Statistical Analysis of Product Warranty Data

J.F. Lawless
University of Waterloo

RR-97-05
September 1997

STATISTICAL ANALYSIS OF PRODUCT WARRANTY DATA

J.F. Lawless
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

Summary

Data bases regarding warranty claims for manufactured products record claims experience and information about concomitant factors. If constructed and maintained properly, warranty data bases may be used for a variety of purposes that include the prediction of future claims, the comparison of claims experience for different groups of products, the estimation of field reliability, and the identification of opportunities for quality and reliability improvement. This paper reviews some methods of analyzing warranty data and of addressing these objectives. Some extensions of previous work and suggestions for future development are included. Examples involving warranty claims for automobiles and refrigerators are considered.

Keywords: aggregate claims; count data; log linear models; prediction; robust estimation.

1. Introduction

Manufacturers of products which are sold with warranty coverage usually collect fairly comprehensive data on warranty claims and costs. This includes the time and place of manufacture of the product, the date of sale, and the time and type of problem which lead to the claim. Covariates concerning product usage or other factors may also be recorded. If constructed and maintained properly, warranty data bases may be used to predict future claims, to compare claims experience for different groups of products, and to study variations in claims relative to factors such as time and place of manufacture, or usage environment. In some circumstances warranty data may also be used to estimate the field reliability of products and to identify opportunities for the improvement of quality and reliability.

The analysis of warranty data was historically a rather neglected topic, although there were many papers on the mathematical modelling and design of warranties (e.g. Blischke and Murthy 1996). Although some of the problems associated with warranty data are similar to problems involving aggregate claims in insurance and other areas (e.g. Taylor 1986), it is only recently that many warranty data issues have been studied. Starting with Suzuki (1985 a,b), there has been a good deal of development over the past 10-15 years (e.g. Kalbfleisch, Lawless and Robinson 1991, Robinson and McDonald 1991, Kalbfleisch and Lawless 1996). The objectives of the present paper are to review and extend this work, and to indicate areas for further investigation.

Certain features of warranty claims data create interesting statistical problems. For illustration we briefly introduce two examples which will be discussed further in Section 4. The first involves North American automobiles. In this case the manufacturer knows the dates of sale for all cars sold up to a given time, and has a record of warranty claims made for each vehicle. However, because warranties have dual time and mileage limitations (e.g. 2 years and 24,000 miles), the manufacturer is not aware of exactly how many vehicles are still under warranty at a given time. In addition, delays may occur in the recording of claims in their warranty data base, so the manufacturer may not be aware of a substantial number of recent claims at any given time.

The second illustration involves refrigerators. In this case the warranty coverage is for a fixed period after sale. However, at a given time the exact number of units sold and most of the dates of sale are not known by the manufacturer, since retailers and purchasers are not compelled to submit this information. Dates of sale and other information are obtained on most units only when a claim is made.

This paper focuses on two important areas. The first is age-based (or age-specific) analysis of warranty claims and costs, in which claims are related to the time since a product was sold, or entered service. This forms the basis for the estimation, prediction and comparison of claims across different product groups or across time, and is developed and illustrated in Sections 2-5. The second topic concerns the estimation of failure distributions or rates from warranty data, and is discussed in Section 6. If the causes of claims are diagnosed and reported accurately and there is accurate information about concomitant variables, the methods described there may be used to assess field reliability and make comparisons with engineering predictions and reliability goals. They can also provide ideas for reliability improvement and aid in the design of warranty, maintenance and parts replacement programs.

Complex products experience warranty claims for many different reasons, and in practice claims are often subdivided by type and studied separately. In this paper we typically consider a generic single type of claim, but it should be clear that multiple types can be considered. Relationships between different types of claims are not discussed in any detail, but we comment on this topic in Section 7, which concludes the paper. Other topics that are left out of the main body of the paper for the sake of brevity are also mentioned in Section 7. Very simple statistical tools such as Pareto charts showing the relative cost, frequency, or importance of different types of claims are not discussed in the paper, but are clearly valuable.

2. Age-Based Claims Analysis

For most products the rate of occurrence of warranty claims varies with the time since the product was sold or entered service; we refer to the time since sale as the "age" of the product. It is consequently informative to analyze claims as a function of age, bearing in mind that other factors may also need to be examined. Methodology in this area has been given by Kalbfleisch, Lawless and Robinson (1991), Robinson and McDonald (1991), and Kalbfleisch and Lawless (1996). In this and the next section we review their techniques, give some extensions, and note topics that deserve further study.

2.1 Notation and Assumptions

We focus first on the frequency of claims, and consider costs later. Suppose that product units $i = 1, 2, \cdots, N$ are sold over calendar days $0, 1, \cdots, \tau$ and let $N(d) \geq 0$ denote the number of units sold on day d. Assume that the data include all claims reported to the manufacturer by calendar day T. Let d_i denote the day of sale for unit i and define

$$n_i^T(a) \quad = \quad \text{number of claims at age } a \text{ for unit } i\text{, reported by day } T.$$

Age is measured in days. The upper limit for a is determined by the warranty plan and by convention claims made on products either before or on the day of sale are considered made at age 0. It is also convenient to define aggregate claim counts; we let $n^T(d, a)$ be the total reported number of claims at age a for those units sold on day d, or

$$n^T(d, a) = \sum_{i:d_i = d} n_i^T(a) \quad (a \geq 0, d \geq 0, a + d \leq T).$$

We also define $n^{*^T}(t, a) = n^T(t - a, a)$, which is the total claims on day t for units sold on day $t - a$.

We assume that there is a conceptual population of units for which those observed are representative, and define

$$\lambda(a) \;=\; \text{expected number of claims for a unit at age } a \quad (a = 0, 1, \cdots)$$
$$\Lambda(a) \;=\; \sum_{u=0}^{a} \lambda(u) = \text{expected number of claims per unit up to age } a$$

It should be noted that all units sold are not necessarily under warranty up to the same age; with automobiles, for example, there may be both age and mileage limits on coverage. The $\lambda(a)$'s represent the expected number of claims per unit sold, and not the expected number per unit still under warranty at age a. In addition, if a failed unit is replaced with a new one under a warranty, the age of the unit continues to be the time since sale of the original unit. These definitions are most relevant for the analysis of aggregate claims and costs.

There are often delays in reporting claims to the manufacturer, so that the reported number of claims by day T is less than the actual number. To handle this we assume the distribution of reporting delays is stable over time, defining

$$f(r) = \Pr\ (\text{a claims is reported } r \text{ days after it occurs}) \qquad r = 0, 1, 2, \cdots$$

and $F(r) = f(0) + \cdots + f(r)$. The term "reported" here refers to the point at which the record of the claim is entered into the data base used for analysis. Extensions to deal with non-stationary reporting delays are straightforward but will not be considered.

Simple age-based analysis, described in subsection 2.2, is founded on the estimation of $\lambda(a)$ and $\Lambda(a)$ for homogeneous populations of product units. Later, extensions to deal with explanatory variables are considered. Thus for the time being we ignore restrictions concerning day-of-the week or other calendar time effects in the occurrence and reporting of claims.

2.2 Analysis of Claim Frequencies

We assume that

$$E\left\{n_i^T(a)\right\} = \lambda(a)F\left(T - d_i - a\right) \tag{2.1}$$

and, for now, that the $F(r)$'s are known. Kalbfleisch et al. (1991) describe their estimation. In many applications reporting delays longer than two months are rare, and most claims are reported within a month. When no reporting delays occur, we set $F(r) = 1$ for all $r \geq 0$; then many of the estimates and expressions below simplify considerably.

Under (2.1) the expected values of $n^{*T}(t,a)$ and $n^{T}(d,a)$ are respectively $\mu^{*T}(t,a) = N(t-a)\lambda(a)F(T-t)$ and $\mu^{T}(d,a) = N(d)\lambda(a)F(T-d-a)$. An obvious moment estimate for $\lambda(a)$ is

$$\hat{\lambda}(a) = \frac{n^{T}(a)}{R^{T}(a)} \tag{2.2}$$

where

$$n^{T}(a) = \sum_{t=a}^{T} n^{*T}(t,a) = \sum_{d=0}^{T-a} n^{T}(d,a) \tag{2.3}$$

is the total number of age a claims reported up to day T, and

$$R^{T}(a) = \sum_{d=0}^{T-a} N(d)F(T-d-a). \tag{2.4}$$

The estimates (2.2) and the corresponding estimates

$$\hat{\Lambda}(a) = \sum_{u=0}^{a} \hat{\lambda}(u) \tag{2.5}$$

are very useful; examples are given in Section 4.

