PRT-210

How Regression Methods Impact Uncertainty Results

Presented at the
International Cost Estimating and Analysis Association (ICEAA)
Professional and Training Workshop
Atlanta Georgia
7-10 June 2016

Dr. Boyan Jonov
bjonov@tecolote.com

Dr. Shu-Ping Hu
shu@tecolote.com

Alfred Smith
asmith@tecolote.com

13 May 2016

ABSTRACT

This paper was motivated by analysts who noticed that the same dataset could obtain a similar cost estimating relationship (CER) using different regression methods, but would yield a different uncertainty analysis based on the regression method used. While the point estimate would be similar, how the point estimate was interpreted (mean, median, mode, something else) and the uncertainty distribution assumption and construction were different. This paper is the result of a study we performed to provide a comprehensive discussion and a systematic approach to bring consistency to the application of uncertainty in our cost models (including MUPE and ZMPE).

Log transformation and weighted least squares are commonly used to develop multiplicative error cost estimating relationships. We objectively compare the two regression techniques and provide a sound defense of log-linear ordinary least squares (LOLS) to counter arguments against its use. We then demonstrate how the uncertainty modeling can vary substantially based on the selected regression method even when the point estimates are quite close. Lastly, we establish the criteria that will lead to a justifiable uncertainty assignment when using LOLS.

1 BACKGROUND

In this section, we review the multiplicative and additive error models as well as the LOLS, MUPE, and ZMPE regression techniques. We also provide basic definitions and establish some terminology.

1.1 ADDITIVE AND MULTIPLICATIVE ERROR MODELS

The given dataset on which regression is performed will be denoted as $(x_i, y_i)_{i=1}^{n}$. Here x_i represents the *cost drivers* and y_i – the *observed costs*. The *hypothetical equation* that models the relationship between cost and cost drivers is given by:

$$y = f(x, \beta).$$

The unknown parameters $\beta = (\beta_1, \dots, \beta_p)$ are to be solved by the regression process and the corresponding estimates will be denoted as $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$. The *predicted cost* is given by $\hat{y} = f(x, \hat{\beta})$.

The difference between the observed and the hypothesized cost at the i^{th} data point, i.e. $y_i - f(x_i, \beta)$, will be referred to as the *residual* error. The fundamental difference between the multiplicative and additive error models is based on the decision to model the magnitude of the residual error as a constant value through the entire data range (additive) or as a value varying proportionally to the level of the hypothetical equation (multiplicative). The latter is the most intuitively correct for most cost analysis applications, but we will investigate both.

The additive error model is given by the equation

$$y_i = f(x_i, \beta) + \varepsilon_i$$

where ε_i is the error associated the observed cost at the i^{th} data point. The error ε_i is assumed to have a mean 0 and a variance σ^2, for each i .The implication of this model is that the residual error is constant throughout the entire data range.[1]

The objective of the regression analysis in the additive error model setting is to solve for the coefficients β that minimize the sum of squared errors:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - f(x_i, \beta)\big)^2$$

The measure of how well the derived predicted costs $\hat{y}_i$ approximate the observed costs is called standard error of estimated (SEE):

$$SEE = \sqrt{\sum_{i=1}^{n} \frac{1}{n-p} (y_i - \hat{y}_i)^2} \qquad (1)$$

where n is the number of data points in the sample and p is the number of coefficients estimated in the hypothetical equation. For the multiplicative error model, we have:

$$y_i = f(x_i, \beta) * \varepsilon_i$$

The assumption for the error term ε_i is that it has mean 1 and variance σ^2 (for each i). The error term can be further broken down into the form:

$$\varepsilon_i = e_i - 1.$$

The variable e_i is referred to as the generalized error term with mean 0 and variance σ^2. Furthermore, the generalized error can be expressed as:

$$e_i = \frac{y_i - f(x_i, \beta)}{f(x_i, \beta)}$$

The interpretation of the above expression is that the residual error is proportional in magnitude to the hypothetical equation (rather than being constant as in the additive error model). The cost estimating community prefers the multiplicative error model because practice shows that observations that are bigger in magnitude tend to produce a proportionally bigger error in absolute terms.

