Empirical Software Engineering, 8, 139–161, 2003.

A Further Empirical Investigation of the Relationship Between MRE and Project Size

ERIK STENSRUD erik.stensrud@bi.no
Norwegian School of Management

TRON FOSS tron.foss@bi.no
Norwegian School of Management

BARBARA KITCHENHAM b.a.kitchenham@cs.keele.ac.uk
Keele University

INGUNN MYRTVEIT ingunn.myrtveit@bi.no
Norwegian School of Management

Editor: Lionel Briand

Abstract. The mean magnitude of relative error, MMRE, is the *de facto* standard evaluation criterion to assess the accuracy of software project prediction models. The fundamental metric of MMRE is MRE, a relative residual error. For MMRE to be a meaningful summary statistic, it is a necessary, but not sufficient, condition that MRE and project size are uncorrelated. Except for two previous conference studies done by the same authors, it has never been empirically validated that MRE and project size really are uncorrelated. In this paper, we extend the previous studies using the same data sets as before: Albrecht, Kemerer, Finnish, DMR and Accenture-ERP. Unlike the previous studies, we plot MRE against the predicted effort rather than against the actual effort and, in so doing, we obtain very different results from the previous studies. The results of this study suggest that MRE and project size are uncorrelated, which apparently is contradictory to the previous results where we found a negative correlation. The explanation for these seemingly contradictory results is presented in this study.

Keywords: Software engineering, project cost estimation, evaluation metrics, magnitude of relative error, empirical validation

1. Introduction

The mean magnitude of relative error, MMRE, is the *de facto* standard evaluation criterion to assess the accuracy of software prediction models (Briand and Wieczorek, 2001). MMRE is a summary statistic, i.e., a single number, aggregating the fundamental metric MRE, a relative residual error. For MMRE to be a meaningful summary statistic for software project prediction models, it is a necessary, but not sufficient, condition that MRE and project size are uncorrelated. Were they negatively or positively correlated, MMRE would be a poor measure of the MRE to expect for a large or a small project in the data set. Rather, MMRE would only be representative of the average-size projects in the data set.

Furthermore, were MRE and project size negatively correlated, and at the same time, the residual and project size uncorrelated, the rationale for preferring a relative residual to the (non-relative) residual would be non-existent.

MMRE is used for two kinds of assessments (at least). One purpose of MMRE is to select between competing prediction models: The model that obtains the lowest MMRE is preferred. Another purpose is to provide a quantitative measure of the uncertainty of a prediction. (Where a low MMRE is taken to mean low uncertainty or inaccuracy.) In this study, it is the latter use of MMRE that is of interest because if MRE and project size are negatively or positively correlated, MMRE taken over all projects will be a rather imprecise indication of the MRE to expect for, say, a large project. (Where "large" and "small" refer to relative sizes within a data set and not to any absolute sizes.) The magnitude of relative error, MRE, is defined as follows (Conte et al., 1986):

$$\text{MRE} = \frac{|y - \hat{y}|}{y}, \; y = \text{actual and } \hat{y} = \text{prediction}$$

The reasons for the popularity of MRE is that it is independent of units and that it apparently can be used to evaluate all types of prediction systems. It is believed that it can be applied equally well to assess regression models as well as AI-inspired machine learning type models such as case-based reasoning (CBR), estimation by analogy (EBA), classification and regression trees (CART) and artificial neural net (ANN) prediction models. (For an overview of the various models see, for example, Briand and Wieczorek, 2001.)

Also, the concept of relative error makes intuitive sense to software researchers and practitioners alike. At a first glance, it may seem reasonable that we should be able to predict the effort of small and large projects with the same relative error rather than the same absolute error. For example, it seems reasonable that if we are able to predict a small project, say, $x = 10$ personmonths (PM) within a 1 PM error (10%), we should also be able to predict a large project, say, $x = 100$ PM within a 10 PM error (also 10%). It seems less reasonable that we also ought to predict the large project within the same absolute error as the small project, i.e., a 1 PM error for both. (Of course, it is not always the case that a customer would accept the same relative error for small and large projects alike.) In other words, we implicitly assume that MRE and project size are uncorrelated, and that project data sets are heteroscedastic, exhibiting an increasing variance. (For an introduction to basic statistical terms, e.g. heteroscedasticity, see, for example, Gujarati, 1995.)

Depending on the characteristics of software project data sets with respect to homoscedasticity (or the opposite: heteroscedasticity), there are in theory three possible outcomes regarding the correlation between MRE and project size: no correlation, negative correlation, or positive correlation. In this study, we investigate whether or not MRE and project size are uncorrelated. We use the predicted effort, $\hat{y}$, as the project size measure. An empirical study is justified because there is no theory or established truth regarding the relationship between MRE and project size i.e., whether they are uncorrelated, negatively correlated, or positively correlated.

First, it is not obvious that MRE and $\hat{y}$ are uncorrelated. On the contrary, from Rosenberg's (1997) study, one may be misled to believe that MRE and $\hat{y}$ must always be negatively correlated since MRE includes a $1/y$ term (which is close to a $1/\hat{y}$ term). Rosenberg conjectures that "since there must be a negative correlation between values of X and 1/X, it follows that the correlation between X (e.g. SLOC) and Y/X (e.g. defects/SLOC) must also be negative". It is true that X and 1/X are negatively correlated, but it is not generally true that X and Y/X are always negatively correlated. Rosenberg's conjecture holds in the special case when X and Y are uncorrelated, but is does not hold for *all* cases when X and Y are correlated. In fact, Rosenberg adds as a reservation to his general conjecture: "if Y is non-linearly related to X, then this will be true whenever the exponent is not greater than unity (i.e., whenever defects are *growing at most linearly* in the number of lines of code)." (Our italics.) We have no good theoretical reasons to dismiss the possibility that the nominator in MRE, $(y - \hat{y})$, "grows more than linearly". In other words, that the heteroscedasticity grows exponentially. Therefore, contrary to Rosenberg's conjecture, it is no more obvious that MRE and $\hat{y}$ are (negatively) correlated either. Had it been obvious, we believe a relative residual like MRE would never have been preferred to the (non-relative) residual.

As we have already stated, we may distinguish between (at least) three types of correlations between MRE and $\hat{y}$:

1. Negative correlation.

2. No correlation.

3. Positive correlation.

Case 2 corresponds to a data set with the variance increasing in a fashion similar to pattern B in Figure 1 whereas case 3 corresponds to a data set with variance similar to pattern A.