Variance estimates for $\hat{\Lambda}(a)$ and confidence limits for $\Lambda(a)$ are desirable when comparing different groups of products, for example, units manufactured over different periods or at different locations. The estimates (2.2) and (2.5) are maximum likelihood estimates under a Poisson model with the counts $n^{T}(d,a)$ independent, but extra-Poisson variation is typically evident. Assuming that checks of residuals $n^{T}(d,a) - \hat{\mu}^{T}(d,a)$, where $\hat{\mu}^{T}(d,a) = N(d)\hat{\lambda}(a)F(T-d-a)$, do not indicate any problem with the assumed form of $\mu^{T}(d,a)$, variance estimation can be approached in several ways. Kalbfleisch et al. (1991) and Kalbfleisch and Lawless (1996) use specifications for $\text{Var}\{n^{T}(d,a)\}$ that incorporate extra-Poisson variation. An alternative approach, taken here, is to specify a model at the product unit level.

One line of attack is to use (2.1) and robust methods of variance estimation developed by Lawless and Nadeau (1995). A straightforward extension of the results in their Section 2 yields the estimate

$$\widehat{\text{Var}}\left\{\hat{\Lambda}(a)\right\} = \sum_{i=1}^{N}\left\{\sum_{u=0}^{\min(T-d_i,a)} \frac{n_i^T(u) - \hat{\lambda}(u)F\left(T-d_i-u\right)}{R^T(u)}\right\}^2 . \tag{2.6}$$

This estimate is robust in the sense that it does not require any strong assumptions about the claims processes for product units. Approximate $2(\alpha - .5)$ confidence limits for $\Lambda(a)$ of the form $\hat{\Lambda}(a) \pm z_\alpha \widehat{\text{Var}}\left\{\hat{\Lambda}(a)\right\}^{1/2}$, where z_α is the standard normal α quantile, generally have good coverage properties with this method when N is moderately large. However, (2.6) requires claims data on individual units, and not just aggregate claim frequencies $n^T(d, a)$; if N is very large it requires considerable, though trivial, computation; finally, it is not easily adapted to deal with the prediction of claims, as discussed in Section 4.

We sketch an alternative approach which uses only aggregate claims. Denote

$$\begin{aligned} n_i(a) &= \text{number of claims on unit } i \text{ at age } a \text{ (reported or not)} \\ \overline{n}_i^T(a) &= n_i(a) - n_i^T(a), \end{aligned}$$

so that $\overline{n}_i^T(a)$ is the number of age a claims on unit i not reported by day T. Note that $E\{n_i(a)\} = \lambda(a)$ and $E\{\overline{n}_i^T(a)\} = \lambda(a)\overline{F}(T - d_i - a)$, where $\overline{F}(r) = 1 - F(r)$. We assume as well that reporting delays on claims on distinct days are independent, and

$$n_i^T(a)|n_i(a) \sim \quad \text{Binomial} \quad \left(n_i(a), F(T - d_i - a)\right). \tag{2.7}$$

We further denote

$$\sigma_{ab} = \text{cov}\left\{n_i(a), n_i(b)\right\}. \tag{2.8}$$

Writing $F_{ia} = F(T - d_i - a)$ and $\overline{F}_{ia} = 1 - F_{ia}$, we then have

$$\text{cov}\left\{n_i^T(a), n_i^T(b)\right\} = F_{ia}F_{ib}\sigma_{ab} + I(a = b)\lambda(a)F_{ia}\overline{F}_{ia} \tag{2.9}$$

$$\text{cov}\left\{n_i^T(a), \overline{n}_i^T(b)\right\} = F_{ia}\overline{F}_{ib}\sigma_{ab} - I(a = b)\lambda(a)F_{ia}\overline{F}_{ia}. \tag{2.10}$$

From (2.9), variances for the $\hat{\lambda}(a)$'s and $\hat{\Lambda}(a)$'s are readily obtained. Define

$$R^T(a, b) = \sum_{i=1}^{N} F_{ia}F_{ib} = \sum_{d=0}^{T-\max(a,b)} N(d)F(T - d - a)F(T - d - b)$$

$$\overline{R}^T(a,b) = \sum_{i=1}^{N} F_{ia}\overline{F}_{ib} = \sum_{d=0}^{T-\max(a,b)} N(d)F(T-d-a)\overline{F}(T-d-b).$$

Then, assuming independence of claims and reporting across units $i = 1, \cdots, N$, we have

$$\begin{aligned} \text{Cov}\left\{\hat{\lambda}(a), \hat{\lambda}(b)\right\} &= \frac{1}{R^T(a)R^T(b)} \sum_{i=1}^{N} \text{Cov}\left\{n_i^T(a), n_i^T(b)\right\} \\ &= \frac{R^T(a,b)\sigma_{ab} + I(a=b)\lambda(a)\overline{R}^T(a,a)}{R^T(a)R^T(b)} \qquad (2.11) \\ \text{Var}\left\{\hat{\Lambda}(a)\right\} &= \sum_{a=0}^{u}\sum_{b=0}^{u} \text{Cov}\left\{\hat{\lambda}(a), \hat{\lambda}(b)\right\} \qquad (2.12) \end{aligned}$$

Various specifications for σ_{ab} can be entertained. For example, if we were to consider a mixed Poisson model (Lawless 1987) where we associate with each unit an unobservable random variable with mean 1 and variance σ^2, we would obtain $\sigma_{aa} = \lambda(a) + \sigma^2\lambda(a)^2$, $\sigma_{ab} = \sigma^2\lambda(a)\lambda(b)$ for $a \neq b$. An approach that assumes less is to simply allow the σ_{ab}'s to be arbitrary variances and covariances. In any case, we need to estimate the unknown parameters σ_{ab} in order to obtain estimates for (2.11) and (2.12). Quadratic moment estimating functions can be used for this purpose: for example, under the model given above, where

$$\sigma_{ab} = \sigma^2\lambda(a)\lambda(b) + I(a=b)\lambda(a), \qquad (2.13)$$

we could consider estimating functions of the form

$$\sum_{d}\sum_{a} w(d,a)\left\{\left[n^T(d,a) - \hat{\mu}^T(d,a)\right]^2 - \hat{v}(d,a;\sigma)\right\} = 0, \qquad (2.14)$$

where $\hat{v}(d,a;\sigma) = \hat{\text{Var}}\left\{n^T(d,a)\right\} = N(d)F(T-d-a)^2\hat{\lambda}(a)^2\sigma^2 + N(d)F(T-d-a)\hat{\lambda}(a)$, and the $w(d,a)$'s are weights. The practical aspects of estimating variance parameters has not been examined in any detail, though theoretical properties of quadratic estimating functions are clear. The choices $w(d,a) = 1$ and $w(d,a) = \hat{v}(d,a;\sigma)^{-1}$ are often used.

The use of (2.6) or of (2.13) combined with (2.14) tend to give similar variance estimates for $\hat{\Lambda}(a)$ in many situations. A simpler approach given by Kalbfleisch and Lawless (1996) also gives similar results in many cases. This makes the assumptions that $\text{Var}\left\{n^T(d,a)\right\} =$

$\sigma^2\mu^T(d,a)$ and that $n^T(d,a)$'s are mutually independent. These assumptions are not very plausible when enough units suffer reporting delays, but appear to work reasonably well. They give the simple result

$$\text{Var}\left\{\hat{\Lambda}(a)\right\} = \sigma^2 \sum_{u=0}^{a} \frac{\lambda(u)}{R^T(u)}. \tag{2.15}$$

The variance parameter σ^2 can be estimated from equations of the form (2.14) with $\hat{v}(d,a;\sigma) = \sigma^2\hat{\mu}^T(d,a)$. Experience suggests that the weights $w(d,a) = \sigma^{-4}\hat{\mu}^T(d,a)^2$ work well, though the choices $w(d,a) = 1$ and $w(d,a) = \hat{v}(d,a;\sigma)^{-1}$ are also often used. Dean (1992) compares methods in a simpler context involving independent univariate counts.

The variance estimates given here assume that the $N(d)$'s and $F(r)$'s are both known. Uncertainty about their values is discussed in Section 3.2.

2.3 Analysis of Costs

Age-based analysis of warranty costs may be handled by a simple extension of the methods above. We stratify claims into groups $k = 1, 2, \cdots, K$ according to cost and let $C(k)$ be the (average) cost of a claim in group k. Let $\lambda_k(a)$ denote the expected number of claims of cost $C(k)$ for a unit at age a. Then, corresponding to (2.2), we have the estimate

$$\hat{\lambda}_k(a) = \frac{n^T_{(k)}(a)}{R^T(a)} \tag{2.16}$$

where, in an obvious notation, $n^T_{(k)}(a) = \sum_{d=0}^{T-a} n^T_{(k)}(d,a)$ is the number of claims of cost $C(k)$ and age a reported up to time T. The cumulative warranty cost per unit up to age a is estimated by

$$\widehat{CC}(a) = \sum_{k=1}^{K} C(k)\hat{\Lambda}_k(a) \tag{2.17}$$

where $\hat{\Lambda}_k(a) = \sum_{u=0}^{a} \hat{\lambda}_k(u)$. It is assumed that the same reporting delay distribution applies to claims of different costs. This assumption can be relaxed but generally has little effect.

