Modeling Loss Development with Micro Data

by Daniel R. Corro

Abstract:

Actuaries have long since recognized the value of survival analysis for calculating case reserves. While there is also a patent connection between setting case reserves and loss development, the tools of survival analysis have been largely ignored in building loss development models. This may be explained in part from history: data storage and computation limitations have traditionally restricted loss development models to aggregated data unsuited to the analysis of individual lives. Until fairly recently, actuarial mathematics has followed an unnecessarily restrictive interpretation of survival analysis.

The thesis of this paper is that comparatively recent advances in data processing and in survival analysis theory can be exploited to provide an alternative approach to loss development. Computerized insurance data files now enable automatic production of loss and premium "triangles" directly from individual claim and individual policy rating class exposure data. That suggests building development models directly from micro-data. Moreover, much of that data is transactional; making it a natural fit to survival analysis models.

The idea is to regard paid losses on open claims as "right-censored" along the lines in which incomplete information is handled in the accelerated failure time models of bio-statistics and engineering. It is no longer the claimant that represents a "life" but the claim itself, with "death" or "failure" corresponding to claim closure. Also, the paper discusses the use of paid dollars—as well as time—to parameterize the progression from claim emergence to claim closure.

The discussion argues that, under this setup, the "expectation of life" plays the role of "case reserve". The paper considers the application of this case reserve to "develop" paid losses to "ultimate incurred losses". The main result of the paper is the proof that this "ultimate loss" model has the correct mean, namely the same mean as the accelerated failure time model, but without the complicating presence of censored observations.

I. Introduction

Actuaries have long since recognized the value of survival analysis for calculating reserves for lifetime pensions or related benefits. This is consistent with the usual identification within actuarial mathematics of a "life" with the life of the beneficiary—this paper argues that that is an unnecessarily restrictive interpretation of survival analysis

While there is a patent connection between setting case reserves and loss development, the tools of survival analysis have been largely ignored in building loss development models. Loss development models are built from "triangles" of aggregated premiums or losses. Historically, that aggregation represented a necessary interim step in devising a model to track how premium and loss data "matures". With the advent of computerized data files, it is possible to build the triangles on demand directly from individual claim and exposure data. That same capability opens the door to building other development models directly from micro data.

Recent advances in loss development modeling take advantage of weighted regression models. In fact, the more convincing argument in favor of using regression methods for loss development is more algebraic than conceptual. The models successfully unify traditionally different loss development formulas. They also enable a more systematic review of "residuals" and provide confidence intervals about the factor estimates. That ability to attach a confidence interval is clearly very important. It has, however, led some to promote regression based loss development models as inherently superior due to the ability of (weighted) OLS (ordinary least squares) regression to deal with uncertainty. That argument would be more convincing if the models actually were built from modeling an underlying stochastic process. It is not at all apparent that the "uncertainty" of loss development estimates is the same as that customarily measured in OLS. Indeed, the loss development regression equations make no pretense of including all the requisite explanatory variables. It is probably more accurate to regard the information provided by the residuals and the statistical tests of the model parameters as measuring goodness of fit rather than of occult uncertainty.

For the purpose of this paper, there are two noteworthy points as regards the use of regression models for loss development: (1) the inability of the loss development regression models to incorporate explanatory variables and (2) the inability of OLS to handle censored data.

The last thirty years has seen significant advancements in survival analysis, much of it focussed on improving its ability to handle explanatory variables within survival time models. While not the focus of the discussion, the model presented in this paper is readily adapted to include explanatory variables and/or to group the underlying data into strata.

The second point is more central to this paper. The very reason that reported loss figures require development is the fact that the underlying premium or loss data is incomplete. Survival analysis is designed to handle censored observations. This paper shows how a

survival time model can be used to determine a "case reserve" for all "censored", i.e. open, claims. It is proven that this reserve provides a means to "develop" each loss and that the resulting "uncensored" distribution has the same expected value (average cost per claim) as that specified by the survival time model. From this perspective, it is the uncensored "developed" loss data that is more naturally suited to OLS models.

This suggests that while using (weighted) OLS models to develop losses can unify the calculations it does not attack the heart of the issue, which is the presence of immature data. It also ignores the fundamental fact that computerized insurance data files now enable automatic production of loss and premium "triangles" directly from individual claim and individual policy rating class exposure data. That capability argues for building development models directly from micro-data. Moreover, much of that data is transactional; making it a natural fit to survival analysis models.

The idea is to regard paid losses on open claims as "right-censored" along the lines in which incomplete information is handled in the survival time models used in bio-statistics and engineering. It is no longer the beneficiary who represents a "life" but the claim itself, with "death" or "failure" corresponding to claim closure. Of course, one can still use time to track the claim from its reporting to closure. In that case the survival time models can be used to study claim duration. This paper also suggests the use of paid dollars—as an alternative to time—to parameterize the progression from claim emergence to claim closure. Under that setup, the "expectation of life" plays the role of "case reserve".

The paper considers the application of this case reserve to "develop" paid losses to "ultimate incurred losses". The idea is to begin with censored data, i.e. claim data that includes paid to date on both open and closed cases. The paid amounts on open cases are converted to their ultimate incurred value as the sum of payments to date plus the case reserve with claim status changing from censored = "open" to uncensored = "closed". This yields a data set of uncensored claim data. The main result of the paper is the proof that this "ultimate loss" model has the same mean cost per case as the survival time model but without the complicating presence of censored observations. In particular, aggregate reserves can be derived from simple aggregation of the case reserves. These aggregate reserves must conform to historical loss development patterns and are readily allocated to any claim subset, in particular to any sub-line of insurance or rating class group. The details are worked out in two generic cases: Section II handles discrete data and Section III presents a continuous model. Section IV concludes with some general comments and suggestions for future study.

