

MINIMUM-PERCENTAGE-ERROR REGRESSION UNDER ZERO-BIAS CONSTRAINTS

Stephen A. Book and Norman Y. Lao
The Aerospace Corporation
El Segundo, CA 90245

ABSTRACT

Classical least-squares regression imposes severe requirements on analysts who want to derive functional relationships between dependent y and independent x variables, forcing the analyst to model the error as additive when the relationship is linear ($y = a+bx$) or logarithmic ($y = a + b \log x$) but as multiplicative when the relationship is exponential ($y = ax^b$) or power ($y = ab^x$). This severely restricts his or her ability to optimally model natural phenomena. "General-error regression," taking advantage of modern computing capability and advanced numerical analysis techniques, offers the analyst the choice of minimizing additive or multiplicative error regardless of the functional form of the relationship. It turns out, though, that relationships derived by minimizing percentage (i.e., multiplicative) error contain significant positive bias (i.e., they tend to overestimate the actual values of the dependent variable). In the recent past, the method of iteratively reweighted least squares has been applied to yield zero-bias relationships at some cost in the magnitude of the standard error. In this report the general-error regression problem is instead formulated as a constrained nonlinear optimization problem, with percentage standard error of estimation optimized (i.e., minimized), subject to percentage bias being zero. Naturally, the percentage error will be somewhat larger than it would be if the bias were unconstrained (one cannot serve two masters!), but in general not as large as given by iteratively reweighted least squares, so zero bias is paid for by a small increase in standard error.

INTRODUCTION

Cost-estimating relationships (CERs) comprising Version 7 (August 1994) of the Air Force's *Unmanned Space Vehicle Cost Model* (USCM-7) have been statistically derived from historical cost data using "general-error" regression. Such CERs are usually expressed in the form of linear or curvilinear regression equations that predict cost (the dependent variable) as a function of one or more "cost drivers" (independent variables). Because the range of cost data behind a given CER will span one or more orders of magnitude, the correct choice of error model is "multiplicative" (a percentage of the estimate) rather than "additive" (a specific number of dollars). Unfortunately, when classical least-squares regression, or "ordinary least squares" (OLS), is used to derive functional relationships between dependent y and independent x variables, the analyst must model the error as additive when the relationship is linear ($y = a+bx$) or logarithmic ($y = a + b \log x$) but as multiplicative when the relationship is exponential ($y = ax^b$) or power ($y = ab^x$). In the pre-computer age, when explicit formulas were used in the linear case to calculate the coefficients a and b from the data, the latter two curvilinear relationships were derived suboptimally (i.e., with a larger standard error of the estimate than necessary) by applying ordinary least-squares regression to the logarithms of the data points. In addition to inducing a larger than required error of estimation, this technique tended to yield relationships having significant negative bias (i.e., they underestimated the actual value of the dependent variable). Furthermore, when the coefficients of nonlinear forms are derived by taking logarithms of both sides and reducing the formulation to $\log(y) = \log(a) + b \log(x) + \log(E)$, the error of estimation is expressed in meaningless units ("log dollars"), so that the quality of the nonlinear form cannot easily be compared with the quality of the linear form, whose error of estimation is expressed in "dollars."

Using modern computing capability and advanced numerical analysis techniques in place of applying ordinary least squares to logarithmically-transformed data, The Aerospace Corporation developed "general-error regression" in order to derive functional relationships having optimal (i.e., minimum possible) error of estimation, while allowing the analyst to choose to minimize additive error or multiplicative error regardless of whether the functional relationship turns out to be linear or nonlinear. An additional advantage turned out to be that previously unavailable functional forms (most prominently $y =$

Approved for public release; distribution is unlimited.

$a+bx^c$) can be fit to the data when appropriate. Unfortunately, as was shown in 1993 by Tecolote Research Inc., functional forms derived by minimizing percentage (i.e., multiplicative) error act with significant positive bias (i.e., they tend to overestimate the actual values of the dependent variable). As a solution to the problem of bias, Tecolote suggested the technique of “iteratively reweighted least squares” (IRLS). See Reference 11 and Appendix C of Reference 15 for details. This report, on the other hand, proposes to formulate the general-error regression problem as a constrained nonlinear optimization problem, the constraint being that the percentage bias of the functional relationship be zero. In particular, percentage standard error of estimation is optimized (i.e., minimized), subject to the percentage bias being zero, with the resulting standard percentage error somewhat larger than it would be if the bias were unconstrained, but in general somewhat smaller than given by IRLS.

