

R² vs. r²

Dr. Shu-Ping Hu
Tecolote Research, Inc.
5266 Hollister Ave., Ste. 301
Santa Barbara, CA 93111

ABSTRACT

Cost estimating relationships (CER) with multiplicative error assumptions are commonly used in cost analysis. Consequently, we need to apply appropriate statistical measures to evaluate a CER's quality when developing multiplicative error CERs such as the Minimum-Unbiased-Percentage Error (MUPE) and Minimum-Percentage Error under the Zero-Percentage Bias (ZMPE) CERs.

Generalized R-squared (GRSQ, also denoted by the symbol r^2) is commonly used for measuring the quality of a nonlinear CER. GRSQ is defined to be the square of Pearson's correlation coefficient between the actual observations and the CER predicted values (see Reference 9). Many statistical analysts believe GRSQ is an appropriate analog to measure the proportion of the variation explained by a nonlinear CER (see Reference 7), including the MUPE and ZMPE CERs; some even use it to measure the appropriateness of shape of a CER.

The adjusted R² in unit space is a frequently used alternative measure for CER quality. This statistic translates the error sum of squares (SSE) from the absolute scale to the relative scale. This metric is used to measure how well the CER-predicted costs match the actual data set.

There have been academic concerns over the years about the relevance of using adjusted R² and Pearson's r². For example, some insist that adjusted R², calculated by the traditional formula, has no value as a metric except for ordinary least squares (OLS); others argue that Pearson's r² does not measure how well the estimate matches the database actuals for nonlinear CERs. This paper discusses these concerns and examines the properties of these statistics, along with the pros and cons of using each for CER development. In addition, this paper proposes (1) a modified adjusted R² for evaluating MUPE CERs and (2) a modified GRSQ to correct for degrees of freedom.

INTRODUCTION

We will briefly introduce four different methodologies for fitting multiplicative error models; we will start with the Log-Error method.

Log-Error Method. The multiplicative error model is generally stated as follows:

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) \varepsilon_i \quad \text{for } i = 1, \dots, n \quad (1)$$

where:

- n = sample size
- Y_i = observed cost of the i^{th} data point, $i = 1$ to n
- $f(\mathbf{x}_i, \boldsymbol{\beta})$ = the value of the hypothesized equation at the i^{th} data point
- $\boldsymbol{\beta}$ = vector of coefficients to be estimated by the regression equation
- \mathbf{x}_i = vector of cost driver variables at the i^{th} data point
- ε_i = error term

If the multiplicative error term (ε_i) is further assumed to follow a log-normal distribution with a mean of 0 and variance σ^2 in log space, then the error can be measured by the following:

$$e_i = \ln(\varepsilon_i) = \ln(Y_i) - \ln(f(\mathbf{x}_i, \boldsymbol{\beta})) \quad (2)$$

where “ln” stands for nature logarithmic function. The objective is then to minimize the sum of squared e_i s (i.e., $(\sum(\ln(\varepsilon_i))^2)$). If the transformed function is linear in log space, then OLS can be applied in log space to derive a solution for $\boldsymbol{\beta}$. If not, we need to apply the nonlinear regression technique to derive a solution.

Although a least squares optimization in log space produces an unbiased estimator in log space for log-linear models, the estimator is no longer unbiased when transformed back to unit space (see References 3, 11, and 14). However, the magnitude of the bias can be corrected with a simple factor if the errors are distributed normally in log space (see References 3 and 11). Because of this shortcoming, the Minimum-Percentage-Error (MPE) and Minimum-Unbiased-Percentage-Error (MUPE) methods are recommended for modeling multiplicative error directly in unit space.

MPE and MUPE Methods. The general specification for a MPE model, as well as a MUPE model, is the same as given above (Equation 1), except that the error term is assumed to have a mean of 1 and variance σ^2 . Based upon this assumption of a multiplicative model, a generalized error term is defined by

$$e_i = \frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})}{f(\mathbf{x}_i, \boldsymbol{\beta})} \quad (3)$$

where e_i now has a mean of 0 and variance σ^2 .

This percentage error differs from the traditional percentage error in the denominator, where predicted cost instead of actual cost is used as the baseline. The optimization objective is to find the coefficient vector $\boldsymbol{\beta}$ that minimizes the sum of squared e_i s:

$$\text{Minimize } \sum_{i=1}^n \left(\frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})}{f(\mathbf{x}_i, \boldsymbol{\beta})} \right)^2 = \sum_{i=1}^n e_i^2 \quad (4)$$

The MPE method produces a regression equation for the condition that the function (i.e., $f(\mathbf{x}_i, \boldsymbol{\beta})$) in both the numerator and denominator are solved simultaneously (see Reference 10). Due to the simultaneous minimization, the resultant MPE equation is biased high (see Reference 7 for details). To eliminate the bias, the MUPE method solves for the function in the numerator separately from the function in the denominator through an iterative process,

$$\text{Minimize } \sum_{i=1}^n \left(\frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta}_k)}{f(\mathbf{x}_i, \boldsymbol{\beta}_{k-1})} \right)^2 = \sum_{i=1}^n \left(\frac{y_i - f_k(\mathbf{x}_i)}{f_{k-1}(\mathbf{x}_i)} \right)^2 \quad (5)$$

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

Note that the weighting factor of each residual in the current iteration is equal to the reciprocal of the predicted value from the previous iteration. Since the denominator in Equation 5 is kept fixed

throughout the iteration process, the MUPE technique turns out to be a weighted least squares with an additive error. This optimization technique (Equation 5) is commonly referred to as Iteratively Reweighted Least Squares (IRLS, see References 12 and 13). The corresponding standard error of estimate for the MUPE CER is commonly termed multiplicative error or standard percent error (SPE):

$$SPE = \sqrt{\sum_{i=1}^n ((y_i - \hat{y}_i) / \hat{y}_i)^2 / (n - p)} \quad (6)$$

Note that \hat{y}_i is the predicted value in unit space for the i^{th} data point. The MUPE CER provides consistent estimates of the parameters and has zero proportional error for all points in the data set. See Reference 5 or 8 for detailed descriptions of the MUPE method.

