Cutting the Gordian Knot: Maximum Likelihood Estimation for Regression of Log Normal Error

Christian Smart, Ph.D., CCEA
Director, Cost Estimating and Analysis
Missile Defense Agency
christian.smart@mda.mil

Abstract

Nonlinear equations are commonly used in cost estimating relationships (CERs) for government projects. Three primary methods commonly used in the cost estimating community for calculating nonlinear CERs from historical data are log-transformed ordinary least squares (LOLS), iteratively reweighted least squares/minimum unbiased percentage error (IRLS/MUPE), and minimum percent error with zero percent bias (MPE-ZPB). Of these three, LOLS is the oldest, since it is not computationally intensive. Indeed, the parameters can be calculated using a hand calculator.

LOLS involves performing ordinary least squares (OLS) on a log-transformation of a power equation. The use of transformations in LOLS has led to several criticisms, such as that it is not optimal, and there are issues with transforming the data. Newer methods, such as IRLS/MUPE and MPE-ZPB have been developed that do not involve the use of transformations. The criticism regarding the use of transformations has some merit. However the claim that LOLS is not optimal is not accurate. When the CER residuals are lognormally distributed, LOLS is a maximum likelihood estimate of the median.

In this paper we provide theoretical reasons for why the residuals of CERs should be lognormally distributed, as well as ample empirical evidence for actual CERs used in practice. In addition we provide information about the use of the lognormal to model risk in other industries, such as health care and property insurance.

To overcome the issue with transformations we introduce a new method for calculating CERs with lognormal residuals using untransformed maximum likelihood estimation. This method is also optimal for lognormal residuals. It calculates the mean of the distribution directly without need for a correction factor, thus avoiding the drawbacks inherent in LOLS.

Introduction

In 333 B.C., Alexander the Great, early in his conquest of the known world (at that time), reached the city of Gordium. When shown an intricate knot with its ends hidden, Alexander sliced through the knot with his sword. “Cutting the Gordian knot” is thus a direct and simple solution to a complicated problem. In this paper, we present a direct approach to developing unbiased, optimal estimates of the mean in the case that the errors are lognormally distributed. This avoids the drawbacks inherent in log transformed ordinary least squares (LOLS).

Nonlinear equations are commonly used in cost estimating relationships (CERs) for government projects. Three primary methods commonly used in the cost estimating community for calculating nonlinear CERs from historical data are LOLS, iteratively reweighted least squares/minimum unbiased percentage error (IRLS/MUPE), and minimum percent error with zero percent bias (MPE-ZPB). Of these three, LOLS is the oldest, since it is not computationally intensive. Indeed, the parameters can be calculated using a hand calculator.

The pedigree and simplicity of LOLS have led to the perception that this method is antiquated, and should be replaced by more modern, computationally intensive methods such as IRLS/MUPE or MPE-ZPB. In log-transforming the data we are estimating "log-dollars." The transformed estimate is unbiased in log-space, but is biased once we transform the equation back to unit space. LOLS is estimating the median of a lognormal, which is less than the mean, so LOLS is a biased estimator of the mean. The bias is low, so if we are trying to estimate the mean, LOLS will underestimate that value.

Dr. Steve Book along with others (Book and Young 1995, 1997; Book and Lao 1996; Book 2006), developed MPE-ZPB as an alternative to LOLS.

In this paper we provide theoretical reasons for why the residuals should be lognormally distributed, as well as ample empirical evidence for actual CERs used in practice. In addition we provide information about the use of the lognormal to model risk in other industries, such as health care and property insurance.

We begin by providing a discussion of CER residuals and methods for modeling them using maximum likelihood. Each of the methods in wide use today – LOLS, MPE-ZPB, and IRLS/MUPE – have a connection to maximum likelihood estimation. We show that for the case of lognormally distributed residuals, LOLS is an optimal method for estimating the median. We then provide a theoretical foundation for why we should expect CER residuals to be lognormally distributed. We further provide empirical evidence in support of the lognormal, in the form of analysis of residuals for CERs from the NASA/Air Force Cost Model (NAFCOM). We also provide a comparison of modeling methods used in other industries, including insurance.

Since there is strong evidence that CER residuals are lognormally distributed, and there are issues with LOLS, we use the maximum likelihood method on the non-transformed equation to circumvent these issues. We provide three examples comparing the various methods.

Model Development and Maximum Likelihood Estimation

Our goal is to use historical data to predict the cost of future programs and projects. It is important when developing models to limit our choices, since given enough models to choose from there will be at least one model that appears to fit the data well, but will not help us effectively predict future cost. For example given ***n*** data points, we can perfectly predict the past by fitting ***n-1*** parameters. However, doing so will capture many idiosyncrasies in the historical data that are not likely to be repeated in the future, a

phenomenon referred to as over fitting. Experience is a useful guide in limiting the universe of choices.

There are many models from which to choose. The analyst has many decisions to make in selecting a model, such as whether to use an explicit equation form, or, for example, a decision tree. An example of a decision tree might be something like "if the total dry weight of a spacecraft is more than 1,000 lbs., my estimate is $100 million, otherwise it is $25 million." Or the analyst might select a model that estimates the costs based on user inputs, such as the NASA/Air Force Cost Model (NAFCOM). Using a mathematical equation to estimate the cost using one or more predictive cost drivers, such as weight, is a traditional approach. Also, if one selects to use an equation to estimate cost, the analyst must also decide what type of equation to use, whether linear or nonlinear.

Through data and experience, it is largely agreed in the NASA and DoD cost community that costs do not typically follow a linear pattern. Rather they tend to vary nonlinearly in relation to the cost drivers that are typically selected. In particular, the power law equation, or some variation of it, has been widely adopted. The power law equation has the form

$$Y = aX^b$$

In this case Y typically represents cost in $, but can also represent effort (hours, full-time equivalents). The variable X typically represents weight or some other performance parameter. The equation can also be modified to accommodate multiple cost drivers. The value of the b parameter in the power equation is usually less than 1, indicating economies of scale in design and production. This model has been found to do a good job explaining the relationship between cost and cost drivers, including weight and other performance characteristics, for a wide variety of spacecraft and other programs. For example, if the equation has the form

$$Estimated\ Cost = 1.5 \cdot Weight^{0.5}$$

then as weight doubles, cost is increased by a factor equal to the square root of weight, rather than a simple linear relationship. Such equations are so commonly used they are referred to as "cost estimating relationships," or CERs.

If an equation form is selected, the form of the residuals between predicted and actual costs must also be chosen. For example, given an equation of the form

$$Y = a + bX$$

and a set of data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

the residuals are defined as:

$$\varepsilon_i = Y_i - (a + bX_i) = Actual - Estimated$$

This is also referred to as the "error" term since it is the difference between the actual cost and the estimated cost. Residuals or "errors" are an important consideration in modeling since they often drive the methods used for parameter calculation. For example, linear regression finds the "best fit" by finding the parameters ***a*** and ***b*** that minimize the sum of the squares of the residuals

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left(Y_i - (a + bX_i)\right)^2 = \sum_{i=1}^{n} \left(Actual_i - Estimated_i\right)^2 \quad .$$

This method was first developed by the mathematicians Legendre and Gauss in the early 19th century, who used it to predict the orbits of heavenly bodies using observed data. Francis Galton later applied this technique to find linear predictive relationships between various phenomena, such as the relationship between the heights of fathers and sons. Galton found a positive correlation between these heights but found a tendency to return or "regress" toward the average height, hence the term "regression analysis"

The residuals of the power equation can either be additive or multiplicative.

Additive residuals have the form

$$Y = aX^b + \varepsilon$$

while multiplicative residuals have the form

$$Y = aX^b \varepsilon.$$

Multiplicative residuals are more appropriate for the spacecraft and defense industry in most applications because of wide variations in size, scope, and scale of the systems that are estimated. For example, if historical data ranges from $50 million to $1 billion, it is better to analyze percentage differences, since this provides a more meaningful comparison of accuracy than absolute dollar values. As a result we are primarily interested in the percentage difference between actual and estimated costs, not the absolute difference. See Figure 1 for a graphical comparison of multiplicative and additive errors. The commonly-used regression techniques considered in this paper are all based on the multiplicative error assumption.

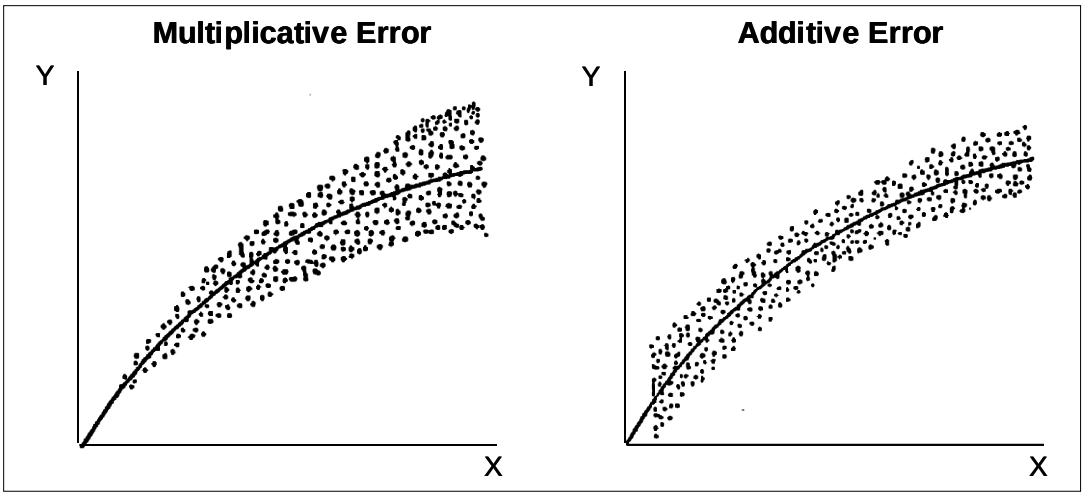

Figure 1. Multiplicative Vs. Additive Errors (Eskew and Lawler 1994).

