METHODS FOR THE ANALYSIS AND PREDICTION OF WARRANTY CLAIMS

J.D. Kalbfleisch
J.F. Lawless
and
J.A. Robinson

IIQP Research Report
RR-90-01

January 1990

METHODS FOR THE ANALYSIS AND PREDICTION OF WARRANTY CLAIMS

J.D. Kalbfleisch and J.F. Lawless

Department of Statistics and Actuarial Science
University of Waterloo

J.A. Robinson

General Motors Research Laboratories

ABSTRACT

This article discusses methods whereby reports of warranty claims can be used to estimate the expected number of warranty repairs per unit in service as a function of the time in service. These estimates are adjusted for delays or lags corresponding to the time from the warranty repair until it is entered into the data base used for analysis. Forecasts of the number and cost of warranty repairs on the collection of all units in service are also developed along with standard errors for these forecasts. The methods are based on a log linear Poisson model for numbers of warranty claims. Both the case of a known distribution of reporting lag and simultaneous estimation of that distribution are considered. The use of residuals for model checking, extensions to allow for extra Poisson variation, and the estimation of warranty costs are also considered.

Key words and phrases: warranty data, reporting lags, log linear Poisson models, field reliability, prediction.

1. Introduction

With products under warranty, manufacturers often collect detailed claims data. In this article, we discuss use of these data for the prediction of future warranty claims and the making of comparisons of claim rates and costs for different product lines, different components of a product, units manufactured at different times, and so on. One practical problem which often arises and which we address is the presence of reporting delays between the time a claim occurs and the time that it is entered into the data base used for prediction and analysis.

For convenience, we consider the product to be automobiles. The methods we present, however, are widely applicable and also have use in the analysis of repair data in organizations which operate fleets of a particular type of unit.

The basic statistical model we consider is as follows: Suppose that cars enter service (are sold) on days $x : 0 \leq x \leq \tau^*$ and that once a car is in service, the number of repairs t days later (at age t) is assumed to be Poisson (λ_t), $(t = 0, 1, ...,$ independently). In most applications λ_t is small and can be interpreted as the probability of a repair at age t. Let f_l be the probability that the repair claim enters the data base used for analysis (is 'reported') l days after it takes place $(l = 0, 1, ...)$. Suppose that N_x cars are put into service on day x and let n_{xtl} be the number of repairs at age t and with a report lag l, for cars put into service on day x. The distribution of n_{xtl} is

$$n_{xtl} \sim \text{Poisson}(\mu_{xtl}) \tag{1.1}$$

where $\mu_{xtl} = N_x \lambda_t f_l$. It is convenient also to define $\Lambda_t = \sum_{u=0}^{t} \lambda_u$, the expected number of repairs for a car up to and including age t.

Individual cars almost certainly have varying repair rates, but the counts n_{xtl} obtained from the superposition of the repair processes of many cars should be close to Poisson when repair rates are small. Extensions that allow for extra-Poisson variation are discussed in Section 4. The parameter λ_t in (1.1) can be interpreted as the marginal or average rate of

repair at age t. Note that λ_t is assumed to be independent of when the car was manufactured or put into service and f_l is assumed independent of when the repair occurred. These assumptions can be checked and modified if necessary; we return to this point in Sections 4 and 7.

Our primary objective is the prediction of the (eventual) average number of repairs at age t, or up to age t, for cars put into service over the period $0, 1, ..., \tau$. These are respectively

$$m(t) = \frac{\sum_{x=0}^{\tau} \sum_{l=0}^{\infty} n_{xtl}}{\sum_{x=0}^{\tau} N_x} \qquad t = 0, 1, ... \tag{1.2}$$

and

$$M(t) = \sum_{u=0}^{t} m(u). \tag{1.3}$$

Note that these are simple measures of quality and that the total number of repairs to age t relates directly to cost. We suppose that data are available over the time period 0 to T. Thus, all claims reported by day T are included so that the counts n_{xtl} for x, t, l such that $0 \leq x + t + l \leq T$ are observed. Estimation of $m(t)$ and $M(t)$ is then a prediction problem; we predict those n_{xtl}'s in (1.1) and (1.2) that are not yet observed. Figure 1 portrays the situation when there are no reporting lags (i.e. $f_0 = 1$), and Figure 2 the general situation.

The prediction problem for warranty repairs is considered in Section 2. Section 3 introduces alterations for grouped data. Section 4 outlines methods of model checking and more refined statistical analysis of the claims data. Section 5 deals with the prediction and analysis of warranty claim costs, as opposed to the number of claims. Section 6 contains examples and Section 7 concludes the paper with a discussion of additional problems.

The methods presented here are similar in spirit, although different in detail, than methods used for other problems where reporting lags are important. For examples involving the reporting of AIDS cases see Kalbfleisch and Lawless (1989abc) and for examples in insurance, Kaminsky (1987). Simple Poisson and mixed Poisson models have been used often in reliability problems (e.g. Ascher and Feingold, 1978, Lawless, 1987, Nelson, 1988) but the use of more complex Poisson models, as in the present context, is new.

2. Estimating Repair Rates and the Number of Warranty Claims

2.1 Reporting Lag Probabilities Known

To begin, suppose that the probabilities f_l are known and let $F_l = f_0 + ... + f_l$. Throughout it is also supposed that N_x, the number of cars entering service on day x, is known for $x = 0, ..., \min(\tau, T)$ where T is the current date. The data comprise the frequencies n_{xtl}, where $x + t + l \leq T$ and $x \leq \tau$, and give rise to the likelihood

$$L = \prod_{\substack{x+t+l \leq T \\ x \leq \tau}} \prod \prod e^{-N_x \lambda_t f_l} (N_x \lambda_t f_l)^{n_{xtl}} / n_{xtl}! \ . \tag{2.1}$$

The maximum likelihood estimates obtained from (2.1) are

$$\hat{\lambda}_t = n_{.t.} / R_{T-t} \qquad t = 0, ..., T \tag{2.2}$$

where

$$n_{.t.} = \sum_{\substack{x+l \leq T-t \\ x \leq \tau}} \sum n_{xtl} \tag{2.3}$$

is the total number of repairs which have been observed on cars of age t days, and

$$R_{T-t} = \sum_{x=0}^{\min(\tau, T-t)} N_x F_{T-t-x} \tag{2.4}$$

is an adjusted count of the number at risk at day t. Note that N_x, the number of cars entering service on day x, is adjusted by a multiplicative factor which is the probability that, for a car in this group, a repair at age t would be reported by time T.

