한국통계학회 논문집 제1권1호
1994년 12월 pp. 10-19

A Study on Change-Points in System Reliability[1]

Kwang Mo Jeong[2]

Abstract

We study the change-point problem in the context of system reliability models. The maximum likelihood estimators are obtained based on the Jelinski and Moranda model. To find the approximate distribution of the change-point estimator, we suggest a parametric bootstrap method in which the estimators are substituted in the assumed model. Through an example we illustrate the proposed method.

1. Introduction

Consider a sequence of independent random variables $X_1, X_2, \cdots, X_n$ such that $X_1, \cdots, X_\tau$ are identically distributed with distribution function F, and $X_{\tau+1}, \cdots, X_n$ are distributed with a different distribution function G, then τ is called a *change-point* in the usual sense. The change-point τ is unknown and must be estimated from the data.

Many authors have studied change-point problems. Among others, Hinkley (1970) and Hinkley and Hinkley (1970) suggested maximum likelihood estimators (MLEs) when F and G are from the same parametric family. Also, Smith (1975) suggested a Bayesian approach and Cobb (1978) studied conditional approach under the assumption that F and G are known to belong to parametric families. On the other hand, Pettitt (1979) and Carlstein (1988) discussed nonparametric estimation of the change-points.

Hinkley and Schechtman (1987) suggested conditional bootstrap methods in the mean shift model. Pham and Nguyen (1991) discussed theoretical distribution of bootstrap estimators of change-points in hazard rate. Further, Joseph and Wolfson (1992) applied bootstrap methods to the multi-path change-points problem where several independent sequences are considered simultaneously.

In this paper, we are concerned with the change-point problem in software reliability models. Many authors studied the estimation and prediction for a software reliablity model with no change-points. Littlewood (1981) proposed a stochastic reliability growth model, and Joe and Reid (1985) suggested the estimation of the number of faults in a system. In software

1)This paper was supported (in part) by Pusan National University Oversea Research Fund, 1993
2) Department of Statistics, Pusan National University, Kumjung Ku, Pusan 609-735, KOREA

reliability models a more complicated situation occurs because the total number of faults are usually unknown.

Zhao (1993) discussed the maximum likelihood estimation of change-points and number of faults. But, it is not easy to obtain the asymptotic distribution of change-point estimator because the explicit form of the MLE does not exist in general. We consider the so called Jelinski and Moranda (JM) model for system reliability which has only one change-point. We also suggest bootstrap confidence intervals for change-points.

In Section 2 we consider a change-point model in software reliability with some assumptions which are a little different from those in Zhao (1993). The maximum likelihood estimation and bootstrap confidence intervals are discussed in Section 3. In Section 4, we also apply the proposed method to the well-known failure time data of Musa (1979). Finally in Section 5 we summarize the results and comment about further researches and the restriction of this paper.

2. Change-Point Models in Software Reliability

In this section we consider a change-point model in software reliability which is a little different from the ususal change-point models previously studied. In software reliability the JM model has been used extensively in the literature.

The times to failures in software, resulting from initial N bugs, are assumed to be independent exponential random variables with a common failure rate ϕ. The model assumes that, on failure, the bug responsible is removed and that no other bugs are affected by this process. With its assumptions of a fixed number of bugs, independence, and the common constant failure rate for each bug the JM model has been criticised for being over simplistic. Even so, the JM model has received considerable attention and has been of interest in its own right. We refer the readers to Littlewood (1981) and Wright and Hazelhurst (1988) for a more detailed discussion of JM model.

In software reliability we execute the software in a specific environment and debug it by detecting the faults, where the initial number of faults N is usually unknown. The JM model assumes that each failure caused by a fault occurs independently and randomly. When a fault occurs it is corrected instantaneously and it does not affect the other occurrences of faults.

The JM model for system reliability can be modified by considering changes in failure rate because the running environment may not be homogeneous and can be changed with the human learning process. Zhao (1993) introduced the change-point problem in JM reliability models. In this section we slightly modify the assumptions in Zhao (1993) and specify the change-point models in more detail.

Let $X_1, X_2, \cdots, X_n$ denote the interfailure times, of the sequential failures, in a life testing situation. We make the following assumptions.

