International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 7 (2018) pp. 5065-5068

Estimation of Goodness of Fit of the HGM Model and Comparing it with PARETO Type II Model

I.Siva Aryama[1], Dr. A. Srisaila[2], SK.Khasim Sherif [3], Dasari Akash[4]

Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Andhra Pradesh, India.

Abstract

As the usage of software is growing rapidly, accessing the software reliability is a critical task in development of a software system. So, many Software Reliability Growth Models (SRGM) are used in order to decide upon the reliable/unreliable of the developed software very quickly. The Hyper exponential growth model which contains two parameters is adopted for interval domain data based on Non Homogeneous Poisson Process (NHPP) which is used in assessing the reliability of developed software[1]. The parameters are estimated using Maximum Likelihood Estimator. In this project, we will compare the goodness of fit of our model with other models by applying some criteria like PRR.

Keywords: Maximum Likelihood Estimator, PRR.

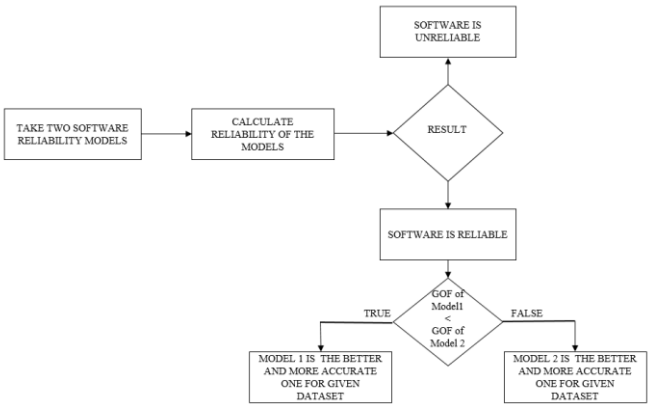
INTRODUCTION

Software Reliability is most important and most measurable aspect of software quality and it is very customer oriented. The user will likewise be benefited by software reliability measure, user is essentially worried about the failure free operation[2] of the system. Reliability is the probability of success or the probability that the system will perform its intended function under specified design limits. Its measurement and management technologies employed during the software lifecycle are essential for producing and maintaining quality/reliable software systems. Reliability is the likelihood that a product or part will work appropriately for a predetermined duration of time (design life) under the design operating conditions, (for example, temperature, volt, and so on.) without failure. Many Software Reliability Growth Models (SRGMs)[3] have been developed to extraordinarily encourage engineers in measuring the development of dependability as software is being enhanced and different statistical models have been proposed to get to the product software reliability.

PROBLEM STATEMENT

Software reliability growth models based on Non Homogeneous Poisson process (NHPP) seems to be most commonly used because of their simplicity. If the selected model does not fit the collected software testing data relatively well, we would expect a low prediction ability of this model and the decision-makings based on the analysis of this model would be far from what is considered to be optimal decision. Simply we estimate "how good the model is" by estimating the goodness of fit.

PROPOSED SYSTEM ARCHITECTURE



MODELS USED

Here we are comparing two models. They are Hyper Exponential Growth Model and Pareto Type II Model.

HYPER EXPONENTIAL GROWTH MODEL :

The Hyper Exponential Growth Model is the non-homogenous Poisson process (NHPP) based on software reliability growth model. At that point Non-homogeneous Poisson process (NHPP) based (SRGM) are turned out to be very effective in practical software reliability engineering. The main issue in the NHPP model is to focus a fitting mean value function to denote the expected number of failures in interval domain data. Model parameters can be assessed by utilizing maximum likelihood estimator (MLE)[4]. The hyper exponential growth model is based on the assumption that a program has a number of clusters of modules, each having a different initial number of errors and different failure rate[5].

$$m(t) = \sum_{i=1}^{n} a_i[1 - e^{-b_i t}]$$

Estimating the parameters:

The information are given for the cumulative number of identified errors in a given time interval $(0, t_i)$ where i= 1, 2. .., n and $0 < t1 < t2 < \ldots <$, then the log likelihood (LLF) takes on the following form :

$$\mathrm{LLF} = \sum_{i=1}^{n} (y_i - y_{i-1}) \log[(m(t_i)) - m(t_{i-1})] - m(t_n)$$

Substituting the value of m(t) in LLF and partial derivation of LLF w.r.t 'a' gives,

$$a_j = \frac{\sum_{i=1}^{n}(y_i - y_{i-1})}{(1 - e^{-b_j t_n})}$$

By substituting the value of 'a' in LLF

and partially differentiating with respect 'b' and equating to zero

$$g(b) = \sum_{i=1}^{n}(\mathrm{y_i} - \mathrm{y_{i-1}})[\sum_{j=1}^{n}[\frac{t_n e^{-b_j t_n}}{(1 - e^{-b_j t_n})} + \frac{-t_{i-1}e^{-b_j t_{i-1}} - t_i e^{-b_j t_i}}{e^{-b_j t_{i-1}} - e^{-b_j t_i}}]]$$

