UNIVERSIDADE DE SÃO PAULO

INFERENCE FOR THE SOFTWARE RELIABILITY USING IMPERFECT RECAPTURE DEBUGGING MODEL

JOSEMAR RODRIGUES
JOSÉ GALVÃO LEITE

Nº 28

NOTAS

Instituto de Ciências Matemáticas de São Carlos

Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2577

INFERENCE FOR THE SOFTWARE RELIABILITY USING IMPERFECT RECAPTURE DEBUGGING MODEL

JOSEMAR RODRIGUES
JOSÉ GALVÃO LEITE

Nº 28

NOTAS DO ICMSC
Série Estatística

São Carlos
Mar./1996

Inference for the software reliability using imperfect recapture debugging model

Josemar Rodrigues
ICMSC-USP-São Carlos-13560-970
Brazil
and
José Galvão Leite
IME-USP - São Paulo-05389-970
Brazil

Abstract

In this paper we describe a likelihood and Bayes approaches for the recapture debugging design proposed by Nayak (1988) to get information about the number of faults, N, in a reliability system. It is considered that the debugging is only successful with known probability p. It is shown that the mle of N depend not only of the frequencies of detected bugs but also of the times between bugs. The sensibility of the posterior distribution of N with respect to p and the influence of the time data are numerically considered via variational distance and Kullback-Leibler divergence. Also, it is shown that the recapture debugging is a necessary and sufficient condition for the existence of the posterior distribution of N when an improprer prior is imposed.

Key words: Bayesian and likelihood inferences, Jelinski-Moranda model, Φ-divergence.

1. Introduction:

Let N be unknown number of errors in a piece of a computer software. An important problem in software reliability is the estimation of N from the data. The Jelinski-Moranda model (1972), , henceforth known as the JM model, was the first software reliability model widely used and the main motivation for various alternative models. The JM model asssumes that the times between bugs $W_i(i=1,\ldots,N)$ are independent exponential variables with failure rate proportional to $N-i+1$, which is the number of bugs remain in the software. Also, he assumed a perfect debugging after each failure, that is , the error is removed without inserting any additional errors. In this paper, we suppose that the software is tested for a fixed time τ and the sucessive failure times $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$, are observed , where n is also a random variable. Also, the JM model is adopted but with an imperfect and recapture debugging as described by Goel and Okumoto (1978) and Nayak (1988), respectively. As in Nayak, suppose that the N bugs are in different unknown areas $A_1, \ldots, A_N$ and that testing counters are used to count the number of times that each area is accessed. In order to formulate our model , let us consider the following notations:

.Notations:

1. n=number of discovered bugs in the package after running τ times unit ($n \geq 1$).
2. Given n and $T_{(1)} \leq T_{(2)} \leq \ldots . T_{(n)} \leq \tau$, sucessives failure times, let $W_i = T_{(i)} - T_{(i-1)}$; $i = 1,\ldots,n+1$, where $T_{(o)} = 0$ and $W_{n+1} = T_{(n+1)} - T_{(n)}$ is censored because $W_{n+1} > \tau - T_{(n)}$.
3. N=unknown number of errors.
4. p= probability of fixing a ith fault when it is encountered, $i = 1,\ldots . n$ (known).
5. M_i = number of times that the package return to the i-th observed area.
6. $M = \sum_{i=1}^{n} M_i$.

Given the above notations, we formulate the imperfect recapture model , based on the censored data $W_1, W_2, \cdots, W_n, W_{n+1}$, as

(1) $W_i | N, \Lambda \approx Exp\{\Lambda[N - p(i-1)]\}, i = 1,\ldots,n+1.$

(2) $M_i \approx Poisson(\Lambda(\tau - T_{(i)})), i = 1,\ldots . n.$ (1)

The first step of the model (1) corresponds to the Goel and Okumoto's Imperfect Debugging Model (1978).