General-error regression can be implemented in a number of commercial software packages, including Microsoft's Excel spreadsheet using the Excel Solver routine, which handles complex nonlinear problems by building a worksheet with multiple changing cells. Several numerical examples are provided to illustrate the relative magnitudes of standard error and bias. In any specific case, the analyst has the option to select the minimum-percentage-error relationship (typically with positive bias) or the zero-bias relationship (generally with suboptimal error of estimation).

WHY PERCENTAGE ERROR?

In USCM-7 (Reference 15), all errors of estimation, both standard errors and bias errors, are expressed in percentage terms, not in dollar values. There are two practical benefits of this that accrue to the cost estimator, for whose use the document was produced. The first benefit of expressing cost-estimating error in percentage terms is stability of meaning across a wide range of programs, time periods, and estimating situations. A percentage error of, say 30%, retains its meaning whether a \$10,000 component or a \$10,000,000,000 program is being estimated. A standard error expressed in dollars, say \$59,425, is an extremely huge error when estimating a \$10,000 component, but is much less significant when reported in connection with a \$10,000,000,000 program. Even in cases that are not so extreme, a standard error expressed in dollars quite often makes a CER virtually unusable at the low end of its data range, where relative magnitudes of the estimate and its standard error are inconsistent.

While “standard error of the estimate” is the root-mean-square (RMS) of all percentage errors made in estimating points of the data base (a “one-sigma” number that bounds probable cost within an interval surrounding the estimate), “net-percentage bias” is the algebraic sum, including positive values and negative values, of all percentage errors made in estimating points of the data base. Net percentage bias is a measure of balance between percentage overestimates and underestimates of data-base actuals. The second practical benefit is the fact that a constant dollar-value expression of bias would not be as informative as estimating the error in percentage terms, because a particular amount of dollars of bias would not have the same meaning at every point of the cost range.

THE MULTIPLICATIVE-ERROR MODEL

OLS regression, either linear or nonlinear, has been applied in the past to historical-cost data in order to derive CERs. A fundamental assumption of OLS is that the error be additive. More precisely, each observed value of cost is assumed to be a function of cost-driving parameters plus a random error term that does not depend on the parameters. Unfortunately, this assumption is not always valid. A case in point is where the values of “actual” costs in the data base change by an order of magnitude or more as a function of the parameters, in which case the random error is more realistically considered to be proportional to the magnitude of the cost, thereby effectively depending on the parameters. In such a case it is often more realistic to assume a multiplicative error model. This type of situation has been dealt with in the past by taking logarithms of both sides and then applying additive-error linear regression. An alternate “ad-hoc” method is described in Reference 9; however, both are suboptimal in the least-squares sense. Other discussions of theoretical and practical difficulties of working with the logarithmic-transformation method can be found in References 3, 6, 9, 10, 13, and 18. This procedure also unnecessarily binds one to a specific class of regression-equation forms (see Reference 2), and it is far from clear that the appropriate forecasting error is the one that is being minimized. Reference 5 reports on a Monte Carlo study of the general question of additive vs. multiplicative error.

General-error regression is designed to fill this gap. It allows the user to specify, given historical-cost data, whether an additive or multiplicative error model is to be used in deriving the least-squares CERs. And it allows the user to select an appropriate functional form of the CER independently of his or her choice of error model. In the past, straight-line OLS regression forced the choice of an additive-error model, while particular curvilinear forms (namely, $y = ax^b$ and $y = ab^x$) required the assumption of a multiplicative-error model. In general-error regression, the choice of functional form is essentially unrestricted. As well the OLS-compatible forms $y = ax^b$ and $y = ab^x$, now available are forms such as $y = a+bx^c$, $y = a + bc^x$, $y = a+bx+c \sqrt{x}$, etc. The range of available forms is unlimited.

