined as the mean squared proportional error.

Weighted-Iterative (GERM-ITER)

If we now take the same error term used in the immediate above example, but now subject it to an iterative optimization, it can be stated mathematically as

$$\text{MIN } \Sigma [(y_i - P)/P_0]^2$$

for the first iteration where P_0 is some initial guess at P . The result of this optimization is

$$P_1 = \bar{y}$$

where P_1 represents the first iteration. The next iteration is thus expressed as

$$\text{MIN } \Sigma [(y_i - P)/P_1]^2$$

and the result will be

$$P_2 = \bar{y}$$

Since $P_1 = P_2 = \bar{y}$, the optimization has converged with the result

$$P = \bar{y},$$

which is unbiased and, when substituted in the original error term formulation, has preserved the desired multiplicative error.

Weighted by the Observation

For the multiplicative error form created by weighting by the observation, the optimization can be mathematically stated as

$$\text{MIN } \Sigma [(y_i - P)/y_i]^2$$

The closed-form solution of this optimization for the simple case results in an inequality which holds

$$P \leq \bar{y}$$

where the equal sign only comes into play if the variance about \bar{y} is zero. (See Appendix A for this proof.)

Log-error

The optimization of the log-error formulation can be stated as

$$\text{MIN } \Sigma (\log y_i - \log P)^2$$

For any symmetric density distribution, the result of this optimization is

$$P \leq \bar{y}$$

For the particular assumption of a log-normal distribution

$$\bar{y} \approx P * e^{\sigma^2/2} \rightarrow \sigma^2/2$$

where σ^2 is defined as before and the term $e^{\sigma^2/2}$ has been often utilized as a theoretical correction factor in log-error regressions. (See Appendix D for the derivation of this correction factor.)

THEORETICAL ANALYSES WITH A FACTOR EQUATION

Further closed-form analyses are possible for optimizations of factor equations (i.e., equations of the form $y = a*x$) with weighted-simultaneous and weighted-iterative multiplicative error forms. We have been unable to perform closed-form analyses of more complicated equation forms (e.g., $y = a*x^b$, $y = a*x_1^b*x_2^c$, etc.) with these multiplicative error forms. The case for the factor equation can be stated as

$$y = a*x*\epsilon$$

where the multiplicative error term, ϵ , has expectation

$$E(\epsilon) = 1$$

and variance, σ^2 .

Weighted-simultaneous (GERM-SIMUL)

It is convenient for the closed-form analysis to transform y_i by

$$z_i = y_i/x_i$$

Then the optimization is stated mathematically as

$$\text{MIN } \Sigma [y_i/(a*x_i) - 1]^2$$

or as $\text{MIN } \{1/a^2 * \Sigma z_i^2 - 2/a * \Sigma z_i + n\}$.

The minimization seeks a predicted value for, a , and the result is

$$a = \bar{z} + \sigma_z^2/\bar{z}$$

which is identical to the form for the weighted-simultaneous optimization in the simple case. The expected value of z_i is \bar{z} since the x_i are not random variables in the transformation. Thus, the weighted-simultaneous optimization of the factor equation is biased high.

Weighted-Iterative (GERM-ITER)

For the weighted-iterative error form it is not necessary to perform the $z_i = y_i/x_i$ transformation to conduct the analysis. The optimization can be stated mathematically for the first iteration as

$$\text{MIN } \Sigma [(y_i - a*x_i)/f_i]^2$$

where all the f_i are set equal to 1 as the initial guess. The result of this first iteration produces

$$a_1 = \Sigma(x_i*y_i)/\Sigma x_i^2.$$

For iteration 2, let the $f_i = a_1 * x_i$ in the optimization function above and the result of iteration number 2 is

$$a_2 = \Sigma(y_i/x_i)/n$$

which is equal to \bar{z} as defined above. Finally, a third iteration is performed with

$$f_i = a_2 * x_i$$

The result for the third iteration is

$$a_3 = \bar{z},$$

or exactly the same value as was found for the second iteration. Since $a_3 = a_2$, $f_j(x_i) = f_{j-1}(x_i)$ demonstrating convergence (here j designates the iteration number). Furthermore, the process has converged on the value for $a = \bar{z}$, and therefore the predictor is unbiased.

ANALYSIS OF EXAMPLES AND SIMULATIONS

The theoretical investigations were by necessity limited to simple cases. Most practical applications of regression analyses deal with more complex equation forms and, in the cost analysis community, limited data sets. Accordingly, it is important to study and compare optimization techniques involving real data with the newly hypothesized multiplicative error term.

Three examples from the USCM7 data base have been abstracted to test the theoretical findings comparing GERM-SIMUL and GERM-ITER optimizations. The three examples are:

A data set for traveling wave tube amplifiers, 8 observations

A data set for digital electronics, 16 observations

A data set for tracking, telemetry, & control radio frequency (TT&C RF) distribution systems, 13 observations

For each data set, engineering hypotheses were generated and tested by regression analyses to determine which equation forms and parameters provided the greatest explanatory power. In all cases, due to the range in the values of the observations, multiplicative error terms were hypothesized. For this investigation, these three data sets were tested by the two new optimizations and compared.

The results of these tests are presented in Tables 1, 2, and 3. For each table, the observations are listed by name and the observed value in column 1. (Since all three CERs deal with recurring costs, the observed values are the theoretical first unit cost [T1] in thousands of dollars.) Columns 2 and 3 list the calculated results from the two regression equations, first for the GERM-SIMUL and then the GERM-ITER. Columns 4 and 5 list the calculated residuals for the GERM-SIMUL, first in terms of dollars and then in proportional error, by dividing column 4 by the predicted value in column 2. Similarly, columns 6 and 7 show the dollar and proportional residuals for the GERM-ITER regression. Column 8 shows the calculated differences between the two values (column 2 - column 3) and these differences are transformed into proportional errors shown in column 9 by dividing by the predicted value of the GERM-ITER regression, column 3. Column 10 simply calculates the squared value of the proportional errors shown in column 9. The average of each column is shown at the bottom row of the table. It should be noted here that the objective function for the GERM-SIMUL regression is the sum of the squared errors shown in column 5. Similarly, the objective function for the GERM-ITER regression is the sum of the squared errors shown in column 7.

One condition that must be met, if an estimator is to be considered unbiased, is that the algebraic sum of its percentage-error residuals must be zero (true for GERM-ITER but not for GERM-SIMUL as shown by the averages for columns 5 and 7). This condition is not sufficient to claim the estimator is unbiased, as this property must be demonstrated all along the function as well, something we cannot show by these examples. However, when it is not met, it is sufficient to state that the estimator is biased (the case here for GERM-SIMUL regression).