2. Likelihood Estimation (p: known)

Under the model (1) , the likelihood function is given by

$$L(N,\Lambda) = (n+1)\prod_{i=1}^{n} \Lambda[N - p(i-1)] \exp\left\{-\Lambda \sum_{i=1}^{n+1}[N - p(i-1)]W_i^*\right\} \prod_{i=1}^{n}\left[\Lambda(\tau - T_{(i)})\right]^{M_i} .$$

$$. \exp\left\{-\Lambda\left(n\tau - \sum_{i=1}^{n} T_{(i)}\right)\right\} \quad \text{for} \quad N \geq n,\ \Lambda > 0, \qquad (2)$$

where $W_i^* = \begin{cases} W_i, i=1,\ldots,n \\ \tau - T_{(n)}, i = n+1 \end{cases}$.

The next result will be useful to relate our likelihood approach with results obtained by Nayak (1988).

Theorem 1:

$$\sum_{i=1}^{n+1}\left[N-p(i-1)\right]W_i^{*}=p\sum_{i=1}^{n}T_{(i)}+(N-np)\tau. \qquad (3)$$

Proof:

It is not difficult to prove that

$$\sum_{i=1}^{n}\left[N-p(i-1)\right]W_i=p\sum_{i=1}^{n}T_{(i)}+(N-np)T_{(n)}.$$

The result follows if we add the quantity $(N-np)W_{n+1}^{*}$ in the above expression.

Given N, the mle of Λ is expressed as

$$\hat{\Lambda}=\frac{n+M}{N\tau+(1-p)\left[n\tau-\sum_{i=1}^{n}T_{(i)}\right]} \qquad (4).$$

In order to help to understand our results let us consider some remarks:

1. If $p=1$ our likelihood coincides with Nayak's likelihood; this is a possible justification why the time data does not provide any extra information about the parameters in Nayak's paper.
2. If $0<p<1$ does the time data provide any extra information about the parameter involved in the model? This question and the sensibility with respect to the choice of p will be addressed from the Bayesian point of view via Kullback-Leibler divergence and variational distance.

   From the likelihood point of view the information about N can be outlined in the following way:

The Profile likelihood for N:

$$L_{*}(N)=\frac{\prod_{i=1}^{n}\left[N-p(i-1)\right]}{\left[N\tau+(1-p)\left(n\tau-\sum_{i=1}^{n}T_{(i)}\right)\right]^{n+M}}, \qquad (5)$$

for $N\geq n$.

Using the same procedure introduced by Nayak (1988) we can obtain, from our likelihood function , the mle of N which is given by: Let

$$g(N)=\left[1-\frac{1}{N+(1-p)v}\right]^{n+M}+1-\prod_{i=1}^{n}\left[1-\frac{1}{N-p(i-1)}\right], \text{ for } N\geq n+1,$$

where $v=\dfrac{n\tau-\sum_{i=1}^{n}T_{(i)}}{\tau}$. (6)

If $g(n+1)<1$ then $\hat{N}=n$. If $g(n+1)>1$ then $\hat{N}=k$ if and only if $h^{*}_{n,k+1}<M<h^{*}_{n,k}$, where

$$h^{*}_{n,k}=\frac{\sum_{i=1}^{n}\ln\left(1-\frac{1}{k-p(i-1)}\right)}{\ln\left\{1-\frac{1}{k+(1-p)v}\right\}}-n.$$

If $p=1\Rightarrow h^{*}_{n,k}=h_{n,k}$ and the mle of Nayak follows (see Nayak's paper for the definition of $h_{n,k}$).

3. The Bayes approach :

The Bayesian approach for N and Λ involves assigning a prior distribution for N and Λ, and using the information summarized in the likelihood function and the Bayes theorem to obtain the posterior distributions. Again, we are concerned about the parameter N. The likelihood function can be rewritten as

$$L(N,\Lambda)\propto\Lambda^{n+M}\prod_{i=1}^{n}[N-p(i-1)]\exp\left\{-\Lambda\left[N\tau+(1-p)\left(n\tau-\sum_{i=1}^{n}T_{(i)}\right)\right]\right\}, \quad (7)$$

for $N\geq n$.

