ELSEVIER

Analysis of Lognormal Survival Data

RAMESH C. GUPTA
Department of Mathematics and Statistics, University of Maine, Orono, Maine 04469-5752

AND

NANDINI KANNAN AND APARNA RAYCHAUDHURI
Division of Mathematics, Computer Science and Statistics, University of Texas, San Antonio, Texas 78249

Received 20 November 1995; revised 29 May 1996

ABSTRACT

The failure rate and the mean residual life function (MRLF) of a lognormal distribution are known to be nonmonotonic. It is of interest to study the point at which the monotonicity changes (the change point). In this article we study the change points of the failure rate and the MRLF for the lognormal distribution. It is shown that the change points are the solutions of certain nonlinear equations. We apply these results to estimate the change points for survival data on guinea pigs given by Bjerkedal. The standard deviation of the estimate is obtained using bootstrap and jackknife methods. Finally confidence bands for the failure rate and the MRLF are also provided to illustrate the behavior of the estimates. © *Elsevier Science Inc., 1997*

1. INTRODUCTION

Survival and failure time data are frequently modeled by either increasing or decreasing failure rate distributions. This may be inappropriate where the course of disease is such that mortality reaches a peak after some finite period and then declines slowly. For example, in a study of curability of breast cancer, Langlands et al. [1] found that the peak mortality occurred after about 3 years. Bennett [2] analyzed the data from the Veterans Administration lung cancer trial presented by Prentice [3] and showed that the empirical failure rates for both low- and high-performance status (PS) groups are nonmonotonic [3]. Therefore it is important to analyze such data sets with appropriate models like the lognormal, inverse Gaussian, loglogistic, and Burr type XII.

Several investigators, such as Osgood [4], Feinleib and MacMohan [5], and Feinleib [6] observed that the distribution of the survival time of

655 Avenue of the Americas, New York, NY 10010
0025-5564/97/$17.00
PII S0025-5564(96)00133-2

different diseases, such as Hodgkin's disease and chronic leukemia, may be closely approximated by a lognormal distribution. Since the survival times are markedly skewed to the right, the logarithms of survival times are approximately normally distributed. Horner [7] showed that the distribution of age at the onset of Alzheimer's disease followed the lognormal distribution. The model is also useful for competing risk lifetimes related to some types of surgery where one experiences lower failure rate before surgery, rising to a peak during or shortly after surgery and then trailing off to 0. However, it may be mentioned that the lognormal will give a poor fit if the death takes place due to the competing risk of general aging. In this case, the failure rate will not trail off to 0.

In this article, we consider the lognormal model to study the change point of the failure rate (hazard function) and the mean residual life function. Here the change point refers to the age at which the gradient of the failure rate or mean residual life function changes sign (direction). The studies described in the previous paragraph indicate that it is important to estimate these change points.

The failure rate of the lognormal distribution increases initially to a maximum and then decreases to zero as time approaches infinity. On the other hand, the mean residual life function (MRLF) exhibits a reverse behavior: it decreases initially and then increases. For a general discussion of such a behavior, see [8].

The purpose of the present article is to estimate the change point of the failure rate and the MRLF. In Section 2 we present the model and some of its structural properties, including the expressions for the failure rate and MRLF. In Section 3, a data set consisting of survival times of guinea pigs injected with different amounts of tubercle bacilli is analyzed. The estimation of change points is described in Section 4. The change points of failure rate and the MRLF are estimated by the maximum likelihood method. To obtain the standard error of the estimates, we resort to resampling techniques like the bootstrap and the jackknife. Finally, confidence bands for the failure rate and the MRLF are also provided to illustrate the behavior of the estimates.

2. THE MODEL

Let X be a random variable having a lognormal distribution. The probability density function and the survival function are given by

$$f_x(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-(1/2\sigma^2)(\ln t - \mu)^2}, \qquad t > 0, \sigma > 0, \tag{1}$$

$$\bar{F}(t) = 1 - \Phi\left[\frac{\ln(at)}{\sigma}\right], \tag{2}$$

where $a = e^{-\mu}$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

The failure rate is given by

$$r(t) = \frac{\left(1/\sqrt{2\pi}\, t\sigma\right) e^{-(\ln(at))^2/2\sigma^2}}{1 - \Phi[\ln(at)/\sigma]}. \tag{3}$$

See [9]. Note that if X follows lognormal distribution then $Y = \ln X$ has a normal distribution. Thus the mean and variance of X are given by

$$E(X) = e^{\mu + \sigma^2/2}$$
$$\operatorname{Var}(X) = 2^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}.$$

The median is e^{μ} and the mode $e^{\mu - \sigma^2}$.

