Survival Models for Actuarial Work

Stephen J. Richards

December 18, 2011

Abstract

This article gives an overview of survival models and their advantages for actuarial work, specifically for modelling mortality and longevity risk. Data collected and analysed by actuaries is particularly suitable for survival modelling — payment systems and policy-administration systems contain detailed individual data, and most portfolios are essentially a longitudinal study with continuous recruitment. The role of financial liability means that data is usually updated on at least as timely a basis as the statistics collected for medical trials, where survival models are routinely used. Survival models make the best use of available data, without the loss of information inherent in q_x models. By modelling at the level of the individual, the thorny actuarial problem of duplicates can be addressed at source, and not in the model itself. Furthermore, survival models can more easily cope with the challenges of competing risks than q_x models.

1 Introduction

This article was written at the request of the CMI, although it is not a CMI note and the information and opinions contained here are those of the author alone. The note outlines basic features of survival models, and why they are particularly useful for actuarial work. Since the subject of survival models is of wider interest to actuaries, an electronic copy of the article is freely available at www.longevitas.co.uk/site/informationmatrix/. References are given at the end of this article, while readers of the electronic version will find direct hotlinks to relevant background material on the Internet.

It is not the purpose of this note to document the various modelling methodologies used by the CMI, either past or present. However, we note in passing that the CMI has modelled the force of mortality, μ_x, for over three decades now, beginning with the graduation of the 80 series mortality tables[1].

Before proceeding to more detailed discussion of survival models, it is worth highlighting to actuaries that survival models are neither obscure, new or untested. Early work on estimating the survival function was done over six decades ago by Kaplan and Meier (1958)[6], a paper which ranks as one of the most commonly cited in scientific literature. Similarly, Cox (1972)[2] introduced the idea of the proportional hazards model, which is the cornerstone of parametric survival modelling. Survival models are widely used in modern medical statistics and experimental trials — see Collett (2003)[1]. The use of survival models is therefore widespread and uncontroversial.

The author has been studying and applying survival models to actuarial problems since 2006. This article covers the basic concepts of survival models, but it also highlights the specific features which make them relevant to actuaries.

2 Basic terminology and concepts

A survival model seeks to describe ${}_tp_x$, i.e. the probability of survival from age x to age $x+t$. This is described by non-actuaries as the *survivor function*, and is more usually denoted as $S_x(t)$, i.e. the probability of surviving t years from age x at outset.

Define ${}_hq_x$ as the probability that a life aged x dies before age $x+h$. The force of mortality is therefore defined as:

$$\mu_x = \lim_{h \to 0^+} \frac{{}_hq_x}{h} \tag{1}$$

A model for the continuous survival curve is directly equivalent to a model for the force of mortality, μ_x, due to the standard identity:

$${}_tp_x = S_x(t) = \exp\left(- \int_0^t \mu_{x+s} ds \right) \tag{2}$$

[1]Source: A. D. Wilkie, personal communication.

where the integral $H_x(t) = \int_0^t \mu_{x+s} ds$ is known to statisticians as the *integrated hazard function*[2].

An important feature of survival models is that they model mortality at the level of the individual, not of the group. An individual i enters observation at age x_i, and is observed for time t_i. An indicator variable, d_i, shows whether the event of interest has occurred at age $x_i + t_i$ or not: $d_i = 1$ for death and $d_i = 0$ for survival in the case of a mortality investigation, but d_i could be an indicator for a critical-illness claim, a policy lapse or any other demographic risk of interest.

The annual rate of mortality, q_x, can be exactly calculated from the force of mortality as follows:

$$q_x = 1 - \exp\left(-\int_0^1 \mu_{x+s}\right) ds \tag{3}$$

The converse does not hold true, however, as the force of mortality, μ_x, can only be approximated from the annual rate of mortality, q_x. A common simple approximation is as follows:

$$\mu_{x+\frac{1}{2}} = -\log_e (1 - q_x) \tag{4}$$

where $\log_e$ denotes a logarithm to base e.