In terms of what works in practice, given that the error is bounded below since neither the estimate nor the actual cost can be less than zero, but there is an infinite amount of room on the upside, we should expect positive skew in our error distributions. The lognormal distribution is a natural choice for modeling positive skew. Don Mackenzie (Mackenzie et al. 2008) has given empirical evidence in support of the lognormal for modeling the residuals. This is in accordance with what I have found with respect to the NASA/Air Force Cost Model, for which the lognormal distribution has been shown fits the CER residuals for almost all subsystems which have a sufficient number of data points to provide a meaningful test. An example of a CER developed for NAFCOM will be discussed in the next section. The gamma distribution also has positive skew, which makes it an appealing choice as well.

For the power equation with multiplicative residuals, i.e.,

$$Y = aX^b\varepsilon$$

the estimates vary based on the variation of the residual

$$\varepsilon = \frac{Y}{aX^b}$$

It's also common to adjust this to treat ε as a percentage, i.e., set

$$Y = aX^b(1+\varepsilon)$$

$$\varepsilon = \frac{aX^b - Y}{aX^b} = \frac{Estimate - Actual}{Estimate}$$

In other words the actual cost is equal to the estimate plus or minus a percentage of the estimate. If the estimate is greater than the actual cost the residual is greater than zero. If the estimate is less than the actual the residual is less than zero. Note the lack of

symmetry. For estimates above the actual, the maximum value of the residual is ***1***, and for estimates below the actual, the minimum value has no bound!
See Figure 2 for an illustration of multiplicative residuals for a subsystem CER in the NASA/Air Force Cost Model.

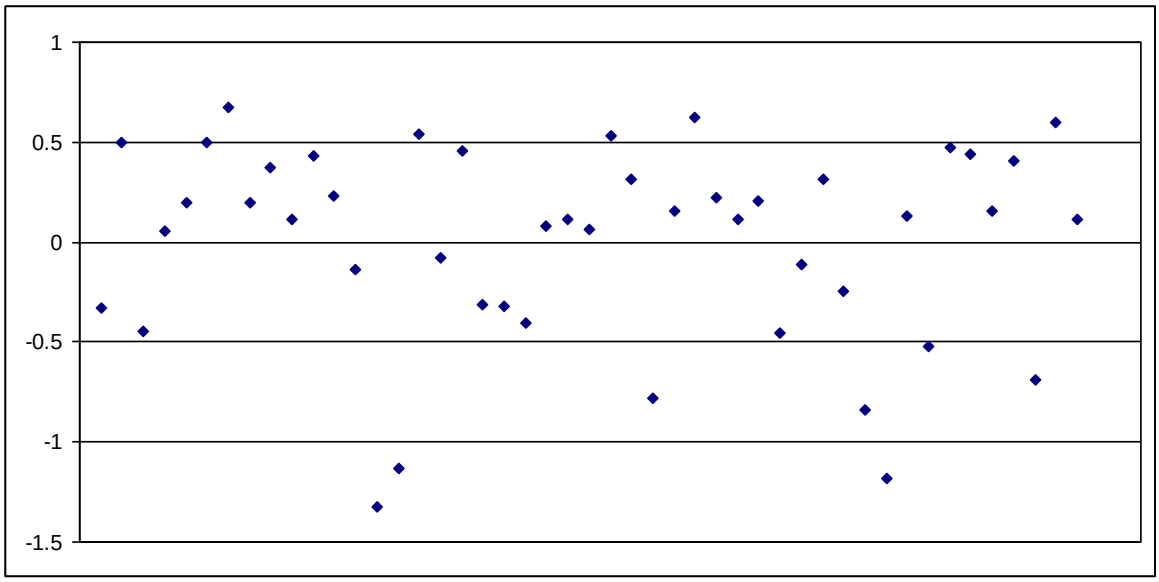

Figure 2. Multiplicative Residuals for a Subsystem CER in the NASA/Air Force Cost Model (NAFCOM).

For a "good" model, the cost drivers explain all (or most) of the variation in the historical data that can be explained. Therefore it is typically assumed that any remaining variation is random, either due to non-repeatable random phenomena (e.g., test failures) that are truly random phenomena and will not help predict future cost. The multiplicative residuals that represent this unexplained variation are thus treated as random variables. Thus after choosing the equation, and choosing the type of residual, we have to make yet a third choice, namely, the type of distribution that this unexplained variation follows. For CER development, residuals are typically assumed to follow normal, lognormal, gamma, or they are treated without making such an assumption (non-parametric).

The normal distribution is the most common probability distribution. Many random phenomena follow this distribution. It is also known as the "bell curve," due to its symmetry and small tails. If cost is a sum of many random independent phenomena, the central limit theorem indicates this may be the appropriate distribution. See Figure 3 for a depiction of a normal distribution.

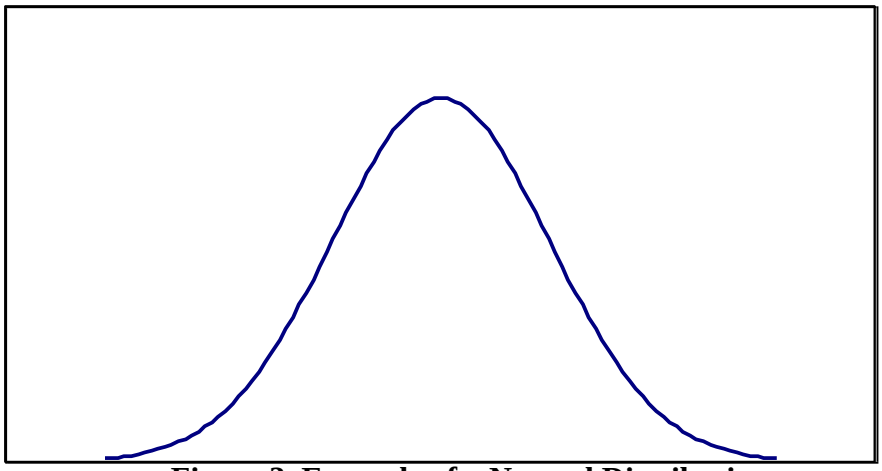

Figure 3. Example of a Normal Distribution.

The lognormal distribution is a skewed distribution. If ***X*** is lognormally distributed, ***y = ln(x)*** is normally distributed. The lognormal has a heavier right tail than the normal distribution, is bounded below by zero, and unbounded above. If cost is a function of multiplicative factors, for example, test failures cause a percentage increase in cost rather than a fixed amount increase, project costs are likely to be lognormally distributed. This is a multiplicative analogue to the central limit theorem. These aspects make the lognormal appealing for cost modeling. See Figure 4 for a graphical depiction of a lognormal distribution.

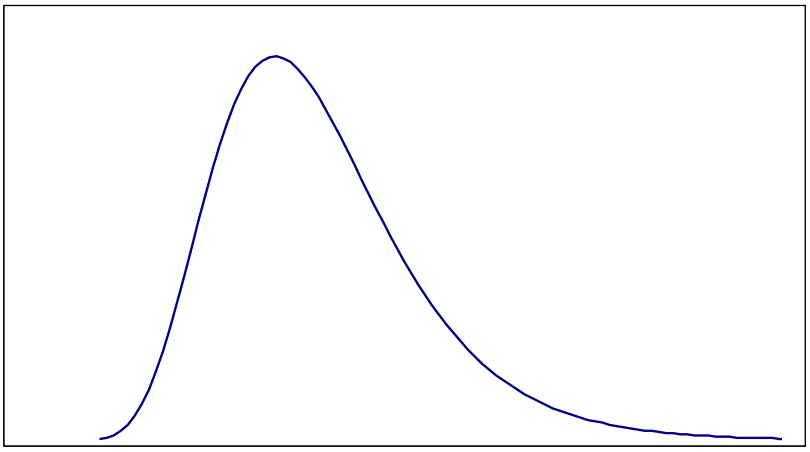

Figure 4. Example of a Lognormal Distribution.

The gamma distribution is a flexible distribution. It can to some extent resemble a lognormal, and can also resemble an exponential distribution. Indeed the gamma distribution is the sum of independent exponential distributions, so the exponential is a special case of a gamma distribution. See Figure 5 for representations of gamma distributions.

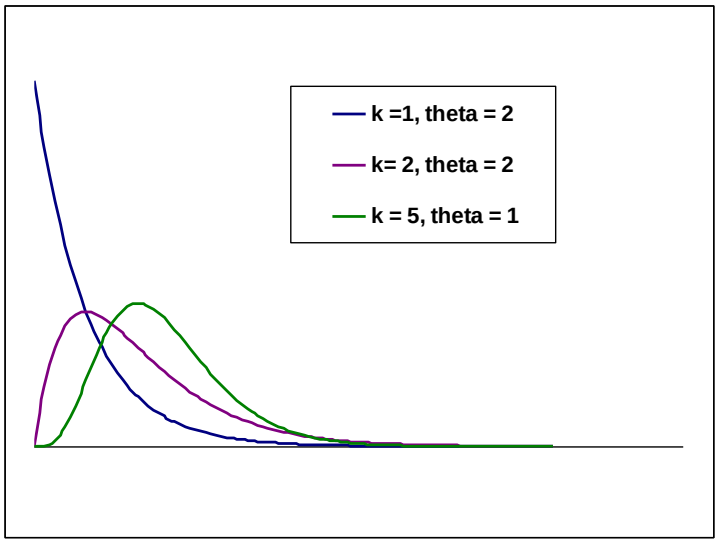

Figure 5. Examples of a Gamma Distributions.