The corresponding estimates of $m(t)$ and $M(t)$ are

$$\hat{m}(t) = \hat{\lambda}_t, \quad \hat{M}(t) = \sum_{u=0}^{t} \hat{\lambda}_u = \hat{\Lambda}_t \tag{2.5}$$

and to obtain prediction limits, we consider the variation in $m(t) - \hat{m}(t)$ or $M(t) - \hat{M}(t)$. Let $N = \sum_0^\tau N_x$ be the total number of cars entering service by day τ. We presume that N is known. It then follows easily that

$$E\{m(t) - \hat{m}(t)\} = 0$$

and, in the Appendix, it is shown that

$$var\{m(t) - \hat{m}(t)\} = \left(\frac{N - R_{T-t}}{NR_{T-t}}\right)\lambda_t. \tag{2.6}$$

It follows immediately that $E\{M(t) - \hat{M}(t)\} = 0$ and

$$var\{M(t) - \hat{M}(t)\} = \sum_{u=0}^{t}\left(\frac{N - R_{T-u}}{NR_{T-u}}\right)\lambda_u. \tag{2.7}$$

Variance estimates are obtained by replacing λ_u with $\hat{\lambda}_u$ in (2.7) and a normal approximation provides approximate confidence intervals for $M(t)$. For large samples, this approximation is very accurate; in small sample problems, more accurate approximations could be derived.
REMARKS:

1. If there is no reporting lag or if it is ignored, then estimates are given by the formulas above with all of the F_l's ($l = 0, 1, ...$) equal to one. In this case, R_{T-t} is the total number of cars in service that have an age of at least t at time T. If the reporting lag is ignored when there is a delay in reporting, the estimates of λ_t (see (2.2)) are biased downwards, as are predictions of claims. This is particularly serious for early predictions; see Section 6.

2. Note from (2.4) that if $T - t \geq \tau$ and $F_l = 1$ for $l \geq T - t - \tau$ then $N = R_{T-t}$ and, by (2.6), $var\{m(t) - \hat{m}(t)\} = 0$. In this case, $m(t) = \hat{m}(t)$ is fully known.

3. If $T < \tau$ then to estimate the variances (2.6) or (2.7) we have to estimate $N_{T+1} + ... + N_\tau$, i.e. we need to know the number of cars entering service up to day τ. To avoid this, one can always select $\tau \leq T$, i.e. estimate eventual claims only for cars having entered service by the current date.

4. Estimation of the eventual number of claims $m(t)$ or $M(t)$ is a finite-sample prediction problem for Poisson random variables. In the numerator of (1.2), the n_{xtl}'s with $x + t + l \leq T$ are observed by the current time T and the remaining n_{xtl}'s are future values to be predicted.

5. Confidence levels associated with prediction intervals refer to repeated occurrences of the entire process. As a consequence, sequences of predictions which are made as new data are reported are not independent. This is a well known problem with multiple prediction statements but does not materially affect their usefulness.

6. We have assumed that there is no reporting lag associated with the N_x's. If there are small reporting lags then a simple adjustment is to scale up the N_x values reported close to the current date T using estimates of the lag probabilities. Another approach is to eliminate from the set of repairs considered any that are on cars not yet reported as in service. If this is done, and the delay in reporting cars entering service is not related to the failure rates, then estimation of $\hat{\lambda}_t$ would still be valid.

7. If the warranty involves both a mileage and age limit, $\hat{\lambda}_t$ will be an underestimate of the true repair rate at age t. This is because some of the cars will have exceeded the mileage limit; they are still counted in the denominator of $\hat{\lambda}_t$ but cannot contribute to the numerator. The bias will be small for small t but increase as t increases. It should be noted, however, that $\hat{\lambda}_t$ is a valid estimate of the probability of a warranty claim at age t and so predictions of eventual warranty claims are still valid. Similarly, random events that result in the temporary or permanent withdrawal of cars from service do not affect the validity of warranty claim estimates.

8. The formula (2.2) for $\hat{\lambda}_t$ adjusts for the reporting lag by discounting the number at risk for an age t repair using the known distribution of lags as in (2.4). It might be argued that the need for adjustment when reporting lags are present arises not from an error in the nominal number at risk $r_{T-t} = \sum_{x=0}^{\min(\tau, T-t)} N_x$, but rather from an

under-reporting in the number of repairs at age t. In communicating the adjustment to nonstatisticians, therefore, there may be advantage to rewriting the estimate as

$$\hat{\lambda}_t = \hat{n}_{.t.}/r_{T-t} \tag{2.8}$$

where $\hat{n}_{.t.}$ is the estimated number of age t repairs (both reported and unreported) that have occurred prior to time T. It is clear that

$$\hat{n}_{.t.} = \frac{r_{T-t}}{R_{T-t}} n_{.t.} \quad . \tag{2.9}$$

The estimate (2.8) could be obtained directly by use of an E-M algorithm as discussed in Dempster et al. (1977). At the E step we calculate

$$E\left\{\sum_{l=0}^{\infty} n_{xtl} | n_{xtl}, x+t+l \leq T; \lambda_t^{(0)}\right\} = n_{xt.} + \sum_{l=T-x-t+1}^{\infty} N_x \lambda_t^{(0)} f_l \tag{2.10}$$

and, at the M step, the updated estimates are obtained as

$$\lambda_t^{(1)} = \left\{ n_{.t.} + \sum_{x=0}^{\min(\tau, T-t)} N_x \lambda_t^{(0)} (1 - F_{T-x-t}) \right\} / r_{T-t}. \tag{2.11}$$

It is easily seen that $\hat{\lambda}_t$ is a fixed point of (2.11) and (2.10) converges to (2.9).

2.2 Concurrent Estimation of the Reporting Lag Distribution

It is possible to estimate simultaneously the reporting lag distribution $\{f_l\}$. For this purpose, we maximize (2.1) jointly with respect to $\{\lambda_t\}$ and $\{f_l\}$. The likelihood equations are

$$\frac{\partial \log L}{\partial \lambda_t} = \frac{n_{.t.}}{\lambda_t} - \sum\sum_{\substack{x+l \leq T-t \\ x \leq \tau}} N_x f_l = 0 \tag{2.12}$$

$$\frac{\partial \log L}{\partial f_l} = \frac{n_{..l}}{f_l} - \sum\sum_{\substack{x+t \leq T-l \\ x \leq \tau}} N_x \lambda_t = 0 \tag{2.13}$$

where $n_{.t.}$ is as defined earlier, and $n_{..l}$ is the number of observed lags of duration l days.

From the form of the likelihood (2.1) it is clear that $\{\lambda_t\}$ and $\{f_l\}$ can be estimated only up to a constant of proportionality c since $\{c\lambda_t\}$, $\{c^{-1}f_l\}$ have the same likelihood as $\{\lambda_t\}$, $\{f_l\}$. Kalbfleisch and Lawless (1989a) discuss this point in a similar problem arising in a different context. To obtain a unique solution to (2.12) and (2.13) we assume that $F_T = \sum_{l=0}^{T} f_l = 1$. In most applications, T is large enough that this is a reasonable assumption; if T were not large, an alternative approach would be to assume that F_T were known.

