Nonparametric Estimation of a Lifetime Distribution When Censoring Times are Missing

X. Joan Hu and Jerald F. Lawless
University of Waterloo
and
Kazuyuki Suzuki
University of Electro-Communications - Tokyo

I.I.Q.P. Research Report
RR-96-04
July 1996

Nonparametric Estimation of A Lifetime Distribution When Censoring Times are Missing

X. Joan Hu[1], Jerald F. Lawless[1] and Kazuyuki Suzuki[2]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

[2]Department of Communications and System Engineering, the University of Electro-Communications, Chofu-City, Tokyo 182, Japan

SUMMARY

We consider data sets for which lifetimes associated with the units in a population are observed if they occur within certain time intervals, but for which lengths of the time intervals, or censoring times of unfailed units, are missing. We consider nonparametric estimation of the lifetime distribution for the population from such data; a maximum likelihood estimator and a simple moment estimator are obtained. An example involving automobile warranty data is discussed at some length.

Keywords: Incomplete data; Moment estimates; Warranty reports.

1 Introduction

There are several contexts in the analysis of failure time or lifetime data where censoring times for unfailed units are missing. The area that motivated the current research concerns the estimation of failure time distributions or rates from product warranty data. If under warranty a product may experience a certain type of event, or "failure", then we can estimate the distribution of time to failure (or the intensity function for recurrent events) over the warranty period from warranty reports. However, we typically have to deal with missing censoring times, as we now discuss.

Suppose that T_i is the time to failure for product unit i in a population of M manufactured units. In some applications T_i is measured in real time since the time of sale of the unit. For many types of product the manufacturer does not know the date of sale for most units, and therefore the censoring time (*i.e.*, the elapsed time between the sale of the item and when the data are assembled) for most unfailed items is unknown. For units that fail under warranty the failure time and the potential censoring time are known, because the date of sale is verified as part of the warranty claims process.

Similar problems arise when T_i is some type of usage, or operational time. A familiar example is in connection with automobiles, where T_i represents the mileage at failure. In that case the censoring time is the minimum of the vehicle's current mileage and the mileage which it passes out of the warranty plan. The exact censoring times are in general unknown for all vehicles, but they may be estimated for cars experiencing a failure, since the date of sale, date of failure, and mileage at failure are all known. An application involving automobile warranty data is discussed at some length in Section 7.

Suppose that the lifetime variable T has distribution function $F(t) = \mathrm{P}(T \leq t)$ and that the population of M units has independent lifetimes $t_1, \ldots, t_M$ generated from that distribution. There are also censoring times $\tau_1, \ldots, \tau_M$ associated with the units, where the τ_i's are independent of each other, with common distribution function $G(\tau) = \mathrm{P}(\mathcal{T}_i \leq \tau)$. The observed data are as follows: if $t_i \leq \tau_i$ we observe t_i (and possibly also τ_i), but if $t_i > \tau_i$ we know only that fact and not the value of τ_i or t_i. Our objective is to estimate the distribution $F(t)$ from such data, avoiding any parametric assumptions.

Suzuki (1985), Kalbfleisch and Lawless (1988), and Hu and Lawless (1996a) discuss the use of supplementary followup samples of unfailed units as a way to compensate for the missing censoring times. In many circumstances it is possible to estimate the censoring time distribution, however, and this provides another approach. We present in this paper two nonparametric estimation methods for the case in which the censoring time distribution $G(\tau)$

is known, or at least estimated from other sources. The main assumptions that we make initially are (i) the number of product units M in service is known, (ii) all failures are reported under the data collecting scheme, (iii) the censoring time distribution is known, and (iv) censoring times are statistically independent of failure times. The assumptions are discussed further in the paper, and ways to handle departures from them are presented.

Section 2 of the paper presents nonparametric maximum likelihood and moment estimators for $F(t)$, assuming that $G(\tau)$ is known. Section 3 gives reports on a small simulation study comparing the two estimators. Section 4 consider cases where $G(\tau)$ is estimated, and Section 5 examines the assumption of independence of failure and censoring times. Section 6 presents a detailed example involving automobiles. Section 7 outlines extension of the methodology to deal with multiple failure modes and recurrent events, and Section 8 presents some concluding remarks.

2 Maximum Likelihood and Simple Moment Estimators

A nonparametric method of lifetime distribution estimation was previously given by Suzuki (1988). However, Suzuki and Kasashima (1993) showed that method was inferior to maximum likelihood so we will not discuss it here.

To develop nonparametric estimators it is convenient and customary to work with discrete distributions; finite sample estimates of continuous distribution functions $F(t)$ are discrete anyway, and may be obtained from the discrete-time framework. Thus we assume that lifetime T and censoring time $\mathcal{T}$ may each take on values $1, 2, \ldots$, and $f(t) = \mathrm{P}(T_i = t)$, $g(\tau) = \mathrm{P}(\mathcal{T}_i = \tau)$. The corresponding distribution functions are $F(t) = f(1) + \ldots + f(t)$ and $G(\tau) = g(1) + \ldots + g(\tau)$. In this section, we assume that T_i and $\mathcal{T}_i$ are independent, and that $G(\tau)$ is known.

2.1 Maximum Likelihood Estimation

With known population size M and censoring time distribution $G(\tau)$, the likelihood function based on the probability of the observed data for the population is of the familiar censored data form (Lawless 1982, Chapter 1),

$$\prod_{t_i \leq \tau_i} f(t_i) \prod_{t_i > \tau_i} \mathrm{P}(T_i > \tau_i). \tag{1}$$

The difference with the usual situation is that τ_i is not observed for the unfailed units. Thus, treating it as a random variable with distribution $G(\cdot)$,

$$\begin{aligned} \mathrm{P}(T_i > \tau_i) &= \sum_{\tau=1}^{\tau_{max}} [1 - \sum_{t \leq \tau} f(t)] g(\tau) \\ &= 1 - \sum_{t=1}^{\tau_{max}} f(t)\bar{G}(t), \end{aligned}$$

where $\bar{G}(\tau) = \mathrm{P}(\mathcal{T}_i \geq \tau)$ and $\tau_{max} = \sup\{\tau : \bar{G}(\tau) > 0\}$. We assume that $\tau_{max} < \infty$, which is rather unrestrictive in practice.

The likelihood function (1) may be written as

$$\prod_{t=1}^{\tau_{max}} f(t)^{n_t} [1 - \sum_{t=1}^{\tau_{max}} f(t)\bar{G}(t)]^{M-m} \tag{2}$$

with $n_t = \#\{t_i : \ t_i \leq \tau_i, t_i = t\}$ and $m = \#\{i : t_i \leq \tau_i\} = \sum_t n_t$, where $\#A$ represents the number of elements in set A. Estimates of $f(t), t = 1, \ldots, \tau_{max}$, can be obtained by maximizing (2) under the constraints $f(t) \geq 0$ and $f(1) + \ldots + f(\tau_{max}) = F(\tau_{max}) \leq 1$. To do this we set $c = 1 - F(\tau_{max})$ and note that $\sum_{t=1}^{\tau_{max}} f(t) + c = 1$, where $c \geq 0$. To maximize (2) with respect to $f(1), \ldots, f(\tau_{max})$ and c under the constraints we consider

$$l(\underline{f}, c, \lambda) = \sum_{t=1}^{\tau_{max}} n_t \log f(t) + (M-m) \log[1 - \sum_{t=1}^{\tau_{max}} f(t)\bar{G}(t)] + \lambda[\sum_{t=1}^{\tau_{max}} f(t) + c - 1],$$

where λ is a Lagrange multiplier. The equations

$$\frac{\partial l}{\partial f(t)} = \frac{n_t}{f(t)} - (M-m)\frac{\bar{G}(t)}{1 - \sum_{s=1}^{\tau_{max}} f(s)\bar{G}(s)} + \lambda = 0, \quad t = 1, \ldots, \tau_{max},$$

$$\frac{\partial l}{\partial c} = \lambda = 0, \ \frac{\partial l}{\partial \lambda} = \sum_{s=1}^{\tau_{max}} f(s) + c - 1 = 0,$$

give the *m.l.e.*'s of the $f(t)$'s as

$$\hat{f}_{ML}(t) = \frac{n_t}{M\bar{G}(t)}, \ t = 1, \ldots, \tau_{max}, \tag{3}$$

provided that $\hat{f}_{ML}(1) + \ldots + \hat{f}_{ML}(\tau_{max}) \leq 1$. This is virtually always the case when $F(\tau_{max})$ is not too close to one, which is satisfied in most applications. In the warranty reports context, for example, τ_{max} is the maximum failure time observable and not larger than the warranty time limit, and the probability that a unit fails while under warranty is considerably less than one. If the constraint is not met, then doubt may be cast on the validity of the assumed function $G(\tau)$. When the estimates (3) sum to slightly over 1, a reasonable approach is to simply rescale them so they sum to 1.

