ELSEVIER

Reliability Engineering and System Safety 72 (2001) 1–8

RELIABILITY ENGINEERING & SYSTEM SAFETY

www.elsevier.com/locate/ress

Field data analyses with additional after-warranty failure data ☆

Y.S. Oh, D.S. Bai*

Department of Industrial Engineering, Korea Advanced Institute of Science and Technology, Gusong-dong 373-1, Yusong-gu, Taejon 305-701, South Korea

Received 28 September 1999; accepted 18 May 2000

Abstract

This paper proposes methods of estimating the lifetime distribution for situations where additional field data can be gathered after the warranty expires in a parametric time to failure distribution. For satisfactory inference about parameters involved, it is desirable to incorporate these after-warranty data in the analysis. It is assumed that after-warranty data are reported with probability $p_1(<1)$, while within-warranty data are reported with probability 1. Methods of obtaining maximum likelihood estimators are outlined, their asymptotic properties are studied, and specific formulas for Weibull distribution are obtained. An estimation procedure using the expectation and maximization algorithm is also proposed when the reporting probability is unknown. Simulation studies are performed to investigate the properties of the estimates. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Field data; After-warranty data; Reporting probability; EM algorithm

1. Introduction

Life data analyses are commonly used to estimate the lifetime distribution of a product and to obtain information on the life characteristics such as reliability, failure rate, percentile and mean time to failure, etc. This information is then used in developing new products or improving the reliability of existing products, in designing burn-in and warranty programs, and in planning the supply of replacement parts. Life data can be obtained from life testing at laboratory, warranty claims or failure-records in the field use. Most of previous works on life data analyses utilized laboratory life test data. However, many times field data is superior to laboratory data because it captures actual usage profiles and the combined environmental exposures that are difficult to simulate in the laboratory and it is more likely to observe longer time-to-failures. Thus it is important to develop procedures for the collection and analyses of the field data [1].

Field data can be obtained from requests for repair when item failures occur within warranty period. For example, when an automobile fails during the warranty period, its owner will seek to have it repaired by the manufacturer. In this way, the manufacturer can accumulate records of failure times, failure modes, models and user characteristics, etc. Several authors considered the problems of analyzing such data. Suzuki [2,3] proposed parametric and nonparametric methods of estimating lifetime distribution from field failure data with supplementary information about censoring times obtained from following up a portion of the products that survive warranty time. Kalbfleisch and Lawless [4] suggested procedures for the collection of field data and used a regression model to estimate lifetime distribution from field failure data with supplementary information about covariates. Hu and Lawless [5] developed an estimation procedure with supplementary information about covariates and censoring times. Kalbfleisch et al. [6] proposed methods of analysis and prediction of warranty claims with reporting delays, and Hu et al. [7] extended Hu and Lawless [5] to the nonparametric case. Jones and Hayes [8] proposed practical methods of analyzing field data from large databases and David and Kieron [9] presented methods of testing the suitability of the exponential distribution for grouped field data.

All these works on the field data analyses use only failure and supplementary data within warranty period. A large amount of after-warranty failure data, however, become available as well as within-warranty failure data since consumers often have their repairs done by the original manufacturer even after the warranty expires. Moreover, there are many time-dependent failure mechanisms (wear-out) that are very unlikely to occur within the warranty period because of the time required for the physical mechanism to lead to a failure. Parameter estimates based

* Corresponding author. Tel.: +82-42-869-3111; fax: +82-42-869-3110. *E-mail address:* dsbai@mail.kaist.ac.kr (D.S. Bai).

☆ This research was supported by the Korea Sanhak Foundation.

PII: S0951-8320(00)00056-9

Nomenclature

N	total number of items
$\boldsymbol{\theta}$	column vector of parameters of lifetime distribution
$f(t;\boldsymbol{\theta})$	probability density function(pdf) of lifetime distribution
$S(t;\boldsymbol{\theta})$	survival function (sf) of lifetime distribution
T_1	warranty time
T_2	pre-specified time for analysis ($T_2 > T_1$)
p_1	reporting probability; probability that an item failed in $(T_1, T_2]$ is reported
D_1	set of items failed and reported in $(0, T_1]$
D_2	set of items failed and reported in $(T_1, T_2]$
D_3	set of items not reported in $(0, T_2]$
D_4	set of items failed but not reported in $(T_1, T_2]$
D_5	set of items not failed in $(0, T_2]$
n_j	number of items in D_j, $j = 1, 2, 3, 4, 5$.

only on warranty data could potentially over-estimate long-term reliability or under-estimate the number of failures. Therefore, if after-warranty data are available, it is desirable to incorporate these data with within-warranty data for remedying this deficiency even when after-warranty data is randomly reported. In fact, automotive or electronic companies maintain large service departments and have much after-warranty failure-data that are not fully utilized.