Now that any of a wide range of functional forms may be combined with either of the two error models (additive or multiplicative), it is incumbent upon the cost analyst to choose the best pairing of functional form and error model that is consistent with engineering economics and historical-cost data. The decision made in the case of USCM-7 was to use the multiplicative-error model throughout the analysis and to let the choice of functional form be dictated by engineering and data considerations. It is felt that the multiplicative model, incorporating uniform percentage error of estimation across the entire cost range, reflects reality better than does a uniform dollar amount of error across that range. In cases where the dollar range is sufficiently narrow as to make the uniform-dollar-error assumption tenable, the percentage-error assumption is also adequate to model the reality.

Details of only the two-dimensional case are described in this report, but the procedures can be (and have been) easily generalized to higher dimensions, such generalizations having been used in USCM-7 to derive higher-dimensional CERs where appropriate. In the two-dimensional case, each observation consists of a deterministic cost-driving parameter (x) and a stochastic estimated cost (y). Both linear and nonlinear fits are considered, the theory being impervious to any specific form of the regression equation. The error definition is as follows:

$$\text{Multiplicative Error} = (\text{Actual} - \text{Predicted}) \div \text{Predicted}$$

The following are inputs to the mathematical computations:

- n = Number of data points (observations) in sample
- x_i = Value of cost-driving parameter for each data point, $i = 1, K, n$
- y_i = Observed value of cost for each data point $i = 1, K, n$
- m = Number of numerical coefficients in model ($m < n$).

The term “coefficient” in this context includes numerically constant exponents, as well as numerically constant multiplicative factors. Then

$$y = f(x, \underline{a})$$

is the regression function, y of x , to be fit to the historical cost data, and

$$\underline{a} = (a_1, K, a_m)$$

is the coefficient vector to be determined by mathematical optimization. If $m = n$, the parameter vector \underline{a} is determined exactly (“interpolation”), regardless of the form of $f(x, \underline{a})$. If $m > n$, the parameter vector cannot be determined uniquely.

The multiplicative error model on which USCM-7 CER-development is based has the probabilistic structure

$$Y_i = f(x_i, \underline{a}) \varepsilon_i, \quad i = 1, K, n,$$

where ε_i is a random error such that

$$E(\varepsilon_i) = 1,$$

$$\text{Var}(\varepsilon_i) = \sigma_M^2,$$

and σ_M^2 represents a constant (independent of x_i) multiplicative-error dispersion around 1. Otherwise, as in the additive-error model, the probability distribution of ε_i is arbitrary. Mean and variance of Y_i are, respectively,

$$E(Y_i) = f(x_i, \underline{a}) \quad E(\varepsilon_i) = f(x_i, \underline{a})$$

$$Var(Y_i) = f^2(x_i, \underline{a}) \quad Var(\varepsilon_i) = f^2(x_i, \underline{a}) \sigma_M^2$$

Note that, while the expected values are the same as those of the additive-error model, the variance in the multiplicative-error model depends on the cost-driving parameter, growing in magnitude as the dependent variable (in our context, cost) grows.

MINIMIZING PERCENTAGE ERROR OF ESTIMATION

In the multiplicative-error model, as before one sample observation y_i of Y_i corresponds to each x_i , but in this case the sample error ε_i equals the ratio of y_i to $E(Y_i)$. Thus,

$$\varepsilon_i = \frac{y_i}{E(Y_i)} = \frac{y_i}{f(x_i, \underline{a})}$$

where $\varepsilon_i = 1$ for all i indicates no prediction error. In this case, the least-squares problem can be formulated to find the parameter vector \underline{a} that minimizes the sum of squared relative deviations from the predictions:

$$SSD_M^2 = \sum_{i=1}^n (\varepsilon_i - 1)^2 = \sum_{i=1}^n \left(\frac{y_i}{f(x_i, \underline{a})} - 1 \right)^2 = \sum_{i=1}^n \left(\frac{y_i - f(x_i, \underline{a})}{f(x_i, \underline{a})} \right)^2 \quad (1)$$

This operation minimizes the sum of squares of the percentage errors for the multiplicative-error model. The far-right-hand representation expresses SSD_M^2 in terms of “relative error” and allows the least-squares multiplicative error model to be interpreted as minimizing the sum of squares of relative errors.