Again partially differentiating with respect 'b' and equating to '0'

$$g'(b) = \sum_{i=1}^{n}(\mathrm{y_i} - \mathrm{y_{i-1}})[\sum_{j=1}^{n}[\frac{-t_n^2 e^{-b_j t_n}}{(1 - e^{-b_j t_n})^2} + \frac{t_{i-1}^2 e^{-b_j t_{i-1}} - t_i^2 e^{-b_j t_i}}{e^{-b_j t_{i-1}} - e^{-b_j t_i}} - \frac{(t_i e^{-b_j t_i} - t_{i-1}e^{-b_j t_{i-1}})^2}{(e^{-b_j t_{i-1}} - e^{-b_j t_i})^2}]]$$

b value is estimated using Newton Raphson Method i.e.

$$\mathrm{b_{n+1}} = \mathrm{b_n} - \frac{\mathrm{g(b)}}{\mathrm{g'(b)}}$$

The values of 'b' in the above specified equations can be obtained using Newton Raphson Method. Solving the above equations simultaneously, yields the point estimates of the parameters a, b. These equations are to be solved iteratively and their solutions in turn when substituted in the log likelihood equation of 'a' would give analytical solution for the MLE of 'a'. The values of b are obtained by applying numerical methods.

PARETO TYPE II MODEL

In our present examination we study and evaluate the performance of the SRGM in view of NHPP with mean value function as given by

$$m(t) = a[1 - \frac{c^{\mathrm{b}}}{(t+c)^b}]$$

Estimating the parameters

$$\mathrm{Log\ L} = \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}})\log[(m(\mathrm{t_i})) - \mathrm{m(t_{i-1})}] - \mathrm{m(t_k)}$$

Substituting the value of m(t) in LLF and partial derivation of LLF w.r.t 'a' gives,

$$a = \sum_{i=1}^{k}(\mathrm{n_i} - \mathrm{n_{i-1}}).\frac{(t_k+c)^b}{(t_k+c)^b - \mathrm{c^b}}$$

Just like in HGM after partial differentiation w.r.t 'b' followed by 'c' we get,

$$g(b) = \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}})\,[0 + \log c - \log(t_{i-1} + c) - \log(t_i + c) + \frac{1}{[(t_i + c)^b - (t_{i-1} + c)^b]}[(t_i + c)^b . log(t_i + c) - (t_{i-1} + c)^b].\log(t_{i-1} + c)]] - 0 + a\left(\frac{c}{(t_k + c)}\right)^{\mathrm{b}} . log\frac{c}{(t_k + c)}\backslash$$

\

$$\mathrm{gI(b)} = \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}})\,.2(t_{i-1} + 1)^b](t_i + 1)^b]\log(\mathrm{t_i} + 1)\log\left[\frac{(\mathrm{t_{i-1}} + 1)}{\mathrm{t_i} + 1}\right] + \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}}).\log(\mathrm{t_k} + 1)$$

$$g(c) = \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}}).[\frac{\mathrm{b}}{\mathrm{c}} - \frac{\mathrm{b}}{(\mathrm{t_{i-1}}+\mathrm{c})} - \frac{\mathrm{b}}{(\mathrm{t_i}+\mathrm{c})} + \mathrm{b}.\frac{(t_i+c)^{b-1} - t_{i-1}+c)^{b-1}}{(t_i+c)^b - t_{i-1}+c)^b} + ab\frac{c}{(\mathrm{t_k}+\mathrm{c})}\frac{\mathrm{t_{k+c}} - \mathrm{c}}{(t_k+c)^2}$$.

$$g^I(c) = \sum_{i=1}^{k}(\mathrm{y_i} - \mathrm{y_{i-1}}).\left[-\frac{1}{\mathrm{c}^2} + \frac{1}{(\mathrm{t_{i-1}}+\mathrm{c})^2} + \frac{1}{(\mathrm{t_i}+\mathrm{c})^2}\right] - \sum_{\mathrm{i=1}}^{\mathrm{k}}(\mathrm{y_i} - \mathrm{y_{i-1}}).\frac{1}{(\mathrm{t_k}+\mathrm{c})^2}$$

The estimations of "b" and "c" in the above equations can be gotten utilizing Newton Raphson Method. Solving the above equations all the while yields the point assessments of the parameters b and c. These equations are to be explained iteratively and their answers in turn gives estimation of "a".

