VOL. 13, NO. 7, APRIL 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

AN EMPIRICAL APPROACH TO PREDICT SOFTWARE DEVELOPMENT EFFORT USING LINEAR REGRESSION MODEL

M. Pramod Kumar[1], M. Babu Reddy[2] and A. Lakshmi Lavanya[3]
[1]Krishna University, SVEC, Tadepalligudem, Andhra Pradesh, India
[2]Department of Computer Science, KRISHNA University, Machilipatnam, Andhra Pradesh, India
[3]Department of Computer Science Engineering, SVEC, Tadepalligudem, Andhra Pradesh, India
E-Mail: mukthipudi@gmail.com

ABSTRACT

To develop qualitative software products in a given schedule it is necessary to estimate the software effort. The effort estimation is helpful to the project managers to determine effort required to complete qualitative software products. In this paper we proposed a novel linear regression model based on software size metric. We applied proposed regression model on 60 real time BPO projects. The proposed regression model considers a linear relationship between size and effort. To validate the accuracy of regression equation, we calculated R^2 value. We used MMER and PRED(x) to calculate error rate and made comparisons with standard models. The proposed model has shown much closed and better results against some standard estimation models.

Keywords: effort estimation, linear regression, KLOC, MMRE, PRED.

1. INTRODUCTION

The problems faced by project managers in managing and controlling software projects are overrun of effort estimation. It surely affects project designers to make correct decisions and leading to the failure of entire software project development. For the last decade there is a growth in demand for software development team to build quality application software in the competitive global markets. The total investment of application software development and maintenance has been overrun for the past ten years. As analyzed by [1] [2], the embellish effort schedule is varied substantially from 41% to 258%, and the total investment over-prediction is from 97% to 151%. Besides, there are researches conducted by US government agencies admit that 50% of these projects have been beat the estimated costs while in the case of 46% of those developed projects were useless, 60% of the software project development beats the original scheduled completion time [3]. The results indicated that not only the importance of planning and controlling over software projects, but also the early effort estimation of software project development is important too. The next section explains background and related research for our project. In section 3 describes the proposed methodology. Section 4 describes experimental results and section 5 describes conclusion.

2. BACKGROUND AND RELATED WORK

The below section describes the main concepts used in this paper which includes empirical parametric estimation models, linear regression and evaluation criteria.

2.1 Empirical estimation models

These models mainly rely on the experience gained on previously developed projects. They used size and effort value in the form of explicit function, by applying regression analysis methods like linear and exponential dependence. Generally effort is expressed in values such as man-hour or man-day.

Many empirical parametric models have been developed over time resulting in attempts to determine the best among them which would be stood standard. Once such model is determined, one should have sufficient valid data to be able to establish the relation between estimated value and independent parameters in the model. One of the popular known empirical parametric models developed hitherto is the form of

$$\text{Effort} = A\,(\text{LOC})\,B \quad (1)$$

Where Effort is in the form of man-months to implement the system and LOC is source line of code to be developed. This model had been investigated by Walston and Felix, Bailey and basil, Boehm (1981, 2004) [4] [5] as the basis for COCOMO model. Walston and Felix have reached the value for b<1 and other researchers have reached values>1.Kitchenham (1992), Kitchenham and Pearl Brereton [6] (2010) in his research concludes that in the prevailing number of cases the value of b is close enough to 1 which justifies consideration of introducing linear dependence between the number of code lines and effort. In order to verify this theory, Banker [7], Chang and Kemerer (1994), used model form:

$$\text{Effort} = a + b\text{LOC} + c\,\text{LOC}\,2 \quad (2)$$

The best known and most widely used metrics among them is function points metrics. A number of researchers, among them Albrecht and Gaffney (1983), (Kemerer 1987), Kemerer (1993) Matson, Barret and Mellichamp (1994), [8] [9] have examined models form:

$$\text{Effort} = a + b\,\text{FP} \quad (3)$$

Where Effort (man-months) is work required for the realization of the system and FP is the number of function points.

Basili and Panlilip Yap [10] (1985), (Basili *et al.,* 1996), suggest a model that would be based on counting the number of pages of the system documentation.

Effort = a + b Number of pages Documentation (4)

Brownlow (1994) [11] researches effort estimation model that can be applied on object-oriented system analysis and design. It is based on the number of objects and services of the system. The form of the model researched is following

Effort=a+b Number of objects +c Number of Services (5)

Conte, Dunsmore i Shen [12] (1989) introduce COPMO model, which also connects effort, value and number of personnel members. This model is based on a presumption that the total effort required for system development can be divided into individual team member effort plus effort required for coordination of their labor. The derived model has following form:

Effort = a + b LOC + Average Team member Number^d (6)

Where average team member number is calculated as quotient of effort and total project duration.