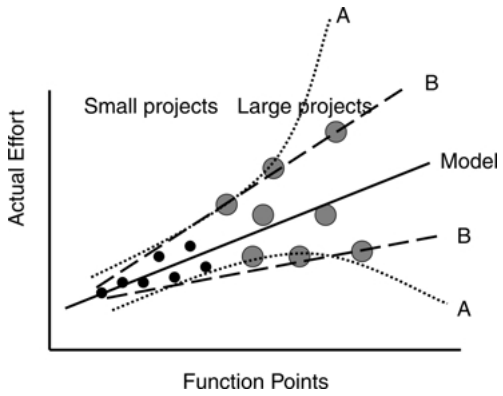

Figure 1. Heteroscedastic data with two different patterns of increasing variance (A) and (B), respectively.

Negative Correlation Let us assume a model for the effort y and a prediction method for $\hat{y}$ such that the prediction error $y - \hat{y}$ for a given value $\hat{y} = \hat{y}_0$ has a distribution that is independent of $\hat{y}_0$. Then $y - \hat{y} \mid \hat{y} = \hat{y}_0$ is a stochastic variable ε, and we therefore get

$$y = \hat{y}_0 + \varepsilon$$

In this case,

$$\text{MRE} \mid \hat{y} = \hat{y}_0 = \frac{|\hat{y} - \hat{y}_0|}{y} = \frac{|\varepsilon|}{\hat{y}_0 + \varepsilon}$$

We observe that $\text{MRE} \mid \hat{y} = \hat{y}_0$ is a decreasing function of $\hat{y}_0$ (negative correlation) for any value of ε. From this, it follows that $E(\text{MRE} \mid \hat{y} = \hat{y}_0)$ is also a decreasing function of $\hat{y}_0$. In other words, if the distribution of ε is independent of $\hat{y}$, then the expected value of MRE will also decrease with $\hat{y}$.

Let us next assume a model for the effort y and a prediction method for $\hat{y}$ such that the prediction error $y - \hat{y}$ for a given value $\hat{y} = \hat{y}_0$ has a distribution of the form $\hat{y}_0^k \varepsilon$ where ε is a stochastic variable that does not take $\hat{y}_0$ as a parameter. Here, $k \geq 0$ is a constant. That is, $y - \hat{y} \mid \hat{y} = \hat{y}_0$ is a stochastic variable $\hat{y}_0^k \varepsilon$, and we therefore get

$$y = \hat{y}_0 + \hat{y}_0^k \varepsilon$$

In this case,

$$\text{MRE} \mid \hat{y} = \hat{y}_0 = \frac{|\hat{y} - \hat{y}_0|}{y} = \frac{|\hat{y}_0^k \varepsilon|}{\hat{y}_0 + \hat{y}_0^k \varepsilon} = \frac{\hat{y}_0^k |\varepsilon|}{\hat{y}_0 (1 + \hat{y}_0^{k-1} \varepsilon)} = \frac{\hat{y}_0^{k-1} |\varepsilon|}{1 + \hat{y}_0^{k-1} \varepsilon} = \frac{|\varepsilon|}{\hat{y}_0^{1-k} + \varepsilon}$$

We observe that $\text{MRE} \mid \hat{y} = \hat{y}_0$ is a decreasing function of $\hat{y}_0$ (negative correlation) for $0 \leq k < 1$, and for any value of ε. From this, it follows that $E(\text{MRE} \mid \hat{y} = \hat{y}_0)$ is also a decreasing function of $\hat{y}_0$. ($k = 0$ covers the first case, i.e., where $\hat{y}_0^k \varepsilon = \varepsilon$.)

No Correlation We also observe that $\text{MRE} \mid \hat{y} = \hat{y}_0$ is neither a decreasing nor an increasing function of $\hat{y}_0$ for $k = 1$, and for any value of ε. From this, it follows that $E(\text{MRE} \mid \hat{y} = \hat{y}_0)$ is independent of $\hat{y}_0$ (uncorrelated).

Positive Correlation Finally, we observe that $\text{MRE} \mid \hat{y} = \hat{y}_0$ is an increasing function of $\hat{y}_0$ (positive correlation) for $1 < k$, and for any value of ε. From this, it follows that $E(\text{MRE} \mid \hat{y} = \hat{y}_0)$ is also an increasing function of $\hat{y}_0$.

To our knowledge, the relationship between MRE and project size has never been empirically investigated except for two previous conference studies done by the same authors where the results suggested that MRE and project size are negatively correlated (Stensrud et al., 2002; Foss et al., 2001a). Specifically, it has never been empirically validated that MRE is uncorrelated with project size, which is a

necessary, but not sufficient, condition to defend the use of MMRE as an evaluation metric.

In the two previous studies, MRE was plotted against actual effort, y. The results suggested that MRE was negatively correlated with y. In this study, we plot MRE against the predicted effort, $\hat{y}$, and obtain different results, namely that MRE and $\hat{y}$ seem uncorrelated. We discuss why it is more appropriate to plot MRE against $\hat{y}$ (predicted effort) rather than against y (actual effort) in the Discussion section, and thus, we argue that the results of this study are of more use and interest than the results of the two previous studies. We redo the latter of the two studies (Stensrud et al., 2002) using the same five data sets: Albrecht, Kemerer, Finnish, DMR and Accenture-ERP. The first four data sets presumably contain custom software development (CSD) projects whereas the last contains ERP projects.

The paper is organized as follows. Section 2 presents related work. Section 3 presents the data sets used in the study. Section 4 describes the validation methodology. Section 5 presents the results of the regression analysis. The regression models are required to obtain residuals $(y - \hat{y})$ and, consequently, to derive MRE values. Section 6 presents the main results on MRE. In Section 7, we discuss the use of y versus $\hat{y}$ as the project size measure. In Section 8, we briefly list other evaluation criteria that could be objects of future research, and we conclude in Section 9.

2. Related Work

Evaluation of the evaluation metrics themselves seems to have received little attention since Conte et al. (1986) publicized MMRE. The only related work we are aware of is Miyazaki et al. (1994), Kitchenham et al. (2001), and Foss et al. (2002). All of them discuss issues related to the use of MMRE, but none of them have investigated the implicit assumption that MRE and $\hat{y}$ are uncorrelated.

Foss et al. (2002) examine whether MMRE is a reliable measure when used to select between two competing linear prediction models. It documents that MMRE is unreliable for this purpose. The reason is there is a high probability of selecting the model with the worst fit to the data. In particular, a model which underfits (produces too low estimates) will likely obtain a lower MMRE (depending on the degree of underfit) than a model with a perfect fit (the true model).