Assumption 1. The total number of faults (or bugs) N is unknown but finite.

Assumption 2. The failure rate ϕ changes after an unknown time point τ.

Assumption 3. The interfailure times $X_1, X_2, \cdots, X_n$ are mutually independent.

Comment Zhao (1993) assumed that the sequence $\{X_1, \cdots, X_\tau\}$ is independent on the sequence $\{X_{\tau+1}, \cdots, X_n\}$ but this assumption is not sufficient to define the likelihood function and MLEs. It seems that Zhao (1993) missed some theoretical aspects of the problem.

Under the above assumptions, we can formulate the JM model with one change-point. The *ith* interfailure time X_i is distributed with exponential probability density function

$$f(x_i \mid \lambda_i) = \lambda_i \exp(-\lambda_i x_i) \tag{2.1}$$

where

$$\lambda_i = \begin{cases} (N-i+1)\phi_1 & \text{for } i = 1, \cdots, \tau \\ (N-i+1)\phi_2 & \text{for } i = \tau+1, \cdots, n \end{cases}$$

The number of observations n is predetermined according to a censoring plan before the data is collected. Here we also note that each fault in a computer program has the same size and the removal of faults reduces the failure density by the same amount. The failure rate changes after the first τ faults are detected. In hardware reliability where the testing equipments are replaced or repaired during the lifetime testing the situation is similar to the software reliability model except that the parameter N is usually known.

3. Estimation of Change-Point

3.1. Maximum Likelihood Estimation

Under the model (2.1) we can write the log-likelihood function based on the observations $x_1, \cdots, x_n$ as follows.

$$\begin{aligned} l(\theta) = & \sum_{i=1}^{n} \log(N-i+1) + \tau \log(\phi_1) - \phi_1 \sum_{i=1}^{\tau}(N-i+1)x_i \\ & +(n-\tau)\log(\phi_2) - \phi_2 \sum_{i=\tau+1}^{n}(N-i+1)x_i \end{aligned} \tag{3.1}$$

where $\theta = (\phi_1, \phi_2, \tau, N)$.

For given τ and N we obtain the following estimators of ϕ_1 and ϕ_2 by differentiating the likelihood function with respect to ϕ_1 and ϕ_2.

$$\hat{\phi}_1 = \frac{\tau}{\sum_{i=1}^{\tau}(N-i+1)x_i} \tag{3.2}$$

$$\hat{\phi}_2 = \frac{n-\tau}{\sum_{i=\tau+1}^{n}(N-i+1)x_i} \tag{3.3}$$

If we substitute (3.2) and (3.3) in the equation (3.1), the profile likelihood function can be rewritten as

$$\begin{aligned} l(\tau,\ N) \propto\ & \sum_{i=1}^{n}\log(N-i+1)+\tau\log(\tau/\sum_{i=1}^{\tau}(N-i+1)x_i) \\ & +\ (n-\tau)\log((n-\tau)/\sum_{i=\tau+1}^{n}(N-i+1)x_i) \end{aligned} \tag{3.4}$$

We find the MLEs of τ and N which maximize, simultaneously, the profile likelihood function (3.4). As pointed out by many authors the MLE of N may be infinite and $\hat{\phi}_1$ and $\hat{\phi}_2$ equal to zero. This situation is very undesirable and we need to confine N in a certain range of integer values which has practical interpretation. The maximization of (3.4) proceeds as follows. The unknown τ may vary from 1 to $n-1$ and we also let N vary from n to a certain upper bound N_0 which can be predetermined. For every grid point $(\tau,\ N)$ in the parameter space we find the MLEs maximizing the likelihood (3.4). Zhao (1993) obtaind the MLE of N by differentiating with respect to N. In this case the estimator is not necessarily integer valued and which may reduce its practical meaning.

Here we are mainly concerned with the change-point τ. We note that the MLE $\hat{\tau}$ of τ can only be found in implicit form and it is not easy to obtain the asymptotic distribution of $\hat{\tau}$. Joseph and Wolfson (1992) suggested a bootstrap method to estimate the distribution of $\hat{\tau}$ under the assumption of multi-path sequences of observations. But the assumption of multi-path sequences is so strong that it limits its application.