3 Censoring

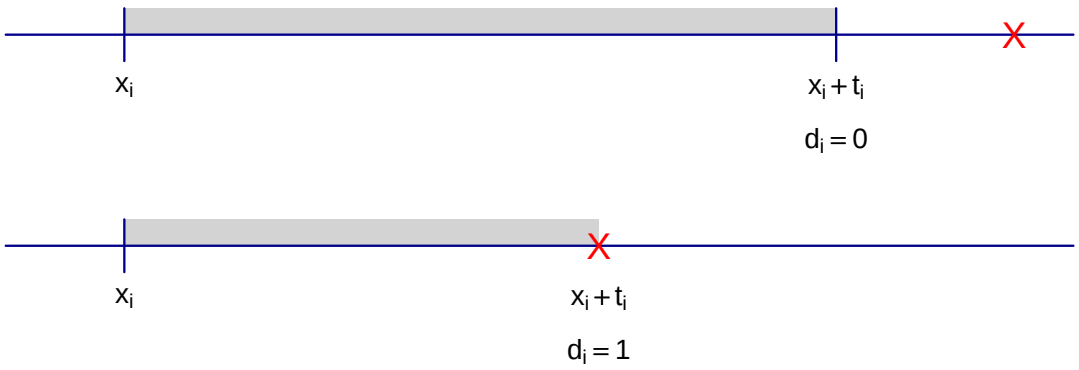
A defining feature of survival models is the concept of censoring, i.e. where incomplete information is available. Censoring is one of the reasons why survival analysis has to be approached differently from other statistical inference.

Censoring is where an individual is unobserved for a period of time. The most common kind of censoring in survival-model work is *right censoring*, i.e. where not everybody in a mortality study is dead at the point of analysis ($d_i = 0$). Another type of censoring is where an individual is known to have died before a particular date, but not precisely when. A third form of censoring is *interval censoring*. This is where the event is known to have occurred between two times, t_1 and t_2, but where precise knowledge of when the event occurred is unavailable. An example in a medical study would be where a participant last attended a consultation date at time t_1, but who did not attend a subsequent consultation at time t_2 and it was subsequently established that the participant had died. The observation for that participant would be interval-censored between t_1 and t_2. Interval censoring involves a loss of information, and also occurs when modelling q_x instead of μ_x, a subject we will return to in the next section.

A related but quite different phenomenon is *truncation*. Most data which actuaries encounter is left-truncated, because members of pension schemes and holders of insurance

[2]As might be inferred from the name of $H_x(t)$, statisticians refer to the force of mortality as the *hazard rate*. Engineers have a similar concept, which they call the *failure rate*.

Figure 1: Diagram of survival-model setup. The time observed, t_i, is shown in grey, while deaths are marked with a cross, $\times$. Since people do not usually enter into life-insurance contracts at birth, observations are left-truncated, i.e. lives start being observed at age $x_i > 0$. The upper case is an example of right-censored data as death happens after the end of the observation period. Figure reproduced from Richards (2008)[9].



contracts typically only enter a portfolio at an adult age[3]. The distinction can be explained as follows:

- A left-censored observation is where death is known to have occurred prior to a particular date, but the date is not known exactly.
- A left-truncated observation is where the exposure prior to the entry age cannot be used.

Left-truncation is a particular practical challenge for survival models, and in many software implementations only a handful of models can cope with left-truncation. The ability to deal with left-truncation is critical for actuarial work, and Richards (2008, 2010)[9][10] tabulates the integrated hazard functions for left-truncated models.