The Mean Residual Life Function (life expectancy) of any nonnegative random variable X is defined as

$$\begin{aligned}\mu(t) &= E(X - t \mid X > t)\\ &= \frac{\int_t^\infty \bar{F}(x)\, dx}{\bar{F}(t)}.\end{aligned}$$

The failure rate and the MRLF are connected by the relation

$$r(t) = \frac{1 + \mu'(t)}{\mu(t)}. \tag{4}$$

The MRLF of X, the lognormal random variable, is derived as

$$\begin{aligned}\mu(t) &= E(X - t \mid X > t)\\ &= E(e^Y \mid Y > \ln t) - t,\end{aligned} \tag{5}$$

where $Y = \ln X$ follows normal distribution with mean μ and variance σ^2. The conditional moment-generating function of Y given $Y > b$ is

$$\begin{aligned}E(e^{sY} \mid Y > b) &= \frac{\int_b^\infty e^{sy}\left(1/\sqrt{2\pi}\,\sigma\right) e^{-(y-\mu)^2/2\sigma^2}\, dy}{P(Y > b)}\\ &= e^{\mu s + \sigma^2 s^2/2} \frac{\left[1 - \Phi((b - \mu - \sigma^2 s)/\sigma)\right]}{1 - \Phi((b - \mu)/\sigma)}.\end{aligned} \tag{6}$$

Thus, the MRLF of X is given by

$$\mu(t) = e^{\mu+\sigma^2/2}\frac{[1-\Phi((\ln t-\mu-\sigma^2)/\sigma)]}{1-\Phi((\ln t-\mu)/\sigma)} - t. \tag{7}$$

As mentioned earlier, $r(t)$ increases initially to a maximum and then decreases to zero as time approaches infinity. This fact can be proved by using the technique developed by Glaser [10], which is briefly outlined as follows: Define $\nu(t) = -f'(t)/f(t)$. Then $r(t)$ is of the above type U (defined later) if

a. there exists $t_0 > 0$ such that $\nu'(t) > 0$ for all $t \in (0,t_0), \nu'(t_0) = 0$, and $\nu'(t) < 0$ for all $t > t_0$ and

b. $\text{limit}_{t\to 0} f(t) = 0$.

To obtain the change point t^* of the failure rate, we note that

$$\frac{d}{dt}\ln r(t) = -\nu(t) + r(t).$$

So, the critical point t^* of the failure rate is a solution of

$$\nu(t) = r(t). \tag{8}$$

The MRLF has the shape reverse of the failure rate; i.e., MRLF decreases to its minimum at the point k^* and then steadily increases. This fact can be verified by using the following result of Gupta and Akman [8]. Before stating the theorem, we present a definition for functions of types U and B.

DEFINITION

A function g is said to be of type $U(B)$ if there exists a t^* such that $g^*(t) > (<)0$ for all $t\varepsilon(0,t^*), g'(t^*) = 0$, and $g'(t) < (>)0$ for all $t > t^*$. For details see [10].

THEOREM

If $r(t)$ is of the type U (as in our case), then $\mu(t)$ is of type B (reverse) if $r(0) < 1/\mu$, where μ is the mean.

The expression of $\mu(t)$ is rather complicated for the lognormal distribution. The change point k^* can be obtained as the solution to

$$r(t)\mu(t) = 1. \tag{9}$$

Gupta and Akman [8] have shown that $k^* < t^*$ as follows: Taking the derivatives of Eq. (4) and using the fact that $\mu'(k^*) = 0$ and $\mu''(k^*) > 0$, it follows that $r'(k^*) > 0$. Thus $k^* < t^*$.

