Professional Paper 534 / June 1993

Correct and Incorrect Error Specifications in Statistical Cost Models

Henry L. Eskew • Kletus S. Lawler

Center for Naval Analyses
4401 Ford Avenue • P.O. Box 16268 • Alexandria, VA 22302-0268

Copyright CNA Corporation/Scanned September 2006

Approved for public release; distribution unlimited

The ideas expressed in this paper are those of the authors. The paper does not necessarily represent the views of the Center for Naval Analyses, the Department of the Navy or the Department of Defense.

CORRECT AND INCORRECT ERROR SPECIFICATIONS IN STATISTICAL COST MODELS

Henry L. Eskew
Kletus S. Lawler
Center for Naval Analyses
Alexandria, VA

Abstract

A functional form often employed in cost-estimating relationships is $Y = \alpha X^{\beta}$, where Y and X are the cost and predictor variables, respectively, and α and β are parameters to be estimated from empirical data. Logarithmic transformation of either side of this *intrinsically* linear relationship leads to $Y^* = \alpha^* + \beta X^*$, with the asterisks denoting logs. What typically happens in practice is that ordinary least squares (OLS) regression is applied to a set of observations on Y^* and X^*, generating estimates of α^* and β. From a statistical point of view, however, justification of this process requires the presence of a multiplicative random error term in the untransformed model. In addition, the multiplicative structure gives rise to variance in cost that changes systematically with X. If the error term is additive, the variance in cost will be constant, but the parameters in that model must be estimated by nonlinear least squares (NLS) applied to the untransformed variables. In general, OLS and NLS lead to different parameter estimates and different cost predictions, and frequently analysts are unsure as to which method is correct. This paper reports on a Monte Carlo study of the accuracy of each estimation method under correct and incorrect error specifications. The multi-predictor model that serves as the basis of the study is representative of models with which cost analysts frequently deal. Results suggest that OLS, applied to the logs of the variables, may be the preferred method under either specification, provided the estimate of α is properly developed. This finding, however, is applicable strictly to the issue of parameter estimation. Accuracy of cost prediction requires further analysis.

Background

Cost-estimating relationships are the bread and butter of applied cost analysis. Those formulations typically have some measure of cost (unit, average, or total) as the dependent variable, and one or more predictor variables on the right-hand side. The predictors capture important design, performance, and programmatic dimensions of the item whose cost is to be estimated.

Considering, for simplicity, the case of a single predictor, a functional form often used for the cost relationship is:

$$Y = \alpha X^{\beta},$$

where Y and X are the cost and predictor variables, respectively, and α and β are parameters to be estimated from empirical data. Obtaining accurate estimates of the model's parameters contributes to the accuracy of predictions and is doubly advantageous in that it permits meaningful sensitivity analysis—"what if" drills—to be performed on the design, performance, and programmatic variables of interest. This paper's primary concern is with accuracy of parameter estimation, although some aspects of prediction are addressed in appendix A. A subsequent paper will deal with predictive accuracy in more depth.

The functional form displayed above enjoys widespread acceptance for at least two reasons. First, in many cases analysts have reason to believe that (unit) cost varies nonlinearly with, for instance, weight ($0 < \beta < 1$) or procurement quantity ($-1 < \beta < 0$). And second, the function is *intrinsically* linear, meaning that a logarithmic transformation of both sides produces a linear relationship:

$$Y^* = \alpha^* + \beta X^*,$$

with the asterisks denoting logs.[1]

What frequently happens in practice is that ordinary least squares (OLS) regression is applied to a set of observations on Y^* and X^*, generating estimates of α^* and β. Correlation coefficients are computed, significance tests carried out, and a value of Y^* predicted, conditional on some value of X^*. Computing the antilog of the predicted Y^*, sometimes with an associated prediction interval, usually completes the cost-estimating process.

From a statistical point of view, the extent to which this process can be justified depends on the exact stochastic specification taken for the model. The logarithmic transformation requires the assumption of a multiplicative random error term (V) in the original formulation:

$$Y = \alpha X^{\beta} V,$$

thereby leading, after transformation, to a function with additive error, which is standard in the linear regression model:

[1] A third feature that some analysts find attractive is that a value of zero for X means that Y will also be zero.

$$Y^* = \alpha^* + \beta X^* + V^*.$$

Moreover, for the usual—or at least some of the usual—regression results to hold, a further assumption about the form of V is needed:

$V = \exp(U)$,

with U assumed to be a normally distributed random variable having mean zero and variance σ^2. Thus, V is lognormally distributed.