The multiplicative error regression analysis seeks to minimize the sum of the squares of the generalized error term:

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

[1] This is not to be confused with the confidence or prediction interval which is based on the residual error, but also a function of location within the dataset.

The analogue to the additive error model SEE quantity that measures the "goodness of fit" of the predicted costs $\hat{y}_i$ is referred to as standard percent error (SPE) and is given by:

$$SPE = \sqrt{\sum_{i=1}^{n} \frac{1}{n-p}\left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2} \quad (2)$$

1.2 LOLS, MUPE, AND ZMPE REGRESSIONS

In this paper, we consider three popular optimization techniques: LOLS, MUPE, and ZMPE. They will be applied to the following multiplicative error model:

$$y_i = a x_i^b * \varepsilon_i \quad (3)$$

The LOLS model further assumes that the error term is normally distributed with a mean 0 and a standard deviation σ in log-space, i.e. log-normally distributed in unit space. In the case of MUPE and ZMPE, the error term is assumed to have a mean of 1 and a standard deviation σ in unit space.

LOLS (log-linear ordinary least squares): optimization is performed in log-space and the first step of the process is to take the natural log of each side of equation (3):

$$\ln(y_i) = \ln(a) + b\ln(x_i) + \ln(\varepsilon_i)$$

The above expression in log-space can be regarded a linear additive error model. As a result, OLS can be applied to minimize the sum of squares $\sum(\ln \varepsilon_i)^2$ and to solve for the parameters. The results can be further transformed back to unit space by exponentiation.

MUPE (minimum unbiased percent error): is an iterative optimization technique. At the k^{th} iterative step, MUPE solves for the coefficient β_k that minimizes the quantitiy:

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

where, $\hat{\beta}_{k-1}$ is the coefficient estimate obtained in the previous iteration. The final coefficient solution $\hat{\beta}$ is obtained when the change in the coefficient estimates in successive iteration steps is within a predefined tolerance limit.

A property of the MUPE iterative process is that the cost predictions $\hat{y}_i$ satisfy a zero percentage error (sample bias), i.e.

$$\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{\hat{y}_i} = 0$$

ZMPE (zero bias minimum percent error): is an optimization technique to reduce the percentage error. The ZMPE method imposes zero bias as a constraint in the optimization process. In other words, ZMPE seeks to directly minimize

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

subject to the constraint:

$$\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{\hat{y}_i} = 0$$

2 COMPARISON

The LOLS regression method has been subject to academic concerns and its validity has been questioned. In this section, we defend the LOLS model against common criticism and provide legitimate reasons why it is a relevant optimization choice. Furthermore, we objectively compare the LOLS, MUPE, and ZMPE regression processes by discussing their advantages and disadvantages.

2.1 LOLS PROS AND CONS

We start by addressing the strengths of the LOLS regression. If the assumptions of log-space linearity of the CER and the normal distribution of the error term in log-space are satisfied, then the LOLS process can be regarded as a preferable choice over MUPE and ZMPE for its well established and analytically sound uncertainty assignment process.

The log-linear nature of the hypothetical equation allows for OLS regression to provide coefficient estimates. Unlike the MUPE and ZMPE, the OLS solution for the coefficients is analytical and unique. This is a considerable advantage which completely bypasses reliability issues that MUPE and ZMPE face such as consistency of the coefficient estimates, dependence of the solutions on starting input, convergence and stability of the methods (getting stuck in a local minimum due to choice of starting points), etc. Moreover, LOLS regression is linear in nature, while MUPE and ZMPE rely on nonlinear computations which could easily become tedious and cumbersome to validate.