The theoretical investigations of the simple cases showed that the GERM-SIMUL should be biased high and that the magnitude of this bias is equal to the mean-squared proportional error. This is also demonstrated in the examples by comparing the average of column 9 with that of column 10 (the highlighted boxes). Only in the third example, the CER dealing with TT&C RF distribution components, is the comparison not precise, although it is close. The difference is thought to reside in the fact that the CER contains a dummy-type variable which, in this instance, produces a system in which the proportional errors are not distributed uniformly over the range of the function. This is shown by inspecting the residuals in column 9 of Table 3, which clearly shows two distinct classes for the value of the residuals, one in the range of 0.0246 to 0.0549 and the other in the range of

TABLE 1

TUBE TYPE AMPLIFIERS RECURRING T1
 CURVE EQN: $T1 = A * WEIGHT \wedge B * WPF \wedge C$ ¹¹
s.e.GERM = 0.19

SYSTEMS	DATA POINTS		ITER PREDICTED	ESTIMATING ERRORS				DIFFERENCES		SQUARED ITER PROP RES
	ACTUAL	SIMU (2)		PRED- ACTUAL	PROP ERROR (4)/(2)	PRED- ACTUAL	ITERATIVE PROP ERROR (6)/(3)	DIFF (SIMU,ITER)	PROP DIFF	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
				(2)-(1)	(4)/(2)	(3)-(1)	(6)/(3)	(2)-(3)	(8)/(3)	(7) ²
Sat 1	278.81	263.18	264.76	-15.63	-0.0594	-14.05	-0.0531	-1.58	-0.0059	0.0028
Sat 2	357.79	405.67	405.11	47.88	0.1180	47.32	0.1168	0.56	0.0014	0.0136
Sat 3	368.38	422.64	410.77	54.26	0.1284	42.39	0.1032	11.87	0.0289	0.0106
Sat 4	437.44	479.67	475.56	42.23	0.0880	38.12	0.0802	4.11	0.0086	0.0064
Sat 5	652.31	667.65	653.82	15.34	0.0230	1.51	0.0023	13.83	0.0211	0.0000
Sat 6	823.70	740.21	720.64	-83.49	-0.1128	-103.06	-0.1430	19.57	0.0272	0.0205
Sat 7	1066.73	853.03	819.98	-213.70	-0.2505	-246.75	-0.3009	33.05	0.0403	0.0906
Sat 8	1219.83	1607.32	1514.52	387.49	0.2411	294.69	0.1946	92.80	0.0613	0.0379
AVERAGE	650.62	679.92	658.15	29.30	0.0220	7.52	-0.0000	21.78	0.0229	0.0228

MEAN SQUARED ERROR (PROP) = 0.0228

¹¹WPF stands for weighted power-frequency and is defined as $p^{1/2} f^2$, where p is in watts and f is in gigahertz.

TABLE 2

TT&C DIGITAL ELECTRONICS RECURRING T1

CURVE EQN: $T1 = A * WEIGHT^B * NO_BOXES^C * NO_LINKS^D$

S.e.GERM = 0.24

SYSTEMS	DATA POINTS		ITER PREDICTED	ITER	ESTIMATING ERRORS				DIFFERENCES		SQUARED ITER PROP RES
	ACTUAL	SIMU			PRED- ACTUAL	PROP ERROR	PRED- ACTUAL	ITERATIVE PROP ERROR	SIMU- ITER	PROP DIFF	
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)	(10)
					(2)-(1)	(4)/(2)	(3)-(1)	(6)/(3)	(2)-(3)	(8)/(3)	(7)^2
Sat 1	376.47	461.71	436.31		85.24	0.1846	59.84	0.1372	25.40	0.0582	0.0188
Sat 2	918.85	1323.78	1235.93		404.93	0.3059	317.08	0.2566	87.85	0.0711	0.0658
Sat 3	1128.39	1290.23	1242.94		161.84	0.1254	114.55	0.0922	47.29	0.0380	0.0085
Sat 4	1231.13	1128.63	1051.09		-102.50	-0.0908	-180.04	-0.1713	77.54	0.0738	0.0293
Sat 5	1314.85	1699.42	1601.35		384.57	0.2263	286.50	0.1789	98.07	0.0612	0.0320
Sat 6	1619.62	1735.82	1627.53		116.20	0.0669	7.91	0.0049	108.29	0.0665	0.0000
Sat 7	2045.42	1715.75	1663.94		-329.67	-0.1921	-381.48	-0.2293	51.81	0.0311	0.0526
Sat 8	2051.06	2198.81	2136.04		147.75	0.0672	84.98	0.0398	62.77	0.0294	0.0016
Sat 9	2079.58	1441.18	1347.34		-638.40	-0.4430	-732.24	-0.5435	93.84	0.0697	0.2954
Sat 10	2497.71	2977.53	2912.74		479.82	0.1611	415.03	0.1425	64.79	0.0222	0.0203
Sat 11	2786.09	4480.48	4206.26		1694.39	0.3782	1420.17	0.3376	274.22	0.0652	0.1140
Sat 12	3130.08	3596.54	3539.69		466.46	0.1297	409.61	0.1157	56.85	0.0161	0.0134
Sat 13	3641.96	4334.70	4167.13		692.74	0.1598	525.17	0.1260	167.57	0.0402	0.0159
Sat 14	3989.48	3254.69	3205.28		-734.79	-0.2258	-784.20	-0.2447	49.41	0.0154	0.0599
Sat 15	7008.74	6563.59	6125.73		-445.15	-0.0678	-883.01	-0.1441	437.86	0.0715	0.0208
Sat 16	9028.31	8445.94	8219.00		-582.37	-0.0690	-809.31	-0.0985	226.94	0.0276	0.0097
AVERAGE	2802.98	2915.55	2794.89		112.57	0.0448	-8.09	0.0000	120.66	0.0473	0.0474