The decision to choose or not choose an underlying distribution for the residuals represents another choice. When the data follow an observable pattern, based either on preliminary data analysis or through experience, parametric analysis is preferred. For example, NASA cost data are skewed, which makes intuitive sense, because cost cannot be less than zero, but there is no upper limit. This often leads to the assumption of lognormal or gamma residuals. However when the data do not follow an observable pattern, or there is no reason to assume an underlying pattern in the data, non-parametric analysis may be suitable. This may be the case if data sets are small, or if there is no reason to assume similarity with other available data.

However, if non-parametric techniques are used, the analyst must be careful to ensure models are valid, since techniques may assume an inherent pattern in the data or be similar enough to a parametric technique that the non-parametric version inherits some features of the parametric version.

Another issue with non-parametric techniques is the lack of rich techniques for developing confidence intervals, prediction intervals, covariance matrices, and other useful metrics and methods available for parametric models. Indeed, some statistical techniques do not exist for nonparametric problems. As shown by Bahadur and Savage in their 1956 paper “The Nonexistence of Certain Statistical Procedures in Nonparametric Problems” (Bahadur and Savage 1956), in such cases there is no effective hypothesis test for the population mean, no effective confidence interval for the population mean, and no effective point estimate for the population mean. They also showed that no confidence interval will fit the data well. This makes model validation problematic for non-parametric methods. However, note that parametric techniques do not necessarily involve assuming the residuals follow a particular probability distribution. This assumption can be much weaker, such as assuming finite variance.

There are numerous ways to calculate the parameters of a cost-estimating relationship. One powerful statistical technique commonly used for parameter calculation is the method of maximum likelihood, which is sometimes referred to as maximum likelihood estimation (MLE). MLE is a widely used technique that serves as a unifying framework

for the CER methods we shall discuss in this paper, and is the basis for the new method that we present in this paper. Note that MLE is a frequentist approach that requires large data sets. Some researchers recommend using a least 50 data points when developing CERs using classical methods (Babyak 2004). For smaller data sets we need Bayesian methods. This paper only deals with MLE, for smaller data sets you need to consider Bayesian applications for CER development (Smart 2014).

Let $a_1,\ldots, a_n$ represent the observed data and $x_1,\ldots,x_n$ represent random variables where a_i results from observing the random variable x_i. The likelihood function, which represents the likelihood of obtaining the sample results, is defined as

$$L(\theta)=\prod_{i=1}^{n} Pr(X_i = A_i \mid \theta)$$

The maximum likelihood estimate of θ is the vector that maximizes the likelihood function. This technique is appealing because maximizing the likelihood of finding the true underlying parameters of this distribution is exactly what we hope to accomplish in developing a CER. One major advantage of this technique is that the likelihood function is almost always available. Maximum likelihood uses all the available data, unlike other methods, such as percentile matching and method of moments. Maximum likelihood methods have good statistical properties, like consistency and efficiency. A rich body of statistical theory has been developed for maximum likelihood estimation.

An estimator $\hat{\theta}$ is a uniformly minimum variance unbiased estimator (UMVUE) if it is unbiased and for any true value of θ there is no other unbiased estimator that has a smaller variance. An estimator that is UMVUE is efficient, in that it achieves its theoretical lower bound. This means that the estimated coefficient will likely be closer to the true coefficient than that calculated with another estimator. A (finite) data set is often considered as a random sample from an underlying population. The variance of the coefficients is a decreasing function of the sample size, so for small samples, the variance can be quite large relative to the coefficient. In these cases, the variance of the coefficient, if large, can mean than the estimated coefficient is far away from the true coefficient. Smaller variances mean a quicker convergence to the true underlying population coefficient as the sample size increases, as long as the estimator is consistent. Consider for example two consistent estimators for a coefficient, and suppose that one has variance equal to 50% of the size of the estimate, and the other has variance equal to 100% of the size of the estimate. Suppose for the sake of simplicity that the estimator's mean is equal to the true population coefficient. The coefficient for a single data set can be viewed as one random sample from a Monte Carlo simulation of the coefficient distribution. A single draw drawn for each of these is likelier to be closer to the true coefficient for the distribution with smaller variance. For example, a single Monte Carlo draw for the distribution with smaller variance is 0.89 while the same random draw for the distribution with higher variance results in 1.58. The first estimator is thus 11% below the true coefficient, while the second is 58% greater. Thus small variance is a highly desirable property. Since maximum likelihood estimates are UMVUE, it is desirable to use them whenever their use can be justified. In general parametric models can be seen as

more accurate predictors whenever the hypotheses required for their use can be supported.

Three popular CER methods are log-transformed OLS, MUPE, and MPE-ZPB. All have a connection to maximum likelihood estimation, in the sense that parameter calculation for each of the methods considered can be viewed in the context of maximum likelihood. Maximum likelihood is used together with an assumption about the underlying residuals to calculate the parameters of the CER. Each of the CER techniques we consider has a strong connection to maximum likelihood estimation paired with either the lognormal, normal, or gamma distribution.

The probability density function of a lognormal is defined as

$$p(y,\mu,\theta)=\frac{1}{y\sqrt{2\pi\theta}}e^{-\frac{(\ln y-\mu)^2}{2\theta}}$$

The parameters μ and θ are the mean and variance (respectively) of the transformed lognormal, that is, the associated normal mean and variance.

In log-transformed ordinary least squares, logarithmic transformation is applied to both sides of the power equation $Y=aX^b$, which transforms the equation from a nonlinear equation to a linear one, i.e.,

$$\ln Y=\ln\left(aX^b\right)=\ln a+b\ln X.$$

Ordinary least squares is then applied to the transformed data. The parameters a and b are chosen so that $\hat{\mu}=\ln a+b\ln X$. To get the costs in unit space we exponentiate, to obtain e^{μ}. However, this is not an estimate of the unit space mean. Rather it is the median.

The three most commonly encountered measures of centrality are the mean, median, and mode. The mean is the "expected value," so for a sample of n data points this is

$$\sum_{i=1}^{n}\frac{x_i}{n}$$

The median is the 50th percentile, the point at which half the population is less than this value, and half is greater. The mode is the "most likely" point of the density function, that is, the peak of the distribution. For a normal distribution, the mean, median, and mode are all equal. For a lognormal, the mode is always less than the median, and the median is always less than the mean. Thus for a lognormal, the mean is always greater than the 50th percentile, and can be any percentile greater than the 50th, such as the 90th or 95th percentile. For this reason, the median is a good measure of centrality. That is why it is common to report the median rather than the mean as the "average" of skewed data. Whenever average income or average house price data are reported in the media for example, the average reported is always the median, and for exactly this reason. Some

analysts prefer the median over the mean (Foussier 2008). See Figure 6 for a graphical comparison of the mode, median, and mean of a lognormal distribution.

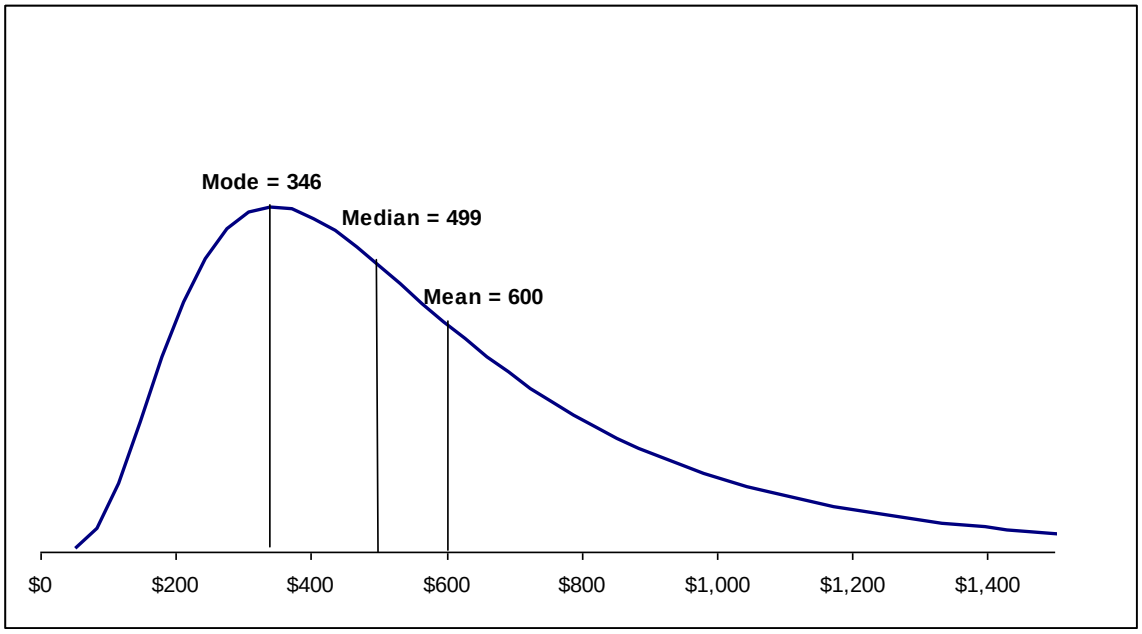

Figure 6. Comparing the Mode, Median, and Mean for a Lognormal Distribution.