The somewhat peculiar properties of this model have been thoroughly examined in the statistical and econometric literature [1 through 3] and were revisited not long ago in a cost analysis setting [4,5]. They are summarized in appendix A. One of those properties, which is directly relevant to the focus of this paper, concerns the conditional variance of Y. This specification with multiplicative error gives rise to a *heteroscedastic* Y, meaning that the conditional variance of Y changes systematically with X.[2] (If $\partial Y/\partial X > 0$, $Var(Y|X)$ will increase with X; otherwise it will decrease.) A conceptual sketch of the increasing variance is shown in panel A of figure 1, where $(0 < \beta < 1)$ is assumed. Panel B depicts the same nonlinear form but with *homoscedastic* (constant) variance. Its corresponding stochastic specification is:

$$Y = \alpha X^{\beta} + W$$

where W is the normally distributed error with mean zero and variance σ^2. Clearly this function cannot be linearized by logarithmic transformation, thus ruling out direct application of OLS regression. Its parameters can be estimated, however, by nonlinear least squares (NLS), a procedure whose properties are well understood and for which computer codes are readily available.[3]

Summing up the discussion thus far, if an analyst believes that the variance in cost will increase (decrease) as cost increases (decreases), the multiplicative error specification is the correct one. That specification also justifies use of standard regression procedures on the logarithms of the variables. (But as appendix A explains, matters are still not entirely straightforward.) On the other hand, constant error variance calls for an additive error term and necessitates parameter estimation by nonlinear least squares applied to the untransformed variables.[4] The central question of this paper is, Does the difference matter? The answer will be sought through a Monte Carlo experiment. Before discussing that experiment, however, we describe what motivated it: a cost-estimating relationship

---

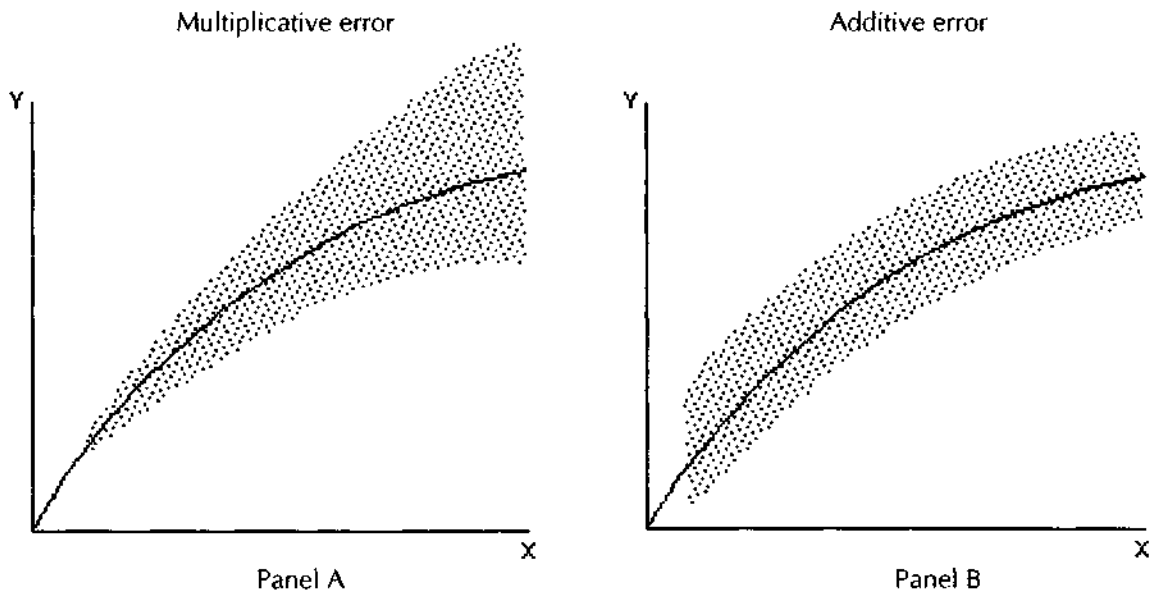
[2]See, for example, [6, p. 121] and also appendix A.

[3]Nonlinear least squares computations reported in this paper were carried out by using the NLIN procedure in the statistics package licensed by the SAS Institute Inc. The iterative search method employed is GAUSS.

[4]By assuming the additive error term to be normally distributed, the nonlinear least squares estimators are also *maximum likelihood* estimators. See [7, p. 262].

whose OLS and NLS parameter estimates differed substantially, leading to correspondingly different predictions of the cost in question.

Figure 1. Alternative error specifications



Motivation for the Monte Carlo study

In *Augustine's Laws*, Norman Augustine put forth several observations that, despite (or perhaps because of) the light-hearted nature of his book, have influenced thinking in the defense analytic community. One that has considerable interest for cost analysts is:

> ...the cost of an individual (tactical) airplane has unwaveringly grown by a factor of four every ten years. This rate of growth seems to be an inherent characteristic of such systems, with the unit cost being most closely correlated with the passage of time rather than with changes in maneuverability, speed, weight, or other technical parameters. [8, p. 140]

This proposition was tested recently at the Center for Naval Analyses [9]. A consistent source of procurement cost data was available for 17 fighter and attack aircraft, a grouping that had been highlighted in Augustine's work. We constructed a regression model with unit-100 flyaway cost (in millions of FY 1990 dollars) as the dependent variable.[5] The predictors were:

[5]Strictly speaking, the variable was *approximate* unit-100 flyaway cost. The unit cost was approximated by the average cost for the lot that included unit 100.

1. A time variable, defined as the year in which the first operational unit was delivered (19xx - 1900)
2. Empty weight (in thousands of pounds)
3. Maximum speed at altitude (in knots)
4. Production rate in the year that included unit 100.

