Why Comparative Effort Prediction Studies may be Invalid

Barbara Kitchenham
School of Computing and Mathematics
Keele University, Keele,
Stoke on Trent, ST5 5BG, UK
b.a.kitchenham@cs.keele.ac.uk

Emilia Mendes
Computer Science Department
University of Auckland
Auckland, New Zealand
emilia@cs.auckland.ac.nz l

ABSTRACT

Background: Many cost estimation papers are based on finding a "new" estimation method, trying out the method on one or two past datasets and "proving" that the new method is better than linear regression. **Aim:** This paper aims to explain why this approach to model comparison is often invalid and to suggest that the PROMISE repository may be making things worse. **Method**: We identify some of the theoretical problems with studies that compare different estimation models. We review some of the commonly used datasets from the viewpoint of the reliability of the data and the validity of the proposed linear regression models. **Discussion points**: It is invalid to select one or two datasets to "prove" the validity of a new technique because we cannot be sure that, of the many published datasets, those chosen are the only ones that favour the new technique. When new models are compared with regression models, researchers need to understand how to use regression analysis appropriately. The use of linear regression presupposes: a linear relationship between dependent and independent variables, no significant outliers, no significant skewness, no relationship between the variance of the dependent variable and the magnitude of the variable. If all these conditions are not true, standard statistical practice is to use a robust regression or transform the data. The logarithmic transformation is appropriate in many cases, and for the Desharnais dataset gives better results than the regression model presented in the PROMISE repository. **Conclusions**: Simplistic studies comparing data intensive methods with linear regression will be scientifically valueless, if the regression techniques are applied incorrectly. They are also suspect if only a small number of datasets are used and the selection of those datasets is not scientifically justified.

Categories and Subject Descriptors

D.2.9 [**managements**]: Cost estimation

General Terms

Economics

Keywords

Effort estimation; model construction; linear regression, model comparison.

1. INTRODUCTION

One of us (Kitchenham) is a frequent reviewer of cost estimation papers. In the past year, she has reviewed three papers (by different authors) that compared an exotic new estimation method with linear regression on some historic datasets. These papers all used the COCOMO dataset [3] and two also used the Desharnais dataset [5]. The percentage Mean Magnitude Relative Error (MMRE) values they report for linear regression models for the two datasets are shown in Table 1. Clearly anything we say about these unpublished studies is hearsay and has little evidential value. However, we also include the linear regression MMRE value for Desharnais dataset reported in the PROMISE repository plus an MMRE for linear regression reported in another published paper [20] and a stepwise linear regression with jack-knife (i.e. leave one out) MMRE reported for the COCOMO dataset [18]. Although the published MMRE values come from relatively old studies, both are referenced in another recent paper [12], as establishing the linear regression model accuracy of the two datasets.

Table 1 Reported values of %MMRE for linear regression models using the Desharnais and COCOMO datasets

Source	Technique	COCOMO 81	Desharnais
Manuscript (MS)1	Multiple Linear Regression (LR)	172.24	58.8
MS 2	Stepwise LR	1540.8	51.3
MS 3	LR (from [18] and [21])	1540 & 520	Not used
PROMISE	LR		65.0
[18]	Stepwise LR (Using Jackknife)	520.71	Not used
[20]	LR	Not used	66

A feature of Table 1 is that the values are inconsistent (particularly for COCOMO) and are all, as we will demonstrate later, incompatible with values obtained by good practice statistical analysis. It might be suggested that papers using invalid models will simply be rejected, so no harm is done. However, they still represent wasted time both for the researchers and reviewers, and furthermore, if the reasons for rejections are not made public other researchers may make similar mistakes.

Other frequently used datasets are the Albrecht-Gaffney dataset, also referred to as the IBMDPS dataset [4] and the International Software Benchmarking Standards Group (ISBSG) dataset. Again, the percentage MMRE values reported to be based on a form of linear regression model, are somewhat inconsistent.