Conte et al. (1986) did not justify the use of a relative error measure. They seemed more concerned about arguing for the need of taking the absolute value (or magnitude) of the relative error to prevent that negative and positive values cancel each other out (in this, they were right). Neither did they justify using a relative error that measures the error relative to the actual rather than relative to the estimate. This is surprising. From a practitioner's view, it would make more sense to measure the error relative to the estimate because the estimate, not the actual, comes first in time. It is the estimate, not the actual, that is known to the client and project team throughout the whole project life-cycle. The actual is not known until project completion. Therefore, the client always judges how well the project (and project

manager) did comparing the result to the estimate. That is, (s)he measures the prediction accuracy relative to the estimate. (This is an argument for plotting MRE against $\hat{y}$.)

Miyazaki et al. (1994) rightly observed that if minimum-MMRE is used as the fitting (calibration) method to fit a function to a data set, the prediction model will have a bad fit to the data producing biased, too low, estimates (''underestimate''). Also, they observed that there is a discrepancy between the evaluation criterion (e.g. MMRE) and the fitting method when using, say, ordinary least squares (OLS) regression. Their solution to the problem was to throw out OLS. They invented their own fitting methods, the least squares of balanced relative errors (LBRS) and the least squares of inverted balanced relative errors (LIRS), and they invented new evaluation criteria consistent with these fitting methods. Still, we believe they committed a major mistake discarding the OLS method. OLS has several desirable statistical properties: OLS estimators are BLUE—best linear, unbiased estimator (Gujarati, 1995). For example, the point estimate of an OLS regression model is the mean. The mean is a well-defined statistic. As opposed to this, Miyazaki et al. (1994) failed to prove any desirable statistical properties for their fitting methods, LBRS and LIRS.

3. Data

All the data sets used in this study have been thoroughly presented in other published studies. Therefore, they are only briefly presented here. References to the other studies are provided for those readers interested in more details on the data. The reason we have chosen these data sets are as follows. They are easily available. They have been used in many other studies in software engineering. Most of them ought to be familiar to the reader. Last, but not least, we believe they are representative of software engineering data sets.

3.1. Accenture-ERP Data

This Accenture-ERP data set consists of 81 ERP projects carried out in a multinational consultant organization in Accenture (formerly Andersen Consulting) from 1990 to 1998. The data include four predictor variables (users, sites, interfaces, modules) and one response variable (effort). The projects span from 320 to 111,420 workdays (Table 1).

The variables are defined in Stensrud (1998). An evaluation of the data quality is found in Myrtveit et al. (2001) (or alternatively, Myrtveit and Stensrud, 1999; Stensrud and Myrtveit, 1998).

Table 1. Descriptive statistics for Accenture-ERP data set.

Variable	N	Mean	Median	StDev	Min	Max
Users	81	999	300	2430	6	17,000
Sites	81	29	5	75	1	500
Interfaces	81	29	15	50	0	270
Modules	81	5	5	2.7	1	13
Effort	81	11,526	7300	15,170	320	111,420

3.2. Data Sets of Custom Software Projects

The Albrecht data set consists of 24 projects carried out in IBM in the 1970s. The data are described in Albrecht and Gaffney (1983). The projects span from 0.5 to 205 personmonths (Table 2).

The Kemerer data set consists of 15 projects. The data are described in Kemerer (1987). The 15 projects span from 23 to 1107 workdays (Table 3).

The Finnish data set consists of 40 projects (Table 4). Two projects have missing data. The data comes from different companies, and the data collection was performed by a single person. The projects span from 460 to 23,000 workhours. More details on the data may be found in Kitchenham and Kansala (1993).

The DMR dataset contains 81 projects (Table 5). For more details see Desharnais (1989). The data comes from a single company but have used three different language types. We have used observations with language types 1 and 2, thus

Table 2. Descriptive statistics for Albrecht data set.

Variable	N	Mean	Median	StDev	Min	Max
FP	24	648	506	488	199	1902
Effort	24	22	11	28	0.5	205

Table 3. Descriptive statistics for Kemerer data set.

Variable	N	Mean	Median	StDev	Min	Max
FP	15	999	993	590	100	2307
Effort	15	219	130	263	23	1107

Table 4. Descriptive statistics for Finnish data set.

Variable	N	Mean	Median	StDev	Min	Max
FP	40	761	638	511	65	1814
Effort	38	7573	5430	6872	460	23,000

Table 5. Descriptive statistics for DMR data set.

Variable	N	Mean	Median	StDev	Min	Max
FP Adj	81	289	255	186	62	1116
Effort	81	5046	3647	4419	546	23,940

discarding observations with language type 3. The projects span from 1116 to 23,940 workhours.

4. Validation Methodology

4.1. Missing Data Procedure

We applied the same missing data procedure to all data sets with missing values. Data sets with missing values include the Accenture_ERP and the Finnish data sets. The Kemerer, DMR and Albrecht data did not have any missing values in the variables we considered.

Projects with missing values were removed using listwise deletion (LD). LD is appropriate provided that the missing data are missing completely at random (MCAR). If the missing data are MCAR, the application of LD will not introduce any bias in the data or the results (Little and Rubin, 1987). In the Accenture-ERP data set, the missing data are likely to be missing at random. (The data set includes active as well as completed projects. Active projects naturally do not have actuals for all variables as some of them are not known until project completion. Most of the missing values are in active projects for the variables Interfaces, Conversions, Modifications and Reports. Cf. Myrtveit et al., 2001, for more details.) Regarding the Finnish data set, only two observations are missing. Therefore, we do not commit a serious error by removing them even if they are not MCAR.

4.2. Model Specifications

We applied two variants of linear regression analysis to fit a function to the observations. The assumption of linearity in the parameters is a good first order approximation. Also, it is feasible to obtain linear relationships between the predictor and response variables by suitable transformations of the scales, e.g. log-linear transforms.

The two regression models that were investigated were ordinary least squares (OLS) and least absolute deviation (LAD). OLS was used as the default method whereas LAD was investigated in addition to OLS for data sets where we suspected heavy tails. LAD is efficient when the distribution of residuals exhibit heavy tails, i.e., the data exhibit high kurtosis (Foss et al., 2001b) whereas OLS is efficient when

the residual is normal. In general, it is prudent to apply more than one method to confirm that the initial model specification and results are reasonable.

In addition to investigating both OLS and LAD, we also examined the data to decide whether a linear or a log-linear model was the most appropriate. A log-linear model (actually, a multiplicative model) forces the line through the origin whereas a linear model does not. Sometimes, it is desirable that the intercept is zero. In our case, we want to plot MRE against the prediction, $\hat{y}$. It is therefore desirable that we avoid negative intercepts to avoid negative predictions. Thus, the log-linear model is our preferred choice.