Confidence limits for $CC(a)$ can be obtained by utilizing variance estimates analogous to those given for $\hat{\Lambda}(a)$ in Section 2.2. Assuming that occurrences of claims of different costs are independent, we have

$$\text{Var}\left\{\hat{C}C(a)\right\} = \sum_{k=1}^{K} C(k)^2 \text{Var}\left\{\hat{\Lambda}_k(a)\right\}.$$

where $\text{Var}\left\{\hat{\Lambda}_k(a)\right\}$ may be estimated by expressions such as (2.6).

It is possible that the occurrence of certain types of claims may not be independent, for example, when two or more parts are physically linked in the product. The effect of this on the variance of age-specific cost curve estimates $\widehat{CC}(a)$ is generally slight. However, if desired, a variance estimate that is robust to non-independence of claims may be obtained by using the approach of Lawless and Nadeau (1995), which gave (2.6). Obviously $\widehat{CC}(a)$ in (2.17) can be written as

$$\widehat{CC}(a) = \sum_{u=0}^{a} \hat{c}c(u) = \sum_{u=0}^{a} \frac{tc(u)}{R^T(u)},$$

where $tc(u)$ is the total cost of all reported claims which occurred at age u. The variance of $\widehat{CC}(a)$ can be estimated by

$$\sum_{i=1}^{N} \left\{ \sum_{u=0}^{\min(T-d_i,a)} \frac{tc_i(u) - \hat{c}c(u)F(T-d_i-u)}{R^T(u)} \right\}^2 \tag{2.18}$$

where $tc_i(u)$ is the total cost of all reported claims on unit i at age u.

3. Refinements to Age-Based Analysis

The methods in the preceding section deal with the ideal situation in which the dates of sale are known almost exactly for all products, as well as the day of occurrence for claims. Often the data are grouped over longer time intervals and the daily numbers of units sold are unknown. In fact, for most products the manufacturer knows when units are shipped to a customer or retail outlet but does not know when they are sold. (Dates of sale are, however, obtained for products that experience a warranty claim.) In this section we consider grouped data and uncertainty about sales or reporting delays. We also discuss the incorporation of covariates or calendar time effects into age-based analysis.

3.1 Grouped Claim Frequency Data

Sometimes claims are grouped by age and time intervals. If the age intervals and time periods are of different lengths or are fairly long, it is best to recognize this in the analysis. Extending the notation in Section 2.1, we write

$$n^{*T}(P,A) = \sum_{t\epsilon P}\sum_{a\epsilon A} n^{*T}(t,a)$$

where P and A represent a time period and age interval, respectively. Suppose now that claims are grouped into age intervals $A_j = [a_{j-1}, a_j)$ with $a_0 = 0 < a_1 < a_2 < \cdots$. In this case we estimate the expected number of claims for that interval,

$$\Lambda\left(A_i\right) = \sum_{a=a_{j-1}}^{a_j-1} \lambda(a). \tag{3.1}$$

A natural estimate of (3.1) is

$$\hat{\Lambda}\left(A_j\right) = \frac{n^T\left(A_j\right)}{R^T\left(A_j\right)}, \tag{3.2}$$

where

$$\begin{aligned} n^T\left(A_j\right) &= \sum_{t=0}^{T} n^{*T}\left(t, A_j\right) \\ R^T\left(A_i\right) &= \frac{1}{a_j - a_{j-1}} \sum_{a=a_{j-1}}^{a_j-1} R^T(a). \end{aligned} \tag{3.3}$$

To motivate (3.2) note that $n^T\left(A_j\right)$ is the number of claims reported on units of all ages $a \in A_j$, and that

$$E\left\{n^T(A)\right\} = \sum_{a=a_{j-1}}^{a_j-1} \lambda(a) R^T(a). \tag{3.4}$$

Assume (with little practical consequence if $a_j - a_{j-1}$ is not too large) that the $\lambda(a)$'s are constant over $[a_{j-1}, a_j)$, and then (3.4) reduces to $\Lambda(A_j)\, R^T(A_j)$.

If sales data are available only in aggregate form for different time periods, then we need to estimate the "exposures" (3.3). The simplest approach is to estimate the daily sales $N(d)$ from the available data and plausible assumptions about sales patterns, so that (2.4) can be used. Dealing with uncertainty in the $N(d)$'s is discussed in subsection 3.2.

Variance estimates for $\hat{\Lambda}(A_j)$'s can be developed by using approaches described in Section 2.2 combined with the assumption that the $\lambda(a)$'s are constant within age intervals A_j. A crude approach based on assumptions leading to (2.15) gives

$$\text{Var}\left\{\hat{\Lambda}(A_j)\right\} = \sigma^2 \frac{\Lambda(A_j)}{R^T(A_j)}, \tag{3.5}$$

with $\hat{\Lambda}(A_j)$'s independent for $j = 1, 2, \cdots$. The variance parameter σ^2 can be estimated by using estimating equations similar to (2.14), with sums over d and a replaced by sums over time and age intervals:

$$\sum_l \sum_j w(l,j) \left\{ \left[n^{*T}(P_l, A_j) - \hat{\mu}^{*T}(P_l, A_j) \right]^2 - \sigma^2 \hat{\mu}^{*T}(P_l, A_j) \right\} = 0 \tag{3.6}$$

where $w(l,j)$ are weights and

$$\hat{\mu}^{*T}(P_l, A_j) = \frac{\hat{\Lambda}(A_j)}{a_j - a_{j-1}} \sum_{a=a_{j-1}}^{a_j - 1} \sum_{t \epsilon P_j} N(t-a) F(T-t).$$

3.2 Estimated Sales and Reporting Delay Probabilities

As noted previously, daily sales totals $N(d)$ often have to be estimated. In addition, reporting delay probabilities $F(r)$ are typically estimated from historical data. Variance estimates for $\hat{\Lambda}(a)$ and other quantities given above have ignored these sources of variation.

If values of $N(d)$ and/or $F(r)$ are estimated then the effect on estimates of $\Lambda(a)$ is that the exposures $R^T(a)$ given by (2.4) are estimated; let us denote them by $\tilde{R}^T(a)$ to indicate this. A detailed analysis that recognized how the $N(d)$'s and $F(r)$'s were estimated would in principle be feasible in some situations, but rather cumbersome. Appealing alternatives are to perform a sensitivity analysis or to use simulation to assess the effect of estimates $\overset{\sim T}{R}(a)$ on estimation of $\Lambda(a)$. With the former approach we select a few sets of $\overset{\sim T}{R}(a)$'s that cover what is considered plausible, and compute $\tilde{\Lambda}(a)$'s and associated variance estimates and confidence intervals for each. This gives a range of estimates, and conveniently expresses our uncertainty. This is a common approach for dealing with unknown sales totals $N(d)$; for example, a manufacturer might select scenarios that assume each unit is sold a specified number of days after it is shipped.

The latter approach above is a little more formal. Suppose that the estimates $\overset{\sim T}{R}(a)$ are approximately unbiased and independent of the claim counts $n^T(d,a)$. Then for (2.5) we have, letting $\tilde{\boldsymbol{R}}_T = \left\{\overset{\sim T}{R}(a), a = 1, 2, \cdots\right\}$,

$$\text{Var}\left\{\hat{\Lambda}(a)\right\} = E\left\{\text{Var}\left(\hat{\Lambda}(a) | \tilde{\boldsymbol{R}}_T\right)\right\} + \text{Var}\left\{\sum_{u=0}^{a} \lambda(u) \frac{\overset{\sim T}{R}(u)}{R^T(u)}\right\}. \tag{3.7}$$

The first term on the right side of (3.7) can be estimated by using estimates such as (2.6), (2.12) or (2.15), evaluated at values $\tilde{\lambda}(a), \overset{\sim T}{R}(a)$. The second term on the right side can be estimated by placing a distribution on $\overset{\sim T}{R}(u)/R^T(u)$ values and using simulation, replacing $\lambda(a)$ with $\hat{\lambda}(a)$. This approach has yet to be explored in any detail and, of course, methods of determining a distribution for values $\overset{\sim T}{R}(a)/R^T(a)$ need to be provided for specific contexts.

3.3 Covariate or Calendar-Time Effects

Claim frequencies may be related to factors such as manufacturing conditions or usage environments, and covariates or calendar-time effects may be incorporated into age-based analysis to deal with them. If only a few separate conditions are of interest the simplest approach is to estimate expected claims separately for each condition, as illustrated in Section 4.1 where claims for automobiles manufactured in different time periods are compared. More generally, covariates can be used to represent various factors, and regression models employed.