The analytical solution of the LOLS' coefficients is a major advantage in the field of uncertainty assignment. Assuming a log-normally distributed error term in unit space, LOLS analytic approach leads to a sound and justifiable uncertainty distribution assignment for the CER result (details provided below). The distribution shape is proven to be log-normal in unit space, and PE location within the distribution is established to be the median. As a result, prediction intervals (PI) can be precisely generated. Unlike the LOLS technique, regressions such as ZMPE provide neither a mathematically proven error distribution type nor a verified location for the PE.

LOLS regression also has the advantage of providing goodness-of-fit measures that are essential for a thorough analysis of the quality of the fit in log-space. As a result, the significance of the coefficients can be analyzed, outliers can be detected, and model flaws can be exposed. Other optimization techniques, such as ZMPE, provide only a limited goodness of fit measures and therefore restricted options for analyzing the fit quality.

Next, we address the concerns raised in reference [1] about the LOLS model. We show that, except for one case, the reported flaws and the criticism are invalid and unjustifiable.

(1) The LOLS objective is to minimize the quantity

$$\sum (\ln y_i - \ln a - b \ln x_i)^2 = \sum (\ln \varepsilon_i)^2$$

Reference [1] regards this as a flaw by stating that minimizing $\sum(\log \varepsilon_i)^2$ is not the same as minimizing the sum of squared error $\sum e_i^2$:

$$\sum \left(y_i - a x_i^b\right)^2 = \sum e_i^2$$

The LOLS optimization process, however, was never intended to minimize $\sum e_i^2$. As a multiplicative error model, LOLS goal is to minimize the sum of squared percentage errors, not the sum of squared absolute errors (which is the objective of additive error models). It is inappropriate to compare the fit measure of different error model when they have different fit criteria.

(2) The second concern from reference [1] is that the log-space error term $\ln(\varepsilon_i)$ is expressed in meaningless units (i.e. log of dollars instead of dollars). An immediate response is that the error term ε_i is never measured in dollars for a multiplicative error model. Instead, ε_i is a unit-less ratio by design. Moreover, the error term $\ln(\varepsilon_i)$ does have a meaningful interpretation. By Taylor series expansion, we have

$$\ln(\varepsilon_i) \approx \varepsilon_i - 1 = \frac{y_i - f(x_i, \beta)}{f(x_i, \beta)}$$

The expression on the right-hand side can be recognized as the percentage error term which is incorporated into the objective function of the multiplicative error model.

(3) The log-space transformation process is further criticized for restricting the CER choice to power forms such as $y = ax^b$. Indeed, the OLS method cannot handle fixed cost equations $y = ax^b + c$ because they are nonlinear in log-space (the distributive nature of the log function is not compatible with additive terms such as c). However, we are not limited to the OLS regression. Multiplicative error models $y = (ax^b + c) * \varepsilon$ still become additive in log-space and they can readily be handled by non-linear optimization. We should keep in mind that the choice of the CER and the error model should be driven by technical grounds and logic and not by the desire to perform a preselected regression technique.

(4) The one shortcoming raised in reference [1] about the LOLS method that we acknowledge is the fact that the LOLS solution is biased in unit space. In log-space, the OLS regression derives an equation with zero bias:

$$\frac{1}{n}\sum_{i=1}^{n}\left(ln\hat{a} + \hat{b}\, lnx_i - lny_i\right) = 0$$

In unit space, however, the corresponding CER has a non-zero proportional error over the dataset:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i - \hat{a}\, x_i^{\hat{b}}\right)}{\hat{a}\, x_i^{\hat{b}}} \neq 0$$

To remove the bias, multiplicative factors have been developed (see reference [4]) so that the corrected LOLS CER estimates the mean in unit space. However, these adjustments are rarely necessary. Many different estimating methods are used throughout the work breakdown structure (WBS). It is not uncommon to have to add a mean, median, mode, or

some percentile because either policy of the method itself yields these types of results. We know the LOLS produces the median. Using that as the point estimate directly and as one point in the log-normal distribution (using the prediction interval for a second point) uniquely defines the uncertainty distribution. It will be the exact same distribution if you chose to adjust the PE to reflect the mean. There would be no impact on the uncertainty simulation. So why complicate the situation by including an adjustment factor?