This formulation departed from Augustine's proposition in three respects. First, we measured cost in constant rather than current (then-year) dollars—the latter being the measure he used. Use of constant dollars permits trends and relationships unique to military aircraft to be ferreted out from those linked to the general increase in manufacturing prices over time. Second, we made no effort to capture maneuverability as a technical characteristic; its quantification is not straightforward and data are not generally available. Finally, we added a fourth variable in the belief that unit cost varies inversely with production rate.[6] These differences notwithstanding, estimation and testing of the model's parameters provides at least a partial test of what Augustine had to say.

The form of this multi-predictor model was specified as:

$$C = \alpha W^{\beta} S^{\lambda} R^{\delta} (1 + \rho)^{T} V,$$

where

C = unit-100 flyaway cost
W = empty weight
S = max speed
R = production rate
T = year of first delivery
V = random error term
$\alpha, \beta, \lambda, \delta, \rho$ = model parameters.

With transformation, this relationship is linear in T and in the logs of all other variables.[7] Its parameters were estimated by applying OLS to that set of variables. Results are in table 1 below.

From a statistical point of view, these results appear to be quite satisfactory. Each parameter estimate is significant at better than the 0.02 level, and the R^2 value (applicable to the log of cost) is reasonably high. (Implications for the Augustinian test are straightforward. Aircraft weight, speed, and production rate each have important influences on cost, and the rate of *real* growth in cost, estimated by subtracting 1.0 from

[6]See, for example, [10, p.66].
[7]The expression $(1 + \rho)^T$ is standard representation of growth in discrete (period) form. Its counterpart in continuous time is $\exp(\rho T)$.

the antilog of the parameter estimate associated with T, is only a bit more than 3 percent per year.)

Table 1. Parameter estimates for linearized cost model ($R^2 = 0.937$)

Variable (parameter)	Parameter estimate	Standard error	t-ratio	p-value
Intercept (α^*)	-4.2831	1.3208	-3.243	0.0071
Weight (β)	0.7128	0.1701	4.191	0.0013
Speed (λ)	0.6188	0.1509	4.101	0.0015
Prod. rate (δ)	-0.3910	0.1425	-2.744	0.0178
Time ($1 + \rho$)*	0.0326	0.0059	5.503	0.0001

Despite the satisfactory appearance of these results, there is no *a priori* way of knowing that the multiplicative error specification is the correct one. Specifying an additive error and employing NLS estimation produces the results shown in table 2.

Table 2. Parameters estimated by nonlinear least squares (pseudo-$R^2 = 0.876$)

Variable (parameter)	Parameter estimate	Asymptotic stnd. error
Intercept (α)	0.0552	0.0922
Weight (β)	0.4739	0.1894
Speed (λ)	0.5247	0.2086
Prod. rate (δ)	-0.3543	0.1820
Time ($1 + \rho$)	1.0315	0.0070

A few clarifying remarks will aid in comparing the results in table 2 with those in table 1. First, in standard linear regression, the sum-of-squares explained by the regression (RSS) partitions as follows:

$$\text{RSS} = \text{TSS} - \text{ESS},$$

where TSS and ESS are total sum-of-squares and error sum-of-squares, respectively. The R^2 statistic is then defined as RSS/TSS or, equivalently, (1 - ESS/TSS). With nonlinear estimation, however, that partitioning is not assured. What has been called pseudo-R^2 here is computed as (1 - ESS/TSS), which is closely analogous to the conventional measure. In comparing the value of that statistic in table 2 with its somewhat higher counterpart in table 1, one should remember that the latter represents the fraction of the *log* of cost that is explained, whereas the former pertains to the untransformed variable.

Second, concerning the format of table 2, the assumption that the additive error is normally distributed implies an asymptotic normal distribution for the NLS estimators, but

the usual small-sample t-tests are not applicable. The estimated asymptotic standard errors are shown, however, and may be compared against their corresponding parameter estimates, and against the standard errors in table 1, to provide some insights into estimation accuracy.[8] Such a comparison suggests, at least on the surface, that the estimates in table 1 are more precise than those in table 2.

Last, the estimate for the intercept in table 1 is in fact an estimate of α^*, the log of α. In table 2, however, the NLS procedure estimates α directly. The antilog of the intercept value in table 1 is 0.0138. While that value, for reasons explained in appendix A, is an upwardly biased estimate of α, it is nevertheless the one typically used in practice. It is also very different—only one-fourth as large—as the estimate of α in table 2. Differences in the other estimates are not nearly as great, but they are by no means negligible.

In light of the above, it is natural to ask: Do these two sets of estimates lead to similar or dissimilar cost predictions? Table 3 posits values of predictor variables for several notional aircraft and shows the predictions of unit-100 flyaway cost generated by the OLS and NLS estimates.

Table 3. Comparative cost predictions and predictor values (costs in millions of FY 1990 dollars)

OLS Pred.	NLS Pred.	Weight	Speed	Prod. rate	Time
76.3	66.4	43.4	500	26	103
80.1	68.8	45.2	516	26	103
97.5	79.1	56.2	552	26	103
48.4	45.3	35.2	526	48	100
41.1	42.1	23.3	575	60	105

The asymmetry among the differences in predicted costs is striking. For the heavier aircraft, the OLS-based values are from 15 to 20 percent higher. That can be attributed in large part to the much higher parameter estimate for the weight variable (0.7128 compared with 0.4739). But for the lighter aircraft, the two sets of costs are very nearly the same. Apparently the higher NLS estimate of α is compensating for its other, smaller estimates.