This paper proposes methods of estimating the lifetime distribution for situations where some additional field data can be gathered after warranty period in a parametric time-to-failure distribution. It is assumed that after-warranty data are reported to the original manufacturer with probability $p_1(<1)$, while within-warranty data are reported with probability 1. The methods of estimating the lifetime distribution from within-warranty failure data and additional after-warranty failure data are proposed in Section 2. Section 3 provides an estimation procedure using the expectation and maximization (EM) algorithm for the situation where the reporting probability is unknown. Simulation results on the properties of the estimates are presented in Section 4.

1.1. Assumptions

The following assumptions will be used in this paper.

1. The time scales of warranty time and failure time are the same.
2. Failure times of reported items are exactly observed with probability 1.
3. A failed item within warranty period is reported with probability 1.
4. A failed item after warranty period is reported with probability $p_1(<1)$.

2. Estimation of lifetime distribution

A failed item within warranty period is reported to the original manufacturer with probability 1. However, after the warranty expires, a failed item is reported with known probability p_1. The case of unknown p_1 is considered in Section 3. The likelihood function can be constructed as follows:

(i) Each failed item within warranty period $(0, T_1]$ contributes a term $f(t_i;\boldsymbol{\theta})$ to the likelihood.
(ii) Each failed item reported in $(T_1, T_2]$ contributes a term $p_1 f(t_i;\boldsymbol{\theta})$.
(iii) Each unreported item up to T_2 either fails in $(T_1, T_2]$ but is not reported (with probability $1 - p_1$) or survives (with probability $S(T_2;\boldsymbol{\theta})$). Its contribution is

$$(1 - p_1)\{S(T_1;\boldsymbol{\theta}) - S(T_2;\boldsymbol{\theta})\} + S(T_2;\boldsymbol{\theta})$$
$$= (1 - p_1)S(T_1;\boldsymbol{\theta}) + p_1 S(T_2;\boldsymbol{\theta}).$$

Using these contributions, the log-likelihood function becomes

$$\log L(\boldsymbol{\theta}) = \sum_{i\in D_1} \log\{f(t_i;\boldsymbol{\theta})\} + \sum_{i\in D_2} [\log p_1 + \log\{f(t_i;\boldsymbol{\theta})\}] + n_3 \log[(1 - p_1)S(T_1;\boldsymbol{\theta}) + p_1 S(T_2;\boldsymbol{\theta})]. \quad (1)$$

If all failures up to T_2 are reported, then $p_1 = 1$ and log-likelihood function (1) reduces to the usual one with censoring time T_2. Similarly, if no failures at interval $(T_1, T_2]$ are reported, then $p_1 = 0$ and formula (1) is simplified to the log-likelihood function with censoring time T_1.

The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained by maximizing Eq. (1), and under regularity conditions of Cramér [10] $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has a limiting multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})$, where

$$\mathbf{I}(\boldsymbol{\theta}) = \lim_{N\to\infty} \frac{1}{N} E\left[-\frac{\partial^2 \log L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right], \quad (2)$$

is the Fisher information matrix and $\boldsymbol{\theta}'$ is the transpose of

Table 1
Failure times

Within-warranty data			After-warranty data			
0.2038	0.5856	0.8382	1.0102	1.3456	1.6182	1.8837
0.2126	0.5966	0.8722	1.0406	1.3909	1.6428	1.9051
0.2279	0.6246	0.8909	1.1793	1.4068	1.6622	1.9139
0.2279	0.6256	0.8909	1.1914	1.4068	1.7211	1.9505
0.3473	0.6486	0.8984	1.2377	1.4261	1.7678	1.9552
0.4900	0.6689	0.9369	1.2456	1.4320	1.7724	1.9945
0.5006	0.6914	0.9541	1.2617	1.4878	1.7857	
0.5128	0.7941	0.9851	1.2810	1.5673	1.7861	
0.5754	0.7949		1.3084	1.5752	1.7971	
0.5829	0.8344		1.3418	1.5937	1.8723	

vector $\boldsymbol{\theta}$. $\mathbf{I}(\boldsymbol{\theta})$ is consistently estimated by the observed Fisher information matrix:

$$\begin{aligned}\mathbf{I}_N(\hat{\boldsymbol{\theta}}) &= -\frac{1}{N}\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= -\frac{1}{N}\left[\sum_{i\in D_1}\frac{\partial^2 \log(f(t_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &\quad -\frac{1}{N}\left[\sum_{i\in D_2}\frac{\partial^2 \log(p_1 f(t_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &\quad -\frac{1}{N}\left[n_3\frac{\partial^2 \log\{(1-p_1)S(T_1;\boldsymbol{\theta})+p_1 S(T_2;\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}\, \partial \hat{\boldsymbol{\theta}}}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\end{aligned} \tag{3}$$