To solve the least-squares problem, we could use Equation (1) to calculate the m partial derivatives of SSD_M^2 with respect to each component $a_j, j = 1, K, M$, of the parameter vector \underline{a} and set them all equal to zero. If possible, we solve the system of simultaneous so-called “normal equations” for \underline{a} to minimize SSD_M^2 , but even if $f(x, \underline{a})$ is linear in $a_j, j = 1, K, m$, the normal equations are not necessarily linear in the a_j s. For nonlinear normal equations, numerical-analysis methods are usually necessary (unless the equations can fortuitously be solved in closed form analytically). The multidimensional Newton-Raphson method is a good technique (Reference 8). In general, the solution to a nonlinear system of equations is not often unique, because the function being minimized may have several “peaks” and “valleys.” Unreasonable MPE solutions must be excluded, and the solution that is most plausible “physically” selected.

To estimate dispersion around the least-squares fit, the standard error of estimate for the multiplicative error model can be defined as follows:

$$SEE_M = \sqrt{\frac{1}{n-m} \sum_{i=1}^n \left\{ \frac{y_i}{f(x_i, \underline{a}_0)} - 1 \right\}^2} = \sqrt{\frac{1}{n-m} \sum_{i=1}^n \left\{ \frac{y_i - f(x_i, \underline{a}_0)}{f(x_i, \underline{a}_0)} \right\}^2}, \quad (2)$$

where this time \underline{a}_0 is the value of \underline{a} that minimizes SSD_M^2 . The far-right expression for SEE_M leads to interpretation of SEE_M as a measure of percentage error made in using the multiplicative error regression formula as a predictor. It would make sense to interpret SEE_M times 100% as the “one-sigma” percentage error made in using $f(x, \underline{a}_0)$ as an estimate of the cost corresponding to the cost-driving parameter x_i .

BIAS

Experience has shown that minimum-percentage-error (MPE) CERs resulting from minimizing the expression in Equation (1) have positive net percentage bias, defined by

$$B_M = \sum_{i=1}^n \left(\frac{f(x_i, \underline{a}) - y_i}{f(x_i, \underline{a})} \right) \quad (3)$$

The reason for this is not yet fully understood (by us), but we believe it may have something to do with the fact that, for the same absolute difference between y_i and $f(x, \underline{a})$, a smaller value of SSD_M^2 will result from choosing $f(x, \underline{a})$ above rather than below y_i . This is due to the fact that $f(x, \underline{a})$ appears in the denominators, and larger denominators lead to lower values of SSD_M^2 . Nevertheless, the magnitude of the net percentage bias is not large for most CERs, typically being around 8%.

REDUCING BIAS BY ITERATIVELY REWEIGHTED LEAST SQUARES

Tecolote Research, Inc. (Reference 11 and Appendix C of Reference 15), the Air Force's prime contractor for development of USCM-7, suggested a method based on IRLS to reduce the bias of MPE CERs at a small cost in standard error. (See also References 1, 12, 16, and 17.) Tecolote's method, referred to in USCM-7 as the "minimum unbiased percentage error" (MUPE) technique, calls for computation of a sequence of CER parameter vectors $\underline{a}_1, \underline{a}_2, \dots, \underline{a}_K$ converging to a parameter vector \underline{a}_0 , that may or may not be optimal with respect to any appropriate criterion other than zero bias. If the functional form $f(x, \underline{a})$ is specified, then successive sequential CER candidates are defined as follows:

$$f(x_i, \underline{a}_{j+1}) = \text{Min} \sum_{i=1}^n \left(\frac{y_i - f(x_i, \underline{a}_j)}{f(x_i, \underline{a}_j)} \right)^2. \quad (4)$$

Notice that only the parameter vector \underline{a} in the numerator is subject to optimization; the denominator is constant with respect to the optimization process, having been selected in the previous iteration.