4.3. MRE Calculations

MRE is defined as (Conte et al., 1986):

$$\text{MRE} = \frac{|y - \hat{y}|}{y}$$

In the cases where we applied a log-linear regression model, we used the following formula to calculate MRE.

$$\text{MRE} = \left|1 - \mathrm{e}^{-residual}\right|$$

(The derivation of this formula is provided in the Appendix.)

4.4. Research Hypothesis and Test Procedure

In a previous study, we investigated whether or not MRE and actual effort, y, are uncorrelated. In this study, we investigate if MRE and predicted effort, $\hat{y}$, are uncorrelated. The rationale for using $\hat{y}$ is provided in the Discussion section. The research hypothesis may be formally stated as follows.

- H0: MRE is uncorrelated with $\hat{y}$.
- H1: MRE is correlated with $\hat{y}$, positively or negatively.

To test the hypothesis, we performed OLS regression analysis with $\hat{y}$ as the independent and MRE as the dependent variable. It should be observed that this procedure is not completely unproblematic: It is difficult to transform the scales of MRE and $\hat{y}$ to comply with the assumptions of regression analysis (in particular linearity and homoscedasticity).

An alternative test procedure, the one that we used in our previous studies, is, unfortunately, not any more attractive. In our previous studies we divided the data set into two subsamples, calculated MMRE for each subsample and used a two-tailed 2-sample t-test (Minitab, www) to test if the two MRE's were significantly

different. The 2-sample t-test accounts for unequal variance and unequal size in the two samples. Still, this test procedure is problematic for two reasons. First, the test requires normal distributions. MRE is not normal. For non-normal distributions, we thus require large sample size in order to call on the central limit theorem. The central limit theorem states that the sample mean will be approximately normal even if the population is not normal, provided n is sufficiently large. As a rule of thumb, "n = large" means $n > 30$. Unfortunately, we do not have sufficiently large sample sizes for three of the data sets (Albrecht, Kemerer, and Finnish).

Since the OLS regression is also problematic as a test procedure, we have added another test, the Park test of heteroscedasticity (Gujarati, 2003, p. 403). The rationale for applying this test is based on the following observation. Testing if MRE and $\hat{y}$ are uncorrelated, i.e., that MRE is constant, is equal to testing if the residual of the log-linear model is homoscedastic, cf. Section 4.3. Therefore, we apply the Park test to the residuals of the log-linear models in Section 5. If the results of the Park test are consistent with the results of the OLS regression test (i.e., MRE vs. $\hat{y}$), we have more evidence on which to draw conclusions. The Park test is a regression analysis with the squared residual as the dependent variable and either $\hat{y}$ or x as the independent variable. (We have used x for the univariate data sets and $\hat{y}$ for the multivariate data set.) If the coefficient of the slope is statistically significant, i.e., $T > 2$, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant ($T < 2$), we may accept the assumption of homoscedasticity, and thus, that MRE and $\hat{y}$ are uncorrelated.

5. Results—Regression Analysis

5.1. Accenture-ERP Regression Model

The data set exhibited a pronounced heteroscedasticity. We therefore applied a log-linear model to this data set. The data also seemed messy (seemingly high kurtosis) by initial visual inspection. Therefore, it seemed natural to investigate OLS as well as LAD. In a previous study, we have investigated both LAD and OLS on this data set concluding that OLS was at least as efficient as LAD (Foss et al., 2001b). We therefore use the OLS model in this study (cf. Table 6). We observe that all coefficients are significant ($T < 2$), and that all are non-negative, as we would expect. Also, the explanatory power is high (R-sq = 73.5%). All in all, it seems like a plausible model.

Residual analysis further suggested that the distribution of residuals of the log-linear model was normal. (The Anderson-Darling test of normality: $p = 0.493$.) The residual analysis did not reveal any non-linearity either (visual inspection of Figure 2). There is no significant multicollinearity ($VIF < 2$).

Table 6. Accenture-ERP OLS log-linear regression model.

Predictor	Coef	SE Coef	T	P
Constant	4.8504	0.2972	16.32	0.000
ln(Users)	0.2745	0.0631	4.35	0.000
ln(Sites)	0.1530	0.0552	2.77	0.007
ln(Intfaces)	0.2889	0.0585	4.94	0.000
ln(Module)	0.7316	0.1639	4.46	0.000
S	R-Sq	R-Sq(adj)		
0.5806	73.5%	72.1%		

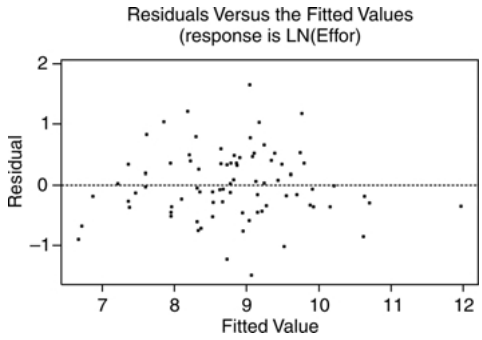

Figure 2. Accenture-ERP, residuals vs. fits.

5.2. Albrecht Regression Model

This data set did not exhibit heteroscedasticity nor high kurtosis. Still, we applied an OLS log-linear model to this data set to force the intercept through the origin, and thus, to avoid having to plot MRE against negative values of predicted effort. (We obtained some negative fits when applying a linear, additive model. The linear model had a negative intercept. This is not reported.) LAD models were not investigated. From Table 7, we observe that the FP coefficient is significant ($T > 2$). Also, the explanatory power is high (R-sq = 73.5%).

Table 7. Albrecht OLS log-linear regression model.

Predictor	Coef	SE Coef	T	P
Constant	− 6.813	1.196	− 5.70	0.000
Ln(FP)	1.4873	0.1905	7.81	0.000
S	R-Sq	R-Sq(adj)		
0.6148	73.5%	72.3%		

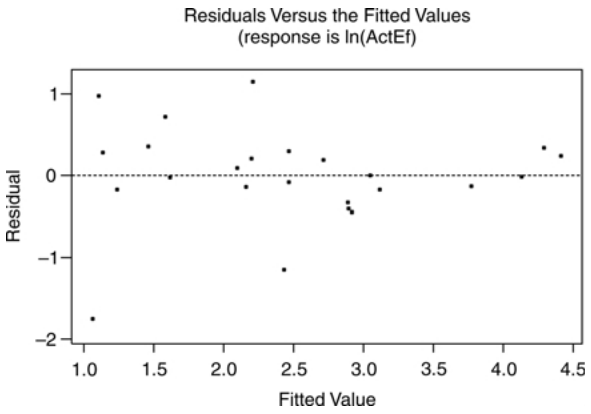

Figure 3. Albrecht, residuals vs. fits.

