Two Step Approach For Software Reliability: Rayleigh

Mrs P. Padmaja,
A.S.R.I.T. College,
Dept. of CS&Engg.
West Godavari.

Dr. G. Krishna Mohan,
Dept. of CS& Engg,
KL University.

Dr. R. Satya Prasad
Dept. of CS&Engg,
Acharya Nagarjuna university.

***Abstract*—Software Reliability Growth Model is a mathematical model of how the software reliability improves as faults are detected and repaired. The performance of SRGM is judged by its ability to fit the software failure data. How good does a mathematical model fit to the data and reliability of software is presented in the current paper. The model under consideration is the, Rayleigh model. A Two step approach is used to estimate the model parameters by combination of Maximum Likelihood Estimation and regression. To assess the performance of the considered Software Reliability Growth Model, we have carried out the parameter estimation on the real software failure data sets.**

***Keywords*—Rayleigh model, Maximum Likelihood Estimation,Least Squares Regression, SRGM, Goodness Of Fit**

I. INTRODUCTION

Software reliability is defined as the probability of failure-free software operation for a specified period of time in a specified environment[1]. Software Reliability Growth Model (SRGM) is a mathematical model of how the software reliability improves as faults are detected and repaired [2]. Among all SRGMs developed so far a large family of stochastic reliability models based on a non-homogeneous poisson process known as NHPP reliability models, has been widely used. Some of them depict exponential growth while others show S-shaped growth depending on nature of growth phenomenon during testing. The success of mathematical modelling approach to reliability evaluation depends heavily upon quality of failure data collected.

However, a problem is the model validation and selection. 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 [3]. The present paper presents a method for model validation.

II. LITERATURE SURVEY

A. NHPP Models

The NHPP group of models provides an analytical framework for describing the software failure phenomenon during testing. They are proved to be quite successful in practical software reliability engineering [4]. They have been built upon various assumptions. If ‘t’ is a continuous random variable with probability density function: $f(t,\theta_1,\theta_2,\ldots,\theta_k)$, and cumulative distribution function: $F(t)$ .where $\theta_1,\theta_2,\ldots,\theta_k$ are k unknown constant parameters. The mathematical relationship between the pdf and cdf is given as: $f(t)=F'(t)$.

Let $N(t)$ be the cumulative number of software failures by time ‘t’. A non-negative integer-valued stochastic process $N(t)$ is called a counting process, if $N(t)$ represents the total number of occurrences of an event in the time interval [0, t] and satisfies these two properties:

1. If $t_1 < t_2$, then $N(t_1) \le N(t_2)$
2. If $t_1 < t_2$, then $N(t_2)-N(t_1)$ is the number of occurrences of the event in the interval $[t_1,t_2]$.

One of the most important counting processes is the Poisson process. A counting process, $N(t)$, is said to be a Poisson process with intensity λ if

1. The initial condition is N(0) = 0
2. The failure process, N(t), has independent increments
3. The number of failures in any time interval of length s has a Poisson distribution with mean λs , that is,

$$P\{N(t+s)-N(t)=n\}=\frac{e^{-\lambda s}(\lambda s)^n}{n!}$$

Describing uncertainty about an infinite collection of random variables one for each value of ‘t’ is called a stochastic counting process denoted by $[N(t),t\ge 0]$. The process $\{N(t),t\ge 0\}$ is assumed to follow a Poisson distribution with characteristic Mean Value Function $m(t)$, representing the expected number of software failures by time ‘t’. Different models can be obtained by using different non decreasing $m(t)$. The derivative of $m(t)$ is called the failure intensity function $\lambda(t)$.

A Poisson process model for describing about the number of software failures in a given time (0, t) is given by the probability equation.

D.

$$P[N(t)=y]=\frac{e^{-m(t)}[m(t)]^{y}}{y!}, \qquad y=0,1,2,...$$

E.