MEAN SQUARED PROP ERROR = 0.0474

Table 3

TT&C RF DISTRIBUTION RECURRING T1

LINEAR EQN: $T1 = A + B * TT\&C\&R\&F_W\&T^{11} + C * ACTIVE^{111}$

$s.e.^{GERM} = .56$

SYSTEM\$	DATA POINTS		ESTIMATING ERRORS				DIFFERENCES		SQUARED ITER PROP RES
	ACTUAL (1)	SIMU PREDICTED (2)	ITER (3)	SIMULTANEOUS PRED-ACTUAL (4)	PROP ERROR (5)	ITERATIVE PRED-ACTUAL (6)	PROP ERROR (7)	DIFF (SIMU, ITER) SIMU-ITER (8)	
			(2)-(1)	(4)-(1)	(4)/(2)	(3)-(1)	(6)/(3)	(8)-(3)	(8)/(3)
Sat 1	2.05	1.95	1.85	-0.10	-0.0504	-0.20	-0.1081	0.10	0.0549
Sat 2	5.35	9.25	8.97	3.90	0.4214	3.62	0.4036	0.28	0.0308
Sat 3	16.93	86.63	53.00	69.70	0.8046	36.07	0.6806	33.63	0.6346
Sat 4	28.66	69.13	35.90	40.47	0.5854	7.24	0.2017	33.23	0.9255
Sat 5	29.24	79.34	45.87	50.10	0.6315	16.63	0.3625	33.47	0.7296
Sat 6	32.85	77.88	44.45	45.03	0.5782	11.60	0.2610	33.43	0.7521
Sat 7	64.64	131.57	96.88	66.93	0.5087	32.24	0.3328	34.69	0.3581
Sat 8	66.22	125.09	122.09	58.87	0.4706	55.87	0.4576	3.00	0.0246
Sat 9	95.42	142.08	107.13	46.66	0.3284	11.71	0.1093	34.95	0.3262
Sat 10	112.23	121.59	118.67	9.36	0.0770	6.44	0.0543	2.92	0.0246
Sat 11	123.09	86.34	52.71	-36.75	-0.4256	-70.38	-1.3352	33.63	0.6381
Sat 12	134.96	114.59	111.83	-20.37	-0.1778	-23.13	-0.2068	2.76	0.0247
Sat 13	218.20	133.32	98.59	-84.88	-0.6366	-119.61	-1.2132	34.73	0.3523
AVERAGE	71.53	90.67	69.07	19.15	0.2396	-2.45	0.0000	21.60	0.3751

MEAN SQUARED PROPORTIONAL ERROR = **0.3470**

¹¹TTVCRF_WT is the weight in lbs of the TT&C RF distribution components.

¹¹¹ACTIVE distinguishes between active components such as ferrites and solid state devices from passive components such as waveguides, etc.

0.32 to 0.92. This example is perhaps better treated as two distinct data sets. All three examples, however, confirm quite well the theoretical finding which states that the bias in the GERM-SIMUL is equal to the mean-squared proportional error. (See Appendix C for derivation of an approximate correction factor for GERM-SIMUL regressions.)

One of the shortcomings of the three previous examples is the small sizes of the data sets. With samples this small, it is difficult to investigate the bias properties of the predictor function throughout its range. To address this difficulty, several simulations were created and investigated. Data sets of 100 observations were created by random sampling from multiplicative-normal error distributions around several interesting functional forms. In each case, a particular functional form was selected as a generating function representing “truth.”

An observation is created by sampling a value for the independent variable from among its range in a uniform fashion. A corresponding error is randomly sampled from a normal distribution of constant variance as represented in proportional error space. The sampled value of the independent variable is used in the function to derive a central value of the dependent observation, to which is added the product of the random error times the value of the dependent variable calculated from the function. This then constitutes a simulated observation for a known generating function with known error properties. By sampling the independent variable uniformly throughout its range, we create a data set that has equal representation throughout the range.

Figures 4 and 5 portray the results of regressions performed with one of the simulated data sets. Figure 4 is the result when the GERM-SIMUL optimization is used for the regression analysis. The generating function is

$$Y = 0.01 * X^2$$

The estimated regression function for GERM-SIMUL is

$$Y_{est} = 0.010193 * X^{2.03102}$$

Although the coefficients of the regression equation appear reasonable, it is readily seen from the plot that it is significantly above the true generating equation. It should be noted that, in

$$Y = A * (X^B)$$

A B

0.010193 2.03102 <-----Regression Coefficients

-18.3872 <-----Sum Proportional Bias

-14.8282 <----- Sum Prop Errors

14.82824 <----- Obj Function

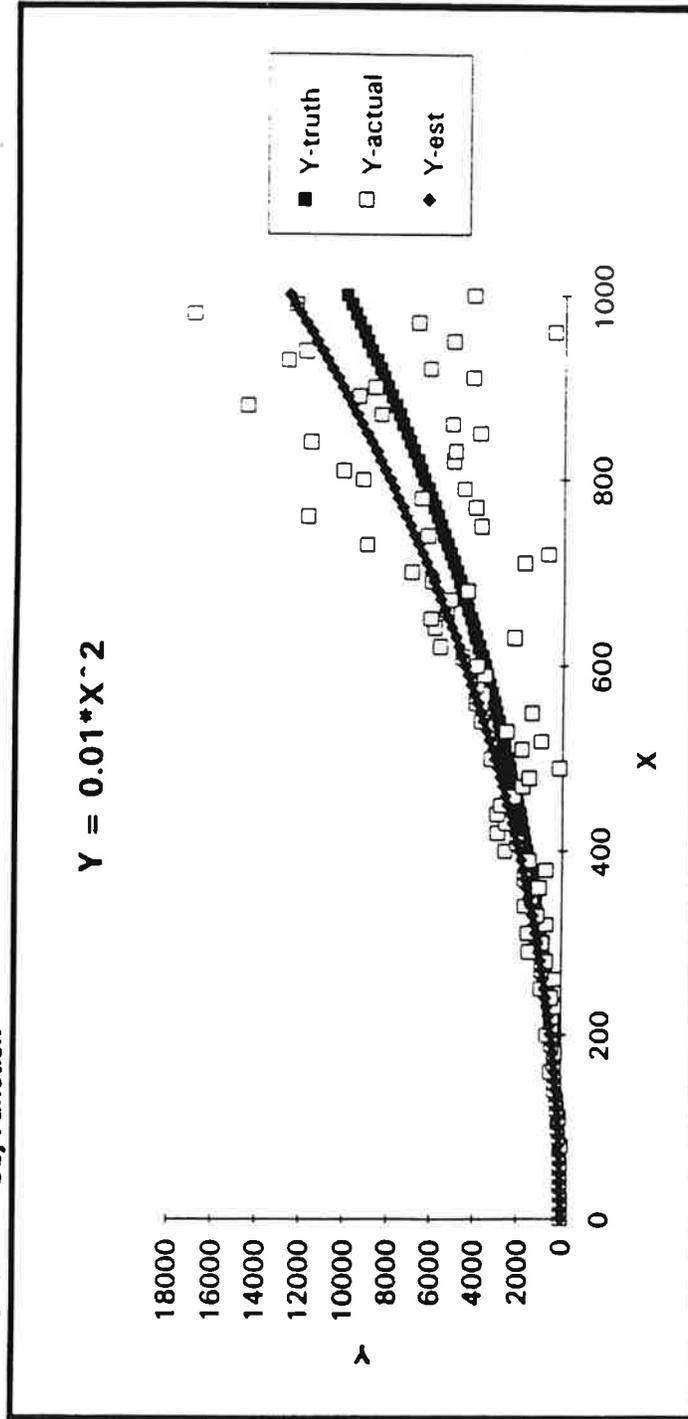


Figure 4. GERM_SIMUL Regression Fit to a Simulated Multiplicative Error Data Set.