3. EXAMPLE: ANALYSIS OF GUINEA PIGS DATA

The data set consists of survival times of guinea pigs injected with different amount of tubercle bacilli and was studied by Bjerkedal [11]. Guinea pigs are known to have high susceptibility to human tuberculosis, which is one of the reasons for choosing this species. We consider only the study in which animals in a single cage are under the same regimen. The regimen number is the common log of the number of bacillary units in 0.5 ml of challenge solution; e.g., regimen 4.3 corresponds to 2.2×10^4 bacillary units per 0.5 ml ($\log(2.2 \times 10^4) = 4.342$).

We considered the data for regimens 4.3 and 6.6. There were 72 observations under each regimen. These data sets did not contain any censored observations. The data sets are given in Table 1. The descriptive statistics for the two data sets are provided in Table 2.

The distribution of survival times was noticeably skewed and the imperical failure rate was found to be nonmonotomic, suggesting the lognormal distribution might be appropriate. We fit a lognormal distribution to both data sets. The $q-q$ plots are given in Fig. 1.

We tested the appropriateness of the lognormal distribution using the Kolmogorov–Smirnov (KS) test. For both regimen levels, the test did not reject the lognormal fit. Table 3 shows the value of the KS test statistic and the corresponding two-sided p-value.

TABLE 1

Survival Times of Guinea Pigs in Days

Regimen 4.3						Regimen 6.6					
10	100	116	153	197	254	12	44	60	70	95	146
33	102	120	159	202	254	15	48	60	72	96	175
44	105	121	160	213	278	22	52	60	73	98	175
56	107	122	163	215	293	24	53	60	75	99	211
59	107	122	163	216	327	24	54	61	76	109	233
72	108	124	168	222	342	32	54	62	76	110	258
74	108	130	171	230	347	32	55	63	81	121	258
77	108	134	172	231	361	33	56	65	83	127	263
92	109	136	176	240	402	34	57	65	84	129	297
93	112	139	183	245	432	38	58	67	85	131	341
96	113	144	195	251	458	38	58	68	87	143	341
100	115	146	196	253	555	43	59	70	91	146	376

TABLE 2

Descriptive Statistics from the Two Data Sets

	Regimen 4.3		Regimen 6.6	
	Mean	St. Dev	Mean	St. Dev
Raw data	176.82	102.73	99.82	80.55

The above information indicates that the difference between the lognormal distribution and the true distribution of survival times is sufficiently insignificant so as to remain undetected. We thus assume that the survival times in both the data sets follow a lognormal distribution.