2.2 MUPE & ZMPE PROS AND CONS

We finish this section by addressing the strengths and weaknesses of the MUPE and ZMPE regression. These two methods have been thoroughly compared and contrasted in reference [6]. A summary of the main observations from that reference are as follows.

MUPE Pros

- The MUPE regression provides estimators with zero percent bias directly in unit space. Unlike the LOLS process, no transformation or correction factors are applied to the CER result.
- For linear CERs, MUPE provides the best linear unbiased estimates (BLUE) solutions for the parameters. Under the same linear assumptions, ZMPE solutions are not BLUE. For nonlinear CERs, MUPE gives consistent estimates for the parameters and mean of the equation. Moreover, the parameter estimates are the maximum likelihood estimators (MLE).
- Under the normality assumption, the MUPE process provides a wider variety of goodness of fit measures than ZMPE to judge the quality of the model. In particular, statistical tools are available to analyze significance level of the coefficient estimates which helps detect model flaws.
- Statistical tools are available to provide prediction interval for MUPE's CER result.

MUPE Cons

- The MUPE regression relies on non-linear optimization which can be a cumbersome process.
- MUPE's iterative process does not always converge.

ZMPE Pros

- Similar to the MUPE method, an unbiased CER result is provided without the need of transformation or adjustment factors.
- ZMPE's standard percent error is reported to be smaller than MUPE's SPE (Remark: This statement is actually not true if the ZMPE's SPE is adjusted to reflect the generalized degree of freedom, which accounts for the regression constraint, as recommended in reference [3]).

ZMPE Cons

- ZMPE's solution finding process can be less reliable than MUPE and far less than LOLS. ZMPE's optimization fails to converge more often as a result of a tendency to being trapped in local minima. Stability of ZMPE's solutions is directly linked to Solver's (or other selected optimization tool) sensitivity to the input starting points.

- The only goodness of fit measures available for the ZMPE regression are the SPE and R2. However, there is not enough information to analyze coefficient significance which further translates into an inability to characterize the statistical significance of the model.
- The location of the CER result with the uncertainty distribution (mean, median, mode, etc.) is not established. Mode of a triangle has been used, but the choice is arbitrary. Neither the distribution shape nor its dispersion can be formally determined. Therefore, confidence and prediction intervals are unavailable.
- Similarly to MUPE, ZMPE optimization relies on non-linear regression which can be a tedious process.

3 UNCERTAINTY

Of the three regression techniques discussed in this paper, the LOLS method leads to the most precise and justifiable uncertainty assignment for the CER result (assuming that one can justify the error is log-normally distributed in unit space). In this section, we provide the mathematical derivation of the uncertainty distribution for the LOLS CER. In particular, we show that the CER result is the median of a log-normal distribution. In contrast, the techniques available for uncertainty assignment to ZMPE's predictors are arbitrary, subjective, and not mathematically supported. In an effort to establish consistency, we propose a systematic approach to assign uncertainty to ZMPE's CER result. We also briefly discuss MUPE's uncertainty assignment process which is analytical in nature but is based on approximation. We finish the section with examples that compare the uncertainty results of the three regressions and we present our conclusions based on the observations.

3.1 LOLS UNCERTAINTY

We will show that the uncertainty distribution around the LOLS CER result is given by the following expression:

$$\text{LOLS CER Uncertainty: } \hat{y}_0 * \hat{\varepsilon}_0 \quad where \quad \hat{\varepsilon}_0 \sim LN(0, \sigma^2\,[1 + \gamma^2(X, x_0)]), \tag{4}$$

where $\hat{y}_0$ is the LOLS predictor at a given cost driver point $x = x_0$. We note that $\hat{y}_0$ is the median of the log-normal distribution in (4) (see reference [4]). We use the following notation:

$$\gamma^2(X, x_0) = \; (1, \ln(x_0))\,(X^T X)^{-1}(1, \ln(x_0))^T$$

$$= \frac{1}{n} + \frac{\left(\ln(x_0) - \overline{\ln(x)}\right)^2}{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2} \tag{5}$$

The above term can be interpreted as a location factor that captures the location of the estimating point x_0 relative to the mean $\overline{\ln(x)}$ of the cost driver sample in log-space.

