APPLICATION OF REGRESSION ANALYSIS IN SOFTWARE QUALITY ESTIMATION

Dr.Lakshmi Vishnu Murthy Tunuguntla, Assistant Professor,Institute of Management Technology, Hyderabad

Dr. Mu.Subrahmanian, HOD, Dept of Management Studies, Easwari Engg College, Chennai, India

Abstract

Estimation is one of the critical activities in the Software Project Management area. The predictability of quality becomes very vital as it impacts the estimates to be submitted to the customer during presales and in the subsequent stages of the product development lifecycle. This paper focuses on using Correlation analysis technique to estimate/predict the quality for software projects. The major objective of this paper is to understand the relationship between the effort spent in various phases and the defects that are generated at the end of each phase. The methodology consisted of major activities like defining the key projects, stakeholders of each of these projects, data requirements, data capturing mechanisms, Data validation, model building etc. The data is collected from the IT organizations in Hyderabad. While collecting the data, lot of challenges like lack of valid data, missing data, sometimes unwillingness to share the data, delays were encountered. However with right follow up and discussions, the data was gathered. After establishing the relationship, The relationship is quantified using regression method by looking at various options to fit the right regression line. Then the reliability of the regression line is estimated and the confidence limits are found. The information is shared with the stakeholders and it is tested with some of the projects to understand if the results are acceptable. The regression line established is periodically revisited to refine further that contributed to developing the Process capability limits satisfying "CMMI level IV" standards.

Introduction

Estimation is one of the critical activities in the Software Project Management area. The predictability of quality becomes very vital as it impacts the estimates to be submitted to the customer during presales and in the subsequent stages of the product development lifecycle. This paper focuses on using Correlation analysis technique to estimate/predict the quality for software projects.

A forecasting technique used to establish the relationship between quantifiable variables. In regression analysis, data on dependent and independent variables is plotted on a scatter graph or diagram, and trends are indicated through a line of best fit. The use of a single independent variable is known as simple regression analysis, while the use of two or more independent variables is called multiple regression analysis.

Samples of Previous Research:

The Correlation and regression analysis is widely used in the industry in various sectors like Financial. An empirical study was conducted to identify the impact of foreign investors funds on the BSE index. Similarly another application used this technique in the real estate sector to use standard appraisal approaches including the market comparison technique as well as the advantages and disadvantages of using multiple regression analysis. So this technique is used in multiple sectors and found very useful in the forecasting.

Objectives of the Study

To understand the relationship between the efforts spent in various phases' development and the defects and test the reliability of the relationship.

Methodology

- Identification of the Requirements
- Identification of the Target Projects
- Identification of the stakeholders of the initiative
- Identification of the data requirements
- Designing the Data Collection Mechanisms
- Piloting of the Data collection mechanisms

Data Collection

- Collection of data from the stakeholders
- Validation of Data
- Consolidation of Data
- Computation
- Computation of the Correlation coefficient and regression Curve

- Selection of the appropriate Regression Curve
- Estimating the Reliability of the Regression Curve
- Testing of the Regression Line

Observations

Consolidated data (Table I):

Project	Effort	Defects	Defects\100 hours of Effort
P1	12614	514	4.074837482
P2	28008	502	1.792345044
P3	1235	1052	85.18218623
P4	3575	121	3.384615385
P5	1231	9	0.731112916
P6	9042	129	1.426675514
P7	11380	200	1.757469244
P8	8232	164	1.992225462
P9	12130	247	2.036273702
P10	15456	250	1.617494824
P11	7136	153	2.144058296
P12	5234	98	1.872372946
P13	4145	83	2.002412545
P14	6754	126	1.865561149
P15	9345	164	1.754949171
P16	7564	132	1.745108408
P17	10245	159	1.551976574
P18	2356	43	1.825127334
P19	4248	68	1.600753296
P20	6353	85	1.337950575
P21	3852	71	1.843198339
P22	5469	98	1.791918084

Refined Data (Table II):

Project	Effort	Defects	Defects\100 hours of Effort
P1			4.074837482
P2	28008	502	1.792345044
🟥	🟥	🟥	85.18218623
P4	3575	121	3.384615385
P5	1231	9	0.731112916
P6	9042	129	1.426675514
P7	11380	200	1.757469244
P8	8232	164	1.992225462
P9	12130	247	2.036273702
P10	15456	250	1.617494824
P11	7136	153	2.144058296
P12	5234	98	1.872372946
P13	4145	83	2.002412545
P14	6754	126	1.865561149
P15	9345	164	1.754949171
P16	7564	132	1.745108408
P17	10245	159	1.551976574
P18	2356	43	1.825127334
P19	4248	68	1.600753296
P20	6353	85	1.337950575
P21	3852	71	1.843198339
P22	5469	98	1.791918084

The outliers are deleted from the rest of the population.