Table 2 Reported values of %MMRE of linear regression models for the Albrecht-Gaffney and ISBSG datasets

Source	Technique	ISBSG	Albrecht-Gaffney
[16]	General Linear model allowing for categorical variables	56.02 (release 7)	27.61
[19]	LR	36 (release 6, 52 projects)	Not used
[4]	LR	Not used	71
[7]	LR	103 (release 8, 33 projects)	43 (training set omitting smallest effort project)
[2]	LR (exponential model)	118 (release 8)	Not used
[20]	LR	Not used	90

This paper discusses why such divergent results occur. In Section 2 we discuss the problems associated with the use of historic datasets to "prove" one technique is better than another. In Section 3 we review the datasets in more detail and suggest statistical models for predicting effort for the Desharnais and COCOMO datasets that are more consistent with the nature of the data than linear regression on the raw data. We discuss our results and present our conclusions in Section 4.

2. COMPARING ESTIMATION METHODS

This section discusses some of the potential problems we have found in papers comparing cost estimation models using historical datasets.

2.1 Dataset Selection

One of the main problems with evaluating techniques using one or two datasets is that no-one can be sure that the specific datasets were not selected because they are the ones that favor the new technique. Mair et al. [15] studied 50 empirical cost estimation papers and point out that there are 31 datasets in the public domain that have been used in cost estimation studies. Given this number of published datasets, it is disappointing that published studies aimed at validating new estimation techniques only use two or three datasets and give no clear rationale for the choice of datasets that they use. Mair et al. also point out that many popular datasets are among the oldest datasets "with potential problems for obsolescence".

2.2 Non-repeatable Results

Another issue raised by Mair et al., and indicated by the examples given in Table 2, is that some datasets grow with time, the ISBSG being a point in case. The ISBSG dataset is also a problem because it contains many missing values, and researchers do not always explain how they chose the specific projects they used in their studies. This means that some papers that use the ISBSG projects (although they are using more up-to-date data) cannot have their results independently validated – which is a violation of basic scientific principles.

2.3 Expertise in competing methods

Researchers proposing new techniques for estimation are usually experts in those techniques. They may not, however, be experts in statistical methods. Some researchers use an MMRE value quoted in the literature; some develop a statistical model themselves. However, if they build a model, we cannot be sure that the regression approach that they use is consistent with good statistical practice (see Section 3.1). If they use an MMRE from the literature, we cannot be sure they have chosen an MMRE value based on a well-performed regression or simply one that is easy for their own technique to outperform. Unless all competing models are built to best current practice, comparisons are invalid.

2.4 Failure to present statistical evidence

In 1999, Stendrud and Myrveit pointed out that it is invalid to suggest that one model technique is better than another without performing some statistical test to confirm that any difference in accuracy statistics is significant [22]. They proposed paired t-tests of the MRE. Later, Kitchenham et al. suggested the use of paired tests of the absolute residuals [11], and recently, Kitchenham et al. discussed the problems the lack of statistical tests cause in meta-analytic studies [10]. Nonetheless, many recent papers still base their conclusion solely on the value of accuracy statistics, usually MMRE, Median MRE or Pred(25) with no statistical tests (e.g. [2], [7], [13], [17]).

Another important issue is whether to compare with predictions based on the entire dataset or predictions based on dividing the data into training and testing datasets. Most researchers agree that the latter technique is better, but if we use anything other than a simple leave-one-out procedures results are not auditable unless the specific dataset partitions are defined. For example, Tronto et al. [23] and Sentas et al. [19], both use exactly the same six-fold cross-validation for the COCOMO dataset proposed by Kitchenham [9]. Furthermore, Kirsopp and Shepperd [8] demonstrate that results may be misleading unless at least 5 different training sets, and preferable more than 20, are used.

Recently, Mittas and Angelis [16] suggested using re-sampling methods (i.e. bootstrapping) to identify the distribution of accuracy statistics. These methods use many different training sets, avoid the need to specify each specific training set, and provide statistical tests for accuracy statistics that do not have known distributions. Personally we prefer permutation sampling to bootstrapping because bootstrapping involves re-sampling with replacement and depends on the dataset being a random sample from a defined population, whereas permutation sampling is based on re-sampling without replacement lessening the dependency on having a random sample.