The nonparametric maximum likelihood estimate of $F(t), t = 1, \ldots, \tau_{max}$, is then $\hat{F}_{ML}(t) = \hat{f}_{ML}(1) + \ldots + \hat{f}_{ML}(t)$. Similar to the discussion for the nonparametric maximum likelihood estimator in Hu and Lawless (1996a), for example, arguments can be given to establish the consistency and asymptotic normality of this estimator. This also follows from the next section.

2.2 A Simple Moment Estimator

We see that the number of observed units with failures at t in (2) is $n_t = \sum_{i=1}^{M} \mathrm{I}(t_i = t, \tau_i \geq t)$, where $\mathrm{I}(A)$ is the indicator of event A (*i.e.*, it equals 1 if A is true and 0 if not), and its expectation is $M\bar{G}(t)f(t)$, $t = 1, 2, \ldots$ A simple moment estimator of $f(t)$,

$$\hat{f}_{SM}(t) = \frac{n_t}{M\bar{G}(t)}, \ t = 1, \ldots, \tau_{max}, \tag{4}$$

is obtained by noting that $\mathrm{E}\{n_t\} = M\bar{G}(t)f(t)$, $t = 1, \ldots, \tau_{max}$. Notice that $\hat{f}_{SM}(t)$ is the same as the nonparametric *m.l.e.* $\hat{f}_{ML}(t)$ (3) in the current

situation. It is easy to see that $\hat{f}_{SM}(t)$ is unbiased with variance

$$\mathrm{Var}\{\hat{f}_{SM}(t)\} = \frac{f(t)}{M\bar{G}(t)}[1 - f(t)\bar{G}(t)],\ t = 1, \ldots, \tau_{max}. \tag{5}$$

The sample variance estimator

$$\begin{aligned} \widehat{\mathrm{Var}}\{\hat{f}_{SM}(t)\} &= \frac{1}{M^2}\sum_{i=1}^{M}\{u_i(t) - \bar{u}(t)\}^2 \\ &= \frac{n_t(M - n_t)}{M^3\bar{G}(t)^2}, \end{aligned}$$

with $u_i(t) = \mathrm{I}(t_i = t, \tau_i \geq t)/\bar{G}(t)$ and $\bar{u}(t) = \sum_{i=1}^{M} u_i(t)/M$, is the same as the consistent estimate for the variance (5) achieved by replacing $f(t)$ with its estimate $\hat{f}_{SM}(t)$.

Then the estimator for $F(t)$, $t = 1, \ldots, \tau_{max}$, based on $\hat{f}_{SM}(\cdot)$ is

$$\hat{F}_{SM}(t) = \sum_{s=1}^{t} \hat{f}_{SM}(s),\ t = 1, \ldots, \tau_{max}. \tag{6}$$

By noting that

$$\mathrm{Cov}\{\hat{f}_{SM}(s_1), \hat{f}_{SM}(s_2)\} = \frac{f(s_1)}{M\bar{G}(s_1)}\left[\mathrm{I}(s_1 = s_2) - f(s_2)\bar{G}(s_1)\right],$$

we have a consistent estimate for the variance of $\hat{F}_{SM}(t)$ as

$$\begin{aligned} \widehat{\mathrm{Var}}\{\hat{F}_{SM}(t)\} &= \sum_{s=1}^{t}\frac{n_s(M - n_s)}{M^3\bar{G}(s)^2} + \sum_{s_1 \neq s_2}\frac{-n_{s_1}n_{s_2}}{M^3\bar{G}(s_1)\bar{G}(s_2)} \\ &= \frac{1}{M^2}\sum_{i=1}^{M}\left\{\sum_{s=1}^{t} u_i(s) - \bar{u}(s)\right\}^2. \end{aligned} \tag{7}$$

2.3 Extension

We generalize the situation above slightly to allow the distribution of $\mathcal{T}_i$ to depend on a discrete covariate or group indicator x_i. Suppose that x_i takes on

values $x_1^0, \ldots, x_K^0$ and is observable. This is useful because with automobiles, for example, the censoring time for a car may depend upon what time it entered service and that time is usually provided by the dealer. We then denote $g_k(\tau) = \mathrm{P}(\mathcal{T}_i = \tau | x_i = x_k^0)$ and $G_k(\tau) = \mathrm{P}(\mathcal{T}_i \leq \tau | x_i = x_k^0)$. Let $M_k = \#\mathcal{P}_k$, with $\mathcal{P}_k = \{i : \ x_i = x_k^0, i = 1, \ldots, M\}$, $k = 1, \ldots, K$. We assume that the distribution of T_i does not depend on x_i now.

With known subpopulation sizes M_k and censoring time distributions $G_k(\tau)$, $k = 1, \ldots, K$, the likelihood function of the failure time distribution based on the data available is

$$\begin{aligned} & \prod_{t_i \leq \tau_i} f(t_i) \prod_{t_i > \tau_i} \mathrm{P}(T > \mathcal{T}_i | x_i) \\ = & \prod_{t=1}^{\tau_{max}} f(t)^{n_t} \prod_{k=1}^{K} [1 - \sum_{s=1}^{\tau_{max}} f(s) \bar{G}_k(s)]^{M_k - m_k}, \end{aligned} \tag{8}$$

where $\bar{G}_k(t) = \mathrm{P}(\mathcal{T}_i \geq \tau | x_i = x_k^0)$, $\tau_{max} = \max_k \sup\{\tau : \bar{G}_k(\tau) > 0\}$, and $m_k = \#\{x_i : \ x_i = x_k^0, t_i \leq \tau_i\}$. A nonparametric *m.l.e.* of $f(t)$ can be obtained similarly to the procedure in Section 2.1. In this case the maximum likelihood equations reduce to

$$\hat{f}_{ML}(t) = n_t \left\{ \sum_{k=1}^{K} \frac{(M_k - m_k) \bar{G}_k(t)}{1 - \sum_{s=1}^{\tau_{max}} \hat{f}_{ML}(s) \bar{G}_k(s)} \right\}^{-1}, \ t = 1, \ldots, \tau_{max}, \tag{9}$$

where

$$1 - \sum_{s=1}^{\tau_{max}} \hat{f}_{ML}(s) \bar{G}_k(s)$$

is an estimate of $\mathrm{P}(T > \tau_i | x_i = x_k^0)$. There is no closed form for $\hat{f}_{ML}(t)$ when $K \geq 2$. Equations (9) can be used to provide an iteration scheme for obtaining the $\hat{f}_{ML}(t)$'s.

We can, however, obtain a simple closed form estimator. The fact that

$$\mathrm{E}\{n_t | x_i, i = 1, \ldots, M\} = \sum_{k=1}^{K} M_k \bar{G}_k(t) f(t), \ t = 1, \ldots, \tau_{max}$$

gives an unbiased estimator of the $f(t)$:

$$\hat{f}_{ESM}(t) = \frac{n_t}{\sum_{k=1}^{K} M_k \bar{G}_k(t)}, \; t = 1, \ldots, \tau_{max}. \tag{10}$$

We have

$$\begin{aligned} & \mathrm{Cov}\{\hat{f}_{ESM}(s_1), \hat{f}_{ESM}(s_2)\} \\ = \; & \frac{f(s_1)}{\sum_{k=1}^{K} M_k \bar{G}_k(s_1)} \left[\mathrm{I}(s_1 = s_2) - f(s_2) \frac{\sum_{k=1}^{K} M_k \bar{G}_k(s_1) \bar{G}_k(s_2)}{\sum_{k=1}^{K} M_k \bar{G}_k(s_2)} \right]. \end{aligned} \tag{11}$$

The variance of $\hat{F}_{ESM}(t) = \sum_{s=1}^{t} \hat{f}_{ESM}(s)$ is

$$\mathrm{Var}\{\hat{F}_{ESM}(t)\} = \sum_{s_1=1}^{t} \sum_{s_2=1}^{t} \mathrm{Cov}\{\hat{f}_{ESM}(s_1), \hat{f}_{ESM}(s_2)\},$$

and can be consistently estimated by

$$\begin{aligned} & \widehat{\mathrm{Var}}\{\hat{F}_{ESM}(t)\} \\ = \; & \sum_{s_1=1}^{t} \sum_{s_2=1}^{t} \frac{n_{s_1}}{\{\sum_{k=1}^{K} M_k \bar{G}_k(s_1)\}^2} \left[\mathrm{I}(s_1 = s_2) - n_{s_2} \frac{\sum_{k=1}^{K} M_k \bar{G}_k(s_1) \bar{G}_k(s_2)}{\{\sum_{k=1}^{K} M_k \bar{G}_k(s_2)\}^2} \right], \end{aligned} \tag{12}$$

obtained through replacing the $f(\cdot)$ in (11) with the estimates $\hat{f}_{ESM}(\cdot)$. It is easily seen that $\sqrt{M}\{\hat{F}_{ESM}(t) - F(t)\}$ has a limiting normal distribution as $M \to \infty$ and that in order to construct tests or confidence intervals it can be treated as normal with mean 0 and variance estimated by (12).