3.2. Bootstrap Confidence Intervals of τ

To obtain approximate confidence intervals for τ, we use a resampling technique called the *parametric bootstrap* where we resample from the probability density function $f(x \mid \hat{\theta})$, where

$f(x \mid \theta)$ is an exponential density given in (2.1). The parametric bootstrap algorithm for finding confidence intervals of change-point can be written as:

Step 1. Find MLEs of $\theta = (\phi_1, \phi_2, \tau, N)$ from the given data by the equations (3.2), (3.3) and (3.4).

Step 2. Resample $x_1^*, \cdots, x_n^*$ from $f(x \mid \hat{\theta})$, where $\hat{\theta}$ denotes the MLE of θ.

Step 3. For the sample $x_1^*, \cdots, x_n^*$, we find the MLE $\hat{\theta}^* = (\hat{\phi}_1^*, \hat{\phi}_2^*, \hat{\tau}^*, \hat{N}^*)$.

Step 4. Repeat Step 2 and Step 3 B times.

Step 5. Compute an approximate $100(1-\alpha)\%$ confidence interval from the empirical distribution of $\hat{\tau}^{*(b)}$, $b = 1, \cdots, B$, where $\hat{\tau}^{*(b)}$ is the MLE of τ based on the b^{th} resampled data.

The number of bootstrap samples B is usually taken to be $B = 200$ or $B = 500$. For a general discussion of bootstrap confidence intervals see Hall (1988). We note that the sampling distribution of $n(\hat{\tau}^* - \hat{\tau})$ approximates the distribution of $n(\hat{\tau} - \tau)$ under some regularity conditions.

4. An Example

We apply the proposed method to interfailure times of a system shown in Fig. 4.1, which were originally given in Musa (1979). The data is assumed to follow an exponential distribution in general. The JM model can be modified by considering a change-point

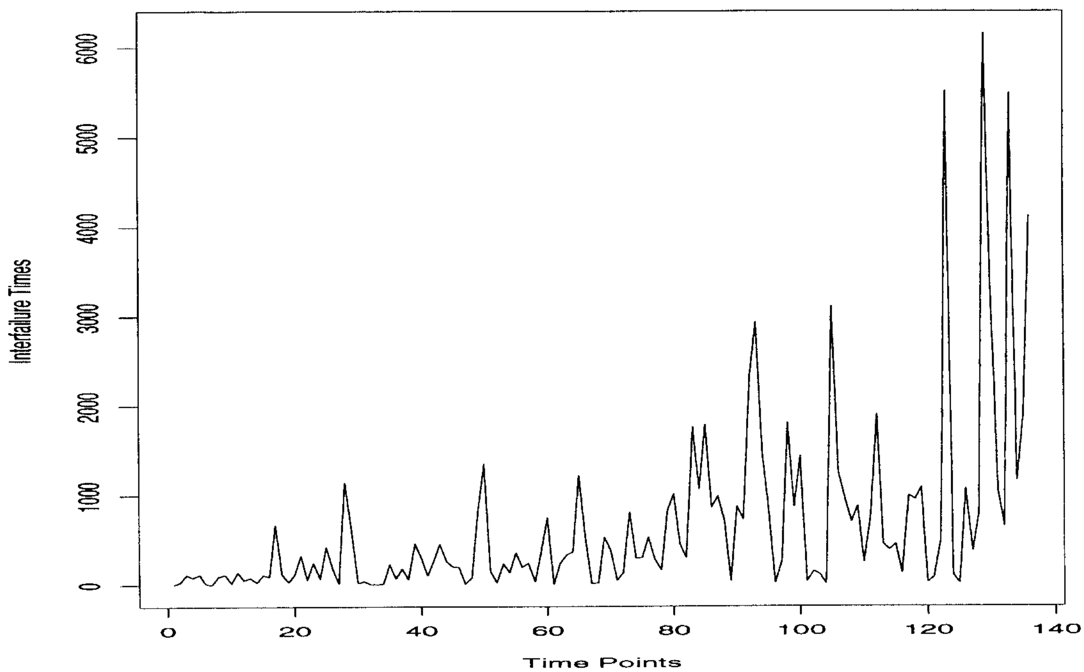

Fig. 4.1. Plot of Interfailure Times of a System

problem in which the failure rate may change after an unknown time point τ. For this data we fit the change-point model in (2.1), where the sample size n = 136 and the total number of faults N being unknown.