2.5 Using MMRE for Model Building and Model Comparison

The MMRE is not a very trustworthy accuracy statistics at the best of times [6]. However, when it is used both as a means of model construction (for example, as the criterion for deciding which factors to include and how many projects to use for prediction in a machine learning situation) and as a means of assessing competing models, it has two problems:

1. Foss et al. [6] have shown that MMRE is optimized by choosing a model that underestimates. So it is possible that a machine learning system may learn to choose a model that underestimates.
2. Comparisons with models that have not optimized on MMRE will be unfair.

2.6 Relevance to Real Estimation Processes

There is seldom any reliable information about the way in which the projects included in a dataset were obtained. For prediction purposes, the data on which a model is built must be representative of the projects for which the model will be used. This is not some fussy "statistical restriction" that can be avoided by use of clever AI techniques; it is a basic fact of science. No cost estimation model (or any other model, come to that) will predict well if it is asked to predict effort for projects that are substantially different in nature to the projects on which the model was built. At the very least real datasets need to take into account of any heterogeneity among the projects by appropriately classifying different types of project such as new and enhancement projects. They should also consider the age of projects – particularly for organizations that are undertaking process improvement initiatives. The action of changing (and, perhaps, improving) will change the nature of future projects, thus undermining any predictive model. Fitting data to a dataset that is static and time independent, doesn't prove that a model building technique will work well in real cost estimation situations.

3. SPECIFIC HISTORICAL DATASETS

The section looks at the Albrecht-Gaffney, Desharnais, and COCOMO datasets in more detail. In this section we report MMRE values not because we agree with MMRE as an accuracy statistic but to compare our analysis results with reported results.

3.1 Desharnais dataset

A scatter plot of the Desharnais dataset is shown Figure 1. This is based on all 81 data points because there are no missing values for effort and adjusted function points. Figure 1 shows not only the relationship between function points and effort but also the impact of the "Language Type" variable, which indicates three types of project: 1 refers to Basic Cobol projects; 2 refers to Advanced Cobol projects; 3 refers to 4GL projects. The variable is sometimes treated as an ordinal scale variable but is in fact a categorical variable and should be treated as such. This dataset shows characteristics that are typical of many software effort datasets:

- Skewness i.e. there are fewer large projects than small projects.
- Heteroscedasticity. The variability of effort increases with project size.
- Outliers. There is one extremely large data value.
- Heterogeneity. It looks as if different types of projects have different relationships between size and effort.

All these factors make it clear that a simple linear regression on the raw data scale is inappropriate. The data need to be transformed before applying linear regression. When this situation arises, it is common practice to normalise the data, i.e. to apply a functional transformation to the data values to make the distribution closer to a normal distribution. A common transformation is to take the natural log (ln), which makes larger values smaller and brings the data values closer to each other.

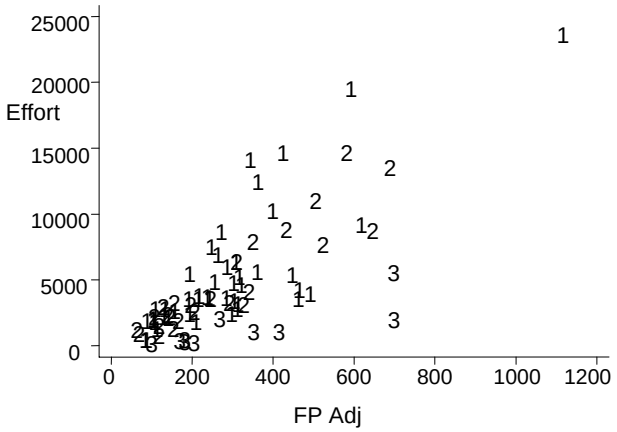

Figure 1 Effort against Adjusted Function Points for Desharnais

The Desharnais dataset contains a number of variables, in addition to effort and adjusted function points, some of which have missing values: Team Experience, Project Manager Experience, Transactions, Entities, Year Completed, Raw Function points and Function point adjustment factor (AF). Most analyses of the dataset include only the 77 projects that have no missing values.