The only unknown in expression (4) is σ, the standard deviation of the observed cost error. However, in practical applications, it can be approximated the by LOLS' SEE in log-space since $E[SEE^2] = \sigma^2$.

To derive the above formula, we start by rewriting the log-linear error model in the form

$$y_i = e^{\beta_0}\, x_i^{\beta_1}\, \varepsilon_i$$

which is more suitable for manipulations in log-space. The error term is assumed to be log-normally distributed in unit space: $\varepsilon_i \sim LN(0, \sigma^2)$. Taking the natural log of each side of the equation, we get:

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \ln(\varepsilon_i) \quad \text{with} \quad \ln(\varepsilon_i) \sim N(0, \sigma^2) \qquad (6)$$

Therefore, the distribution of the dependent variable $\ln(y)$ for a given cost driver $x = x_0$ is given by:

$$\ln(y_0) \sim N\left((1, \ln(x_0)) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \sigma^2 \right)$$

Since equation (6) is a linear additive error model, OLS regression can be applied to solve for the coefficients in log-space. It can be shown that the OLS coefficient estimates and their uncertainty are given by:

$$\hat{\beta} = (X^T X)^{-1} X^T \ln(Y) \quad \text{with} \quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

In the above expression, we use the notation:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \beta_1 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & \ln(x_1) \\ \vdots & \vdots \\ 1 & \ln(x_n) \end{pmatrix}, \qquad \ln(Y) = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_n) \end{pmatrix}$$

with $(x_i, y_i)_{i=1}^n$ being the given dataset.

Consequently, the OLS predictor in log-space is given by:

$$\ln(\hat{y}_0) = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_0) = (1, \ln(x_0)) \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

and it follows a normal distribution:

$$\ln(\hat{y}_0) \sim N\left((1, \ln(x_0)) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\ \sigma^2 * \gamma^2(X, x_0) \right)$$

where γ^2 is given by (5).

The prediction error (the error between the predicted and the observed value) and its distribution are therefore given by:

$$\ln(\hat{\varepsilon}_0) = \ln(y_0) - \ln(\hat{y}_0) \sim N(0,\ \sigma^2 [1 + \gamma^2(X, x_0)])$$

The uncertainty of the predicted cost in log-space is therefore:

$$\ln(\hat{y}_0) + \ln(\hat{\varepsilon}_0) \quad \text{with} \quad \ln(\hat{\varepsilon}_0) \sim N(0,\ \sigma^2 [1 + \gamma^2(X, x_0)])$$

Taking exp() on each side, we arrive at (4).

3.2 MUPE\ZMPE UNCERTAINTY

The uncertainty assignment for the MUPE CER results is based on analysis that involves Taylor series linearization (see reference [5]). It is an analytical method but it involves approximations

and does not have the precision of LOLS' uncertainty closed-form formula. The prediction intervals for MUPE regression can be provided by statistical tools.

For the ZMPE regression there is no established and analytically supported process for CER uncertainty assignment. The shape of the uncertainty distribution is generally unknown and so is the position of the CER result.

To provide some structure and consistency in the uncertainty assignment for the ZMPE CER result, we propose a procedure that uses the statistics of the ratio of observed and predicted cost. In particular, we start by fitting a distribution curve through the points $\left\{\frac{y_1}{\hat{y}_1}, \dots, \frac{y_n}{\hat{y}_n}\right\}$. The exact fitting process is proposed in reference [2] and it takes into account the location of the driver x_0 relative to the other cost drivers. The fitted curve, which we find with the help of statistical tools, represents the error distribution around the CER result $\hat{y}_0$ at given cost driver $x = x_0$. Since we are assuming multiplicative error model, the final uncertainty distribution is given by:

$$\hat{y}_0 * \text{Fitted Distribution Curve } of \left\{\frac{y_1}{\hat{y}_1}, \dots, \frac{y_n}{\hat{y}_n}\right\} \quad (7)$$

3.3 EXAMPLES

The LOLS, MUPE, and ZMPE regressions differ not only by the optimization process that derives the CER result, but also by the methods used to assign uncertainty to the predictors. To quantify the differences in the uncertainty results, we assign the uncertainty distributions presented above on specific data sets and we compare the corresponding 80th percentiles.