All of this is somewhat disquieting. First, uncertainty exists as to the correct error specification. Second, because of that, there is uncertainty as to which set of parameter estimates should be used. And last, the different estimates lead, at least in some cases, to very different cost predictions.[9] A question that arises at this point is: If the correct error

[8]Asymptotic 95-percent confidence intervals may be constructed by adding algebraically, to each parameter estimate, $\pm 1.96s$, where s is the associated asymptotic standard error.

[9]A reviewer of this paper reran the two regressions using the "bi-weight" options in a statistical package called DATA-PLOT. His motivation was to reduce the effects of outliers on the parameter estimates and

specification is additive (multiplicative), and if OLS (NLS) is mistakenly chosen as the method of estimation, what will be the consequences in terms of loss of estimation accuracy? Monte Carlo methods can help answer that question.

The Monte Carlo experiment

The first step in a Monte Carlo study is to postulate the complete model that will serve as the basis of the study. This includes defining the model's variables, specifying its functional form, assuming numerical values for its parameters, and postulating the form and parameters of the probability distribution characterizing the error term. Next, a set of "observations" on the predictor variables—hereafter the *design matrix*—is created so as to be representative, in size and other characteristics, of the application being studied. Then a "trial" is carried out by:

1. Obtaining a random drawing, one for each observation, from the postulated error distribution
2. Generating a set of "observed" values of the dependent variable (Y) by combining each random error algebraically with the associated error-free value of Y that was computed by applying the assumed numerical parameters to the design matrix
3. Estimating a vector of parameter values ($\boldsymbol{p}$) from the simulated data, using each estimation method being evaluated in the study
4. Recording each of the p_j for later summary and comparison with the vector of postulated parameters ($\boldsymbol{\pi}$).

A great many trials are then conducted in succession—typically several hundred. Results of the experiment are summarized and assessed as to estimation accuracy. The traditional measure of accuracy is mean squared error (MSE), which is related to variance (Var) and squared bias (Bias^2) as follows:

$$\begin{array}{ccccc} \underline{\text{MSE}} & \equiv & \underline{\text{Var}} & + & \underline{\text{Bias}^2} \\ E(p_j - \pi_j)^2 & \equiv & E[p_j - E(p_j)]^2 & + & [E(p_j) - \pi_j]^2. \end{array}$$

The model chosen for the experiment is the one described in the preceding section. Absent an error term but with assumed parameter values, it is:

$$C = 0.3W^{0.6}S^{0.5}R^{-0.4}(1.03)^T.$$

Except for the multiplicative constant, which was chosen arbitrarily, the other parameter values are rough averages of the OLS and NLS estimates in tables 1 and 2. The design matrix consists of essentially the same 17 observations on the predictor variables that made up the fighter and attack aircraft data base. Those data are provided in appendix B. In terms of number of observations, number of variables, and degree of independent and

---

predictions. In his results, the parameter estimates and predictions were in closer agreement than those reported above. He concluded, however, that the concerns expressed in the paper remain valid in general.

interdependent variability among the predictors, this model and data base are thought to be representative of those cost analysts frequently encounter.

Insofar as the Monte Carlo study is concerned, the crucial feature of the model is the random error term Two variants are examined: one where the error term is multiplicative and denoted by V, and the other where it is additive and denoted by W. In the first, $ln(V)$ is postulated to be normally distributed with mean zero and variance 0.05. In the second, W is likewise normal; its mean is zero and variance 140. The variance values were chosen to produce R^2 (or pseudo-R^2) statistics in the neighborhood of 0.90.

Model with multiplicative error

Table 4 below, which pertains to the model with multiplicative error, displays results summarized over 500 trials for each specification. There the appropriate estimation method is OLS, applied to the transformed relationship. For each parameter, the table shows (1) its actual (postulated) value, (2) the mean of the 500 estimates, (3) the variance and squared bias of those estimates, and (4) the root-mean-squared error (RMSE) as a fraction of the actual parameter value.[10]

Table 4. Monte Carlo results for the model with multiplicative error

Parameter	Actual value	Mean	Variance	Bias^2	RMSE/Act.
		—OLS	estimates—		
α – (est. 1)	0.3	0.9037	5.5243	0.3644	8.0889
α – (est. 2)	0.3	0.2875	0.7257	0.0002	2.8400
β	0.6	0.6102	0.0517	0.0001	0.3793
λ	0.5	0.5021	0.0231	0.0000	0.3042
δ	-0.4	-0.3983	0.0268	0.0000	-0.5719
$1 + \rho$	1.03	1.0298	0.0000	0.0000	0.0062
R^2	- - -	0.9030	0.0015	- - -	- - -
		—NLS	estimates—		
α	0.3	1.3239	10.8063	1.0483	11.4768
β	0.6	0.6045	0.0766	0.0000	0.4612
λ	0.5	0.5146	0.0435	0.0002	0.4182
δ	-0.4	-0.4301	0.0514	0.0009	-0.6139
$1 + \rho$	1.03	1.0305	0.0001	0.0000	0.0080
R^2	- - -	0.8997	0.0026	- - -	- - -

For the OLS results, which we consider first, two sets of estimates of α are generated. The first is simply the antilog of a^*, where a^* is the estimate of α^*. As noted earlier, this

[10]The authors acknowledge that although the Monte Carlo results are reported to four decimal places, 500 trials are almost certainly too few to provide that level of accuracy.

is the estimate used in practice almost without exception. The second is the antilog of $[a^* - \text{est. } Var(a^*)/2]$.[11] Appendix A establishes this as the theoretically superior estimator. Results in table 4 indicate that its superiority is far from just theoretical. The conventional estimates of α exhibit serious flaws. Their variance is quite large and there is substantial bias. Ordinarily there is no causal connection between variance and bias, but in this case there is.[12] The hypothetical data in table 5 describe the connection on an intuitive level.