A good statistical procedure for analyzing this dataset is to:

1. Use the natural logarithmic transformation for size variables (i.e. adjusted functions points, raw function points, transactions and entities). The logarithmic transformation ensures that the resulting model goes through the origin on the raw data scale. It also caters for both linear and non-linear relationships between size and effort.
2. Treat the three language types as two dummy variables: L1, which takes the value 1 when the language is Basic Cobol, and 0 otherwise; and L2, which takes the value 2 when the language variable is Advanced Cobol, and 0 otherwise.
3. Use stepwise regression to include only those variables that contribute significantly to the model.

Using the SPSS statistical package, this procedure leads to the following model (for the 77 projects with complete data i.e. data values for every variable):

$$\text{Ln(Effort)} = 1.685 + 0.971 \times \text{Ln(AdjFp)} + 1.368 \times L1 + 1.323 \times L2 \quad (1)$$

On the raw data scale this equates to an MMRE of 32.2% (with a median MRE of 27% a Pred(25) of 47% and a leave-one-out MMRE of 34.%). The MMRE is quite different to the regression model MMRE presented in the PROMISE dataset and other unpublished datasets (see Table 1). Note the very similar values for the L1 and L2 parameters confirms that treating Language as an ordinal scale metric is incorrect.

3.2 The Albrecht-Gaffney dataset

A scatter plot of raw function points against effort is shown in Figure 2, where effort is measured in thousand work hours. An interesting feature of this dataset is that it appears possible to deliver about 190 function points for no effort. Perhaps a more realistic interpretation is that at least some of the "projects" in this dataset are enhancement projects but the function point counts are for the total project. As Matson et al. point out it also looks as though the largest four projects are from a rather different distribution to the smaller projects [14].

Since we do not know which of the projects are new and which are enhancements, this does not appear to be a very useful dataset for testing software estimation model construction methods. In addition, a dataset that appears to have a negative intercept will give very poor MMRE values for any simple linear regression model. For example, a simple regression model based on the data shown in Figure 2 gives:

$$\text{Effort} = -16.203 + 0.0596 \times \text{RawFP} \quad (2)$$

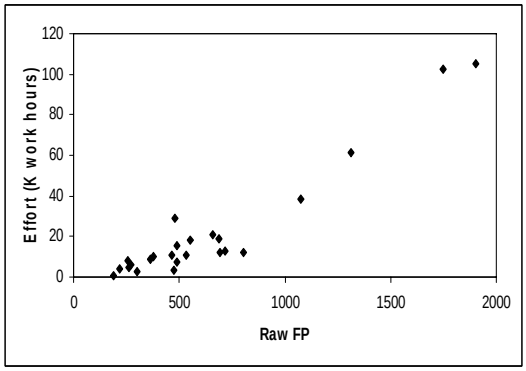

Figure 2 Effort against Raw Function Points for Albrecht-Gaffney

This gives an excellent R^2 value of 0.902, but still leads to a poor MMRE (109%) because the model is not very good for small effort values, when some fitted values become negative. For example, consider the smallest effort value (0.5) which has a Raw FP value of 189, the effort estimate is -4.912, which is not only invalid but leads to an MRE of 10.88 (1088%) for that data point alone. However, using a logarithmic transformation, we get an MMRE of 51.41% and if we omit the smallest value (like [7] and [4]) we get an MMRE of 34.75%. Overall, however, there is too much evidence that the dataset is heterogeneous for a linear regression analysis (or indeed any other modeling technique) to be a sensible option for this dataset.

3.3 COCOMO dataset

Figure 3 shows a scatter plot of the COCOMO dataset (nowadays often referred to as COCOMO-1 or COCOMO81 to distinguish it from the COCOMO-2 dataset). The COCOMO dataset includes both development projects and enhancement projects. The adjusted KSDI value is intended to adjust for re-using existing code in enhancement projects. Estimation models should therefore use adjusted KDSI rather than the total KSDI, which is the actual size of a project.