2.1. Formulas for Weibull model with known reporting probability

If the lifetime t_i follows the Weibull distribution with scale parameter α and shape parameter β then its pdf and sf are

$$f(t_i;\alpha,\beta) = \alpha\beta t_i^{\beta-1}\exp(-\alpha t_i^{\beta}), \qquad t_i > 0, \tag{4}$$

$$S(t_i;\alpha,\beta) = \exp(-\alpha t_i^{\beta}), \qquad t_i > 0, \tag{5}$$

respectively. The log-likelihood function is then

$$\log L(\alpha,\beta) = \sum_{i\in D_1\cup D_2}(\log\alpha + \log\beta + (\beta-1)\log t_i - \alpha t_i^{\beta}) + n_2\log p_1 + n_3\log[(1-p_1)\exp(-\alpha T_1^{\beta}) + p_1\exp(-\alpha T_2^{\beta})]. \tag{6}$$

The first partial derivatives of Eq. (6) are given in Eqs. (A1)and (A2). The MLEs $\hat{\alpha}$ and $\hat{\beta}$ of α and β can be obtained by using a numerical method such as Newton–Raphson method. The observed Fisher information matrix $\mathbf{I}_N(\hat{\alpha},\hat{\beta})$ is

$$\mathbf{I}_N(\hat{\alpha},\hat{\beta}) = -\frac{1}{N}\begin{bmatrix}\dfrac{\partial^2\log L}{\partial\alpha^2} & \dfrac{\partial^2\log L}{\partial\alpha\,\partial\beta} \\ \dfrac{\partial^2\log L}{\partial\alpha\,\partial\beta} & \dfrac{\partial^2\log L}{\partial\beta^2}\end{bmatrix}_{\alpha=\hat{\alpha},\beta=\hat{\beta}}, \tag{7}$$

where the second partial derivatives are given in Eqs. (A3)–(A5).

2.2. Example 1

The proposed method is illustrated with failure data generated from Weibull distribution with parameters:

$$\alpha = 0.05, \qquad \beta = 2.0;$$

$$T_1 = 1.0, \qquad T_2 = 2.0;$$

$$N = 500;$$

$$p_1 = 0.5.$$

Table 1 contains the failure times of reported items; $n_1 = 28$ items failed before T_1, $n_2 = 36$ items failed between T_1 and T_2 were reported, and the remaining $n_3 = 436$ items were not reported until T_2. The MLEs are obtained using formulas (A1) and (A2) and Newton–Raphson method, their asymptotic variances are obtained from Eq. (7). The results are

$$\hat{\alpha} = 0.05778, \qquad \hat{\beta} = 1.9473,$$

$$As\ \mathrm{var}(\hat{\alpha}) = 0.9197\times10^{-4}, \qquad As\ \mathrm{var}(\hat{\beta}) = 0.04534.$$

95% confidence intervals of the parameters are

$$0.03898 \le \alpha \le 0.07658,$$

$$1.5299 \le \beta \le 2.3646.$$

3. Unknown reporting probability

In Section 2, the reporting probability was assumed to be known. That is, it can often be estimated fairly accurately from prior information. The reporting probability of a new product, however, is usually unknown and the method of Section 2 can not be directly used. In this section, we develop a method of estimating lifetime distribution and reporting probability.

When reporting probability p_1 is unknown, the log-likelihood function is the same as Eq. (1) except for the parameter space:

$$\log L(\boldsymbol{\theta}, p_1) = \sum_{i \in D_1} \log\{f(t_i; \boldsymbol{\theta}, p_1)\} + \sum_{i \in D_2} [\log p_1 + \log\{f(t_i; \boldsymbol{\theta}, p_1)\}] + n_3 \log[(1-p_1)S(T_1; \boldsymbol{\theta}, p_1) + p_1 S(T_2, \boldsymbol{\theta}, p_1)]. \tag{8}$$

In the log-likelihood functions (1) and (8), the contribution to the likelihood of set D_3 is used in place of the contributions of sets D_4 and D_5 because n_4 and n_5 are unobservable or missing.

When missing data are involved, the EM algorithm can be used to compute the MLEs iteratively. The EM algorithm always gives the MLEs if the pdf f is unimodal, and is very efficient.

On each iteration of the EM algorithm, there are two steps; the expectation step and the maximization step. In the expectation step, missing data are replaced by their conditional expectations given the observed data. In the maximization step, MLEs of the parameters are computed with the observed data and conditional expectations of the missing data calculated in the expectation step [11].

Initial step: Initial estimates of p_1 and $\boldsymbol{\theta}$ are, respectively, $p_1^{(0)}$ and $\boldsymbol{\theta}^{(0)}$.