Assuming that N and Λ are independent, we consider the following priors :

$$1. \quad \Pi(N) \propto \frac{\Gamma(N+a_1)}{N!(b_1+1)^{N+a_1}} \qquad (8)$$

$$2. \quad \Pi(\Lambda) \propto \Lambda^{a_2-1} \exp\{-b_2\Lambda\}$$

The motivation to consider the Negative-Binomial prior for N (prior 1.) follows from the fact that it is equivalent to take a Poisson distribution for N, given the parameter Θ, and a Gamma prior distribution for Θ.

If $a_2 = b_2 = a_1 = b_1 = 0$, we have an improper prior for N and Λ given by

$$\Pi(\Lambda) \propto \frac{1}{\Lambda} \text{ and } \Pi(N) \propto \frac{1}{N} \Rightarrow \Pi(N,\Lambda) \propto \frac{1}{N\Lambda}.$$

After obtainning the joint posterior of N and Λ and integrating over Λ we have , under the noninformative case , the following posterior for N:

$$\Pi(N|data) \propto \frac{\prod_{i=1}^{n}[N-p(i-1)]}{N\left[N\tau+(1-p)\left(n\tau-\sum_{i=1}^{n}T_{(i)}\right)\right]^{n+M}}, \quad (9)$$

f or $N \geq n$.

It is very important to note that the above posterior distribution is proper if and only if M>0, so, we have here a very good Bayesian justification to consider the recapture procedure proposed by Nayak. The question of the existence of the posterior distribution arises when one imposes improper distributions on the interest parameter as we did before. The recapture debugging procedure proposed by Nayak gives a necessary and sufficient condition for the existence of (9). We suggest to the reader to see Natarajan and McCulloch 's paper to understand how important is the existence of the posterior distribution when applied to other models.
The next result gives a relation between the Bayesian and the likelihood approaches.

Theorem 2:

$$\Pi(N|data) \propto \frac{L_*(N)}{N} \qquad (10)$$

Proof: It follows from (5) and (9).

Using the above result and a similar procedure used before to get the mle of N, we can obtain the following procedure to find the mode, $\hat{N}_0$, of N: Let

$$g^*(N) = \left[1 - \frac{1}{N + (1-p)v}\right]^{n+M} + 1 - \prod_{i=2}^{n}\left(1 - \frac{1}{N - p(i-1)}\right), \text{ for } N \geq n+1$$

and $n \geq 2$. If $n = 1$, then $\hat{N}_o = 1$.

If $g^*(n+1) < 1$ then $\hat{N}_0 = n$. If $g^*(n+1) > 1$ then $\hat{N}_0 = k$ if and only if

$$h^{**}_{n,k+1} < M < h^{**}_{n,k}, \text{ where}$$

$$h^{**}_{n,k} = \frac{\sum_{i=2}^{n} \ln\left(1 - \frac{1}{k - p(i-1)}\right)}{\ln\left\{1 - \frac{1}{k + (1-p)v}\right\}} - n. \qquad (11)$$

4. Robustness considerations of the posterior distribution of N with respect to p and the influence of the time data via divergence measures.

Let $\pi_0 = \pi_0(N|data)$, the posterior distribution of N in (9) with $p = 1$ and π the posterior distribution of N for $0 < p < 1$. Our purpose is to compare the posterior distributions of N, π and π_0, for different values of p, by using the Kullback-Leibler and the variational distance as divergence measures. Also, the influence of the time data is discussed for a numerical data.

Following Csiszár (1967), we can in general define the Φ-divergence between the posteriors π_0 and π as

$$D_\Phi = D(\pi_0, \pi) = \sum_{N \geq n} \Phi\left(\frac{\pi}{\pi_0}\right)\pi_0 \qquad (12)$$

Several choices of Φ are given in Dey and Birmiwal (1994), in this section we choose $\Phi(x) = -\ln(x)$ which defines the KL divergence and $\Phi(x) = \frac{1}{2}|x-1|$ which defines the variational distance (VD) or L_1 norm. Peng and Dey (1995) suggested the use of the variation distance for detecting influential observations on the posterior distributions. In order to compare our posterior distributions , let us consider the following particular data: n=5, τ=1, M=2 and $\sum_{i=1}^{n} T_{(i)} = 1$