From the equations (3.2), (3.3) and (3.4) the MLEs of ϕ_1, ϕ_2, τ and N are routinely obtained as

$$\hat{\tau} = 16, \quad \hat{N} = 145, \quad \hat{\phi}_1 = 1.107742e-04, \quad \hat{\phi}_2 = 2.985331e-05$$

For this data the MLE of N is a unique finite value but in some situations the MLE of N may be infinite and this fact may cause the non-existence of the MLEs of ϕ_1, ϕ_2 and τ. Fig. 4.2 denotes the log-likelihood against change-points when the estimated parameters of ϕ_1, ϕ_2 and N are given. We note that the maximum of the likelihood occurs at $\hat{\tau}$ = 16.

By substituting the MLEs obtained above in the model (2.1) we can perform the bootstrap algorithm of Step 1 through Step 5 in Section 3.2. The empirical distribution of $\hat{\tau}-\tau$ based on B = 500 bootstrap samples is given in Table 4.1. The smoothed empirical

Fig. 4.2. Log-likelihood against the Change-Points

probability density function, based on 500 bootstrap estimators of $\hat{\tau}$, is given in Fig. 4.3. The empirical distribution of $\hat{\tau}$ is skewed to the right as shown in Table 4.1 and Fig. 4.3. The fact that a change-point has likely occurred very early may cause this phenomenon.

From the bootstrap estimators, we can determine the approximate $100(1-\alpha)\%$ confidence intervals (C.I.s) for τ. The 95% and 90% C.I.s are (10, 45) and (13, 29),

Table 4.1. Empirical Distribution of $\hat{\tau}-\tau$ Estimated with B = 500 Samples

$\hat{\tau}-\tau$	≤ -5	-4	-3	-2	-1	0	1	2	3	4	5	≥6
Prob	.05	.01	.016	.036	.078	.364	.156	.082	.066	.026	.018	.108

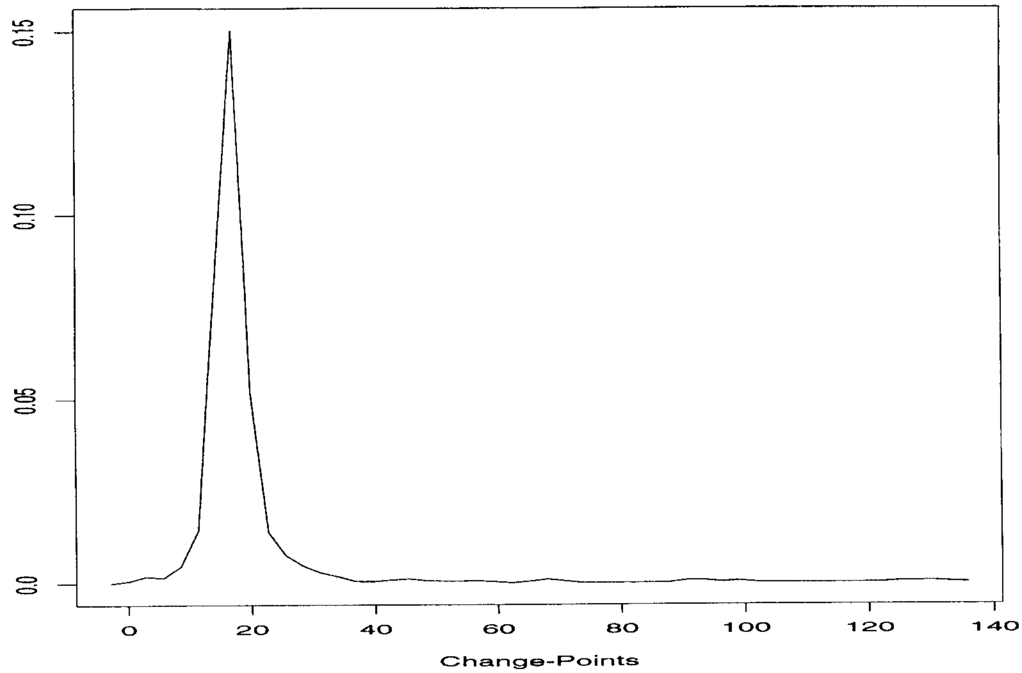

Fig. 4.3. Empirical Probability Density of $\hat{\tau}$ Based on B = 500 Bootstrap Samples

respectively. We recommend the use of 90% C.I. because the 95% C.I. is very wide. For the general discussion of bootstrap C.I.s we refer the readers to Hall (1988).