Example 1 For our first example, we consider a hypothetical dataset of 7 data points. The observed costs were derived using the relation:

$$y_i = 0.07x_i^{1.8}\varepsilon_i\,, \quad \text{with} \quad \varepsilon_i \sim LN(0, \sigma^2), \quad (8)$$

where $i = 1, \dots, 7$. Choosing $\sigma = 0.34$ for the spread of the error term, we construct the following dataset:

Observations	1	2	3	4	5	6	7
X – Cost Driver	7.9	8.2	9.8	11.5	16.4	19.7	23.6
Y – Observed Cost	1.6	3.2	2.3	5.1	7.5	16.3	14.5

The table below summarizes the regression and uncertainty results for the LOLS, MUPE, and ZMPE methods. The first section of the table provides the CER equations of each regression along with the corresponding goodness of fit statistics: SEE (see equation (1)), SPE (see equation (2)), and CV. The remaining part of the table contains the point estimates (PE) and the 80th percentiles (Ptile) of the uncertainty distribution all evaluated at the cost driver value $x_0 = 21$. The percentiles in parenthesis next to MUPE's and ZMPE's PE and Ptile values denote the percent difference relative to the corresponding LOLS' estimates.

LOLS		MUPE		ZMPE	
$y = 0.038x^{1.936}$	SEE: 2.38	$y = 0.042x^{1.913}$	SEE: 2.386	$y = 0.046x^{1.869}$	SEE: 2.343
	SPE: 0.316		SPE: 0.302		SPE: 0.302
	CV: 0.329		CV: 0.330		CV: 0.324

	LOLS		MUPE		ZMPE	
	PE	80 Ptile	PE	80 Ptile	PE	80 Ptile
$x_0 = 21$	13.8	18.4	14.1(2%)	18.6(1%)	13.8(0%)	21.8(18%)

LOLS CER uncertainty is computed by formula (4). By construction, the error of the observed cost here follows log-normal distribution (see (8)). Therefore, the use of (4) is justified here. ZMPE's uncertainty is assigned as described in (7). In this particular case, the fitted distribution curve is a beta curve with parameters $\alpha = 0.2, \beta = 0.29, min = 0.54, max = 1.65$. Finally, MUPE's 80th percentile is computed with statistical tools.

We notice in this example that ZMPE's 80th percentile is 18% bigger than LOLS' while the corresponding point estimates are identical. For MUPE, both the point estimate and the 80th percentile are close to LOLS' values.

Example 2 In our second example, we generate a hypothetical 9-point dataset from the equation:

$$y_i = 64x_i^{0.7}\varepsilon_i\,, \quad \text{with} \quad \varepsilon_i \sim LN(0, \sigma^2),$$

where $i = 1, \dots, 9$ and $\sigma = 0.34$.

Observations	1	2	3	4	5	6	7	8	9
X – Cost Driver	5	5.2	7	10	12	17.8	18	21	25
Y – Observed Cost	205.3	111	225.6	182.2	255.3	523.4	695.8	377	638.5

The corresponding results table is:

LOLS		MUPE		ZMPE	
$y = 33.4x^{0.9}$	SEE: 121.7	$y = 35.5x^{0.892}$	SEE: 120.0	$y = 36.4x^{0.881}$	SEE: 120.2
	SPE: 0.343		SPE: 0.326		SPE: 0.326
	CV: 0.341		CV: 0.336		CV: 0.337

	LOLS		MUPE		ZMPE	
	PE	80 Ptile	PE	80 Ptile	PE	80 Ptile
$x_0 = 22$	539	732	559(4%)	741(1%)	555(3%)	813(11%)

Similar to Example 1, we notice that ZMPE leads to a relatively large (11%) difference in the 80th percentiles compared to LOLS while the corresponding point estimates are only 3% apart. MUPE's and LOLS's 80th percentiles differ by only 1%.