It is straightforward to solve (2.12) and (2.13) subject to the constraint $F_T = 1$ using a fixed point algorithm or a Newton Raphson procedure. However, simpler methods are available to estimate the reporting lag distribution. Let $h_l = f_l/F_l = Pr\{L = l|L \leq l\}$ where L represents the reporting lag. Then h_l is estimated by

$$\tilde{h}_l = n_{..l} / \sum_{l'+s \leq T-l} n^*_{sl'}, \tag{2.14}$$

where $n^*_{sl} = \sum_{x=0}^{\min(\tau, T-l)} n_{x,s-x,l}$ is the number of observed repairs on day s with lag l. The denominator of (2.14) is thus the number of repairs that occur on or before day $T-l$ with a reporting lag of l days or less. The estimate of f_l then is

$$\tilde{f}_l = \tilde{h}_l \prod_{j=l+1}^{T} (1 - \tilde{h}_j), \quad l = 0, ..., T. \tag{2.15}$$

The estimates (2.14) and (2.15) arise as maximum likelihood estimates under a likelihood formed by using data on the truncated distributions of reporting lags. Thus, a repair on day s with report on day $s+l$ contributes the term f_l/F_{T-s} to the likelihood. Such likelihoods have been extensively studied by several authors. See for example Woodroffe (1985), Keiding and Gill (1987), Lagakos et al. (1988) and Kalbfleisch and Lawless (1989ac).

Estimates of the reporting lag can also be restricted to use only a part of the data. In certain instances, for example, one may wish to use only the most recent data in estimating f_l. Restricting the sums for the numerator and denominator of (2.14) to include only frequencies n_{xtl} with $x+t \geq T-a$ would restrict the estimates of f_l to depend only on the repairs observed

within the most recent a days. It is also possible to develop tests for trends in the reporting lag distribution or to adjust, for example, for seasonal effects. These refinements are not considered further here, but see Kalbfleisch and Lawless (1989c).

Variance estimation of $\tilde{f}_l$ can be accomplished using a variant on Greenwood's formula for the life table or Kaplan-Meier estimate; see Lagakos et al. (1988) for details. Variance estimates in (2.6) and (2.7) could be adjusted for uncertainty in the estimation of f_l. For practical purposes, however, the variation in $M(t) - \hat{M}(t)$ can be estimated using (2.7) with f_l replaced by $\tilde{f}_l$. This is a slight underestimate since it does not account for uncertainty in the estimation of f_l.

3. Grouped Data

Sometimes the data are grouped so that only the numbers of repairs for cars with age t lying in various intervals are observed. For example, we may know the total number of claims for cars aged 0-30 days, 31-60 days, etc. For the following discussion we again assume that the reporting lags distribution $\{f_l\}$ is known.

Consider some age interval $t = a$ to b inclusive and consider estimation of $M(a,b) = \sum_{t=a}^{b} \lambda_t$, the average number of repairs per car for the age interval. Using (2.2), we estimate this by

$$\sum_{t=a}^{b} \hat{\lambda}_t = \sum_{t=a}^{b} \frac{n_{.t.}}{R_{T-t}}. \tag{3.1}$$

However, if we observe only $\sum_{t=a}^{b} n_{.t.}$ then (3.1) has to be approximated. A simple approach is to estimate $M(a,b)$ by

$$\tilde{M}(a,b) = \frac{\sum_{t=a}^{b} n_{.t.}}{\tilde{R}(a,b)} \tag{3.2}$$

where $\tilde{R}(a,b)$ is an estimate of the car-days in service for cars aged a to b. Reasonable estimates are

$$\tilde{R}(a,b) = \frac{1}{2}(R_{T-a} + R_{T-b}) \tag{3.3}$$

or

$$\tilde{R}(a,b) = \frac{1}{b-a+1}\sum_{u=a}^{b} R_{T-u}. \tag{3.4}$$

To motivate (3.3) and (3.4), note that by (2.3) and (1.1),

$$E\{\tilde{M}(a,b)\} = \frac{\sum_{t=a}^{b}\lambda_t R_{T-t}}{\tilde{R}(a,b)}. \tag{3.5}$$

If λ_t is constant for $t = a$ to $t = b$, then $M(a,b) = (b-a+1)\lambda_t$ and (3.2) is an unbiased estimate if $\tilde{R}(a,b)$ is given by (3.4); it is also the maximum likelihood estimate under (1.1). If R_{T-t} is linear over $t = a$ to b then (3.4) reduces to (3.3).

If the intervals (a,b) are not too long then λ_t can usually be taken to be constant over the interval as a reasonable approximation, and this leads to (3.4). Note, however, that although the data on the numbers of repairs are allowed to be grouped, daily counts N_x of the number of cars are needed to compute the R_{T-t}'s. If the counts of cars entering service are also grouped, for example monthly, a further adjustment can be made. Suppose that $N(a,b) = \sum_{x=a}^{b} N_x$ is known but not the individual N_x's. If we assume the N_x's are constant over (a,b) then the contribution to R_{T-t} (see (2.4)) from cars entering service in (a,b) is

$$\sum_{x=a}^{b} N_x F_{T-x-t} = N(a,b)\bar{F}(a,b) \tag{3.6}$$

where $\bar{F}(a,b) = (\sum_{x=a}^{b} F_{T-x-t})/(b-a+1)$. In practice, time periods of the same length, say k days, would usually be used to report both number of repairs and new cars entering service, and T would also be a multiple of k. If there are no reporting lags, simple expressions can be given for $\tilde{M}(a,b)$.

Variance estimates for $M(t) - \hat{M}(t)$ and prediction limits for $M(t)$ may be obtained by using (2.7) with λ_u estimated by the average $\hat{\lambda}_t$ for the time interval in which u lies and, if the N_l's are also grouped, by assuming in the calculation of R_{T-u} values in (2.7) that N_x is constant over time intervals. An example is given in Section 6.

If the f_l's are estimated concurrently with the λ_t's, matters are more complicated although methods based on the Poisson model could be developed.

4. Model Checks and Some Extensions

The models employed here are Poisson log linear models and checks on fit can be made in familiar ways. Pearson residuals

$$r_{xtl} = (n_{xtl} - \hat{\mu}_{xtl})/\hat{\mu}_{xtl}^{1/2}, \tag{4.1}$$

where $\hat{\mu}_{xtl} = N_x\hat{\lambda}_t f_l$ when the f_l's are known, and $\hat{\mu}_{xtl} = N_x\hat{\lambda}_t\hat{f}_l$ when they are estimated, may be used to check on (1.1). Provided expected frequencies $\hat{\mu}_{xtl}$ are not too small, the Pearson statistic $P = \sum_x \sum_t \sum_l r_{xtl}^2$ is approximately χ^2 with degrees of freedom equal to the number of distinct counts n_{xtl} minus the number of parameters estimated; the number of distinct counts is readily found to be $\frac{1}{6}[(T+1)(T+2)(T+3) - x_T(x_T+1)(x_T+2)]$, where $x_T = \max(0, T-\tau)$.