Residual analysis suggested that the distribution of residuals of the linear model was not particularly normal. (The Anderson-Darling test of normality: $p = 0.030$.) The residual analysis did not reveal any non-linearity (visual inspection of Figure 3). No observations were removed.

5.3. Kemerer Regression Model

This data set did exhibit some heteroscedasticity, although not pronounced (cf. Figure 4), and low kurtosis. We therefore investigated an OLS log-linear as well as an OLS linear model, and opted for the log-linear model on the same grounds as for Albrecht. LAD was not investigated because visual inspection of residuals did not

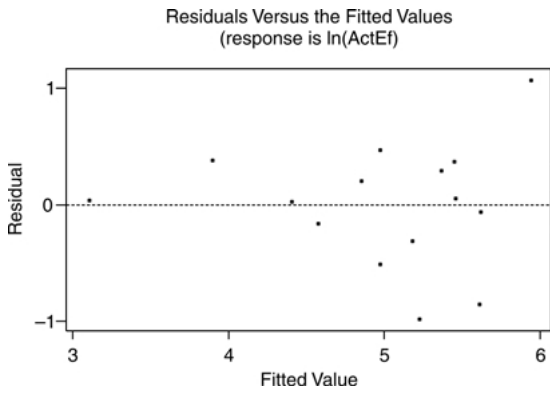

Figure 4. Kemerer, residuals vs. fits.

Table 8. Kemerer OLS log-linear regression model.

Predictor	Coef	SE Coef	T	P
Constant	− 1.055	1.196	− 0.88	0.394
Ln(FP)	0.9039	0.1780	5.08	0.000
S	R-Sq	R-Sq(adj)		
0.5437	66.5%	63.9%		

suggest high kurtosis. From Table 8, we observe that the FP coefficient is significant ($T > 2$). Also, the explanatory power is reasonably high (R-sq = 46.0%).

Residual analysis suggested that the distribution of residuals of the linear model was normal. (The Anderson-Darling test of normality: $p = 0.563$.) The residual analysis did not reveal any non-linearity either (visual inspection of Figure 4). No observations were removed.

5.4. Finnish Regression Model

Two observations had missing data. We applied listwise deletion as the missing data technique, and we thus used 38 out of 40 cases in the analysis. (Two cases with missing values were removed. It should be observed that with such a low percentage of missing data, almost any missing data technique will do.) No outliers were removed.

The data set exhibited a pronounced heteroscedasticity. We therefore applied a log-linear OLS model to this data set. This model was efficient (cf. Table 9). We therefore did not investigate LAD in addition to OLS. From Table 9, we observe that the FP coefficient is significant ($T > 2$). Also, the explanatory power is high (R-sq = 56.2%).

Residual analysis suggested that the distribution of residuals of the log-linear model was normal. (The Anderson-Darling test of normality: $p = 0.176$.) The

Table 9. Finnish OLS log-linear regression model.

Predictor	Coef	SE Coef	T	P
Constant	1.6999	0.9936	1.71	0.096
ln(FP)	1.0528	0.1550	6.79	0.000
S	R-Sq	R-Sq(adj)		
0.79	56.2%	55.0%		

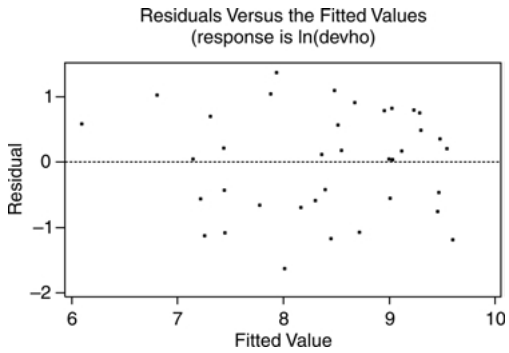

Figure 5. Finnish, residuals vs. fits.

residual analysis did not reveal any non-linearity either (visual inspection of Figure 5).

5.5. DMR Regression Model

The data set exhibited a pronounced heteroscedasticity. We therefore applied a log-linear OLS model to this data set. The log-transformed data seemed to exhibit high kurtosis and skewness: several outlying observations on one side, like a separate group of data. On inspection, all these outlying observations used the same language: language 3. (There are three language types in the data set, language 1, 2 and 3.) Language 3 projects exhibited distinctly different productivity from language 1 and 2 projects. We therefore decided to use projects with language 1 and 2, only. This resulted in a screened data set of 71 observations (out of the initial 81). The screened data did not exhibit high kurtosis, and we therefore used OLS. On this log-transformed data set, OLS was very efficient. We therefore used an OLS log-linear model as the final model.

From Table 10 (T and P columns), we observe that the FP coefficient is significant ($T > 2$). Also, the explanatory power is high (R-sq = 71.5%). Residual analysis suggested that the distribution of residuals of the final log-linear model was normal

Table 10. DMR OLS log-linear regression model.

Predictor	Coef	SE Coef	T	P
Constant	2.9575	0.4123	7.17	0.000
ln(FP adj)	0.98726	0.07495	13.17	0.000
S	R-Sq	R-Sq(adj)		
0.3868	71.5%	71.1%		

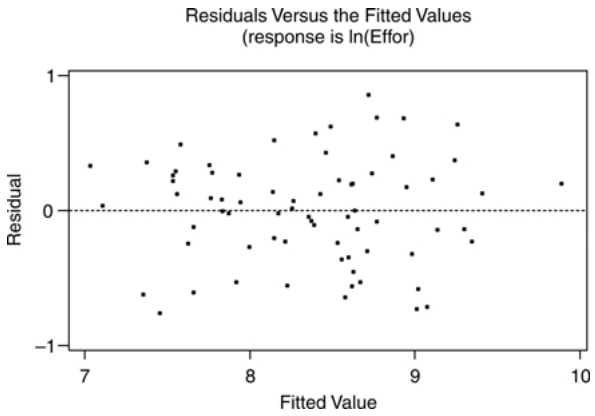

Figure 6. DMR, residuals vs. fits.

(Anderson-Darling test of normality: p-value $= 0.586$). The residual analysis did not reveal any non-linearity either (visual inspection of Figure 6.)