2.2. Linear regression model

Linear regression is a statistical procedure for predicting the value of a dependent variable from an independent variable when the relationship between the variables can be described with a linear model (aka predictors) (Allison, 1984).

A linear regression equation can be written as

$$Y_p = mx + b \quad (7)$$

Where $Y_p$the predicted value of the dependent variable, m is is the slope of the regression line, and b is the Y-intercept of the regression line.

2.3. Evaluation criteria

To assess the accuracy of the proposed estimation model, we have used the most common evaluation criteria.

$$MRE_i=\sum_{i=1}^{N} \frac{ABS(ACTUALEFFORT_i-ESTIMATEDEFFORT_i)}{ACTUALEFFORT_i} \quad (8)$$

MMRE can be achieved through the summation of MER over N estimated observations.

$$MMRE=\frac{1}{N}\sum_{i=1}^{N} MER_i \quad (9)$$

PRED(x): The prediction level used as a complementary criterion to MMER.PRED calculated the ratio of projects MMER that fails into the selected range(x) out of the total projects.

$$PRED(x) = (1-(\frac{\sum_{i=1}^{n} abs(A.EFFORT_i-P.EFFORT_i)}{N})) * 100 \quad (10)$$

Where N is total number of data items in the dataset.

3. REGRESSION MODEL

Linear regression is a statistical method that allows us to summarize and study relationships between two continues variables. One variable is denoted x, the predictor, explanatory or independent variable. The other variable, denoted *y* is regarded as the response, outcome, or dependent variable. Legendre (1805) and Guess (1809) were among the first persons who worked with regression models 200 years ago. Software professionals and project managers use historical data to build regression models. The simple linear regression model is

$$Y_i=\beta_0+ \beta_1*X_i+\varepsilon_i \quad (11)$$

When looking to summarize the relationship between a predictor *x* and a response *y*, we are interested in knowing the population regression

$$Y_i=\beta_0+ \beta_1*X_i \quad (12)$$

3.1. Proposed regression model

Proposed regression model was implemented using 60 BPO projects. We have taken project size as independent variable and effort as dependant variable. If the data is properly distributed the regression equation will be

Effort=β_0+β_1*KLOC (13)

Whereβ_0, β_1 are constants, KLOC means kilo source lines of code. Based on Figure-1 and Figure-2, it can be seen that the data is normally distributed. Thus the linear relationship between effort and size is

Effort = -4.8993+5.4104 * KLOC (14)

The equation (14) of regression model is applied based on the following steps.

Step 1: Collection of data with size and actual effort values
Step 2: perform logarithmic transformation of data if required
Step 3: Applying proposed linear regression model on data
Step 4: Performance evaluation

Collection of data: We have taken 60 BPO projects data from [13]. The data consists of two attributes

namely size in the form KLOC and Effort required to implementing every project.

Data transformation: If the data is not distributed properly, perform logarithmic transformations over data.

Apply regression model: Apply proposed linear regression equation on 60 BPO projects data to get proposed effort

Performance Evaluation: Performance of the proposed model is verified by MMRE and PRED(x) values obtained from test sample

4. EXPERIMENTAL DETAILS

In the proposed research we have collected data from [13] are used. In this table, every row contains details required to develop one project. First column represents project number, second column represents project size in the form of kilo lines of source code (KLOC) and actual effort required expressed in terms of persons-months. Statistical measures of data collected are depicted in Table-2. From this table it can be observed that the projects data sets are normally distributed based on the values of skewness and kurtosis. Hence logarithmic transformations are not needed.

A method was applied to measure the accuracy of the regression equation. For this purpose, the value of the coefficient of determination R^2 was measured. R^2 is the percentage of variation in Effort explained by the variable Size. An acceptable value of R^2 is ≥ 0.5 [14]. The value R^2 reported for the regression model in is 99.78. Approximately 99 % of the variation in Effort can be explained by the variable Size. We found that correlation coefficient is 99.89.This shows that there is strong correlation between Size and Effort.

Table-1. Statistical measures of 60 projects dataset.

Project type	Min	Max	Mean	Median	Standard deviation	Skewness	Kurtosis
60	2.5	20	8.59	7.4	18.27	0.893644	2.941347

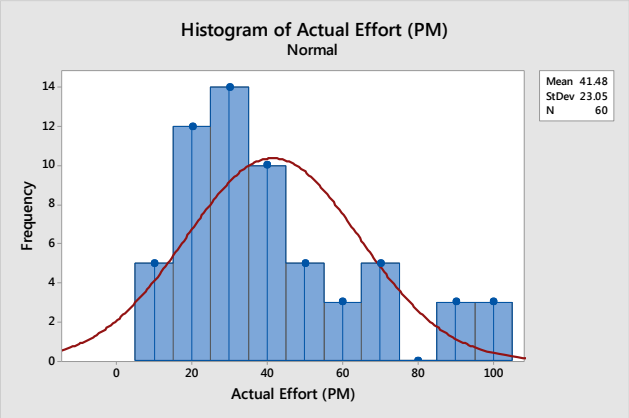

Figure-1. Histogram of actual effort values for 60 projects.

