Statistical Analysis for Activity-Based Software Estimation using Regression Approach

Basavaraj M.J
Perot Systems, EPIP Phase II, Whitefield Industrial Area, Bangalore-560 066
basavaraj.m@ps.net;
basavarajmj@hotmail.com

Dr. K.C Shet
Professor, Computer Department, National Institute of Technology Karnataka, Surathkal
kcshet@nitk.ac.in; kcshet@yahoo.co.uk

ABSTRACT

Application Service Maintenance(ASM) projects mainly use Activity-Based software estimation methodology compared to Function Point or Lines of Code Estimation methodologies[1]. This is due to the nature of execution of ASM Projects differ with nature of execution of development projects. Activity based estimation methodology breaks estimation at each sub activities level of Software Development Life Cycle(SDLC) using work break down structure. Prediction of effort for each micro level activities of SDLC is really challenging one. Regression Analysis has been done for the data collected for some Enhancements of ASM projects. This paper explains how to predict the estimation for each micro level activities of SDLC by statistical analysis using Regression Approach.

Keywords : Software Estimation , Regression,

1. Introduction

ASM Projects generally support Enhancement & Routine maintenance activities. Enhancements are adding features to the existing application[1] and Routine maintenance activities involve supporting of Level-1, Level-2 and Level-3 activities. Estimating of Enhancements is really challenging, since it may not possible to apply full pledged FP or LOC as followed in development projects due to the complexities and interfaces with respect to the base application are occurring in ASM projects.

Many IT companies are adopting Work break down structure(WBS) for estimating efforts for Enhancement. Research challenge lies in how to achieve minimal effort variance for the Enhancements. Authors are addressing this issue by statistical analysis using Regression Approach.

2. Regression Analysis

Regression analysis verifies the dependence of a random variable on other independent variables or predictors[2]. Regression Equation results from mathematical model of their relationship between independent variables and predictors. In addition to the dependent and independent variables, the regression equations generally contain one or more unknown regression parameters (constants), which are estimated from given data. There are two types of regressions, one is linear regression for continuous responses and second one is non-linear regression for discrete responses.

3. Linear Regression Computation

Linear regression is a method of finding the linear equation that should be closest to fitting a collection of data points[3]. Computing of Regression Line[3] has been explained below.

The regression line (least squares line, best-fit line) associated with the points (x1, y1), (x2, y2),…., (xn, yn) is the line that gives the minimum sum-of-squares error (SSE). The regression line is

y = mx + b where m and b are computed as follows.

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{\Sigma y - m(\Sigma x)}{n}$$

"Σ" means "the sum of."

Σx = Sum of the x-values = $x_1 + x_2 + \ldots + x_n$

Σxy = Sum of products = $x_1y_1 + x_2y_2 + \ldots + x_ny_n$

Σx^2 = Sum of the squares of the x-values = $x_1^2 + x_2^2 + \ldots + x_n^2$

$(\Sigma x)^2$ = Square of Σx = Square of the sum of the x-values

n = Number of data points.

A residue is the difference between an observed and predicted value of a function. (A predicted value means a value given by some mathematical model.)

Residue = Observed value - Predicted value

The sum-of-squares error (SSE) when observed data are approximated by a function is given by

SSE = Sum of Squares of Residues
= Sum of $(y_{observed} - y_{predicted})^2$

The smaller SSE, the better the approximating function fits the data.
Coefficient of Correlation r has been calculated by

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{[n(\Sigma x^2) - (\Sigma x)^2]^{0.5}\,[n(\Sigma y^2) - (\Sigma y)^2]^{0.5}}$$

4. Coefficient of Determination

The coefficient of determination (denoted by r^2)[4] is a important output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- An r^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An r^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An r^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An r^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an r^2 of 0.20 means that 20 percent is predictable; and so on.