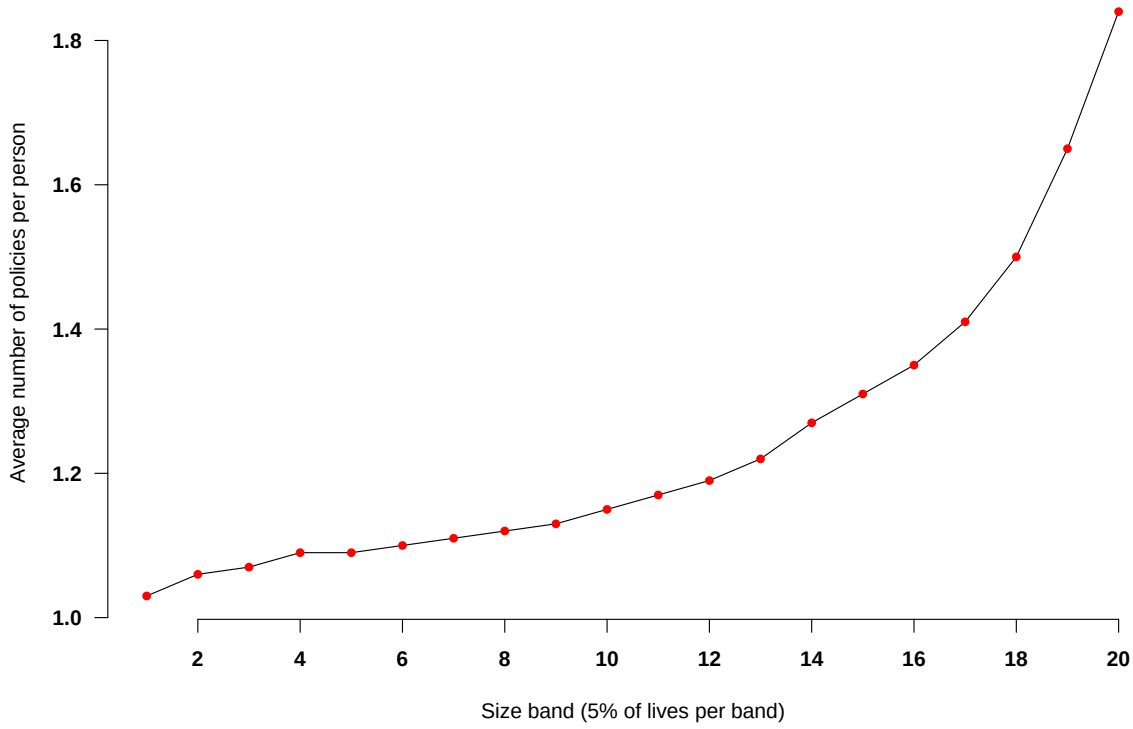
4 Assumptions

For a given individual i, the information pair $\{t_i, d_i\}$ is assumed to be independent of all other information pairs in the data set. An important difference in actuarial applications of survival models compared to elsewhere is the problem of *duplicates*. Actuaries usually deal with data which is held at the level of the insurance policy or benefit record, and individuals can and do have multiple policies, as shown in Figure 2. It is therefore important

[3]Children's pensions do exist in pension schemes, but the point still stands — actuaries do not tend to deal with data on people from birth onwards.

to identify individuals with more than one record, a process known as *deduplication*. Richards (2008)[9] outlines an approach to deduplicating policy-orientated data without requiring a single unique identifier for individuals.

Figure 2: Average number of policies per person in each of equal-sized membership bands ordered by total annual annuity income. Band 1 is the 5% of lives with smallest annual pensions, through to band 20 which is the 5% of lives with the largest annual pensions. Figure reproduced from Richards and Currie (2009)[11].



The ability to deduplicate prior to building any models is a key advantage of survival models — it addresses the problem of duplicates at source. In contrast, historical actuarial analyses dealt with counts which included duplicates, and much effort had to be devoted to restructuring models to cope with the resulting over-dispersion in the counts data — see Daw (1951)[3] and many other actuarial papers on this topic. The duplicate problem is critical when the tendency to have duplicate records is itself correlated with one or more of the risk factors under investigation — Figure 2 shows that the tendency to have duplicate policies is strongly correlated with income, which itself is a risk factor of considerable interest in actuarial work. Not only is it better to deal with the duplicate problem at source, it is in fact more effective to do so. Allowing for duplicates in count data usually involves the simplifying assumption that duplicates are randomly distributed with respect to risk factors, whereas Figure 2 shows that this is not the case.

For a given individual i, the survival time t_i is not independent of the status indicator d_i. It is important that t_i is independent of any mechanism which causes the survival time to be censored (i.e. $d_i = 0$). Censoring when t_i depends on the censoring mechanism is called *informative censoring* and it invalidates the basic criteria for any survival model. This sort of censoring can occur in drug trials, e.g. if treatment were withdrawn from a study participant due to a deterioration in his or her health. However, informative censoring is uncommon in actuarial work.

5 Survival models versus Binomial count models

A survival model is concerned with the future lifetime, T_x, of an individual aged x. T_x is a random variable, and if we are modelling μ_x then T_x is a continuous random variable.

In contrast, other statistical models are concerned with *counts*, often of grouped data. For example, consider the number of deaths in one year amongst a group of E_x identical lives aged x, D_x. D_x can be modelled as a random counting variable with Binomial distribution, i.e. $D_x \sim B(E_x, q_x)$, where inference for q_x follows from the observed number of deaths d_x.

This distinction — modelling survival time instead of modelling counts — introduces one of the advantages of survival models over q_x models, namely fuller use of available data. By modelling counts of events with q_x, we lose information as to when the events actually took place. To illustrate this loss of information, consider the following example from Richards (2008)[9]. Two groups, A and B, each consist of four identical lives alive at the start of the year. During the course of the year one life dies in each group, making the estimated mortality rate, $\hat{q}_A = \hat{q}_B = \frac{1}{4}$ in both cases. However, if the death in group A occurs at the end of January, the estimated force of mortality is $\hat{\mu}_A = \frac{1}{3\frac{1}{12}} = \frac{12}{37}$. If the death in group B occurs at the start of December the estimated force of mortality is $\hat{\mu}_B = \frac{1}{3\frac{11}{12}} = \frac{12}{47}$. As this simple example shows, working with the force of mortality means we can use all the information available. In contrast, working with q-type rates loses information on time of death and is therefore less sophisticated.