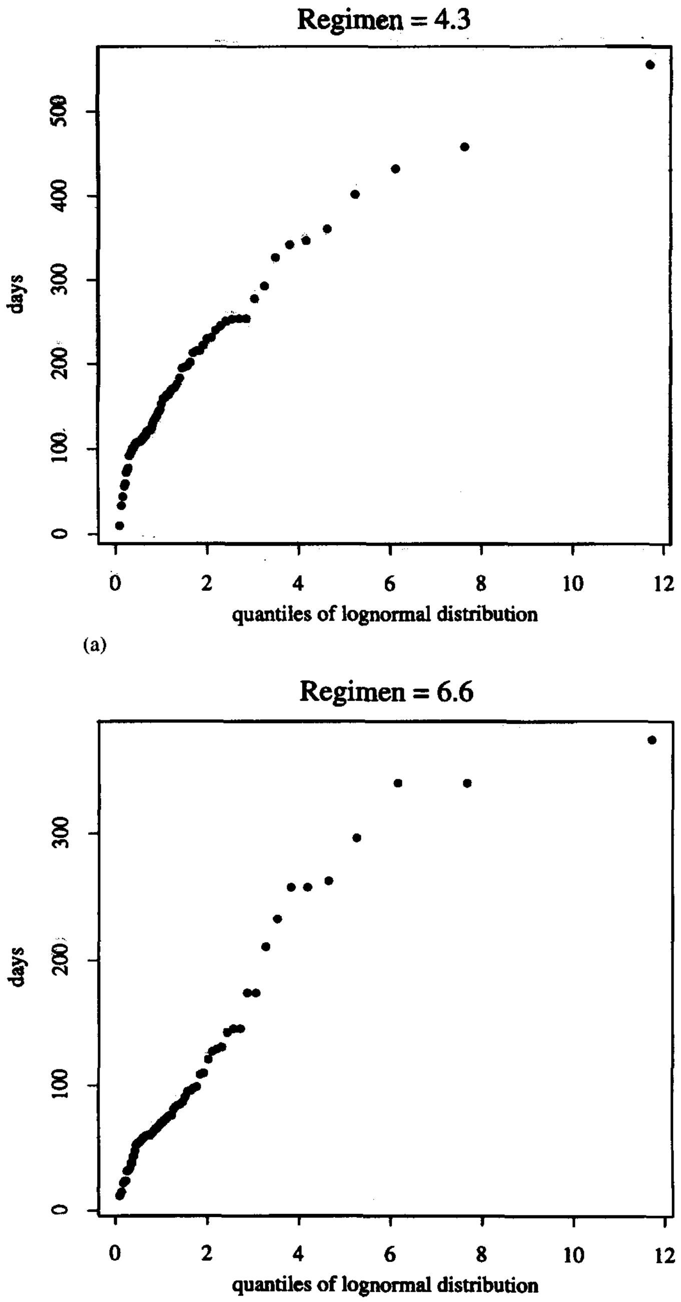
4. ESTIMATION OF THE CHANGE POINT

To compute the estimate of the change point for the failure rate and the MRLF, we first obtain the maximum likelihood estimates of μ and σ from the logarithm of the data. These estimates enable us to obtain the MLEs of the failure rate and the MRLF. The change point of the failure rate is obtained by optimizing the estimate of the failure rate. It is also possible to obtain the change point as a solution of the nonlinear equation (8). The change point for the MRLF is obtained in a similar manner. We could also obtain this as a solution to Eq. (9).

To draw inferences about the change point, we need to estimate the standard deviation. Since $r(t)$ and $\mu(t)$ are complicated functions of the parameters, it is computationally difficult to derive an expression for the variance of the maximum likelihood estimator. We therefore resort to standard resampling procedures like the bootstrap and the jackknife to answer this question. These computer-based methods involve considerable amounts of computation and can answer questions that are far too complicated for traditional analyses as in our case.

These techniques are briefly described here. We select B independent samples of size n (72 in our case), each sample being drawn from the observed data set with replacement. This is called a bootstrap sample. For each bootstrap sample, we compute the change points for the failure rate and MRLF. This is a bootstrap replication of the statistic. Let θ_i denotes the estimate from the ith sample; then the bootstrap estimate of standard deviation is given by

$$\hat{\sigma} = \left[\frac{1}{B-1} \sum_{i=1}^{B} \left(\theta_i - \bar{\theta}\right)^2\right]^{1/2}.$$



(a)

(b)

FIG. 1. *q*-*q* plots for guinea pigs data sets.

TABLE 3

KS Statistics and p-Values from the Two Data Sets

	Regimen 4.3	Regimen 6.6
KS statistic	30.95037	0.8222
p-Value	0.30497	0.4856

For estimating a standard error, B is taken to be between 250 and 500. See [12]. We have taken B to be 500. The average of the bootstrap replications is also provided as an "estimate" of the change point. It may be mentioned that taking B more than 500 did not affect the estimate of σ considerably.

For the jackknife estimate, we generate the so-called "jackknife samples." The ith jackknife sample consists of the data set with the ith observation removed. The change points are estimated from each sample and are called the jackknife replications. If θ_i denotes the estimate from the ith sample, then the jackknife estimate of standard deviation is given by

$$\hat{\sigma} = \left[\frac{n-1}{n} \sum_{i=1}^{n} \left(\theta_1 - \bar{\theta}\right)^2 \right]^{1/2},$$

where n is the number of samples (i.e., 72 in this case) and $\bar{\theta} = \Sigma_{i=1}^{72} \theta_i / n$. The average of the jackknife replications is also provided as an "estimate" of the change point. The results are given in Tables 4 and 5.

