Modeling the Relationship between Software Effort and Size Using Deming Regression

Nikolaos Mittas
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
+302310998236
nmittas@csd.auth.gr

Makrina Viola Kosti
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
+302310998236
mkosti@csd.auth.gr

Vasiliki Argyropoulou
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
+302310998236
vargyrop@csd.auth.gr

Lefteris Angelis
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
+302310998230
lef@csd.auth.gr

ABSTRACT

Background: The relation between software effort and size has been modeled in literature as exponential, in the sense that the natural logarithm of effort is expressed as a linear function of the logarithm of size. The common approach to estimate the parameters of the linear model is ordinary least squares regression which has been extensively applied to various datasets. The least squares estimation takes into account only the error arising from the dependent variable (effort), while the measurement of independent variable (size) is considered free of errors.

Aims: The basis of the study is that in practice the assumption of measuring the size without error is hardly true, since the size of a software project depends on the precision of the tool of measurement and often by the subjectivity of the rater. Moreover, the sizes of projects comprising a dataset have been measured by different measurement tools and this adds another source of variability in the independent variable.

Method: In this paper, we consider a regression technique, known as Deming regression, which takes into account the error in measurement of the independent variable, the size. Deming regression is applied to four publically available datasets in order to model the linear relationship between effort and size and to compare it with ordinary least squares.

Results: Accuracy measures of fitting (MAE, MdAE, MMRE, MdMRE, pred25) are improved by the Deming regression. Comparison of Absolute Errors (AE) by the Wilcoxon test shows significant difference at <0.001 level of significance.

Conclusions: Deming regression is appropriate for datasets where the size is subject to measurement error. However some assumptions on the variances of the measurement errors are arbitrary and need to be studied. Further work is needed for using the Deming regression for effort prediction.

Categories and Subject Descriptors

D.2.9 [**Software Engineering**] Management – *cost estimation*

PROMISE2010, Sep 12-13, 2010. Timisoara, Romania
Copyright 2010 ACM ISBN 978-1-4503-0404-7...$10.00.

General Terms

Algorithms, Management, Measurement

Keywords

Software Cost Estimation, Ordinary Least Squares Regression, Error-in-variables Model, Deming Regression.

1. INTRODUCTION

One of the most important phases for the software project planning process is cost estimation. *Software Cost Estimation* (SCE) is the activity of predicting the effort (or cost) required to develop a new software system. Due to this requirement, there is a large discussion on the relationship between effort and other cost drivers. The most important factor affecting the cost of the projects is the software size ([1], [2]). As literature reveals [3], the researchers have proposed various methods and techniques in order to model the relationship between software effort and size. Most of these studies rely the modeling of the relationship on a software size metric as an input in the model which provides the final estimate for the effort of projects.

Regression Analysis (RA), and especially *Ordinary Least Squares* (OLS) *regression*, is the modeling technique that has attracted the researchers' interest during the past decades [3] since in well-known models (i.e. COCOMO), the researchers have applied a form of regression in order to "capture" the abovementioned relationship. OLS regression explains the relation between the size and effort of projects in the form of an exponential function which can be transformed to linear by the logarithmic transformation. The logarithmic transformation usually normalizes the dependent effort variable.

Despite the popularity of OLS in SCE, this form of RA is subject to several shortcomings since in OLS it is assumed that the values of the independent variable (i.e. the size) are measured without errors. However, for the case of SCE the software size is essentially the result of a counting and estimating process derived from a tool or an expert, either expressed as number of *Source Lines of Code* (SLOC) or *Function Points* (FP). Under this perspective the assumption of error-free measurement is not so realistic and it can be an inhibitory factor for the precision and accuracy of the modeling process. Indeed, the counting procedure of the size of projects in software development depends on the tools used and is affected by subjective decisions which can differ among the practitioners. As a result, there is always the possibility that the estimation results for the same software can vary even between people in the same organization.

In this paper, we explore the possibility of improving the process of modeling the relationship between effort and size of projects through the utilization of *Deming regression* [4]. Deming regression belongs to the general class of *errors-in-variables models* which are more appropriate to apply in situations where random errors exist in the measurements of both the independent and the dependent variable. As it is reasonable to assume measurement errors in the size of software projects, the proposed approach seems to be an alternative and more generally applicable than OLS in the problem of modeling the relationship between effort and size of projects by fitting a regression line.

Trial applications on four publicly available software datasets show the benefits of using Deming regression for the construction of a robust function between effort and size of projects.

The rest of this paper is organized as follows: Section 2 summarizes related work. Section 3, briefly presents the OLS regression technique as applied in SCE. Section 4 describes the proposed Deming regression model. In Section 5, we detail the methodology as applied to the datasets. In Section 6, we present the results of the application and finally in Section 7 we conclude by discussing the results and by providing some directions for future research.

2. RELATED WORK

As we have already mentioned, regression-based approaches significantly dominate in SCE since half of all studies deal with the fitting, improvement or comparison of regression models [3]. There are several forms of regression models, according to the way the functional relationship between the dependent cost variable and the independent cost drivers is estimated. However, *Ordinary Least Squares* (OLS) appears to be one of the most popular techniques. According to OLS, the relationship between cost and size is first expressed as a known explicit function with unknown parameters (the regression coefficients). Then, the parameters are estimated in such a way so as to minimize the sum of squared residuals, i.e. the squared differences between observed and predicted values.

Despite the popularity of OLS and the large amount of studies dealing with this specific form, the researchers have mentioned that there are certain limitations due to the requirement that both the structural model and the error distribution have to be correctly specified. Miyazaki et al. in [5] claim that OLS method does not give proper parameter values evaluated by the criteria commonly used in the field of SCE. For this reason the authors proposed the use of OLS based on relative errors $RE_i = (Y_{A_i} - Y_{E_i}) / Y_{A_i}$, where Y_{E_i} is the estimated value of a dependent variable and Y_{A_i} is the actual value of the cost variable for the i^{th} project, respectively.