The operative element of Expression (4) can be rewritten in the following way:

$$\sum_{i=1}^n \left(\frac{y_i - f(x_i, \underline{a}_j)}{f(x_i, \underline{a}_j)} \right)^2 = \sum_{i=1}^n \frac{1}{f^2(x_i, \underline{a}_j)} (y_i - f(x_i, \underline{a}_j))^2 \quad (5)$$

What is curious about Expression (5) is that the MUPE/IRLS technique is really an additive-error technique. It minimizes a weighted sum of additive squared errors. This apparently is how it manages to reduce the bias to zero. MUPE/IRLS is not truly a multiplicative-error technique. Nevertheless, once a solution is found, the percentage error of estimation and the bias can be calculated and compared with the corresponding statistics for MPE CERs. As noted earlier, the percentage error of a MUPE/IRLS CER will naturally be larger, but its bias will be less, exactly zero in the case of a linear functional form and apparently near zero in other cases.

MPE CERs constitute Section 5 of the USCM-7 document (Reference 15). As a perusal of Section 5 will reveal, MPE CERs tend to have positive average percentage bias (ranging from a low of 1% to a high of 29% at the extremes, with 8% as most typical). What mathematics guarantees about these CERs is that their standard percentage error is as small as possible, consistent with data-base applicability and appropriate technical relationships between cost and cost drivers.

MUPE/IRLS CERs constitute Section 6 of Reference 15. Yet, as the next section of this report will show, the standard errors of MUPE/IRLS CERs are not exactly minimized among the class of unbiased CERs. Perusal of Section 6 will show that these CERs tend to have standard errors somewhat greater than those in Section 5 (ranging from 0% to 23% greater at the extremes, with 6% greater most typical). Percentage bias of MUPE/IRLS CERs does indeed equal zero to the accuracy reported.

TRUE ZERO-BIAS CERs DERIVED BY CONSTRAINED OPTIMIZATION

Obtaining zero percentage bias by the MUPE/IRLS method is intellectually unsatisfying in one major respect: It is not clear (to us) exactly what quantity, quality, or characteristic, if any, is being

optimized by the IRLS procedure. This difficulty is what inspired the ZPB/MPE (zero percentage bias, minimum percentage error) constrained optimization solution. The ZPB/MPE method finds coefficients such that the resulting CER has smallest possible percentage error, *subject to the constraint that* its percentage bias be zero. We know that this *cannot* lead to any *larger* percentage error than that provided by the MUPE/IRLS method, but it *may* lead to possibly *smaller* error. At worst, the error and bias will be the same as that given by MUPE/IRLS; at best, the error will be smaller and the bias will be closer to zero.

FIGURE 1. EXCEL SPREADSHEET BEFORE OPTIMIZATION.
(See Briefing for this Figure)

FIGURE 2. EXCEL SPREADSHEET AFTER OPTIMIZATION.
(See Briefing for this Figure)

As illustrated in Figure 1, the Excel SOLVER routine has the capability to produce a constrained optimization solution. Note that the bias cell is constrained to zero, while the standard percentage error cell is minimized to calculate the coefficients of the ZPB/MPE CER.

Figure 2 illustrates the Excel screen after SOLVER has produced a constrained optimization solution. Compare the magnitudes of the standard percentage error and bias cells, respectively, in Figures 1 and 2.

Results of several case studies on the relationships between MPE, MUPE/IRLS, and ZPB/MPE CERs are reported in the final section of this report below. Note that, in every case, the ZPB/MPE solution is at least as good as the MUPE/IRLS solution in both the standard percentage error and the percentage bias categories, but falls short of the percentage error available with the unconstrained MPE solution. A understanding of the relative magnitudes of the errors being traded off is also provided by the various tables of results.

CASE STUDIES

Figure 3 lists three sets of data points that we will use to illustrate the standard error of the estimate and the bias of candidate CERs derived by all three methods described above. These three data sets are examples of the type that support the USCM-7 model. However, CER quality would be higher for actual USCM-7 CERs because they are not restricted to one cost driver as we are in this report. It is important to remember that the results provided here are not to be considered typical of USCM-7 CERs, but serve only to illustrate the relative quality obtainable by applying the ZPB/MPE procedure instead of the MPE and MUPE/IRLS techniques.