Covariates may be introduced at an individual unit level or at an aggregate claims level. In either case log linear models (e.g. McCullagh and Nelder 1989, Chapter 6) are convenient. For example, to incorporate a calendar time effect (e.g. a seasonal factor) at the aggregate claims level we might consider

$$E\left\{n^{*T}(t,a)\right\} = N(t-a)\lambda(a)h(t;\beta)F(T-t),$$

where $h(t;\beta)$ models the calendar time effect. The parameters $\lambda(a)$ and β may be estimated using quasi-likelihood, say by employing estimating functions based on a model whereby the $n^{*T}(t,a)$'s are mutually independent Poisson random variables. Standard software will readily handle this, but use of the Poisson-based variance estimates for the $\hat{\lambda}(a)$'s and $\hat{\beta}$ is not recommended, since the counts are generally neither approximately Poisson nor independent. Robust sandwich-type variance estimates can be developed, provided that one decides which groups of responses are to be considered independent. For example, rather than assume that the $n^{*T}(t,a)$'s are mutually independent, it might be preferable to assume only that sets $\left\{n^T(d,a); a=1,2,\cdots\right\}$ are independent for $d=1,2,\cdots$. An alternative is to adopt a parametric variance specification, e.g. independent $n^{*T}(t,a)$'s with $\text{Var}\left\{n^{*T}(t,a)\right\} = \sigma^2\mu^{*T}(t,a)$, as in McCullagh and Nelder (1989, Chapter 6).

In terms of individual units, regression models of the form

$$E\left\{n_i^T(a)\right\} = \lambda(a)F(T-d_i-a)e^{\boldsymbol{\beta}'\boldsymbol{z}_i(a)} \tag{3.8}$$

are valuable. The vector $\boldsymbol{z}_i(a)$ is a covariate which may vary across different ages for the unit, for example if it involves calendar time effects. Given assumptions about $\text{Var}\left\{n_i^T(a)\right\}$, model fitting is straightforward though when the number of units is large, aggregation of counts is helpful. Models for aggregate counts such as $n^{*T}(t,a)$ follow from (3.8).

When units of age are small (e.g. days) models such as (3.8) have many parameters and special methods to deal with this are helpful. The methods of Lawless and Nadeau (1995) allow easy calculation of robust variance estimates for $\hat{\Lambda}(a)$ and $\hat{\beta}$ provided one has access to claims on a unit by unit basis. An alternative is to use some form of smoothing with respect to the $\lambda(a)$'s, or to make parametric assumptions about the $\lambda(a)$'s, as in Section 3.4 below.

3.4 Piecewise Constant Claim Rates

A convenient but mild assumption to reduce computational burden is to assume that the $\lambda(a)$'s are piecewise constant. In particular, let us suppose that

$$\lambda(a) = \lambda_j \qquad a_{j-1} \leq a < a_j, \tag{3.9}$$

where $0 = a_0 < a_1 < a_2 < \cdots < a_J$ are pre-specified values. In this case it is easily seen that robust estimates for the λ_j's under the assumption that the $n^T(d,a)$'s have means $N(d)\lambda(a)F(T-d-a)$ are

$$\hat{\lambda}_j = \frac{n^T(A_j)}{R_1^T(A_j)} \qquad j = 1, \cdots, J \tag{3.10}$$

where $A_j = [a_{j-1}, a_j)$, and

$$\begin{aligned} n^T(A_j) &= \sum_d \sum_a n^T(d,a) I(a \epsilon A_j) \\ R_1^T(A_j) &= \sum_d \sum_a N(d) F(T-d-a) I(a \epsilon A_j). \end{aligned}$$

This reproduces (3.2) as an estimate of $\Lambda(A_j)$.

The numerator in (3.10) is the total number of claims in age interval A_j reported by time T. When there are no reporting delays the denominator is the total number of days in service spent in age interval A_j across all units; when there are reporting delays, days in service are discounted. Variance estimates can be obtained as in Section 3.1.

In some industries it is common to estimate warranty or field failure rates based on a model (3.9) with $J = 1$, i.e. it is assumed that the rate is constant with respect to age. That is usually an inappropriate assumption, but a model with J as small as 2,3 or 4 may suffice. For many products, for example, it is adequate to define age intervals to cover early (or "infant mortality"), medium term, and late term failures.

The assumption of piecewise constant $\lambda(a)$'s is likewise a great simplification for regression models as discussed in Section 3.3. In particular, if J is small the calculation of robust variance estimates for the $\tilde{\lambda}_j$'s and regression parameter estimates $\hat{\beta}$ is easy, at least when unit-level data are available (Lawless and Nadeau 1995). Further development for aggregate claims would be useful.

4. Examples

We use two examples to illustrate aspects of age-based claims analysis.

4.1 Automobile Warranty Claims

We consider data on claims for a single system on a particular car model. The warranty coverage at the time was for the minimum of 12,000 miles or one year from sale. More detailed discussions of the data are given by Kalbfleisch et al. (1991) and Kalbfleisch and Lawless (1996).

The data included claims reported up to day $T = 547$ after the first vehicle was sold. By then there were 36,683 cars sold and 5,701 claims had been reported. Sales are reported almost immediately by the car dealers, so the sales total $N(d), d = 0, 1, 2, \cdots$ were all known. Reporting delays were typically between 20 and 60 days, and reporting delay probabilities $F(r)$ were estimated from previous claims experience. Expected claim frequencies $\hat{\lambda}(a)$ were obtained using (2.2) with $R^T(a)$ given by (2.4). An important use of the $\hat{\lambda}(a)$'s and $\hat{\Lambda}(a)$'s is to compare claims for vehicles manufactured in different time periods or locations. Figure 1 shows estimates $\hat{\Lambda}(a)$ for cars manufactured in each of 6 two-month periods, starting with mid-July to mid-September (Period 1) and going to mid-May to mid-July of the following year (Period 6). There were roughly the same number of cars produced in each period. The plots indicate that average claims per vehicle are similar for all periods except number 3 (November-January), for which claims are considerably higher. Figure 1 also shows approximate pointwise 95% confidence limits for $\Lambda(a)$ for Period 3, calculated as $\hat{\Lambda}(a) \pm 1.96\hat{V}(a)^{1/2}$, where $\hat{V}(a)$ is an estimate of $\text{Var}\left\{\hat{\Lambda}(a)\right\}$. The variance estimate was calculated using (2.6), but with no allowance for the fact that the reporting delay probabilities are estimated. Variance estimates based on (2.5) give close to the same confidence limits as those shown in the figure. Confidence limits for the other five periods may be calculated similarly; they overlap each other considerably but do not overlap the limits for $\Lambda(a)$ of Period 3.

Regression methods as outlined in Section 3.3 can be used if formal tests concerning the six production periods are wanted. For example, we may define a covariate vector $\boldsymbol{z}_i = (z_{i1}, \cdots, z_{i5})'$ where $z_{ij} = 1$ if vehicle i was produced in Period j and 0 otherwise. Using the multiplicative model (3.8), we consider the hypothesis $H : \boldsymbol{\beta} = \mathbf{0}$; as Figure 1 indicates, a multiplicative model seems appropriate. We can test H using the Wald statistic $W = \hat{\boldsymbol{\beta}}' V(\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}$, where $V(\hat{\boldsymbol{\beta}})$ is an estimate of the asymptotic variance of $\hat{\boldsymbol{\beta}}$. Lawless and Nadeau (1995) find an observed value $W = 85.6$, with a robust variance estimate for $\hat{\boldsymbol{\beta}}$ developed in that paper. Under H, the distribution of W is approximately $\chi^2_{(5)}$, so there is very strong evidence against H.

4.2 Refrigerator Warranty Claims

An appliance manufacturer ships refrigerators to retailers who then sell them. The manufacturer tracks warranty claims with units stratified according to month of production. However, the exact number of units sold and the dates of sale are not known by the manufacturer because retailers and purchasers are not compelled to submit this information. At best, a fraction of purchasers mail in a warranty card that indicates the place and time of sale, along with the serial number of the product.

The warranty on major components extends for two years from the date of sale. When a claim is submitted the date of sale of the unit and its month of production (from the serial number) are obtained as part of the verification process. There are generally delays of up to a month between the time of a claim and the time it enters the warranty data base.

Consider units produced in a particular month, and let $\lambda(a)$ denote the expected number of claims per unit at age a (days). Given that $N(d)$ units are sold on day d, the expected number of age a claims reported by day T is $\lambda(a)R^T(a)$, where $R^T(a)$ is given by (2.4). However, the $N(d)$'s are unknown so we proceed as in Section 3.2. In this case we can estimate $N(d) = MP(d)$, where M is the number of units produced in the month and $P(d)$ is the fraction sold on day d, from information about the distribution of the time lag between delivery to the retailer and date of sale. In many cases manufacturers do this by making the very naive assumption that all units are sold some fixed time (say 1 month) after being shipped to the retailer. However, it is usually possible to do better, especially for products with seasonal variations in sales.