Table 1. Divergence measures: KL and VD

p	0.10	0.50	0.80	1.00
KL	0.04	0.01	0.00	0.00
VD	0.11	0.06	0.03	0.00

Table 1. shows, under this particular data, that the posterior π is not sensitive to the choice of p and behaves like π_0. Also, the time data is not influential on the posterior π_0. In conclusion, for this data, the posterior distribution π_0 is robust in the sense that we do not to need to worry about imperfect debugging or time data. Using the VD measure and the scale suggested by Peng and Dey (1995), we only observe, for p=0.10, a mild influence of the time data on the posterior distribution of N.

References

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar, 2, 299-318.

Dey D.K. and Birmiwal, L.R. (1994). Robust Bayesian analysis using entropy and divergence measures. Statist. Probab. Lett., 20, 287-294.

Goel, A.L. and Okumoto, K. (1978). An analysis of recurrent software failures on real time control system, Proc. ACM Annu. Tech. Conf. Washington D.C.: ACM, 496-500.

Jelinski, Z. and Moranda , P.B. (1972). Software reliability research. In Statistical Computer Performance Evaluation, ed. W. Freiberger, New York Londom: Academic Press, 465-484.

Nayak, K. T. (1988). Estimating population size by recapture sampling, Biometrika, , 75, 1, 113-120.

Natarajan, R. and McCulloch, C.E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial process, Biometrika, 82, 3, 639-643.

Peng, F. and Dey, D.K. (1995). Bayesian analysis of outlier problems using divergence measures, The Canadian J. of Statistics, 23, 2, 199-213.

Resumo

Neste artigo apresentamos a abordagem Bayesiana e verossimilhança para o planejamento de recaptura de erros em um software proposto por Nayak (1988), com a finalidade de obter informação sobre o número de erros, N, existentes no software. O processo de eliminar o erro é considerado com sucesso somente com probabilidade p. Verificamos que o estimador de m.v. de N depende da frequência dos erros detectados e do tempo entre os sucessivos erros. A sensibilidade da distribuição a posteriori de N em relação a p e a influência dos tempos entre erros são numéricamente analizadas via "variational distance" e a distância de Kullback-Leibler. Também, quando uma priori não informativa é adotada, provamos que o processo de recaptura de erros é uma condição neçessária e suficiente para a existência da distribuição a posteriori de N.

NOTAS DO ICMSC

SÉRIE ESTATÍSTICA

027/96 ACHCAR, J.A.; DEY, D.K.; NIVERTHI, M. - A bayesian approach using nonhomogeneous Poisson process for software reliability models.

026/96 ANDRADE, M.G.; VAL, J.B.R. do - Um método numérico baseado na solução do valor médio para a equação de Helmholtz parte II: Rede triangular.

025/96 ANDRADE, M.G.; VAL, J.B.R. do - Um método numérico baseado na solução do valor médio para a equação de Helmholtz parte I: malha quadrada.

024/96 MAZUCHELLI, J.; ACHCAR, J.A. - Análise bayesiana para modelos de crescimento.

023/96 ANDRADE, M.G.; SOARES, S.; CRUZ Jr., G.; VINHAL, C.D.N - Uma abordagem estocástica para o planejamento a longo prazo da operação de sistema hidrotérmicos.

022/95 ACHCAR, J. - A generalized Moranda software reliability model: a bayesian approach.

021/95 RODRIGUES, J. - Inference for the software reliability using asymmetric loss functions: A hierarchical Bayes approach.

020/95 RODRIGUES, J.; BOLFARINE, H.; CORDEIRO, G.M. - Nonlinear quasi-bayesian theory and the inverse linear regression.

019/95 LEANDO, R.A.; ACHCAR, J.A.; - Generation of bivariate lifetime data assuming the Block & Basu exponential

018/95 ACHCAR, J.A.; - Use of approximate bayesian inference for software reliability