6. Results—MRE

6.1. Accenture-ERP

The scatter plot in Figure 7 of MRE vs. $\hat{y}$ of the Accenture-ERP data set seems to suggest that MRE and $\hat{y}$ are slightly negatively correlated. At least, it does not convincingly suggest that MRE and $\hat{y}$ are totally uncorrelated. Especially, we observe a few very outlying observations in the range $0 < \hat{y} < 15{,}000$ with MRE's > 1.0. We therefore investigated this further by performing an OLS regression analysis. From Table 11, we see that the slope (the coefficient of $\hat{y}$) is not significantly different from zero ($|T| \ll 2$). We are therefore not able to reject H0 based on this test. The results of the Park test ($T = -0.19$) also suggest that we may

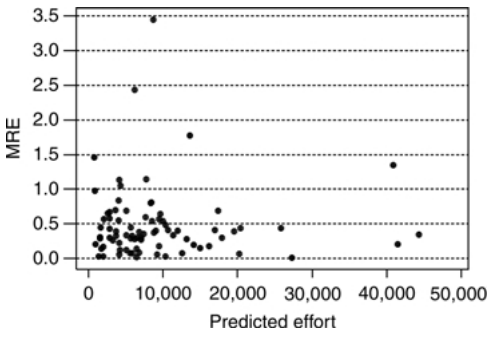

Figure 7. MRE vs. predicted effort, Accenture-ERP.

Table 11. Accenture-ERP OLS regression model of MRE vs. $\hat{y}$.

Predictor	Coef	SE Coef	T	P
Constant	− 0.48851	0.08658	− 5.64	0.000
$\hat{y}$	− 0.00000037	0.00000689	− 0.05	0.957

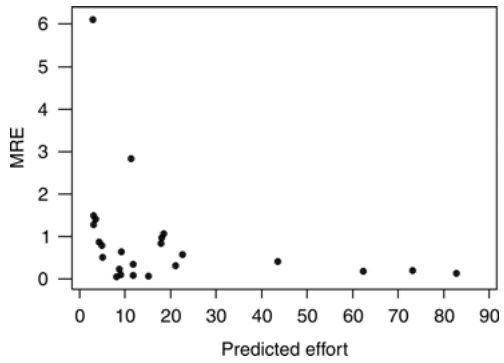

Figure 8. MRE vs. predicted effort, Albrecht.

accept the assumption of homoscedasticity. Thus, the results suggest that we may accept the assumption that MRE and $\hat{y}$ are uncorrelated.

6.2. *Albrecht*

The scatter plot of MRE vs. $\hat{y}$ in Figure 8 suggests that MRE and $\hat{y}$ are negatively correlated. Also, we observe a few very outlying observations in the range $0 < \hat{y} < 15$ with MRE's $\gg 1.0$. We therefore investigated this further by performing an OLS regression analysis. From Table 12, we see that the slope (the coefficient of $\hat{y}$) is not significantly different from zero ($|T| < 2$) although the slope coefficient suggests a decreasing mean. Still, we are not able to reject H0. The results of the Park test ($T = -1.49$) also suggest that we may accept the assumption of homoscedasticity. We observe though a weak negative tendency suggesting that heteroscedasticity may be present in the data (-1.49 is close to -2.00). Thus, the results suggest that we may accept the assumption that MRE and $\hat{y}$ are uncorrelated, but it is not a strong,

Table 12. Albrecht OLS regression model of MRE vs. $\hat{y}$.

Predictor	Coef	SE Coef	T	P
Constant	− 0.7185	0. 2726	2.64	0.015
$\hat{y}$	− 0.009533	0. 009219	− 1.03	0.312

convincing result. We may interpret the results as indicating that MRE and $\hat{y}$ are (weakly) negatively correlated.

6.3. Kemerer

The scatter plot of MRE vs. $\hat{y}$ in Figure 9 suggests that MRE and $\hat{y}$ are positively correlated. This visual observation is, however, not supported by the OLS regression results in Table 13. We see that the slope (the coefficient of $\hat{y}$) is not significantly different from zero ($T = 1.27 < 2.00$) although there seems to be a slight positive tendency. Still, we are not able to reject H0. The results of the Park test ($T = 1.89$) also suggest that we may accept the assumption of homoscedasticity. We observe though a weak tendency suggesting that heteroscedasticity may be present in the data (1.89 is close to 2.00). Thus, the results suggest that we may accept the assumption that MRE and $\hat{y}$ are uncorrelated, but it is not a strong, convincing result. We may also interpret the results as indicating that MRE and $\hat{y}$ are (weakly) positively correlated.

6.4. Finnish

The scatter plot of MRE vs. $\hat{y}$ in Figure 10 suggests that MRE and $\hat{y}$ are uncorrelated. Only, we observe one very outlying observation in the range $0 < \hat{y} < 5000$ with $MRE > 4.0$. This visual observation is supported by the OLS regression results in Table 14. We see that the slope (the coefficient of $\hat{y}$) is not significantly different from zero ($|T| = 0.74 < 2.00$) although there seems to be a slight negative

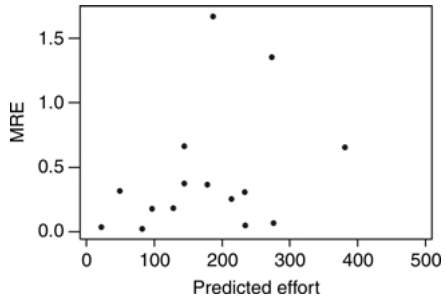

Figure 9. MRE vs. predicted effort, Kemerer.

Table 13. Kemerer OLS regression model of MRE vs. $\hat{y}$.

Predictor	Coef	SE Coef	T	P
Constant	−0.1374	0.2628	0.52	0.610
$\hat{y}$	−0.001674	0.001319	1.27	0.227

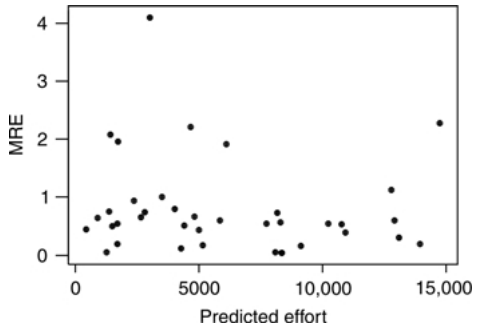

Figure 10. MRE vs. predicted effort, Finnish.

Table 14. Finnish OLS regression model of MRE vs. $\hat{y}$.