Where, $m(t)$ is a finite valued non negative and non decreasing function of $'t'$ called the mean value function. Such a probability model for $N(t)$ is said to be an NHPP model. The mean value function $m(t)$ is the characteristic of the NHPP model.

The NHPP models are further classified into Finite and Infinite failure models. Let 'a' denote the expected number of faults that would be detected given infinite testing time in case of finite failure NHPP models. Then, the mean value function of the finite failure NHPP models can be written as: $m(t)=aF(t)$. The failure intensity function $\lambda(t)$ is given by: $\lambda(t)=aF'(t)$ [5].

B. SRGM

SRGMs are a statistical interpolation of defect detection data by mathematical functions [6]. They have been grouped into two classes of models- Concave and S-shaped. The only way to verify and validate the software is by testing. This involves running the software and checking for unexpected behaviour of the software output [7]. SRGMs are used to estimate the reliability of a software product. In literature, we have several SRGMs developed to monitor the reliability growth during the testing phase of the software development.

C. Model Description: Rayleigh

In recent years the Weibull distribution [8] has become more popular as a reliability function. It is named after the Swedish scientist WaloddiWeibull. The Weibull distribution has a position of importance in the field of reliability and life testing because of its versatility in fitting time-to-failure distributions. Many researchers considered the distribution and worked on it. Some of them are [9], [10], [11]. The three parameters of the Weibull distribution are θ, β and γ. Where θ and β are known as the scale, shape parameters and γ is known as the location parameter. These parameters are always positive. It is probably the most widely used family of failure distributions, mainly because by proper choice of its shape parameter β, it can be used as an Increasing Failure Rate for β > 1 , Decreasing Failure Rate for β < 1, or Constant Failure Rate for β = 1. The Weibull distribution is called Rayleigh distribution at β = 2, γ = 0, and Exponential distribution at β = 1, γ = 0. Software reliability is defined as the probability of failure-free software operation for specified period of time 't' in a specified environment,

$$R(t)=e^{-\left(m(t_i)-m(t_{i-1})\right)}. \quad (1)$$

D. Goodness-of-fit

Model comparison and selection are the most common problems of statistical practice, with numerous procedures for choosing among a set of models proposed in the literature. Goodness-of-fit tests for this process have been proposed by [12]. The AIC is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. AIC deals with the tradeoff between the goodness of fit of the model and the complexity of the model.

$$AIC=-2*L+2*k \quad (2)$$

Where 'k' is the number of parameters in the statistical model, and 'L' is the maximized value of the likelihood function for the estimated model.

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over-fitting.

III. TWO STEP APPROACH FOR PARAMETER ESTIMATION

The main issue in the NHPP model is to determine an appropriate mean value function to denote the expected number of failures experienced up to a certain time point. Method of least squares (LSE) or maximum likelihood (MLE) has been suggested and widely used for estimation of parameters of mathematical models [13]. Non-linear regression is a method of finding a nonlinear model of the relationship between the dependent variable and a set of independent variables. Unlike traditional linear regression, which is restricted to estimating linear models, nonlinear regression can estimate models with arbitrary relationships between independent and dependent variables. The model proposed in this paper is a non-linear and it is difficult to find solution for nonlinear models using simple Least Square method. Therefore, the model has been transformed from non linear to linear. MLE and LSE techniques are used to estimate the model parameters [1], [4]. Sometimes, the likelihood equations are difficult to solve explicitly. In such cases, the parameters are estimated with some numerical iterative methods (Newton Raphson method). On the other hand, LSE, like MLE, applied for small sample sizes and may provide better estimates [14].

A. Algorithm for a 2-step approach of parameter estimation and data as best fit.

- Consider the Cumulative distribution function $F(t)$ and equate to p_i, i.e $F(t) = p_i$,

  where $p_i = \dfrac{i}{n+1}$
- Express the equated equation $F(t) = p_i$ as a linear form, $y = mx + b$.
- Find model parameters of mean value function $m(t)$. Where $m(t) = aF(t)$
  - The initial number of faults $\hat{a}$ is estimated through MLE method. Since, it forms a closed solution.
  - The remaining parameters are estimated through LSE regression approach.
- Find the failure intensity function $\lambda(t) = aF'(t)$
- Find likelihood function L
- Find the Log likelihood function log L. (Which comes to be –ve value.)
- The distribution model with the highest –ve value is the best fit.