ZMPE Method. There is another alternative method to reduce the positive proportional error for MPE CERs and yet maintain the same objective function. Mathematically, it is stated as follows:

$$\begin{aligned} &\text{Minimize } \sum_{i=1}^n \left(\frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})}{f(\mathbf{x}_i, \boldsymbol{\beta})} \right)^2 \\ &\text{Subject to } \sum_{i=1}^n \frac{y_i - f(\mathbf{x}_i, \boldsymbol{\beta})}{f(\mathbf{x}_i, \boldsymbol{\beta})} = 0 \end{aligned} \quad (7)$$

This alternative method (Equation 7) is a “constrained” minimization process. It is commonly referred to as the Zero-Percentage Bias method under MPE, i.e., the ZPB/MPE or ZMPE method by Book and Lao, 1999 (see Reference 6).

In the following sections we will discuss these R^2 -related topics:

- Definitions and Formulas of R^2 , Adjusted R^2 , and GRSQ (r^2)
- Concerns about GRSQ, R^2 and Adjusted R^2
- Analyze R^2 /Adjusted R^2 and GRSQ Using Examples
- Why Is GRSQ Insensitive?
- Modify Adjusted R^2 (for MUPE) and GRSQ (to correct for DF)

DEFINITIONS OF R^2 , ADJUSTED R^2 , AND GRSQ (r^2)

Definition of R^2 . The coefficient of determination (R^2) measures how much total variation about the mean has been explained by the regression equation. By definition, the traditional R^2 is calculated by dividing the regression sum of squares (SSR) by the total sum of squares (SST):

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \quad (8)$$

The SST is the sum of squared deviations of the dependent variable (i.e., cost) about the mean. When the independent variables are not available, the mean is the best estimator. Hence it becomes the basis for measuring the improvement due to regression. The regression sum of squares (i.e., explained variation, denoted by SSR) is defined as the sum of squared deviations of

the CER-predicted values about the mean. In OLS, the mean of the predicted values is the same as the mean of the actual observations.

Another commonly used definition of R^2 is given below:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (9)$$

Equation 9, a more general definition of R^2 , measures the percent difference (percent error) between the total sum of squares and error sum of squares. The error sum of squares (SSE) is the sum of squared deviations between the actual and the estimated costs.

In OLS, R^2 can be calculated by either Equation 8 or Equation 9 because SST equals the sum of SSR and SSE. However, this may not be true for nonlinear regression (including MUPE) because the sum of the cross product term ($\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$) does not necessarily vanish. Also, the fitted equation does not necessarily go through the mean.

Users of OLS CERs are accustomed to describing the quality of fit of a CER in terms of the ratio of the explained variation to the total variation to be explained (i.e., the traditional R^2 measure). In nonlinear regression, however, such a measure is, unfortunately, not readily defined. A measure relatively close to R^2 in the nonlinear case is Equation 9 and we recommend using it for the definition of R^2 . (SSR is not well-defined in the nonlinear case.) Note that we do not use R^2 (Equation 9) to indicate the proportion of the variation in the data set that is accounted for by a CER except for OLS.

Definition of Adjusted R^2 . If data sets are small, the measure of R^2 can over-state the explanatory power of the regression result because it does not take the degrees of freedom (DF) into account. An analyst can increase R^2 by including many independent variables in the model, even if those variables have very little predictive power. In such cases, it is more useful to provide an R^2 that has been corrected for the sampling bias. The adjusted R^2 is R^2 adjusted for degrees of freedom. As shown below, the adjustments are made for the corresponding degrees of freedom of both SSE and SST:

$$Adj. R^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{\sum (y_i - \hat{y}_i)^2/(n-p)}{\sum (y_i - \bar{y})^2/(n-1)} = \frac{MSE_{\bar{y}} - MSE_f}{MSE_{\bar{y}}} \quad (10)$$

In essence, the adjusted R^2 in unit space translates SSE from the absolute scale to the relative scale by 1) comparing SSE to SST and 2) adjusting the degrees of freedom for small samples. It measures how well the CER-predicted costs match the actual data set. The closer the Adjusted R^2 (or R^2) is to one, the closer the estimates match the actual observations.

Both R^2 and adjusted R^2 can be evaluated in either the fit space or the unit space. All evaluations in this paper are done in unit space.

Interpretation of Adjusted R^2 . We can use *Adjusted R^2 in unit space* to compare the CER's performance to the starting point, i.e., the mean square error (MSE) of an average CER when the driver variables are not available. For example, if a CER's estimated variance is 0.1, while the sample variance of the dependent variable (i.e., S_y) is 0.5, then the CER's variance is only 20% of the sample variance. This reduction of variance, 80%, is the Adjusted R^2 in unit space, which is considered to be an "improvement" when applying the CER.

The following formulas relate Adjusted R^2 to R^2 :

$$\begin{aligned} \text{Adjusted } R^2 &= R^2 - (1 - R^2) * (p-1) / (n-p) \\ &= 1 - (1 - R^2) * (n-1) / (n-p) \end{aligned} \quad (11)$$

It follows from the above equation that Adjusted R^2 is bounded above by R^2 .

Definition of GRSQ (r^2). Pearson's correlation coefficient (Pearson's r) calculated between two sets of numbers $\{x_i\}$ and $\{y_i\}$ is defined to be

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (12)$$

where \bar{x} and \bar{y} are the means of $\{x_i\}$ and $\{y_i\}$, respectively, and n is the sample size.