For $y_i=f(x_i,b)\cdot\varepsilon_i$, let β = vector of coefficients of the CER, y_i = actual cost of the ith data point, x_i = vector of cost drivers for the ith data point, and ε_i = residual of the ith data point. The conditional probability density function of Y for each x is $p\left(y|X=x;\beta_0,\beta_1,\ldots,\beta_p,\sigma^2\right)$. Given any data set the probability density of seeing that data is

$$\prod_{i=1}^{n} p\left(y_i \mid x_i;\beta_0,\beta_1,\ldots,\beta_p,\theta\right)=\frac{1}{y\sqrt{2\pi\theta}}e^{-\frac{\left(\ln y-\ln\beta_0-\beta_1\ln X_{i1}-\ldots-\beta_p\ln X_{ip}\right)^2}{2\theta}}$$

Since the logarithm function is monotonically increasing, we can take the log of the likelihood function and maximize that instead, that is,

$$l\left(\beta,\theta\right)=-\frac{n}{2}\ln\theta-\frac{1}{2\theta}\sum_{i=1}^{n}\left(\ln y_i-\ln\beta_0-\beta_1\ln X_{i1}-\ldots-\beta_p\ln X_{ip}\right)^2$$

We ignore constants since they do not affect the maximization. This is the same as minimizing the negative of the log-likelihood function, that is,

$$l\left(\beta,\theta\right)=\frac{n}{2}\ln\theta+\frac{1}{2\theta}\sum_{i=1}^{n}\left(\ln y_i-\ln\beta_0-\beta_1\ln X_{i1}-\ldots-\beta_p\ln X_{ip}\right)^2$$

In order to minimize the likelihood function, we first minimize with respect to θ.

To minimize, we take the partial derivative with respect to θ set equal to zero, and solve for θ. Taking the derivative yields

$$\frac{\partial l}{\partial \theta} = \frac{n}{2\theta} - \frac{1}{2\theta^2}\sum_{i=1}^{n}\left(\ln y_i - n\beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2$$

Setting this equal to zero and solving gives

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\left(\ln y_i - n\beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2}{n}$$

Plugging in the value for θ into the log likelihood function yields

$$l^*(\beta) = \frac{n}{2}\ln\frac{\sum_{i=1}^{n}\left(\ln y_i - n\beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2}{n} + \sum_{i=1}^{n}\ln y_i + \frac{n}{2}$$

Ignoring constants this simplifies to

$$l^*(\beta) = \ln\sum_{i=1}^{n}\left(\ln y_i - n\beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2$$

which is equivalent to minimizing

$$l^*(\beta) = \sum_{i=1}^{n}\left(\ln y_i - \ln\beta_0 - \beta_1 \ln X_{i1} - \ldots - \beta_p \ln X_{ip}\right)^2$$

This is the least squares of the log of the differences between the actual and the estimated costs. Notice the similarity to linear regression. We can replace $\ln\beta_0 + \beta_1 \ln X_{i1} + \ldots + \beta_p \ln X_{ip}$ with any arbitrary function $f(x_i, \beta)$ and get the same result, i.e.,

$$l^*(\beta) = \sum_{i=1}^{n}\left(\ln y_i - f(x_i, \beta)\right)^2$$

Thus we have derived is a generalization of log-transformed ordinary least squares in the context of maximum likelihood.

The parameters can be easily calculated in a spreadsheet. However, note that the maximum likelihood median estimator is more general. In the past this method has merely been viewed as a simplistic way of converting a nonlinear power equation to linear space and applying ordinary least squares to the resulting linear equation (Book and Lao 1996). We have proven instead that any equation form may be used as there is nothing in the derivation that forces a particular equation type to be used. One simply minimizes the sum of the log-squared differences between the actual and the estimated costs. Thus equation forms such as $y=a+bx^c$ can be calculated with this generalized

method, which we term Generalized Maximum likelihood estimation of Lognormal Error, or GMLE (pronounced "Gimli" - apologies to JRR Tolkien).

Although for skewed data, the median is a better representative of a distribution's centrality, the mean is the focus of most statistical estimators. The other two methods we will present estimate the mean, rather than the median of the error distribution, and one criticism often levied on log-transformed OLS is that it is biased low, since the median of a lognormal distribution is always less than its mean. However, this can be corrected, since there is a mathematical relationship between the median and the mean. The mean of a lognormal distribution is $exp(\mu + \sigma^2/2)$ and the median is simply $exp(\mu)$, so the mean is the quantity $exp(\sigma^2/2)$ multiplied the estimate. The only complicating factor is that the population variance is not known with certainty and so it must be estimated using statistical samples. Several methods for estimating this factor have been proposed. A simple one, termed the "Ping" factor (Hu 2005) is

$$exp\left(\left(1-\frac{p}{n}\right)\frac{s^2}{2}\right)$$

where p is the number of parameters, n is the number of data points in the sample, and s^2 is the sample variance.

For the equation $y_i = f(x_i,b)\cdot u_i$, when the residuals are normally distributed, with mean 1 and variance θ, the likelihood function, as demonstrated by Lee (Lee 1997), is

$$L(\beta,\theta) = \frac{exp\left(\frac{-1}{2\theta}\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2\right)}{(2\pi\theta)^{\frac{n}{2}}\prod_{i=1}^{n} f(x_i,\beta)}.$$

The log-likelihood function is thus

$$l(\beta,\theta) = \frac{-1}{2\theta}\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2 - \frac{n}{2}ln(2\pi) - \frac{n}{2}ln\,\theta - \sum_{i=1}^{n} ln\, f(x_i,\beta)$$

Maximizing this expression for θ and then substituting back into $l(\beta,\theta)$ yields the concentrated log-likelihood function

$$l^*(\beta) = -\frac{n}{2}ln\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2 - \sum_{i=1}^{n} ln\, f(x_i,\beta)$$

This is the same as minimizing

$$l^*(\beta) = \frac{n}{2}ln\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2 + \sum_{i=1}^{n} ln\, f(x_i,\beta)$$

As noted in Goldberg and Tuow (2003), this method is very similar to the minimum percent error method developed by Book and Young (1995, 1997), who ignore the final term and instead minimize the sum of squared percentage errors.

The minimum percent error method minimizes

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right)^2$$

Thus the minimum percent error method is a pseudo-likelihood estimator in the case of normally distributed residuals since it is equivalent to minimizing the first term in the concentrated likelihood function. Note that the minimum percent error (MPE) method is biased. Instead of being biased below the mean like with log-transformed OLS, the MPE method is biased high, since one way to make the error term small is to make the estimates large. To correct for this Book and Lao (Book and Lao 1996) introduced a bias constraint. The objective function is the same, but now sample bias is constrained to be zero, that is

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta)}{f(x_i,\beta)}\right) = 0.$$

This method is referred to as MPE-ZPB or ZMPE("Zimpy"). While ostensibly a non-parametric method, it is similar to the normal MLE.

When the residuals follow a gamma distribution, the negative log-likelihood function is

$$l(\beta) = \sum_{i=1}^{n}\left(\frac{y}{f(x_i,\beta)} + \ln f(x_i,\beta)\right).$$

This can be minimized by iteratively minimizing the sum of percent squared errors until the estimates converge, that is

$$\sum_{i=1}^{n}\left(\frac{y_i - f(x_i,\beta_k)}{f(x_i,\beta_{k-1})}\right)^2$$

where k denotes the iteration number. This method was first developed by Nelder (1968) and Wedderburn (1974), who called the method iteratively re-weighted least squares (IRLS). It was re-discovered by Hu in the 1990s, who called it minimum unbiased percentage error (MUPE) (Hu 2005, 2013).

In the case of gamma residuals, IRLS/MUPE is a maximum likelihood estimate. However IRLS/MUPE does not depend upon the assumption of gamma residuals.

The likelihood method was generalized by Wedderburn to consider quasi-likelihood, which has good statistical properties, but only requires a finite variance.

Thus log-transformed OLS, MPE-ZPB, and IRLS/MUPE all share a common connection in maximum likelihood estimation. Log-transformed OLS/GLMLE is a maximum likelihood estimator of the median of lognormally distributed multiplicative residuals. Thus it is a parametric method. MPE-ZPB is a pseudo-likelihood estimator of the mean of normally distributed multiplicative residuals with a bias constraint added. It is not directly parametric but it has parametric properties because in my experience it is typically a good approximation of the normal MLE solution. IRLS/MUPE is a maximum likelihood estimator of the mean of gamma distributed residuals. But it is also more general, since it is a quasi-likelihood parametric method.

Theoretical Foundation for the Use of the Lognormal

We have shown that LOLS is a maximum likelihood estimate when the residuals are lognormally distributed. There is evidence in favor of this hypothesis. In this section we provide a theoretical argument, and then we provide evidence for the use of the lognormal in other applications.

The theoretical argument is that changes in costs over time are proportional to prior costs. This makes sense. Cost is more likely to increase than decrease over time, as evidenced by numerous studies on cost growth that show that over 80% of government projects experience cost growth, and on average increase by over 50% (Smart 2015). So when we talk about cost changes, we almost always mean cost increases. Cost increases often do not result in funding increases in the short term due to funding constraints. Thus cost increases will result in longer schedules. Longer schedules imply a longer period in which the personnel devoted to a project will charge to that particular project. Larger projects have more personnel assigned to a project, meaning that increases in cost will result in a proportional increase in cost.