Predictor	Coef	SE Coef	T	P
Constant	0.9378	0.2374	3.95	0.000
$\hat{y}$	− 0.00002425	0.00003255	− 0.74	0.461

correlation. But this may be due to the single, large outlier. All in all, we are not able to reject H0. The results of the Park test (T = − 0.90) also suggest that we may accept the assumption of homoscedasticity. Thus, the results suggest that we may accept the assumption that MRE and $\hat{y}$ are uncorrelated.

6.5. DMR

The scatter plot of MRE vs. $\hat{y}$ suggests that they are uncorrelated (Figure 11). This visual observation is supported by the OLS regression results in Table 15. We see that the slope (the coefficient of $\hat{y}$) is not significantly different from zero

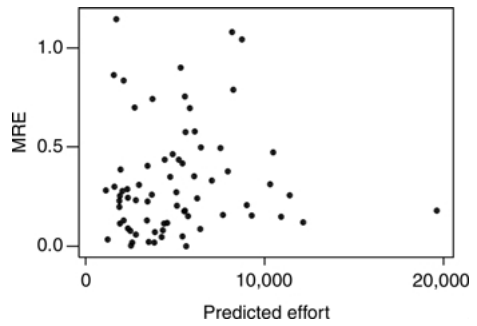

Figure 11. MRE vs. predicted effort, DMR.

Table 15. DMR OLS regression model of MRE vs. $\hat{y}$.

Predictor	Coef	SE Coef	T	P
Constant	0.30173	0.06235	4.84	0.000
$\hat{y}$	0.00000499	0.00001040	0.48	0.633

($T = 0.48 \ll 2$). We are therefore not able to reject H0. The result of the Park test ($T = 0.63$) also suggests that we may accept the assumption of homoscedasticity. Thus, the results suggest that we may accept the assumption that MRE and $\hat{y}$ are uncorrelated.

7. Discussion

In this section, we argue for the use of $\hat{y}$ (predicted effort) rather than y (actual effort) as the project size measure. We also explain why we obtain different results, i.e., why MRE and y can be negatively correlated while MRE and $\hat{y}$ are uncorrelated for the same data set.

In this study, we have used the predicted effort to deem a project small or large whereas in our previous studies, we used actual effort. Since it makes a big difference in the results whether we choose y or $\hat{y}$ along the x-axis, we need to find a good reason to prefer one to the other and an explanation of why the results depend on this choice.

The only good reason for preferring y as the x-axis is the following: Intuitively, in our previous studies, we found it natural to plot MRE against y because MRE uses the y as divisor. In MRE, it is the actual, not the prediction, that is the first-class citizen. This was our reason for using it in our previous studies. There are, however, several better reasons for preferring $\hat{y}$ as the x-axis:

i. In residual analysis, it is common to plot residuals against $\hat{y}$ instead of against y. Thus, copying residual analysis practices, it seems we ought to plot MRE against $\hat{y}$.

ii. The $\hat{y}$ axis is equivalent to the predictor variable axis (e.g. FP). The y axis is not, due to the stochastic error term. That is, we would observe the same pattern of correlation between $\hat{y}$ and MRE and between FP and MRE.

iii. Another argument in support for using $\hat{y}$ on the x-axis is that if I want to assess the precision for my current prediction (i.e., a project for which actual effort is not yet known), I can only do that if I use the estimate as the x-axis.

Finally, we need to explain why we observe a negative correlation between MRE and y, but not between MRE and $\hat{y}$. The explanation has to do with the fact that

MRE is asymmetric and not with the fact that MRE includes a $1/y$ term. The explanation is as follows. Projects with small actuals tend to be below the regression line whereas projects with large actuals tend to be above. Since MRE is asymmetric, a project with a small y and negative residual, $y - \hat{y}$, will easily obtain a large MRE, i.e., MRE 100% (For example, let $y = 0.01$ and $\hat{y} = 1$. Then, $|y - \hat{y}| = 0.99$. Consequently, MRE $= 0.99/0.01 = 9900\%$) whereas a positive residual will always result in an MRE $< 100\%$. (For example, let $y = 1.99$ and $\hat{y} = 1$. Then, the absolute value of the residual is $|y - \hat{y}| = 0.99$, the same as before. Unlike the large MRE in the previous example, MRE $= 0.99/1 = 99\%$.) To reiterate, a finite value divided by a very small value will result in a large value. Thus, when plotting MRE against y, the large MRE values will tend to occur for small values of y, i.e., towards the left of the x-axis, and vice versa for small MRE values. To conclude, we actually create an (artificial) effect when plotting MRE against y because of the asymmetry in MRE, and not because there automatically is a negative correlation between y and x/y regardless of x. The problem with the asymmetry of MRE we avoid when plotting MRE against $\hat{y}$.

8. Alternative Evaluation Criteria for Further Research

One of our objections to MRE is that the absolute residual is divided by the actual rather than the estimate. From a practitioner's perspective, it would make more sense to use a measure of performance relative to the estimate rather than to the actual because, from a project life-cycle perspective, the estimate always precedes the actual. Also, from personal experience (one of the authors is an experienced software project manager), we know it is common to measure estimating performance relative to the estimate.

In this section, we briefly present alternative evaluation criteria to consider. It is beyond the scope of this paper to explore these measures in depth as we have done with MRE. Rather, this is a topic of further research.

Kitchenham et al. (2001) propose using the mean or median EMRE (the magnitude of relative error relative to the estimate) which, unlike MRE, uses the estimate as the divisor, not the actual.

Miyazaki et al. (1994) propose BRE (the balanced relative error) and IBRE (the inverted balanced relative error).

As for traditional statistics, it is more common to use the residual rather than a relative error measure like MRE as the basic metric. A common summary statistic in statistical science is MSE (the minimum mean-square-error) estimator. MSE measures the dispersion around the true value of the residual (see, for example, a textbook like Gujarati, 1995).

Another evaluation criterion, the correlation coefficient (r for univariate and R for multivariate cases), is commonly used in statistics to evaluate regression models. It measures the correlation between actuals and estimates.

9. Conclusions

It has never been empirically validated that MRE is uncorrelated with (or independent of) project size. This is a necessary, but not sufficient, condition to defend its use as an evaluation metric for prediction systems. Were MRE and project size negatively correlated, and at the same time, were the (non-relative) residual and project size uncorrelated, then we ought to use the (non-relative) residual rather than a relative residual like MRE. Summary statistics like the mean MRE, i.e., MMRE, assume that MRE is uncorrelated with project size. Otherwise, this summary statistic, the mean, is not very meaningful. For example, were MRE and project size negatively correlated, the MMRE statistic would be meaningful only to medium-sized projects, but not for small or large projects in the data set.