Table 5. Hypothetical estimators of $ln(\varpi)$, where $\varpi = 1$

	First estimator			Second estimator		
	w_1^*	$p(w_1^*)$	$\exp(w_1^*)$	w_2^*	$p(w_2^*)$	$\exp(w_2^*)$
	1	0.25	2.718	2	0.25	7.389
	0	0.50	1.000	0	0.50	1.000
	-1	0.25	0.368	-2	0.25	0.135
Mean	0.0	---	1.272	0.0	---	2.381
Variance	0.5	---	0.764	2.0	---	8.484
Bias^2	0.0	---	0.074	0.0	---	1.907

Imagine that the parameter of interest in the table is $\varpi = 1$. What is being estimated, however, is its logarithm, $ln(\varpi) = \varpi^* = 0$. The table gives probability distributions of two hypothetical estimators of ϖ^*, labeled w_1^* and w_2^*, with w_2^* having the larger variance. Both w_1^* and w_2^* are unbiased estimators of ϖ^* in that $E(w_1^*) = E(w_2^*) = \varpi^* = 0$. But it is the antilogs of the estimates that are typically computed, and the antilogs constitute a biased estimator of ϖ, with the bias increasing systematically as variance increases. Only in the special case of zero variance is there no bias. Interestingly, OLS estimates of the log of $(1 + \rho)$ in table 4 border on that case. The conditioning of the design matrix is such that the variance of the antilogs of those estimates—from the Monte Carlo results— is zero through four decimal places. Consequently the antilogs produce very precise estimates of that parameter. The remaining parameters are estimated by OLS with virtually no bias, as statistical theory leads us to expect. However, their root-mean-squared errors are 30 to 40 percent as large as the actual values, a reminder that unbiasedness and estimation precision are not the same thing.

Turning to the results generated by nonlinear least squares—an inappropriate estimation method for the model with multiplicative error—the estimates of α are considerably worse than either set of OLS estimates. There is a very large variance and a large positive bias. That is attributable in brief to the "inflated" values of cost produced by the multiplicative error. The other parameters are estimated essentially without bias, and $(1 + \rho)$ is again

[11]The estimated variance of a^* is the conventional one: the square of the estimated standard error of a^* or, equivalently, the first element off the main diagonal of the estimated variance-covariance matrix.
[12]The method of *ridge regression,* where the parameter estimates are biased in such a way as to achieve substantial variance reduction, is one notable exception to this statement.

estimated with considerable precision, but the relative RMS errors are roughly one-third larger than their OLS counterparts. In short, nonlinear least squares is demonstrated to be an inferior estimation method for this model, as expected. We note also that the mean R^2 statistics are virtually identical for OLS and NLS. This is somewhat discouraging because it suggests that, when doubt exists, R^2 cannot be relied on to confirm OLS as the correct estimation method.

Model with additive error

Results in table 6 are generated by the model having additive error, with NLS being the appropriate estimation method. We note first that here, unlike the results with multiplicative error, the mean R^2 value is much higher for the appropriate method. Then, a quick comparison of the corresponding error ratios confirms, with one glaring exception, what was expected: NLS performs better. Except for the estimates of α, the OLS estimates tend to be biased upward and exhibit error ratios roughly twice as large as the NLS ratios. (Note, however, that the OLS error ratios are not appreciably different from what they were in table 4, and that the ratio for the preferred estimate of α is even lower than its table 4 value.)

Table 6. Monte Carlo results for the model with additive error

Parameter	Actual value	Mean	Variance	Bias^2	RMSE/Act.
		—OLS estimates—			
α – (est. 1)	0.3	1.1699	16.5138	0.7567	13.8525
α – (est. 2)	0.3	0.1808	0.3447	0.0142	1.9969
β	0.6	0.6216	0.0575	0.0004	0.4014
λ	0.5	0.5226	0.0396	0.0005	0.4000
δ	-0.4	-0.4360	0.0422	0.0013	-0.5214
$1+\rho$	1.03	1.0316	0.0001	0.0000	0.0079
R^2	---	0.8740	0.0043	---	---
		— NLS estimates—			
α	0.3	0.4994	0.9407	0.0397	3.3005
β	0.6	0.6049	0.0157	0.0000	0.2090
λ	0.5	0.5014	0.0172	0.0000	0.2623
δ	-0.4	-0.4033	0.0131	0.0000	-0.2863
$1+\rho$	1.03	1.0302	0.0000	0.0000	0.0043
R^2	---	0.9442	0.0005	---	---

The situation with respect to α is surprising. There is no theoretical reason to expect the amount of upward bias in the NLS estimates that emerged, and certainly no reason to expect an RMS error half-again as large as that for the preferred OLS estimate. Because this was such an unexpected result, it was pursued further from a number of perspectives.