Mathematically the change in cost from time ***t–1*** to time ***t*** can be represented as

$$\boldsymbol{X_t - X_{t-1} = \epsilon_t X_{t-1}}$$

where the $\boldsymbol{\epsilon_t's}$ are mutually independent and independent of X_{t-1}. Rearranging, we have that

$$\boldsymbol{\frac{X_t - X_{t-1}}{X_{t-1}} = \epsilon_t.}$$

Summing over ***t*** we find that

$$\boldsymbol{\sum_{t=1}^{n} \frac{X_t - X_{t-1}}{X_{t-1}} = \sum_{t=1}^{n} \epsilon_t.}$$

Proportional changes can be approximated as

$$\boldsymbol{\sum_{t=1}^{n} \frac{X_t - X_{t-1}}{X_{t-1}} \approx \int_{X_0}^{X_n} \frac{dX}{X} = ln(X_n) - ln(X_0)}$$

Thus

$$\boldsymbol{ln(X_n) - ln(X_0) \approx \sum_{t=1}^{n} \epsilon_t}$$

Rearranging terms we find that

$$\boldsymbol{ln(X_n) \approx ln(X_0) + \sum_{t=1}^{n} \epsilon_t}$$

According to the Central Limit Theorem the sum of many random variables is normally distributed. Thus for large values of ***n***, $\boldsymbol{ln(X_n)}$ is normally distributed. Thus by definition $\boldsymbol{X_n}$ is lognormally distributed.

The Use of the Lognormal in Other Industries

The lognormal has been widely used in cost analysis for decades. It is also widely used in other industries. The book *Statistical Rules of Thumb* recommends the use of the lognormal in environmental studies (van Belle 2008).

The analogy with cost estimating in insurance is "loss modeling." In insurance parlance, a "loss" is the amount of a loss experienced by a policyholder. Parametric models are used to estimate both loss size and claim frequency. As cited in *Modelling Extremal Events* (Embrechts et al., 2003), Seal noted that "Types of distributions of independent claim sizes are…limited, for apart from the Pareto and lognormal distributions, we are not aware that any has been fitted successfully to actual claim sizes in actuarial history." (Seal 1983). Embrechts and his co-authors note that this is an extreme statement, but also state that many years later this point still stands, and cite other studies from the 1990s that make similar statements. More recently, the lognormal has also been used in ratemaking and reserve setting, a process not unlike cost risk analysis. Fu and Moncher's 2004 presentation (Fu and Moncher 2004) reports that the gamma and lognormal are the most widely used distributions in loss modeling. They mention 31 recent papers that reported the use of lognormal distributions and 37 that reported the use of gamma distributions for residual modeling. Fu and Moncher also study the normal distribution, but find the lognormal and gamma much better for modeling skewed, positive data, like "loss" and "cost." They recommend against use of normal distribution for modeling skewed data because the normal distribution is symmetric. Ismail and Jemain (2009) also report the widespread use of the lognormal and gamma distributions in loss modeling as of 2009.

The lognormal has been widely used in loss distribution modeling in studies from the 1960s through the 2000s, including fire losses, auto losses, hurricane losses, and property insurance losses (Kleiber and Klotz 2003).

Costs are modeled parametrically in health care and labor economics as well (Manning and Mullahy 2001; Basu et al., 2004; Manning et al., 2005; Gallin 2004). Both log-transformed OLS and IRLS/MUPE are widely used. No mention was found of a normal MLE or MPE-ZPB type method.

Empirical Evidence for the Lognormal

If a maximum likelihood method has been used, we need to check to see if residuals fit the assumed shape. Fit is used here in the negative sense. We can never truly prove anything statistically. We can use data to disprove conjectures but the best we can hope for when we make hypotheses and test them, these tests will fail to disprove or reject our hypotheses. As the philosopher of science Karl Popper once remarked "Our knowledge can only be finite, while our ignorance must necessarily be infinite." Three commonly used tests for the goodness of fit of a distribution are chi-square, Kolmogorov-Smirnov (K-S), and Anderson-Darling (A-D). Chi-square and Kolmogorov-Smirnov are both simple and easy to compute. Anderson-Darling is more powerful and considered a good test for departure from normality. Anderson-Darling gives more weight to the tails of the distribution, while chi-square gives more weight to low probability intervals.

The K-S test statistic D is the maximum difference, in absolute value, between the empirical and fitted distributions,

$$\boldsymbol{max\,|F_n(x) - F^*(x;\theta)|}$$

where $F_n(x)$ represents the empirical distribution, and $F^*(x;\theta)$ is the fitted distribution with parameter θ. The maximum is evaluated for all sample data points.

The Anderson-Darling test statistic integrates the difference between the empirical distribution and the fitted distribution function over the entire range. It weighs the difference by the reciprocal of the variance. The formula for the Anderson-Darling statistic is

$$\boldsymbol{A^2 = n\int_0^\infty \frac{\left(F_n(x) - F^*(x)\right)^2}{F^*(x)\left(1 - F^*(x)\right)} f^*(x)dx}$$

We focus our attention on assessing the fit of the residuals to the A-D and K-S tests. The NASA/Air Force Cost Model (NAFCOM) contains numerous subsystem-level and component-level CERs for spacecraft. These CERs are all LOLS CERs. We assess the fit of 32 nonrecurring and recurring CERs that have sufficient data points to fit a distribution to the residuals. The subsystems and components assessed are displayed in Tables 1 and 2.

Subsystem	Number of Data Points
Attitude Control	72
Communications, Command and Data Handling	77
Electric Power	78
Reaction Control Subsystem	62
Solid Rocket Motor/Apogee Kick Motor	26
Structures and Mechanisms	121
Thermal Control	114

Table 1. Subsystem NAFCOM CERs Fit to Residual Distributions.

Component	Number of Data Points
Amplifier	19
Antenna	21
Battery	25
Command and Data Handling	31
Computer	17
Power Distribution	34
Tank	21
Transmitter	22
Transponder	18

Table 2. Component-Level NAFCOM CERs Fit to Residual Distributions.

For the 16 subsystems and components in the table, there are CERs for both nonrecurring and recurring costs. Thus the residuals for 32 CERs were fit to continuous distribution. The Crystal Ball add-in for MS Excel was used to perform these fits.

The critical value for the Anderson-Darling varies by distribution. For the normal and lognormal distributions the 10% critical value is 0.752 and the 5% critical value is 0.631. The K-S critical values depend on the sample size. The 10% critical value is $\frac{1.22}{\sqrt{n}}$ and the 5% critical value is $\frac{1.36}{\sqrt{n}}$. The critical vale is the probability of rejecting a true hypothesis. Statistics textbooks typically cite 5% as the standard critical value. (van Belle 2003).

Crystal Ball includes a wide variety of continuous and discrete distributions, including the beta, lognormal, normal, gamma, and many others, including the maximum extreme (i.e., Gumbel) distribution.

For the 32 CERs, using the Anderson-Darling test, the lognormal is not rejected at the 10% critical value for 30 out of 32 of the CERs, and is not rejected at the 5% critical value for 31 out of 32 of the CERs. Using the Kolmogorov-Smirnov test, the lognormal is not rejected at the 10% critical value for any of the 32 CERs. Using the chi-square test, the lognormal is not rejected at the 10% critical value for 30 of 32 CERs, and is not rejected at the 5% critical value for 31 of 32 CERs.

The two CERs that are rejected for a lognormal fit using the 10% criteria are also rejected for all the other continuous distributions included in Crystal Ball. These two CERs are the design and development CER for communications, command, and data handling, and the theoretical flight unit CER for batteries. In order to compare the fit for these two with another case for which the fit is very good, we turn to the important visual comparison, sometimes referred to as the "eyeball test."

The lognormal is a good fit for the design and development CER for the attitude control subsystem. It has the lowest A-D test statistic and the lowest K-S test statistic among all of the continuous distributions included in Crystal Ball. The maximum extreme (Gumbel) distribution and the gamma distribution provide good fits as well. See Figure 7 for a graphical comparison of the fits with the empirical distribution.

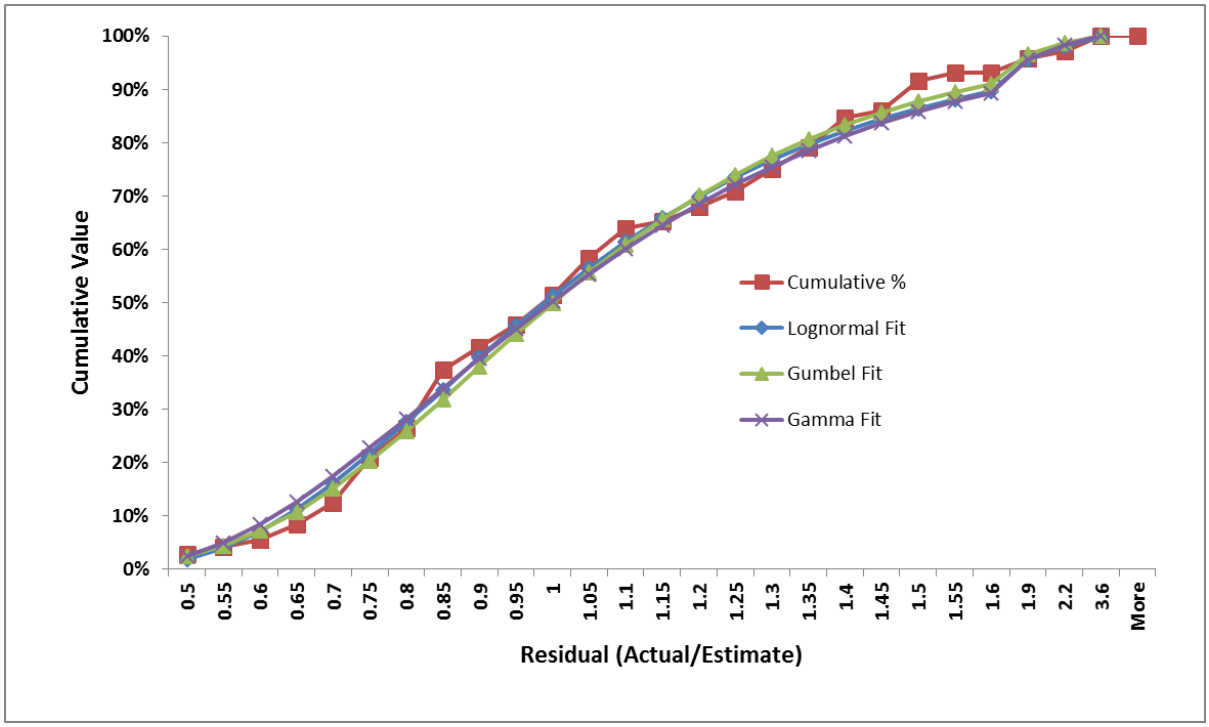

Figure 7. Graphical Comparison of Residuals for Attitude Determination and Control Subsystem with Three Fitted Distributions.