Although the method could solve the OLS problem of inconsistency between the calibration method and the evaluation criteria for calibrated models, it was not robust enough to solve the problem of outliers [6]. Miyazaki et al. in [6] point out that the robustness of the statistical method that is used for the description of the relationship between the effort and the size of projects is very crucial since software data tend to be more inconsistent than data in other field of statistics. One reason for the inconsistent collection of the data is the lack of standardization in software terminology since there is no universal definition even for the number of lines of code. Furthermore, the collection of the data, especially for large projects, is based both on software tools and manpower and this fact alone increases the complexity and ambiguity of the counting process and thus, the variability of the modeling process between effort and size. Finally, the authors in [6] note the important role of development standardization on the quality of the estimation process. In order to overcome the abovementioned problems they proposed the use of a method called the *Least Squares of Inverted Balanced Relative Errors* (LIRS), whereas they also concluded that a model with many parameters and independent variables is difficult to calibrate and for this reason simple models with a few independent variables are more practical than models with many parameters.

In [7], the authors examine the problem of *heteroscedasticity* that is the situation in which the variance of the dependent variable varies across the data. Heteroscedasticity complicates analysis because OLS is based on the assumption of constant variance. The presence of heteroscedasticity in software project data is usually related to the structure of software measurement or metric that vary in size from a few lines of codes to thousands [7]. The authors also point out that the usage of two specific forms of regression *(Least Median of Squares*-LMS and *Least Trimmed Squares*-LTS) provide robust estimation when there are outliers in the data but they do not provide accurate estimation under heteroscedasticity. The aforementioned techniques were also the topic of interest in [8], in which three statistical techniques (OLS, LMS and LTS) and a neural network were evaluated on various dataset characteristics. The conclusion derived from the analysis is that no modeling method is the best in every case due to the different dataset characteristics.

Pickard et al. [9] investigate the efficacy of different data analysis techniques including *Robust Regression* (RR) and *Least Absolute Deviation* (LAD) on simulated software data with different characteristics (i.e. skewness, unstable variance, outliers and combination of these characteristics) concluding that there is not a global best model but different dataset characteristics may favor different modeling techniques.

Foss et al. [10] claim that although OLS is the de facto technique in SCE, software dataset may however exhibit certain characteristics that do not always comply with the requirements of OLS. They also remark the inaccurate software datasets due to the cost and difficulty of gathering data. Moreover, they remark that when a practitioner observes an unusual value, it is difficult to determine whether it is an outlying point or just an erroneous value due to a reporting or measurement error. In order to overcome the restriction of OLS to handle these values, they proposed the usage of *Least Absolute Deviation* (LAD) regression. Although they expected LAD to be more efficient than OLS on a specific dataset with errors and outliers, they reached to the conclusion that the analysis did not offer some support for LAD.

In a recent study [11], the researchers proposed a robust technique for calibrating the parameters of COCOMO model by imposing constraints on the model. They tried different objective functions (i.e. sum of squared errors, sum of absolute errors and sum of magnitude relative errors), whereas the constraints imposed non-negative coefficients and an upper limit of magnitude relative error. The proposed technique was compared with OLS and two other types of constrained regression, namely the *Lasso* and *Ridge* regression. Although the results seem to be encouraging, there are also certain limitations since the forms of constrained regressions may be trapped in a local optimum and return a sub-optimal solution.

Summarizing some of the findings of the literature, we have to conclude that the researchers recognize the important role of OLS technique on the difficult task of modeling the relationship between the effort and size of the projects. On the other hand, this form of regression is not free of restrictions and limitations and there is an ongoing research in the utilization of more robust techniques. Although there have been introduced many variations of regression, there is always an existing factor that may result in incorrect regression coefficients and this is the imprecision in the measurement of the size (independent variable) of projects.

As the most important independent variable in most SCE models is the size, accurate measures of size of the deliverables of a software project at early stages of the development process will allow the accurate evaluation of the relationship between the deliverables and the cost (effort or duration) required to accomplish them [12]. Despite the fact that there is an ongoing research on software metrics, only two software size metrics are widely used in practice, which are the number of SLOC of the deliverable project and the number of FP.

SLOC is the oldest metric for measuring project effort, whereas it is utilized by two well-known cost estimation models, the Putnam's *Software LIfecycle Management* (SLIM) and Boehm's COnstructive COst MOdel (COCOMO). The most significant advantage of SLOC is that it is directly related to the software to be developed. On the other hand, SLOC cannot be accurately measured from the early stages of the development process [13]. Furthermore, there is also noted an obstacle to the standardization of counting SLOC since there is even no standard definition of what a line code is. In order to overcome this limitation, organizations give certain directions to formalize the counting process but there is always the risk for errors in the size measurement, due to the fact that new programming languages are introduced.

Due to the crucial drawbacks of the SLOC metric, Albrecht [14] developed an alternative software size metric (FP). This metric is essentially the weighted sum of five different factors (inputs, outputs, logic files, inquires and interfaces) that are related with the user requirements. According to the inspirer, FPs would be much easier to measure than SLOC at the early stages of project lifecycle; whereas the programmers could also compare different systems since FP counting is free of language specification. Despite the popularity of FPs in different software applications, there are also critics concerning the reliability of the method. The first one regards the objectivity of the measurement process, since two different practitioners performing an FP count for the same project would generate different results [12]. Another problem arises from the fact that there are many variations of the original method developed by Albrecht, posing difficulties in the data collection. Finally, the method cannot be easily automated and thus CASE tools are difficult to be developed in order to automatically collect data.

From what we have discussed above, OLS method seems to be an appropriate technique to model the relationship between effort and size of projects when the measurements of size are counted with precision and are error-free. But this is not the usual case in SCE, since the measurement of size is a complicated task containing subjectivity and high variability. The problem of inadequate measurements can have a significant impact on SCE which can lead to wrong managerial decisions on the scheduling, monitoring and controlling of the whole development process.

Having in mind all these topics, we can conclude that there is a practical need for a modeling technique that is able to incorporate the measurement errors of the project sizes in the building of a robust relationship between the effort and size. We therefore suggest Deming regression which seems to be an appropriate technique, able to address the abovementioned problems related to the errors occurring during the counting process of size.

3. ORDINARY LEAST SQUARES REGRESSION

Describing the general modeling process in SCE, we denote by Y the real random dependent variable representing in our context the cost of projects (usually expressed by the effort) and by X the real random independent variable representing the size of projects. The historical dataset that is used to model the relationship between effort and size consists of a sequence of observations $(x_1, y_1),...,(x_n, y_n)$, where each project is represented by a pair of numbers $x_i, y_i \in \Re$.