Systematic departures from the fitted model will often be detectable by scrutinizing the residuals (4.1) or related residuals obtained by grouping the data. The assumed independence of λ_t from x can be examined using the residuals based on $n_{xt.} = \sum_{l=0}^{T-t-x} n_{xtl}$,

$$r_{xt} = (n_{xt.} - \hat{\mu}_{xt.})/\hat{\mu}_{xt.}^{1/2}. \tag{4.2}$$

For example, cumulative sums over t of the residuals (4.2) for fixed x, or of related residuals based on intervals of x values, can be used to detect variations in the λ_t's according to the time at which the cars entered service. It would also be useful to compare observed and expected counts for groups based on monthly intervals for each of x and t.

Formal tests of various aspects of the mean specification μ_{xtl} in (1.1) can be developed; the most straightforward approach is to use likelihood ratio tests based on (2.1). We leave a more thorough discussion of this and related problems to a future article. We remark also that in applications where counts are small, it is preferable to use deviance or Anscombe residuals rather than Pearson residuals. See Pierce and Schafer (1986) or McCullagh and Nelder (1989) for these residuals and further discussion.

Residuals (4.1) can also be used to assess the assumptions that the n_{xtl}'s are independent Poisson variates with $E(n_{xtl}) = var(n_{xtl}) = \mu_{xtl}$. In particular, plots of r_{xtl} versus $\hat{\mu}_{xtl}$

and normal probability plots of r_{xtl} provide graphical methods for detecting overdispersion, where $var(n_{xtl}) > \mu_{xtl}$. We discuss analyses in which extra Poisson variation is allowed.

It is well known that provided the mean specification is correct, the Poisson analyses above give consistent estimates of the unknown parameters (in this case, the claim rates λ_t) when there is extra Poisson variation, but the variance estimates under the Poisson are too small. Unobservable heterogeneity is one way in which extra Poisson variation can arise, and since one car can give rise to several claims, this would also give rise to correlated counts. Suppose, for example, that the ith car placed in service at time x has repair rate $\alpha_i\lambda_t$ where the α_i's are taken to be independent random variables with $E(\alpha_i) = 1$, $var(\alpha_i) = \delta$; note that λ_t retains its interpretation as the average age t repair rate. It is supposed that, conditional on α_i, the number of repairs for this car at age t and with reporting lag l is $n_{xtli} \sim \text{Poisson}(\alpha_i\lambda_t f_l)$ where $i = 1, ..., N_x$. The α_i's incorporate heterogeneity from various sources including variation in environment, manufacturing, user characteristics and so on; this is a standard method of incorporating extra Poisson variation (e.g. Lawless, 1987). Note that $n_{xtl} = \sum n_{xtli}$. Straightforward calculation shows that

$$E(n_{xtl}) = N_x\lambda_t f_l \tag{4.3}$$

$$var(n_{xtl}) = N_x\lambda_t f_l(1 + \delta\lambda_t f_l) \tag{4.4}$$

$$cov(n_{xtl}, n_{xsk}) = N_x\delta\lambda_s\lambda_t f_l f_k \tag{4.5}$$

and $cov(n_{xtl}, n_{ysk}) = 0$ for $x \neq y$ and all t, l, s and k. For simplicity, we consider the case in which the f_l's are known.

Estimation of the λ_t's can be accomplished by using generalized estimating equations (or quasi-likelihood equations) for the λ_t's that utilize (4.4) and (4.5). An alternative and only slightly less efficient approach utilizes the Poisson estimates already derived, but adjusts variance estimation using the structure in (4.4) and (4.5). Technically, the Poisson estimating equations (2.12) are still unbiased for zero in the mixture model and the estimates $\hat{\lambda}_t =$

$n_{.t.}/R_{T-t}$ are consistent. But using (4.4) and (4.5) we find that

$$cov(n_{.t.}, n_{.s.}) = \delta\lambda_s\lambda_t \sum_{x=0}^{\min(T-t,T-s,\tau)} N_x F_{T-t-x} F_{T-s-x} \quad s \neq t \tag{4.6}$$

$$var(n_{.t.}) = \lambda_t \sum_{x=0}^{\min(T-t,\tau)} N_x F_{T-x}(1 + \delta\lambda_t F_{T-x}). \tag{4.7}$$

It follows that $var(\hat{\lambda}_t) > \lambda_t$ and the $\hat{\lambda}_t$'s are correlated.

The parameter δ can be estimated from an auxiliary moment equation such as

$$\sum_x \sum_t \sum_l \frac{(n_{xtl} - \hat{\mu}_{xtl})^2}{\hat{\mu}_{xtl}(1 + \hat{\mu}_{xtl}\delta/N_x)} = d \tag{4.8}$$

where $d = \frac{1}{6}[(T+1)(T+2)(T+3) - x_T(x_{T+1})(x_T+2)]$ is the number of distinct terms in the triple sum, as noted earlier in this section, $\hat{\mu}_{xtl} = N_x\hat{\lambda}_t f_l$, and the denominators in the left hand side of (4.8) are based on (4.4).

Estimates of $var(\hat{\lambda}_t)$ are obtained by inserting estimates $\hat{\lambda}_t$ and $\hat{\delta}$ in (4.7). The problem of obtaining variance estimates for $M(t) - \hat{M}(t)$, in order to get prediction limits for $M(t)$, is rather more involved because of the correlations among n_{xtl}'s that have the same x value. The Appendix develops the exact variance of $M(t) - \hat{M}(t)$.

5. Costs of Warranty Claims

In many instances, the total cost of warranty claims is of interest. We assume here that claim costs are indexed by $c = 1, 2, ..., m$ and $r(c)$ is the cost of a claim in the cth group. The amount of grouping desirable will depend on the particular application, but note that actual dollar amounts could be used.

To investigate estimation of the costs, suppose that $\lambda_t^{(c)}$ is the expected number of claims of cost $r(c)$ for a car at age t and, in an obvious notation, suppose that $n_{xtl}^{(c)} \sim$ Poisson$(N_x\lambda_t^{(c)}f_l)$, independently for $x, t, l \geq 0$. Note that this assumes that the reporting lag distribution is independent of the cost of the claim, but that the distribution of claim

costs can be age dependent. It would be possible to allow f_l to be dependent on cost, but we shall look at this simple model.