$Y = A * (X^B)$
A **B**
 0.008545 2.034012 <-----Regression Coefficients

 -4.15916 <-----Sum Proportional Bias
 -1.2E-05 <----- Sum Prop Errors
 17.37972 <----- Obj Function

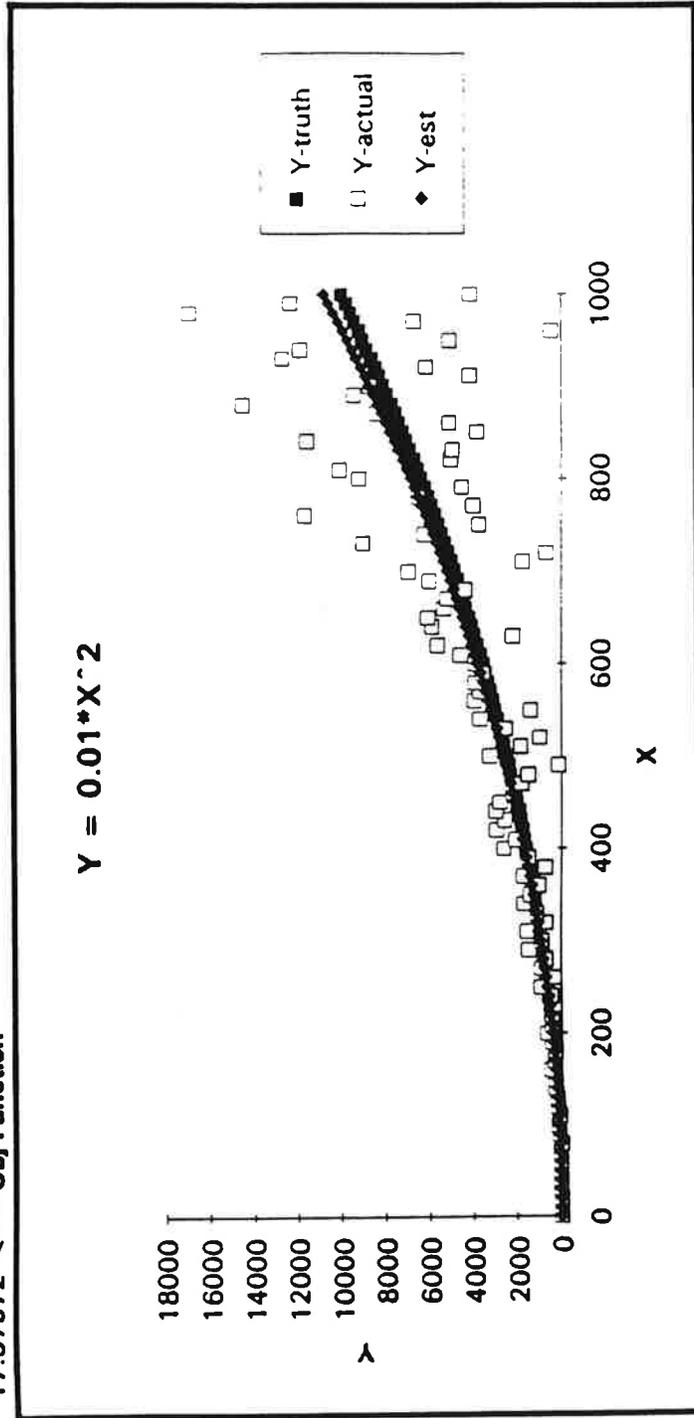


Figure 5. GERM_ITER Regression Fit to a Simulated Multiplicative Error Data Set.

generating the simulated data points, a constraint was applied to prevent any negative cost values, which would tend to raise the level of the estimated function above the generating equation. However, the estimated exponent more than compensates for the constraint such that the net effect is to produce a high estimate. (Note that at the maximum value of the abscissa, $x=1000$, the difference between 1000^2 and $1000^{2.03}$ is 23%.)

Figure 5 presents the results of the simulation for the GERM-ITER regression for the corresponding generated simulated set of data points. The result is

$$Y_{\text{est}} = 0.008545 * X^{2.034015}$$

The figure clearly shows that GERM-ITER has removed most of the bias apparent in Figure 4, and when allowance is made for the effect of the constraint noted above, GERM-ITER appears to be well within the sampling error. Further, note that the value of the objective function (sum-squared-error) is given in Figure 5 as 17.37972, which if divided by the sample size of 100 gives a variance, uncorrected for degrees of freedom, of 0.173797. This represents, according to the theoretical analysis, the estimated percent bias in the GERM-SIMUL regression. Since the two regressions produce nearly the same estimate of exponent, the difference in the levels of the two functions is represented by the differences in the estimated coefficients. Note that the coefficient in Figure 5 of 0.008545, when multiplied by 1.173797, very nearly equals the coefficient of 0.010193 in Figure 4 ($1.173797 * 0.008545 = 0.01003$).

Other simulations with varying equation forms have demonstrated the same property, namely, that the percent difference between GERM-SIMUL and GERM-ITER is always very nearly equal to the proportional variance in the data set. The results depart from this general property when data sets are very small and the influence of a single point can be significant in estimating both coefficients and exponents. However, when the data sets are large, as in the case of the simulations, it appears that the effect of GERM-SIMUL is to raise the level of the function by a constant percentage throughout the range of the data without modifying the exponent. This is what one would expect based on the theoretical investigations with simple cases. (It can be argued that at every point within the range of the independent variable(s) a sample of data can be treated as a sample of a univariate function, and hence the result should be governed by the result found in the theoretical investigation.

Thus, at every point throughout the range the function should be raised by a constant percentage given by the value of the variance.)

CONSIDERATIONS OF GOODNESS OF FIT

The discussion above has treated the matter of bias in regression equations without much comment as to its consequences. Clearly, the analyst wants to know to what degree he should pursue the objective of eliminating bias in his results. Perhaps the most important characteristic of bias of the type discussed above is its effect on measuring, statistically, the quality of the fitted regression equation. Ordinarily, the analyst relies on a standard set of goodness-of-fit statistics such as “students t,” “F,” standard error, and the coefficient of determination or “ R^2 .”

Of these, the standard error is a direct consequence of the objective function used in the least-squares optimization, and is used directly to compare results using the same objective function. A standard error that itself is biased, such as in the case of GERM-SIMUL, cannot be used to compare to unbiased standard errors even if they both purport to model multiplicative error. The same is true of comparisons of standard error measurements between log-error optimizations and GERM-ITER optimizations, although the differences here are of lesser order.