Data are often grouped by month or week, so we have a scenario like that in Table 1. The sales numbers N_{ij} are estimated by the manufacturer, and they allow estimation of the weekly sales totals. When data are grouped by week, as here, it is generally adequate to treat weeks as the basic time unit (i.e. replace "days" with "weeks" in the methodology presented earlier); there is little be be gained, and a considerable increase in computation, in using the formulas for grouped data in Section 3.1.

Table 1. Estimated Sales Patterns

		Week Sold					
Week Shipped	Number Shipped	1	2	3	4	5	6
1	M_1	N_{11}^*	N_{12}	N_{13}	N_{14}	N_{15}	N_{16}
2	M_2		N_{22}	N_{23}	N_{24}	N_{25}	N_{26}
3	M_3			N_{31}	N_{32}	N_{33}	N_{34}
4	M_4				N_{41}	N_{42}	N_{43}
		$N(1)$	$N(2)$	$N(3)$	$N(4)$	$N(5)$	$N(6)$

N_{ij}^* = number of units shipped in Week i and sold in Week j.

As discussed in Section 3.2 we may wish to assess the effect of uncertainty in the $R^T(a)$ values on estimates $\hat{\Lambda}(a)$ or associated confidence intervals.

5. Prediction of Claims

Forecasts of warranty claims and costs are of considerable importance to manufacturers. Age-based claims analysis provides a convenient approach, but we switch the focus from expected claim counts for a hypothetical infinite population of units, as in Section 2 and 3, to the specific finite population that is manufactured over the period of interest. For example, automobile makers may wish to estimate the average per vehicle warranty cost for the vehicles produced in a given model year, or to forecast warranty expenses for a model year on a month by month basis, starting at the time of production of the first vehicles. The finite population approach below may also in some situations be preferred for the estimation or comparison of average or total claims, as illustrated in Section 4.

Kalbfleisch et al. (1991) presented methods of prediction which we outline and extend here. Consider a finite population of $N = \sum_{d=0}^{\tau} N(d)$ units sold over the time period $(0, \tau)$, where $N(d)$ is the number sold on day d. The actual average number of claims per unit at age a is

$$m(a) = \frac{1}{N} \sum_{d=0}^{\tau} n(d,a) \qquad a = 0, 1, 2, \cdots$$

and $M(a) = \sum_{u=0}^{a} m(u)$ is the average number of claims per unit up to age a. Using notation similar to that of Section 2.3, we also have

$$M_k(a) = \frac{1}{N} \sum_{u=0}^{a} \sum_{d=0}^{\tau} n_{(k)}(d,u)$$

as the average number of claims of cost $C(k)$ up to age a, per unit sold, and

$$CM(a) = \sum_{k=1}^{K} C(k) M_k(a)$$

as the average total cost per unit. If data on claims to time T are available then the $m(a)$'s are only partially known, because $n(d,a)$ is as yet unobserved if $d + a > T$, and $N(d)$ is also unobserved if $\tau > T$. Some claims may also be unobserved due to reporting delays. Thus, estimation of the $m(a)$'s $M(a)$'s or $M_k(a)$'s reduces to prediction of the $n(d,a)$'s and $n_{(k)}(d,a)$'s based on data currently available.

In the notation of Sections 2.2 and 2.3, and considering one type of claim, we have

$$\begin{aligned} m(a) &= \frac{1}{N}\left\{\sum_{d=0}^{\tau} n^T(d,a) + \sum_{d=0}^{\tau} \overline{n}^T(d,a)\right\} \\ &= \frac{1}{N}\left\{\sum_{i=1}^{N} n_i^T(a) + \sum_{i=1}^{N} \overline{n}_i^T(a)\right\}. \end{aligned} \tag{5.1}$$

A natural point estimate of $m(a)$ is

$$\hat{m}(a) = \frac{1}{N}\left\{\sum_{d=0}^{\tau} n^T(d,a) + \sum_{d=0}^{\tau} N(d)\hat{\lambda}(a)\overline{F}(T-d-a)\right\}.$$

Since $n^T(d,a) = \hat{\lambda}(a)R^T(a)$, $f(r) = 0$ for $r < 0$, and $N(d) = 0$ for $d > \tau$, this reduces to

$$\hat{m}(a) = \hat{\lambda}(a),$$

where $\hat{\lambda}(a)$ is based on the data observed up to time T. Thus we also take $\hat{M}(a) = \hat{\Lambda}(a)$. If we wish to predict claims for ages a that are greater than T, estimates $\hat{\lambda}(a)$ have to be based on sources other than the current data.

To obtain prediction limits for $M(a)$ we require estimates of $\mathrm{Var}\left\{\hat{M}(a) - M(a)\right\} = V_M(a)$; approximate $2(\alpha - .5)$ limits are given by $\hat{M}(a) \pm z_\alpha \hat{V}_M(a)^{1/2}$, where z_α is the standard normal α quantile. Kalbfleisch and Lawless (1996) and Kalbfleisch et al. (1991) provide variance calculations under specific assumptions. We note that

$$\begin{aligned} \mathrm{Var}\ \{m(a) - \hat{m}(a)\} &= \mathrm{Var}\ \left\{\sum_{d=0}^{\tau} \overline{n}^T(d,a) - \sum_{d=0}^{\tau} N(d)\hat{\lambda}(a)\overline{F}(T-d-a)\right\} \\ &= \frac{1}{N}\mathrm{Var}\ \left\{\sum_{d=0}^{\tau} \overline{n}^T(d,a) - \left(N - R^T(a)\right)\hat{\lambda}(a)\right\}. \end{aligned} \tag{5.2}$$

To evaluate (5.2) we need $\mathrm{Var}\{\overline{n}^T(d,a)\}$ and $\mathrm{Cov}\left\{\overline{n}^T(d,a), \hat{\lambda}(a)\right\}$, and to further obtain $\mathrm{Cov}\{m(a) - \hat{m}(a), m(b) - \hat{m}(b)\}$ for $a \neq b$ we also need $\mathrm{Cov}\left\{n^T(a), n^T(b)\right\}$ and $\mathrm{Cov}\left\{n^T(a), \overline{n}^T(b)\right\}$. These are straightforward but tedious to write down under various variance models. For example, under the model (2.7)-(2.8), they may be obtained directly from expressions (2.9) and (2.10). Note that if all units sold have reached age a and if no further reporting delays are possible, then $m(a) = \hat{m}(a)$ and the right side of (5.2) equals zero.

As in the case of $\text{Var}\left\{\hat{\Lambda}(a)\right\}$ in Section 2.2, a crude variance model yields very simple estimates analogous to (2.15) that appear to work well in many situations. The (rather implausible) assumptions are that $\text{Var}\left\{n^T(d,a)\right\} = \sigma^2 N(d)F(T-d-a)\lambda(a)$, $\text{Var}\left\{\overline{n}^T(d,a)\right\} = \sigma^2 N(d)\overline{F}(T-d-a)\lambda(a)$, and that the $n^T(d,a)$'s and $\overline{n}^T(d,a)$'s are all mutually independent. Then it is easily shown from (5.2) that

$$\text{Var }\{m(a) - \hat{m}(a)\} = \sigma^2 \left\{\frac{N - R^T(a)}{NR^T(a)}\right\} \lambda(a),$$

and that

$$\text{Var }\left\{\hat{M}(a) - M(a) =\right\} = \sum_{u=0}^{a} \text{Var}\left\{m(u) - \hat{m}(u)\right\}.$$

This estimate was given by Kalbfleisch and Lawless (1996), who also provide a corresponding result for grouped data:

$$\text{Var }\{\hat{m}(A_j) - m(A_j)\} = \sigma^2 \left\{\frac{N - R^T(A_j)}{NR^T(A_j)}\right\} \Lambda(A_j).$$

The development above provides methods for the age-specific prediction of claims for a finite population of units. One may also wish to predict total claims and costs over calendar time. The number of age a claims at day t is just $n^*(t,a)$, with

$$E\{n^*(t,a)\} = \mu^*(t,a) = N(t-a)\lambda(a). \tag{5.3}$$

The total claims at day t is

$$n^*(t) = \sum_{a=0}^{t} n^*(t,a).$$

A point estimate (prediction) for $n^*(t)$ is given by using $\hat{n}^*(t,a) = N(t-a)\hat{\lambda}(a)$, for a day t in the future; values $N(t-a)$ may also have to be estimated. If t is a day in the past, but not all claims on day t may yet have been reported, then a point estimate is based on $n^*(t,a) = n^{*T}(t,a) + N(t-a)\overline{F}(T-t)\hat{\lambda}(a)$. Prediction intervals may be developed through estimation of $\text{Var}\{\hat{n}^*(t) - n^*(t)\}$.