Estimation of the asymptotic variance of $\hat{f}_{ML}(t)$ may be obtained through the standard procedure for a maximum likelihood estimator. The information matrix is now

$$\mathrm{INFO}(\underline{f}) = -\frac{1}{M} \left(\frac{\partial^2 l}{\partial f(s_1) \partial f(s_2)} \right)_{\tau_{max} \times \tau_{max}}, \tag{13}$$

where the elements are

$$\frac{\partial^2 l}{\partial f(s_1) \partial f(s_2)} = -\mathrm{I}(s_1 = s_2) \frac{n_{s_1}}{f(s_1)^2} - \sum_{k=1}^{K} (M_k - m_k) \frac{\bar{G}_k(s_1) \bar{G}_k(s_2)}{\{1 - \sum_{s=1}^{\tau_{max}} f(s) \bar{G}_k(s)\}^2},$$

$s_1, s_2 = 1, \ldots, \tau_{max}$. Thus, an estimate of the asymptotic variance of $\hat{F}_{ML}(t) = \sum_{s=1}^{t} \hat{f}_{ML}(s)$ may be obtained as well. However, to implement the procedure is very complex when τ_{max} is large because calculation of large matrices has to be involved. In addition, if censoring is heavy, INFO($\underline{f}$) could often be singular and the method can not be applied. The discussion in Hu and Lawless (1996a) for estimation of asymptotic variance of nonparametric maximum likelihood estimators applies here.

3 Simulation

Comparison of the nonparametric maximum likelihood and moment estimators presented in Section 2 was investigated through a simulation study. Being motivated by warranty data, we chose the following simulation setup. Consider a product with an one-year warranty; suppose that there are $M = 4000$ units sold within a year, and the warranty data have been collected over one and a half years since the first unit was sold (we take its sale time as zero). Suppose further that the times of these units to their first failures are independent from each other, identically Weibull distributed, and independent of their sale times, while the sale times are uniformly distributed over the one year period. The censoring time associated with unit i is now $\tau_i = \min(1, 1.5 - x_i)$ year, where x_i is its sale time and $\tau_{max} = 1$.

We generated sale times x_i, $i = 1, \ldots, M$, from the uniform distribution over $(0, 1]$, and failure times t_i from the Weibull distribution

$$f(t) = \frac{\delta}{\alpha}(\frac{t}{\alpha})^{\delta-1} \exp\{-(\frac{t}{\alpha})^{\delta}\}$$

with $\delta = 2.0$ and $\alpha = 3.95, 1.85$. We chose the values of the parameters to make the simulation realistic. The values $m/M = 0.05, 0.20$, respectively, and allow us to study situations with heavy and moderately heavy censoring. From the simulated data, we evaluated the three nonparametric estimates of $F(t) = 1 - \exp\{-(t/\alpha)^{\delta}\}$, $t \in (0, 1]$: (i) $\hat{F}_{SM}(t)$, based on (4); (ii)$\hat{F}_{ESM}(t)$,

	F(0.5)		F(0.8)	
methods	$\alpha = 3.95$	$\alpha = 1.85$	$\alpha = 3.95$	$\alpha = 1.85$
true value	.015895	.070442	.040189	.170550
$\hat{F}_{SM}$	.015844 (.002017)	.070463 (.004016)	.040227 (.003271)	.171054 (.006321)
$\hat{F}_{ESM}$	.015844 (.002017)	.070463 (.004016)	.040225 (.003270)	.171045 (.006296)
$\hat{F}_{ML}$	.015843 (.002015)	.070460 (.004016)	.040223 (.003263)	.171006 (.006294)

Table 1: Sample Means and Standard Errors of the Estimates for F(0.5) and F(0.8)

based on (10) and assuming the numbers of units sold in each quarter of the year (M_1, M_2, M_3, M_4) are known; (iii)$\hat{F}_{ML}$, the *m.l.e.* based on (9) for the same stratified-sample situation as (ii). The estimates were obtained through discretizing the time interval $(0, 1]$. That is, we divided the time period into 120 small intervals, $((k-1)/120, k/120]$, $k = 1, \ldots, 120$, and assigned variables having values in the kth interval to the value $k/120$.

We used Splus for generating the random variables needed and all the computing. The maximum likelihood estimate was evaluated by using the iteration procedure based on (9); we took $\tilde{f}^{(0)}(k/120) = f(k/120)$ and terminated the iterations when $\sum_{k=1}^{120} |\tilde{f}^{(j+1)}(k/120) - \tilde{f}^{(j)}(k/120)| < 0.0001$, where $\tilde{f}^{(j)}(t)$ is the jth iterate toward $\hat{f}_{ML}(t)$.

Table 1 presents the sample means and standard errors (in brackets) of estimates for $F(0.5)$ and $F(0.8)$ based on the three estimators from 1000 simulation repetitions, for each case $\alpha = 3.95, 1.85$. There is essentially no difference in the estimators. We also show estimates of $F(t)$ from a single simulated sample in Figure 1, and estimates of $F(0.5)$ and $F(0.8)$ and their estimated standard deviations (in brackets) in Table 2. The estimated standard deviations were based on (7), (12) and (13) respectively. Comparing the two tables, we see (7) and (12) estimate the variances of $\hat{F}_{SM}(t)$ and $\hat{F}_{ESM}(t)$ well.

	F(0.5)		F(0.8)	
methods	$\alpha = 3.95$	$\alpha = 1.85$	$\alpha = 3.95$	$\alpha = 1.85$
true value	.015895	.070442	.040189	.170550
$\hat{F}_{SM}$	.016750 (.001999)	.065000 (.003987)	.042109 (.003238)	.172678 (.006424)
$\hat{F}_{ESM}$	.016750 (.002000)	.065000 (.003987)	.042019 (.003261)	.172319 (.006428)
$\hat{F}_{ML}$	.016752 (.001984)	.064967 (.003973)	.042052 (.003343)	.172151 (.006431)

Table 2: Estimates for F(0.5) and F(0.8) and Their Estimated Standard Deviations

(Figure 1 is inserted here)

Figure 1: Estimates of the Lifetime Distribution Function
(a) $\alpha = 3.95$; (b) $\alpha = 1.85$

This simulation suggests that there is almost no difference in the three estimators, and they all estimate $F(t)$ well in the situations we consider.

4 Effect of Estimating The Censoring Time Distribution

The estimation procedures in Section 2 assume the censoring time distribution $G(\tau)$ is known. In most practical situations, however, $G(\tau)$ is estimated, or only roughly known. Hu and Lawless (1996b) investigate likelihood based parametric estimation for this situation; their approach can be extended to a nonparametric setting. We here focus on the extension of the simple moment estimator in Section 2.2; the estimator in Section 2.3 can be extended similarly. An example is presented in Section 6.

Suppose that $\tilde{G}(\tau)$ is consistent. In that case the estimates analogous to

(4),

$$\tilde{f}_{SM}(t) = \frac{n_t}{M\tilde{\bar{G}}(t)},\ t = 1, \ldots, \tau_{max}, \tag{14}$$

and to $\hat{F}_{SM}(t)$ in (6),

$$\tilde{F}_{SM}(t) = \sum_{s=1}^{t} \tilde{f}_{SM}(s),\ t = 1, \ldots, \tau_{max}, \tag{15}$$

are both consistent.