Under the assumption that the f_l's are known, the maximum likelihood estimate of $\lambda_t^{(c)}$ is

$$\hat{\lambda}_t^{(c)} = n_{.t.}^{(c)} / R_{T-t} \tag{5.1}$$

where R_{T-t} is given by (2.4). Let $m^{(c)}(t)$ and $M^{(c)}(t)$ be the natural extensions of (1.2) and (1.3) representing the average numbers of claims of cost $r(c)$ at age t and up to age t for cars put in service over the period $0, 1, ..., \tau$. Clearly, $m^{(c)}(t)$ and $M^{(c)}(t)$ are estimated with $\hat{m}^{(c)}(t)$ and $\hat{M}^{(c)}(t)$ analogous to (2.5) with variance estimates obtained from (2.7). The average cost of all claims up to age t for cars put in service in $0, 1, ..., \tau$ is

$$K(t) = \sum_{c=1}^{m} r(c) M^{(c)}(t)$$

and it is easily seen that

$$var(\hat{K}(t) - K(t)) = \sum_{c=1}^{m} r^2(c) var(\hat{M}^{(c)}(t) - M^{(c)}(t))$$

where $var(\hat{M}^{(c)}(t) - M^{(c)}(t))$ is obtained from (2.7).

Separate estimation of rates as in (5.1) would also be relevant if c were to index the type rather than the cost of the repair. Results in this section can be extended to allow simultaneous estimation of $\{f_l\}$ or to accommodate extra Poisson variation, but these topics are not pursued here.

6. Examples

We begin with an example which illustrates what happens when reporting lags are ignored.

In the situation described in Section 2, failure to incorporate reporting lags would lead to estimates

$$\tilde{\lambda}_t = \frac{n_{.t.}}{\sum_{x=0}^{\min(T-t,\tau)} N_x}. \tag{6.1}$$

If, however, reporting lags are actually present as described in Section 2.1 then $E(n_{.t.}) = \lambda_t R_{T-t}$, where R_{T-t} is given by (2.4). Thus estimates of λ_t are biased:

$$E(\tilde{\lambda}_t) = A_{T-t}\lambda_t, \quad A_{T-t} = \frac{R_{T-t}}{\sum_{x=0}^{\min(T-t,\tau)} N_x}. \tag{6.2}$$

Estimates of Λ_t and $M(t)$ are subsequently also biased.

As an illustrative but realistic example, we consider a situation where units are introduced over a one year period and estimates of repair rates up to two years are desired. With time measured in days we have $\tau = 364$ and $t = 0, 1, ..., 729$. Consider a situation where the N_x's are equal and where the true rates are $\lambda_t = .002$ $(t = 0, 1, ..., 364)$ and $\lambda_t = .001$ $(t = 365, ..., 729)$. Suppose finally that reporting lags are distributed over 0 to 59 days with probabilities $f_l = 1/30$ for $l = 20, ..., 39$ days and $f_l = 1/120$ for $l = 0, ..., 19$ and $40, ..., 59$ days.

Figure 3 shows expected values for estimates $\tilde{\Lambda}_t = \sum_{u=0}^{t} \tilde{\lambda}_u$ based on (6.1), for data that accrue for each of $T = 3, 6, 12$ and 24 months. The uppermost curve shows the true values of Λ_t for t going from 0 to 729 days. As the other curves indicate, when the reporting lag is ignored the cumulative claim rate tends to be underestimated. The underestimation is severe when T is smaller, especially for values of t close to T. In other words, failure to adjust for the reporting lag leads to severe underestimation of cumulative repair rates, especially from data for the first six to nine months or so after units first begin to enter service.

Incorporation of reporting lag probabilities corrects the problem just seen. For illustration we consider a second artificial example that is, however, broadly realistic at least for automobiles. We generated data for the situation described above, with $N_x = 100$ cars introduced into service on days $x = 0, 1, ..., 364$. With the f_l's assumed known, estimates of λ_t were computed, based on the data available up to various times T. The first three columns of Table 1 show some numerical results for $T = 364$ days (12 months). Figure 4 portrays the estimates $\hat{\Lambda}_t = \hat{M}(t)$ and 95% confidence limits for $M(t)$ of the form $\hat{M}(t) \pm 1.96\hat{V}(t)^{1/2}$, where $\hat{V}(t)$ is the variance estimate obtained when the $\hat{\lambda}_u$'s are inserted into (2.7).

If the $n_{.t.}$'s are grouped, we have to use (3.2). The last five columns of Table 1 show some

calculations for the case where the $n_{.t.}$'s are grouped into totals corresponding to 30 or 31 day age intervals. The estimates $\tilde{\Lambda}_t$ obtained from the grouped data are very close to those from the ungrouped data ($\hat{\Lambda}_t$).

The results given here assume knowledge of the reporting lag probabilities. If these are estimated from the data at hand, the estimates of Λ_t change very little. A more serious point in practice is that the reporting lag distribution may not be stationary, and we hope in a future communication to be able to illustrate these and other problems for some real automobile warranty data, using the methods discussed in Section 4.

Table 1. Estimation of Λ_t's at $T = 364$ Days

t	R_{T-t}	$\hat{\Lambda}_t$	(a,b)	$\tilde{R}(a,b)^1$	$\sum_a^b n_{.t.}$	$\tilde{M}(a,b)$	$\tilde{\Lambda}_t{}^2$
0	33,550	.0024					
30	30,550	.0643	0-30	32,050	2064	.0644	.0644
60	27,550	.1239	31-60	29,000	1726	.0595	.1239
90	24,550	.1850	61-90	26,000	1586	.0610	.1849
121	21,450	.2468	91-121	22,950	1416	.0617	.2466
151	18,450	.3036	122-151	19,900	1132	.0569	.3035
181	15,450	.3661	152-181	16,900	1053	.0623	.3658
211	12,450	.4281	182-211	13,900	870	.0626	.4284
242	9450	.4882	212-242	10,850	648	.0597	.4881
272	6352	.5479	243-272	7,800	470	.0603	.5484
303	3250	.6099	273-303	4,750	294	.0619	.6103
333	635	.6836	304-333	1808.5	136	.0752	.6855
364	.83	.7498	334-364	169.73	10	.0589	.7444

[1]Based on (3.4) [2]$\tilde{\Lambda}_t$ based on the $\tilde{M}(a,b)$'s.