Generalized R-squared (GRSQ, also denoted by the symbol r^2) is commonly used for measuring the quality of a nonlinear CER. By definition, GRSQ is the square of Pearson's correlation coefficient between the actual observations (y_i) and CER predicted values (\hat{y}_i):

$$GRSQ = r^2(y, \hat{y}) = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) \right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \quad (13)$$

where \bar{y} and $\bar{\hat{y}}$ are the averages of the actual and CER-predicted values, respectively.

Note that Pearson's correlation coefficient between two sets of numbers is a measure of the linear association between them. It measures the degree to which two sets of data move together in a linear manner. A high positive correlation indicates a strong direct linear movement and a high negative correlation represents a strong inverse relationship. Pearson's correlation coefficient ranges between -1.0 and $+1.0$, where -1 indicates a perfect inverse relationship, 0 indicates no correlation, and $+1$ indicates a perfect positive relationship.

The value of GRSQ can be shown to be equal to the R^2 measure in OLS, i.e.,

$$GRSQ = r^2(y, \hat{y}) = R^2 = 1 - SSE/SST \quad \text{for OLS CERs} \quad (14)$$

However, GRSQ does not indicate the goodness of a CER if it is not generated by OLS. This is because Pearson's correlation coefficient does not provide insight into the deviation between the estimated and the actual for nonlinear CERs. Also, this measure is insensitive to different fitting methods and different equation forms. Detailed analyses and discussion will be given below.

General Cautions on R^2 , Adjusted R^2 , and GRSQ. The measures R^2 , Adjusted R^2 , and GRSQ defined above are all computed in unit space, so they are all predictive measures and are used to evaluate a model's actual predictive power. However, they may be highly influenced by extreme values, which will make unreasonable relationships look very attractive. Therefore, sole reliance on any one of them should be avoided and additional data analysis and residual examination are strongly recommended.

CONCERNS ABOUT R^2 AND ADJUSTED R^2

There have been several papers and articles discussing the pitfalls and concerns about using R^2 in regression analysis. Below are a few criticisms about R^2 and Adjusted R^2 raised recently:

- R^2 , as well as Adjusted R^2 , has no value as a metric in cases other than OLS.
- The formulas of R^2 and Adjusted R^2 are inapplicable.
- Many good CERs may be dismissed when using Adjusted R^2 because they might have a negative Adjusted R^2 .

We believe the main problem with R^2 , as well as Adjusted R^2 , is **confusion** because (1) R^2 can be calculated by two different formulas (see Equations 8 and 9) and (2) R^2 can be used for many different types of equations, including the nonlinear ones. In addition, two different spaces (i.e., domains) are commonly used to evaluate this statistic: one is the **fit** space; the other the **unit** space. The fit space is referring to the domain where the regression equation is derived by the optimization technique, while the unit space denotes the domain of the dependent variable. The fit and unit spaces are the same for regression equations using OLS.

All confusion can be avoided as long as people understand how R^2 and Adjusted R^2 are calculated and applied. Based upon the definitions given above, both statistics are well-defined and applicable: R^2 measures the percent difference between SST and SSE; Adjusted R^2 measures the percent difference between the CER's estimated variance and the sample variance of the dependent variable. Additionally, we do not use R^2 or Adjusted R^2 to indicate the proportion of the variation explained by a MUPE CER.

Just as the "Adjusted R^2 evaluated in fit space," the "Adjusted R^2 in unit space" can go below zero. When this happens, it means that the model developer has generated a CER, which has more variation in SSE than the starting point (i.e., SST) when the drivers are not available. This is certainly a warning flag. We recommend reviewing the residual plot and percentage error table for outliers when it occurs (see Reference 4). Other driver variables or different functional forms should also be explored.

We recommend the following statistics for review when developing MUPE CERs:

Statistics for Review
SPE (Multiplicative Error)
Approximated T-Stats

Adjusted R^2 Pearson's r^2 (i.e., GRSQ) MAD of % Errors RMS of % Errors
--

Note that the first two statistics are fit measures (the approximated t-stats are used to evaluate the significance levels of the regression coefficients); the remainder are predictive measures. Since several fit, as well as predictive measures, are examined when developing CERs, it is not possible to reject a CER just because of its negative Adjusted R^2 . Consequently, this warning flag (to indicate a negative Adjusted R^2) cannot possibly lead to the rejection of a number of good CERs.

CONCERNS ABOUT GRSQ (r^2)

The use of GRSQ has become very popular in recent years. Besides the standard error of the CER, GRSQ is probably the **only** statistic reported for non-OLS CERs in the cost community.