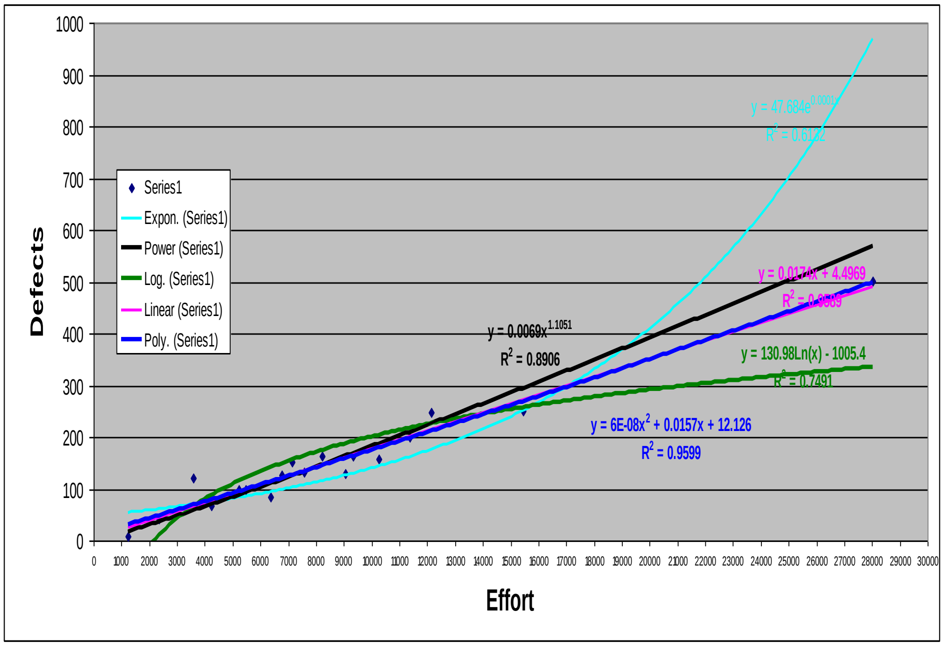
Vol 1 Issue 1 September 2013
www.eecmbajournal.in

Computation & Selection of the Regression Curve

With the help of above data, various types of curves have been tried and they are tabulated as follows:

Type of Curve	Regression Equation of the Curve	Correlation Coefficient
Exponential	$y = 47.684e^{0.0001x}$	$R^2= 0.6132$
Polynomial	$y = 6E\text{-}08x^2 + 0.0157x + 12.126$	$R^2= 0.9599$
Linear	$y = 0.0174x + 4.\ 4969$	$R^2= 0.9589$
Logarithmic	$y = 130.98Ln(x) – 1005.4$	$R^2= 0.7491$
Power	$y = y = 0.0069x^{1.1051}$	$R^2= 0.8906$

Analysis



Using the above data five types of curves are identified as described in the table above. The five curves are given different colors and the respective equation and the Correlation coefficient is given the same color. It is observed that out of all the five curves, the highest correlation coefficient is found for Polynomial equation i.e 95.99%.

Estimating the Reliability of the Regression Curve

Standard Error Computation

Project	Effort (x)	Defects (y)	Estimated Y using equation(defect estimate)	y-y^	(y-y^)^2	Standard error = Sqrt (∑ [(y-y^)^2]/no of data points- no of variables)	Std err *t
P1							
P2	28008	502	498.9184838	3.081516	9.495742	14.94477862	26.18325214
				0	0		
P4	3575	121	69.0203375	51.97966	2701.885		
P5	1231	9	31.54362166	-22.5436	508.2149		
P6	9042	129	158.9908658	-29.9909	899.452		
P7	11380	200	198.562264	1.437736	2.067085		
P8	8232	164	145.4343494	18.56565	344.6834		
P9	12130	247	211.395214	35.60479	1267.701		
P10	15456	250	269.1184762		0		
P11	7136	153	127.2165498	25.78345	664.7863		
P12	5234	98	95.94348536	2.056515	4.229252		
P13	4145	83	78.2333615	4.766639	22.72084		
P14	6754	126	120.900791	5.099209	26.00193		
P15	9345	164	164.0822415	-0.08224	0.006764		
P16	7564	132	134.3136458	-2.31365	5.352957		
P17	10245	159	179.2701015	-20.2701	410.877		
P18	2356	43	49.44824416	-6.44824	41.57985		
P19	4248	68	79.90233024	-11.9023	141.6655		
P20	6353	85	114.2897365	-29.2897	857.8887		
P21	3852	71	73.49267424	-2.49267	6.213425		
P22	5469	98	99.78389766	-1.7839	3.182291		
Total					3796.889		

The reliability is estimated as follows:

For a 90 % confidence level and 17 degrees of freedom the "t" value from Student T distribution is 1.752:

Upper Limit: Defect Estimate + Std error * t

Lower Limit: Defect Estimate - Std error * t

Results

For a project size of 10,245 hours, the defect estimated by the regression line is 179.27. The actual value is 159 and for a reliability level of 90%

Upper limit = 205.45, Lower Limit = 153

So the Regression line is predicting the values well within the limits.

References

1. A research project on impact of FII on capital market (An Empirical Study On Indian Capital Markets) By Purnendra
2. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation
3. Journal of Real Estate Practice and Education, 2004 by Benjamin, John D, Guttery, Randall S, Sirmans, C F