DESCRIPTION OF DATASETS

Phase I Dataset

Week Index	Exposure time (cum system test hours)(t_i)	Fault (f_i)	Cum. Fault (f_i)
1	356	1	1
2	712	0	1
3	1068	1	2
4	1424	1	3
5	1780	2	5
6	2136	0	5
7	2492	0	5
8	2848	3	8
9	3204	1	9
10	3560	2	11
11	3916	2	13
12	4272	2	15
13	4628	4	19
14	4984	0	19
15	5340	3	22
16	5696	0	22
17	6052	1	23
18	6408	1	24
19	6764	0	24
20	7120	0	24
21	7476	2	26

Phase 1 test data

Phase II Dataset

Week Index	Exposure Time (t_i)	Fault (f_i)	Cum. Fault (f_i)
1	416	3	3
2	832	1	4
3	1248	0	4
4	1664	3	7
5	2080	2	9
6	2496	0	9
7	2912	1	10
8	3328	3	13
9	3744	4	17
10	4160	2	19
11	4576	4	23
12	4992	2	25
13	5408	5	30
14	5824	2	32
15	6240	4	36
16	6656	1	37
17	7072	2	39
18	7488	0	39
19	7904	0	39
20	8320	3	42
21	8736	1	43

Phase 2 test data

RELIABILITY CALCULATION

$$R(S/X) = e^{-[(m(s+x)-m(s)]}$$

Where,

S -Time between Errors, X- Cumulative Time

HGM

DATASETS	a	b
PHASE I	62.1517	0.0233
PHASE II	1.2235	0.0052

$$m(t) = \sum_{i=1}^{n} a_i[1 - e^{-b_i t}]$$

Substitute a, b values of each dataset in m(t) and calculate the reliability

DATASETS	S	X	Reliability
PHASE I	356	7476	0.513
PHASE II	416	8736	0.8687

Pareto Type II Model

Software Product	Estimate 'a'	Estimate 'b'	Estimate 'c'
PHASE I	37.120867	0.962019	3396.7586
PHASE II	59.398002	0.961882	3969.2460

$$m(t) = a[1 - \frac{c^{b}}{(t+c)^{b}}]$$

Substitute a, b, c values of each dataset in m(t) and calculate Reliability

Data Sets	S	X	Reliability
PHASE I	7476	2848	0.087978
PHASE II	8736	2080	0.071912

GOODNESS OF FIT

The goodness of fit of a statistical model describes how well it fits a set of observations. If the selected model does not fit the collected software testing data relatively well, we would expect a low prediction ability of this model and the decision-makings based on the analysis of this model would be far from what is considered to be optimal decision.

The **Predictive-Ratio Risk (PRR)** represents the distance of the model estimates from the actual data against the model estimates and is defined as

$$\text{PRR} = \sum_{i=1}^{n} \left(\frac{\widehat{m}(t_i) - y_i}{\widehat{m}(t_i)}\right)^2$$

Where m(t) represents the estimated expected number of faults detected by time t; yi represents the observation value;

For the goodness-of-fit criteria described above, the smaller the value, the better is the goodness of fit for the software reliability model.

6.1 GOODNESS OF FIT CALCULATION

Data Sets	HGM	PARETO MODEL
PHASE I	13.094918	4.561045
PHASE II	9.318	3.091478

So, by the above two comparisons it can be concluded that PARETO TYPE II MODEL better fits for the given data than HGM MODEL as the PRR values of PARETO MODEL are lower than HGM MODEL.

CONCLUSION

Software Reliability is a critical and most measurable aspect of software quality and it is very customer oriented. Many Software Reliability Growth Models (SRGMs) have been developed to extraordinarily encourage engineers in measuring the development of dependability as software is being enhanced and different statistical models have been proposed to get to the product software reliability. So, we have taken two SRGM's i.e. the HGM Model and the PARETO TYPE II Model to apply them on two given datasets and find out the reliability of the given software dataset. We got two different sets of values of reliability for the given datasets and therefore to exactly know which model is the best suited one for given software datasets we estimated the Goodness of Fit values of both the models on the given datasets and therefore concluded that PARETO TYPE II Model is better fitted model than HGM Model as it has lower Goodness of fit values than HGM Model.

REFERENCES

[1] Agresti, A (1990) Categorical Data Analysis. Wiley, New York , ANSI/IEEE, (1991) "Standard Glossary of Software Engineering Terminology", STD-729 ANSI/IEEE.

[2] Kimura, M., Yamada, S., Osaki, S.,"Statistical Software reliability prediction and its applicability based on mean time between failures". Mathematical and Computer Modeling Volume 22, Issues 10-12, 1995. 149-155.Koutras, M.V., Bersimis, S., Maravelakis,P.E., "Statistical .

[3] Koutras, M.V., Bersimis, S., Maravelakis,P.E., "Statistical process control using shewart control charts with supplementary Runs rules" Springer Science + Business media 9: 2007. 207-224.

[4] MacGregor, J.F., Kourti, T., "Statistical process control of multivariate processes". Control Engineering Practice Volume 3, Issue 3, March 1995, 403-414.

[5] Anderson T, Lee P (1980) Fault Tolerance: Principles and Practices, *Prentice-Hall*, Englewood Cliffs.