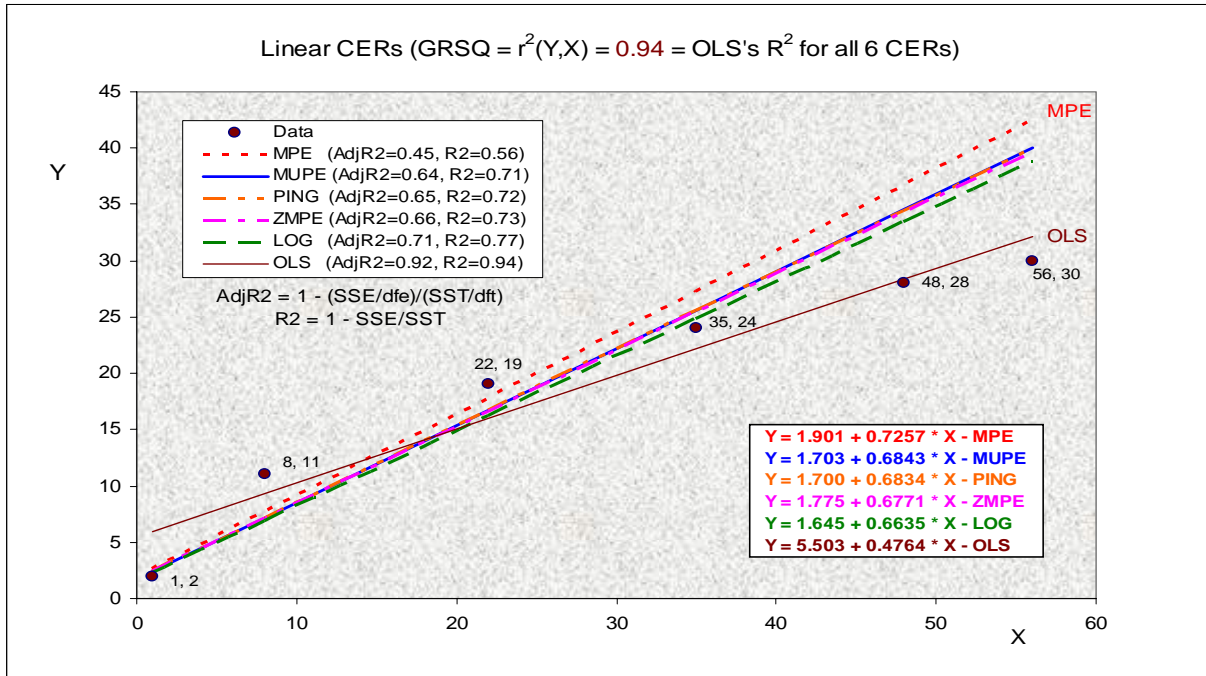
Many Analysts Mistake GRSQ for Traditional R^2 . Many statistical analysts believe GRSQ is a “coefficient of determination” analog for nonlinear CERs, including the MUPE, MPE, ZMPE, and Log-Error CERs. It is stated that “Pearson’s r^2 is the R^2 value that measures how well estimates match the database actuals to which they correspond” (see Reference 7). Therefore, this statistic has been used to measure the proportion of the variation explained by a nonlinear CER. For example, many believe that for a non-linear CER, a 90% GRSQ implies this CER has explained 90% of the variation in the data set. This is certainly an erroneous interpretation. Although GRSQ is equivalent to the traditional R^2 for the OLS CERs, it is not true for other types of CERs. Furthermore, GRSQ does not have good predictive power if the CER is not generated by OLS. This is because GRSQ only measures the linear association between the actual observations and the predicted values, not the difference between them. See the examples in the Analysis Section for the shortcomings of using GRSQ.

Recommend Using the Symbol r^2 (not R^2) for GRSQ to Avoid Possible Confusion. Since Pearson’s correlation coefficient is commonly denoted by the letter r (not R), the symbol r^2 is suggested to represent GRSQ. Detailed analysis of GRSQ using examples is given below.

ANALYZE R^2 /ADJUSTED R^2 AND GRSQ USING EXAMPLES

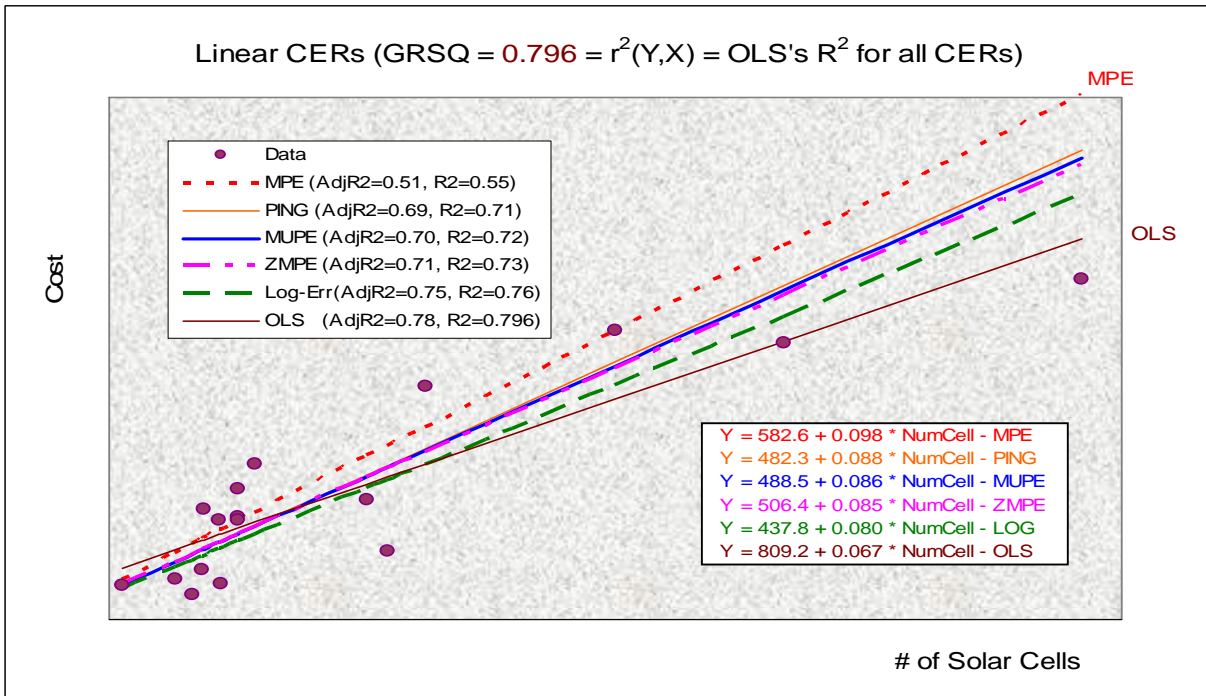
In this section, we will use illustrative examples to analyze the behaviors of R^2 /Adjusted R^2 and GRSQ. These examples are used to demonstrate that GRSQ is not only insensitive to different fitting methods but also insensitive to different equation forms.