It is observed that in all cases, the standard deviations of the estimates obtained from the bootstrap samples are slightly lower than those obtained from the jackknife procedure. These results are in agreement with theoretical results on the bootstrap and the jackknife. Also, the maximum likelihood estimates of both the change points are

TABLE 4

Change Point Estimation of Failure Rates

	Regimen 4.3		Regimen 6.6	
	Estimate	St. Dev	Estimate	St. Dev
Jackknife	210.19	48.05	92.71	14.03
Bootstrap	217.47	41.87	95.23	13.82
Max. likelihood	210.080		92.6744	

TABLE 5

Change Point Estimation of Mean Residual Lifetime Function

	Regimen 4.3		Regimen 6.6	
	Estimate	St. Dev	Estimate	St. Dev
Jackknife	146.50	50.95	58.47	13.33
Bootstrap	156.50	40.96	59.95	13.28
Max. likelihood	146.34		58.44	

slightly lower than the corresponding jackknife and bootstrap estimates. It should also be noted that the change point of the failure rate exceeds that of the MRLF as has been indicated before.

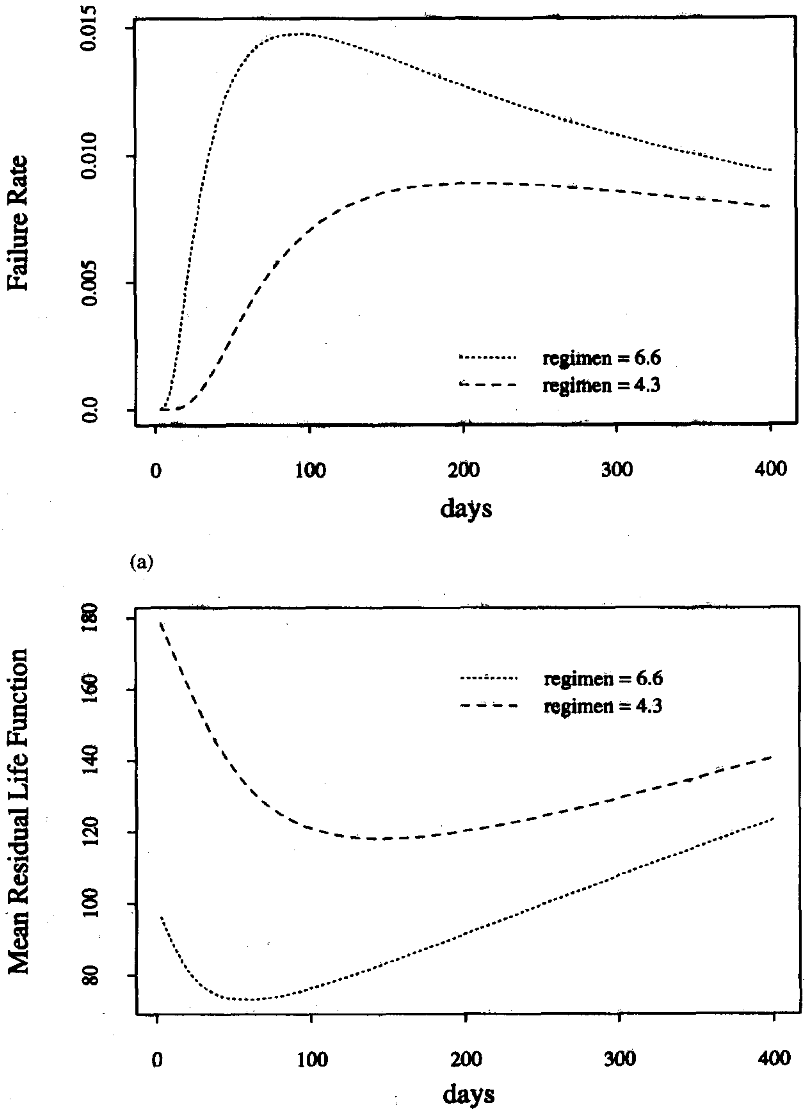
The failure rate and MRLF for both data sets using the maximum likelihood estimates of parameters are displayed in Fig. 2.

We also obtained 90% pointwise confidence bands for the failure rate and MRLF for both regimens. These bands are created from 500 bootstrap samples. For each sample, all three functions are evaluated at several points. For each of these points, the 5th and 95th percentiles of these estimates provide the lower and upper bound respectively. The confidence bands for regimen 4.3 and regimen 6.6 are displayed in Figs. 3 and 4 respectively.