Expectation step (k): Missing variable n_4 follows binomial distribution $b(n_3, p_3^{(k)})$, where

$$p_3^{(k)} = \frac{(1-p_1^{(k-1)})\{S(T_1; \boldsymbol{\theta}^{(k-1)}) - S(T_2; \boldsymbol{\theta}^{(k-1)})\}}{(1-p_1^{(k-1)})S(T_1; \boldsymbol{\theta}^{(k-1)}) + p_1^{(k-1)} S(T_2; \boldsymbol{\theta}^{(k-1)})} \tag{9}$$

In Eq. (9), the denominator is the probability that an item is not reported up to T_2 and the numerator is the probability that an item failed in $(T_1, T_2]$ is not reported. Thus, using $p_3^{(k)}$ the expectation of n_4 given n_3 at the kth iteration is

$$E^{(k)}(n_4) = n_3 \times p_3^{(k)}. \tag{10}$$

Similarly, the expectation of n_5 is

$$E^{(k)}(n_5) = n_3 - E^{(k)}(n_4). \tag{11}$$

Maximization step (k): The log-likelihood (8) is derived using observable n_3 in place of missing n_4 and n_5. Here the values of $E^{(k)}(n_4)$ and $E^{(k)}(n_5)$ obtained in formulae (10) and (11) can be used in deriving the contributions to the likelihood of sets D_4 and D_5. The kth step log-likelihood function now becomes

$$\log L^{(k)} = \sum_{i \in D_1} \log\{f(t_1; \boldsymbol{\theta}, p_1)\} + \sum_{i \in D_2} [\log p_1 + \log\{f(t_i; \boldsymbol{\theta}, p_1)\}] + E^{(k)}(n_4) \cdot [\log(1-p_1) + \log\{S(T_1; \boldsymbol{\theta}, p_1) - S(T_2; \boldsymbol{\theta}, p_1)\}] + E^{(k)}(n_5) \cdot \log\{S(T_2; \boldsymbol{\theta}, p_1)\}. \tag{12}$$

The values $(p_1^{(k)}, \boldsymbol{\theta}^{(k)})$ of $(p_1, \boldsymbol{\theta})$ that maximize log-likelihood are chosen. In particular

$$p_1^{(k)} = \frac{n_2}{n_2 + E^{(k)}(n_4)}, \tag{13}$$

$\boldsymbol{\theta}^{(k)}$ can be obtained by the method of Section 2.

As the iteration of the expectation and the maximization steps progresses, the estimates of p_1 and $\boldsymbol{\theta}$ converge to the stationary solutions. If the likelihood function is unimodal, the stationary solution of the algorithm is the unique MLE [12].

For the asymptotic variances of the MLEs, the method in Section 2 can be directly used. We can obtain the observed Fisher information matrix with the second derivatives of Eq. (8) and the MLEs in EM algorithm, the observed Fisher information matrix $\mathbf{I}_N(\hat{\alpha}, \hat{\beta}, \hat{p}_1)$ is

$$\mathbf{I}_N(\hat{\alpha}, \hat{\beta}, \hat{p}_1) = -\frac{1}{N} \begin{bmatrix} \dfrac{\partial^2 \log L}{\partial \alpha^2} & \dfrac{\partial^2 \log L}{\partial \alpha\, \partial \beta} & \dfrac{\partial^2 \log L}{\partial \alpha\, \partial p_1} \\ \dfrac{\partial^2 \log L}{\partial \alpha\, \partial \beta} & \dfrac{\partial^2 \log L}{\partial \beta^2} & \dfrac{\partial^2 \log L}{\partial \beta\, \partial p_1} \\ \dfrac{\partial^2 \log L}{\partial \alpha\, \partial p_1} & \dfrac{\partial^2 \log L}{\partial \beta\, \partial p_1} & \dfrac{\partial^2 \log L}{\partial p_1^2} \end{bmatrix}_{\alpha=\hat{\alpha}, \beta=\hat{\beta}, p_1=\hat{p}_1}, \tag{14}$$

where the second partial derivatives are given in Appendix A.

3.1. Example 2

The EM algorithm is illustrated with the data of Example 1 when p_1 is unknown.

Initial step: Initial the estimates of p_1, $\boldsymbol{\theta} = (\alpha, \beta)$ are chosen. Let $p_1^{(0)} = 0.4$, $\alpha^{(0)} = 0.1$, $\beta^{(0)} = 2.5$.

Expectation step (1): With the initial estimates, $p_3^{(1)}$, $E^{(1)}(n_4)$ and $E^{(1)}(n_5)$ are computed from Eqs. (9)–(11), respectively:

$$p_3^{(1)} = \frac{\{1-p_1^{(0)}\}\{S(T_1; \boldsymbol{\theta}^{(0)}) - S(T_2; \boldsymbol{\theta}^{(0)})\}}{p_1^{(0)} S(T_2; \boldsymbol{\theta}^{(0)}) + \{1-p_1^{(0)}\} S(T_1; \boldsymbol{\theta}^{(0)})} = 0.7656,$$

Table 2
Performances of the MLEs (p_1 known)