GRSQ Insensitive to Different Fitting Methods. The first example (Example 1) is obtained from Reference 1. There are only six data points in this example and the x-y pairs are given by $\{(1, 2), (8, 11), (22, 19), (35, 24), (48, 28), (56, 30)\}$. Simple linear CERs are generated using five different methods: MPE, MUPE, ZMPE, Log-Error, and OLS. As an excursion, we also applied the PING Factor to the log-error CER to adjust the CER result to produce the mean in unit space. (See Reference 3 for detailed descriptions of the PING Factor.) As shown by Graph 1, *Adjusted R^2 in unit space* varies from 0.45 to 0.92 under these five different methods. However, the GRSQ measure is equal to 0.94 for all these equations. Clearly GRSQ does not reflect the deviations between the actual and predicted values.



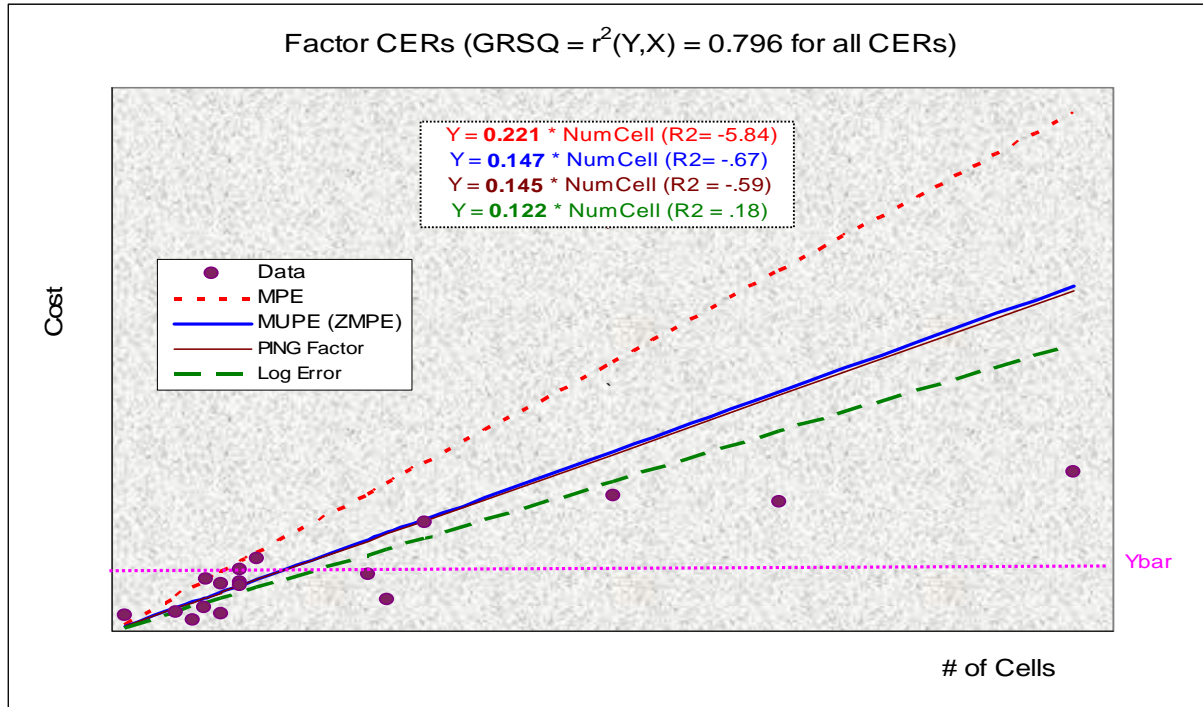
Graph 1: Comparisons of R^2 /Adjusted R^2 and GRSQ for Six Linear CERs

The next example is based on the USCM7 EPS generation data set. Again, we used the same five methods as described above to fit a simple linear CER (i.e., Cost = a + b*Number of Solar Cells) to this data set plus a PING Factor excursion. As shown in Graph 2 below, GRSQ is 0.80 for all six linear equations, while R^2 and Adjusted R^2 in unit space vary from 0.55 to 0.80 and from 0.51 to 0.78, respectively, under the different methods. Clearly GRSQ does not measure how well the estimates track to the actual observations if the CER is not derived by OLS.



Graph 2: Comparisons of R^2 /Adjusted R^2 and GRSQ (r^2) for Six Linear CERs

Similarly, when fitting a simple factor equation to the above data set, the GRSQ measure is again equal to 0.80 for all factor equations regardless of the fitting methods. However, *Adjusted R² in unit space* varies from -5.84 to 0.18 under these different methods. See Graph 3 below for details. (In this example, we only compare the multiplicative methods for the factor equation, so the OLS case is not listed.) Clearly GRSQ does not reflect the deviations between the actual and predicted values because GRSQ remains the same (i.e., 0.8) for all equations graphed in Graph 1 and Graph 2. GRSQ does not have good explanatory power for these linear and factor equations.



Graph 3: Comparisons of R^2 /Adjusted R^2 and GRSQ (r^2) for Five Factor CERs

GRSQ Insensitive to Different Equation Forms. Now let us examine various exponents of the regressed equations. Based upon all five examples given in Reference 1, most of the GRSQ numbers are about the same, regardless of the equation forms. (See Table 1 through Table 5 in Reference 1 for details.) We just list two tables here to review the exponent change in the fitted equations.