Coefficient of determination [5] is given by

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.
$\sigma_x = \mathrm{sqrt}\,[\, \Sigma (x_i - x)^2 / N \,]$
$\sigma_y = \mathrm{sqrt}\,[\, \Sigma (y_i - y)^2 / N \,]$

5. Activity Based Estimation – Regression Approach

We have collected a past data for 30 Enhancements which were already delivered from ASM project from “ABC” CMM Level5 Company for Regression Analysis Purpose. Due to maintaining the confidentiality of company name, data and client , exact names have not been disclosed. One of the Enhancement data collected out of those 30 Enhancements is listed below for illustration for Regression analysis approach.

Enhancement – Stage wise effort details			
Stages	Estimated Efforts[x]	Actual Efforts [y]	**Predicted Value**
Analysis & Query Resolution	90	90	**87.5887**
Design	35	30	**32.6449**
Coding	125	124	**122.553**
Testing	20	7	**17.6602**
Project Management	4	0	**1.67658**
Quality Assurance	4	0	**1.67658**
Reviews	12	9.5	**9.6684**
User Acceptance Testing	10	0	**7.67045**
Onsite Coordination	0	6	**-2.31933**
Estimate and SOW	0	10	**-2.31933**

Fig 1 : Enhancement data

Predicted values arrived from the equations mentioned above

y = 0.998978 x + -2.31933

Here m = 0.998978 and b = -2.31933

r = 0.987753

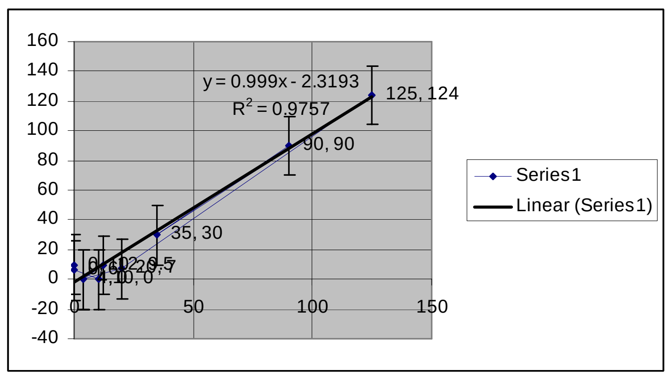

Fig 2 : Scatter Graph for Regression analysis

From the equation y = 0.998978 x + -2.31933, we can predict the actual efforts for any stage by knowing the value for estimated efforts for that respective stage.
Vertical Bars represent the difference of predicted efforts and actual efforts in the graph.

6 Conclusion

By using Regression analysis effectively, we can predict the efforts for actual specific component for any stage of SDLC by knowing the estimated efforts for that stage. We can do the same exercise for many enhancements across the different technologies within the ASM project and across the ASM projects. By doing the trend analysis for predicted actual efforts for each stage with respect to estimated efforts for that stage, we can achieve the minimal effort variance while estimating subsequent enhancements.

6. References

[1] Capers Jones, "Estimating Software Costs", Tata McGraw Hill, Edition 2005
[2] http://en.wikipedia.org/wiki/Regression_analysis#Introduction
[3] http://people.hofstra.edu/faculty/Stefan_Waner/RealWorld/tutorialsf0/frames1_5.html
[4] http://stattrek.com/Help/Glossary.aspx?Target=Coefficient%20of%20determination
[5] http://stattrek.com/AP-Statistics-1/Regression-Example.aspx
[6] Marc I. Kellner, Raymond J Madachy and David M. Raffo," Software Process Simulation Modelling : Why ? What? How? ", Journal of Systems and Software, Vol. 46, No. 2/3 (15 April 1999)
[7] Capers Jones, Assuring Productivity and Quality Applied Software Measurements
[8] Function Points Counting Practices Manual Release 4.1.1, IFPUG
[9] J. Brian Dreger ,"Function Point Analysis", Prentice Hall
[10] Roger S. Pressman, "Function Point Analysis – A Practitioner's Approach" , McGraw Hill
[11] Capers Jones, Assuring Productivity and Quality Applied Software Measurements