Our goal is to find a regression function $f(x_i)$ which will serve for building a model of the form

$$y_i = f(x_i) + \varepsilon_i \qquad (i = 1,..., n) \qquad (1)$$

We assume that the random errors ε_i are independent with zero mean. Parametric estimation techniques assume that the function $f(x_i)$ of Eq. (1) can be estimated by a linear expression of x_i,

$$f(x_i) = \beta_0 + \beta_1 x_i \qquad (2)$$

where β_0 is the *constant* and β_1 the *slope* of the regression line, the unknown parameters or the regression coefficients that have to be estimated. When the relationship between the dependent and the independent variable is not linear, we assume that a simple transformation (like the logarithmic) can always result to a model of the form

$$y_i = \beta_0 + \beta_1 x_\iota + \varepsilon_i \,. \qquad (3)$$

As we already mentioned, the most common method for estimating the unknown regression coefficients is OLS, where the estimation is based on the minimization of the overall *Sum of Squared Residuals* (SSR).

$$SSR = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (4)$$

As illustrated in Figure 1, OLS minimizes the sum of the squared lengths of the linear segments, vertical to the x-axis, from the data points to the regression line. The estimation of the parameters β_0 and β_1 based on OLS is given by the following equations:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2} \qquad (5)$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \tag{6}$$

where

$$\bar{y} = \frac{1}{n}\sum\nolimits_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i \tag{7}$$

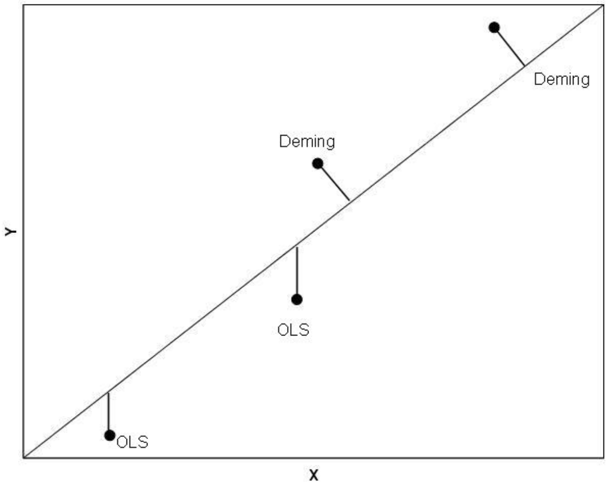

Figure 1. OLS projects the points onto the line in the vertical to the x-axis direction; Deming regression projects the points onto the line at an angle determined by λ (here $\lambda = 1$)

The fitting of OLS model presupposes that (a) the relationship between effort and size is linear and (b) the residuals ε_i are normally distributed and uncorrelated with the predictor. These two assumptions are usually addressed in the case of software project data by using the logarithmic transformations of the effort (dependent) variable and the size (independent) variables [15, 16].

4. DEMING REGRESSION

Except from the abovementioned assumptions of OLS, there is also a significant one that it is not usually taken into account when building a model between the effort and size of software projects. This assumption is that the independent variable X is measured without error.

As we have described in Sections 1 and 2, this is not realistic in SCE since the size, measured either by SLOC or FP, is not a reliable measurement since different practitioners estimating the same application could obtain different results. Having in mind that OLS analysis assumes that only the Y measurements are associated with random measurement errors, the slope estimate becomes biased. For this reason, a practitioner has to apply a more appropriate methodology taking into account random errors in both the size (X) and effort (Y) of software projects.

Deming regression is a form of *errors-in-variables* model which tries to fit the best line on a dataset assuming measurement errors for both sets of measurements (X) and (Y) [4].

Let us assume that the available historical data (x_i, y_i) are erroneously measured observations of the true, but unknown, values (X_i, Y_i). The measured value is likely to deviate from the true value by some small "random" amount $(\varepsilon_i, \delta_i)$. For a given dataset $(x_1, y_1),...,(x_n, y_n)$ the equations describing the model are

$$x_i = X_i + \varepsilon_i \tag{8}$$

$$y_i = Y_i + \delta_i \tag{9}$$

The error terms ε_i and δ_i are assumed to be independent normal variables with zero mean values. In order to estimate the regression line by the Deming methodology, it is necessary to evaluate or assign a value to the ratio λ of the variances of ε and δ for the x and y, respectively.

$$\lambda = \frac{S_{e_x}^2}{S_{\delta_y}^2} \tag{10}$$

The objective of Deming regression is to find the best fitting line

$$Y_i = \beta_0 + \beta_1 X_\iota \tag{11}$$

such that the weighted SSR of the model is minimized.

$$\begin{gathered} SSR = \sum\nolimits_{i=1}^{n} \left(\frac{{\varepsilon_\iota}^2}{S_{\varepsilon_x}^2} + \frac{{\delta_\iota}^2}{S_{\delta_y}^2}\right) = \\ \sum\nolimits_{i=1}^{n} ((y_i - \beta_0 - \beta_1 X_i)^2 + \lambda (x_i - X_i)^2) \end{gathered} \tag{12}$$

The value of λ determines the angle in which the points are projected onto the line in order to minimize SSR [17]. By setting $\lambda = 1$, the results of Deming regression is equal to the results of *orthogonal regression* which takes into account the distance of each data point from the line (Figure 1).

The final estimates of the parameters by Deming regression are given by

$$\hat{\beta}_1 = \frac{s_{yy} - \lambda s_{xx} + \sqrt{(s_{yy} - \lambda s_{xx})^2 + 4\lambda {s^2}_{xy}}}{2 s_{xy}} \tag{13}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{14}$$

$$X_i = x_i + \frac{\hat{\beta}_1}{{\hat{\beta}_1}^2 + \lambda}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{15}$$

where

$$s_{xx} = \frac{1}{n-1}\sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2 \tag{16}$$

$$s_{yy} = \frac{1}{n-1}\sum\nolimits_{i=1}^{n} (y_i - \bar{y})^2 \tag{17}$$

$$s_{xy} = \frac{1}{n-1}\sum\nolimits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{18}$$

It must be noted that the value of λ is not known in advance and it can be estimated only by designing a study where multiple measurements of the same cases (projects) are taken and these measurements vary. Since in historical data such information is not available, we take in our applications the simplest orthogonal case, $\lambda = 1$ which usually gives better results than OLS even if the λ is misspecified [17]. An interesting property of this type of regression is this: if we solve the equation derived with respect to *X*, the new equation will be equivalent to that obtained by considering the *X* variable as dependent and the *Y* variable as independent. In other words, Deming's method always results in one line, whether *X* or *Y* is used as the independent variable. This property does not hold for OLS.

