Model Prediction using Correlation In Contrast With QQ-Plot

Kotte Sandeep, R. Satya Prasad

Abstract: ***In this paper, we mainly focused on developing reliable, efficient and error free software products by following practices. The software which follows proper testing techniques has less failure rate. To analyze the information or data, it should be understood first that is better to be done with the best suited model. Selection of the best suited model isn't a simple task, and there are different methods in it. One such method is qq-plot, however, it's not a quantified measure and time overwhelming process. We proposed a qq-plot and quantified measure "correlation factor" to demonstrate its use to choose the best fit model for information among different models being referred to.***

Index Terms: Correlation factor, QQ-Plot, Best fit, LPETM, MLE, HPP, NHPP

I. INTRODUCTION

Engineering the info for work its behavior is a sort of standard task in multiple fields like analytics, production and quality control, etc. Information attached enumeration processes extensively converted to prevalence of events based mostly on the frequency of your time intervals. The foremost distinguished method attached enumeration is Poisson method. Poisson model the info in question are often investigated to look at its pattern and nature once it's fitted into the most effective suited obtainable model. To create a distribution from the info, initially the most effective suited model should be selected.

II. QQ-PLOT AND ITS LIMITATIONS

The qq plot is a graphical approach for deciding whether or not the given two sets of quantiles follow common distribution. If each set of quantiles came from a similar distribution, we should always see the points forming a line that is roughly straight.

Generally involve investigating of events inside a notable distinct time intervals. The Poisson models supported frequency of your time intervals are classified into Homogenous Poisson process (HPP) and Non-Homogenous Poisson process (NHPP) [1]. HPP models wherever time interval for the prevalence of targeted events is same and NHPP has distinct time intervals for the prevalence of events [10]. Here to decide the distribution which best suits the data, multiple qq-plots must be plotted and these-plots now act as probability plots. We replace the first dataset with the quantiles of a theoretical distribution. The graphical plotting is visually good and clearly provides the main points to choose whether or not the distribution model most accurately fits or not, once the linear difference in distribution is significantly visually clear and recognizable.

In Table 1, we treated each point its own quantile and add the same number of quantiles to the normal curve as you created for the data. Suppose if there are 25 lines dividing the data into equal sized groups. For the normal curve, "equal sized groups" means that there is an equal probability of observing a value within each group. This means that groups on the edge must be wider to account for the lower probability of observing a value out there and groups in the middle are narrower, since there is a higher probability of observing a value in that region[11]. Now see how well the dots fit a straight line. If the data were normally distributed, most of the points would be on the line. In Table 1: First we need to sort sample values of x in ascending order and compute the rank proportion of each value using prob_i=(i-0.5)/n. In the fourth column theoretical Z score based on prob_i with the NORMSINV (prob_i) function. In the fifth column we compute standardized values of data. In Figure 1, we plotted first data set with the quantiles of a theoretical distribution

Table I: QQ-Plot Quantiles, Percentiles and Standardized Values [6].

i	x	Percentiles prob_i= (i-0.5)/n	Z Quantiles for percentiles	standardized values for data (x-avg(x1:25)) /stdev(x)
1	66	0.02	-2.05	-2.22
2	75	0.06	-1.55	-1.75
3	83	0.1	-1.28	-1.33
4	84	0.14	-1.08	-1.28
5	91	0.18	-0.92	-0.92
6	94	0.22	-0.77	-0.76
7	99	0.26	-0.64	-0.50
8	102	0.3	-0.52	-0.34
9	103	0.34	-0.41	-0.29
10	103	0.38	-0.31	-0.29
11	104	0.42	-0.20	-0.24
12	107	0.46	-0.10	-0.08

Revised Manuscript Received on July 05, 2019

Kotte Sandeep, Assistant Professor, Dept of CSE, Dhanekula Institute of Engineering & Technology, Vijayawada, A.P, India.