5. Conclusion

We considered a change-point problem for the JM model in system reliability. It can be justified because the environment of a system may depend on several factors as the testing time proceeds. The MLEs are obtained in a routine way but can only be obtained implicitly. So, it is not easy to find the asymptotic distribution of the estimators.

We suggested a bootstrap technique to obtain the approximate distribution of the change-point estimator. The bootstrap algorithm is summarized as follows. First, we obtain the MLEs of parameters in the assumed model. We generate a bootstrap sample from the distribution with estimated parameters and find bootstrap estimator for each bootstrap sample. We repeat the above process B times and determine the approximate C.I.s of true change-point.

We applied a parametric bootstrap method to the interfailure times of a system given in

Musa (1979). After finding the MLEs we performed the bootstrap algorithm $B = 500$ times. In this case the empirical distribution of $\hat{\tau}$ is skewed to the right. We can also determine the approximate C.I.s but they can be quite wide. The adjustment of skewness of C.I.s is not treated in this paper and we refer the readers to Hall (1992).

As discussed in other papers the estimator of N, number of faults in a software, may be infinite. We did not discuss the problem in detail when the esimator of N is infinite. The study on the asymptotic distribution of $\hat{\tau}$ and the generalization of JM model is remained as further researches.

References

[1] Carlstein, E. (1988), Nonparametric Change-Point Estimation, *The Annals of Statistics*, 2, 188-197.

[2] Cobb, G. W. (1978), The Problem of the Nile: Conditional Solution to a Change-Point Problem, *Biometrika*, 65, 243-251.

[3] Hall, P. (1988), Theoretical Comparison of Bootstrap Confidence Intervals, *The Annals of Statistics*, 16, 927-953.

[4] Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

[5] Hinkley, D. V. (1970), Inference about the Change-Point in a sequence of Random variables, *Biometrika*, 57, 1-16.

[6] Hinkley, D. V. and Hinkley, E. A. (1970), Inference about the Change-Point in a Sequence of Binomial Variables, *Biometrika*, 57, 477-488.

[7] Hinkley, D. V. and Schechtman, E. (1987), Conditional Bootstrap Methods in the Mean Shift Model, *Biometrika*, 74, 85-93.

[8] Joe, H. and Reid, N. (1985), Estimating the Number of Faults in a System, *Journal of the American Statistical Association*, 80, 222-226.

[9] Joseph, L. and Wolfson, D. B. (1992), Estimation in Multi-Path Change-Point Problem, *Communications in Statistics*, Theory and Method, 21, 897-913.

[10] Littlewood, B. (1981), Stochastic Reliability Growth: A Model for Fault-Removal in Computer-Programs and Hardware-Design, *IEEE Transactions and Reliability*, R-30, 312-320.

[11] Musa, J. D. (1979), *Software Reliability Data*, RADC, New York.

[12] Pettitt, A. N. (1979), A Non-Parametric Approach to the Change-Point Problem, *Applied Statistics*, 28, 126-135.

[13] Pham, T. D. and Nguyen, H. T. (1991), Bootstrapping the Change-Point of a Hazard Rate, *Rapport De Recherche, Center National de la Recherche Scientifique.*

[14] Smith, A. F. M. (1975), A Bayesian Approach to Inference about a Change-Point in a

Sequence of Random Variables, *Biometrika*, 62, 407-416.

[15] Wright, D. E. and Hazelhurst, C. E. (1988), Estimation and Prediction for a Simple Software Reliability Model, *The Statistician*, 37, 319-325.

[16] Zhao, M. (1993), Change-Point Problems in Software and Hardware Reliability, *Communications in Statistics, Theory and Methods*, 22(3), 757-768.