FIGURE 3. THREE SAMPLE DATA SETS.

| Example 1 Data | | Example 2 Data | | Example 3 Data | |
|----------------|-------|----------------|-------|----------------|------|
| y | x | y | x | y | x |
| 357.79 | 6.90 | 2045.42 | 38.59 | 134.96 | 4.18 |
| 823.70 | 11.79 | 1619.62 | 28.92 | 2.05 | 0.32 |
| 652.31 | 10.23 | 2079.58 | 23.30 | 5.35 | 0.57 |
| 278.81 | 6.74 | 918.85 | 21.11 | 64.64 | 2.34 |
| 1066.73 | 16.70 | 1231.13 | 17.54 | 32.85 | 0.50 |
| 437.44 | 8.05 | 3641.96 | 27.60 | 95.42 | 2.70 |
| 1219.83 | 23.46 | 1314.85 | 16.20 | 66.22 | 4.54 |
| 368.38 | 16.50 | 1128.39 | 34.89 | 112.23 | 4.42 |
| | | 3989.48 | 46.61 | 29.24 | 0.55 |
| | | 3130.08 | 65.90 | 123.09 | 0.79 |
| | | 376.47 | 14.63 | 28.66 | 0.20 |
| | | 9028.31 | 50.10 | 16.93 | 0.80 |
| | | 2786.09 | 38.10 | 218.20 | 2.40 |
| | | 2497.71 | 73.21 | | |
| | | 2051.06 | 64.81 | | |
| | | 7008.74 | 41.60 | | |

Figures 4 and 5 list, respectively, standard percentage errors of the estimate and bias of each of six CERs of different forms optimally selected by each of the three methods we have been discussing using Example 1 data.

FIGURE 4. PERCENTAGE ERRORS OF CER FORMS FIT TO EXAMPLE 1 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|---------|-------------|---------|---------|
| $y = bx$ | 27.814% | 28.806% | 28.806% | 29.143% |
| $y = a + bx$ | 29.539% | 31.040% | 30.555% | 31.464% |
| $y = a + b \log x$ | 26.884% | 28.268% | 27.644% | 28.204% |
| $y = bc^x$ | 34.135% | 35.793% | 35.732% | 35.904% |
| $y = bx^c$ | 29.932% | 31.270% | 30.992% | 31.430% |
| $y = a + bx^c$ | 29.839% | 30.438% | 29.994% | 31.341% |

Note that the ZPB/MPE percentage error never exceeds the corresponding MUPE/IRLS percentage error and is sometimes quite a bit smaller. Take note also of the nonzero bias of the MPE CERs and the increase in standard error required to bring the bias to zero. For purposes of comparison only, we have calculated and listed in the far-right column the percentage error and bias of the MPE CER that is constrained to zero additive bias, the so-called ZAB/MPE CER. This CER will always have larger percentage error than the ZPB/MPE CER, as well as nonzero percentage bias.

FIGURE 5. PERCENTAGE BIAS OF CER FORMS FIT TO EXAMPLE 1 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|--------|-------------|---------|---------|
| $y = bx$ | 6.770% | 0.000% | 0.000% | -1.090% |
| $y = a + bx$ | 6.551% | 0.000% | 0.000% | -1.439% |
| $y = a + b \log x$ | 5.415% | 0.000% | 0.000% | -1.128% |
| $y = bc^x$ | 8.739% | 0.000% | 0.000% | 0.393% |
| $y = bx^c$ | 6.720% | 0.000% | 0.000% | -0.479% |
| $y = a + bx^c$ | 5.501% | 0.000% | 0.000% | -1.177% |

Figures 6 and 7 list, respectively, the standard percentage error of the estimate and the bias of each of six CERs of different forms optimally selected by each of the three methods we have been discussing using Example 2 data.. As before, the ZPB/MPE percentage errors never exceed the corresponding MUPE/IRLS percentage errors and are sometimes quite a bit smaller. Note also the nonzero bias of the MPE CERs and the increase in standard error required to bring the bias to zero.