II. Discrete Model.

Let $0 < t_1 < t_2 < \ldots < t_N \leq 1$ be a series of discrete "times". Assume there are f_i observations at time t_i of which $f_{0,i}$ are "censored" and $f_{1,i}$ are observed "failures". We have $f_i = f_{0,i} + f_{1,i}, 1 \leq i \leq N$ and let $n = \sum_{i=1}^{N} f_i$ be the number of all observations.

For the purpose of this paper, there are two examples to keep in mind, in both the censored observations represent open claims and the failures represent closed claims. In one interpretation the t_i represent duration in time from either report date (or perhaps the accident date) to a current evaluation date for open claims and to the date of closure for closed claims. This interpretation is appropriate when studying claim duration. In the second interpretation, the t_i represent the paid loss at a current evaluation for open claims and the final incurred loss for closed claims. The second interpretation is the one promoted for using a survival time model to assign case reserves and to model loss development.

These interpretations require that all paid amounts be indexed to a common—presumably the current--purchasing power (inflation adjustment). In the case when the coverage terms may vary over the time frame, duration and paid amounts should also be adjusted to a common—presumably the current and applicable—terms of coverage (benefit on-level adjustment). One of the purported advantages of age to age factors are typically derived from losses arising under common (or nearly common) coverage terms and so the on-level adjustment can be is assumed to cancel out. There is also the question of whether and how to deal with any trend over and above inflation or changes in coverage. While key to any practical application of the method, these issues lie beyond the scope of this paper.

We first make some general observations and develop our notation ignoring the ability to identify censored data. Begin by recalling the usual survival function $S(t_i)$. For this, set $f_0 = t_0 = 0$ and define:

$$start_i = n - \sum_{j=0}^{i-1} f_j = \sum_{j=i}^{N} f_j, \quad 0 \le i \le N+1$$

$$stop_i = start_{i+1} \quad 0 \le i \le N$$

Note that $start_0 = start_1 = stop_0 = n$, $start_N = f_N$ and $stop_N = start_{N+1} = 0$. Set $S(0) = S(t_0) = 1$ and define recursively:

$$S(t_i) = S(t_{i-1})p_i \text{ where } p_i = \frac{stop_i}{start_i}$$

which is the usual survival function, inasmuch as the ratio represents the empirical probability that an observation "survives" the i-th interval conditional upon its surviving to the beginning of the interval—ignoring all censoring and interpreting all observations as observed "failures" or "deaths".

Observe that $S(t_i) = \frac{stop_i}{n}$; indeed $S(t_0) = 1 = \frac{n}{n}$, and by induction on i:

$$S(t_i) = S(t_{i-1}) \cdot \frac{stop_i}{start_i} = \frac{stop_{i-1}}{n} \cdot \frac{stop_i}{start_i} = \frac{start_i}{n} \cdot \frac{stop_i}{start_i} = \frac{stop_i}{n},$$

as claimed. Let T denote the random variable of the distribution of the (uncensored) "failures" at times t_i. It is well known that the expected value of T can be determined from the survival function:

$$E(T) = \mu = \frac{1}{n}\sum_{i=1}^{n} f_i t_i = \sum_{i=1}^{N} S(t_{i-1}) \cdot (t_i - t_{i-1}) = t_1 + \sum_{i=1}^{N-1} S(t_i) \cdot (t_{i+1} - t_i)$$

In this case, this is readily verified directly; indeed, for any nonnegative integer k:

$$\sum_{i=1}^{N} S(t_{i-1}) \cdot ({t_i}^k - {t_{i-1}}^k) = \sum_{i=1}^{N} \frac{stop_{i-1}}{n} \cdot ({t_i}^k - {t_{i-1}}^k) = \frac{1}{n}\sum_{i=1}^{N} start_i \cdot ({t_i}^k - {t_{i-1}}^k):$$

$$= \frac{1}{n}\sum_{i=1}^{N}\sum_{j=i}^{N} f_j \cdot ({t_i}^k - {t_{i-1}}^k) = \frac{1}{n}\left(\sum_{\substack{1 \le i,\, j \le N \\ j \ge i}} f_j \cdot {t_i}^k + \sum_{\substack{1 \le i,\, j \le N \\ j \ge i}} f_j \cdot {t_{i-1}}^k \right)$$

$$= \frac{1}{n}\left(\sum_{\substack{1 \le i,\, j \le N \\ j \ge i}} f_j \cdot {t_i}^k + \sum_{\substack{1 \le i,\, j \le N \\ j > i}} f_j \cdot {t_i}^k \right) = \frac{1}{n}\left(\sum_{i=1}^{N} f_i \cdot {t_i}^k \right) = E(T^k)$$

Consider next what this implies when restricted to those observations with time $\ge t_i$ for a fixed observed time t_i. Using transparent notation, for that subset of observations with times $\hat{t}_k = t_{i+k}$ we have:

$$\hat{S}(\hat{t}_k) = \frac{S(t_{i+k})}{S(t_i)} \quad k = 0,1,\ldots \hat{N} = N - i$$

$$\hat{\mu} = \frac{1}{start_i}\sum_{j=i}^{N} f_j t_j = \hat{t}_0 + \sum_{k=1}^{\hat{N}} \hat{S}(\hat{t}_k)(\hat{t}_k - \hat{t}_{k-1}) = t_i + \rho_i$$

$$\text{where } \rho_i = \sum_{k=0}^{N-i} \frac{S(t_{i+k})}{S(t_i)}(t_{i+k+1} - t_{i+k}) = \sum_{j=i}^{N} \frac{S(t_j)}{S(t_i)}(t_{j+1} - t_j)$$

Here, ρ_t can be interpreted as the "expectation of life". We make the convention that $t_{N+1} = t_N$; note that $\rho_N = 0$ and $\rho_0 = \mu$.