R. Satya Prasad, Professor & Head, Dept of CSE, Acharya Nagarjuna University, Guntur, A.P, India.

13	112	0.5	0.00	0.18
14	114	0.54	0.10	0.29
15	115	0.58	0.20	0.34
16	117	0.62	0.31	0.44
17	118	0.66	0.41	0.50
18	118	0.7	0.52	0.50
19	122	0.74	0.64	0.70
20	124	0.78	0.77	0.81
21	125	0.82	0.92	0.86
22	127	0.86	1.08	0.97
23	131	0.9	1.28	1.17
24	133	0.94	1.55	1.28
25	146	0.98	2.05	1.96

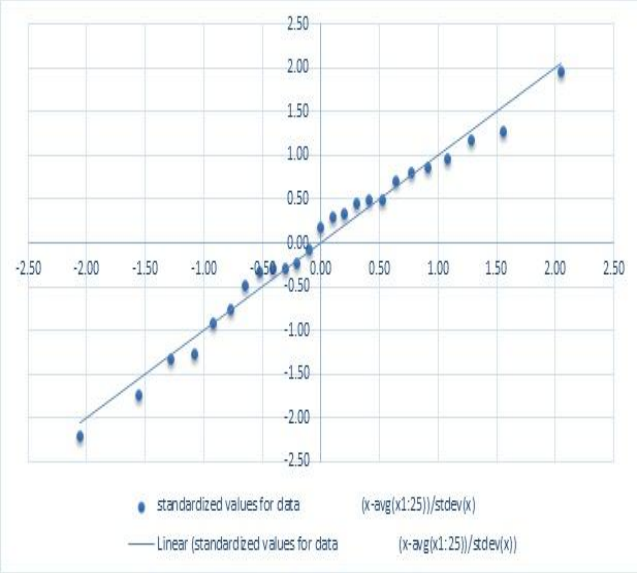

Figure 1: QQ-Plot Plotting from Table 1

In Figure 2, we plotted two different theoretical distributions using sample data from Table 1. For better understanding, two plots are shown in Figure 2.The major limitation, where still, it's visually not that convenient to choose best suited model for the data.

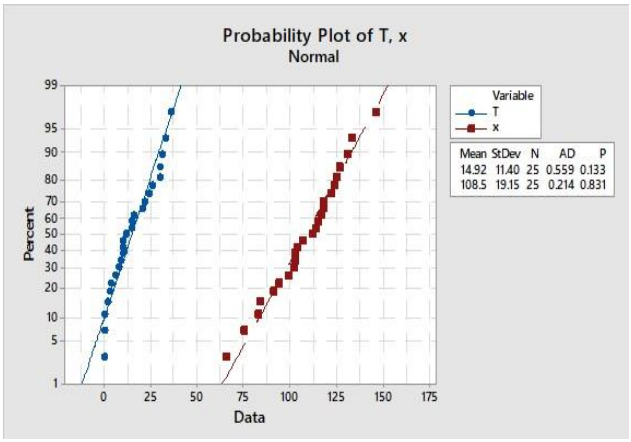

Figure 2: QQ-Plot using two different theoretical distributions.

III. ALTERNATIVE METHOD FOR GOODNESS OF FIT

The author comprehensively exercised distinctive methods to comprehend this circumstance and proposes the thought of utilizing Correlation Factor as an alternative approach rather than plotting the quantiles. The qq plot is a graphical approach for deciding whether the given two sets of quantiles follow common distribution, only the visual perception is the deciding factor. The way toward utilizing correlation factor gives a quantified measure, as well as resolves the possible tie in distinguishing one of numerous apparently similarly close distributions for a dataset through qq-plot[4].To use the correlation factor, the data sources are same as that of qq-plots. In any case, the weight of plotting is overcome and quantified measure is accomplished. The quantified measure accomplished for various models being referred to are compared and the solution is effectively obtained [4].

The following are common to decide the best fit model through qq-plot or through correlation factor. Consider the information for which the model must be chosen and check whether there are any missing values, if so fill them with suitable mean values. Firstly, estimate the unknown parameters using Maximum Likelihood Estimation (MLE), then generate theoretical and distribution quantiles [11]. When both the quantiles are determined, the qq-plot considers plotting of the quantiles, and obtain graphical measure while the correlation factor decides the correlation between the quantiles and thus obtains the quantified measure.

IV. ESTIMATION OF UNKOWN PARAMETERS OF GROUPED DATA USING MAXIMUM LIKELIHOOD ESTIMATION METHOD

Estimation of parameters is very influential in foreseeing the software reliability. After finishing up the analytical solution for the mean value function m(t) for the specific model, the MLE method is enforced for accomplishing the parameter estimation. The intension of MLE is to determine the parameters that magnify the probability of the fragment data. They yield estimators with good statistical factors [1][2][3][5]. MLE techniques are flexible, adaptable and can be utilized to distinct models. To assess the software reliability, the unknown parameters 'a' and 'b' is to be treasured and they are to be anticipated utilizing the failure data of the software fragment data [7].