FIGURE 6. PERCENTAGE ERRORS OF CER FORMS FIT TO EXAMPLE 2 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|---------|-------------|---------|---------|
| $y = bx$ | 53.851% | 63.109% | 63.109% | 63.866% |
| $y = a + bx$ | 52.258% | 61.390% | 59.902% | 65.638% |
| $y = a + b \log x$ | 53.275% | 60.978% | 58.181% | 60.722% |
| $y = bc^x$ | 56.791% | 71.589% | 67.032% | 74.124% |
| $y = bx^c$ | 53.321% | 64.011% | 61.519% | 66.108% |
| $y = a + bx^c$ | 53.125% | 61.270% | 60.461% | 63.478% |

FIGURE 7. PERCENTAGE BIAS OF CER FORMS FIT TO EXAMPLE 2 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|---------|-------------|---------|---------|
| $y = bx$ | 27.197% | 0.000% | 0.000% | -1.182% |
| $y = a + bx$ | 23.892% | 0.000% | 0.000% | -3.891% |
| $y = a + b \log x$ | 19.080% | 0.000% | 0.000% | -1.364% |
| $y = bc^x$ | 28.222% | 0.001% | 0.000% | -0.491% |
| $y = bx^c$ | 24.876% | 0.000% | 0.000% | -1.212% |
| $y = a + bx^c$ | 22.864% | 0.010% | 0.000% | -1.671% |

Figures 8 and 9 list, respectively, the standard percentage error of the estimate and the bias of each of six CERs of different forms, optimally selected by each of the three methods we have been discussing using Example 3 data.. Again, the ZPB/MPE percentage errors never exceed the corresponding MUPE/IRLS percentage errors and are sometimes quite a bit smaller. Continue to note the nonzero bias of the MPE CERs and the increase in standard error required to bring the bias to zero.

FIGURE 8. PERCENTAGE ERRORS OF CER FORMS FIT TO EXAMPLE 3 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|---------|-------------|---------|----------|
| $y = bx$ | 69.711% | 93.878% | 93.878% | 134.220% |
| $y = a + bx$ | 68.176% | 88.529% | 87.527% | 82.492% |
| $y = a + b \log x$ | 63.393% | 78.072% | 78.034% | 73.032% |
| $y = bc^x$ | 69.578% | 95.767% | 90.554% | 85.440% |
| $y = bx^c$ | 65.260% | 83.208% | 81.599% | 76.095% |
| $y = a + bx^c$ | 66.426% | 87.059% | 82.786% | 76.605% |

FIGURE 9. PERCENTAGE BIAS OF CER FORMS FIT TO EXAMPLE 3 DATA.

| FUNCTION | MPE | MUPE (IRLS) | ZPB/MPE | ZAB/MPE |
|--------------------|---------|-------------|---------|----------|
| $y = bx$ | 44.866% | 0.000% | 0.000% | -36.971% |
| $y = a + bx$ | 39.347% | 0.000% | 0.000% | 6.785% |
| $y = a + b \log x$ | 34.018% | 0.000% | 0.000% | 7.564% |
| $y = bc^x$ | 40.963% | 0.000% | 0.000% | 5.962% |
| $y = bx^c$ | 36.037% | 0.000% | 0.000% | 8.070% |
| $y = a + bx^c$ | 33.940% | 0.000% | 0.000% | 7.557% |

SUMMARY

General-error regression separates the problem of whether estimating error should be additive (expressed as a uniform dollar value across the board) or multiplicative (expressed as a percentage of the estimate) from the problem of whether the functional relationship is linear or nonlinear. It turns out, though, that relationships derived by minimizing percentage (i.e., multiplicative) error contain significant positive bias (i.e., they tend to overestimate the actual values of the dependent variable). Because functional relationships cannot be optimized with respect to more than one criterion, the analyst must decide whether to insist upon minimum possible percentage error or to accept an increase in percentage error in trade for a reduction in bias. The method of iteratively reweighted least squares appears to be useful in this respect, but the fact that it does not seem to be optimal in any particular respect leaves room for a more intellectually satisfying solution. In this report the general-error regression problem is formulated as a constrained nonlinear optimization problem, with percentage standard error of estimation optimized (i.e., minimized), subject to percentage bias being zero. Naturally, the percentage error turns out to be somewhat larger than it would be if the bias were unconstrained, but in general not as large as that given by iteratively reweighted least squares. In short, zero bias is achieved with the smallest possible increase in standard error.