The results suggest that we may accept the assumption that MRE and project size are uncorrelated. Investigating five software project data sets, we did not find sufficient empirical evidence to reject the assumption that MRE and project size are uncorrelated, i.e., that MRE exhibits constant mean. For the two largest data sets (Accenture-ERP and DMR), the results clearly point in this direction. For the three smallest data sets (Albrecht, Kemerer, and Finnish), however, the results are less clear-cut. For the latter, one may suspect that there is a (weak) correlation between MRE and project size. However, whereas the correlation is negative for Albrecht and Finnish, it is positive for Kemerer. Since these three data sets are small compared to the other two, we believe it is reasonable to pay more attention to the results of the two larger data sets.

To summarize, from this investigation, we have no reasons to dismiss the use of MRE and MMRE. However, in this paper, we have investigated one single property of MRE, only: the correlation between MRE and project size. We would like to repeat, however, that it is a necessary, but not sufficient, condition that MRE and project size are uncorrelated. There are a number of other requirements that need to be met on order to ensure we have a reliable evaluation metric. Other studies by Kitchenham et al. (2001) and by Foss et al. (2002) suggest there are other, serious flaws with MMRE. In those studies, the conclusions partly discourage the use of MRE and MMRE and propose other evaluation criteria. It should be observed that there is no contradiction between those studies and this study. Together, these three studies state that some properties of MRE are acceptable (this study), and that other properties are not (the other two studies). Since the other studies suggest that MRE and MMRE seem to possess some undesirable properties, it therefore still seems like a good idea to investigate other evaluation criteria such as those presented in Section 8 and to be cautious using MMRE as an evaluation metric.

Appendix: Calculation of MRE in Log-Linear Regression Models

This appendix shows how the formula for calculating MRE is derived when one applies a log-linear (or log-log) regression model to predict effort. Suppose the log-

linear model, with y = actual effort, is

$$\ln y = \ln \alpha + \beta \ln X + \ln u$$

Then, predicted effort (or rather the predicted ln-effort) is

$$\ln \hat{y} = a + b \ln X \tag{1}$$

Let the residual be given by

$$\textit{residual} = \ln y - \ln \hat{y} \text{ which is equal to residual} = \ln \left(\frac{y}{\hat{y}} \right)$$

This may be transformed to

$$e^{\textit{residual}} = \frac{y}{\hat{y}} \text{ or alternatively } e^{-\textit{residual}} = \frac{\hat{y}}{y}$$

Thus,

$$1 - e^{-\textit{residual}} = \frac{y - \hat{y}}{y} \tag{2}$$

By definition, MRE is

$$\text{MRE} = \left| \frac{y - \hat{y}}{y} \right| \tag{3}$$

From (2) and (3) we may restate MRE

$$\text{MRE} = \left| 1 - e^{-\textit{residual}} \right|$$

Acknowledgment

The Finnish data were collected by Hannu Maki from the TIEKE organization.

References

Albrecht, A. J., and Gaffney, J. R. 1983. Software function, source lines of code, and development effort prediction: a software science validation. *IEEE Trans. Software Eng.* 9(6): 639–648.

Briand, L. C., and Wieczorek, I. 2001. Resource modeling in software engineering. In *Encyclopedia of Software Engineering* 2 volume set, J. Marciniak (ed.) Wiley.

Conte, S. D., Dunsmore, H. E., and Shen, V. Y. 1986. *Software Engineering Metrics and Models*. Menlo Park, CA: Benjamin-Cummings.

Desharnais, J. M. 1989. analyze statistique de la productivite des projects de developpement en informatique a partir de la technique des points de fonction. Master's thesis, Universite du Quebec a Montreal.

Foss, T., Myrtveit, I., and Stensrud, E. 2001a. MRE and heteroscedasticity: An empirical validation of the assumption of homoscedasticity of the magnitude of relative error. *Proc. ESCOM*. The Netherlands: Shaker Publishing BV, 157–164.

Foss, T., Myrtveit, I., and Stensrud, E. 2001b. A comparison of LAD and OLS regression for effort prediction of software projects. *Proc. ESCOM*. The Netherlands: Shaker Publishing BV, 9–15.

Foss, T., Stensrud, E., Kitchenham, B., and Myrtveit, I. 2002. A Simulation Study of the Model Evaluation Criterion MMRE. Discussion paper 3/2002, Norwegian School of Management.

Gujarati, D. N. 1995. *Basic Econometrics*. Third Edition. New York: McGraw-Hill.

Gujarati, D. N. 2003. *Basic Econometrics*. Fourth Edition. New York: McGraw-Hill.

Kemerer, C. F. 1987. An empirical validation of software cost estimation models. *Comm. ACM* 30(5): 416–429.

Kitchenham, B. A., MacDonell, S. G., Pickard, L. M., and Shepperd, M. J. 2001. What accuracy statistics really measure. *IEE Proceedings Software* 148(3): 81–85.

Kitchenham, B. A., and Kansala, K. 1993. Inter-item correlations among function points. *Proc. First METRICS*. Los Alamitos, California: IEEE Computer Society Press, 11–14.

Little, R. J. A., and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Minitab, Minitab Statistical Software, release 13, http://www.minitab.com.

Miyazaki, Y., Terakado, M., Ozaki, K., and Nozaki, H. 1994. Robust regression for developing software estimation models. *Journal of Systems and Software* 27: 3–16.

Myrtveit, I., Stensrud, E., and Olsson, U. 2001. Analysing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. *IEEE Trans. Software Eng.* 27(11): 999–1013.

Myrtveit, I., and Stensrud, E. 1999. A controlled experiment to assess the benefits of estimating with analogy and regression models. *IEEE Trans. Software Eng*. 25(4): 510–525.

Rosenberg, J. 1997. Some misconceptions about lines of code. *Proc. METRICS'97*. Los Alamitos CA: IEEE Computer Society Press, 137–142.

Stensrud, E., Foss, T., Kitchenham, B., and Myrtveit, I. 2002. An empirical validation of the relationship between the magnitude of relative error and project size. *Proc. METRICS'02*. Los Alamitos CA: IEEE Computer Society Press, 3–12.

Stensrud, E., and Myrtveit, I. 1998. Human performance estimating with analogy and regression models: an empirical validation. *Proc. METRICS'98*. Los Alamitos CA: IEEE Computer Society Press, 205–213.

Stensrud, E. 1998. Estimating with enhanced object points vs. function points. *Proc. COCOMO'13*. B. Boehm and R. Madachy (eds.), Los Angeles CA: USC Center for Software Engineering.