The use of the remaining goodness-of-fit statistics must observe many more constraints. The use of “t” and “F” depends on a normal distribution for the error term and unbiased estimates of the coefficients, characteristics that are commonly assumed in OLS regression analysis. Also, it is necessary that the regression function be unbiased, which requires that all the significant driver variables have been included. If a significant driver is omitted, the analyst cannot rely on the error term being randomly distributed. To the degree that it is not, the measures offered for “t” and “F” will be in error.

When the regression optimization is non-linear, the use of the “t” and “F” values is less precise, even if all the significant drivers have been included. For non-linear regressions it is difficult to establish that the regression equation is necessarily unbiased, and for that reason most statistical packages show “asymptotic” values for these statistics. Still, the assumption of a normal distribution of the error term must be satisfied.

Figure 6 summarizes the applications of goodness-of-fit measures for most regression analysis circumstances. It is understood, that in all cases, the analyst has succeeded in including all the significant driver variables such that the error residuals can be described as random. It should be noted that under no circumstances can goodness-of-fit measures be used to evaluate a GERM-SIMUL regression result.

OTHER CONSIDERATIONS

- **When can goodness-of-fit statistics be applied to judge significance of model? (“t”, “F”, “R²”)**

When analyst can assume a normal distribution and the regression curve is unbiased.

		Under Normal Assumption	
		Goodness of Fit	Asymptotic Goodness of Fit
Linear Regression	OLS	Yes	
	OLS in log space	Yes	
Non-Linear Regression	Least-Squares GERM_SIM		No
	Least-Squares GERM_ITER		Yes
	Least Squares in log-space		Yes

Figure 6

SUMMARY

This investigation of multiplicative regression techniques provides the basis for the following two conclusions.

First, it is recognized that typical past practices in modeling multiplicative error terms were restricted to techniques with known bias problems. The log-error technique is perhaps the best known in its most used log-linear form. However, the problems in the use of log-error regressions, whether linear or non-linear, do not involve the optimization mathematics but the bias on transforming a logarithmic result back to unit space. If one can reasonably assume that the error term in log-space is normally distributed (i.e., log-normal), this regression technique has much to offer. It should be noted here that a log-normal distribution is probably a very good analog for the distribution of cost errors. The log-normal distribution implicitly restricts costs to positive values, something not true when one assumes a normal distribution in unit space. Also, the log-normal distribution when viewed in unit space is skewed to the high side, a characteristic that appears common in cost data. If the analyst has confidence in the log-normal assumption, the use of the log-error term provides useful goodness-of-fit measures, proper only for evaluations in log-space. To the degree that the log-normal assumption is valid, a theoretical correction factor can be applied to the function transformed to unit space to provide an approximately unbiased result.

Second, we have shown that the error term originally defined by Book/Young, when coupled to an iterative optimization technique, is capable of producing unbiased regression results with multiplicative error. This is true independent of the error term distribution, providing all the significant driver variables have been included, such that the error residuals can be accurately described as random. If it can be confidently assumed that the error term is normally distributed, the regression calculated measures of goodness-of-fit can be applied as well. No correction factor is required. This new technique represents a valuable addition to the regression tool box, and for that purpose, we have designated it the **Multiplicative Unbiased Regression Technique (MURT)**.

REFERENCES

1. Philip H. Young, "GERM: Generalized Error Regression Model," Aerospace Corporation, presented to the 25th Annual DoD Cost Analysis Symposium, Leesburg, VA, 4-6 Sep 1991.
2. "Unmanned Spacecraft Model, 7th Edition (USCM7)," USAF, SSC (in preparation).

APPENDIX A

LOW BIAS IN WEIGHTED LEAST SQUARES

Prop: The traditional weighted least squares method using the reciprocal of the actual observations as the weighting factors will produce a biased low estimator.

We now provide a proof of the above proposition in a univariate case.

If a simple multiplicative error model is given by

$$Y_i = f \cdot \epsilon_i \quad \text{for } i = 1, \dots, n, \quad \text{eqn (1)}$$

where f is a population constant, n is the sample size, and the error terms are ϵ_i 's which are independently identically distributed (i.i.d.) random variables with mean value of 1 and variance σ^2 . Then the corresponding weighted least squares objective function using $1/y_i^2$ as the weighting factors will be given by

$$F = \sum_1^n \left(\frac{y_i - f}{y_i} \right)^2 \quad \text{eqn (2)}$$

The closed-form solution for f derived from minimizing equation (2) is

$$\hat{f} = \frac{\sum_1^n \frac{1}{y_i}}{\sum_1^n \frac{1}{y_i^2}} \quad \text{eqn (3)}$$

Now it is required to prove that

$$\hat{f} = \frac{\sum_1^n \frac{1}{y_i}}{\sum_1^n \frac{1}{y_i^2}} \leq \bar{y} \quad \text{for } y_1 > 0, \dots, y_n > 0 \quad \text{eqn (4)}$$

where the equality occurs if the y 's are all equal.

Let's define $z_i = 1/y_i$ for $i = 1, \dots, n$, then

$$\begin{aligned}
 \hat{f} &= \frac{\sum_1^n z_i}{\sum_1^n z_i^2} = \frac{n \bar{z}}{\sum_1^n (z_i - \bar{z})^2 + n \bar{z}^2} \\
 &= \frac{1}{\bar{z} + S_z^2 / \bar{z}} \\
 &= \frac{1}{\bar{z} (1 + S_z^2 / \bar{z}^2)} \\
 &\leq \frac{1}{\bar{z}} \\
 &= \frac{n}{\sum_1^n \frac{1}{y_i}}
 \end{aligned}
 \tag{5}$$

where

$$S_z^2 = \sum_1^n (z_i - \bar{z})^2 / n. \quad \text{uncorrected sample variance of the z's.}$$

Now we need to show that

$$\frac{n}{\sum_1^n \frac{1}{y_i}} \leq \frac{\sum_1^n y_i}{n} = \bar{y} \quad \text{for } y_1 > 0, \dots, y_n > 0
 \tag{6}$$

i.e.,

$$\left(\sum_1^n \frac{1}{y_i} \right) \left(\sum_1^n y_i \right) \geq n^2 \quad \text{for } y_1 > 0, \dots, y_n > 0
 \tag{7}$$

The above equation can be proved by Math Induction. For $n=1$, it is trivial that equation (7) holds. For $n=2$:

$$\begin{aligned}
\left(\frac{1}{y_1} + \frac{1}{y_2}\right)(y_1 + y_2) &= 2 + \frac{y_1}{y_2} + \frac{y_2}{y_1} \\
&= 2 + \frac{y_1^2 + y_2^2}{y_1 y_2} \\
&= 2 + \frac{(y_1 - y_2)^2 + 2y_1 y_2}{y_1 y_2} \\
&\geq 2 + \frac{2y_1 y_2}{y_1 y_2} \\
&= 2^2
\end{aligned}
\tag{8}$$