ACKNOWLEDGMENTS

The authors would like to thank Jonathan F. Binkley for his detailed study of the mathematics of bias vs. standard error and his review of the iteratively reweighted least squares method. We would also like to thank Richard H. Lucas for his development of a PC-compatible C++ stand-alone piece of software that implements general-error regression and gives us a validation capability independent of Excel. We would also like to thank Marie Calamaria and Sandra E. McCarthy for their excellent work in typing this technical material.

REFERENCES

1. Bickel, P.J. and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, Inc., 1977 pages 132-141.
2. Book, S.A. and P.H. Young, "Optimality Considerations Related to the USCM-6 'Ping Factor'," ICA/NES National Conference, Los Angeles, CA, 20-22 June 1990, 40 briefing charts.
3. Bradu, D. and Y. Mundlak, "Estimating in Lognormal Linear Models," *Journal of the American Statistical Association*, Vol. 65 (March 1970), pages 198-211.
4. Draper, N.R. and H. Smith, *Applied Regression Analysis*, (2nd Edition), New York: John Wiley, 1981, pages 43-45, 90-98.
5. Eskew, H.L. and K.S. Lawler, "Correct and Incorrect Error Specifications in Statistical Cost Models," *Journal of Cost Analysis*, Spring 1994, pages 105-123.

6. Eskew, H.L., "Tutorial on Log-Linear Regression," *National Estimator*, Spring 1994, pages 10-13.
7. Farnum, N.R., "Improving the Relative Error of Estimation," *The American Statistician*, Vol. 44 (November 1990), pages 288-289.
8. Gallant, A.R., *Nonlinear Statistical Models*, John Wiley & Sons, 1987, page 473.
9. Heien, D.M., "A Note on Log-Linear Regression," *Journal of the American Statistical Association*, Vol. 63 (September 1968), pages 1034-1038.
10. Hu, S.P. and A.R. Sjøvold, "Error Corrections for Unbiased Log-Linear Least Square Estimates," Tecolote Research, Inc., Report No. TR-006/1, October 1987, page 51.
11. Hu, S.P. and A.R. Sjøvold, "Multiplicative Error Regression Techniques," Tecolote Research, Inc., 1994, 31 pages. (Also available as Appendix C of Reference 17.)
12. Jennrich, R.I. and R.H. Moore, "Maximum Likelihood Estimation by Means of Nonlinear Least Squares," American Statistical Association, *Proceedings Statistical Computing Section*, 1975, pages 57-65.
13. Miller, D.M., "Reducing Transformation Bias in Curve Fitting," *The American Statistician*, Vol. 38 (May 1984), pages 124-126.
14. Neyman, J. and E.L. Scott, "Correction for Bias Introduced by a Transformation of Variables," *Annals of Mathematical Statistics*, Vol. 31 (September 1960), pages 643-655.
15. Nguyen, P., N. Lozzi, W. Galang, et al., *Unmanned Space Vehicle Cost Model, Seventh Edition*, U.S. Air Force Space and Missile Systems Center (SMC/FMC), 2430 E. El Segundo Blvd., Suite 2010, Los Angeles AFB, CA 90245-4687, August 1994, xvi + 452 pages.
16. SAS Institute Inc., *SAS/Statistical User's Guide, Volume 2*, 1990, pages 1135-1193.
17. Seber, G.A.F, and C.J. Wild, *Nonlinear Regression*, John Wiley & Sons, 1989, pages 37, 46, 86-88.

This paper appeared in the literature as follows: "Minimum-Percentage-Error Regression under Zero-Bias Constraints", *Proceedings of the Fourth Annual U.S. Army Conference on Applied Statistics, 21-23 October 1998*, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.