B. ML (Maximum Likelihood) Estimation

The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability of the sample data. The method of maximum likelihood is considered to be more robust and yields estimators with good statistical properties. In other words, MLE methods are versatile and apply to many models and to different types of data. Although the methodology for MLE is simple, the implementation is mathematically intense. Using today's computer power, however, mathematical complexity is not a big obstacle. If we conduct an experiment and obtain N independent observations, $t_1, t_2, \ldots, t_N$. The likelihood function [15] may be given by the following product:

$$L(t_1, t_2, \ldots, t_N \mid \theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{N} f(t_i; \theta_1, \theta_2, \ldots, \theta_k) \qquad (3)$$

Likelihood function by using λ(t) is:

$$L = e^{-m(t_n)} \prod_{i=1}^{n} \lambda(t_i) \qquad (4)$$

Log Likelihood function for ungrouped data [5]

$$\log L = \sum_{i=1}^{n} \log[\lambda(t_i)] - m(t_n) \qquad (5)$$

The maximum likelihood estimators (MLE) of $\theta_1, \theta_2, \ldots, \theta_k$ are obtained by maximizing L or Λ, where Λ is ln L . By maximizing Λ, which is much easier to work with than L, the maximum likelihood estimators (MLE) of $\theta_1, \theta_2, \ldots, \theta_k$ are the simultaneous solutions of k equations such as: $\dfrac{\partial(\Lambda)}{\partial \theta_j} = 0$, j=1,2,…,k.The parameters 'a' and 'b' are estimated as follows. The parameter 'b' is estimated by iterative Newton Raphson Method, which is substituted in finding 'a'.

C. LS (Least Square) estimation

LSE is a popular technique and widely used in many fields for function fit and parameter estimation [16]. The least squares method finds values of the parameters such that the sum of the squares of the difference between the fitting function and the experimental data is minimized. Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X.

The Least Squares Regression Line

Linear regression finds the straight line, called the least squares regression line that best represents observations in a bivariate data set. Given a random sample of observations, the population regression line is estimated by: $\hat{y} = bx + a$. where ,'a' is a constant, 'b' is the regression coefficient and 'x' is the value of the independent variable, and '$\hat{y}$' is the predicted value of the dependent variable. The least square method defines the estimate of these parameters as the values which minimize the sum of the squares between the measurements and the model. Which amounts to minimizing the expression: $E = \sum_i (Y_i - \hat{Y}_i)^2$.

Taking the derivative of E with respect to 'a' and 'b' and equating them to zero gives the following set of equations (called the normal equations):

$$\frac{\partial E}{\partial a} = 2Na + 2b\sum X_i - 2\sum Y_i = 0, \text{ and}$$

$$\frac{\partial E}{\partial b} = 2b\sum X_i^2 + 2a\sum X_i - 2\sum Y_i X_i = 0$$

The solutions of 'a' and 'b' are obtained by solving the above equations. Where,

$$a = \bar{Y} - b\bar{X} \quad \text{and} \qquad (6)$$

$$b = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}. \qquad (7)$$

IV. Illustration.

A. Procedure to find parameter ‘a’ using MLE.

The likelihood function of Rayleigh is given as, $LogL = \log\left[e^{-a\left[1-e^{-(bt_n)^2}\right]} \prod_{i=1}^{n} 2ab^2 t_i e^{-(bt)^2} \right]$ (8)

Taking the natural logarithm on both sides, The Log Likelihood function is given as:

$$LogL = \sum_{i=1}^{n} \log(2ab^2 t_i e^{-(bt)^2}) - a[1-e^{-(bt_n)^2}] \quad (9)$$