It is clear that the COCOMO dataset has similar problems to the Desharnais dataset, so the same method of analysis is required. However, the COCOMO dataset has additional problems because the values for the cost driver adjustment factors do not map exactly to the cost drivers adjustment factors used in the COCOMO model. Most researchers, who use the COCOMO data set, appear to use the adjustment factors as reported in the COCOMO data set (although by no means all, see for example [9] and [23]). For purposes of comparison, therefore, we built the following model using the Adjusted KSDI (after transformation to the logarithmic scale), all fifteen adjustment factors as reported in the dataset, and the Mode parameter which is treated as two dummy variables (EMB which takes a value 1 if the mode is embedded otherwise 0; and ORG which takes the value 1 if the mode is organic, otherwise 0):

$$\text{Ln(Effort)} = -13.728 + 71 \times \text{Ln(AKDSI)} + 1.556 \times \text{Stor} + 2.731 \times \text{Sched} + 1.442 \times \text{PCap} + 3.596 \times \text{VExp} + 0.353 \times \text{EMB} + 1.848 \times \text{ACap} + 1.065 \times \text{ModP} + 0.737 \times \text{Cplx} \quad (3)$$

On the raw data scale this model gives a percentage MMRE of 24.9% (with an Median MRE of 21% and a Pred(25) of 65). The "leave-one-out" MMRE is 31.34%

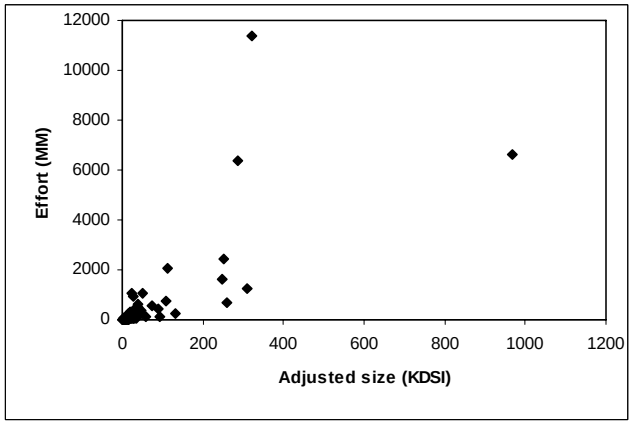

Figure 3 Effort against Size for COCOMO

4. DISCUSSION AND CONCLUSIONS

Overall, we can see numerous problems with many comparative cost estimation studies. Furthermore, in our opinion, the PROMISE repository is making matters worse by encouraging the use of the Desharnais and COCOMO datasets at the expense of other published datasets and publishing an invalid regression model along with the Desharnais data. We suggest that PROMISE might be better advised to:

1. Act as a portal by providing links to papers publishing datasets not as a data repository.

2. Provide links to best practice regression study results that can act as benchmarks for specific datasets.

3. Publicize known problems with existing datasets.

We believe there is need to establish guidelines for acceptable comparative cost estimation studies. In particular:

1. Researchers proposing the adoption of new estimation methods should at least confirm that they outperform statistical models.

2. Benchmark statistical models must be based on good statistical practice not simplistic application of statistical algorithms (see Section 3.1). Furthermore, although not discussed the issue in this paper, models should undergo sensitivity analysis.

3. Researchers should use appropriate statistical tests when alternative models are compared, and we believe the re-sampling approach has great potential value [16].

4. Researchers who want to investigate competing techniques rigorously should make a point of explaining how each of the competing techniques is applied (see for example [16], [19], or [23]).

5. Researchers should explicitly justify their selection of test datasets, particularly the use of old datasets.

6. When live datasets such as the ISBSG are used, researchers should make explicit how they selected the projects they used, both for the training and the validation datasets.

5. ACKNOWLEDGMENTS

Barbara Kitchenham's research is supported by UK Engineering and Physical Sciences Research Council project EP/E046983/1.

6. REFERENCES

[1] Albrecht, A.J. and Gaffney, J.E. (1983) Software function, sources lines of code and development effort: a software science validation. IEEE TSE, 6, pp 639-648.