We now take into account the ability to identify censored data, which in the two interpretations amounts to taking into account claim status. The survival time model is constructed using probabilities of failure determined based on observed failures. Censored observations are regarded as survivors up to their observed time of departure and are ignored in the subsequent intervals (heretofore they were indifferently handled the same as any other failure).

Analogous to the above, we define:

$$stop_{1,i} = start_i - f_{1,i} = \sum_{j=i}^{N} f_j - f_{1,i} = f_{0,i} + \sum_{j=i+1}^{N} f_j = f_{0,i} + start_{i+1}$$

$$start_{1,i} = start_i$$

As before set $S_1(0) = S_1(t_0) = 1$ and define recursively:

$$S_1(t_i) = S_1(t_{i-1}) \cdot \frac{stop_{1,i}}{start_{1,i}} = S_1(t_{i-1}) \cdot p_{1,i} \text{ where } p_{1,i} = \frac{stop_{1,i}}{start_{1,i}}.$$

Motivated by the above, we further define

$$\rho_{1,i} = \sum_{j=i}^{N} \frac{S_1(t_j)}{S_1(t_i)} (t_{j+1} - t_j)$$

Now clearly the function $S_1(t_i)$ can be viewed as the survival function for a survival time model with random variable denoted here by $\tilde{T}$. With transparent notation, we have in particular that:

$$\rho_{1,0} = E(\tilde{T}) = \tilde{\mu}$$

This survival time model takes into account the censored nature of the data and so for our interpretations this mean is a better estimate of claim duration or average claim costs than could normally be obtained from simple descriptive statistics derived from the claim data. This mean corresponds to the expected claim duration or cost at claim closure. When the t_i are interpreted as paid costs, this mean corresponds to paid at closure, i.e. incurred, and this survival time model can be regarded as a formula for paid loss development.

There is another evident way to use this model to "develop" paid losses to incurred losses. Define

$$\hat{t}_i = t_i + \rho_{1,i} = t_i + \sum_{j=i}^{N-1} \frac{S_1(t_j)}{S_1(t_i)}(t_{j+1} - t_j) \le t_i + \sum_{j=i}^{N-1}(t_{j+1} - t_j) = t_N \le 1$$

It is natural to consider the effect of replacing the $f_{0,i}$ censored observations with time t_i with the same number of "uncensored" observations at (usually later) time $\hat{t}_i$. This results in a distribution with random variable we denote by $\hat{T}$. The main result of this section is:

Proposition I.1: *For this discrete model* $\tilde{\mu} = \hat{\mu}$

Proof: We have:

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\hat{t}_i + f_{1,i}t_i - \frac{1}{n}\sum_{i=1}^{N} f_{0,i}t_i + f_{1,i}t_i$$

$$= \frac{1}{n}\left(\sum_{i=1}^{N} f_{0,i}(\hat{t}_i - t_i) + f_{1,i}(t_i - t_i)\right) = \frac{1}{n}\sum_{i=1}^{N} f_{0,i} \cdot \rho_{1,i} = \frac{1}{n}\sum_{i=1}^{N-1} f_{0,i} \cdot \rho_{1,i}$$

A simple induction on $q=0, 1, 2, \ldots, N-i$ shows that:

$$\frac{S_1(t_{i+q})}{S_1(t_i)} = \prod_{k=1}^{q} \frac{stop_{1,i+k}}{start_{1,i+k}} = \prod_{k=1}^{q} p_{1,i+k}$$

It then follows that

$$\rho_{1,i} = \sum_{j=i}^{N-1} \frac{S_1(t_j)}{S_1(t_i)}(t_{j+1} - t_j) = \sum_{q=0}^{N-1-i} \frac{S_1(t_{i+q})}{S_1(t_i)}(t_{i+q+1} - t_{i+q}) = \sum_{q=0}^{N-1-i} \prod_{k=1}^{q} p_{1,i+k}(t_{i+q+1} - t_{i+q})$$

and so

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{N-1} f_{0,i} \cdot \rho_{1,i} = \frac{1}{n}\sum_{i=1}^{N-1} f_{0,i} \cdot \sum_{q=0}^{N-1-i} \prod_{k=1}^{q} p_{1,i+k}(t_{i+q+1} - t_{i+q})$$

On the other hand, both $\tilde{\mu}$ and μ can be determined from their corresponding survival functions, which implies:

$$\tilde{\mu} - \mu = \sum_{a=1}^{N} S_1(t_{a-1}) \cdot (t_a - t_{a-1}) - \sum_{a=1}^{N} S(t_{a-1}) \cdot (t_a - t_{a-1}) = \sum_{a=1}^{N} (S_1(t_{a-1}) - S(t_{a-1})) \cdot (t_a - t_{a-1})$$