Table 3 in “The Trouble with R2” (Sample size = 8)

Parameter or Measure	OLS Linear $y = a + bx$	Multiplicative Error Linear $y = a + bx$	Log-Log Fit $y = ax^b$	Direct Nonlinear Fit $y = ax^b$
a	60.717	-89.829	62.889	48.267
b	47.018	64.842	0.905	1.063
GRSQ	0.615	0.615	0.616	0.614
$R^2 = 1 - SSE/SST$	0.615	0.478	0.603	0.527

In the table above, the exponent ranges from 0.905 to 1.063, while the GRSQ measures remain the same around 0.62.

Table 5 in “The Trouble with R2” (Sample size = 13)

Parameter or Measure	OLS Linear $y = a + bx$	Multiplicative Error Linear $y = a + bx$	Log-Log Fit $y = ax^b$	Direct Nonlinear Fit $y = ax^b$
a	31.408	41.282	35.411	84.896
b	21.454	39.107	0.872	0.517
GRSQ	0.317	0.317	0.334	0.378
$R^2 = 1 - SSE/SST$	0.317	-0.398	0.261	-0.003

In the table above, the exponent is between 0.517 and 1, while the GRSQ measures range from 0.32 to 0.38. In other words, when the exponent is almost doubled (a 100% change), there is only a 16% change in GRSQ. Also, the last CER listed above does not appear to be the best one, but its GRSQ is the highest among the four. From these examples, it is clear that GRSQ is not a sensitive measure with respect to the exponent change of the driver variable and it is not a good predictive measure either.

WHY IS GRSQ INSENSITIVE?

As we know, if a CER is not regressed by OLS, GRSQ does not detect the difference between the actual and predicted values since GRSQ only measures the linear association between them. We will also use an important property of GRSQ for additional analysis.

Pearson’s r and GRSQ Are Invariant under Linear Transformations. By the definition given in Equation 12, we can easily prove that Pearson’s correlation coefficient (Pearson’s r) is invariant under any linear transformation, namely, $r(x,y) = r(a+bx, c+dy)$, where $\{x_i\}$ and $\{y_i\}$ are two sets of numbers of size n and a, b, c, and d are all constant.

$$\begin{aligned}
 r(a + bx, c + dy) &= \frac{\sum_{i=1}^n (a + bx_i - a - b\bar{x})(c + dy_i - c - d\bar{y})}{\sqrt{\sum_{i=1}^n (a + bx_i - a - b\bar{x})^2} \sqrt{\sum_{i=1}^n (c + dy_i - c - d\bar{y})^2}} \\
 &= \frac{bd \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{bd \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = r(x, y)
 \end{aligned} \tag{15}$$

The above equality is the so-called invariance property.

Now let us consider using a triad CER with a multiplicative error term to explain the variation in the dependent variable such as $y = (a + b \cdot x^c) \varepsilon$. Since GRSQ (r^2) is invariant under any linear transformations, we can further interpret GRSQ as follows:

$$\text{GRSQ} = r^2(y, \hat{y}) = r^2(y, \hat{a} + \hat{b} \cdot x^{\hat{c}}) = r^2(y, x^{\hat{c}}) \quad (16)$$

(The hat indicates the predicted value.) Therefore, the fixed-cost term (a), the slope coefficient (b), and the error term (ε) are not used in the computation of GRSQ. If several different methods generate a similar sensitivity parameter, i.e., the exponent, then GRSQ should be similar among these CERs regardless of the following:

- whether or not the error term is additive or multiplicative,
- the size and the sign of the fixed-cost and slope terms, and
- the fitting method applied.

If an analyst further assumes a certain power of the driver variable by engineering logic (or prior information), then inevitably GRSQ would be a constant for all CERs regardless of the methods. This is exactly the reason that GRSQ is a fixed number for all the linear and factor CERs in the examples above. To be more specific, $\text{GRSQ} = r^2(y, \hat{y}) = r^2(y, \hat{a} + \hat{b} \cdot x) = r^2(y, x)$ for simple linear and factor CERs regardless of the methods. For linear CERs with multiple drivers, however, GRSQ may not be uniquely determined by the pairwise correlation coefficients unless all drivers are uncorrelated, i.e.,

$$\text{GRSQ} = r^2(y, x_1) + r^2(y, x_2) + r^2(y, x_3) + \dots \quad \text{if drivers are uncorrelated} \quad (17)$$

Invariance Is Not a Desirable Property. Invariance may seem to be a valuable characteristic at first, but is in fact a detrimental property for GRSQ. As explained above, GRSQ, i.e., $r^2(y, \hat{y})$, remains the same when you multiply, divide, add, and/or subtract your estimate (\hat{y}) by any amount, which is certainly not desirable. (The invariance property partly explains why GRSQ is insensitive to different fitting methods and CER forms.) Note that neither R^2 nor Adjusted R^2 is invariant under linear transformations whether or not the CER is linear or non-linear.

As for why GRSQ is not sensitive to the exponent change in the CER, we believe the reason is that GRSQ only measures the linear association between the observed and predicted values instead of the actual deviations between them.