5. CONCLUSION

We have estimated the change points of failure rate and MRLF for a log normal distribution. The change points are computed for two data sets. To estimate the standard deviation of these estimated change points, jackknife and bootstrap resampling techniques are employed. In general, jackknife standard deviation estimates tend to be more conservative. For the data with the guinea pigs, we observed that the change point for the failure rate for regimen 4.3 was greater than that for regimen 6.6. This seems to indicate that an infection initiated with a large number of bacilli progresses more rapidly, the peak mortality occurring much earlier for the higher regimen. It may be mentioned that similar methods can be used in the censored data case. The likelihood function can be written accordingly.

The authors are thankful to the referees for some useful comments, which enhanced the presentation.



(a)

(b)

FIG. 2. Failure rate and MRLF for the two data sets.

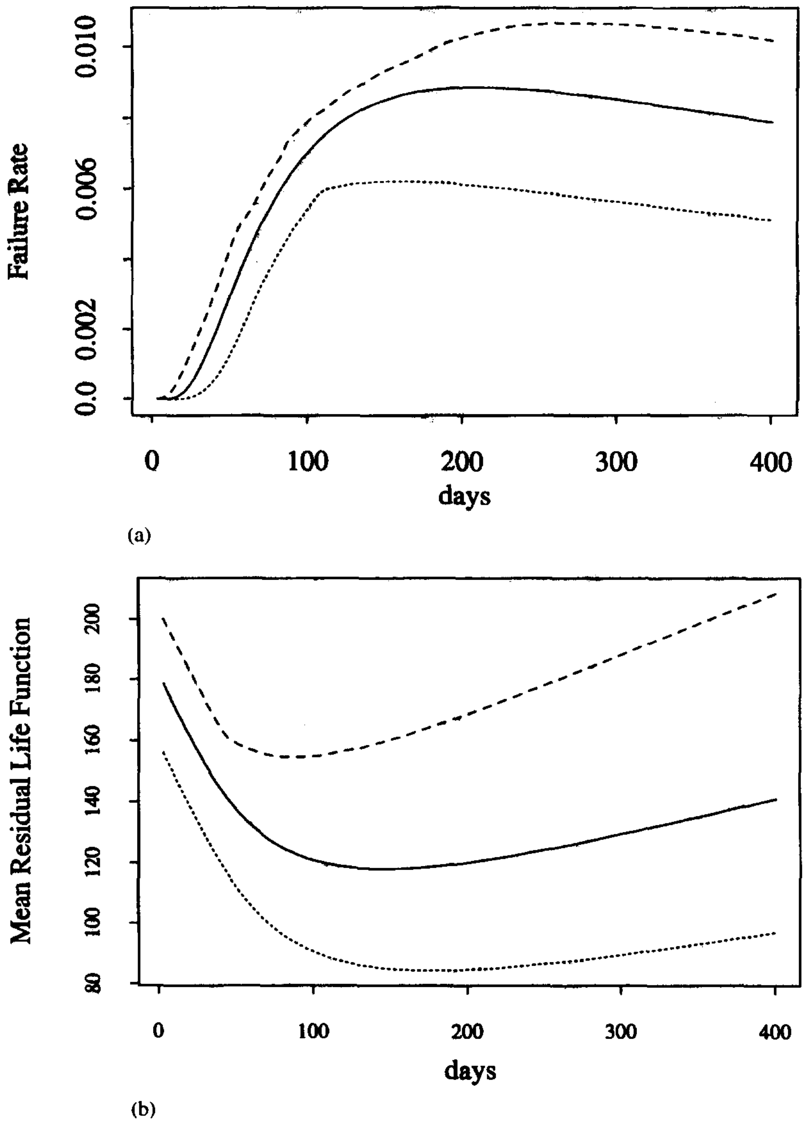


FIG. 3. Confidence bands for failure rate and MRLF (regimen 4.3).