Set $d_a = S_1(t_a) - S(t_a)$ then we have:

$$d_a = S_1(t_a) - S(t_a) = p_{1,a}S_1(t_{a-1}) - p_a S(t_a - 1) = \frac{stop_{1,a}}{start_a} S_1(t_{a-1}) - \frac{stop_a}{start_a} S(t_a - 1)$$

$$\Rightarrow start_a d_a = stop_{1,a}\left(d_{a-1} + S_1(t_{a-1})\right) - stop_a S(t_a) = stop_{1,a} d_{a-1} + S_1(t_{a-1})(stop_{1,a} - stop_a)$$

$$= stop_{1,a} d_{a-1} + S_1(t_{a-1}) f_{0,a} = stop_{1,a} d_{a-1} + stop_{a-1} \frac{f_{0,a}}{n} = stop_{1,a} d_{a-1} + start_a \frac{f_{0,a}}{n}$$

$$\Rightarrow d_a = p_{1,a} d_{a-1} + \frac{f_{0,a}}{n}$$

We claim that:

$$d_a = \frac{1}{n} \sum_{b=1}^{a} f_{0,b} \prod_{c=b+1}^{a} p_{1,c}$$

Since $d_0 = 0$ the formula holds vacuously for $a = 0$ and for $a = 1$:

$$d_1 = S_1(t_1) - S(t_1) = p_{1,1} - p_1 = \frac{n - f_{1,1}}{n} - \frac{n - f_1}{n} = \frac{f_1 - f_{1,1}}{n} = \frac{f_{0,1}}{n}$$

Proceeding by induction on a, assuming the formula holds for a-1:

$$d_a = p_{1,a} d_{a-1} + \frac{f_{0,a}}{n} = p_{1,a}\left(\frac{1}{n} \sum_{b=1}^{a-1} f_{0,b} \prod_{c=b+1}^{a-1} p_{1,c}\right) + \frac{f_{0,a}}{n}$$

$$= \frac{1}{n}\left(\sum_{b=1}^{a-1} f_{0,b} \prod_{c=b+1}^{a} p_{1,c} + f_{0,a}\right) = \frac{1}{n}\left(\sum_{b=1}^{a} f_{0,b} \prod_{c=b+1}^{a} p_{1,c}\right)$$

as required. We find that:

$$\tilde{\mu} - \mu = \sum_{i=1}^{N} d_{a-1} \cdot (t_a - t_{a-1}) = \sum_{i=1}^{N}\left(\frac{1}{n} \sum_{b=1}^{a-1} f_{0,b} \prod_{c=b+1}^{a-1} p_{1,c}\right) \cdot (t_a - t_{a-1})$$

$$= \frac{1}{n} \sum_{a=1}^{N} \sum_{b=1}^{a-1} f_{0,b} \prod_{c=b+1}^{a-1} p_{1,c} \cdot (t_a - t_{a-1}) = \frac{1}{n} \sum_{b=1}^{N-1} \sum_{a=b+1}^{N} f_{0,b} \prod_{c=b+1}^{a-1} p_{1,c} \cdot (t_a - t_{a-1})$$

making the change of index variables:

$$q = a - b - 1, \quad q = 0,1,\ldots,N-b-1; \qquad k = c - b, \quad k = 1,2,\ldots,a-b-1 = q$$

$$\tilde{\mu} - \mu = \frac{1}{n}\sum_{b=1}^{N-1}\sum_{q=0}^{N-b-1} f_{0,b}\prod_{k=1}^{q} p_{1,b+k}\cdot(t_{q+b+1} - t_{q+b}) = \hat{\mu} - \mu$$

completing the proof of Proposition II.1

III. Continuous Model.

Let $f(t)$ denote a positive, real-valued (Lebesgue) integrable function on (0,1) satisfying:

$$\int_0^1 f(t)dt = 1 - p \quad \text{where} \quad 0 \le p \le 1$$

and define:

$$S(t) = 1 - \int_0^t f(s)ds \text{; note that } 1 = S(0) \ge S(t) \ge S(1) = p \quad \text{for} \quad t \in (0,1).$$

To eliminate some uninteresting degenerate cases (which can be readily avoided by rescaling t), we assume that $1 > S(t) > p$ for $t \in (0,1)$.

As is customary, we refer to $S(t)$ as the *survival function*, $f(t)$ as the *probability density function [PDF]* and t as "time". We also let T denote the random variable for the distribution of survival times and $\mu = E(T)$ the mean duration. Survival analysis refers to the following function:

$$h(t) = \frac{f(t)}{S(t)} \quad , \quad t \in (0,1)$$

as the *hazard rate function* or sometimes as the *force of mortality*. The hazard rate function measures the instantaneous rate of failure at time t and can be expressed as a limit of conditional probabilities:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \le T < t + \Delta t \mid T \ge t\}}{\Delta t}$$

There are many well-known relationships and interpretations of these functions—refer to Allison[1] for a particularly succinct discussion. It is convenient to recall that setting

$$g(t) = \int_0^t h(s)ds \text{ then } S(t) = e^{-g(t)} \text{ for } 0 \le t \le 1.$$

We will make extensive use of the following:

Proposition III.1: *For any positive integer n:*

$$E(T^n) = n\int_0^1 t^{n-1}S(t)dt\,.$$

Proof: The proof is a straightforward integration by parts:

$$u = -S(t) \quad du = f(t)dt; \quad v = t^n \quad dv = nt^{n-1}dt$$

$$E(T^n) = \int_0^1 t^n f(t)dt + p = \int_0^1 v\,du + p = uv\Big]_0^1 - \int_0^1 u\,dv + p$$

$$= -t^n S(t)\Big]_0^1 + n\int_0^1 t^{n-1}S(t)dt + p = -p + n\int_0^1 t^{n-1}S(t)dt + p = n\int_0^1 t^{n-1}S(t)dt$$

completing the proof.