Example 3 We start with the following 13 data points:

Obs.	1	2	3	4	5	6	7	8	9	10	11	12	13
X	40	50	75	75	75	100	100	240	250	300	550	670	780
Y	10	45	50	70	65	100	90	120	100	80	200	230	300

Remark: The observed cost in this example is not assumed to be generated from any specific hypothetical equation and error term distribution.

The results are given below:

LOLS		MUPE		ZMPE	
$y = 2.059x^{0.7336}$	SEE: 27.19	$y = 3.047x^{0.67}$	SEE: 27.277	$y = 4.359x^{0.6}$	SEE: 30.19
	SPE: 0.392		SPE: 0.337		SPE: 0.329
	CV: 0.242		CV: 0.243		CV: 0.269

	LOLS		MUPE		ZMPE	
	PE	80 Ptile	PE	80 Ptile	PE	80 Ptile
$x_0 = 500$	196	297	196(-0%)	259(-12%)	181(-7%)	244(-17%)

In this case both MUPE's and ZMPE's 80th percentiles differ substantially from LOLS value (- 12% and -17%, correspondingly). At the same time, the point estimates of MUPE and LOLS are identical.

3.4 CONCLUSION

As the three examples from the previous section demonstrate, LOLS, MUPE, and ZMPE regressions can lead to substantially different uncertainty results even in cases when the point estimates are not too far apart. In the first two examples, the error of the observed cost is ***known*** to be log-normally distributed. Therefore, LOLS uncertainty results are sound and mathematically justified. Other non-linear regressions, such as ZMPE, do not have an established uncertainty assignment procedure and the corresponding uncertainty results are not as reliable. As a result, whenever there is enough evidence that observed cost is log-normally distributed, we recommend using LOLS regression.

For the data set in the last example, we have no solid reasons to believe that the observed cost is generated from a log-normally distributed error term. Thus, LOLS uncertainty assignment can no longer be regarded as more precise compared to ZMPE and MUPE. In such cases, we still have

confidence in the LOLS method because it is stable and the CER results are analytically sound. However, there are no compelling reasons to reject MUPE or ZMPE regressions in these circumstances. Just beware that they are sensitive to the assigned starting position and the uncertainty assignment is an assumption. Some prefer MUPE and ZMPE because the CER results have a zero percentage bias without the need of correction factors, but as stated earlier that is actually of no significance when it comes to uncertainty assignment.

APPENDIX A
References

1. Book, S., "Significant Reasons to Eschew Log-Log OLS Regression when Deriving Estimating Relationships," 2012 ISPA/SCEA Joint Annual Conference, Orlando, FL, 26-29 June.
2. Hu, S., "Fit, Rather Than Assume, a CER Error Distribution," 2013 ICEAA Annual Conference, New Orleans, LA, 18-21 June 2013
3. Hu, S., "Generalized Degrees of Freedom(GDF)," 2015 ICEAA Annual Conference, San Diego, CA, June 9-12.
4. Hu, Shu-Ping. 2005. "The Impact of Using Log CERs Outside the Data Range and PING Factor." Paper presented at Joint ISPA/SCEA International Conference, Denver, CO, June 14-17.
5. Hu, S., "Prediction Interval Analysis for Nonlinear Equations," 2006 Annual SCEA International Conference, Tysons Corner, VA, 13-16 June 2006
6. Hu, S. and A. Smith, "Why ZMPE When You Can MUPE," 6th Joint Annual ISPA/SCEA International Conference, New Orleans, LA, 12-15 June 2007.