True values			MLEs		Biases		MSEs	
α	β	p_1	$\hat{\alpha}$	$\hat{\beta}$	bias($\hat{\alpha}$)×10^{-4}	bias($\hat{\beta}$)×10^{-4}	MSE($\hat{\alpha}$)×10^{-4}	MSE($\hat{\beta}$)×10^{-2}
0.05	1.0	0.2	0.0497	1.0263	−2.5865	2.6270	0.8956	3.0994
		0.5	0.0497	1.0226	−3.4750	2.2572	0.8020	2.5100
		0.8	0.0497	1.0222	−3.0807	2.2189	0.7393	2.2548
	2.0	0.2	0.0496	2.0274	−4.1936	2.7384	0.7961	6.9535
		0.5	0.0497	2.0158	−3.4888	1.5760	0.6821	4.5243
		0.8	0.0498	2.0152	−2.2311	1.5152	0.5822	3.5819
	3.0	0.2	0.0500	3.0089	0.4007	0.8943	0.7763	9.2992
		0.5	0.0500	3.0086	0.4529	0.8570	0.6183	5.5102
		0.8	0.0498	3.0158	−2.2185	1.5753	0.5630	4.7564
0.1	1.0	0.2	0.0996	1.0163	−4.4732	1.6292	1.7564	1.6336
		0.5	0.0992	1.0100	−7.7375	0.9979	1.5480	1.2656
		0.8	0.0994	1.0097	−6.4710	0.9702	1.4233	1.0672
	2.0	0.2	0.0998	2.0104	−1.7628	1.0376	1.6203	3.9673
		0.5	0.0997	2.0072	−3.1903	0.7161	1.3724	2.5753
		0.8	0.0997	2.0058	−2.5189	0.5779	1.2095	2.0338
	3.0	0.2	0.0997	3.0149	−3.4001	1.4857	1.6677	5.9479
		0.5	0.0996	3.0136	−3.7434	1.3567	1.3287	3.3336
		0.8	0.0998	3.0122	−2.0053	1.2189	1.0761	2.2509
0.2	1.0	0.2	0.1991	1.0047	−8.5084	0.4671	3.7536	0.8448
		0.5	0.1992	1.0052	−7.6930	0.5195	3.3695	0.6960
		0.8	0.1991	1.0042	−8.7533	0.4165	3.0912	0.5986
	2.0	0.2	0.1993	2.0083	−6.9808	0.8273	3.9269	2.5570
		0.5	0.1996	2.0090	−4.2799	0.8955	3.3720	1.7263
		0.8	0.1996	2.0079	−3.7460	0.7884	2.9608	1.2713
	3.0	0.2	0.1988	3.0226	−12.1245	2.2554	3.5881	5.3038
		0.5	0.1989	3.0186	−10.9775	1.8633	3.1117	3.0199
		0.8	0.1991	3.0170	−9.3901	1.7007	2.6568	1.8610

$$E^{(1)}(n_4) = n_3 \times p_3^{(1)} = 333.8,$$

$$E^{(1)}(n_5) = n_3 - E^{(1)}(n_4) = 102.2.$$

Maximization step (1): With $E^{(1)}(n_4)$ and $E^{(1)}(n_5)$, $\alpha^{(1)}$, $\beta^{(1)}$ and $p_1^{(1)}$ that maximize Eq. (12) are computed as $\alpha^{(1)} = 0.08282$, $\beta^{(1)} = 4.2309$, $p_1^{(1)} = 0.09736$. Computations are iterated until inequalities $|\alpha^{(k+1)} - \alpha^{(k)}| < 10^{-5}$, $|\beta^{(k+1)} - \beta^{(k)}| < 10^{-5}$ and $|p_1^{(k+1)} - p_1^{(k)}| < 10^{-5}$ are satisfied. The stationary solution

$$\hat{\alpha} = 0.05769, \qquad \hat{\beta} = 1.7818, \qquad \hat{p}_1 = 0.5026,$$

are obtained after 55 iterations and 9 s on a Pentium II PC with 300 MHz.

For the asymptotic variances, we use the observed Fisher information matrix $\mathbf{I}_N(\hat{\alpha}, \hat{\beta}, \hat{p}_1)$, which is the negative of the second partial derivatives of the log-likelihood. Using Eq. (14), we obtain

$$\mathbf{I}_N(\hat{\alpha}, \hat{\beta}, \hat{p}_1) = \frac{1}{500}\begin{bmatrix} 18354.6538 & 527.7261 & 1234.6337 \\ 527.7261 & 37.2583 & 66.6975 \\ 1234.6337 & 66.6975 & 154.0746 \end{bmatrix}_{\alpha=\hat{\alpha},\beta=\hat{\beta},p_1=\hat{p}_1},$$

and the asymptotic variances are

$$As\ \mathrm{var}(\hat{\alpha}) = 0.1181 \times 10^{-3}, \qquad As\ \mathrm{var}(\hat{\beta}) = 0.1193,$$

$$As\ \mathrm{var}(\hat{p}_1) = 0.03711.$$

These results are similar to those in Example 1.