Behavior of the estimator $\tilde{f}_{SM}(t)$ depends on how well $\tilde{G}(\tau)$ estimates $G(\tau)$, and whether $\tilde{G}(\tau)$ is related to the primary data, *i.e.*, the n_t's. In this paper, we assume $\tilde{G}(\tau)$ is independent of the primary data. The covariance of the $\tilde{f}_{SM}(t)$'s is then

$$\begin{aligned}
& \mathrm{Cov}\{\tilde{f}_{SM}(s_1), \tilde{f}_{SM}(s_2)\} \\
= \ & \mathrm{E}\left\{\mathrm{Cov}[\tilde{f}_{SM}(s_1), \tilde{f}_{SM}(s_2)|\tilde{G}(\tau)]\right\} + \mathrm{Cov}\left\{\mathrm{E}[\tilde{f}_{SM}(s_1)|\tilde{G}(\tau)], \mathrm{E}[\tilde{f}_{SM}(s_2)|\tilde{G}(\tau)]\right\} \\
= \ & \mathrm{E}\left\{\frac{f(s_1)\bar{G}(s_1)[\mathrm{I}(s_1 = s_2) - f(s_2)\bar{G}(s_2)]}{M\tilde{\bar{G}}(s_1)\tilde{\bar{G}}(s_2)}\right\} + \mathrm{Cov}\left\{\frac{\bar{G}(s_1)f(s_1)}{\tilde{\bar{G}}(s_1)}, \frac{\bar{G}(s_2)f(s_2)}{\tilde{\bar{G}}(s_2)}\right\},
\end{aligned} \tag{16}$$

which can be estimated by

$$\frac{\tilde{f}_{SM}(s_1)}{M\tilde{\bar{G}}(s_2)}[\mathrm{I}(s_1 = s_2) - \tilde{f}_{SM}(s_2)\tilde{\bar{G}}(s_2)] + \frac{\tilde{f}_{SM}(s_1)\tilde{f}_{SM}(s_2)}{\tilde{\bar{G}}(s_1)\tilde{\bar{G}}(s_2)}\widehat{\mathrm{Cov}}\left\{\tilde{\bar{G}}(s_1), \tilde{\bar{G}}(s_2)\right\}, \tag{17}$$

assuming an estimate $\widehat{\mathrm{Cov}}\left\{\tilde{\bar{G}}(s_1), \tilde{\bar{G}}(s_2)\right\}$ is available. The second term in (17) accounts for variation due to $G(\tau)$ having been estimated. Then, an estimate for the variance of $\tilde{F}_{SM}(t)$ can be obtained from

$$\mathrm{Var}\{\tilde{F}_{SM}(t)\} = \sum_{s_1=1}^{t}\sum_{s_2=1}^{t} \mathrm{Cov}\{\tilde{f}_{SM}(s_1), \tilde{f}_{SM}(s_2)\}.$$

In Section 6, we will discuss this further based on the example there. We remark that the discussion above can be extended to the situation in Section 2.3 with a little modification.

5 Non-independent Censoring

Section 2 assumes that censoring times $\mathcal{T}_1, \ldots, \mathcal{T}_M$ are independent of lifetimes $T_1, \ldots, T_M$. This assumption may sometimes be questionable: for example, if the lifetime of an automobile component depends on both the age of the car and the number of miles it is driven, then the fact that warranty plans have age and mileage limitations (*e.g.*, two years and 24000 miles) implies a dependence between T_i and $\mathcal{T}_i$. Our objective here is to briefly consider the effect of non-independent censoring on the estimator of Section 2.2. We also present a version of the simple moment estimator for a special case in this situation.

5.1 Effect on $\hat{f}_{SM}$

The estimator $\hat{f}_{SM}$ of (4) can be written in the form

$$\hat{f}_{SM}(t) = \frac{\sum_{i=1}^{M} \mathrm{I}(t_i = t, \tau_i \geq t)}{\sum_{i=1}^{M} \mathrm{P}(\mathcal{T}_i \geq t)}.$$

Now $\mathrm{E}\{\mathrm{I}(t_i = t, \tau_i \geq t)\} = f(t)\mathrm{P}(\mathcal{T}_i \geq t | T_i = t)$, so if $\mathrm{P}(\mathcal{T}_i \geq t | T_i = t) \neq \mathrm{P}(\mathcal{T}_i \geq t)$, then $\hat{f}_{SM}(t)$ is biased, with

$$\mathrm{E}\{\hat{f}_{SM}(t)\} = f(t)\left\{\frac{\sum_{i=1}^{M} \mathrm{P}(\mathcal{T}_i \geq t | T_i = t)}{\sum_{i=1}^{M} \mathrm{P}(\mathcal{T}_i \geq t)}\right\}. \tag{18}$$

The extent of the bias may be assessed by hypothesizing models for the dependence of T_i and $\mathcal{T}_i$, and in many cases we may find that the bracketed term in (18) is close to one. If it is not, there may be little motivation to estimate the marginal distribution $f(t)$; what is needed instead is a model that accounts for the dependence of T and $\mathcal{T}$. With automobiles, this usually means that a lifetime model which incorporates both age and mileage is needed. Lawless *et al.* (1995) discuss such models and indicate how to test independence of lifetime and censoring time from automobile warranty data. If there is a serious concern in a practical situation about dependence, then such methods should be employed.

5.2 A Special Case

In some situations T_i and $\mathcal{T}_i$ are related only through a covariate (or covariates), say x_i, such that T_i and $\mathcal{T}_i$ are independent given x_i, $i = 1, \ldots, M$. This is considered in different contexts by Kalbfleisch and Lawless (1991), and Hu and Lawless (1996b). We extend the model of Section 2.3 slightly to deal with this.

As in Section 2.3, we suppose that x_i takes on values x_k^0, $k = 1, \ldots, K$, and is observed for every unit i. Then

$$\hat{f}_{SM}(t|x_k^0) = \frac{n_{t,k}}{M_k \bar{G}_k(t)}, \; t = 1, \ldots, \tau_{max} \tag{19}$$

is an unbiased estimator of $f(t|x_k^0)$, $k = 1, \ldots, K$, where $n_{t,k} = \#\{t_i : \; t_i \leq \tau_i, t_i = t, x_i = x_k^0\}$. Noting that $f(t) = \sum_{k=1}^{K} f(t|x_k^0)\mathrm{P}(X = x_k^0)$, we have an estimator for $f(t)$, and also for $F(t)$, provided $\mathrm{P}(X = x_k^0)$ is known or estimated, $k = 1, \ldots, K$. The changing pattern of $\hat{F}_{SM}(t|x_k^0) = \sum_{s=1}^{t} \hat{f}_{SM}(s|x_k^0)$ when the value of x_k^0 varies may help us see how lifetime is related to censoring time. If the dependences between T_i and x_i, and $\mathcal{T}_i$ and x_i can be specified parametrically, we can see how the dependence of the failure time and the censoring time affects the simple moment estimator from (18). Parametric models also allow us to handle continuous covariates. Hu and Lawless (1996b) consider this approach.

For a slightly different situation where only the x_i's associated with units having observed failures can be observed, we may consider the estimator for $f(t|x_k^0)$,

$$\tilde{f}(t|x_k^0) = \frac{n_{t,k}}{\tilde{M}_k \bar{G}_k(t)}, \; t = 1, \ldots, \tau_{max}, \tag{20}$$

if an estimate $\tilde{M}_k$ is available. We address this in Section 6 through the example. The idea may be applied to situations where the number of product units in service M is unknown but there is an estimate for it.

Similarly, as in Sections 2 and 3, we can consider variance estimation of $\hat{F}(t|x_k^0)$, $t = 1, \ldots, \tau_{max}$ and $k = 1, \ldots, K$.

6 An Example

Some real warranty data for a specific system on a particular car model are considered for illustration. The data include warranty claims from 823 cars among 8394 cars produced during a two-month period. The warranty plan in question was for one year or 12000 miles; the data collection was over the first 18 months after the first car was sold. We examine the distribution of the time to the first failure (claim) of the cars. For illustration we consider both "time" as mileage in miles and as age in years (*i.e.*, real time) although for engineering purposes mileage is more relevant.

Let t_i and τ_i, $i = 1, \ldots, M = 8394$ be the first failure times and censoring times, respectively, and let s_i denote the time of sale for car i, where the first car sold has a sale time of zero. Real time will be expressed in years, and the warranty data therefore include followup of cars up to time 1.5 years. The censoring time τ_i for car i may be described as follows. Let the age of the car (*i.e.*, time since the car was sold) when it reaches 12000 miles be a_i years, and define $u_i = 12000/a_i$ as the average mileage accumulation rate over the age interval $(0, a_i]$, in miles per year. Then for the case where t_i and τ_i represent age in real time (*i.e.*, years since sale), $\tau_i = \min(1.5 - s_i, 1, 12000/u_i)$. In the case where t_i and τ_i represent mileage, and $\tau_i = \min(\min[1.5 - s_i, 1]u_i, 12000)$.