Prediction of costs may also be handled through these methods, as indicated at the start of this section. When the objective is to provide forecasts of costs over calendar time, however, it is simpler to consider expected cost curves analogous to (5.3), i.e. to let $C^*(t)$ be the total cost of claims on day t, and to take

$$E\{C^*(t)\} = \sum_{a=0}^{t} N(t-a)cc(a), \tag{5.4}$$

where $cc(a)$ is the average warranty cost per unit at age a. We may extend (5.4) to include covariates or calendar time effects.

Prediction methodology and, in particular, methods of setting prediction limits, deserve investigation. These should have the property that as all of the units in the target population (e.g. automobiles sold over a specific time period) enter service and age, prediction limits for total claims or average claims per unit shrink. Early in the sales period one has to rely on informed guesses about sales and age-based claims, and a Bayesian framework for prediction of costs is therefore attractive. Robinson and McDonald (1991, Section 4) examine these issues. Chen et al. (1996) also discuss warranty claim forecasting.

6. Estimation of Field Reliability

Reliability and durability are important dimensions of the quality of a product. Manufacturers obtain information about field reliability from a variety of sources that include field tracking or followup studies, customer surveys, and warranty data. Robinson and McDonald (1991) and Lawless and Kalbfleisch (1992) provide some general discussion. Because followup studies and surveys are relatively expensive there is considerable interest in using information from warranty data to estimate replacement or failure rates and to provide ideas for reliability improvement.

A fundamental problem is that warranty claim records often do not identify the source of a problem correctly or accurately enough to be useful for engineering reliability assessment. This must be overcome by an effective identification and reporting process. Even then, there remain some interesting inferential problems. In particular, data are generally missing for units that do not experience a failure under warranty: dates of sale, usage information, and covariate information are typically obtained when a unit experiences a warranty claim, but are not available for units with no claims. Such information is not missing completely at random, and biased inference may result if it is not dealt with properly.

Most warranties cover the early life of a product and usually provide little direct information about longer term reliability or durability. Early life reliability is important as a dimension of quality and as a key factor in determining the costs of a warranty plan, but extrapolations well beyond the range of existing data must be treated cautiously.

We now discuss these issues in the context of failure time distributions and recurrent events, and indicate some methods of analysis.

6.1 Estimation of a Failure Time Distribution

Our objective here is to estimate the early portion of a failure time, or time-to-event, distribution. This is rather different than in the aggregate warranty claims analysis of preceding sections, because we wish to assess unit-to-unit variability and to identify factors that explain some of that variability. For example, with multiple or recurrent events it is of interest whether events occur across units in a fairly homogeneous way, or whether a small fraction of units account for the vast majority of events. In addition, we may want to examine usage or other factors in relation to failure.

To start, let T denote the time to "failure" for a product unit; time here could be either chronological (calendar) time or some type of usage time, such as miles driven for automobiles. Suppose that T is recorded if failure occurs while the unit is under warranty; for convenience we ignore reporting delays. If N units have entered service up to the current date, let $t_1, \cdots, t_n$ denote the failure times of units reported to fail, and let $\tau_{n+1}, \cdots, \tau_N$ denote the censoring times for the remaining units. Note that τ_i depends on the date of sale of the unit, the current date, the terms of the warranty coverage, and possibly the usage of the unit. For example, if a warranty covers a product for A days after purchase, then $\tau_i = \min(A, D_i)$, where D_i is the number of days between the purchase date and the current date.

Suppose that T has probability density function $f(t;\theta)$ and survivor function $S(t;\theta) = \Pr(T > t;\theta)$ where θ is a vector of parameters to be estimated and for simplicity we ignore covariates. If $\tau_{n+1}, \cdots, \tau_N$ are known then the familiar censored data likelihood function

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta) \prod_{i=n+1}^{N} S(\tau_i;\theta) \tag{6.1}$$

allows estimation of θ (e.g. Lawless 1982, Chapter 2). However, $\tau_{n+1}, \cdots, \tau_N$ are generally unknown for most or all of the product units not having failed under warranty. For example, with products such as appliances most customers and retailers do not report dates of sale for each unit; manufacturers honour a warranty claim if the date of sale can be validated when the claim is made. With time scales such as mileage for cars and for more complex warranty plans, censoring times are not available even if the date of sale happens to be known. Likewise, covariates recorded when warranty claims are made will usually be missing for units with no claim.

If the τ_i's in (6.1) are missing, some alternative must be sought. One approach is to use only $t_1, \cdots, t_n$, conditioning on the fact that $t_i \le \tau_i$ in each case; this assumes that $\tau_1, \cdots, \tau_n$ are known, which is usually the case. However, this approach is informative only about the conditional distribution of T, given that $T \le \tau_{\max}$, the largest censoring time observed. It is quite uninformative about the unconditional distribution of T (Kalbfleisch and Lawless 1988, Hu and Lawless 1996a) and for recurrent events (see Section 6.2) is very susceptible to model misspecification (Hu and Lawless 1996b). It is thus desirable to incorporate information about N and $\tau_{n+1}, \cdots, \tau_N$. This has been studied in a sequence of papers beginning with Suzuki (1985ab), and two main approaches have been proposed.

The first approach is to randomly select a followup sample F of m units from the $N - n$ that have not failed under warranty and to obtain censoring times (and any covariate values) for each unit; the data from this sample are used to "estimate" the second term on the right side of (6.1). For example, the weighted pseudo log likelihood function

$$\widetilde{\ell}\,(\theta) = \sum_{i=1}^{n} \log f(t_i;\theta) + \frac{N-n}{m} \sum_{i \epsilon F} \log S(\tau_i;\theta) \tag{6.2}$$

wherein we estimate θ by maximizing $\widetilde{\ell}\,(\theta)$, has been shown to be very effective (Suzuki 1985ab, 1987, Kalbfleisch and Lawless 1988, Hu and Lawless 1996a). Strictly speaking (6.2) requires knowledge of the exact value of N, the number of units in service by the current date. However, this method performs well if N is known reasonably accurately.

The second approach is to estimate the distribution of the censoring times $\tau_1, \cdots, \tau_N$ for the population of units in service. The motivation for this is to notice that if $\tau_{n+1}, \cdots, \tau_N$ are not observed then the observed likelihood function is, instead of (6.1), proportional to

$$L_1(\theta) = \prod_{i=1}^{n} f\,(t_i;\theta) \prod_{i=n+1}^{N} \Pr\ (T_i > \tau_i)\,, \tag{6.3}$$

where both T_i and τ_i are random variables. If T_i and τ_i are independent and τ_i has distribution function $G_i(\tau)$ then

$$\Pr\,(T_i > \tau_i) = \int_0^\infty S\,(\tau;\theta)\, dG_i(\tau).$$

We allow $G_i(\tau)$ to vary across $i = 1, \cdots, N$ because there may be observable covariates that are informative about τ_i. For example, the date of manufacture affects the date of sale of a unit, and hence τ_i; for automobiles, the date of sale conveys information about τ_i in the case where the "time" scale for failure is mileage.

If N and the $G_i(\tau)$'s are known, we can estimate θ by maximizing (6.3). Usually, however, the $G_i(\tau)$'s are estimated, or known only approximately, and in that case we should allow confidence intervals to reflect this. Hu and Lawless (1997) provide methods for parametric models $f(t;\theta)$ and Hu, Lawless and Suzuki (1997) consider nonparametric estimation. The latter authors note that in the discrete time case where $t = 0, 1, 2, \cdots$, and where τ_i and T_i are independent,

$$\widetilde{f}\,(t) = \frac{d(t)}{\sum_{i=1}^{N} \overline{G}_i(t)} \tag{6.4}$$

estimates $f(t) = \Pr(T = t)$, where $d(t)$ is the total number of failures observed at time t, and $\overline{G}_i(t) = \Pr(\tau_i > t)$. This gives the simple estimate $\tilde{F}(t) = \tilde{f}(0) + \cdots + \tilde{f}(t)$.

The methods based on (6.2) extend easily to deal with covariates that are missing for unfailed units but which may be observed for the followup sample. Those based on (6.3) may also be extended, but this is slightly more complicated (Hu and Lawless 1997).

Example: Car Warranty Data

Lawless, Hu and Cao (1995) discuss an example involving warranty data on about 8400 automobiles in some detail, and we merely summarize the results. Figure 2 shows pointwise .95 confidence interval estimates of the survivor function $S(t) = \Pr(T > t)$, where the time scale t is mileage; since the warranty coverage was for 12 months or 12,000 miles, estimates up to only $t =$ 12,000 miles are shown. Three sets of intervals are portrayed: the "nonparametric" is based on (6.4) and estimates of $\overline{G}_i(t)$; the "parametric" is based on (6.3) and estimates of $G_i(t)$, in which a parametric Weibull distribution was used for T, with mileage accumulation rate as a covariate; the "truncated data only" used only the units with observed failures. Lawless et al. (1995) found that failure time (with time = mileage) is more or less independent of the mileage accumulation rate, and Figure 2 shows that the parametric and nonparametric estimates are in good agreement. It also indicates that the truncated failure data on their own are very uninformative, as discussed above.