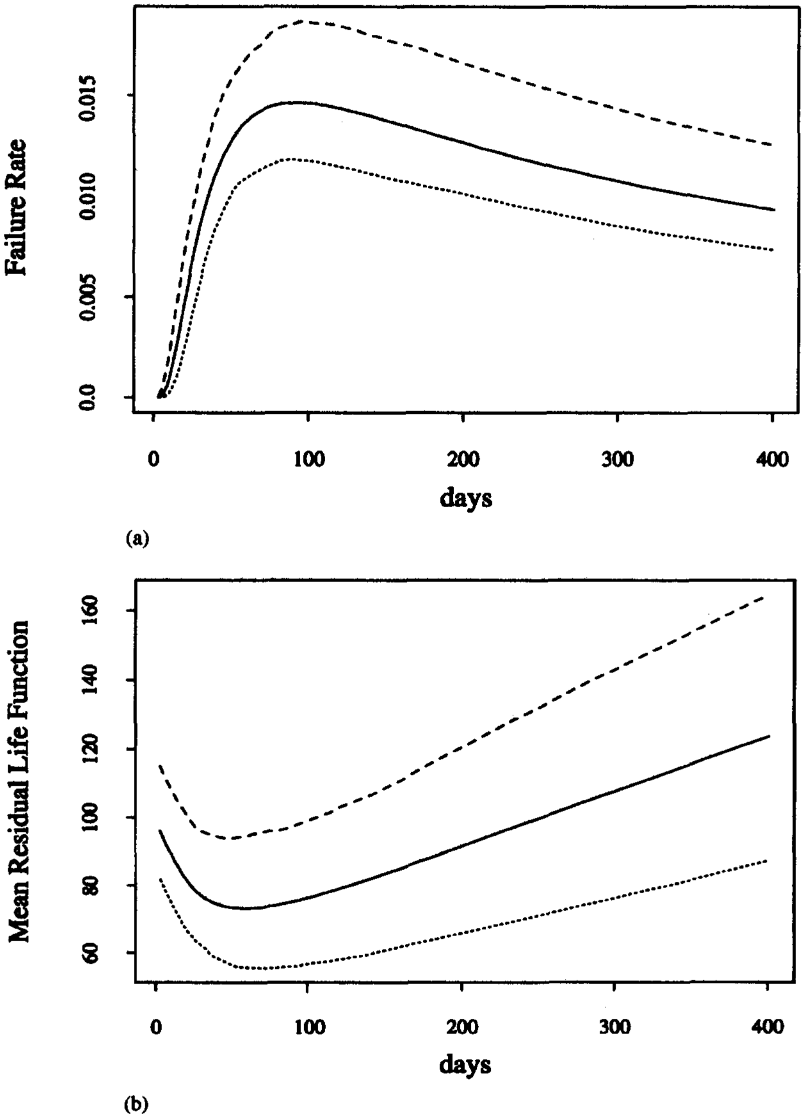


FIG. 4. Confidence bands for failure rate and MRLF (regimen 6.6).

REFERENCES

1 A. O. Langlands, S. J. Pocock, G. R. Kerr, and S. M. Gore, Long term survival of patients with breast cancer: A study of curability of the disease. *Brit. Med. J.*, 1247–1251 (1979).

2 S. Bennett, Loglogistic regression models for survival data. *Appl. Statist.* 32(2): 165–171 (1983).

3 R. L. Prentice, Exponential survivals with censoring and explanatory variables. *Biometrika* 60:279–288 (1973).

4 E. W. Osgood, Methods for analyzing survival data illustrated by Hodgkin's disease. *Amer. J. Med.* 24(24):40–47 (1958).

5 M. Feinleib and B. MacMohan, Variation in the duration of survival of patients with chronic leukamia. *Blood*, 332–349 (1960).

6 M. Feinleib, A method of analyzing log-normality distributed survival data with incomplete followup. *J. Amer. Statist. Assoc.* 55:534–545 (1960).

7 R. D. Horner, Age at onset of Alzheimer's disease: Clue to the relative importance of etiologic factors? *Amer. J. Epidemiol.* 126:409–414 (1987).

8 R. C. Gupta and O. Akman, Mean residual life functions for certain type of non-monotonic ageing. *Stochastic Models* 11(1):219–225 (1995).

9 Elisa T. Lee, *Statistical Methods for Survival Data Analysis*, Wiley, New York (1992).

10 R. E. Glaser, Bathtub and related failure rate characterizations. *J. Amer. Statist. Assoc.* 75:667–672 (1980).

11 T. Bjerkedal, Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilii. *Amer. J. Hyg.* 72:130–148 (1960).

12 B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall (1993).