To see a comparison of the lognormal fit versus the empirical distribution for the communications, command, and data handling nonrecurring CER, see Figure 8. The lognormal distribution has the lowest test statistic for both the A-D and K-S tests of

goodness of fit. Note that this fit is rejected at the 10% critical value, but not rejected at the 5 % critical value.

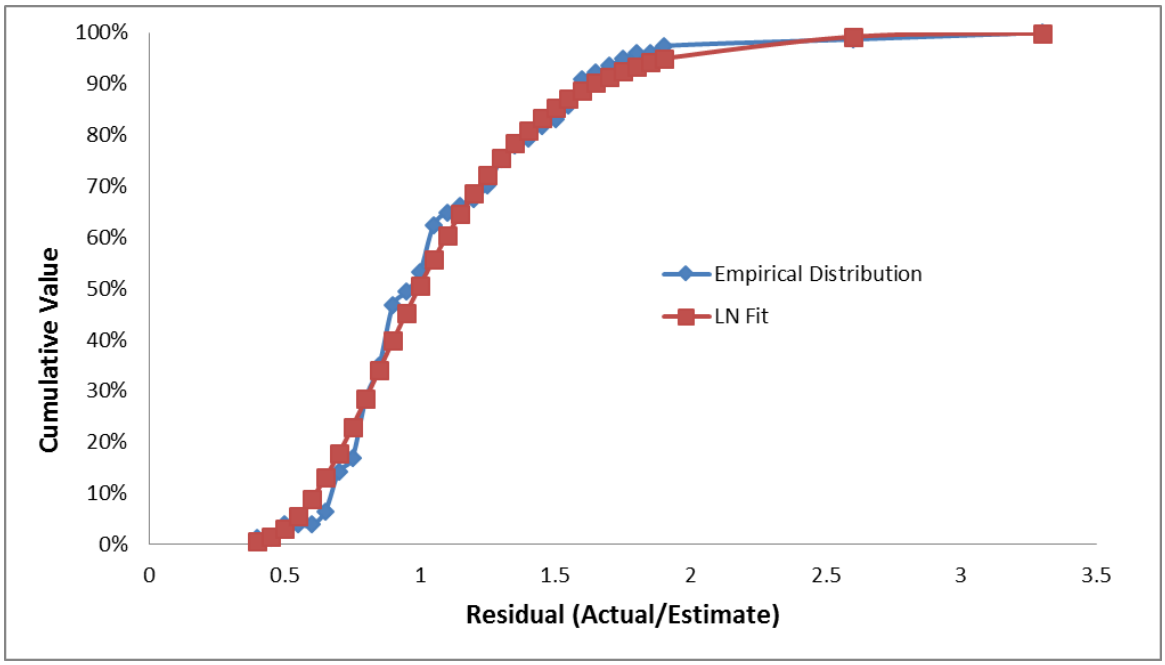

Figure 8. Graphical Comparison of Residuals for Communications, Command and Data Handling Subsystem.

From the eyeball test this still appears to be a decent fit, confirming the rule of thumb that the 5% critical value is the primary discriminator.

The only CER that does not fit any distribution at the 5% critical value is the theoretical flight unit CER for batteries. For this component, the lognormal has the lowest test statistic value for the A-D and K-S tests. We can clearly see from the eyeball test that the lognormal is not a good fit (and thus the other distributions will be even worse, since they have worse test statistics). This is largely due to one outlier. Without this outlier, the CER residuals fit a lognormal distribution well. For this outlier the actual cost is approximately six times as much as the estimate. This merits additional investigation into the outlier to determine if there is a reason why this is the case, or if there is an error in the data. See Figure 9 for a graphical comparison of the empirical residuals to the lognormal fit for batteries.

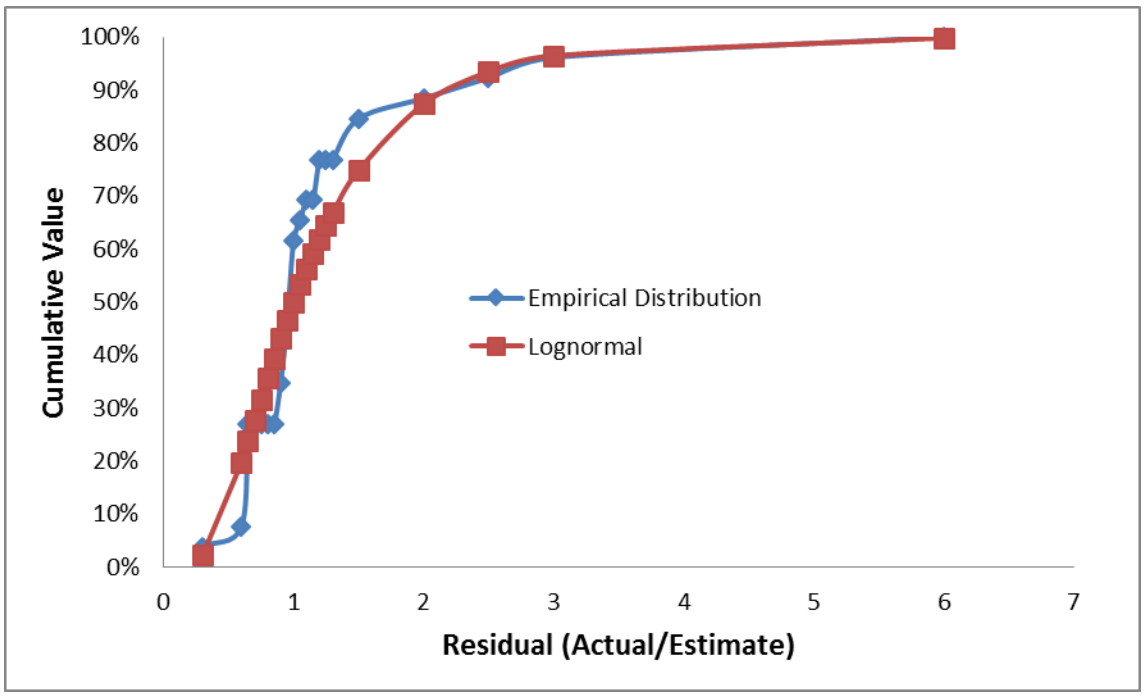

Figure 9. Graphical Comparison of Residuals for Batteries.

Don Mackenzie (Mackenzie et al. 2008) has given empirical evidence in support of the lognormal for modeling the residuals.

In a 2015 paper, Smart showed that cost growth for 289 DoD and NASA programs closely follows a lognormal distribution (Smart 2015).

The Joint Agency Cost and Schedule Risk and Uncertainty Handbook recommends the lognormal distribution as a default for modeling cost risk if no other information is known about the shape of the uncertainty distribution. (Thomas and Smith 2014)

A New Method for CER Development

We have provided theoretical and empirical evidence that the lognormal is a proper choice for CER residuals. However, the optimal method for modeling that is currently in use, LOLS, is subject to significant criticism, which involves essentially two facts. Specifically, because of the transformation LOLS estimates the median of the lognormal, and the method is not optimal (Garvey et al., 2016)

.

There is good reason to measure the median, as we have discussed, for a single project. However, in the vast majority of cases, projects are typically not budgeted for in isolation, but are part of a larger portfolio. While the sum of the means of a portfolio of projects is the portfolio mean, the percentiles of a distribution do not add. For the 50th percentile of a lognormal, since it is always less than the mean, the sums of the 50th percentiles is less than the overall 50th percentile. So estimating at the median is problematic. In practice it is much better to estimate at the mean or at another, higher risk measure. As discussed there are factors to adjust the CER estimate to the mean value.

However, these are approximations and are not accurate outside of the input data range (NCCA 2016).

Since LOLS is minimizing the error of the log of cost, which is not a "meaningful measure," then the claim gas been made that LOLS is not optimal. We have shown the LOLS is a maximum likelihood estimate of the median of the lognormal error, so it is an optimal method.