7. Discussion

A number of extensions of these methods would merit further investigation:

1. The Poisson model could be extended in the usual way to allow for covariance analysis. For the ith car put into service on day x, the mean number of repairs at age t with reporting lag l can be modelled as

$$\log \mu_{xtli} = \log \mu_{xtl}^{(0)} + \mathbf{z}_i'\beta$$

where $\mathbf{z}_i$ is a vector of regression variables and β is a vector of regression parameters specifying, for example, manufacturing characteristics or time of manufacture, model line, etc. and, if available for all individuals, use or environmental factors that apply when the car is put in service. Note that

$$\log \mu_{xtl}^{(0)} = \log f_l + \log \lambda_t$$

so that failure rates are here allowed to depend on z_i. The lack of dependence of reporting lags on x and z_i is important here and needs to be checked. In most instances, when the covariates take only a relatively small set of values, stratification could be used. Thus, if $N_x^{(z)}$ is the number of cars with covariate value z placed in service on day x and $\{f_l^{(z)}\}$ is the corresponding distribution of reporting lags, a natural model would be

$$\log \mu_{xtl}(z) = \log\{N_x^{(z)} f_l^{(z)}\} + \log \lambda_t + \mathbf{z}'\beta.$$

Standard software for Poisson regression (e.g. GLIM) could be used for analysis.
If $\mathbf{z}$ includes covariates that correspond to use characteristics, they will typically not be observed for all cars put into service. Supplementary sampling will then be needed to estimate regression coefficients. Kalbfleisch and Lawless (1988) discuss some of the issues involved.

2. In some instances, parametric models provide a useful alternative to the nonparametric approach discussed in this paper. This is particularly true when the total number of items in service is relatively small.

3. The use of mileage as well as time in service is an important area for further investigation. It is to be expected that rates of failure at age t will also depend on mileage accumulated, and it would be useful to incorporate this in the analysis. As well, when warranties depend on both age and mileage, the methods discussed here do not estimate actual failure rates (see Remark 6 in Section 2.1). The use of models which relate failure rates to mileage would be an important generalization. Here again, supplementary information on the mileages accumulated for the population of cars in service is important to attack this problem.

4. The summary of warranty claims in terms of the age of the car is useful for reporting to engineering or in assessing reliability problems. This is especially true in considering various failure types. For accounting purposes, it is often more useful to summarize warranty repairs or costs in terms of calendar time. In this, the aim of estimation is the total cost incurred due to warranty claims up to some specified calendar time. This problem can be approached with methods similar to those discussed above.

5. In some instances, multiple visits are required to remedy the same basic failure. A possible model would allow failures to occur as a Poisson process and then a subsequent process generating visits to correct the failure. If the warranty data set had information that allowed one to identify multiple visits to address a single problem, models of this sort could be used and analyzed.

Appendix

Mean and variance of $m(t) - \hat{m}(t)$ and $M(t) - \hat{M}(t)$ when the f_l's are known

From (1.2) and (2.3), one can write

$$m(t) = N^{-1}\{n_{.t.} + \sum\sum_{\substack{x+l>T-t \\ x\leq\tau}} n_{xtl}\} = m_1(t) + m_2(t)$$

and, at time T, $m_1(t)$ is known whereas $m_2(t)$ must be estimated or predicted. Similarly, $M(t) = M_1(t) + M_2(t) = \sum_{u=0}^{t}\{m_1(u) + m_2(u)\}$ where $M_1(t)$ is known. Now

$$\begin{aligned}\hat{m}(t) &= N^{-1}\{n_{.t.} + \sum\sum_{\substack{x+l>T-t \\ x\leq\tau}} N_x\hat{\lambda}_t f_l\} \\ &= m_1(t) + \hat{m}_2(t).\end{aligned}$$

Thus $var(m(t) - \hat{m}(t)) = var(m_2(t) - \hat{m}_2(t)) = var(m_2(t)) + var(\hat{m}_2(t))$ since $m_2(t)$ and $\hat{m}_2(t)$ are independent. Straightforward calculation shows that

$$var(m_2(t)) = N^{-2} \sum\sum_{\substack{x+l>T-t \\ x\leq\tau}} N_t\lambda_t f_l = N^{-2}[N - R_{T-t}]\lambda_t$$

and

$$var(\hat{m}_2(t)) = N^{-2}(N - R_{T-t})^2\lambda_t/R_{T-t}.$$

This yields the result (2.6) and (2.7) follows since the terms $m(u) - \hat{m}(u)$ are independent.

When there is extra Poisson variation, the estimation of variances is more complex. Note however that

$$M_2(t) = N^{-1}\sum_{u=0}^{t} \sum\sum_{\substack{x+l>T-u \\ x\leq\tau}} n_{xul}$$

$$\hat{M}_2(t) = N^{-1} \sum_{u=0}^{t} \sum\sum_{\substack{x+l>T-u \\ x \le \tau}} \hat{\lambda}_u N_x f_l = N^{-1} \sum_{u=0}^{t} \hat{\lambda}_u (N - R_{T-u}).$$

Here $M_2(t)$ and $\hat{M}_2(t)$ are not independent. We note, however, that $\hat{\lambda}_u = n_{.u.}/R_{T-u}$ and that

$$M_2(t) - \hat{M}_2(t) = N^{-1} \sum_{x=0}^{\tau} \sum_{u=0}^{t} \sum_{l=0}^{\infty} a_{xul} n_{xul}$$

where $a_{xul} = -(N - R_{T-u})/R_{T-u}$ if $x+u+l \le T$ and 1 if $x+u+l > T$. Now the formulae in (4.4) and (4.5) can be brought to bear. Some algebra shows that

$$var(M_2(t) - \hat{M}_2(t)) = N^{-2} \sum_{x=0}^{\tau} N_x \{\delta(\Lambda_t - B_x) + \Lambda_t - D_x\}$$

where $B_x = N \sum_{u=0}^{\min(T-t,x)} \{\lambda_u F_{T-x-u}/R_{T-u}\}$ and

$$D_x = \sum_{u=0}^{\min(T-t,x)} [1 - (\frac{N - R_{T-u}}{R_{T-u}})^2] \lambda_u F_{T-x-u}.$$

Acknowledgment

We would like to thank Professor R.J. MacKay for helpful discussions and comments on this manuscript. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to J.D. Kalbfleisch and J.F. Lawless and by a grant from the Manufacturing Research Corporation of Ontario.

References

Ascher, H. and Feingold, H. (1978). Is there repair after failure? In *Proceedings of the 1978 Annual Reliability and Maintainability Symposium (IEEE)*, 190-197.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1-22.

Kalbfleisch, J.D. and Lawless, J.F. (1988). Estimation of reliability from field performance studies (with discussion). *Technometrics 30*, 365-88.

Kalbfleisch, J.D. and Lawless, J.F. (1989a). Inference based on retrospective ascertainment. An analysis of the data on transfusion related AIDS. *Journal of the American Statistical Association, 84*, 360-72.

Kalbfleisch, J.D. and Lawless, J.F. (1989b). Estimating the incubation time distribution and expected number of cases for transfusion-associated acquired immune deficiency syndrome. *Transfusion, 29*, 672-676.