First, to ensure that the result was not due to anomalies in this particular set of 500 trials, the experiment was repeated with additional sets of 500. All results were virtually identical. Then, hypothesizing that the difficulty might stem from the true parameter value's being relatively close to zero, we increased the value of α, first to 0.8 and then to 3.0, making proportional increases in the variance of the random error term.[13] Again, the same pattern of bias and estimation error emerged. Finally we examined frequency distributions of the 500 estimates of α and the other parameters as well. For α, the distributions are extremely right-skewed. Their medians are relatively close to the true parameters, but—as with all very right-skewed distributions—the means are well to the right of the medians, thus producing the upward bias. This suggests that the sample size here (17) is much too small for the property of asymptotic normality (of the estimators) to be manifest. That is not the case, however, with the distributions of the estimates of the other parameters. Those distributions center on the true values of the parameters and are Gaussian in shape.

Were it not for the finding pertaining to α, we would have concluded from the Monte Carlo experiment—without great fanfare—that the appropriate estimation method clearly outperforms its alternative. The obvious implication of that conclusion would be that analysts should endeavor, through examination of residual plots and by formal tests for heteroscedasticity, to ascertain the form of the random error embedded in any given data set. To the extent that a person is interested only in obtaining accurate estimates of the parameters other than α—for use in, say, system design or policy analysis—that implication remains valid. In cases where it is important to obtain an accurate estimate of α, our results suggest that applying ordinary least squares to the logarithms of the variables, and using $\exp[a^* - \text{est.var.}(a^*)/2]$ to estimate α, may be a better bet in either case. It is demonstrably better when the errors are multiplicative, and it seems to be a *safer* alternative even with additive errors, given the problems with the NLS estimates of α.

What may be more likely, however, is that the statistical work is being done to construct a mechanism for predicting some as-yet-unknown cost. Appendix A reviews relevant issues and procedures pertaining to prediction with the multiplicative-error model. But a larger question concerning prediction is one that directly parallels the parameter-estimation focus of this paper: If either OLS or NLS is used when in fact the other is called for, what will be the consequences? The authors are planning further Monte Carlo experiments centering on predictive accuracy in partial pursuit of the answer to this question.

---

[13]Had the error variance not been increased proportionally, the "noise" in the cost variable would tend to be negligible, producing R^2 values approaching 1.0. The parameter estimation process would then lose its statistical content, with there being only one vector of parameter values—the true values—that would constitute a "best fit."

Concluding remarks

As noted earlier, the model selected for study here—17 observations and 4 predictor variables, with R^2 values in the 0.90 range and data drawn from design and performance characteristics of actual aircraft—is thought to be representative of much of the applied work in statistical cost analysis. We consider it unlikely that small perturbations to the model's specification, dimensions, or data characteristics would materially alter the Monte Carlo results. In fact, we examined a few such perturbations, confirming a degree of robustness in the results. Nevertheless, universality can never be claimed for the outcome of a Monte Carlo experiment. What we have here can best be thought of as a single case study. With several such studies in hand, each dealing with the same general topic but differing in such a way as to broaden or deepen the topic's exploration, generalizations might then be justified. We hope sincerely that other researchers will be inclined in that direction.

Appendix A

Properties of the log-linear regression model

This appendix summarizes mean and variance properties and examines issues relating to parameter estimation and prediction with the linear-in-logs model described at the outset of the paper:

$$Y = \alpha X^{\beta} V,$$

where Y and X are the dependent and predictor variables, respectively, α and β the parameters to be estimated, and V a multiplicative random error term.[14] The model in logarithmic form is:

$$Y^* = \alpha^* + \beta X^* + V^*,$$

with the asterisks denoting logs. The exposition here is drawn from [1 through 3, 6].

The standard assumption is that V is lognormally distributed, i.e.,

$$V = \exp(U),$$

where U is normal with mean zero and variance σ^2. U is also assumed to be stochastically independent of X. Interest typically centers on the conditional mean of Y or, in the context of this paper, *expected cost*, conditional on some value of X. Thus we note:

$$E(Y|X) = \alpha X^{\beta} E(V|X) = \alpha X^{\beta} E(V).$$

But because of the assumption made about the form of V—an assumption that justifies direct application of ordinary least squares (OLS) to the logarithms of the variables—the expectation of V is not unity but instead, from the properties of the lognormal distribution:

$$E(V) = \exp[E(U) + \sigma^2/2].$$

Because $E(U) = 0$ by assumption, the conditional mean of Y becomes:

$$E(Y|X) = \alpha X^{\beta} \exp(\sigma^2/2)$$

[14]Economists refer to this functional form as *Cobb-Douglas*. It was the form adopted in the pioneering studies of production by C.W. Cobb and Paul H. Douglas. An interesting and sometimes attractive feature of the model is its property of *constant elasticity*, i.e., $(\partial Y/\partial X)(X/Y) = \beta$.