Another notable point is that as we can see from Eq. (11), the true values of the dependent variable *Y* are estimated after estimating not only the regression coefficients, but also the true values of the independent variable *X* using the expression in Eq. (15).

5. METHODOLOGY

In order to evaluate the fitting accuracy of Deming regression in comparison to that of OLS, we utilized two well-known error measures from the literature ([18, 19]). Based on the actual Y_A and the estimated value Y_E, we calculate for each project i ($i = 1,...,n$) of the dataset

1. The *absolute error* (AE)
2. The *magnitude of relative error* (MRE)

Through the abovementioned local measures of errors (the term "local" refers to the individual error generated by the effort estimation of a single project) (Table 1), the global accuracy measures for the model can be evaluated by the mean and median values of all local errors (Table 2).

Table 1. Local accuracy measures

$AE_i = \left\| Y_{A_i} - Y_{E_i} \right\|$	$MRE_i = \frac{\left\| Y_{Ai} - Y_{Ei} \right\|}{Y_{Ai}}$

Table 2. Global accuracy measures

$MAE = \frac{1}{n}\sum_{i=1}^{n} AE_i$	$MdAE = median\{AE_i\}$
$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$	$MdMRE = median\{MRE_i\}$
$pred25 = \frac{\#(\text{projects with } MRE \leq 0.25)}{\#(\text{projects})}$	

The whole distributions of the derived AEs are used in order to perform a statistical test, so as to compare the performances of the two comparative models. Due to the fact that AEs are usually non-normally distributed, highly skewed with many outliers, we finally apply the non-parametric Wilcoxon sign rank test for paired samples [19, 20]. Regarding the graphical comparison, we select to indicatively present the boxplots of AEs since the formal comparison testing is based on these magnitudes of errors.

Moreover, we also use in our analysis a graphical tool for visual comparison of the models, the *Regression Error Characteristic* (REC) curves. REC curves proposed by Bi and Benett [21] provide a visualization tool for comparison of comparative models, analogous to that of *Receiver Operating Characteristic* (ROC) curves in classification problems. Briefly, REC curves plot simultaneously the *Cumulative Distribution Functions* (CDF) of the errors (y -axis), obtained by different models, offering a graphical technique to estimate the probability of the error to be less or equal than the respective x -axis value. Due to the fact that REC curves are very informative taking into account the whole error distribution of the errors, Mittas and Angelis [22, 23] introduced REC curves in SCE, proposing a throughout visualization framework that can be used by projects managers in order to reinforce their knowledge about the validity and precision of alternative models.

Finally, we present the differences of Deming and OLS lines and the fitting accuracy of each model through scatter plots of the points (x_i, y_i) for each software dataset.

6. APPLICATION TO DATASETS

In this section, we present the results of Deming and OLS regressions described earlier as applied to four datasets publicly available datasets, known from the literature. The size in these datasets is measured either by SLOC or FP.

Describing briefly the building process of both models, the variables are initially checked for normality using the *One-Sample Kolmogorov-Smirnov Test* (K-S test) and *Q-Q* plots to ensure the assumptions related to using these models are satisfied. As a result, both the dependent *effort* and *size* variables are transformed to a natural logarithmic scale providing the new variables *lneffort* and *lnsize*, respectively. Finally, the form of both Deming and OLS regressions is given in Eq. (19). We have also to emphasize that for the evaluation of the accuracy measures as described in Section 5, the model is re-transformed to the initial scale through the utilization of exponential transformation.

$$\ln effort = b_0 + b_1 \ln size \tag{19}$$

6.1 The Desharnais dataset

The first dataset used in our study is the Desharnais dataset which contains data of 77 completed software projects from a Canadian Software house [24]. The project size measured with FPs ranges from 73 FPs to 1127 FPs and the project effort ranges from 546 person hours to 23940 person hours.

The estimation of the regression coefficients (intercept and slope), for both OLS and Deming regressions, are given in the first row of Table 3. Based on the estimated parameters, we construct the regression lines for both models (Figure 2). As we can observe from Table 3, the intercept and slope estimated by the Deming regression appear to be significantly different from those estimated by OLS. The resulting lines in Figure 2 are quite different. What we can note from these lines is that the Deming regression has a better fitting to some outliers appearing at the lower left part of the scatter plot.

The overall performance of the two comparative models is presented in Table 4. We can see that the Deming model outperforms OLS in terms of all the accuracy measures. In the last column of Table 4, we also provide the percentage of the improvement in all measures of Deming regression compared to that of OLS. The improvement achieved by Deming regression ranges from 67.60% (MdMRE) up to 148.17% (pred25).

In order to examine the distributions of local accuracy measures, we indicatively present the boxplots for AEs (Figure 3). The boxplots show that Deming regression has generally better behavior than OLS. The box length and tails of Deming are significantly smaller than OLS regression, whereas the outliers are less extreme compared to the outliers of OLS. Furthermore, OLS presents large variability since the interquantile range (height of the box) is large.

Indeed, the abovementioned conclusions are supported by the conduction of the non-parameteric Wilcoxon test between the AEs of the comparative models (first row of Table 5), which indicates that there is a statistically significant difference (p<0.001) between the distributions of AEs obtained by Deming and the OLS models.