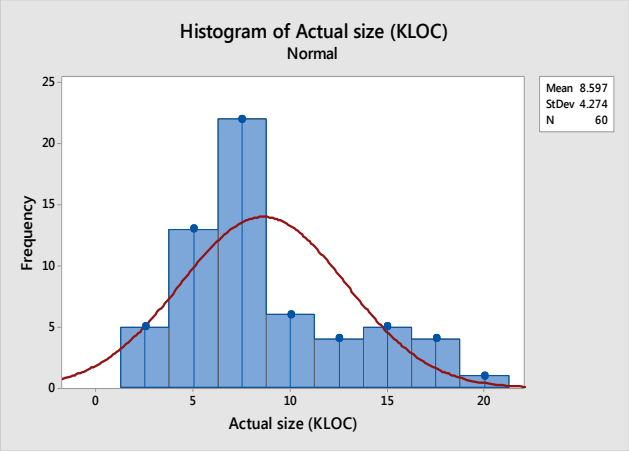

Figure-2. Histogram of size values for 60 projects.

Table-2. Estimation models used in related works.

Related Paper	Estimated model
COCOMO-II[15]	$Effort=2.9*kloc^{1.10}$
Bailey-Basil[16]	$Effort=5.5*kloc^{1.16}$
SEL[18,19]	$Effort=1.4*size^{0.93}$
Doty(for kloc>9)[17]	$Effort=5.288*kloc^{1.047}$
H Patra[13]	$Effort=EAF*2.9*(kloc)^{1.2}$

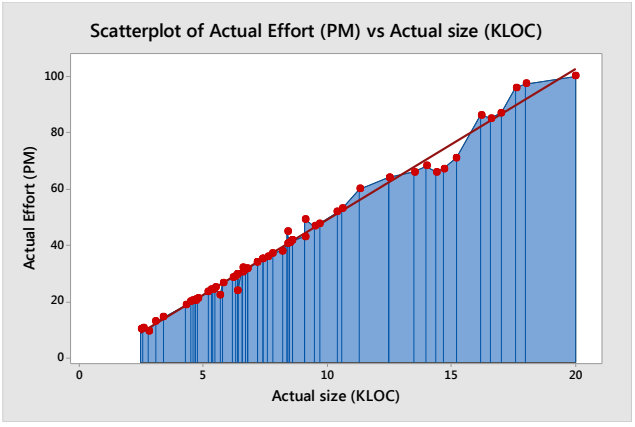

Figure-3. Scatter plot for actual size and actual effort.

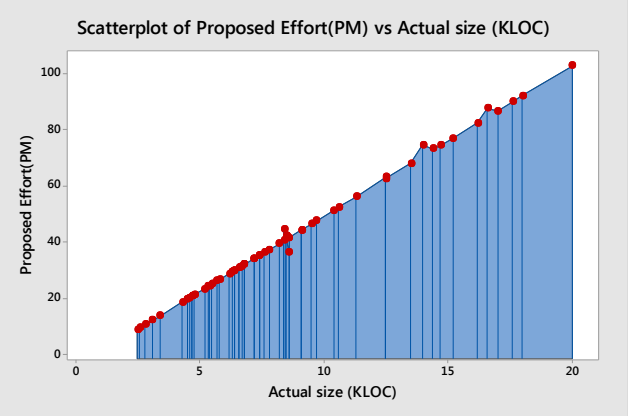

Figure-4. Scatter plot of actual size and proposed effort.

4.1. Model evaluation and discussion

This section presents the evaluation of regression model based on MMRE and PRED. The regression model was evaluated using 60 datasets the criteria uses MMRE, PRED (0.50), PRED (20) and PRED (10).The below table 6 shows evaluation values of proposed and existing models.

4.2. Comparison and analysis of results

On the basis of results obtained, the estimated effort using various efforts estimation models are compared. We compared our proposed regression model with standard effort estimation models shown in Table 3 on same data set. For estimation models shown in the Table-3, we calculated MMRE and PRED. Performance comparisons made in the form results. Table-4 shows MMRE and PRED values for different software effort estimation models over 60 BPO projects. It can be observed that the obtained results from the proposed model provides better prediction accuracy values than results obtained from estimation models given in Table-3.

Table-3. Comparisons of errors and prediction accuracy values using proposed regression model.