The values of t_i's are observed only for those cars with $t_i \leq \tau_i$, *i.e.*, for the $m = 823$ cars with their warranty claims recorded. Although the sale dates s_i's are known for all 8394 cars, the values of τ_i's are not. If we are willing to make the simplifying assumption that mileage accumulation is linear over $(0, a_i]$, then u_i may be evaluated for cars that fail, since the mileage as well as the age at failure is recorded. In this case we would thus have the censoring times τ_i's for the cars which fail, but not for those which do not. The simple estimators used here do not require any censoring times.

A customer survey of 607 cars of the same type and approximate geographic location as those in the warranty data base was taken, wherein the approximate mileages at one year were obtained for each car. We assume

(Figure 2 is inserted here)

Figure 2: Approximate 95% Confidence Intervals of Failure-Time Distribution ("time"=mileage)

that mileage accumulation occurs at a constant rate u_i for car i over the first year after sale; this is obviously an oversimplification but is satisfactory for practical purposes in this case. Sale date and mileage accumulation rate can reasonably be assumed independent, and we know

$$\bar{G}(\tau) = \mathrm{I}(\tau \leq 1)\mathrm{P}(1.5 - s_i \geq \tau, U_i \leq \frac{12000}{\tau})$$

in the "time" equals age case and

$$\bar{G}(\tau) = \mathrm{I}(\tau \leq 12000)\mathrm{P}(U_i \min[1.5 - s_i, 1] \geq \tau)$$

in the "time" equals mileage case. Then we can estimate the distribution of censoring time $G(\tau)$ in the warranty data base population by using the empirical distribution of sale dates s_i $(i = 1, \ldots, 8394)$ along with the empirical distribution of U_i based on the customer survey. The moment estimate (6) may thus be computed, and is shown in Figure 2, for the case where failure "times" are measured in miles. Figure 2 also shows approximate pointwise 95% confidence intervals for the failure time distribution function $F(t)$, obtained as $\hat{F}_{SM}(t) \pm 1.96\hat{V}(t)^{\frac{1}{2}}$, where $\hat{V}(t)$ is the estimated variance of $\hat{F}_{SM}(t)$ given by (7). These intervals are based on the fact that as M increases, the distribution of $[\hat{F}_{SM}(t) - F(t)]\hat{V}(t)^{-\frac{1}{2}}$ approaches a standard normal distribution. Two sets of confidence limits are shown: intervals I use the variance estimate (7), which assume that $G(\tau)$ is known; intervals II are based on (17), and account for the fact that $G(\tau)$ has been estimated by using the car survey. The second set of intervals are considerably wider and provide a more valid assessment of uncertainty. We could similarly produce estimates of the failure time distribution in terms of car age.

We remark that an alternative approach is to stratify cars according to their time of sale and then to use the approach in Section 2.3. This produces

an estimate of $F(t)$ that is indistinguishable from that in Figure 2 in this case.

It is possible here that failure may be related to both age (time since sale) and mileage. To investigate this we formed a covariate x based on mileage accumulation rates, as follows. We divided mileage rates into 5 classes: $(0, 6000], (6000, 12000], (12000, 18000], (18000, 24000], (24000, \infty)$ miles per year, and let $x = k$ denote the kth class ($k = 1, \ldots, 5$). The numbers of failures for cars in the five classes are $92, 266, 245, 109, 111$, respectively. From the customer survey of 607 cars and the car sales data, we estimated the censoring time distributions $G_k(\tau) = \mathrm{P}(\mathcal{T}_i \leq \tau | x_i = k)$, $k = 1, \ldots, 5$ through

$$\bar{G}_k(\tau) = \mathrm{P}(\min[1.5 - s_i, 1] \geq \tau)\mathrm{P}(U_i \leq \frac{12000}{\tau} | x_i = k)$$

for the age case and

$$\bar{G}_k(\tau) = \mathrm{I}(\tau \leq 12000) \int \mathrm{P}(\min[1.5 - s_i, 1] \geq \frac{\tau}{u}) d\mathrm{P}(U_i \leq u | x_i = k)$$

for the mileage case; from the survey data alone we estimated $\mathrm{P}(k) = \mathrm{P}(X_i = k)$. The numbers of cars from the survey sample falling into the five groups are $96, 271, 148, 53, 39$, respectively. Finally, we imputed a value of u_i, and thus x_i, for each car that experienced a failure under warranty by dividing the mileage at failure by the age at failure. We then estimate $F_k(t) = \mathrm{P}(T_i \leq t | x_i = k)$ as

$$\tilde{F}_k(t) = \sum_{s=1}^{t} \frac{n_{s,k}}{\tilde{M}_k \bar{\tilde{G}}_k(s)}, \quad t = 1, \ldots, \tau_{max,k}, \tag{21}$$

where $n_{s,k} = \#\{i : t_i = s, x_i = x_k^0, \tau_i \geq s\}$, $\tilde{M}_k = M\tilde{\mathrm{P}}(x_k^0)$ with $M = 8,394$, and $\tau_{max,k} = \sup\{\tau : \bar{\tilde{G}}_k(\tau) > 0\}$, $k = 1, \ldots, 5$. Estimates of $\tilde{F}_k(t)$, $k = 1, \ldots, 5$ are presented in Figures 3(a) and 3(b) for the cases where failure time is measured as car age and car mileage, respectively. Bearing in mind that the estimates are not very precise, in part because the estimates of $\tilde{G}_k(\tau)$ and $\tilde{P}(k)$ are based on rather small samples, Figure 3 suggests that failure times measured in miles do not depend much upon the mileage accumulation

(Figure 3 is inserted here)

Figure 3: Estimates of Failure-Time Distributions with Different Usage Rates (a) "time" = age, (b) "time" = mileage.

rate, but that failure times measured as car age do. This suggests that mileage is the more relevant time scale for this type of failure. Lawless *et al.* (1995) reached a similar conclusion by using parametric models for failure that incorporate both age and mileage as factors.

7 Recurrent Events and Multiple Failure Modes

Products under warranty are usually repairable systems in which there are multiple types of failure which may occur more than once. The problem discussed in this paper can be studied in this broader context, and methods based on maximum likelihood and on moment estimation may be developed. We will merely mention the main ideas, which are discussed elsewhere.

The modeling of recurrent events often uses Poisson or renewal processes (Ascher and Feingold 1984; Lawless 1995). More generally, the mean and rate functions for the recurrent events or failures are of interest. They are defined as follows: let $N_i(t)$ denote the number of events occurring on unit i over the time interval $(0, t]$. Then $\Lambda(t) = \mathrm{E}\{N_i(t)\}$ is called the mean function and $\lambda(t) = d\Lambda(t)/dt$ is called the rate (or rate of occurrence) function. If the recurrent events follow a Poisson process then $\lambda(t)$ is also the intensity function.

In the case of recurrent events the "censoring" time τ_i refers to the time period $(0, \tau_i]$ over which unit i is observed. Hu and Lawless (1996a) discuss maximum likelihood estimation uner a Poisson model when censoring times are missing for units not experiencing any failures. They also present a moment estimator for $\lambda(t)$ that is analogous to the ones given for failure time distributions in Section 2.2 and 2.3, and is of exactly the same form as

(4),

$$\hat{\lambda}_{SM}(t) = \frac{n_t}{M\bar{G}(t)},\ t = 1, \ldots, \tau_{max},$$

where now, however, n_t is the total number of recurrent events observed at time t across all product units. Hu and Lawless (1996a) give variance estimates for $\hat{\lambda}_{SM}(t)$ and $\hat{\Lambda}_{SM}(t) = \hat{\lambda}_{SM}(1) + \ldots + \hat{\lambda}_{SM}(t)$ and discuss their properties.