The importance of this in practice is that given a set of estimates for α^* and β, denoted by a^* and b, and given some predictor value X_0, the quantity $\exp[a^* + b(\ln X_0)]$ constitutes an estimate of the conditional *median* of Y, not its conditional mean.

Before turning to the matter of parameter estimation, we examine the conditional variance of Y. It is convenient to do so by first subtracting the conditional mean of Y from Y:

$$Y - E(Y|X) = \alpha X^{\beta} V - \alpha X^{\beta} E(V) = \alpha X^{\beta}[V - E(V)],$$

which makes it possible to write:

$$E[Y - E(Y|X)]^2 = E\{\alpha X^{\beta}[V - E(V)]\}^2 = (\alpha X^{\beta})^2 E[V - E(V)]^2.$$

Thus the conditional variance of Y will vary with X, being proportional to the square of the conditional median of Y, with $E[V - E(V)]^2$ being the variance of the multiplicative error.[15] This is the *heteroscedasticity* (in Y, not V) that was noted earlier in the paper.

Parameter estimation

We consider application of ordinary least squares to the transformed model:

$$Y^* = \alpha^* + \beta X^* + U.$$

This formulation satisfies all assumptions of the standard linear regression model and, by the Gauss-Markov theorem, OLS estimators of α^* and β are unbiased and have minimum variance among the class of linear unbiased estimators. If there were multiple predictor variables, the same statement would apply to the parameters associated with each. At issue, however, is the matter of obtaining an estimate of the original α that has the same desirable properties.

We denote the least-squares estimator of α^* by a^* and observe that, because a^* is normally distributed, $\exp(a^*)$ is lognormally distributed. From the earlier discussion of lognormal variables, we see that

$$E[\exp(a^*)] = \exp[E(a^*) + Var(a^*)/2] = \exp[\alpha^* + Var(a^*)/2] = \alpha\exp[Var(a^*)/2].$$

To recap in words, the expected value of the antilog of the estimated intercept is not equal to α but instead to the product of α and $\exp[Var(a^*)/2]$. Thus, the antilog of a^*—the estimate of α that is typically used in practice—is biased, with the bias being upward because $\exp[Var(a^*)/2]$ is greater than 1.0. If $Var(a^*)$ were known with certainty, the bias could be eliminated simply by subtracting $Var(a^*)/2$ from a^* before taking its antilog, i.e.,

[15]The variance of $V = \exp(U)$, with U having mean zero and variance σ^2, is $\exp(\sigma^2)[\exp(\sigma^2) - 1]$.

$$E\{\exp[a^* - Var(a^*)/2]\} = \exp(\alpha^*) = \alpha.$$

However, the $Var(a^*\}$can only be estimated because it involves the unknown σ^2. Nevertheless, subtracting one-half the *estimated* variance of a^*—in practice, one-half the square of the estimated standard error of that parameter estimate—will very nearly eliminate the bias.[16]

Prediction

A prediction of Y, conditional on X_0, is actually an estimator of $\exp(\alpha^* + \beta X_0^*)\exp(\sigma^2/2)$. The natural or "naive" estimator of $\exp(\alpha^* + \beta^*)$, the conditional median of Y, is $\exp(y)$, where $y = a^* + b(\ln X_0)$. But that estimator is biased upward, for the same reason that $\exp(a^*)$ is an upwardly biased estimator of α. In particular,

$$E[\exp(y)] = \exp[E(y) + Var(y)/2] = \exp[\alpha^* + \beta X_0^* + Var(y)/2].$$

To deal with that, we first require the estimated variance of y, which we denote by s_y^2:

$$Var(y) = Var(\boldsymbol{cb}') = s_y^2 = \boldsymbol{c}s^2(\boldsymbol{X'X})^{-1}\boldsymbol{c}',$$

where $\boldsymbol{c}$ is the row vector of prediction values—$(1 \;\; \ln X_0)$ in this case— $\boldsymbol{c}'$ is its transpose, $\boldsymbol{b}'$ is the transpose of the row vector $(a^* \;\; b)$, and $s^2(\boldsymbol{X'X})^{-1}$ is the estimated variance-covariance matrix of a^* and b.[17] Multiplying $\exp(y)$ by $\exp(-s_y^2/2)$ removes, or essentially removes, the bias. What remains is the median-to-mean adjustment. That is accomplished by multiplying the preceding result by $\exp(s^2/2)$. In sum, a prediction of the conditional mean of Y is given by:

$$Y_{pred} = \exp[a^* + b(\ln X_0)]\{\exp[(s^2/2) - (s_y^2/2)]\}.$$

As explained earlier, this predictor is very nearly but not exactly unbiased.

---

[16]It turns out that an exactly unbiased estimator of α is given by $\exp(a^*)F_\alpha$, where F_α is the sum of an infinite series that is a function of three statistics: the number of degrees of freedom in the regression, the first element off the main diagonal of the inverse of the raw moment matrix, and s^2. Similarly, an exactly unbiased estimator of $\exp(\sigma^2/2)$ involves a nearly identical series, but with the series being a function of only the number of degrees of freedom and s^2. See [1, 2, 3, and 4]. But the results in table 4 of this paper and the tabulated values of the series in [3] suggest that when $\exp[a^* - \text{est. } Var(a^*)/2]$ is used to estimate α, the remaining bias tends to be negligible.