Let 'n' be the time instances where the first, second,….K^{th} faults in the software are encountered. If T_k is the total time of the K^{th} failure, 't_k' is an observation of random variable T_k and 'n' such similar failures are successively recorded. The combined probability of such failure time handles $t_1,t_2,\ldots\ldots t_n$ is given by the Likelihood function as

$$L = e^{-m(t_n)} \cdot \prod_{k=1}^{n} m'(t_k) \qquad (1)$$

The logarithmic application of the Equation (1) would result a log likelihood function and is given in Equation (2).

$$LogL = \sum_{i=1}^{k}[(n_i - n_{i-1}).\log(m(t_i) - m(t_{i-1}))] - m(t_k) \quad (2)$$

The MLE is highlighted to maximize L and estimate the values of 'a' and 'b'. The process to maximize is by applying partial derivation with respective to the unknown variables and equate to zero to obtain a close form for the required variable. If the closed form is not destined, then the variable can be evaluated using Newton Raphson Method. In this manner 'a' and 'b' would be solutions of the equations.

$$\frac{\partial \log L}{\partial a} = 0 , \frac{\partial \log L}{\partial b} = 0 , \frac{\partial^2 \log L}{\partial b^2} = 0$$

The mean value function m(t) of LPETM is given

$$m(t) = a \log(1 + bt) \quad (3)$$

Implanting the Equations for m(t), λ(t) given by (1) & (2) Equation in Equation 5 and executing the aforementioned process and with the aid of few combined simplifications, we get closure form for variable 'a' in terms of 'b'.

$$a = \frac{(n_k - n_0)}{\log(1 + bt_k)} \quad (4)$$

$$g(b) = (n_k - n_0).\sum_{i=1}^{k}\left[\frac{(t_i - t_{i-1})}{\log\frac{(1+bt_i)}{(1+bt_{i-1})}(1+bt_i)(1+bt_{i-1})}\right] - \frac{(n_k - n_0)}{\log(1+bt_k)}\frac{t_k}{(1+bt_k)} \quad (5)$$

$$g^1(b) = (n_k - n_0).\sum_{i=1}^{k}(-1).\log\left[\frac{(1+bt_i)}{(1+bt_{i-1})}\right]^{-2}.\frac{(t_i - t_{i-1})^2}{(1+bt_{i-1})^2}.\frac{t_i.t_{i-1}}{(1+bt_i)^2} - \frac{(n_k - n_0)}{\log(1+bt_k)^2}.\frac{t_k^2}{(1+bt_k)^2} \quad (6)$$

Where $b = \frac{g(b)}{g^1(b)}$

V. GENERATING THEORETICAL QUANTILES

To generate theoretical quantiles, the cumulative probabilities are computed first. The cumulative probabilities can be figured by taking the ratio of cumulative count of failure data to the total sum of failure data for each record. The acquired cumulative probability is equated to the mean value function of distribution to acquire the model quartiles. Say k_i is cumulative probability and is to be equated to mean value function m(t) of each model and here it is demonstrated for LPETM and solve for t_i where 'i' is the interval measure.

$$t_i = \frac{1}{b}(e^x - 1) \quad (7)$$

This results in the theoretical quantiles break even within amount to the distribution quantiles. To check whether the opted LPETM suits the data, calculation of correlation factor eases the task [3][5].

VI. COORELATION FACTOR

The correlation coefficient always varies between -1 to +1. The larger the absolute estimation of the coefficient, the stronger the relationship between the variables. When it is closer to '1', it indicates a perfect linear relationship. When it is closer to '0', the impact of one variable over another is insignificant. The correlation is estimated for the two variables x and y of n values as

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}} \quad (8)$$

The correlation calculation is eased through the python by using np.corrcoef() function. The same process is repeated for all the models considered for investigation. Here we consider LPETM, HLD, and GO models for the same dataset [5][8][9]. This process is worked out on different datasets and a sample dataset which suits all three models are considered.

A. Results and Interpretation

In Table 2, sample dataset contains 35 weeks Failure data, from that we calculated cumulative failure data and cumulative probabilities. Cumulative probabilities are nothing but cumulative failure data by the sum of cumulative failure data.

Length of failure data is 35
Cumulative failure data is 148

From Equation 5 and 6 we calculated b value.

Where $b = \frac{g(b)}{g^1(b)}$ = b value is 0.043217

From Equation 4, we calculated a value is 150.87113. Substituting cumulative probabilities into Equation 7 to get model quantiles. Applying correlation coefficient Equation 8 on model quantiles to get correlation factor. This correlation is 0.99706396, a strong positive correlation [[1, 0.99706396]. The same process above is repeated for the three models LPETM, HLD, AND GO.