Hence equation (7) is true for $n=2$. Now we assume that equation (7) is true for $n = k-1$, i.e.,

$$\left(\sum_1^{k-1} \frac{1}{y_i}\right) \left(\sum_1^{k-1} y_i\right) \geq (k-1)^2
\tag{9}$$

Then we need to show that equation (7) is also true for $n=k$. Since it follows from both equations (8) and (9) that

$$\begin{aligned}
\left(\sum_1^k \frac{1}{y_i}\right) \left(\sum_1^k y_i\right) &= \left[\left(\sum_1^{k-1} \frac{1}{y_i}\right) + \frac{1}{y_k}\right] \left[\left(\sum_1^{k-1} y_i\right) + y_k\right] \\
&= \left(\sum_1^{k-1} \frac{1}{y_i}\right) \left(\sum_1^{k-1} y_i\right) + 1 + \left(\sum_1^{k-1} \frac{y_k}{y_i}\right) \left(\sum_1^{k-1} \frac{y_i}{y_k}\right) \\
&\geq (k-1)^2 + 1 + \sum_{i=1}^{k-1} \left(\frac{y_k}{y_i} + \frac{y_i}{y_k}\right)^\dagger \\
&\geq k^2 - 2k + 2 + (k-1)2 \\
&= k^2 - 2k + 2 + 2k - 2 \\
&= k^2
\end{aligned}
\tag{10}$$

[†] Each term in the summation is greater than or equal to 2 by equation (8).

Therefore, we claim that equation (7) (as well as equation (4)) is proved by Math Induction. Equation (4) can be easily extended to the factor equation case.

If a factor model is given by

$$Y_i = a x_i \epsilon_i \quad \text{for } i = 1, \dots, n \quad \text{eqn (11)}$$

where a is a constant factor, x_i is the independent variable for the i th data point and ϵ_i 's are as defined previously.

The weighted least squares solution for minimizing the objection function

$$F = \sum_1^n \left(\frac{y_i - a x_i}{y_i} \right)^2 \quad \text{eqn (12)}$$

is given by

$$\hat{a} = \frac{\sum_1^n \left(\frac{x_i}{y_i} \right)}{\sum_1^n \left(\frac{x_i}{y_i} \right)^2} = \frac{\sum_1^n \frac{1}{z_i}}{\sum_1^n \left(\frac{1}{z_i} \right)^2}, \quad \text{eqn (13)}$$

where $z_i = y_i/x_i$ for $i = 1, \dots, n$. Then it follows from the same argument as given above, that

$$\hat{a} \leq \bar{z} \quad \text{eqn (14)}$$

For complicated equation forms, it can be shown by simulation results that the weighted least squares solution (weighting by the actual observation) will still be biased low.

APPENDIX B
GERM FACTOR

It is proved in the text that in a simple multiplicative error model:

$$Y_i = f \epsilon_i \quad \text{for } i = 1, 2, \dots, n, \quad \text{eqn (1)}$$

where all the terms are as defined in Appendix A, the GERM-SIMUL solution is

$$\hat{f}_G = \bar{y} + S_y^2 / \bar{y} \quad . \quad \text{eqn (2)}$$

Then the uncorrected sample variance under GERM-SIMUL methodology, which estimates the variance σ^2 of the ϵ 's, is given by

$$\begin{aligned} \hat{\sigma}_G^2 &= \sum_1^n \left(\frac{y_i - \hat{f}_G}{\hat{f}_G} \right)^2 / n \\ &= \frac{\sum_1^n (y_i - \bar{y} - S_y^2 / \bar{y})^2}{n \bar{y}^2 (1 + S_y^2 / \bar{y}^2)^2} \\ &= \frac{n S_y^2 (1 + S_y^2 / \bar{y}^2)}{n \bar{y}^2 (1 + S_y^2 / \bar{y}^2)^2} \quad \text{eqn (3)} \\ &= \frac{S_y^2}{\bar{y}^2} \left(\frac{1}{1 + S_y^2 / \bar{y}^2} \right) \\ &= \frac{\hat{\sigma}^2}{1 + \hat{\sigma}^2} \left(= \frac{cv^2}{1 + cv^2} \right)^\dagger, \end{aligned}$$

[†] $\hat{\sigma}^2$ is also denoted by the term, coefficient of variation (cv).

where $\hat{\sigma}^2 = S_y^2 / \bar{y}^2$ is the uncorrected sample variance over the mean squared (in percent squared). The above equation can be equivalently stated as

$$\hat{\sigma}^2 = \frac{\hat{\sigma}_G^2}{1 - \hat{\sigma}_G^2} \quad \text{eqn (4)}$$

It can be shown that $\hat{\sigma}^2$ is an unbiased estimator of the population variance, σ^2 , as the sample size n gets sufficiently large. Since it follows by equation (2) that the magnitude of the bias in GERM_SIMUL solution is approximately $\hat{\sigma}^2$:

$$\frac{\hat{f}_G}{\bar{y}} = 1 + \frac{S_y^2}{\bar{y}^2} = 1 + \hat{\sigma}^2 \quad \text{eqn (5)}$$

Therefore, the downward correction factor to adjust GERM_SIMUL solution is given by

$$\begin{aligned} \text{GERM FACTOR} &= \frac{1}{1 + \hat{\sigma}^2} \\ &= \frac{1}{1 + \frac{\hat{\sigma}_G^2}{1 - \hat{\sigma}_G^2}} \quad \text{(from eqn (4))} \\ &= 1 - \hat{\sigma}_G^2 \quad \text{eqn (6)} \end{aligned}$$

In other words, the GERM FACTOR (given above) can be applied to the CERs that are already generated by GERM_SIMUL method, to approximately eliminate the upward bias in the level of the function.

$$\hat{f}_G * (1 - \hat{\sigma}_G^2) \approx \bar{y} \quad \text{eqn (7)}$$

And the downward bias in $\hat{\sigma}_G^2$ can also be removed by dividing the above-mentioned GERM FACTOR (see equation (4)).

$$\hat{\sigma}_G^2 / (1 - \hat{\sigma}_G^2) \approx \hat{\sigma}^2 \quad \text{eqn (8)}$$

This conclusion can be easily extended to a factor equation model

$$y_i = ax_i \epsilon_i \quad \text{for } i = 1, \dots, n,$$

where all the terms are as defined in Appendix A. The corresponding GERM_SIMUL solution is given by

$$\hat{a} = \bar{z} + S_z^2 / \bar{z} \quad , \quad \text{eqn (9)}$$

where $z_i = y_i/x_i$ for $i = 1, \dots, n$, and S_z^2 is the uncorrected sample variance of the z's. The same conclusion of the GERM FACTOR holds by similar arguments.