Table 3. Regression coefficients for OLS and Deming regression for all datasets

	OLS		Deming	
Dataset	intercept	slope	intercept	slope
Desharnais	2.993	0.929	-2.212	1.868
COCOMO81	1.204	1.106	0.243	1.404
Maxwell	3.517	0.827	2.088	1.065
NASA93	1.977	0.920	1.277	1.107

Table 4. Accuracy measures for the Desharnais dataset

	OLS	**Deming**	**Improvement (%)**
MAE	2101.13	614.25	70.77%
MdAE	1107.46	335.59	69.70%
MMRE (%)	66.88	15.28	77.15%
MdMRE (%)	34.88	11.30	67.60%
pred25 (%)	35.06	87.01	148.17%

Table 5. Significance of the Wilcoxon tests for the AEs values for all datasets

Dataset	**p-values of Wilcoxon tests**
Desharnais	<0.001
COCOMO81	<0.001
Maxwell	<0.001
NASA93	<0.001

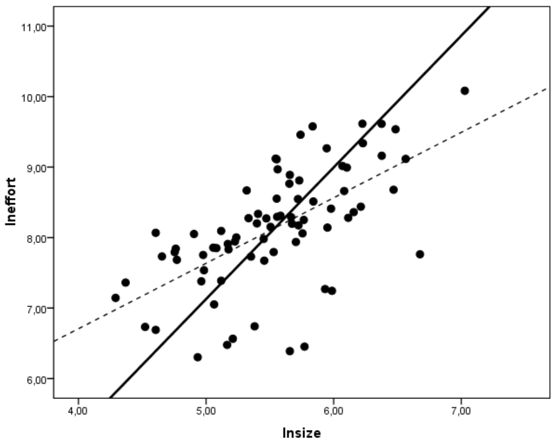

Figure 2. Deming (solid line) vs. OLS (dashed line) models for the Desharnais dataset

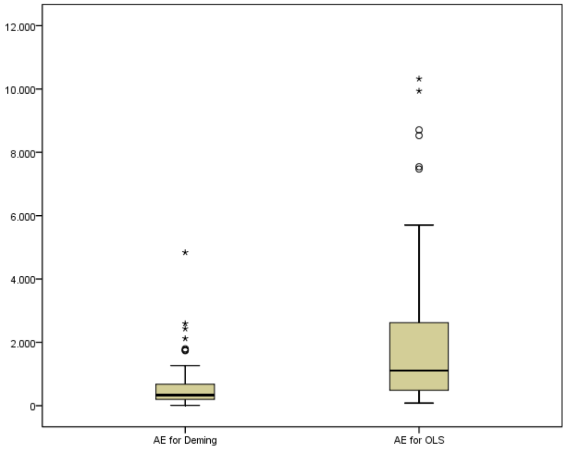

Figure 3. Desharnais dataset: Boxplots of the AE values for Deming and OLS models

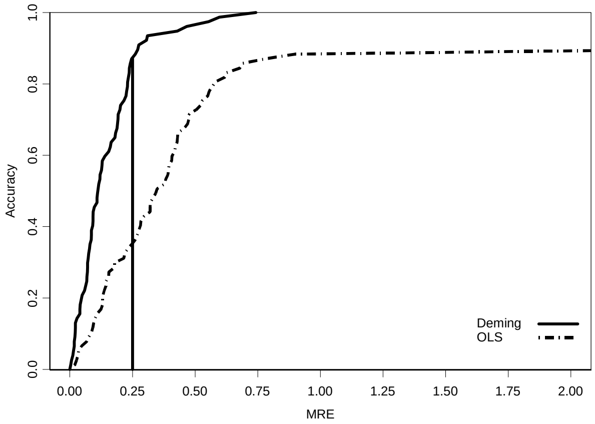

Figure 4. Desharnais dataset: REC curves of the MRE values for Deming and OLS models

Finally, we also present the REC curves (Figure 4) for the two comparative models based on the MRE values of each model. The interpretation of such a plot is generally simple [22, 23]. More precisely, if a curve is placed in higher position with respect to the other curve in the plot, then the corresponding model outperforms and thus a model performs well if the REC curve climbs rapidly towards the upper left corner. It is clear that Deming regression, represented by the solid line, dominates OLS over the whole range of possible MREs.

Another property of REC curves is that they present certain interesting geometrical characteristics, which they can utilized in order to evaluate graphically certain statistics such as the pred25 measure. The values of pred25 for the two models are visualized by drawing first a reference vertical line from 0.25 of the *x* -axis and then from the intersecting point of the REC curve, a horizontal line which meets the accuracy axis. From the relative positions of the two pred25 values, we can infer that Deming regression achieves the highest (and hence the best) pred25 measure.

REC curves can also be used for the identification of extreme errors. When these outliers are present, the top of the REC curve will be flat and will not reach 1 until MRE values become high. For example, in Figure 4, we can see that the MRE REC curve for OLS does not reach 1. This fact is a consequence of the presence of few projects producing MREs higher than 200% and exceed the upper limit (2.0) of the *x* -axis.

6.2 The COCOMO81 dataset

The COCOMO81 dataset is a public domain database that has been utilized in the calibration of the well-known COCOMO algorithmic model [1]. The sizes for projects are SLOC measurements with a minimum of 1.98 SLOC to a maximum of 1150 SLOC, whereas the effort ranges from 1.38 up to 11400 calendar months with a mean of 682.74 calendar months.

As we can observe from Figure 5, the Deming model seems to be fitted better than OLS to the COCOMO81 data and achieves again the best accuracy measures with the improvement to vary between 51.20% and 158.32% (Table 6). Moreover, the distribution of AEs for OLS (Figure 6) has high variability compared to the corresponding distribution obtained by the Deming regression. The hypothesis test conducted for AEs demonstrates that there is indeed a statistical significant difference between these distributions (second row of Table 5).

In Figure 7, we can notice that Deming regression clearly dominates, whereas for MRE values higher than 75% the REC curve of OLS is flat due to extreme values and does not reach 1, until the error becomes extremely high.