Fix t and restrict attention to values of time $w > t$. From Proposition III.1, the expectation of life at time t, given survival to time t, is just:

$$\rho(t) = \int_t^1 \frac{S(w)}{S(t)}dw = \frac{\int_t^1 S(w)dw}{S(t)}.$$

and observe that

$$\rho(0) = \mu,\ \rho(1) = 0 \quad \text{and} \quad t < t + \rho(t) < 1 \text{ for } t \in (0,1).$$

For the case of interest in this paper, we regard some observations as "censored" (e.g. open cases); more precisely, assume that the PDF can be split into two continuous functions:

$$f(t) = f_u(t) + f_1(t) \text{ where } f_u(t) = \text{censored and } f_1(t) = \text{uncensored on } [0,1]$$

The two associated survival time models, taking into account the presence of censored data, are most readily defined via their hazard functions:

$$h_i(t) = \frac{f_i(t)}{S(t)} \quad , \quad i \in \{0,1\}, t \in (0,1) \text{ and so } h = h_0 + h_1.$$

This symmetry illustrates that censored data can be viewed as being subject to a double decrement—one decrement as "failure" and a second as "censure". Define

$$g_i(t) = \int_0^t h_i(w)dw \text{ with survival function } S_i(t) = e^{-g_{i1}(t)} \text{ for } 0 \le t \le 1 \text{ and set } p_i = S_i(1).$$

Note that $S = S_0 S_1$, in particular $p = p_0 p_1$.

Recall that in one of the model interpretations suggested here, the "censored" observations are open claims with *paid loss* = t and the observed "failures" are closed claims with *paid loss=incurred loss* = t. The survival time model with $i=1$ presents a convenient way of specifying that, as with the discrete case, case closures make up the numerator of the conditional probability used to specify the hazard rate function. In particular the probability p_1 measures the likelihood of the (normalized and limited) incurred cost of a claim equaling the per claim loss limit value 1, taking into account the presence of open claims. Define

$$\rho_i(t) = \frac{\int_t^1 S_i(w)dw}{S_i(t)}.$$

For the survival time model with $i=1$, this can be regarded as a case reserve applicable to open claims with *paid loss* = t. Indeed, let $\tilde{T}$ denote the random variable of the distribution defined by the survival function S_1. From Proposition II.1, the mean is

$$\tilde{\mu} = E(\tilde{T}) = \int_0^1 S_1(t)dt$$

which can be regarded as an estimate for the average incurred cost per claim.

As in the discrete case, we are interested in what happens if we replace all the open claims with their estimated incurred loss, traditionally defined as paid plus case reserve:

$$\hat{t} = t + \rho_1(t).$$

Let $\hat{T}$ denote the corresponding random variable and observe that under this construction the expected incurred cost per case is:

$$\hat{\mu} = E(\hat{T}) = \int_0^1 (f_0(t)\hat{t} + f_1(t)t)dt + p$$

The key point is that the $\hat{T}$ distribution is uniformly subject only to one decrement, "failure" = case closure, i.e. involves no censored data, making it a preferable candidate for conventional statistical analysis, in particular OLS cost models. It is reasonable to expect that $\hat{\mu} = \tilde{\mu}$; indeed, we have:

Proposition II.2: *For this continuous model* $\tilde{\mu} = \hat{\mu}$

Proof: The proof is similar to that for the discrete case. Begin with the observation that:

$$\hat{\mu} - \mu = \int_0^1 (f_0(t)(\hat{t} - t) + f_1(t)(t - t)dt + (p - p) = \int_0^1 f_0(t)\rho_1(t)dt$$

On the other hand, note that:

$$\frac{dS_1}{dt} = \frac{de^{-g_1}}{dt} = -e^{-g_1}\frac{dg_1}{dt} = -S_1 h_1 = \frac{-f_1 S_1}{S}$$

and so:

$$\frac{d(1 - \frac{S}{S_1})}{dt} = -\frac{d\frac{S}{S_1}}{dt} = -\left[\frac{S_1(-f) - S\left(\frac{-f_1 S_1}{S}\right)}{S_1^2}\right] = \frac{f - f_1}{S_1} = \frac{f_0}{S_1}$$

$$\text{Since } S(0) = S_1(0) = 1 \text{ the FTC} \Rightarrow \left[1 - \frac{S}{S_1}\right](t) = \int_0^t \frac{f_0(w)}{S_1(w)}dw$$

and we find that:

$$\tilde{\mu} - \mu = \int_0^1 (S_1 - S)dt = \int_0^1 S_1\left(1 - \frac{S}{S_1}\right)dt = \int_0^1 S_1(t)\int_0^t \frac{f_0(w)}{S_1(w)}dwdt = \int_0^1 udv = [uv]_0^1 - \int_0^1 vdu$$

where

$$u(t) = \int_0^t \frac{f_0(w)}{S_1(w)}dw \quad u(0) = 0 \quad du = \frac{f_0(t)}{S_1(t)}dt$$

$$dv = S_1(t)dt \quad v(t) = -\int_t^1 S_1(w)dw \quad v(1) = 0$$

It follows that

$$\tilde{\mu} - \mu = \int_0^1 \left(\int_t^1 S_1(w)dw \right) \frac{f_0(t)}{S_1(t)} dt = \int_0^1 \rho_1(t) f_0(t) dt = \hat{\mu} - \mu \Rightarrow \tilde{\mu} = \hat{\mu}$$

completing the proof.