For complicated equation forms, the same GERM FACTOR can be verified by simulation results.

APPENDIX C

INTERESTING PROPERTY UNDER GERM

The average of proportional errors using GERM_SIMUL methodology is equal to its uncorrected sample variance. In mathematical notations:

$$\sum_1^n \left(\frac{f_i - y_i}{f_i} \right) / n = \sum_1^n \left(\frac{f_i - y_i}{f_i} \right)^2 / n , \quad \text{eqn (1)}$$

where y_i is the actual i th observation, f_i is the i th fitted value by GERM_SIMUL method, and n is the sample size.

This property holds even in some complicated model forms. For the illustrative purpose, we now provide a proof for the univariate case; namely,

$$\sum_1^n \left(\frac{f - y_i}{f} \right) / n = \sum_1^n \left(\frac{f - y_i}{f} \right)^2 / n , \quad \text{eqn (2)}$$

where $f = \bar{y} + S_y^2 / \bar{y}$ by GERM_SIMUL method.

The proof of equation (2) is given as follows: The left-hand side of equation (2) can be simplified as

$$\begin{aligned} \sum_1^n \left(\frac{f - y_i}{f} \right) / n &= \sum_1^n \left(\frac{\bar{y} + S_y^2 / \bar{y} - y_i}{\bar{y} + S_y^2 / \bar{y}} \right) / n \\ &= \sum_1^n \frac{S_y^2 / \bar{y}}{n(\bar{y} + S_y^2 / \bar{y})} \\ &= \frac{S_y^2 / \bar{y}^2}{1 + S_y^2 / \bar{y}^2} \\ &= \frac{\hat{\sigma}^2}{1 + \hat{\sigma}^2} \\ &= \hat{\sigma}_G^2 , \end{aligned}$$

(by equation (3)
in Appendix B)

which is the right-hand side of equation (2) by definition.

APPENDIX D

PING FACTOR

For a number of CERs in the cost analysis field, the error of an individual observation (e.g., cost) appears to be approximately proportional to the magnitude of the observation. In such cases, it is appropriate to hypothesize a multiplicative error term for the CER. One common practice in the past is to work in the log space by taking natural logs of both the dependent variable and the equation form. When the transformed equation is linear in log space, OLS can be applied to derive a Best Linear Unbiased Estimator (BLUE) in log space which is also the Maximum Likelihood Estimator (MLE) in log space. If the equation form is not log-linear, the log transform can still be used to model a multiplicative error, but the optimization will be non-linear least squares. The non-linear (iterative) solution thus derived will be asymptotically unbiased in log space. Although the mean and median are the same for the log-linear CERs in log space, when transforming the equation back to unit space the mean and median differ, with the CER predicting closer to the median instead of the mean. Therefore, the direct translation of the equation back to unit space tends to underestimate the mean value of the original population, and the Ping Factor, a multiplicative correction factor, is used to adjust for this bias and the sampling bias for estimating the median in unit space.

Let us propose a statistical hypothesis with a multiplicative error term. If a multiple log-linear regression model is given by the following:

$$\begin{aligned}
 Y_i &= e^{\alpha} X_{1i}^{\beta_1} X_{2i}^{\beta_2} \dots X_{pi}^{\beta_p} \epsilon_i, & \text{for } i = 1, \dots, n, \\
 &= e^{\underline{\ln(X_i)}' \underline{\beta}} \epsilon_i
 \end{aligned}
 \tag{1}$$

where ϵ_i 's are i.i.d. random variables associated with $\text{LN}(0, \sigma^2)$, α , β_1 , ..., β_p , and σ^2 are unknown parameters, X_{1i} , X_{2i} , ..., X_{pi} are the independent variable values for the i th data point, $\underline{\ln(X_i)}^t$ is a row vector of the X_i 's in log space (i.e., $\underline{\ln(X_i)}^t = (1, \ln X_{1i}, \dots, \ln X_{pi})$), $\underline{\beta}$ is a column vector of the α and β 's (i.e., $\underline{\beta}^t = (\alpha, \beta_1, \dots, \beta_p)$), p is the total number of independent variables in the model, and n is the sample size. The above model can be equivalently stated as

$$\begin{aligned}
 \ln(Y_i) &= \alpha + \beta_1 \ln(X_{1i}) + \dots + \beta_p \ln(X_{pi}) + \ln(\epsilon_i) \\
 &= \underline{\ln(X_i)}^t \underline{\beta} & \text{for } i = 1, \dots, n,
 \end{aligned}
 \tag{2}$$

where $\ln(\epsilon_i)$ is distributed as $N(0, \sigma^2)$ (for $i = 1, \dots, n$). In other words, the conditional distributions of Y at a given x value, say $X = \underline{x_0}$, in both log and unit space are given respectively by the following equations:

$$\ln(Y/X=x_o) = \underline{\ln(x_o)}' \underline{\beta} \sim N(\underline{\ln(x_o)}' \underline{\beta}, \sigma^2) \quad \text{eqn (3)}$$

$$Y/X=x_o = e^{\underline{\ln(x_o)}' \underline{\beta}} \sim LN(\underline{\ln(x_o)}' \underline{\beta}, \sigma^2) \quad \text{eqn (4)}$$

where $\underline{\ln(x_o)}'$ ($= (1, \ln(x_{1o}), \dots, \ln(x_{po}))$) is a vector of given driver values in log space. It follows from equation (3) that the conditional mean and median value of Y (in log space) at this given value, $\underline{x_o}$, are both equal to

$$E(\ln(Y/X=x_o)) = \underline{\ln(x_o)}' \underline{\beta} = \mu_L = M_L \quad \text{eqn (5)}$$

However, it can be easily shown that there is a difference between the conditional mean and median values of Y (at the given $\underline{x_o}$) in unit space.

$$E(Y/X=x_o) = e^{\underline{\ln(x_o)}' \underline{\beta} + \sigma^2/2} = \mu_A \quad \text{eqn (6)}$$

$$\text{Median}(Y/X=x_o) = e^{\underline{\ln(x_o)}' \underline{\beta}} = M_A \quad \text{eqn (7)}$$

And $e^{\sigma^2/2}$ can be regarded as a factor explaining the difference between μ_A and M_A ; namely,

$$\mu_A / M_A = e^{\sigma^2/2} \quad \text{eqn (8)}^\dagger$$

If the classical assumptions hold as explained in equation (1), the OLS method can generate an unbiased estimator of Y in log space (at any given x value, say $X=\underline{x_o}$), which also follows a normal distribution.