Multiple failure modes may also be dealt with. For simplicity we consider two modes A and B and the case of failure times; recurrent events can also be considered. Let T_i^A and T_i^B represent the times to failure of modes A and B, respectively, let $f_A(t) = \mathrm{P}(T_i^A = t)$ and $f_B(t) = \mathrm{P}(T_i^B = t)$ denote the marginal probability functions, and let $f_{AB}(s,t) = \mathrm{P}(T_i^A = s, T_i^B = t)$ denote the joint probability function of T_i^A and T_i^B. Under the assumption of independent censoring times τ_i, the following are unbiased estimates of $f^A(t)$ and $f^B(t)$:

$$\hat{f}_A(t) = \frac{n^A(t)}{M\bar{G}(t)},\ \hat{f}_B(t) = \frac{n^B(t)}{M\bar{G}(t)}, \tag{22}$$

where $n^A(t) = \sum_{i=1}^M \mathrm{I}(T_i^A = t, \tau_i \geq t)$ and $n^B(t) = \sum_{i=1}^M \mathrm{I}(T_i^B = t, \tau_i \geq t)$, and once again we assume $\bar{G}(\tau) = \mathrm{P}(\mathcal{T}_i \geq \tau)$ is known. It is also possible to give a simple moment estimator of $f_{AB}(s,t)$:

$$\hat{f}_{AB}(s,t) = \frac{n^{AB}(s,t)}{M\bar{G}(s \vee t)}, \tag{23}$$

where $n^{AB}(s,t) = \sum_{i=1}^M \mathrm{I}(T_i^A = s, T_i^B = t, \tau_i \geq s \vee t)$ and $s \vee t$ denotes the maximum of s and t. However, in applications where the probability of a failure of any given mode is fairly small over the observation period, the probability of getting failures on two or more modes is usually very small, and so (23) may not be very precise. In many situations it may be adequate simply to consider the different failure modes separately, in which case the estimates (22) are all that is needed. Variance estimates are then given by the expressions for $\hat{f}_{SM}(t)$ in Section 4.2. If, however, we wish to gain insight

into how failure times for different modes are related, (23) can be used. If this is too imprecise to be useful then one can adopt a parametric model to get more precise (but model-dependent) estimates.

The preceding discussion of multiple failure modes assumes that when a failure of one type occurs it does not preclude failures of other types. In some situations the failure modes may be competing, so that this does happen. Suzuki *et al.* (1996) discuss estimation for multiple failure mode problems in detail.

8 Comments and Recommendations

When censoring times are missing, standard methods of estimating lifetime distributions are not available. However, if the censoring time distribution $G(\tau)$ is known or estimated from additional data then either maximum likelihood or moment estimation may be used to obtain nonparametric estimates. The methods in this paper depend on the validity of the assumed $G(\tau)$, and it is important in practice to be confident that $G(\tau)$ is suitable. We also recommend the use of confidence limits for the lifetime distribution that account for uncertainty in $G(\tau)$. When $G(\tau)$ is estimated, the method of Section 4 can be employed. If standard errors for the estimate of $G(\tau)$ are not available, we recommend varying $G(\tau)$ in a sensible way around the estimate and examining the range of confidence limits obtained.

The estimates in Section 2 also require that the censoring times be independent of lifetimes. This can be a problem for some applications. As shown in Sections 5 and 6, we can often handle dependent censoring by utilizing a covariate x such that lifetimes and censoring times are roughly independent conditional on x. In the case of automobile warranty data the mileage accumulation (or usage) rate fulfills this function.

If censoring times are available for units that fail then inferences about the lifetime distribution may be obtained by considering the distribution of

t_i given that $t_i \leq \tau_i$ for failed units. This gives the truncated data likelihood function

$$L_T = \prod_{t_i \leq \tau_i} \frac{f(t_i)}{F(\tau_i)} \tag{24}$$

instead of (1). It is well known that for parametric models $f(t;\theta)$ the likelihood (24) gives much less precise estimates of θ than do methods which use information about the censoring times for unfailed units (Kalbfleisch and Lawless 1988; Hu and Lawless 1996b). Thus, the use of (1) to estimate θ with a parametric model would be much preferred to (24). The same holds true for nonparametric estimation of $f(t)$; Kalbfleisch and Lawless (1991) discuss nonparametric estimation based on (24) but the methods of this paper are to be preferred. The price, of course, is that a good estimate of the censoring time distribution $G(\tau)$ must be obtained.

Finally, in contexts such as manufacturing one may sometimes wish to make estimates for a finite population of units. The estimates of $f(t)$ given here can be used to do this. In principle, finite population corrections to variance estimates can be made but given the large size of the typical populations, it makes no practical difference if these are ignored.

ACKNOWLEDGEMENTS

This research was supported in part by grants to the second author from General Motors Canada, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada, and by a fellowship to the first author.

REFERENCES

Ascher, H. and Feingold, H. (1984) *Repairable System Reliability.* New York: Marcel Dekker.

Hu, X.J. (1995). *Estimation from Truncated Data with Supplementary Information, with Application to Field Reliability.* Ph.D. Thesis, University of Waterloo, Canada.

Hu, X.J. and Lawless, J.F. (1996a) "Estimation of Rate and Mean Functions from Truncated Recurrent Event Data," *Journal of the American Statistical Association* **91**, 300-310.

Hu, X.J. and Lawless, J.F. (1996b) "Estimation from Truncated Lifetime Data with Supplementary Information on Covariates and Censoring Times," To appear in *Biometrika.*

Kalbfleisch, J.D. and Lawless, J.F. (1988). "Estimation of Reliability from Field Performance Studies (with discussion)," *Technometrics* **30**, 365-388.

Kalbfleisch, J.D. and Lawless, J.F. (1991). "Regression Models for Right Truncated Data with Applications to AIDs Incubation Times and Reporting Lags," *Statistica Sinica* **1**, 19-32.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* New York: John Wiley and Sons.

Lawless, J.F. (1995). "The analysis of recurrent events for multiple Subjects," *Applied Statistics* **44**, 487-498.

Lawless, J.F., Hu, X.J, and Cao, J. (1995). "Methods for the Estimation of Failure Distributions and Rates from Automobile Warranty Data," *Lifetime Data Analysis* **1** 227-240.

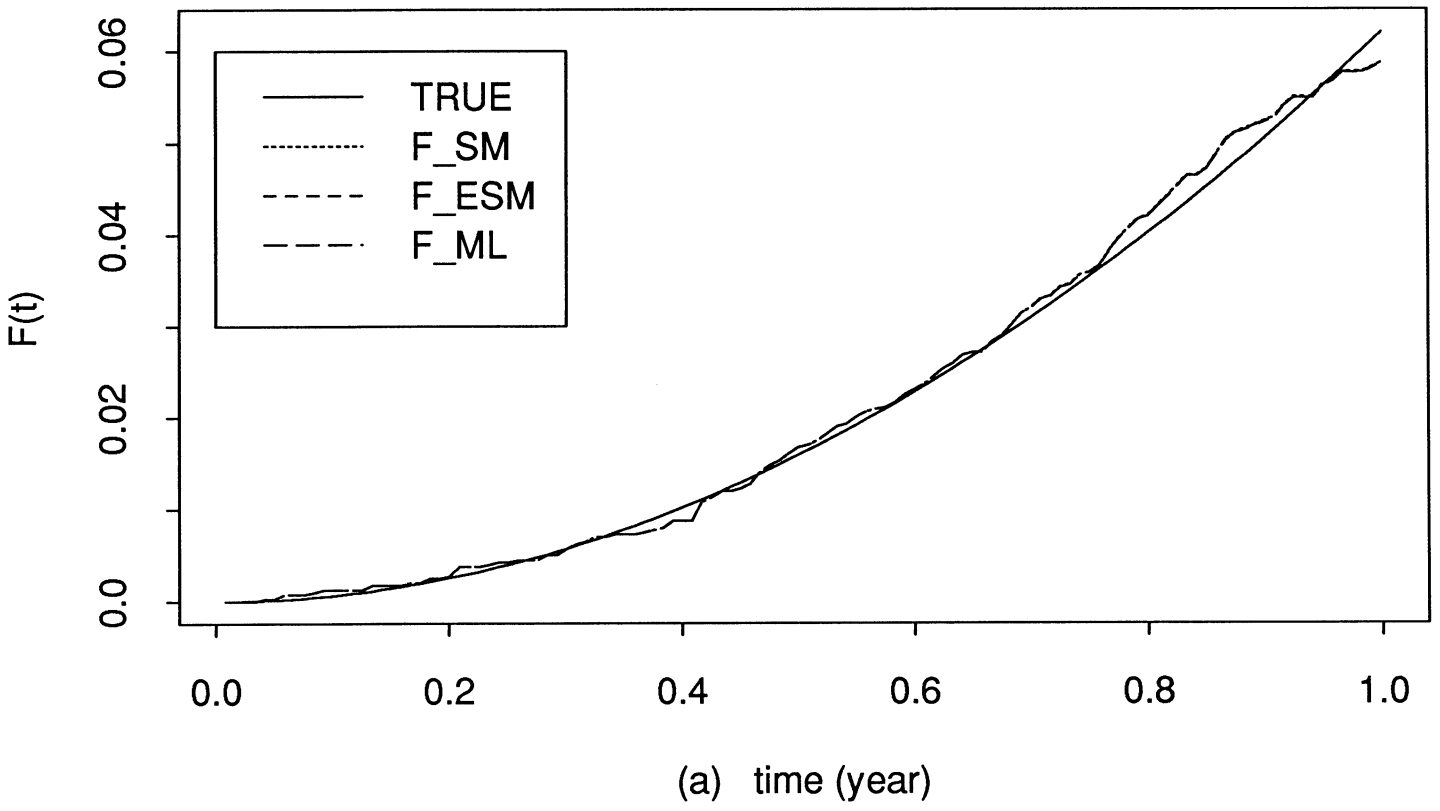
Lawless, J.F. and Kalbfleisch, J.D. (1992). "Some Issues in the Collection and Analysis of Field Reliability Data," *Survival Analysis: State of the Art*, Goel, 141-152. Dordrecht: Kluwer.