Of the two major issues, one is serious. We would like to estimate the mean of a lognormal distribution so we propose just doing that directly and avoid transformation. That is we model lognormal residuals without transformation using the same technique, MLE.. This is computationally intensive, but very easy to implement. We term this method Maximum likelihood estimation Regression for Log Normal error (MRLN or "Merlin"). We begin with the lognormal likelihood, as before, but now we are going to estimate the mean of the equation, rather than the median. We are estimating the power equation:

$$\boldsymbol{Y = \beta_0 X_1^{\beta_1} \dots X_p^{\beta_p}}$$

The mean of a lognormal density function is

$$\boldsymbol{e^{\mu+\frac{\theta}{2}}}$$

Thus for the i[th] observation, we set

$$\boldsymbol{e^{\mu_i+\frac{\theta}{2}} = \beta_0 X_{i1}^{\beta_1} \dots X_{ip}^{\beta_p}}$$

Taking log transformation of both sides of the above equation, we find

$$\boldsymbol{\mu_i + \frac{\theta}{2} = ln\beta_0 + \beta_1 lnX_{i1} + \cdots + \beta_p lnX_{ip}}$$

Therefore,

$$\boldsymbol{\mu_i = ln\beta_0 + \beta_1 lnX_{i1} + \cdots + \beta_p lnX_{ip} - \frac{\theta}{2} = ln\beta_0 + \sum_{j=1}^{p} \beta_i lnX_{ip} - \frac{\theta}{2}}$$

Recall that the likelihood for a lognormal is given by

$$\boldsymbol{L(\mu, \theta) = \prod_{i=1}^{n} \frac{1}{y_i\sqrt{2\pi\theta}} e^{-\frac{(lny_i-\mu_i)^2}{2\theta}}}$$

The log-likelihood is thus (ignoring constants)

$$l(\mu,\theta) = -\frac{1}{2\theta}\sum_{i=1}^{n}(lny_i - \mu_i)^2 - \sum_{i=1}^{n} lny_i - \frac{n}{2}ln\theta$$

We substitute for μ to obtain

$$l(\beta_0,\beta_1,\dots,\beta_p,\theta) = -\frac{1}{2\theta}\sum_{i=1}^{n}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_i lnX_{ip} + \frac{\theta}{2}\right)^2 - \sum_{i=1}^{n} lny_i - \frac{n}{2}ln\theta$$

Ignoring constants and rearranging we obtain

$$l(\beta_0,\beta_1,\dots,\beta_p,\theta) = -\frac{n}{2}ln\theta - \frac{1}{2\theta}\sum_{i=1}^{n}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_j lnX_{ij} + \frac{\theta}{2}\right)^2$$

Taking partial derivatives with respect to the parameters, we obtain

$$\frac{\partial l}{\partial\theta} = -\frac{n}{2\theta} - \frac{n}{8} + \frac{\sum_{i=1}^{n}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_j lnX_{ij}\right)^2}{2\theta^2}$$

$$\frac{\partial l}{\partial\beta_0} = -\frac{\sum_{i=1}^{n}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_j lnX_{ij}\right)}{\beta_0\theta}$$

For k = ***1,…,p,***

$$\frac{\partial l}{\partial\beta_k} = -\frac{\sum_{i=1}^{n} lnX_{ik}\left(lny_i - ln\beta_0 - \sum_{j=1}^{p}\beta_j lnX_{ij}\right)}{\theta}$$

There won't typically be a closed form solution form solution for the roots of these equations (unlike LOLS), so we will need a numerical iterative routine to solve, such as the Newton-Raphson algorithm. The Newton-Raphson method was published in Joseph Raphson's *Analysis Aequationum Universalis* in 1690. While tedious, the tools to calculate nonlinear least squares have existed before the development of the least squares method by Carl Gauss in the early 19th century.

We can utilize Excel's solver routine to minimize the negative of the log likelihood. The statistical programming language R also provides the capability to calculate maximum likelihood estimates.

We consider three examples. The examples are intended to compare the methods and not to be used for predicting future costs, so we do not do any training/testing splits of the data or cross-validation.

For our first example, we look at 121 data points for spacecraft weight and development cost. See Figure 10 for a graphical comparison.

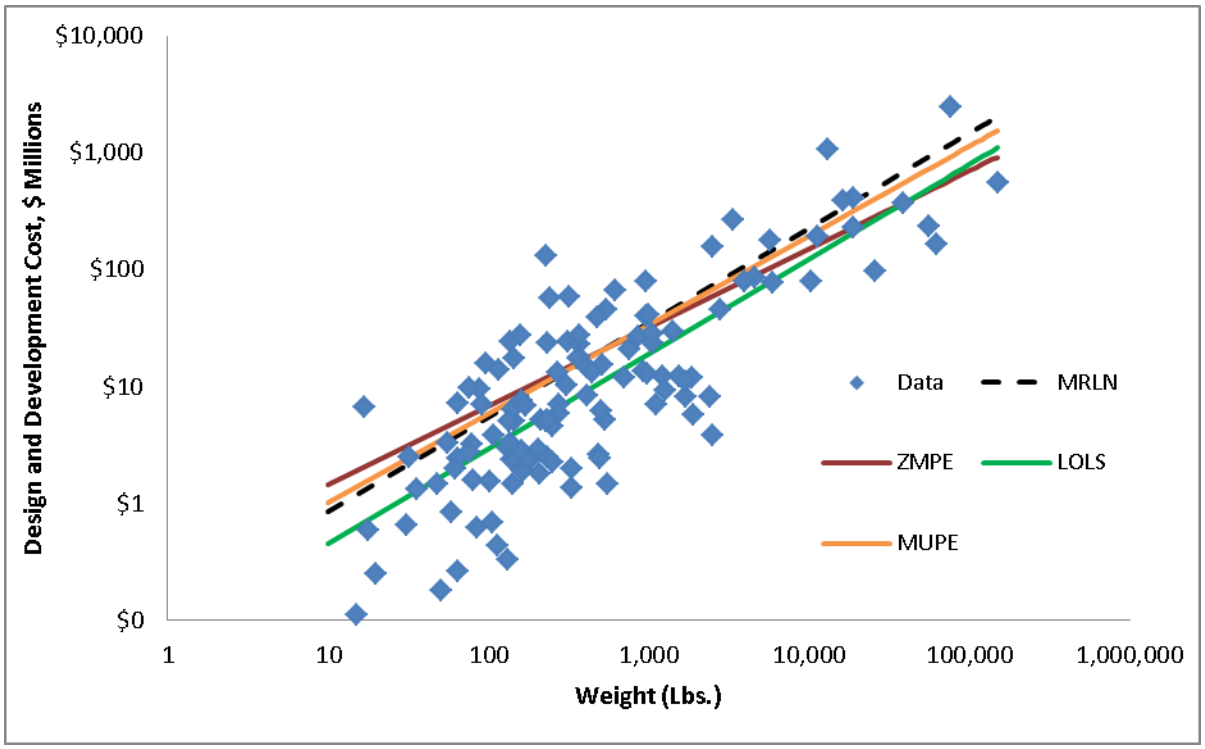

Figure 10. Design and Development Costs for the Structures Subsystem of 121 Spacecraft, Weights, and CERs.

Two commonly used measures of goodness-of-fit are Pearson's R^2 and standard percent errors. Keep in mind that these CER methods are nonlinear ones. As such, the linear R^2 commonly discussed in textbooks is problematic, since in the nonlinear case, this measure can be negative. As a measure of goodness-of-fit that does not suffer from this issue, Book and Young (1995) proposed a measure, Pearson's R^2, that circumvents this issue by calculating the square of the Pearson correlation coefficient. This is similar to the coefficient of determination and has a similar interpretation. Higher values are desired, and the metric ranges from 0 to 1. "Good" CERs often have Pearson's R^2 values above 70%. A second goodness of fit measure that measures deviations away from the fit is standard percent error. The typical linear standard error doesn't work well for the nonlinear multiplicative error case, since we are interested in percentage deviations from the actual cost. The standard percent error is a nonlinear analog to the regression standard error, and is defined as

$$\textit{Standard Percent Error} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\left[\frac{y_i - f(x_i)}{f(x_i)}\right]^2} \times 100\%,$$

where n is the sample size, and k is the number of fitted coefficients. In this case, lower values are desired. These values can be quite large, even for CERs with Pearson R^2s

above 90%. A standard percent error for spacecraft CERs below 30% is considered excellent (and rare). Note the similarity of this metric to the MPE-ZPB objective function..

The parameters for these fits are provided in Table 3.

Method	β_0	β_1	Pearson's R^2	Std % Error
MRLN	**0.13**	**0.81**	**38%**	**146%**
LOLS	**0.07**	**0.81**	**38%**	**283%**
MUPE	**0.17**	**0.76**	**39%**	**143%**
ZMPE	**0.31**	**0.67**	**40.5%**	**140%**

Table 3. Comparison of CER Fits and Goodness-of-Fit Metrics.

The four methods have similar Pearson's R^2 values. In terms of standard percent error, he direct estimation of the mean of the lognormal error has by far the lowest value. This is because the residuals follow a lognormal distribution. A fit of the residuals finds that the lognormal distribution has the lowest K-S, Chi-Square, and Anderson-Darling test values and is not rejected at the 10% critical values for any of the three tests. The gamma distribution ranks second in critical values for all three tests but is rejected at the 5% critical value for the A-D test, but not rejected at the 5% critical value for the K-S test or the Chi-Square test.

The standard errors for MRLN, MUPE, and ZMPE are all similar. MUPE and MRLN have similar trends. A visual inspection seems to indicate that these two follow the trend in the data better than ZMPE.

For our second example we consider 62 reaction control system data points for spacecraft.