Table 6. Accuracy measures for COCOMO81 dataset

	OLS	Deming	Improvement (%)
MAE	455.37	222.22	51.20%
MdAE	63.90	22.47	64.84%
MMRE (%)	137.38	32.99	75.99%
MdMRE (%)	63.97	26.31	58.87%
pred25 (%)	19.05	49.21	158.32%

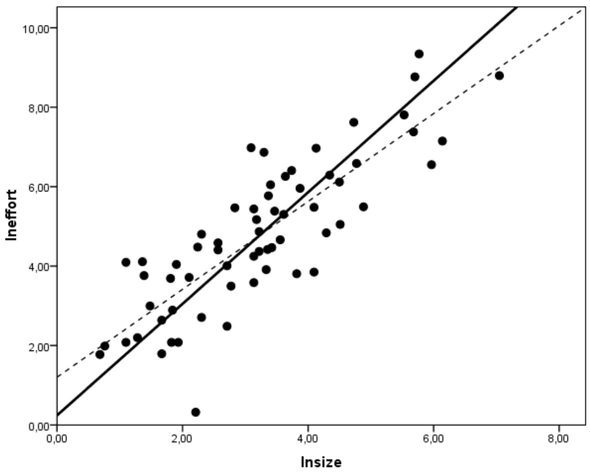

Figure 5. Deming (solid line) vs. OLS (dashed line) models for the COCOMO81 dataset

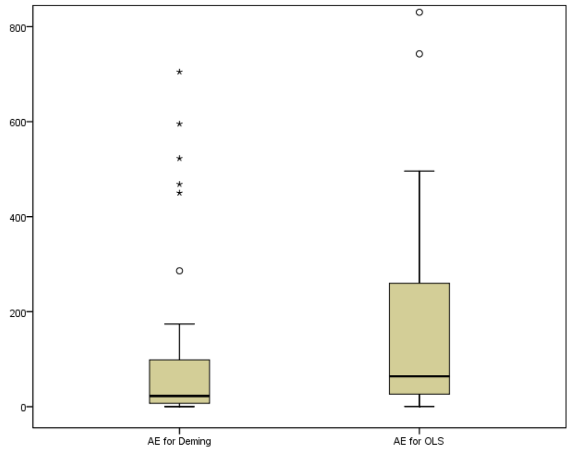

Figure 6. COCOMO81 dataset: Boxplots of the AE values for Deming and OLS models

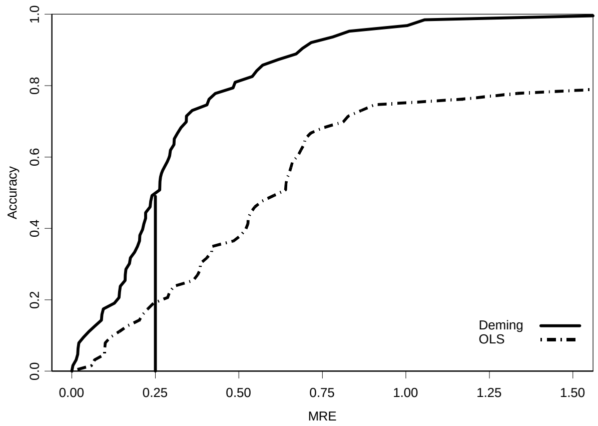

Figure 7. COCOMO81 dataset: REC curves of the MRE values for Deming and OLS models

6.3 The Maxwell dataset

The third dataset used in our applications contains 63 projects from a commercial Finnish bank [25]. Due to the existence of an outlying project, we built both models on 62 projects. The dependent variable (effort) is essentially a measurement of the work carried out by the software supplier from specification until delivery ranging from 583 up to 63694 hours. The independent variable (size) is measured through FP with a minimum value of 48 FP to a maximum of 3634 FP.

Table 7. Accuracy measures for Maxwell dataset

	OLS	Deming	Improvement (%)
MAE	3766.83	1856.38	50.72%
MdAE	1997.54	1068.19	46.52%
MMRE (%)	55.33	25.46	53.99%
MdMRE (%)	45.22	22.67	49.87%
pred25 (%)	20.97	56.45	169.19%

The fitted values of Deming regression (Figure 8) give the best results in all the accuracy measures (Table 7). The decreased global errors show that the improvement ranges from 46.52% up to 169.19%. The distribution of AEs for OLS (Figure 9) presents high variability with a long upper tail. Again, the Wilcoxon test (third row of Table 5) signifies a statistical significant difference between the distributions of AEs obtained by the Deming and OLS models. Figure 10 also depicts the large difference between the pred25 of Deming and the pred25 of the OLS model. Furthermore, the REC curve of Deming regression significantly outperforms the corresponding REC curve of OLS model since the solid line climbs rapidly to 1.

6.4 The NASA93 dataset

The final dataset is the NASA93 containing 93 projects from different centers [26]. The dependent variable is the actual effort measured in person months, whereas the independent variable is the equivalent physical 1000 lines of source code (SLOC).

Regarding the graphical inspection of the regression lines (Figure 11), we can observe that the comparative models do not present very different values for their parameters (fourth row of Table 3). This fact is also depicted from the accuracy measures (Table 8) in which it is clear that the improvement of global accuracy measures due to the utilization of Deming model is the smallest compared with the corresponding improvements of the previous three datasets.

Table 8. Accuracy measures for NASA93 dataset

	OLS	Deming	Improvement (%)
MAE	346.51	198.94	42.59%
MdAE	70.34	34.21	51.36%
MMRE (%)	65.79	26.77	59.31%
MdMRE (%)	36.08	16.02	55.60%
pred25 (%)	33.33	64.52	93.58%

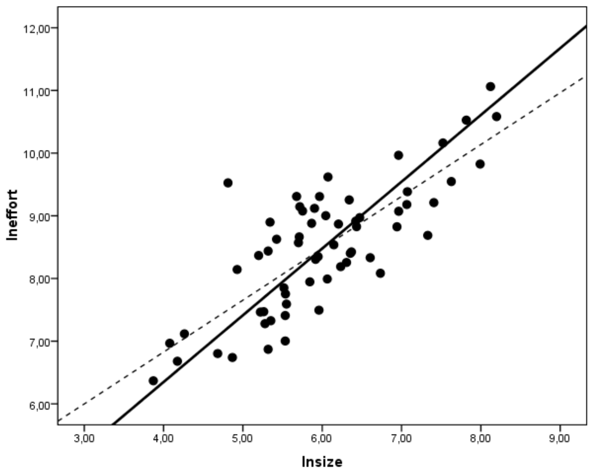

Figure 8. Deming (solid line) vs. OLS (dashed line) models for the Maxwell dataset