An alternative notation setup is sometimes helpful; suppose the PDF $f>0$ on (0,1), then we can define the ratio of censored observations as a function of t:

$$\alpha(t) = \frac{f_0(t)}{f(t)}$$

We then have:

$$f_0(t) = \alpha(t) f(t) \quad f_1(t) = (1 - \alpha(t)) f(t)$$

By the intermediate value theorem for integrals, we may define a function $\zeta(t)$ (not necessarily unique) satisfying:

$$\int_0^t \alpha(w) h(w)\, dw = \alpha(\zeta(t)) \int_0^t h(w)\, dw \quad 0 \le \zeta(t) \le t \le 1$$

and it is readily verified that

$$S_0(t) = S(t)^{\alpha(\zeta(t))} \qquad S_1(t) = S(t)^{1-\alpha(\zeta(t))}$$

The following examples may help put the notation and results into perspective.

Example III.1 Consider first what happens when the proportion of censored observations is constant for all values of t:

$$f_0(t) = \alpha f(t) \quad f_1(t) = (1 - \alpha) f(t) \quad \text{for some constant } \alpha \in (0,1)$$

then

$$h_0(t) = \alpha h(t) \quad h_1(t) = (1 - \alpha) h(t) \quad S_0(t) = S(t)^{\alpha} \quad S_1(t) = S(t)^{1-\alpha}$$

It follows that the survival models for both censored and uncensored decrements are related as proportional hazard shifts from the pooled data and, for that matter, from one another.

It is instructive to consider a very simple concrete example:

Example III.2 Consider the case:

$$f(t) \equiv 1 \quad p = 0 \quad f_0(t) = t \quad f_1(t) = 1 - t$$

The reader can readily verify, in turn, that:

$$S(t) = 1 - t \quad h_1(t) \equiv 1 \quad S_1(t) = e^{-t} \quad S_0(t) = e^t (1 - t)$$

$$\rho_0(t) = \frac{t + e^{1-t} - 2}{1 - t} \qquad \rho_1(t) = 1 - e^{t-1}$$

$$\mu = \frac{1}{2} \qquad \tilde{\mu} = \hat{\mu} \approx 0.632$$

$$E(\tilde{T}^2) \approx 0.528 > 0.520 \approx E(\hat{T}^2)$$
$$E(\tilde{T}^3) \approx 0.482 > 0.461 \approx E(\hat{T}^3)$$

We have proven that in this setup, the distribution determined from applying a case reserve to "develop" or "uncensor" the data has the same mean as that of the survival time model distribution. The above example shows that in the continuous model, that technique does **not** produce the same distribution as the survival time model.

The percentage of censored observations increases with t in Example III.2, which is counterintuitive for the insurance models in which failure refers to case closure and t represents payment duration or losses paid to evaluation. The symmetry of the setup makes it easy to provide an example in which the percentage of cases closing increases with duration or payment amount:

Example III.3 Switch roles of censored and uncensored in Example III.2:

$$f(t) \equiv 1 \quad p = 0 \quad f_0(t) = 1 - t \quad f_1(t) = t$$

The reader can readily verify, in turn, that:

$$\mu = \frac{1}{2} \qquad \tilde{\mu} = \hat{\mu} \approx 0.718$$

$$E(\tilde{T}^2) \approx 0.563 > 0.548 \approx E(\hat{T}^2)$$
$$E(\tilde{T}^3) \approx 0.465 > 0.432 \approx E(\hat{T}^3)$$

As a final note, the Appendix provides some additional findings, mostly directed toward how the higher moments of $\tilde{T}$ and $\hat{T}$ compare. We have just observed in the examples that they are different.

Section IV: Further Study

The paper provides a straightforward blueprint for using micro-level paid claim data to determine case reserves and, by aggregation, bulk reserves. Conversely, it is clear how to itemize the resulting bulk reserve by line or rating class by reference to the classification of the individual claims. Obviously, there are a number of missing items that this simplistic approach does not address (IBNR for example, as well as the issues of inflation, coverage changes and trend that were noted above). Nevertheless, there are three natural applications: (1) as a potential test for reserve adequacy and (2), conversely, as a way to assign case reserves and also (3) as an alternative to loss development triangles when incorporating information on open cases.

As to (1), the basic point is that the method does not demand any assumptions as regards the structure of the survival function. Moreover, the implied loss development pattern is based on "objective" paid data and in particular does not rely on case reserves. This would seem to provide an objective test, one that is self-correcting over time and whose demanding data and computational requirements are now met by current technology. The basic result of this paper is that the suggestion is unbiased, in a technical sense ($\tilde{\mu} = \hat{\mu}$), and therefore has potential application for testing the adequacy of loss reserves.

As to (2), it was noted that significant advancements have been achieved fairly recently in the area of survival time models, especially as regards their ability to handle stratified data and to incorporate explanatory variables, including "time"-dependent interventions. The approach here focuses on aggregate loss levels, as that is the more relevant to loss development. The paper does not consider how good a case reserve $\rho_1(t)$ is on an individual claim basis. Incorporating claimant demographics and other claim characteristics into the survival time model may provide a convenient and useful alternative to more traditional tabular reserve methods. This is an area worthy of future study.

Application (3) simply reiterates the method of the paper: exploit the ability of survival time models to accept censored data to "develop" an uncensored claim data model. This developed data may be better suited for claims analysis using traditional statistical analysis, including OLS. On the other hand, the ability of survival time models to accommodate explanatory variables suggests their potential for providing more than just a simple fit to data points and for revealing relationships that may genuinely help explain loss development patterns.