4. Numerical study

We investigate, by Monte Carlo simulation, the properties of the estimators in terms of the bias and mean square error (MSE) for a Weibull distribution. One thousand estimates were computed, and deviations and squared deviations of the estimate from true value were averaged to obtain

Table 3
Performances of the MLEs (p_1 unknown)

True values			MLEs			Biases			MSEs		
α	β	p_1	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_1$	bias($\hat{\alpha}$)×10^{-4}	bias($\hat{\beta}$)×10^{-2}	bias($\hat{p}_1$)×10^{-2}	MSE($\hat{\alpha}$)×10^{-4}	MSE($\hat{\beta}$)×10^{-2}	MSE($\hat{p}_1$)$-\times 10^{-2}$
0.05	1.0	0.2	0.0504	1.0492	0.1990	3.5149	4.9176	−0.1044	1.0062	4.8030	1.5285
		0.5	0.0504	1.0507	0.4963	4.1841	5.0682	−0.3658	0.9843	4.5659	4.2689
		0.8	0.0510	1.0625	0.7462	9.7199	6.2509	−5.3767	0.9111	4.2087	4.5533
	2.0	0.2	0.0503	2.1027	0.1988	2.9360	10.2674	−0.1173	0.9577	16.7430	0.8497
		0.5	0.0505	2.0946	0.4911	4.6981	9.4620	−0.8928	0.9190	15.6255	2.9835
		0.8	0.0510	2.1107	0.7475	10.4245	11.0679	−5.2465	0.8682	11.7073	3.5529
	3.0	0.2	0.0506	3.1452	0.2025	5.8192	14.5213	0.2546	0.9786	40.5923	0.9082
		0.5	0.0508	3.1282	0.4891	8.4753	12.8164	−1.0882	0.8720	28.5553	2.3491
		0.8	0.0511	3.1262	0.7556	10.9956	12.6210	−4.4385	0.7622	20.2234	2.4643
0.1	1.0	0.2	0.1001	1.0261	0.2010	0.6082	2.6098	0.0999	1.9817	2.2885	0.7179
		0.5	0.1001	1.0253	0.4966	1.2437	2.5334	−0.3426	1.9517	2.1417	2.3282
		0.8	0.1007	1.0307	0.7728	6.8437	3.0747	−2.7168	1.7870	1.8948	2.7571
	2.0	0.2	0.1005	2.0463	0.1999	4.9172	4.6252	−0.0074	1.9653	7.9466	0.4042
		0.5	0.1005	2.0447	0.4978	4.6028	4.4705	−0.2195	1.8519	7.5731	1.5839
		0.8	0.1014	2.0647	0.7669	14.0179	6.4699	−3.3127	1.7617	6.2319	2.0370
	3.0	0.2	0.1000	3.0709	0.2031	−0.2500	7.0944	0.3116	2.0063	17.8891	0.3555
		0.5	0.1002	3.0804	0.4999	2.4882	8.0439	−0.0114	1.7685	16.3654	1.3121
		0.8	0.1006	3.0992	0.7741	5.7698	9.9167	−2.5852	1.5125	11.0876	1.2793
0.2	1.0	0.2	0.1995	1.0082	0.2008	−4.5157	0.8201	0.0831	4.1815	1.0276	0.3638
		0.5	0.1996	1.0089	0.5048	−4.1971	0.8910	0.4789	4.1051	0.9879	1.1605
		0.8	0.2000	1.0117	0.7944	0.1546	1.1662	−0.5604	3.8606	0.9065	1.5886
	2.0	0.2	0.1997	2.0217	0.2010	−2.6796	2.1662	0.1044	4.6123	3.9680	0.1838
		0.5	0.1998	2.0199	0.5035	−2.0113	1.9942	0.3510	4.3554	3.3951	0.6425
		0.8	0.2008	2.0342	0.7902	7.6356	3.4205	−0.9762	4.0272	3.2840	0.9775
	3.0	0.2	0.1983	3.0261	0.2038	−16.6381	2.6107	0.3786	3.9148	7.7889	0.1018
		0.5	0.1985	3.0215	0.5066	−15.0537	2.1548	0.6614	3.5437	6.4752	0.3528
		0.8	0.1990	3.0340	0.8009	−9.9988	3.4019	0.0864	3.1413	4.9251	0.3941

(estimated) bias and MSE for the following parameters:

$N = 500,$

$T_1 = 1.0,$

$T_2 = 2.0;$

$\alpha = 0.05, 0.1, 0.2;$

$\beta = 1.0, 2.0, 3.0;$

$p_1 = 0.2, 0.5, 0.8.$

The results of the simulation for p_1 known and unknown cases are, respectively, shown in Tables 2 and 3. One can see from the these tables that (i) the (estimated) biases and MSEs of the MLEs are found to be very small, (ii) the (estimated) MSEs for the case of known p_1 are always smaller than those for the case of unknown p_1, (iii) the (estimated) MSEs decrease as p_1 increases.