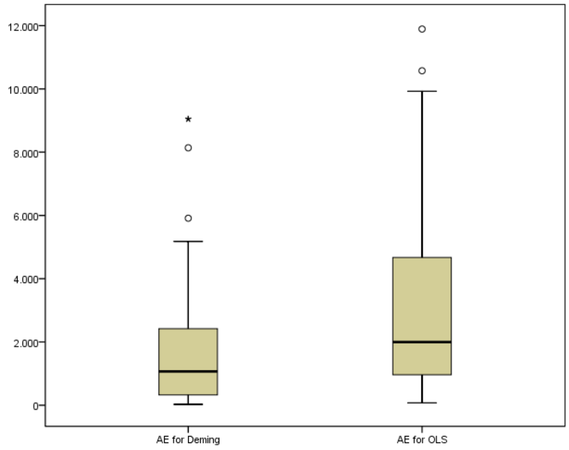

Figure 9. Maxwell dataset: Boxplots of the AE values for Deming and OLS models

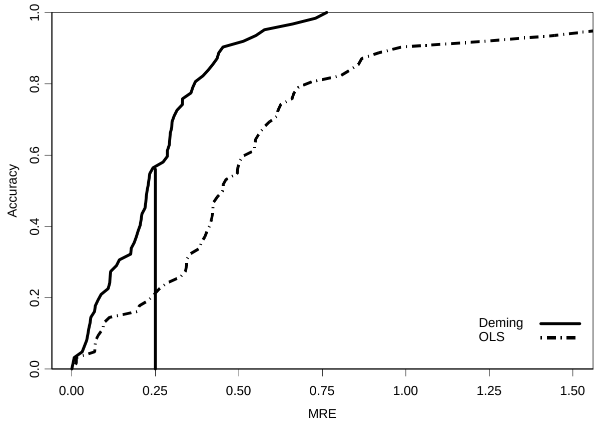

Figure 10. Maxwell dataset: REC curves of the MRE values for Deming and OLS models

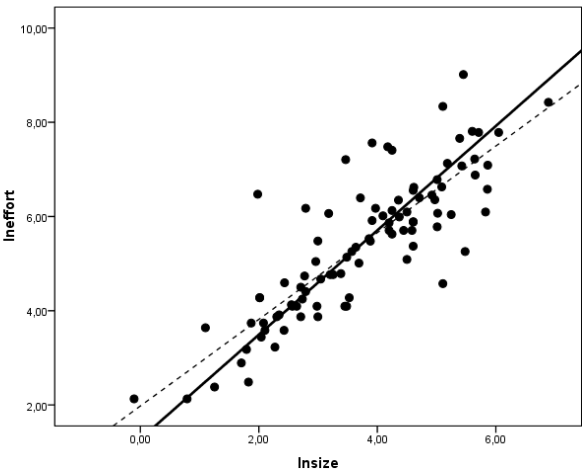

Figure 11. Deming (solid line) vs. OLS (dashed line) models NASA93 dataset

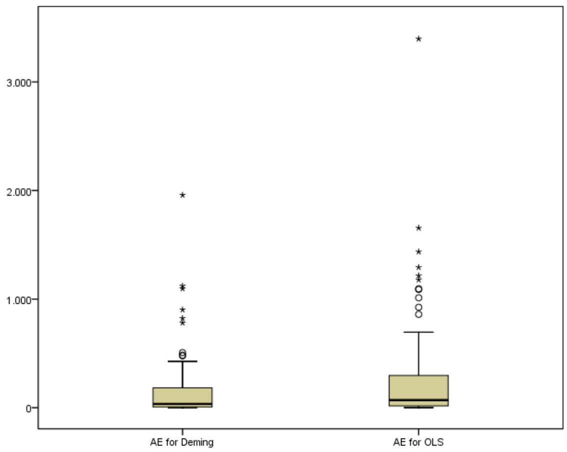

Figure 12. NASA93 dataset: Boxplots of the AE values for Deming and OLS models

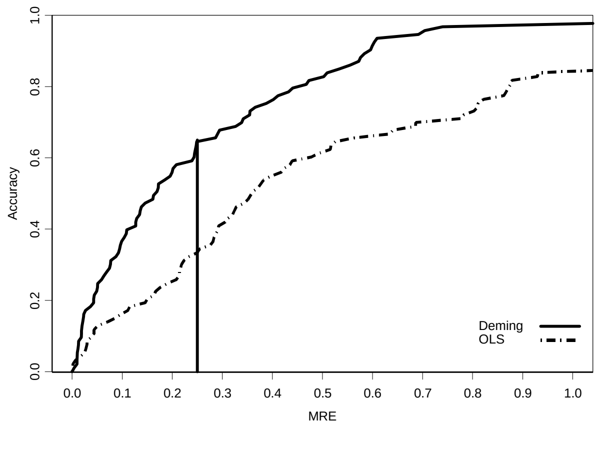

Figure 13. NASA93 dataset: REC curves of the MRE values for Deming and OLS models

The examination of the boxplots for AEs (Figure 12) reveals that the errors computed from OLS present slightly higher variability compared with those of Deming's model. However, the Wilcoxon test shows that there is statistically significant difference between the AEs of the comparative models (fourth row of Table 5). Finally, the inspection of MRE REC curves (Figure 13) shows that the Deming's REC curve significantly dominates the corresponding REC curve for OLS with the latter to present extremely high variability and high errors.

7. DISCUSSION AND CONCLUSIONS

In this paper, we considered the problem of modeling the relationship between the software effort and size. The problem seems to be a trivial statistical problem addressed by a form of regression. Usually in such cases we can find easily a model using OLS which is the most common method of regression not only in SCE but in most research fields. However, this relationship is the base for any cost estimation model involving several variables and is therefore very important to describe it accurately.

The main idea behind this paper was the fact that the usual OLS is applied under the assumption that the observed values of the variables are measurements which coincide with the true values of the quantities they are supposed to measure. This assumption is not realistic in the context of SCE, since our datasets most usually contain heterogeneous projects, not only with respect to their nature, but also in the way they were measured. The dependence of the measurements of size and effort on the tools and the human judgment is widely recognized and there is generally a belief that the same project can give different measurements when measured by different people and different tools. Therefore, there is always the possibility for the practitioners to neglect a systematic error in their models. These ideas led us to consider statistical methodologies which take into account the possible error in the measurements.

More specifically, we investigated and we suggest a new technique for estimating the regression coefficients, namely Deming regression, for the modeling of the relationship between the effort and size of software projects. Deming regression presupposes that besides the error in the dependent variable, there is also an additional source of error related with the measurement of the independent variable. Under this assumption, it is considered as more robust method than OLS.