Kalbfleisch, J.D. and Lawless, J.F. (1989c). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *University of Waterloo Technical Report, STAT-89-23.*

Kaminsky, K.S. (1987). Prediction of IBNR claim counts by modelling the distribution of report lags. *Insurance: Mathematics and Economics, 6*, 151-159.

Keiding, N. and Gill, R. (1987). Random truncation models and Markov processes. University of Copenhagen Statistical Research Unit Research Report 87/3.

Lagakos, S.W., Barraj, L.M. and DeGruttola, V. (1988). Nonparametric analysis of truncated survival data, with applications to AIDS. *Biometrika, 75*, 515-523.

Lawless, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association, 82*, 805-815.

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

Nelson, W. (1988). Graphical analysis of system repair data. *Journal of Quality Technology, 20*, 24-35.

Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association, 81*, 977-986.

Woodroffe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics, 13*, 163-177.

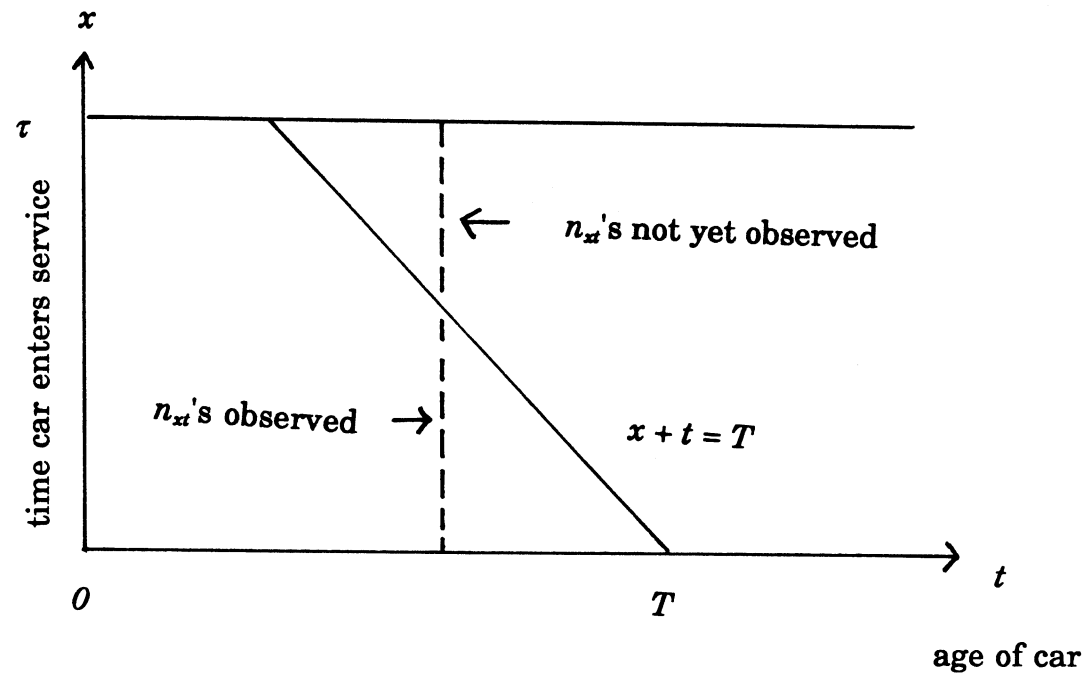

Figure 1

Claim numbers n_{xt} up to day T (no reporting lags)