As a final observation, we have presented two survival time models for claims analysis, one based on time and measuring claim duration and a second based on dollar payments and estimating costs. An advantage of age-to age paid loss development factors is that they not only provide the ability to develop available paid data to estimate its ultimate cost, they in fact provide the pay-out pattern which produces that estimate. For many purposes, e.g. portfolio management and rate of return analysis, this is key. This paper has focussed separately on building time and money survival curves. However, it is likely that the same insurance files would be used to build one data set underlying both survival curves. In theory, then, the time and money survival curves can be correlated. We challenge an interested reader to determine whether, and if so how, the mechanical approach to loss "development" described here can be refined to yield age-to-age development factors.

References:

Allison, Paul D., Survival Analysis Using the SAS® System: A Practical Guide, *The SAS Institute, Inc., 1995.*

APPENDIX

As regards higher moments $k=2,3...$ for the continuous case, we have, as before:

$$E(\tilde{T}^k) - E(T^k) = k\int_0^1 t^{k-1} S_1(t)\left(1 - \frac{S(t)}{S_1(t)}\right)dt = k\int_0^1 t^{k-1} S_1(t)\int_0^t \frac{f_0(w)}{S_1(w)}dwdt$$

$$= k\int_0^1 t^{k-1} S_1(t)\int_0^t \frac{f_0(w)}{S_1(w)}dwdt = k\int_0^1 udv = k[uv]_0^1 - k\int_0^1 vdu$$

$$u(t) = t^{k-1}\int_0^t \frac{f_0(w)}{S_1(w)}dw \qquad u(0) = 0 \qquad \frac{du}{dt} = t^{k-1}\frac{f_0(t)}{S_1(t)} + \left[\int_0^t \frac{f_0(w)}{S_1(w)}dw\right](k-1)t^{k-2}$$

$$dv = S_1(t)dt \quad v(t) = -\int_t^1 S_1(w)dw \quad v(1) = 0$$

$$= -k\int_0^1 vdu = k\int_0^1\left(\int_t^1 S_1(w)dw\right)t^{k-1}\frac{f_0(t)}{S_1(t)}dt + k(k-1)\int_0^1\left(\int_t^1 S_1(w)dw\right)\left[\int_0^t \frac{f_0(w)}{S_1(w)}dw\right]t^{k-2}dt$$

$$= k\int_0^1 f_0(t)\rho_1(t)t^{k-1}dt + k(k-1)\int_0^1 \frac{\int_t^1 S_1(w)dw}{S_1(t)}(S_1(t))\left(1 - \frac{S(t)}{S_1(t)}\right)t^{k-2}dt$$

$$= k\int_0^1 f_0(t)\rho_1(t)t^{k-1}dt + k(k-1)\int_0^1 \rho_1(t)t^{k-2}(S_1(t) - S(t))dt$$

On the other hand,

$$E(\hat{T}^k) - E(T^k) = \int_0^1 (f_0(t)(\hat{t}^k - t^k) + f_1(t)(t^k - t^k)dt + (p - p) = \int_0^1 f_0(t)\left((t + \rho_1(t))^k - t^k\right)dt$$

$$= \int_0^1 f_0(t)\left(\sum_{j=0}^{k}\binom{k}{j}t^j\rho_1(t)^{k-j} - t^k\right)dt = \int_0^1 f_0(t)\left(\sum_{j=0}^{k-1}\binom{k}{j}t^j\rho_1(t)^{k-j}\right)dt$$

$$= k\int_0^1 f_0(t)t^{k-1}\rho_1(t)dt + k(k-1)\int_0^1 f_0(t)\rho_1(t)\left(\sum_{j=0}^{k-2}\frac{(k-2)!}{j!(k-j)!}t^j\rho_1(t)^{k-1-j}\right)dt$$

$$\Rightarrow E(\tilde{T}^k) - E(\hat{T}^k) = k(k-1)\int_0^1 \rho_1(t)\left(t^{k-2}(S_1(t) - S(t)) - f_0(t)\left(\sum_{i=0}^{k-2}\frac{(k-2)!}{j!(k-j)!}t^j \rho_1(t)^{k-1-j}\right)\right)dt$$

In particular:

$$\tilde{\sigma}^2 - \hat{\sigma}^2 = \left(E(\tilde{T}^2) - \tilde{\mu}^2\right) - \left(E(\hat{T}^2) - \tilde{\mu}^2\right) = E(\tilde{T}^2) - E(\hat{T}^2) = 2\int_0^1 \rho_1(t)\left(S_1(t) - S(t) - \frac{f_0(t)\rho_1(t)}{2}\right)dt$$

For the discrete case, consider the distribution with the same $f_{j,i}$ but with the t_i replaced with t_i^2 The survival probabilities depend only upon the $f_{j,i}$ and so the proof of Proposition I.1 shows that

$$E(\tilde{T}^2) - E(T^2) = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\varphi_{1,i} \text{ where } \varphi_{1,i} = \sum_{j=1}^{N-1}\frac{S_1(t_j)}{S_1(t_i)}(t_{j+1}^2 - t_j^2)$$

On the other hand,

$$E(\hat{T}^2) - E(T^2) = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\left(t_i + \rho_{1,i}\right)^2 - t_i^2 = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\rho_{1,i}\left(2t_i + \rho_{1,i}\right)$$

$$\Rightarrow \quad \tilde{\sigma}^2 - \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\left(t_i + \rho_{1,i}\right)^2 - t_i^2 = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\left(\varphi_{1,i} - 2\rho_{1,i}t_i + \rho_{1,i}^2\right)$$