4.1. Effects of the reporting probability

We also investigate the effect of reporting probability p_1 on the estimation accuracy. Tables 4 and 5, respectively, show MSEs of the MLEs for p_1 known and unknown

Table 4
Effects of the reporting probability (p_1 known)

	MSEs	
p_1	MSE($\hat{\alpha}$)$\times 10^{-4}$	MSE($\hat{\beta}$) $\times 10^{-2}$
*	1.9884	8.9304
0.1	1.7435	5.2514
0.2	1.6203	3.9673
0.3	1.5225	3.3287
0.5	1.3724	2.5753
0.7	1.2572	2.1608
0.8	1.2095	2.0338
0.9	1.1601	1.8877

Table 5
Effects of the reporting probability (p_1 unknown)

	MSEs		
p_1	MSE($\hat{\alpha}$) $\times 10^{-4}$	MSE($\hat{\beta}$) $\times 10^{-2}$	MSE($\hat{p}_1$) $\times 10^{-2}$
0.1	1.9889	8.6247	0.1509
0.2	1.9653	7.9466	0.4042
0.3	1.9317	7.3045	0.7379
0.5	1.8519	7.5731	1.5839
0.7	1.7694	6.9160	2.0335
0.8	1.7617	6.2319	2.0370
0.9	1.6772	5.6481	2.0194

cases when $\alpha = 0.1$ and $N = 500$. MSEs of $\hat{\alpha}$ and $\hat{\beta}$ become smaller as p_1 increases. In Table 4, $p_1 = *$ is the case of using only within-warranty data. The MSE of $\hat{\alpha}$ (or $\hat{\beta}$) when $p_1 = *$ is larger than the MSE of $\hat{\alpha}$ (or $\hat{\beta}$) obtained with additional after-warranty data whether reporting probability is known or not.

5. Conclusions

We have proposed methods of estimating the lifetime distributions incorporating additional after-warranty failure data in a parametric time-to-failure distribution. Expectation and maximization algorithms are used to estimate the lifetime distribution and the reporting probability simultaneously when p_1 is unknown. Although we used a Weibull distribution to demonstrate the methods, it can be used for any parametric lifetime distribution. Monte Carlo simulations show that the proposed methods give fairly accurate estimates even in the incomplete reporting situation, and the estimators obtained with additional after-warranty data perform better than estimators with only within-warranty data. Therefore, whenever after-warranty data are available, it is better to incorporate these data with within-warranty data.

Field data, however, can often be messy and unusable. Specially, (a) field data does not always isolate the failed component because a line-replaceable-unit containing many components is discarded after failure, (b) it is very difficult to separate actual failures from components that were damaged by adjacent failed components or by user abuse, and (c) failure times are not always faithfully recorded because there are no elapsed time meters attached to many components, or even systems. Therefore the proposed methods need cautions in that they apply to the field data that has been carefully collected when it has been reported.

Acknowledgements

The authors are grateful to the anonymous reviewers for their valuable comments which led to an improvement of this paper.

Appendix A. Partial derivatives

The first partial derivatives of Eq. (6) with respective to α and β are

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i \in D_1 \cup D_2} \left(\frac{1}{\alpha} - t_i^{\beta} \right) + n_3 \frac{P_\alpha}{P}, \tag{A1}$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i \in D_1 \cup D_2} \left(\frac{1}{\beta} + \log t_i - \alpha t_i^{\beta} \log t_i \right) + n_3 \frac{P_\beta}{P}, \tag{A2}$$

where

$$P = p_1 \exp(-\alpha T_2^{\beta}) + (1 - p_1) \exp(-\alpha T_1^{\beta}).$$

$$P_\alpha = \frac{\partial P}{\partial \alpha} = -p_1 T_2^{\beta} \exp(-\alpha T_2^{\beta}) - (1 - p_1) T_1^{\beta} \exp(-\alpha T_1^{\beta}),$$

$$P_\beta = \frac{\partial P}{\partial \beta} = -p_1 \alpha T_2^{\beta} \log T_2 \exp(-\alpha T_2^{\beta}) - (1 - p_1) \alpha T_1^{\beta} \log T_1 \exp(-\alpha T_1^{\beta}).$$