[17]The scalar analog to this variance computation is as follows: If P is a random variable and k a constant in $Q = kP$, then $Var(Q) = Var(kP) = k^2Var(P)$.

Summary numerical example

The hypothetical example below will tie the preceding discussion together and demonstrate the use of standard regression output in applying the relevant concepts. Consider that the log of Y has been regressed on the log of X, with all assumptions underlying that process being satisfied, and that the following results were obtained:

Parameter (estimator)	Numerical estimate	Estimated variance-covariance matrix
α^* (a^*)	4.8386	0.1855 -0.0511
β (b)	0.6715	-0.0511 0.0147
σ^2 (s^2)	0.1260	—

Imagine first that we are interested in obtaining an unbiased—or nearly unbiased—estimate of α, which we denote by a. In the context of this example, that parameter might represent unit-one cost in a total cost function, with X representing aggregate procurement quantity. That estimate is given by:

$$a = \exp[a^* - est.Var(a^*)/2] = \exp[4.8386 - 0.1855/2] = 115.1.$$

Note that this estimate is some 9 percent lower than the naive or uncorrected estimate, $\exp(a^*) = \exp(4.8386) = 126.3$.

Our interest, however, is more likely to be in predicting the mean value of Y, conditional on some value of X_0. Let that value of X_0 equal 1000. The uncorrected prediction, which is also a prediction of the conditional *median* of Y, is simply:

$$y = \exp[a^* + b(\ln X_0)] = \exp[4.8386 + 0.6715(6.9078)] = 13{,}058.4.$$

That prediction must then be corrected for (1) the median-to-mean adjustment, and (2) the bias embedded in the original median prediction. The first is accomplished by multiplying y by $\exp(s^2/2)$, and the second by multiplying that result by $\exp(-s_y^2/2)$. Recall that (s_y^2) is formed by pre-multiplying the estimated variance-covariance matrix by the row vector of prediction values, and then post-multiplying that result by the prediction vector transposed to column form. The computations to obtain (s_y^2) are:

$$s_y^2 = (1 \quad 6.9078) \begin{pmatrix} 0.1855 & -0.0511 \\ -0.0511 & 0.0147 \end{pmatrix} \begin{pmatrix} 1 \\ 6.9078 \end{pmatrix} = 0.1810.$$

The final, corrected prediction is:

$$Y_{pred} = y\{\exp[(s^2/2) - (s_y^2/2)]\} = 13{,}058.4\exp(0.1260/2 - 0.1810/2) = 12{,}704.2$$

This result is 3 percent lower than the naive prediction, but that outcome is strictly a function of the hypothetical data being used. For instance, had the prediction value of X been 100 instead of 1000, $s^2/2$ would have been larger than $s_y^2/2$, and the corrected prediction would have exceeded the naive one. In general, however, the differences, whether plus or minus, are not likely to exceed a few percentage points.

Appendix B

The Design Matrix and Error-Free Cost Values

Cost	Weight	Speed	Production Rate	Time
33.190	18.400	545	94	55
65.669	25.298	530	48	63
35.272	15.497	595	157	66
53.929	19.856	423	100	75
83.448	27.424	1283	72	60
165.152	39.037	1170	48	72
128.644	25.800	1434	72	74
88.562	16.234	1265	105	78
125.830	23.050	990	63	80
29.606	21.052	521	100	50
30.820	20.638	775	180	53
79.079	28.494	950	48	57
36.146	19.350	680	108	55
34.807	12.760	1150	153	58
76.335	25.855	1195	65	58
91.605	24.861	1153	40	59
131.444	46.172	1262	83	67

References

1. Dale M. Heien. "A Note on Log-Linear Regression." *Journal of the American Statistical Association.* Vol. 63, 1034-1038, 1968

2. Arthur S. Goldberger. "The Interpretation and Estimation of Cobb-Douglas Functions." *Econometrica.* Vol. 35, 464-472, 1968

3. Dan Bradu and Yair Mundlak. "Estimation in Lognormal Linear Models." *Journal of the American Statistical Association.* Vol. 65, 198-211, 1970

4.. A. Book and P.H. Young. "Optimality Considerations Related to the USCM-6 `Ping Factor'." Presentation to the ICA/NES 1990 National Conference, June 1990

5. Philip H. Young. "GERM: Generalized-Error Regression Model." Presentation to the 25th Annual DOD Cost Symposium, September 1991

6. Arthur S. Goldberger. *Topics in Regression Analysis.* New York: The Macmillan Company, 1968

7. Robert S. Pindyck and Daniel L. Rubinfeld. *Econometric Models and Economic Forecasts.* 2nd ed. New York: McGraw-Hill Book Co., 1991

8. Normal R. Augustine. *Augustine's Laws.* New York: Viking Penguin Inc., 1986

9. CNA Professional Paper 524, *Aircraft Cost and Development Program Length: An Empirical Analysis*, by Henry L. Eskew, Dec 1992

10. Douglas Moses. "Learning Curve and Rate Adjustment Models: Comparative Predictive Accuracy Under Varying Conditions" in Roland Kankey and Jane Robbins. *Cost Analysis and Estimating, Shifting U.S. Priorities.* New York: Springer-Verlag New York Inc., 1991