$$\ln(\hat{Y}/X=x_o) = \underline{\ln(x_o)}' \hat{\underline{\beta}} \sim N(\underline{\ln(x_o)}' \underline{\beta}, r_o \sigma^2) \quad \text{eqn (9)}$$

where $r_o = \underline{\ln(x_o)}(\underline{X}'\underline{X})^{-1}\underline{\ln(x_o)}'$, \underline{X} is the design matrix in log space, and $\underline{\ln(x_o)}$ is, as explained above, a row vector of given independent variable values in log space. Therefore, the distribution of the direct translation of the above equation into unit space is then given by

$$\hat{Y}/X=x_o = e^{\ln \hat{Y}/X=x_o} = e^{\underline{\ln(x_o)}' \hat{\underline{\beta}}} \sim LN(\underline{\ln(x_o)}' \underline{\beta}, r_o \sigma^2), \quad \text{eqn (10)}$$

[†] It is clear from equation (8) that the direct translation of the solution from log space to unit space will not be a good estimator of the population mean, μ_A , if the difference between μ_A and M_A is not negligible.

with the mean value equal to

$$E\left(e^{\ln \hat{Y}/X=x_0}\right) = e^{\frac{(\ln(x_0) - \bar{\ln}(x)})^2}{SSx} + r_0 \sigma^2/2} \quad \text{eqn (11)}$$

Then the net bias for estimating the mean value in unit space at a given x_0 vector is

$$e^{(1-r_0) \sigma^2/2} \quad \text{eqn (12)}$$

Let's use two simple examples to illustrate the above equation. In the univariate case (where $T = e^{\bar{\ln}(Y)}$), the population mean is underestimated by an amount

$$e^{\sigma^2/2}, \quad (\text{a transformation bias}) \quad \text{eqn (13)}$$

and the true median is overestimated by an amount

$$e^{\sigma^2/2n}, \quad (\text{a sampling bias}) \quad \text{eqn (14)}$$

Thus, the net bias can be expressed by the factor

$$e^{(1-1/n) \sigma^2/2} \quad \text{eqn (15)}$$

For the one independent variable model, the net bias is given by

$$e^{\left(1 - \frac{1}{n} - \frac{(\ln(x_0) - \bar{\ln}(x))^2}{SSx}\right) \cdot \frac{\sigma^2}{2}} \quad \text{eqn (16)}$$

However, the estimator of the population variance σ^2 is s^2 ,^{††} which on the average overestimates σ^2 in unit space since by Jensen's inequality:

$$E(e^{s^2/2}) \geq e^{E(s^2/2)} = e^{\sigma^2/2} \quad \text{eqn (17)}$$

Therefore, it is necessary to develop a function to generate an unbiased estimator for the net bias factor given in equation (12). Such a function, g , is given below

$$g(t) = 1 + \frac{t}{1!} + \frac{(n-k) t^2}{2!(n-k+2)} + \frac{(n-k)^2 t^3}{3!(n-k+2)(n-k+4)} + \dots, \quad \text{eqn (18)}$$

[†] SSx is the sum of squared deviations (in x) about the mean in log space and x_0 is a given x value. From this equation, it is clear that the estimating bias associated with a given point, i.e., x_0 , depends on how far x_0 is from the center of the data base. The minimum bias occurs when $\ln(x_0) = \bar{\ln}(x)$.

^{††} s stands for standard error of estimate in log space.

where

- $t = (1 - r_o) s^2/2$
- $k =$ total number of coefficients to be estimated,
- $n =$ total number of data points, and
- $s =$ standard error of estimate in log space.

The above defined function, g , has a property that

$$E(g(a \cdot s^2)) = e^{a \cdot s^2/2}, \quad \text{for any real number } a \quad \text{eqn (19)}$$

Hence the unbiased estimator for equation (6) is given by

$$e^{\hat{a}} X_{1o}^{\hat{\beta}_1} \cdots X_{po}^{\hat{\beta}_p} \cdot g((1 - r_o) s^2/2) \quad , \quad \text{eqn (20)}$$

where $g((1 - r_o) s^2/2)$ is the so-called net correction factor. Since r_o has to be evaluated at each different x level which is of little realistic use, we now use k/n as an approximation of r_o for any given x value, i.e., the value of t to be used in the function $g(t)$ is

$$t = \left(1 - \frac{k}{n}\right) \frac{s^2}{2} \quad , \quad \text{eqn (21)}$$

The term k/n is the expected value of r_o . In other words, the Ping Factor is a general correction for the level of the function. It is evaluated near the centroid of the data base. If n gets sufficiently large, the Ping Factor can be approximated by

$$PF = e^{\left(1 - \frac{k}{n}\right) \frac{s^2}{2}} \quad \text{eqn (22)}$$

In the exponent of equation (22), the first term is used to adjust the downward bias between the mean and the median (can be regarded as a transformation bias), the second term is used to adjust the upward bias for estimating the median (can be regarded as a sampling bias). Some references consider e^a as a representative of the level of the (conditional) median for the entire function. This could be very misleading if the intercept term is far away from the mass of the data points. In this circumstance, the variance in the estimate of the intercept can be very large, larger in fact than the population variance of the regression line. To use the median value of Y at the intercept, (i.e., T_1), as the level of the function and apply it to correct the upward bias (of the median) of the translated equation can cause the corrected function in some cases to lie outside the range of the data set.

Generally, the Ping Factor should be applied to all equations fit in log space by respecifying each equation's constant term as the product of its original constant and the correction factor. However, caution should be exercised if dummy variables are used to differentiate observations with different attributes (e.g., airborne vs ground based antenna). A CER's predictive capability might not

be improved by applying one adjustment factor to two or more different populations if the individual sample variances associated with these different categorical data are not equal.

A more detailed discussion of the Ping Factor can be found in Tecolote TR-006/2, "Error Corrections for Unbiased Log-Linear Least Squares Estimates."

REFERENCES

1. Goldberger, Arthur S., "The Interpretation and Estimation of Cobb-Douglas Functions," Econometrica, Vol. 35, July-October 1968, pp. 464-472.
2. Dagel, Harold, "Unbiased Estimation of Power-Function Equations," unpublished paper, Washington, DC: Headquarters Naval Material Command, Cost Analysis Division, December 1984.
3. Lundegard, Robert J., "Unbiased Estimation of Learning Curves," letter from Chief of Naval Material to Commander, Naval Air Systems Command, dated 14 February 1985.
4. Dorsett, J. T., letter to Chief of Naval Material, responding to an earlier memorandum from Commander, Naval Air Systems Command, dated 18 January 1985.
5. Duan, Naihua, "Smearing Estimate: A Nonparametric Retransformation Method," Journal of the American Statistical Association, Vol. 78, September 1983, No. 383, pp. 605-610.
6. Hu, Shu-Ping and Arve Sjøvold, "Error Corrections for Unbiased Log-Linear Least Square Estimates," Tecolote Research, Inc. TR-006/2, March 1989.