Taking the Partial derivative with respect to ‘a’ and equating to ‘0’.

$$a = \frac{n}{\left[1-e^{-(bt_n)^2}\right]} \quad (10)$$

Taking the Partial derivative with respect to ‘b’ and equating to‘0’.

$$\therefore g(b) = \frac{2n}{b} - 2b\sum_{i=1}^{n} t_i^2 - \frac{2.n.b.t_n^2.e^{-(bt_n)^2}}{\left(1-e^{-(bt_n)^2}\right)} = 0 \quad (11)$$

Taking the partial derivative again with respect to ‘b’ and equating to ‘0’.

$$g'(b) = 2n\left(\frac{-1}{b^2}\right) - 2\sum_{i=1}^{n} t_i^2 - 2nt_n^2 \left\{ \frac{e^{-(bt_n)^2}}{\left(1-e^{-(bt_n)^2}\right)} - \frac{2b^2 t_n^2.e^{-(bt_n)^2}}{\left(1-e^{-(bt_n)^2}\right)^2} \right\} \quad (12)$$

B. LS Estimation

Procedure to find parameter ‘b’ using regression approach.

- The cumulative distribution function of Rayleigh is, $F(t) = 1 - e^{-\left(\frac{x_i}{\sigma}\right)^2}$. The c.d.f is equated to p_i. Where, $p_i = \frac{i}{n+1}$.
- The equation $F(t) = p_i$ is expressed as a linear form, $Y_i = C.X_i + D$. Where, $Y_i = \log(-\log(1-p_i))$; $X_i = 2.\log(x_i)$ and $D = -2\log\sigma$

$$\hat{C} = \frac{\sum X_i Y_i - n\overline{Y}\,\overline{X}}{\sum X_i^2 - n\overline{X}^2}; \qquad \hat{D} = \overline{Y} - \hat{C}\,\overline{X};$$

$$\sigma = e^{-\hat{D}/\hat{C}}$$

Where, ‘ $1/\sigma$ ’ is nothing but the parameter ‘b’ estimated through regression approach.

TABLE I. Estimation of parameters and log likelihood for, Lyu data.

n=	30
σ =	7.830927
$\hat{b}$ =	0.127698
$\hat{a}$ =	24.086392
$m(t_n)$=	20.362905
$Log\,L$=	-11.047502

TABLE II. Parameters estimated through MLE and Regression

Data Set (no of observations)	Parameters		
	MLE		Regression
	$\hat{a}$	$\hat{b}$	$\hat{b}$
Xie (30)	30.051592	0.003416	0.008178
NTDS (26)	28.851930	0.011827	0.031848
AT&T (22)	23.719656	0.004824	0.003074
SONATA (30)	31.961497	0.000912	0.000345
IBM (15)	19.164356	0.00711	0.003402

The parameters in the table 2 are estimated in two ways. The parameter $\hat{a}$ is estimated through MLE only. Where as, the parameter $\hat{b}$ is estimated once using MLE and and the other one using regression approach.

V. Method of Performance Analysis

The performance of SRGM is judged by its ability to fit the software failure data. The term goodness of fit denotes the question of “How good does a mathematical model fit to the data?”.Inorder to validate the model under study and to assess its performance, experiments on a set of actual software failure data have been performed. The considered model fits more to the data set whose Log Likelihood is most negative. The application of the considered distribution function and its log likelihood on different data sets collected from real world failure data using both the approaches is given as below.

TABLE III. Log likelihood on different data sets.