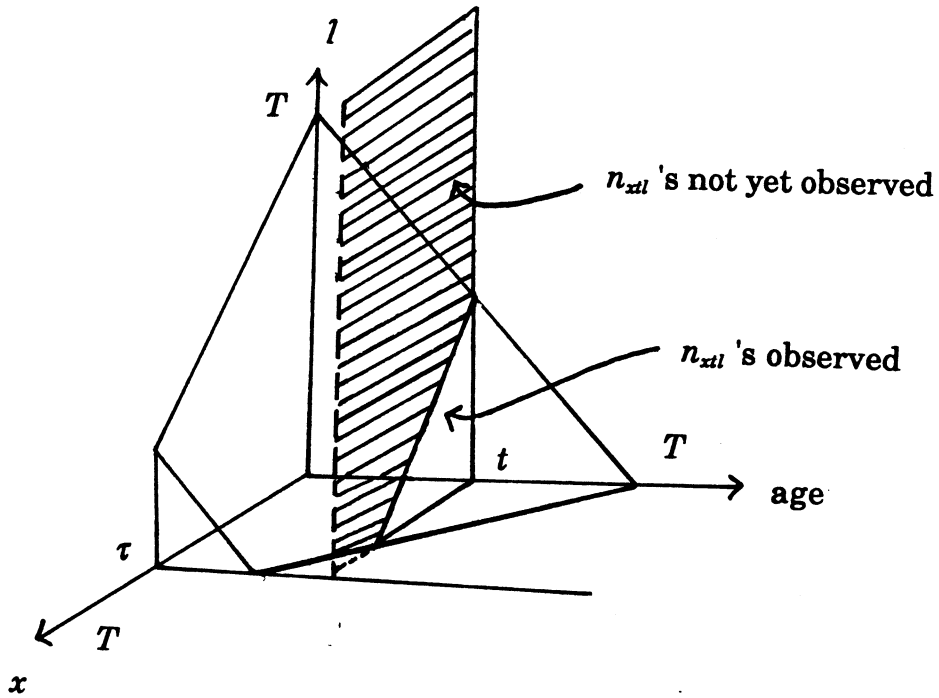

Figure 2

Claim numbers n_{xtl} up to day T

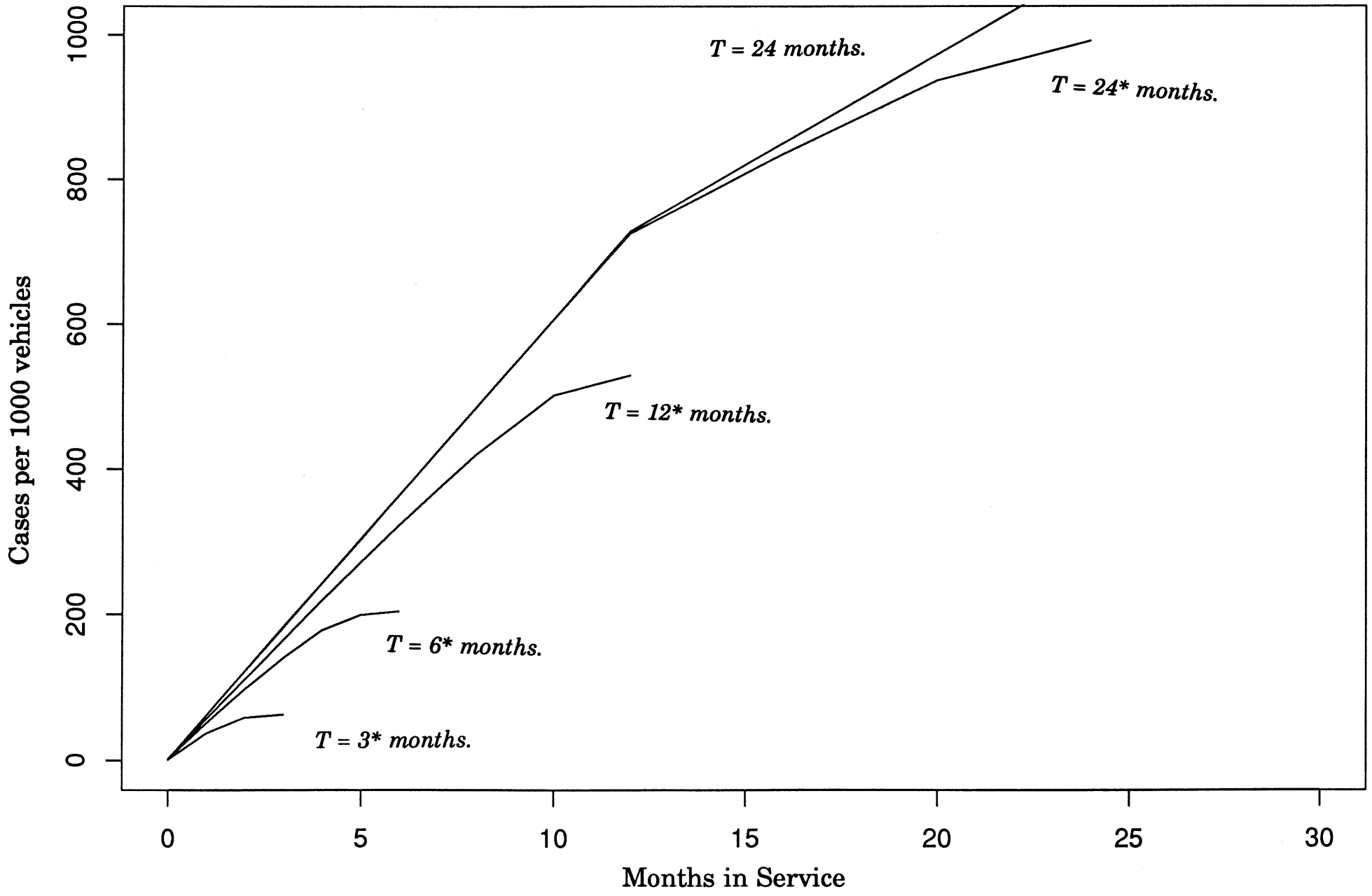

Figure 3: Expected values of $1000\hat{\lambda}_t$ when reporting lags are ignored
(* indicates that reporting delay is ignored)

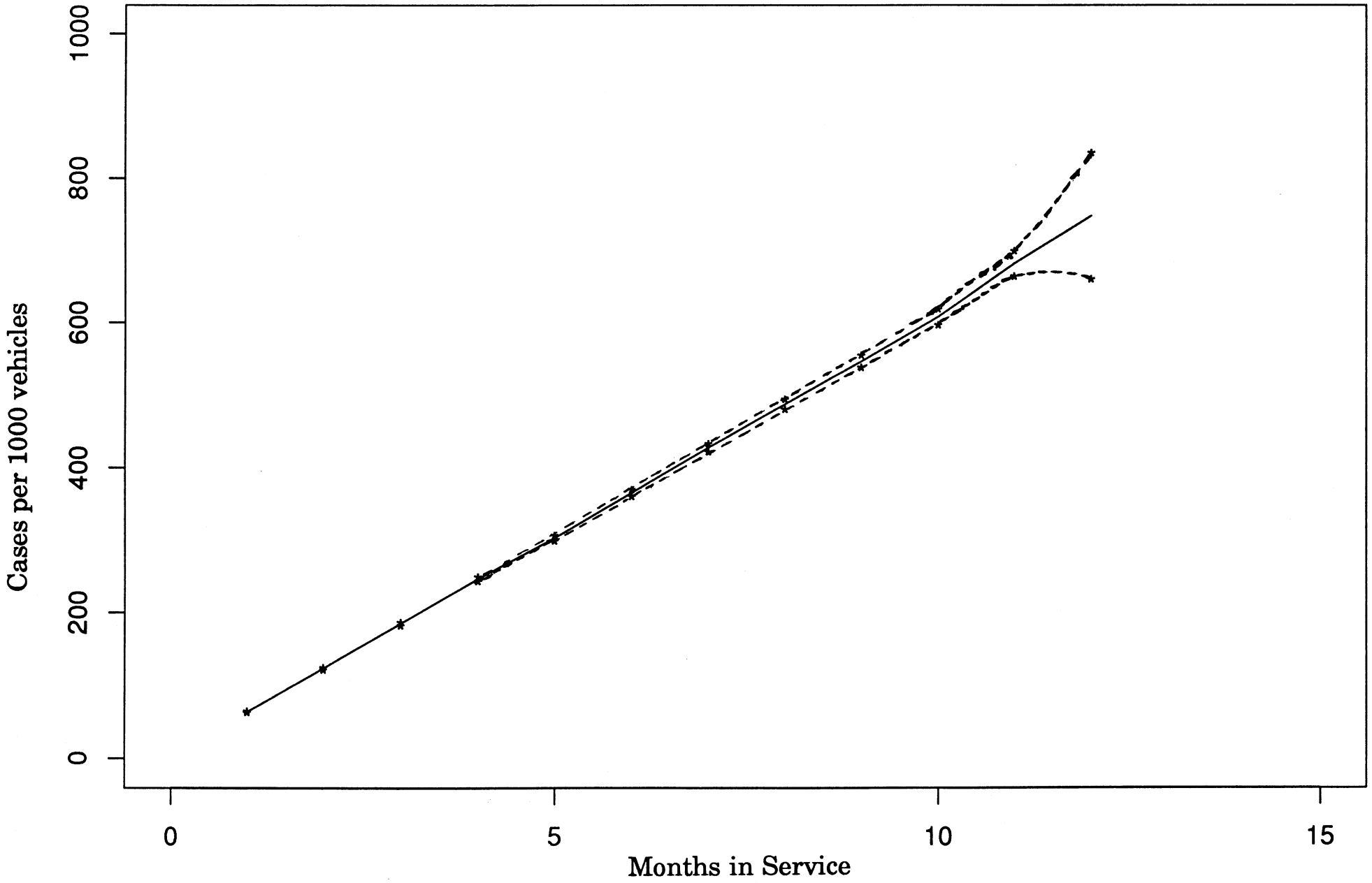

Figure 4: Estimated cumulative claims *1000* $\hat{M}$ *(t)* and 95% prediction limits