6.2 Multiple or Recurrent Event Processes

Product units may experience several claims, or we may wish to break claims into groups or types. In this case we consider $n_i(t)$, the number of events of some specific type for unit i at "time" t. As in Section 6.1, "time" may be age or some measure of cumulative usage. For simplicity we discuss one type of event and take time to be discrete.

For engineering purposes the process $\{n_i(t), t = 0, 1, 2, \cdots\}$ is of interest. Data consist of the times at which events occur or, equivalently, the values of $n_i(t)$ for $t = 0, 1, 2, \cdots$, and possibly some covariates $\boldsymbol{x}_i$. As in Section 6.1, τ_i is the censoring time for unit i, and unless a unit has one or more events before time τ_i, values in $\boldsymbol{x}_i$ and τ_i itself are usually unknown. One approach is to specify a model for the process $\{n_i(t), t \geq 0\}$, for example, that the $n_i(t)$'s are independent Poisson random variables with means $\lambda_i(t;\boldsymbol{\theta})$; then we could estimate $\boldsymbol{\theta}$ through pseudo likelihood approaches analogous to those giving (6.2) or (6.3). Because of the considerable heterogeneity in the usage of products, and their environments, it may be necessary to incorporate covariates or random effects in the models. Model checking is hampered by the typically sparse data and the need to estimate censoring times (and possibly covariate distributions) from supplementary data.

An alternative approach is to model and estimate $E\{n_i(t)\} = \lambda_i(t;\boldsymbol{\theta})$ without further strong assumptions as to the nature of the process $\{n_i(t), t \geq 0\}$. This leads to nonparametric estimators similar to (6.4) in the case where $\lambda_i(t;\theta) = \lambda(t)$:

$$\tilde{\lambda}(t) = \frac{d(t)}{\sum_{i=1}^{N} \overline{G}_i(t)}, \tag{6.5}$$

provided that τ_i and the event process are independent and that estimates of the survivor functions $\overline{G}_i(t)$ for τ_i are available. This has been studied by Hu and Lawless (1996b), who present an example involving the car warranty data of Section 6.1. The values $\tilde{\lambda}(t)$ may be smoothed or used as they are to provide estimates $\tilde{\Lambda}(t) = \sum_{u=0}^{t} \tilde{\lambda}(u)$ of the expected number of events up to time t. Parametric models and extensions to deal with covariates may be based on the estimating function approach of Lawless and Nadeau (1995) and the fact that $E\{n_i^*(t)\} = \lambda_i(t;\boldsymbol{\theta})\overline{G}_i(t)$, where $n_i^*(t) = n_i(t)I(\tau_i > t)$.

Further study of methods for recurrent events would be valuable, including the assessment of heterogeneity across units. It is of considerable interest, for example, whether a small fraction of units generate most of the events. In the absence of observable explanatory variables, mixture models based on unobservable random effects may be considered.

7. Concluding Remarks

The methods of Sections 2-6 deal with warranty data in a wide variety of situations. We conclude with some comments on enhancements and additional problems.

The warranty schemes for a product may render reliability estimation and analysis like that in Section 6 more difficult in some cases. For example, if warranty limits involve usage, then assumptions made in Section 6 about censoring times being independent of failures may be violated. If failure rates for automobiles depend on the mileage accumulation rate, then censoring times on either the "age" or "mileage" time scale are not strictly independent of failure times. Lawless et al. (1995) and Hu and Lawless (1996b) tackle this problem by treating usage factors as covariates, conditional on which failure and censoring times may be assumed independent. Since for field reliability analysis usage is generally an important factor, models both for the dependence of failure on usage and for variations in usage across the field population of product units are important. Further study of this area would be valuable.

For certain types of products, other methods for analyzing repeated claims than those in Section 6.2 may sometimes be useful. For example, if a claim results in a product or component being replaced with a new one, renewal process models are sometimes used. The case of so-called "free replacement warranties" has received a good deal of attention; see Frees (1986, 1988) and various chapters of Blischke and Murthy (1996), which discusses many types of warranties.

Monitoring warranty claims is an ever-present issue for manufacturers. The methods of age-based analysis provide tools for doing this, but procedures for when to take action in the face of apparently elevated claims require study. Multiplicity problems due to large numbers of product lines or product components are not easily dealt with.

Interactions or relationships between different types of claims are sometimes of interest; for example, uncertainty about the source of a problem may lead to the replacement of two or more components when only one is truly involved. The methods in this paper deal with models for individual types of claims. It is possible to extend them in a reasonably straightforward way to deal simultaneously with multiple claim types (e.g. Hu et al. 1997). Chukova and Dimitrov (1996) discuss modelling of complex product failures.

Finally, although our focus has been on the analysis of warranty data, many of the issues discussed arise more generally for field failure or field return data, and the methods presented are of quite broad applicability. Robinson and McDonald (1991), Lawless and Kalbfleisch (1992) and Suzuki (1995) survey field reliability analysis.

8. References

Blischke, W.R. and Murthy, D.N.P. (1996). *Product Warranty Handbook* (editors). New York: Marcel Dekker

Chen, J., Lynn, N.J. and Singpurwalla, N.D. (1996). Forecasting Warranty Claims, Chapter 31 in *Product Warranty Handbook*, eds. W.R. Blischke and D.N.P. Murthy, New York: Marcel Dekker.

Chukova, S. and Dimitrov, B. (1996). Warranty Analysis for Complex Systems, Chapter 22 in *Product Warranty Handbook*, eds. W.R. Blischke and D.N.P. Murthy, New York: Marcel Dekker.

Dean, C.B. (1992). Estimating equations for mixed Poisson models. In *Estimating Functions*, pp. 35-46. Editor, V.P. Godambe. Oxford University Press.

Frees, E.W. (1986). Warranty Analysis and Renewal Function Estimation. *Naval Res. Logist. Quart. 33*, 361-372.

Frees, E.W. (1988). Estimating the Cost of a Warranty. *J. Bus. Econ. Statist. 6*, 79-86.

Hu, X.J. and Lawless, J.F. (1996a). "Estimation from Truncated Lifetime Data with Supplementary Information on Covariates and Censoring Times". *Biometrika, 83*, 747-761.

Hu, X.J. and Lawless, J.F. (1997). Pseudolikelihood Estimation in a Class of Problems With Response-Related Missing Covariates. *Canadian Journal of Statistics, 25,* 125-142.

Hu, X.J. and Lawless, J.F. (1996b). "Estimation of Rate and Mean Functions from Truncated Recurrent Event Data". *J. Amer. Statist. Assoc., 91*, 300-310.

Hu, X.J., Lawless, J.F. and Suzuki, K. (1997). "Nonparametric Estimation of a Lifetime Distribution when Censoring Times are Missing". To appear in *Technometrics.*

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. (1996). "Statistical Analysis of Warranty Claims Data". Chapter 9 in *Product Warranty Handbook*, eds. W.R. Blischke and D.N.P. Murthy, New York: Marcel Dekker.

Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). "Methods for the Analysis and Prediction of Warranty Claims". *Technometrics, 33*, 273-285.

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

Lawless, J.F. (1987). Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics, 15,* 209-225.

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*, eds. J.P. Klein and P.K. Goel, Dordrecht: Kluwer, pp. 141-152.

Lawless, J.F. and Nadeau, J.C. (1995). "Some Simple Robust Methods for the Analysis of Recurrent Events", *Technometrics 37*, 158-168.

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Robinson, J.A. and McDonald, G.C. (1991). "Issues Related to Field Reliability and Warranty Data", in *Data Quality Control: Theory and Pragmatics.* Eds. G.E. Liepins and V.R.R. Uppuluri, New York: Marcel Dekker.

Suzuki, K. (1985a). "Estimation Method of Lifetime Based on the Record of Failures During the Warranty Period". *Journal of the American Statistical Association, 80*, 66-72.

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

Suzuki, K. (1987). "Analysis of Field Failure Data from a Nonhomogeneous Poisson Process". *Reports on Statistical Applications in Research, 34*, 6-15.

Suzuki, K. (1995). "Role of Field Performance Data and its Analysis", in *Recent Advances in Life-Testing and Reliability*, N. Balakrishnan , ed., Boca Raton: CRC Press.

Taylor, G. (1986). *Claims Reserving in Non-Life Insurance.* Amsterdam: North Holland.