Suzuki, K. (1985) "Estimation of Lifetime Parameters from Incomplete Field Data," *Technometrics* **27**, 263-271.

Suzuki, K. and Kasashima, T. (1993). "Estimation of Lifetime Distribution from Incomplete Field Data with Different Observational Periods," *Technical Report, University of Electro-Communications, Department of Commun. & Systems Eng., Japan*, UEC-CAS-93-02.

Suzuki, K., Adachi, K., Hu, X.J. and Lawless, J.F. (1996). "Analysis of Incomplete Field Failure Data with Multiple Failure Modes Using Information on the Censoring Distribution," (in preparation).

FIGURE 1


TRUE
F_SM
F_ESM
F_ML
F(t)
0.0
0.02
0.04
0.06
0.0
0.2
0.4
0.6
0.8
1.0
(a) time (year)

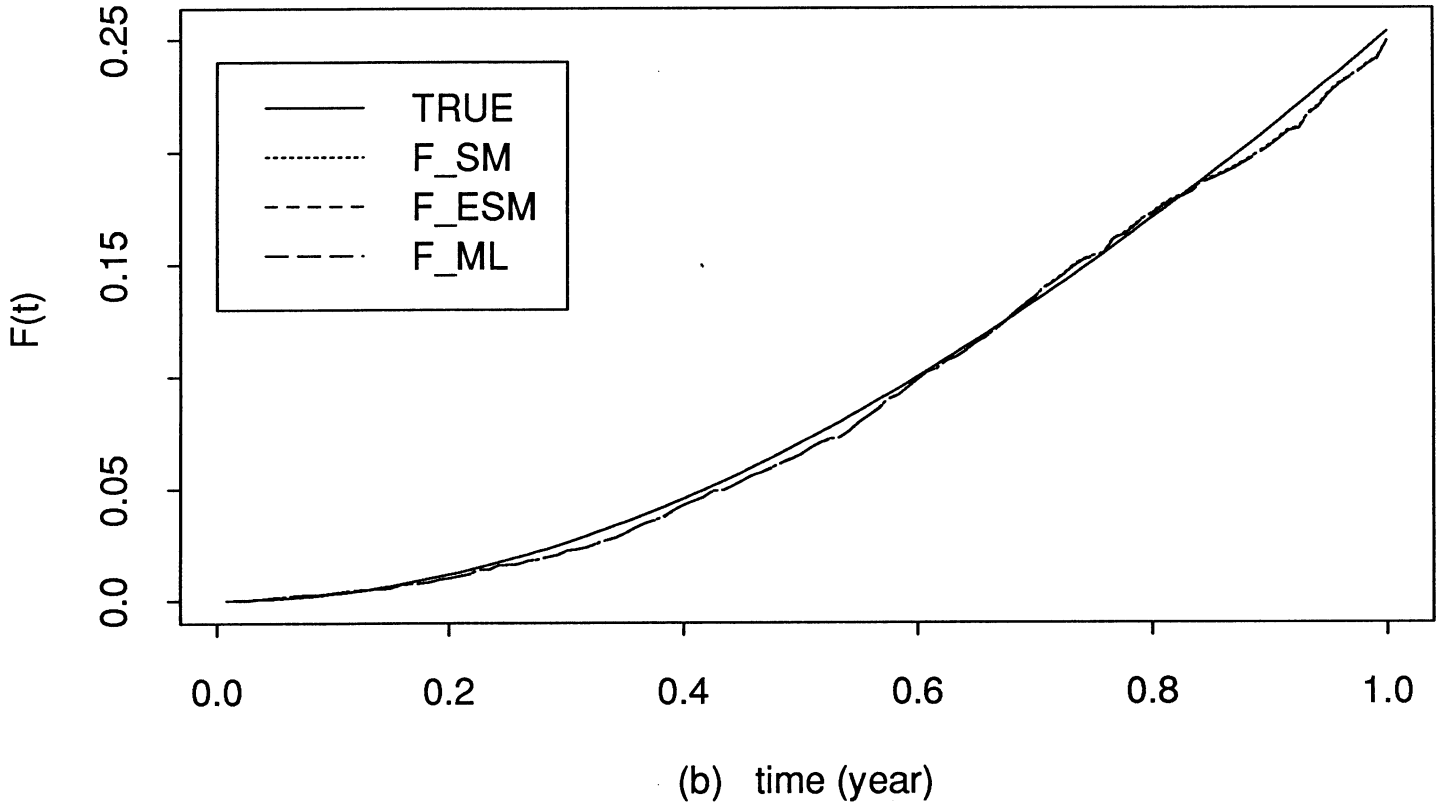

TRUE
F_SM
F_ESM
F_ML
F(t)
0.0
0.05
0.15
0.25
0.0
0.2
0.4
0.6
0.8
1.0
(b) time (year)