Data Set (no of observations)	Log L (MLE)	Log L (Two step)	AIC (Two step)	R(25/t) (Two step)
Xie (30)	- 131.125396	- 128.663198	261.326397	0.170701
NTDS (26)	-90.462571	-72.342582	148.685165	0.993564
AT&T (22)	- 113.484963	- 121.244882	246.489764	0.269103
SONATA (30)	- 157.979313	- 233.444039	470.888079	0.977816
IBM (15)	-63.122147	-72.829640	149.659279	0.502120

The Log Likelihood of the MLE approach on data sets Xie and NTDS are more negative when compared to the Log Likelihood of the Two step approach. Where as , for the data sets of AT&T, SONATA and IBM are more negative in the two step approach when compared to the MLE approach. We prefer to use two step approach when compared to the pure MLE approach to find the reliability of the models over the considered data sets. Since, in many of the cases two step approach is exhibiting most negative log likelihood.

VI. CONCLUSION

To validate the proposed approach, the parameter estimation is carried out on the data sets collected from [1], [17], [5], [18]. Parameters of the model are estimated by MLE and the linear regression least squares method using cumulative failure data against time. Out of the data sets that were considered, the model under consideration best fits the data of SONATA using two step approach. Since, it is having the highest negative value for the log likelihood. The reliability of the model using two step method for various data sets are given in table 3. The reliability of the model over NTDS data after 25 units of time is high among the data sets which were considered.

REFERENCES

[1] Lyu, M.R., Handbook of Software Reliability Engineering, McGraw-Hill, New York. 1996.

[2] Quadri, S.M.K and Ahmad, N. Software Reliability Growth modelling with new modified Weibull testing-effort and optimal release policy, International Journal of Computer Applications, 2010; Vol.6, No.12.

[3] Xie, M., Yang, B. Gaudoin, O. Regression goodness-of-fit Test for Software Reliability Model Validation, ISSRE and Chillarege Corp, 2001.

[4] Musa, J.D., Iannino, A., Okumoto, K. Software Reliability: Measurement, Prediction, Application, New York: McGraw-Hill, 1987.

[5] Pham. H., System software reliability, Springer, 2006.

[6] Wood, A., Software Reliability Growth Models, Tandem Computers, Technical report 96.1, 1996.

[7] Kapur, P.K., Sunil kumar, K., Prashant, J. Ompal, S. Incorporating concept of two types of imperfect debugging for developing flexible software reliability growth model in distributed development environment, Journal of Technology and Engineering sciences, 2009; Vol.1, No.1; Jan-Jun.

[8] Weibull, W. A Statistical distribution function of wide Applicability, Journal of Applied Mechanics, 1951; 18; 293-297.

[9] Kao, J. H. K. Computer Methods for estimating Weibull parameters in Reliability Studies, Transactions of IRE-Reliability and Quality Control, 1958; 13; 15-22.

[10] Dubey, Satya D. On some statistical inferences for Weibull laws, J. Amer. Statist. Assoc., 1963; 58; 549.

[11] Menon, M.V. Estimation of the shape and scale parameters of the Weibull distribution, Technometrics, 1963; 5; 175-182.

[12] Rigdon, S. E. and Basu, A. P. The Power Law Process: a Model for the Reliability of Repairable Systems. Journal of Quality Technology, 1989; 21(4); pp.251- 259.

[13] Kapur, P.K., Gupta, D., Gupta, A. Jha, P.C. Effect of Introduction of Fault and Imperfect Debugging on Release Time, Ratio Mathematica, 2008; 18; pp. 62-90.

[14] Huang, C.Y, Kuo, S.Y. Analysis of incorporating logistic testing effort function into software reliability modelling, IEEE Transactions on Reliability, 2002; Vol.51, No. 3; pp. 261-270.

[15] Pham. H. Handbook Of Reliability Engineering, Springer, 2003.

[16] Liu, J., Function based Nonlinear Least Squares and application to Jelinski-Moranda Software Reliability Model, stat. ME, 25th August, 2011.

[17] Xie, M., Goh. T.N., Ranjan.P., “Some effective control chart procedures for reliability monitoring” - Reliability engineering and System Safety, 2002; 77; 143 -150.

[18] Ashoka. M. Sonata Software Limited Data Set, Bangalore, 2010.