Acknowledgements

This research was supported in part by grants from the Manufacturing Research Corporation of Ontario, the Natural Sciences and Engineering Research Council of Canada, and General Motors of Canada. The author thanks Diane Gibbons, Joan Hu, Jack Kalbfleisch, Jock MacKay, Claude Nadeau, Jeff Robinson and Lara Wolfson for assistance and comments on various aspects of the work.

Legends for Figures

Figure 1. Age-based cumulative automobile warranty claims.

Figure 2. Confidence intervals for survival function of miles to first warranty claim.

Figure 1

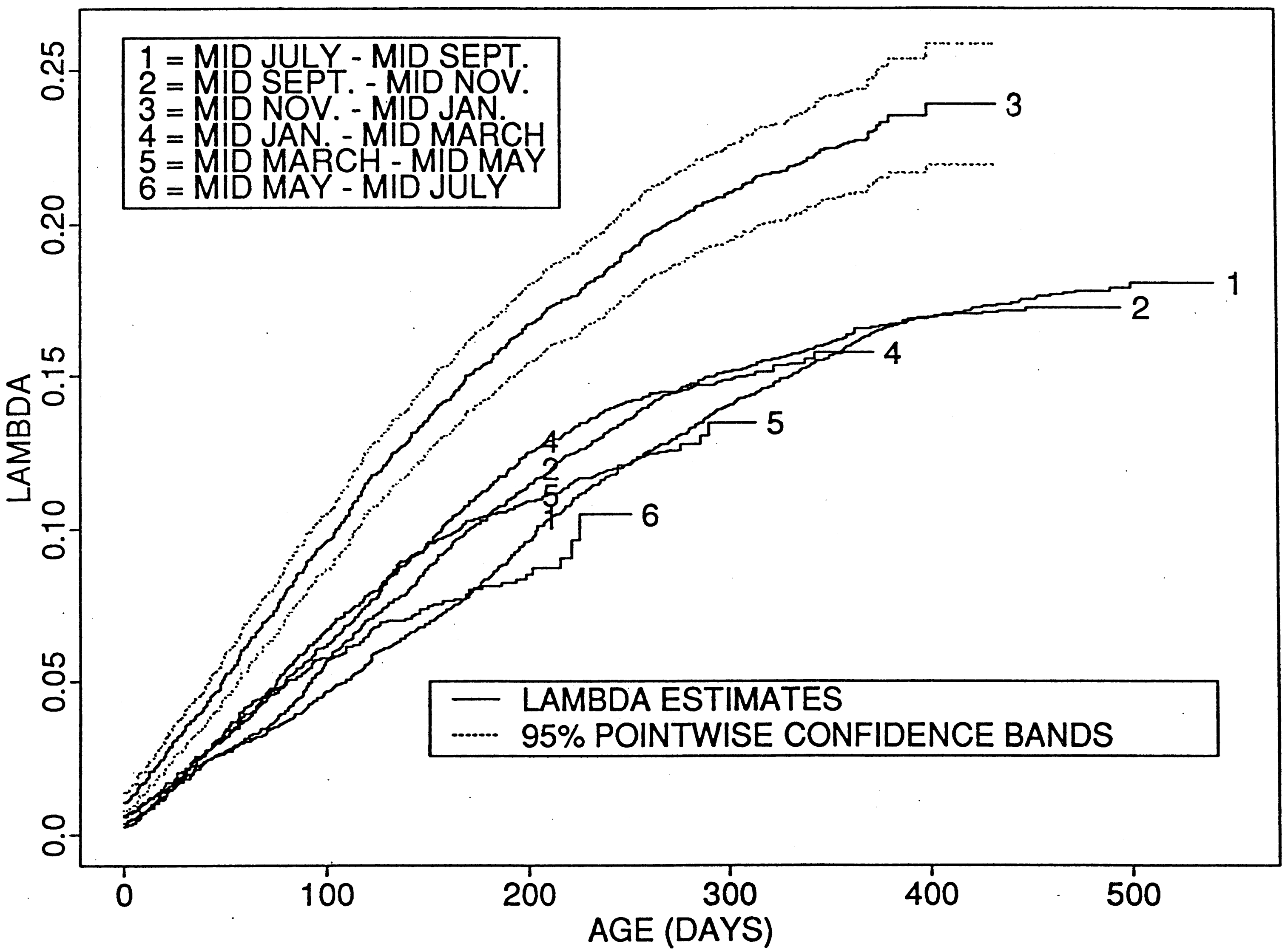
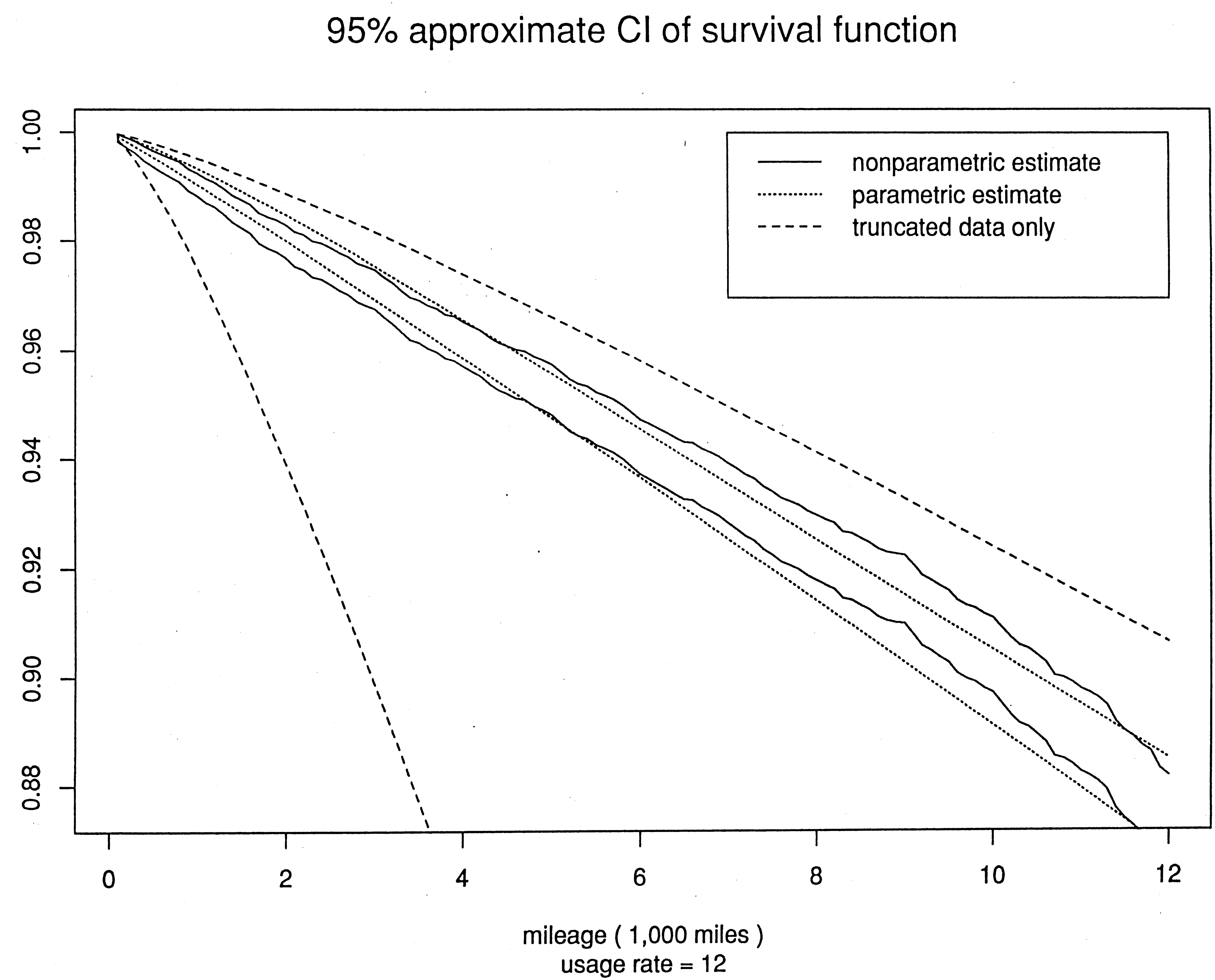

1 = MID JULY - MID SEPT.
2 = MID SEPT. - MID NOV.
3 = MID NOV. - MID JAN.
4 = MID JAN. - MID MARCH
5 = MID MARCH - MID MAY
6 = MID MAY - MID JULY
LAMBDA ESTIMATES
95% POINTWISE CONFIDENCE BANDS
LAMBDA
AGE (DAYS)
0.0
0.05
0.10
0.15
0.20
0.25
0
100
200
300
400
500
1
2
3
4
5
6

figure 2



95% approximate CI of survival function
nonparametric estimate
parametric estimate
truncated data only
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
2
4
6
8
10
12
mileage (1,000 miles)
usage rate = 12