	Proposed	Bailey	SEL	COCOMO	Doty	H Patra
MMRE	0.035	0.633	0.737	0.237	0.249	0.069
PRED(50)	100	1.06	0	100	90	100
PRED(25)	100	0	0	63	61	100
PRED(10)	83	0	0	4	4	75

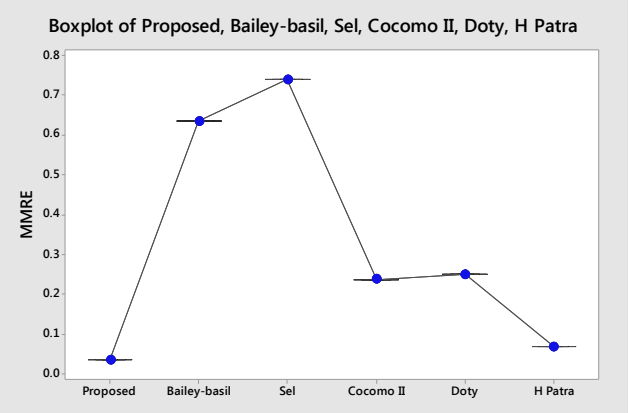

Figure-5. MMRE Comparisons of Estimation models.

5. CONCLUSIONS

In this paper, we proposed a linear regression model to estimate software effort. We have applied a linear regression model on 60 real time BPO data sets. We compared proposed regression model against existing effort estimation models like COCOMO, DOTY, SEL, BAILY-BASIL, H-PATRA. We used MMRE, PRED (.50) and PRED (.25) and PRED (.1) efficiency measures to assess the accuracy of proposed regression model. We observed that the proposed regression model has shown low MMRE and high PRED when compared to other models. As a result we believe that the proposed regression model can be used to estimate software effort. Future work will focus on to use the regression models with soft computing techniques on huge real time data set.

REFERENCES

[1] Norris K. P. 1971. The Accuracy of Project Cost and Duration Estimates in Industrial R&D. R&D Management. 2(1): 25-36.

[2] Murmann Philipp A. 1994. Expected Development Time Reductions in the German Mechanical Engineering Industry. Journal of Product innovation Management. 11: 236-252.

[3] David Consulting Group. 2012. Project Estimating. DCG Corporate Office, Paoli, 2007: http:davidconsultinggroup.com/training/estimation.aspx (January, 2017).

[4] M.PRAMOD KUMAR, M.BABUREDDY, A. Lakshmi Lavanya. A Survey Report on Software Effort Estimation models and ... - IRJCS.

[5] Barbara Kitchenham and Pearl Brereton. 2010. Problems Adopting Metrics from Other Disciplines, Workshop on Emerging Trends in Software Metrics. pp 1-7.

[6] Banker R. D., Chang H. and Kemerer C. F. 1994. Evidence of Economies of Scale in Software Development, Information and Software Technology. 36(5): 275-282.

[7] Albrecht A. J. and Gaffney J. E. 1983. Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation. IEEE Transactions of Soft. Engineering. 9(6): 639-648.

[8] Kemerer C.F. 1993. Reliability of Function Points Measurement: A Field Experiment. Communications of the ACM. pp. 85-97.

[9] Basili V.R. and Panlilio-Yap N.M. 1985. Finding Relatioships between Effort and Other Variables in the SEL, Proceedings of the IEEE Compsac, Chicago.

[10] Brownlow L. S. 1994. Early Estimating in the Object-Oriented Analysis and Design Environment, Proceedings of the European Software Cost Modelling Conference. Ivrea, Italy.

[11] Conte S. D., Dunsmore H. E. & Shen V. Y. 1989. Software engineering metrics and models. San. Francisco: Benjamin Cummings. Software engineering metrics and models.

[12] Patra H.P. & Rajniish K & Panda U.S. 2015. A new software cost estimation model for small software organizations: An empirical approach. International Journal of Applied Engineering Research. 10: 36076-36082.

[13] Humphrey W.S. 1995. A discipline for software engineering. Addison Wesley.

[14] Y. Singh, K.K. Aggarwal. Software Engineering Third edition, New Age International Publisher Limited New Delhi

[15] S. Devnani- Chulani. 1999. Bayesian Analysis of Software Cost and Quality Models. Faculty of the Graduate School, University Of Southern California.

[16] O. Benediktsson and D. Dalcher. 2003. Effort Estimation in incremental Software Development. IEEE Proc. Software. 150(6): 351-357.

[17] Yeong-Seok Seo, Kyung-A Yoon, Doo-Hwan Bae. 2008. An Empirical analysis of software effort estimation with outlier elimination. Proceedings of the 4th International workshop on Predictor models in Software Engineering, ACM 2008 (ISBN: 978-1-60558-036-4/08).

[18] Pichai Jodpimai, Peraphon. 2009. Advanced Virtual and Intelligent Computing Centre, Analysis of Effort Estimation based on Software Project Models. IEEE.