Define $\sigma_{i,j} = \dfrac{S_1(t_j)}{S_1(t_i)}$ $0 \le i \le j \le N$. We leave to the interested reader the verification that

$$\rho_{1,i} = -t_i + \sum_{j=i+1}^{N}(\sigma_{i,j-1} - \sigma_{i,j})t_j$$

and that

$$\tilde{\sigma}^2 - \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{N} f_{0,i}\alpha_i$$

where

$$\alpha_i = \varphi_{1,i} - 2\rho_{1,i}t_i + \rho_{1,i}^2$$

$$= \sum_{j=i+1}^{N}(\sigma_{i,j-1} - \sigma_{i,j})(1 - \sigma_{i,j-1} + \sigma_{i,i})t_j^2 - 2\sum_{i+1 \le j < k \le N}(\sigma_{i,j-1} - \sigma_{i,j})(\sigma_{i,k-1} - \sigma_{i,k})t_j t_k$$

In the examples, observe that $\tilde{\sigma}^2 \geq \hat{\sigma}^2$. The author would appreciate a direct proof of this along the lines of the calculations presented in the paper. Intuitively, this makes sense since replacing the unknown, but presumably variable, lifetimes of observations censored at age x with a single value = x+ *expectation of life at age x* should result in a population with smaller variance. This is a consequence of the following:

Proposition A.1: *Let* $\{x_1,...x_m, x_{m+1},...,x_n\}$ *be any set of real numbers and* $\{f_1, f_2,...., f_n\}$ *be a set of positive real numbers with* $\sum_{i=1}^{n} f_i = 1$ *and* $\mu = \sum_{i=1}^{n} f_i x_i$

Let α *be any real number and define* $\hat{x}_i = \begin{cases} x_i & 1 \leq i \leq m \\ \rho & m+1 \leq n \end{cases}$ *then*

i) $\mu = \sum_{i=1}^{n} f_i \hat{x}_i \quad \Leftrightarrow \quad \rho = \dfrac{\sum_{i=m+1}^{n} f_i x_i}{\sum_{i=m+1}^{n} f_i}$

ii) $\mu = \sum_{i=1}^{n} f_i \hat{x}_i \quad \Rightarrow \sum_{i=1}^{n} f_i (\hat{x}_i - \mu)^2 \quad \leq \sum_{i=1}^{n} f_i (x_i - \mu)^2$

Proof: i)

$$\mu = \sum_{i=1}^{n} f_i \hat{x}_i \Leftrightarrow \sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{n} f_i \hat{x}_i \Leftrightarrow \sum_{i=m+1}^{n} f_i x_i = \sum_{i=m+1}^{n} f_i \hat{x}_i = \rho \sum_{i=m+1}^{n} f_i \Leftrightarrow \rho = \frac{\sum_{i=m+1}^{n} f_i x_i}{\sum_{i=m+1}^{n} f_i}$$

ii) Consider first the case $\rho = \dfrac{\sum_{i=m+1}^{n} f_i x_i}{\sum_{i=m+1}^{n} f_i} = \sum_{i=m+1}^{n} f_i x_i = 0$, then clearly:

$$\sum_{i=1}^{n} f_i (x_i - \mu)^2 \geq \sum_{i=1}^{n} f_i (\hat{x}_i - \mu)^2 \Leftrightarrow \sum_{i=m+1}^{n} f_i (x_i - \mu)^2 \geq \sum_{i=m+1}^{m} f_i (\hat{x}_i - \mu)^2$$

Observe that $\mu = \sum_{i=1}^{n} f_i x_i = \sum_{i=1}^{m} f_i x_i + \sum_{i=m+1}^{n} f_i x_i = \sum_{i=1}^{m} f_i x_i + 0 = \sum_{i=1}^{m} f_i x_i$:

$$\sum_{i=m+1}^{m} f_i (x_i - \mu)^2 = \sum_{i=m+1}^{m} f_i (x_i^2 - 2\mu x_i + \mu^2) = \sum_{i=m+1}^{m} f_i x_i^2 - 2\mu \sum_{i=m+1}^{m} f_i x_i + \mu^2 \sum_{i=m+1}^{m} f_i$$

$$= \sum_{i=m+1}^{m} f_i x_i^2 + \mu^2 \sum_{i=m+1}^{m} f_i \geq \mu^2 \sum_{i=m+1}^{m} f_i = \sum_{i=m+1}^{m} f_i (\hat{x}_i - \mu)^2$$

which establishes the result in the case $\rho = 0$. For the general case, let

$$\beta = \frac{\sum_{i=m+1}^{n} f_i x_i}{\sum_{i=m+1}^{n} f_i}, \quad y_i = x_i - \beta, \hat{y}_i = \hat{x}_i - \beta$$

then

$$\frac{\sum_{i=m+1}^{n} f_i y_i}{\sum_{i=m+1}^{n} f_i} = \frac{\sum_{i=m+1}^{n} f_i (x_i - \beta)}{\sum_{i=m+1}^{n} f_i} = \frac{\sum_{i=m+1}^{n} f_i x_i - \beta \sum_{i=m+1}^{n} f_i}{\sum_{i=m+1}^{n} f_i} = \frac{\sum_{i=m+1}^{n} f_i x_i - \sum_{i=m+1}^{n} f_i x_i}{\sum_{i=m+1}^{n} f_i} = 0$$

and we conclude that

$$\sum_{i=1}^{n} f_i y_i = \mu - \beta = \sum_{i=1}^{n} f_i \hat{y}_i \Rightarrow \sum_{i=1}^{n} f_i (\hat{x}_i - \mu)^2 = \sum_{i=1}^{n} f_i (\hat{y}_i - (\mu - \beta))^2 \leq \sum_{i=1}^{n} f_i (y_i - (\mu - \beta))^2 = \sum_{i=1}^{n} f_i (x_i - \mu)^2$$

completing the proof.