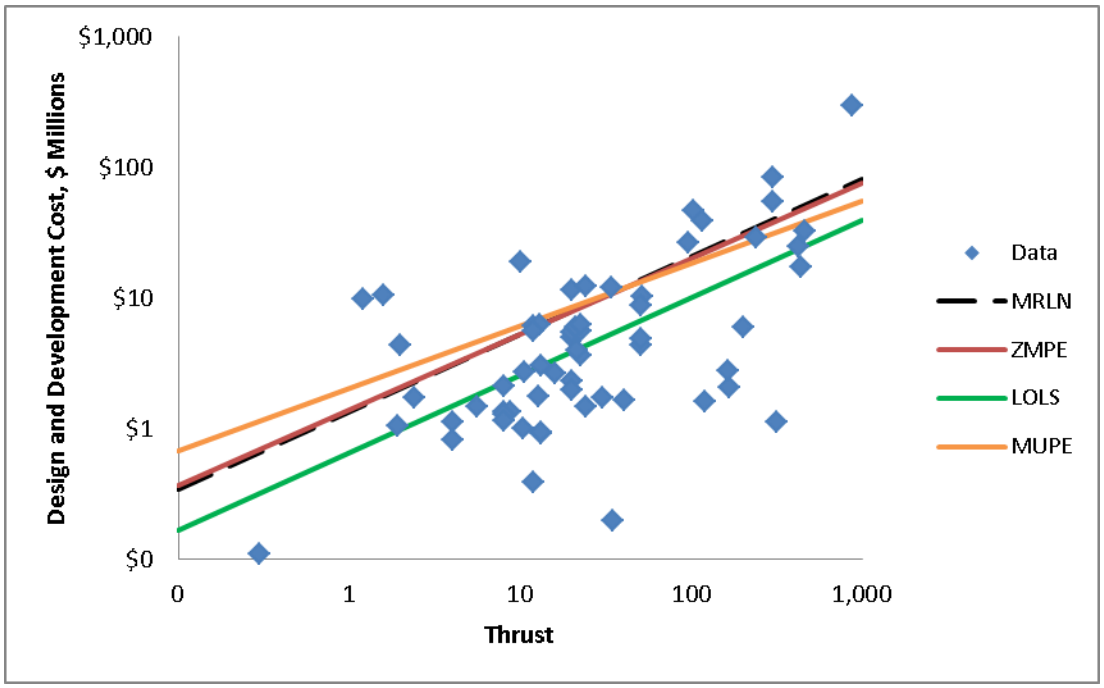

Figure 11. Design and Development Costs and Thrust Values for the Reaction Control Subsystem of 62 Spacecraft, Weights, and CERs.

Method	β_0	β_1	Pearson's R^2	Std % Error
MRLN	**1.35**	**0.59**	**54%**	**144%**
LOLS	**0.65**	**0.59**	**54%**	**315%**
MUPE	**2.04**	**0.48**	**47%**	**157%**
ZMPE	**1.38**	**0.57**	**53%**	**147%**

Table 4. Comparison of CER Fits and Goodness-of-Fit Metrics.

The direct estimation of the lognormal mean again has the lowest standard error and in this case has the highest standard error. The MUPE fit is the most different. I found that the algorithm did not really converge using Excel Solver, but bounced back and forth between the parameter in Table 4 and (0.37,0.92), but the standard error is much smaller for the parameters reported in the table. Again, for this case, the residuals closely follow a lognormal – we fail to reject the lognormal hypothesis at the 5% critical value for any of the three tests provided in Crystal Ball. The Weibull has the next best critical values, followed by the gamma, but we reject the hypothesis that the residuals follow a Weibull or gamma at the 5% and 10% critical values with the Anderson-Darling test.

Our third example is 77 command, control, and data handling data points for spacecraft. We use two independent variables – weight and % new design to predict the cost of the design and development of this subsystem.

Method	β_0	β_1	β_2	Pearson's R^2	Std % Error
MRLN	**0.45**	**1.96**	**0.96**	**97%**	**63%**
LOLS	**0.39**	**1.96**	**0.96**	**97%**	**75%**
MUPE	**0.51**	**1.98**	**0.94**	**97%**	**62%**
ZMPE	**0.65**	**2.03**	**0.89**	**96%**	**61%**

Table 5. Comparison of CER Fits and Goodness-of-Fit Metrics.

All four methods provide similar results and similar fits. The data are better behaved with fewer outliers. The lognormal is again the best fit according to all three criteria and is not rejected at the 5% critical value. The gamma is rejected at the 5% and 10% critical values for the Anderson-Darling test but is not rejected at the 5% critical value for the K-S test.

Summary

Log-transformed OLS was the first nonlinear CER method widely used in estimating costs for government projects. This method has been heavily criticized in recent years for being antiquated. However, we have shown that there are important properties that make LOLS an optimal method. In the case of lognormally distributed residuals, LOLs is an MLE. MLEs have important statistical properties, such as efficiency and consistent, as well as minimum variance. We provided a theoretical argument, evidence from other industries, and empirical evidence in the form of NAFCOM data so argue that CER residuals are lognormally distributed. We also discussed two other commonly used methods for CER development – ZMPE and MUPE – and discussed their connection to MLE.

However, LOLS does have some shortcomings, which are due to underestimating the mean (bias) and the interpretation of the transformation in log space. In order to overcome these, we have introduced the application of the maximum likelihood estimation method directly to lognormal error in unit space. This yields a new method for CER development, which we have termed Maximum likelihood estimation Regression of the Log Normal (MRLN or "Merlin"). We have provided examples comparing the results of this method with ZMPE, MUPE, and LOLS.

References

1. Babyak, M.A., "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," Psychosomatic Medicine, 66, 411-421, 2004.
2. Bahadur, R.R., and L.J. Savage, "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," Annals of Mathematical Statistics 27, 1956, pp 1115-1122.
3. Basu, A. et al., "Comparing Alternative Models: Log Vs. Cox Proportional Hazard," Health Economics, 2004.
4. Book, S.A., and P.H. Young, "General-Error Regression for USCM-7 CER Development," The Aerospace Corporation, El Segundo, CA, 1995
5. Book, S.A., and P.H. Young, "General-Error Regression for Deriving Cost-Estimating Relationships," Journal of Cost Analysis, Fall 1997, pp. 1-28.
6. Book, S.A., and N.Y. Lao, "Deriving Minimum-Percentage-Error CERs Under Zero-Bias Constraints," The Aerospace Corporation, El Segundo, CA, July 1996.
7. Book, S.A., "IRLS/MUPE CERs Are Not MPE-ZPB CERs," Presented at the International Society for Parametric Analysts Annual Conference, Seattle, WA, May 23-26, 2006.
8. Book, S.A., and P.H. Young, "The Trouble with R^2," The Journal of Parametrics, Vol. 26, No. 1, Summer 2006, pp. 87-112.
9. Embrechts, P, C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Fourth Printing, Berlin, 2003.
10. Eskew, H.L. and K.S. Lawler, "Correct and Incorrect Error Specifications in Statistical Cost Models," Journal of Cost Analysis, Spring 1994.
11. Foussier, P. M. and P. Foussier, "Should We Use the Median Instead of the OLS?," Parametric World, Fall 2008, pp. 8-11.
12. Fu, L., and R. Moncher, "Severity Distributions for GLMs: Gamma or Lognormal?", 2004 CAS Spring Meeting, Colorado Springs, CO, 2004.
13. Gallin, J.H., "Net Migration and State Labor Market Dynamics," Journal of Labor Economics, 2004.
14. Garvey, P. R., Book, S.A., and Covert, R.P., *Probability Methods for Cost Uncertainty: A Systems Engineering Perspective*, CRC Press, Boca Raton, FL, 2016.
15. Goldberg, M.S., and A.E. Tuow, *Statistical Methods Learning Curves and Cost Analysis*, Institute for Operations Research and Management Sciences, Linthicum, MD, 2003.
16. Hu, S., "The Impact of Using Log-Error CERs Outside the Data Range and Ping Factor," Presented at the Annual Joint ISPA-SCEA Conference, Denver, CO, June, 2005.
17. Hu, S., "Fit, Rather Than Assume, a CER Error Distribution," Presented at the Annual ICEAA Professional Development and Training Workshop, New Orleans, LA, June 2013.
18. Ismail, N., and A.A. Jemain, "Comparison of Minimum Bias and Maximum Likelihood Methods for Claim Severity," Casualty Actuarial Society E-Forum, Winter 2009.
19. Kleiber, C., and S. Klotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley-Interscience, Hoboken, NJ, 2003.
20. Klugman, S.A., et al., *Loss Models*, 3rd Ed., John Wiley & Sons, Hoboken, 2008.

21. Lee, D.A., *The Cost Analyst's Companion*, Logistics Management Institute, McLean, VA, 1997.
22. Mackenzie, D., "Cost Estimating Relationship Variance Study," AIAA Space Conference, Long Beach, CA, 2003.
23. Mackenzie, D., et al., "Top Level Spacecraft Cost Distribution Study," Joint Annual ISPA-SCEA Conference, Noordwijk, May, 2008.
24. Manning, W.G., and J. Mullahy, "Estimating Log Models: To Transform or Not to Transform?", Journal of Health Economics, 2001
25. Manning, W.G., et al., "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data," Journal of Health Economics, 2005.
26. Naval Center for Cost Analysis, *Cost Estimating Relationship Development Handbook*, 2016 (Draft).
27. Nelder, J. A., "Weighted Regression, Quantal Response Data, and Inverse Polynomials," Biometrics, Vol. 24 (1968), pages 979-985.
28. Seal, H.L., "Numerical Probabilities of Ruin When Expected Claim Numbers Are Large," Mitteilungen SVVM, 89-104, 1983.
29. Smart, C., "Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data," presented at the ICEAA Annual Conference, June 2014, Denver, CO.
30. Smart, C.B., "Covered with Oil: Incorporating Realism in Cost Risk Analysis," Journal of Cost Analysis and Parametrics, 8:3, 186-205, 2015.
31. Thomas, D. and A. Smith, *Joint Agency Cost Schedule Risk and Uncertainty Handbook*, 2014.
32. van Belle, G., *Statistical Rules of Thumb*, 2nd ed., Wiley, Hoboken New Jersey, 2008.
33. Wedderburn, R.W.M., "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," Biometrika, Vol. 61, Number 3 (1974), pages 439-447.