FIGURE 2

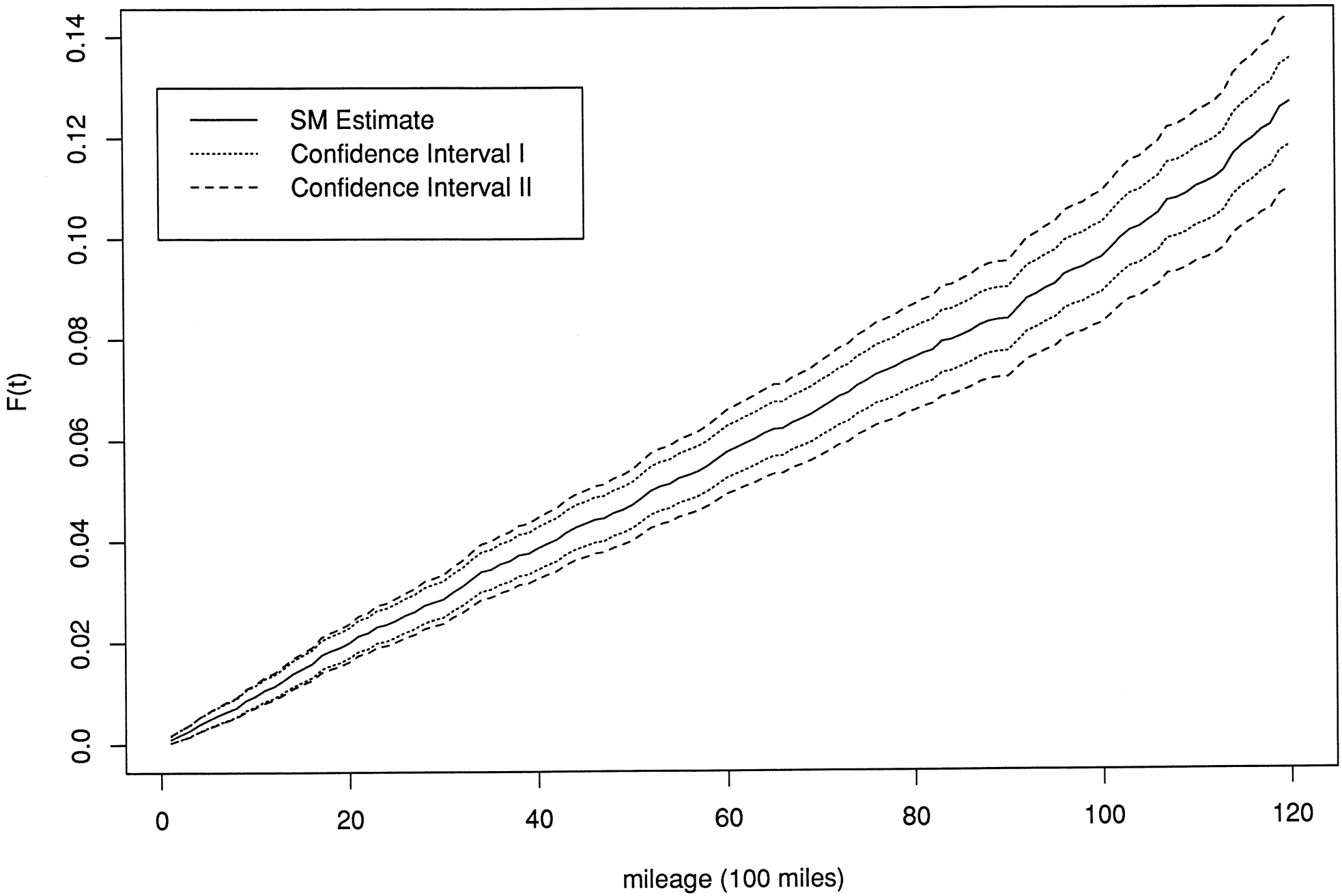
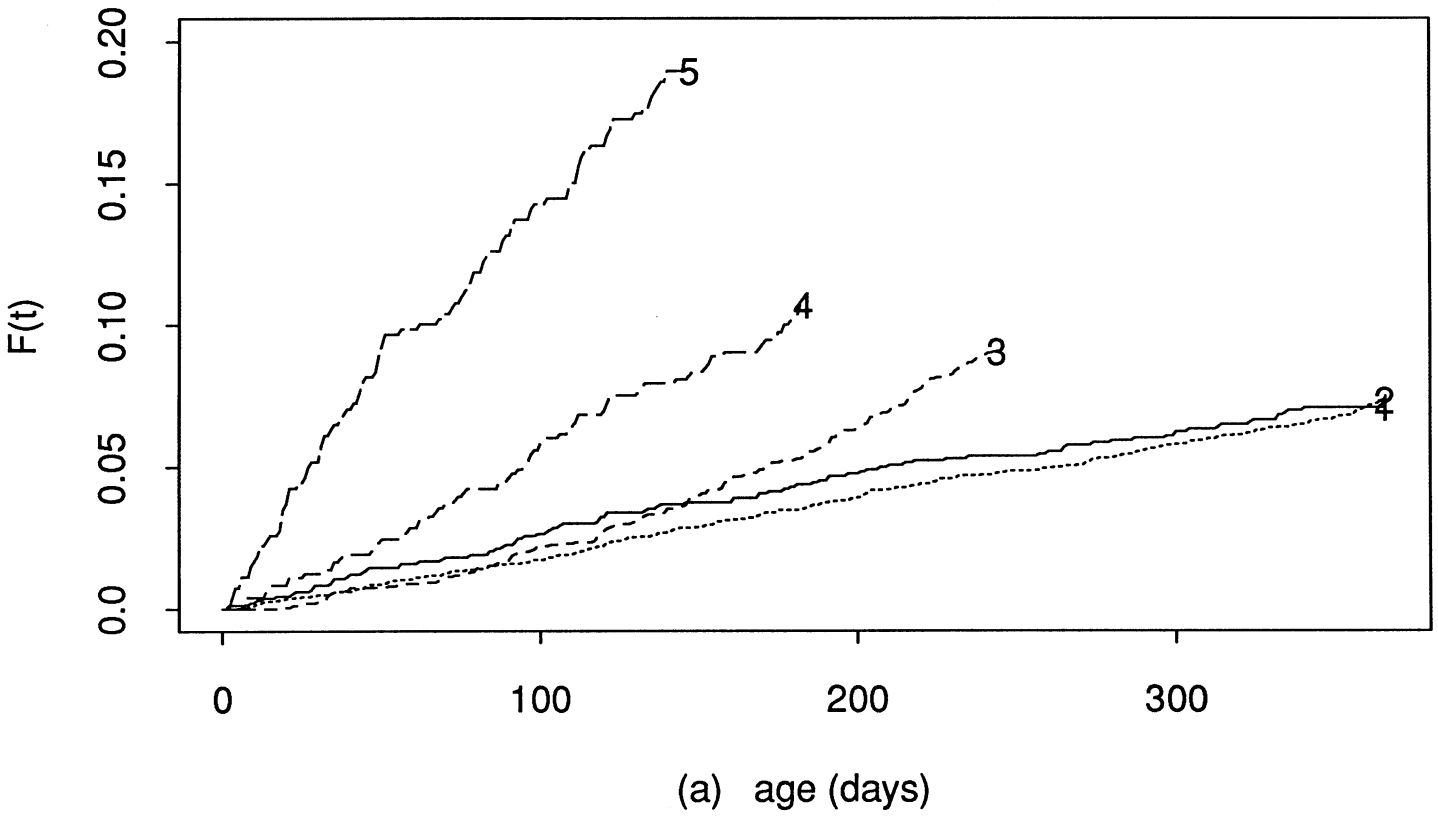
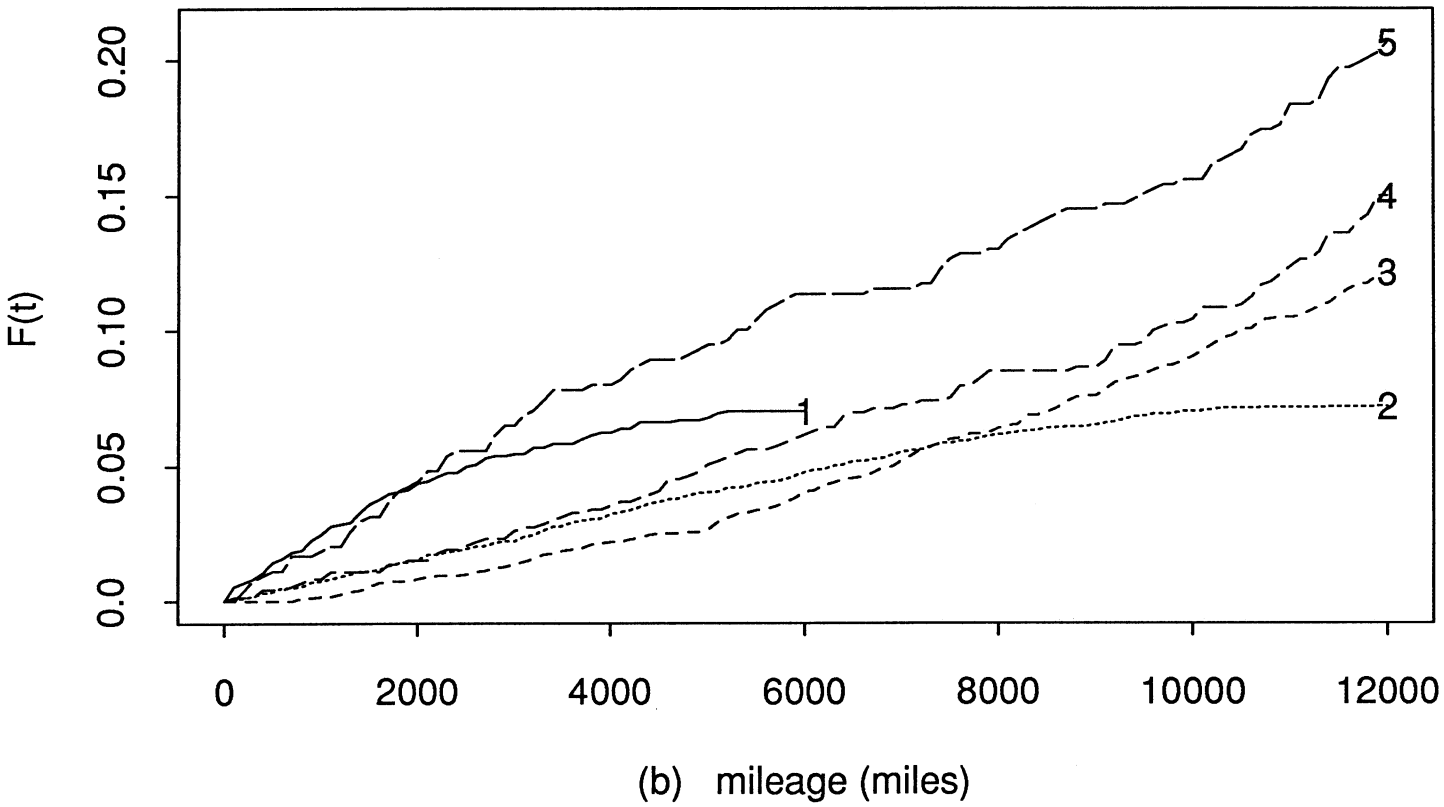

SM Estimate
Confidence Interval I
Confidence Interval II
F(t)
0.0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
20
40
60
80
100
120
mileage (100 miles)

FIGURE 3



(a) age (days)



(b) mileage (miles)