Table II: Sample Data on No of errors Encountered during corresponding Week [6].

Week No	Failure Data	Cumulative Failure Data	Cumulative Probabilities
1	9	9	0.060811
2	4	13	0.087838
3	7	20	0.135135
4	6	26	0.175676
5	5	31	0.209459
6	3	34	0.22973
7	2	36	0.243243
8	5	41	0.277027
9	4	45	0.304054
10	2	47	0.317568
11	4	51	0.344595

12	7	58	0.391892
13	5	63	0.425676
14	3	66	0.445946
15	3	69	0.466216
16	3	72	0.486486
17	4	76	0.513514
18	10	86	0.581081
19	3	89	0.601351
20	1	90	0.608108
21	2	92	0.621622
22	4	96	0.648649
23	5	101	0.682432
24	2	103	0.695946
25	2	105	0.709459
26	3	108	0.72973
27	6	114	0.77027
28	3	117	0.790541
29	1	118	0.797297
30	1	119	0.804054
31	4	123	0.831081
32	3	126	0.851351
33	2	128	0.864865
34	11	139	0.939189
35	9	148	1

VII. CONCLUSION

The fundamental focal point of this prime is to demonstrate a helpful furthermore, simple procedure to propose model best suits the information out of various models by evaluating the closeness of the model to the information through figuring the correlation and rather than graphical plot (QQ-plot).

REFERENCES

1. H. Pham, System software reliability, Springer, 2006.
2. R Satya Prasad,K Ramchand H Rao and R.R. L Kantham,"Software Reliability Measuring using Modified Maximum Likelihood Estimation and SPC" International Journal of Computer Applications, vol-21, Number 7, pp.1-5 Article1, May 2011.
3. R.satyaprasad, "Half Logistic Software Reliability Growth Model",Ph.D. Thesis,2007.
4. Donald J. Wheeler, "Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts" Handbook. Statistical Process Controls, Inc., SPC Press, Knoxville, TN, 2004.
5. R. Satya Prasad, K Sowmya and R Mahesh, "Monitoring Software Failure Process using Half Logistic Distribution" International Journal of Computer Applications, vol-145(4), pp.1-8, July 2016
6. https://data.world/datasets/open-data
7. A.L.Goel andK. Okumoto, "Time-Dependent Error-Detection Rate Model for Software Reliability and Other Performance Measures". IEEE Transl. on Reliability vol-R–28 Issue. 3 pp. 206– 211, Aug, 1979.
8. R Satya Prasadand D. Haritha, "Discovery of Reliable Software using GOM on Interval Domain Data", International Journal of Computer Applications vol. 32 Issue. 5, pp. 7-12, October 2011
9. R. Satya Prasad, D. Haritha and R.Sindhura "Assessing Reliable Software using SPRT based on LPETM", International Journal of Computer Applications vol. 47 Issue.19 pp. 6-12, June 2012.
10. Mekhala Sridevi Sameera, Dr. T. Anuradha, Dr. R. Satya Prasad. A survey on software reliability and measurement techniques. International Journal of Emerging Technologies and Innovative Research. 2018; 5(11):379-382.
11. K. Sandeep, R. Satya Prasad "Analysis of Theoretical sampling distributed using combined LPETM and ANOM subgrouping", Journal of Advanced Research in Dynamical and Control Systems (JARDCS-SCOPUS) ISSN: 1943-023X, Volume 11 Issue 1, Pages-467-471, Mar-2019.

AUTHORS PROFILE

Kotte Sandeep (First Author) working as an Assistant Professor, Department of CSE, Dhanekula Institute of Engineering & Technology, Andhra Pradesh. He received his M.S degree from Alpen-Adria Universitat, Klagenfurt, Austria in 2010. He received his B.Tech degree from JNTUH in 2007. He is pursuing Ph.D. in computer science and engineering from Acharya Nagarjuna University. He is having 10 years of experience as an Assistant Professor in Computer Science field. He published 12 research papers in various National and International Journals and 1 Scopus Indexed Journal.

Dr. R. Satya Prasad (Second Author) Received Ph.D. degree in Computer Science in the faculty of engineering in 2007 from Acharya Nagarjuna University, Andhra Pradesh. He received gold medal from Acharya Nagarjuna University for his outstanding performance in Master's Degree. He is currently working as a Professor & Head, in the Department of Computer Science and Engineering, Acharya Nagarjuna University, Andhra Pradesh. His current research is focused on Software Reliability and software Engineering. He published 155 research papers in International Journals & Scopus.