This greater efficiency of models for μ_x confers a benefit over q_x models in all scenarios, although the benefit will be modest in most cases when a data set is very large. An exception is likely to be where the modelled intensities are very high, where most of the useful information will be contained in the times of events. A simple example would be a care-annuity portfolio, where life expectancy might be of the order of around three years. Where the expected time to an event is relatively short, the information loss in using q_x could be considerable.

6 Survival models versus Poisson count models

We note that a model for μ_x is not synonymous with a survival model — we could also have a Poisson model for D_x based on the central exposure, E^c_x, i.e. $D_x \sim \text{Poisson}(E^c_x \mu_x)$. There are several results linking Poisson count models to survival models. Indeed, some of

the so-called *accelerated failure-time* survival models are fitted using the GLM algorithm for Poisson counts after a suitable data transformation.

It is sometimes the case that data is sometimes only available in an aggregated form[4]. Under such circumstances a survival model for individuals is out of the question and a count-based model is the only possible solution. When the data dictate a counts-based model, it is usually preferably to use a Poisson model, rather than a Binomial model.

7 The Kaplan-Meier estimator of the survival function

An important tool in survival analysis is the estimator of the survival function introduced by Kaplan and Meier (1958)[5]. The main features of this are as follows:

- It is a non-parametric model — the Kaplan-Meier estimator of the survival curve requires no parameters to be estimated.
- It is based around q-type rates, but where the interval over which q applies varies along the curve and is determined by the actual data *ex post*, rather than by the analyst building the model.
- As a non-parametric method it can only model the mortality of sub-groups by stratifying the data and building separate curves.
- Although the Kaplan-Meier survival function is non-parametric, it is still a statistical estimator of the survival function and so confidence intervals can be derived.

One wrinkle for actuaries is that the standard Kaplan-Meier estimator is typically defined with reference to the time since a medical study commenced. In actuarial work it makes more sense to define the Kaplan-Meier estimator with respect to age. The following definition from Richards (2010)[10] will work for any portfolio whether it is closed or open to new business:

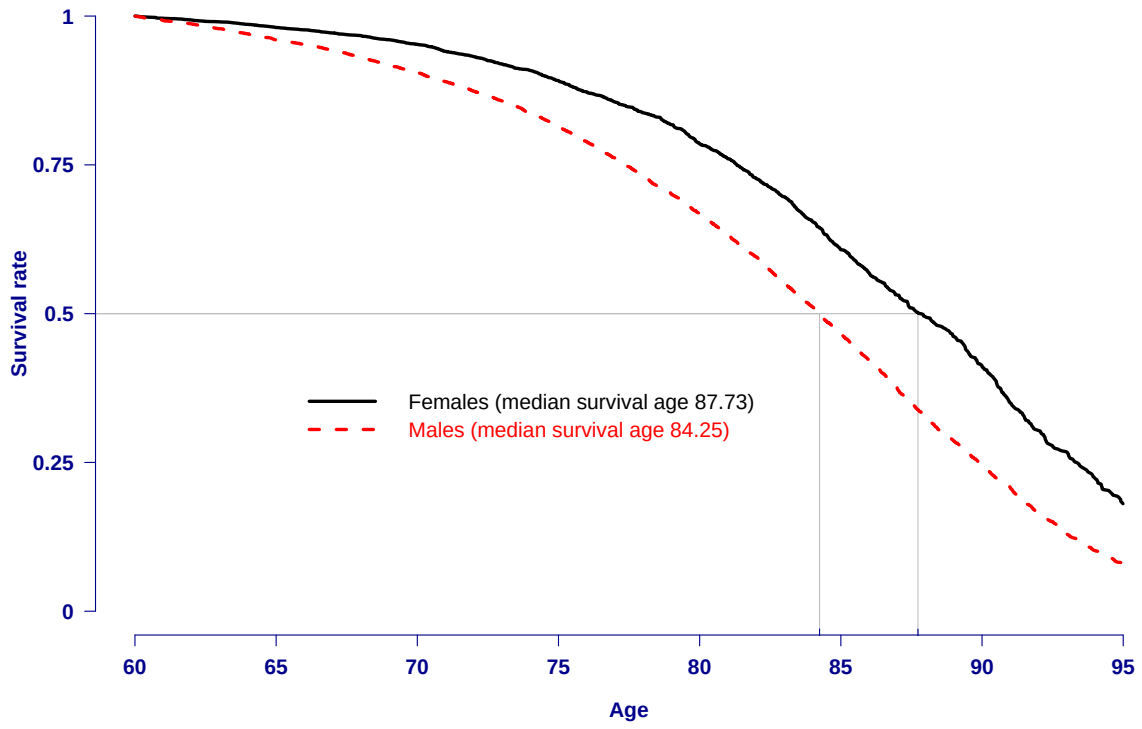
$$ {}_{t_j}p_x = \prod_{i=1}^{j \leq n} \left(1 - \frac{d_{x+t_i}}{l_{x+t_i^-}} \right) \tag{5} $$