The application of Deming regression to four well-known, publicly available datasets showed significant improvement compared to OLS. This improvement is supported by several accuracy measures, graphical inspection and statistical tests. In our applications we used the simplest orthogonal version of Deming regression which assumes that the variances of the errors in the independent and the independent variables are equal. Of course, this may not be generally true; however it is impossible to estimate these variances from historical data without multiple measurements of the same project. Despite this fact, even this simple version, just by taking into account the error in the independent variable, managed to improve significantly the results. Deming regression can be seen as an alternative generalized technique that can be proved quite beneficial in cases where we have reasons to believe that the counting process of the size is characterized by uncertainty due to the subjective decisions of the practitioners and the tools used.

Although there are encouraging results from this first analysis, the method deserves a deeper and thorough study. Some interesting issues came up from our work and deserve further research. First of all, our aim is to focus on the construction of Prediction Intervals which provide an “optimistic” and “pessimistic” guess for the true magnitude of the cost. This seems to be an interesting research topic, since in various studies in SCE the researchers suggest that interval estimation is more realistic than a single point estimate, accounting for both uncertainty and risk. In fact, under the assumption of error in measurement, the point estimate is meaningless in the sense that it expresses not the response to the true size value, but the response to the measured value.

Another issue arisen from the present study is the entrance of more than one explanatory (or independent) variable in the model in order to increase the percent of variability of the effort that is explained by the cost function. In this preliminary study, we investigate the capabilities of the proposed methodology in a simple cost model but there is also the necessity for the inclusion of more cost drivers.

Finally, a very important question needing systematic treatment trough simulation is the examination of the performance of the comparative models to different situations in which the errors of the independent variable ranges from a small amount into a high source of variability.

8. REFERENCES

[1] Boehm, B. 1981. *Software Engineering Economics*. Prentice-Hall, New Jersey.

[2] Boehm, B., Horowitz, E., Madachy, R., Reifer, D., Bradford K., Steece, B., Brown, A., Chulani, S., and Abts, C. 2000. *Software Cost Estimation with COCOMO II*. Prentice Hall, New Jersey.

[3] Jorgensen, M., and Shepperd, M.J. 2007. A systematic review of software development cost estimation studies. *IEEE Trans Softw Eng* 33 (1), 33-53.

[4] Deming, W. 1943. *Statistical adjustment of data*. Wiley, NY (Dover Publications edition, 1985).

[5] Miyazaki, Y., Takanou, A., Nozaki, H., Nakagawa, N., and Okada, K. 1991. Method to estimate parameter values in software prediction models. *Inf Softw Tech* 33 (3) (Apr. 1991), 239-243.

[6] Miyazaki, Y., Terakado, K., Ozaki, K., and Nozaki, H. 1994. Robust regression for developing software estimation models. *J. Syst Softw* 27, 3–16.

[7] Chen, Y.L., and Stromberg, A.J. 1997. Robust estimation in software experiments. ACM SIGSOFT Software Engineering Notes, vol. 22, iss. 4, pp. 60-64.

[8] Gray, A.R., and MacDonell S.G. 1999. Software metrics data analysis-Exploring the relative performance of some commonly used modeling techniques. *Emp Softw Eng* 4 (4), 297-316.

[9] Pickard, L., Kitchenham, B., and Linkman, S. 1999. An investigation of analysis techniques for software datasets. In In *Proceedings of the METRICS 99 Symposium*, IEEE Computer Society, pp. 130-142.

[10] Foss, T., Myrtveit, I., and Stensrud, E. 2001. A comparison of LAD and OLS Regression for Effort Prediction of software projects. In *Proceeding of 12th European Software Control and Metrics Conference*, pp. 9-15.

[11] Nguyen, V., Steece, B., and Boehm, B. 2008. A constrained regression technique for COCOMO calibration. In *Proceedings of the ACM-IEEE 2nd International Symposium on Empirical Software Engineering and Management* (ESEM’08) (Kaiserslautern, Germany, 9-10), pp. 70-79.

[12] Kemerer, C. 1993. Reliability of function points measurement: a field experiment. *Comm of the ACM* 36 (2), 85-97.

[13] Low, G., and Jeffery, D. 1990. Function points in the estimation and evaluation of the software process. *IEEE Trans Softw Eng* 16 (1), 64-71.

[14] Albrecht, A. 1979. Measuring application development productivity. In GUIDE~SHARE: In *Proceedings of the IBM Applications Development Symposium*, pp. 83-92.

[15] Kitchenham, B, and Mendes, E. 2004. A comparison of cross-company and within-company effort estimation models for web applications. In *Proceedings of the Empirical Assessment in Software Engineering*, pp. 47-55.

[16] Mendes, E, and Lokan, C. 2008. Replicating studies on cross- vs single-company effort models using the ISBSG database. *Emp Softw Eng* 13 (1), 3-37.

[17] Linnet, K. 1998. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clin Chem* 44 (5), 1024–1031.

[18] Foss, T., Stensrud, E., Kitchenham, B., and Myrtveit, I. 2003. A simulation study of the model evaluation criterion MMRE. *IEEE Trans Softw Eng* 29 (11), 985-995.

[19] Kitchenham, B., Pickard, L., MacDonell, S., and Shepperd, M. 2001. What accuracy statistics really measure. *IEE Proc. Software* 148 (3), 81-85.

[20] Mittas, N., and Angelis, L. 2008. Comparing cost prediction models by resampling techniques. *J. Syst Softw* 81 (5), 616-632.

[21] Bi, J., and Bennet, K.P. 2003. Regression error characteristics curves. In *Proceedings of the AIII 20th International Conference on Machine Learning*, pp. 43–50.

[22] Mittas, N., and Angelis, L. 2008. Comparing software cost prediction models by a visualization tool. In *Proceedings of the IEEE 34th Euromicro Conference on Software Engineering and Advanced Applications* (SEAA’08), pp. 433–440.

[23] Mittas, N., and Angelis, L. 2010. Visual Comparison of Software Cost Estimation Models by Regression Error Characteristic Analysis. *J. Syst Softw* 83, 621-637.

[24] Desharnais, J. 1989. *Analyse statistique de la productivitie des projets informatique a partie de la technique des point des function*. Masters Thesis. University of Montreal.

[25] Maxwell, K. 2002. *Applied Statistics for Software Managers*. Prentice-Hall, PTR.

[26] NASA93 dataset. 2007. http://promisedata.org/