MODIFY ADJUSTED R^2 FOR MUPE AND GRSQ FOR DF

Modify the Adjusted R^2 for the MUPE CER. Both *Adjusted R^2 in unit space* and GRSQ are predictive measures, which are used to evaluate the predictive capability of the regression equation. Although *Adjusted R^2 in unit space* measures the percent difference between the CER's estimated variance and the sample variance of Y (i.e., cost), the evaluation is done in unit space for CERs modeled by additive errors, namely

$$Y_i = f(x_i, \beta) + \varepsilon_i \quad \text{for } i = 1, \dots, n \quad (18)$$

To be consistent with the underlying hypothesis and the fitting methodology (see Equations 3 and 5), we recommend modifying *Adjusted R^2* for the MUPE CERs:

$$Adj. R^2 \text{ for MUPE} = 1 - \frac{\sum((y_i - \hat{y}_i)/\hat{y}_i)^2/(n-p)}{\sum((y_i - \bar{y})/\bar{y})^2/(n-1)} = \frac{SPE_{\bar{y}}^2 - SPE_f^2}{SPE_{\bar{y}}^2} \quad (19)$$

This modified Adjusted R^2 compares the estimated variances between MUPE and its baseline (i.e., an average CER when driver variables are not available). This comparison is certainly more pertinent to the fitting methodology (Equation 5). For example, if the SPE for a MUPE CER is 0.4, we don't really know whether this number is good or bad (or how good it is). However, if *Adjusted R^2 for MUPE* is calculated to be 0.75, then we know the reduction of variance is 75% when applying this CER. In short, *Adjusted R^2 for MUPE* is a relative measure, which puts MUPE's SPE^2 in perspective.

Note that this modified measure is a hybrid of both fit and predictive measures—it can be used to compare different MUPE CERs but not to compare different methods, e.g., a MUPE CER versus a ZMPE CER. On the other hand, *Adjusted R^2 in unit space* can be compared across different equations and/or fitted by different methods.

Modify GRSQ r^2 for Correcting Degrees of Freedom. As mentioned above, GRSQ is a predictive measure. It measures the linear association between two sets of numbers $\{y_i\}$ and $\{\hat{y}_i\}$ rather than the actual deviations between them. Furthermore, this measure does not take the degrees of freedom or even the sample size into consideration. For example, a GRSQ of 0.75 derived from 30 observations should be more significant than the same GRSQ simply based upon five data points. Unfortunately, neither the sample size nor the degrees of freedom adjustment is provided in the GRSQ formula. We now propose a modified GRSQ to correct for degrees of freedom (DF):

$$GRSQ (DF) = \begin{cases} r^2 - (1 - r^2) * \frac{p-1}{n-p} & \text{if } p > 1 \\ r^2 - (1 - r^2) * \frac{1}{n-1} & \text{if } p = 1 \end{cases} \quad (20)$$

where p = number of estimated coefficients, n = sample size, and r^2 = GRSQ.

When dealing with a simple factor equation, e.g., $p = 1$, the term “ $(p-1)/(n-p)$ ” vanishes (see Equation 20). However, the sample size should still be accounted for when it happens, so a modified term $1/(n-1)$ is used in this case.

CONCLUSIONS

Use GRSQ (r^2) with Caution if the Equation Is Not Regressed by OLS. Although the use of GRSQ has become very popular for nonlinear CERs, use this measure with caution. Based upon the above analysis results for non-OLS equations, it is clear that

1. GRSQ is insensitive to different fitting methods (whether the error is additive or not);
2. GRSQ is insensitive to CER forms, e.g., the exponent change in the driver variables;
3. GRSQ does not have good explanatory power because it measures the linear association between the actual and predicted values rather than the difference between them;
4. GRSQ (a predictive measure) cannot tell if the regressed coefficients are significant;

5. GRSQ cannot detect flaws in the model; and
6. GRSQ should not be used to measure the proportion of the variation explained by non-OLS CERs

Table 1 compares the GRSQ and R^2 /Adjusted R^2 measures for the analyzed examples above.

Table 1: Comparisons of GRSQ (r^2) and R^2 /Adjusted R^2 in Unit Space

Example / Equation	Method	GRSQ (r^2)	R^2	Adj. R^2
#1: $Y = 1.901 + 0.7257 * X$	MPE	0.938	0.557	0.446
#1: $Y = 1.703 + 0.6843 * X$	MUPE	0.938	0.714	0.643
#1: $Y = 1.700 + 0.6834 * X$	PING Factor	0.938	0.717	0.646
#1: $Y = 1.775 + 0.6771 * X$	ZMPE	0.938	0.732	0.665
#1: $Y = 1.645 + 0.6635 * X$	Log Error	0.938	0.772	0.715
#1: $Y = 5.503 + 0.4764 * X$	OLS	0.938	0.938	0.922
#2: $Y = 582.6 + 0.098 * X$	MPE	0.796	0.545	0.515
#2: $Y = 482.3 + 0.088 * X$	PING Factor	0.796	0.706	0.687
#2: $Y = 488.5 + 0.086 * X$	MUPE	0.796	0.721	0.702
#2: $Y = 506.4 + 0.085 * X$	ZMPE	0.796	0.731	0.714
#2: $Y = 437.8 + 0.080 * X$	Log Error	0.796	0.761	0.745
#2: $Y = 809.2 + 0.067 * X$	OLS	0.796	0.796	0.782
#3: $Y = 0.221 * X$	MPE	0.796	-5.84	-5.84
#3: $Y = 0.147 * X$ (SPE=73%)	MUPE	0.796	-0.67	-0.67
#3: $Y = 0.147 * X$ (SPE=73%)	ZMPE	0.796	-0.67	-0.67
#3: $Y = 0.145 * X$	PING Factor	0.796	-0.59	-0.59
#3: $Y = 0.122 * X$	Log Error	0.796	0.18	0.18
#3: $Y = 0.086 * X$	Additive Error	0.796	0.67	0.67
#4: $Y = 60.717 + 47.018 * X$	OLS	0.615	0.615	0.551
#4: $Y = -89.829 + 64.842 * X$	MPE	0.615	0.478	0.391
#4: $Y = 62.889 * X^{0.905}$	Log Error	0.616	0.603	0.537
#4: $Y = 48.267 * X^{1.063}$	MPE	0.614	0.527	0.448
#5: $Y = 31.408 + 21.454 * X$	OLS	0.317	0.317	0.255
#5: $Y = 41.282 + 39.107 * X$	MPE	0.317	-0.398	-0.525
#5: $Y = 35.411 * X^{0.872}$	Log Error	0.334	0.261	0.194
#5: $Y = 84.896 * X^{0.517}$	MPE	0.378	-0.003	-0.094