The second partial derivatives of Eq. (8) with respect to α, β and p_1 are

$$\frac{\partial^2 \log L}{\partial \alpha^2} = -\frac{n_1 + n_2}{\alpha^2} + n_3 \frac{P_{\alpha\alpha} P - P_\alpha^2}{P^2} = -\frac{n_1 + n_2}{\alpha^2} + n_3 \left\{ \frac{p_1(1 - p_1) \exp(-\alpha T_2^{\beta}) \exp(-\alpha T_1^{\beta})(T_1^{\beta} - T_2^{\beta})^2}{P^2} \right\}, \tag{A3}$$

$$\frac{\partial^2 \log L}{\partial \beta^2} = -(n_1 + n_2) \left\{ \frac{1}{\beta^2} + \alpha t_i^{\beta} (\log t_i)^2 \right\} + n_3 \left\{ \frac{P_{\beta\beta} P - P_\beta^2}{P^2} \right\}, \tag{A4}$$

$$\frac{\partial^2 \log L}{\partial \alpha\, \partial \beta} = -(n_1 + n_2)\{t_i^{\beta} \log t_i\} + n_3 \left\{ \frac{P_{\alpha\beta} P - P_\alpha P_\beta}{P^2} \right\}, \tag{A5}$$

$$\frac{\partial^2 \log L}{\partial p_1^2} = -\frac{n_2}{p_1^2} - n_3 \left\{ \frac{P_{p_1}^2}{P^2} \right\}, \tag{A6}$$

$$\frac{\partial^2 \log L}{\partial \alpha\, \partial p_1} = n_3 \left\{ \frac{P_{\alpha p_1} P - P_\alpha P_{p_1}}{P^2} \right\}, \tag{A7}$$

$$\frac{\partial^2 \log L}{\partial \beta\, \partial p_1} = n_3 \left\{ \frac{P_{\beta p_1} P - P_\beta P_{p_1}}{P^2} \right\}, \tag{A8}$$

where

$$P_{\alpha a} = \frac{\partial^2 P}{\partial \alpha^2} = p_1 T_2^{2\beta} \exp(-\alpha T_2^\beta) + (1-p_1) T_1^{2\beta} \exp(-\alpha T_1^\beta),$$

$$P_{\beta\beta} = \frac{\partial^2 P}{\partial \beta^2} = p_1 \alpha T_2^\beta (\log T_2)^2 \exp(-\alpha T_2^\beta)(\alpha T_2^\beta - 1)$$
$$+(1-p_1)\alpha T_1^\beta (\log T_1)^2 \exp(-\alpha T_1^\beta)(\alpha T_1^\beta - 1),$$

$$P_{\alpha\beta} = \frac{\partial^2 P}{\partial \alpha\, \partial \beta} = p_1 T_2^\beta \log T_2 \exp(-\alpha T_2^\beta)(\alpha T_2^\beta - 1)$$
$$+(1-p_1) T_1^\beta \log T_1 \exp(-\alpha T_1^\beta)(\alpha T_1^\beta - 1).$$

$$P_{p_1} = \frac{\partial P}{\partial p_1} = \exp(-\alpha T_2^\beta) - \exp(-\alpha T_1^\beta),$$

$$P_{\alpha p_1} = \frac{\partial^2 P}{\partial \alpha\, \partial p_1} = -T_2^\beta \exp(-\alpha T_2^\beta) + T_1^\beta \exp(-\alpha T_1^\beta),$$

$$P_{\beta p_1} = \frac{\partial^2 P}{\partial \beta\, \partial p_1} = -\alpha T_2^\beta \log T_2 \exp(-\alpha T_2^\beta)$$
$$+ \alpha T_1^\beta \log T_1 \exp(-\alpha T_1^\beta).$$

References

[1] Lawless JF. Statistical methods in reliability. Technometrics 1983;25:305–35.
[2] Suzuki K. Estimation of lifetime parameters from incomplete field data. Technometrics 1985;27:263–72.
[3] Suzuki K. Nonparametric estimation of lifetime distributions from a record of failures and follow-ups. Journal of American Statistical Association 1985;80:68–72.
[4] Kalbfleisch JD, Lawless JF. Estimation of reliability in field-performance studies. Technometrics 1988;30:365–88.
[5] Hu JX, Lawless JF. Estimation from truncated data with supplementary information on covariates and censoring times. Biometrika 1996;83:747–61.
[6] Kalbfleisch JD, Lawless JF, Robinson JA. Methods for the analysis and prediction of warranty claims. Technometrics 1991;33:273–85.
[7] Hu JX, Lawless JF, Suzuki K. Nonparametric estimation of a lifetime distribution when censoring times are missing. Technometrics 1998;40:3–13.
[8] Jones JA, Hayes JA. Use of a field Database for improvement of product reliability. Reliability Engineering and System Safety 1997;55:131–4.
[9] Coit DW, Dey KA. Analysis of grouped data from field-failure reporting systems. Reliability Engineering and System Safety 1999;65:95–101.
[10] Cramér H. Mathematical methods of statistics. Princeton, NJ: Princeton University Press, 1946.
[11] Geoffrey JM, Thriyambakam K. The EM algorithm and extensions. New York: Wiley, 1997.
[12] Wu C. On the convergence property of the EM algorithm. Annals of Statistics 1983;11:95–103.