[2] Aroba, J., Cuadrado-Gallego, J.J., Sicilia, M-Á., Ramos, I. and García-Barriocanal, E. (2008) Segmented software cost estimation models based on fuzzy clustering. JSS, 81 , pp 144-1950

[3] Boehm, B.W. (1981) Software Engineering Economics. Prentice-Hall.

[4] Chiu, N.-H. and Huan, S-J. (2007) The adjusted analogy-based software effort estimation based on similarity differences. JSS, 80, pp 628-640

[5] Desharnais, J-M. (1988) Analyse statistique de la producivité des projets de développment en informatique à partir de la technique des points de fonction. Program de maîtrise en informatique de gestion, Université du Québec à Montréal.

[6] Foss, T, Stensrud, E. Kitchenham, B. and Myrtveit, I. (2003) A Simulation Study of the Model Evaluation Criteria MMRE. IEEE Transactions on Software Engineering, 29(11), pp 985-995.

[7] Huang, S-J. and Chiu, N-H. (2006) Optimization of analogy weights by genetic algorithm for software effort estimation. Information and Software Technology, 48, pp 1034-1045.

[8] Kirsop, C. and Shepperd, M. (2002) Making inferences with small numbers of training sets. IEE Proceedings Software, 149 (5), pp 123-130.

[9] Kitchenham, B.A. (1998) A procedure for analysing unbalanced datasets, IEEE TSE, 24 (4), pp 278-301.

[10] Kitchenham, B., Mendes, E., and Travassos, G. (2007) Cross- vs. Within-Company Cost Estimation Studies: A Systematic Review, IEEE TSE 33(5), pp. 316-329.

[11] Kitchenham, B., Pfleeger, S.L., McColl, B. and Eagan, S. (2002) An empirical study of maintenance and development accuracy. JSS, 64, pp 57-77.

[12] Koch, S. and Mitlöhner, J. (2008) Software Project Effort Estimation with Voting Rules, Decision Support Systems, in press, doi:10.1016/j.dss.2008.12.002.

[13] Kumar, K.V., Ravi, V., Carr, M. And Kiran, N.R. (2008) Software development cost estimation using wavelet neural networks. JSS, 81, pp 1853-1867.

[14] Matson, J.E., Barrett. B.E., and Mellichamp, J.M. (1994) Software Development Cost Estimation using Function Points., IEEE TSE 20 (4), pp 275-287.

[15] Mair, C, Shepperd, M. and Jørgenesen, M. (2005) An analysis of Data Sets used to train and Validate Cost Estimation Prediction Systems, ACM SIGSOFT Software Engineering Notes, 30 (4), pp 1-6.

[16] Mittas, N. and Angelis, L. (2008) Comparing cost prediction models by resampling techniques. JSS, 81, pp 616-623.

[17] Park, H. and Baek, S. (2008) An empirical validation of a neural network model for software effort estimation. Expert systems and applications, 35, pp 929-937.

[18] Samson, Bill, Ellison, D and Duggard, P. (1997) Software Cost estimation using Albus perceptron (CMAC). IST, 30, pp 55-60.

[19] Sentas, P., Angelis, L., Stamelos, I., and Bleris, G. (2005) Software productivity and effort prediction with ordinal regression, IST, 47, pp 17-29.

[20] Shepperd, M. and Schofield, C. (1997) Estimating effort using analogies. IEEE TSE, 23 (11), pp 736-743.

[21] Song, Q., Shepperd, M. and Mair, C. (2005) Using Grey Relational Analysis to Predict Software Effort with Small Data sets. In: Proceedings of the 11th International Symposium on Software Metrics (Metrics 05).

[22] Stensrud, E., and Myrveit, I. (1998) Human performance estimating with analogy and regression models: An empirical validation. In: Proceedings Fifth International Software Metrics Symposium (Metrics '98). IEEE Computer Society Press, Los Alamitos, CA, pp. 205–213.

[23] Tronto, I. F. de B., da Sivla, J. D. S. and Sant'Anna, N. (2008) An investigation of artificial neural networks based on prediction systems in software management. JSS, 81 pp 35-367.