Besides the theoretical proof, Table 1 also shows that for simple linear/factor equations, GRSQ is the same for all regression methods (whether or not the error is additive or multiplicative), while R^2 and Adjusted R^2 vary under different methods. Clearly GRSQ does not reflect the actual deviations between the observed and predicted values. Table 1 also indicates that GRSQ is in fact the traditional R^2 for the OLS equation even though the error term is assumed to be multiplicative. (See the numbers in bright blue in the R^2 column.) This raises two issues: (1) GRSQ is totally insensitive to and does not have good explanatory power for these CERs at all and (2) Why should we use an OLS metric to describe the quality of multiplicative error models? The use of the GRSQ measure in these examples is misleading and incorrect. The invariance property partly explains why GRSQ is insensitive to different regression methods and CER forms.

Cautionary notes on R² and Adjusted R² in unit space

The R² and adjusted R² in unit space are used to evaluate the model’s actual predictive power. While R² is a very common and popular goodness-of-fit measure, it can be misleading for small samples because it does not take the degrees of freedom (DF) into account. (Adjusted R² is a better statistic than R² because of the DF adjustment.) Since both measures (as well as GRSQ) can be highly influenced by extreme values, sole reliance on R² or Adjusted R² should be avoided and additional data analysis and residual/error examination are strongly recommended.

Recommendations:

Use “Adjusted R² for MUPE” To Evaluate MUPE CERs. We recommend using “Adjusted R² for MUPE” to evaluate MUPE CERs because (1) it is more relevant to the fitting method and (2) it puts SPE in perspective.

$$Adj. R^2 \text{ for MUPE} = 1 - \frac{\sum((y_i - \hat{y}_i) / \hat{y}_i)^2 / (n - p)}{\sum((y_i - \bar{y}) / \bar{y})^2 / (n - 1)} = \frac{SPE_{\bar{y}}^2 - SPE_f^2}{SPE_{\bar{y}}^2}$$

Modify GRSQ to Correct for Degrees of Freedom. The GRSQ does not take either the degrees of freedom or the sample size into account. We recommend using the modified GRSQ to reflect the actual degrees of freedom in the data set:

$$GRSQ_{DF} = \begin{cases} r^2 - (1 - r^2) * \frac{p - 1}{n - p} & \text{if } p > 1 \\ r^2 - (1 - r^2) * \frac{1}{n - 1} & \text{if } p = 1 \end{cases}$$

where r² = GRSQ, p = number of estimated coefficients, and n = sample size.

This modified GRSQ can be used to update the correlation analyses for the USCM CERs and the database.

Do Not Rely on a Single Measure for Selecting the Best CER. Beware of using the SPE (or GRSQ) measure alone for selecting CERs. Neither SPE nor GRSQ can be used to determine whether the regression coefficients are significant; they cannot detect model flaws either. Use the fit measures to judge the significance levels of the coefficients (see Reference 2 for details). GRSQ does not have the same statistical meaning and value of the traditional R² except for OLS. “Adjusted R² for MUPE” is a good complement to GRSQ; use them together.

Knowing Your Data Set Is More Important than Comparing Statistical Measures. We recommend using the fit measures to judge the significance of the model (i.e., the quality of the fit) and using the predictive measures (such as GRSQ and Adjusted R²) for further analysis of a potential CER. The combined evaluation between the fit and predictive measures will help select the best equation. Additionally, always review the data set, the residual plot, and percentage error table for outliers, instead of just checking one or two statistical measures.

REFERENCES

1. Book, S. A. and P. H. Young, "The Trouble with R2," International Society of Parametric Analysts, Journal of Parametrics, Vol. XXV, No 1 (Summer 2006), pages 87-112
2. 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
3. Hu, S., "The Impact of Using Log-Error CERs Outside the Data Range and PING Factor," 5th Joint Annual ISPA/SCEA Conference, Broomfield, CO, 14-17 June 2005
4. Nguyen, P., N. Lozzi, et al., "Unmanned Space Vehicle Cost Model, Eighth Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, October 2001
5. Hu, S., "The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development," 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001
6. Book, S. A. and N. Y. Lao, "Minimum-Percentage-Error Regression under Zero-Bias Constraints," Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 Oct 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56
7. Nguyen, P., N. Lozzi, et al., "Unmanned Spacecraft Cost Model, Seventh Edition," U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, August 1994
8. Hu, S. and A. R. Sjovold, "Multiplicative Error Regression Techniques," 62nd MORS Symposium, Colorado Springs, Colorado, 7-9 June 1994
9. Young, P. H., "Generalized Coefficient of Determination," 26th Annual DoD Cost Analysis Symposium, Leesburg, VA, September 1992
10. Young, P. H., "GERM: Generalized Error Regression Model," 25th Annual DoD Cost Analysis Symposium, Leesburg, VA, September 1991
11. Hu, S. and A. R. Sjovold, "Error Corrections for Unbiased Log-Linear Least Square Estimates," TR-006/2, March 1989
12. Seber, G. A. F., and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons, 1989, pages 37, 46, 86-88
13. Weisberg, S., Applied Linear Regression, 2nd Edition," New York: John Wiley & Sons, 1985, pages 87-88
14. Goldberger, A. S., "The Interpretation and Estimation of Cobb-Douglas Functions," Econometrica, Vol. 35, July-Oct 1968, pp. 464-472