where x is the outset age for the survival curve, $\{x+t_i\}$ is the set of n distinct ages at death, $l_{x+t_i^-}$ is the number of lives alive immediately before age $x+t_i$ and d_{x+t_i} is the number of deaths occurring at age $x+t_i$. Implicit in Equation 5 is that the data are ordered by death ages, and that such an approach is only possible with individual-level data, not aggregated data. An example of the Kaplan-Meier estimator is given in Figure 3. The form of Equation 5 explains the alternative description of the Kaplan-Meier estimator as the *product-limit estimator* of the survival function.

[4] Aggregated data is referred to by the CMI as "scheduled data".

[5] A related estimator is due to Fleming and Harrintgon (1991)[4], which produces a survivor function higher than the Kaplan-Meier estimator. The difference between the two estimators is negligible for the size of data sets actuaries typically work with. Note that what Collett (2003)[1] on pages 20–22 refers to as the Nelson-Aalen[8] estimator for the survivor function is in fact the Fleming-Harrington estimator.

Figure 3: Kaplan-Meier survival curves as per Equation 5. Figure reproduced from Richards (2010)[10].



Note that Kaplan-Meier estimator conceptually straddles the concepts of q_x and μ_x. The definition in Equation 5 is clearly based around q_x, but where the discretization is decided by the data *a posteriori*, rather than by the analyst *a priori*. As the number of events in life-office portfolios is typically rather large, the discretisation steps can be quite small and the results quickly look like μ_x due to the relationship in Equation 1. For example, the median age gap for the male lives in Figure 3 is one day, which is the smallest interval possible when using dates to measure survival times (the largest gap is 11 days). For this reason a purist might argue that models fitted in practice are actually for q_x with a daily interval, i.e. ${}_{\frac{1}{365}}q_x$, rather than for μ_x.

A major limitation of the Kaplan-Meier estimator is that it can only model the mortality of sub-groups by stratification, i.e. a separate model for each sub-group. If we have a number of risk factors which we expect to operate independently — say gender, smoker status and distribution channel — then with Kaplan-Meier we have to sub-divide the data into as many sub-groups as there are combinations of risk factors. This quickly becomes a problem for even very large data sets, as some combinations will be much rarer than others. This leads us to discussion of parametric survival models.

8 Parametric survival models — proportional hazards

For actuarial purposes, a far more useful class of survival models is parametric models, i.e. models with one or more parameters to be estimated from data. Such models avoid the stratification problem of the Kaplan-Meier model, and allow a single overarching model to be fitted to the entire data set.

Cox (1972)[2] introduced the idea of a *proportional-hazards model*, where the mortality of an individual i is a proportion e^{α_i} of the baseline hazard, $\mu_0(x)$:

$$\mu_i(x) = e^{\alpha_i}\mu_0(x) \tag{6}$$

At its simplest, the values of α_i might take one of only two values: $\alpha_i = 0$ if individual i belonged to the baseline group, and $\alpha_i = a_{male}$ if the individual belonged to a different group. An example would be a mortality model where females were the baseline ($\alpha_i = 0$ for females) and males had mortality which was a constant proportion of the female hazard ($\alpha_i = a_{male}$ for males). The parameter a would be estimated, and would be the effect of being male on mortality. The interesting thing about the Cox model is that it does not require the estimation of the baseline hazard function, $\mu_0(x)$, and so lends itself to answering the question of whether one group has significantly different mortality from another. It is for this reason that the Cox model is widely used in medical research, since much drug testing is concerned with whether the treatment group has higher survival rates than the control group, and the analyst is not so concerned with the precise shape of the basic hazard function. The Cox model can be expanded to include any number of covariates, however, simply by constructing an individual's α_i from any number of j risk factors, a_j:

$$\alpha_i = \sum_j z_{i,j}a_j \tag{7}$$

where $z_{i,j}$ is an indicator variable taking the value 1 when an individual possesses the risk factor and the value 0 when the individual does not. An example would be where not just the effect of being male was to be measured (a_{male}) but an analyst also wanted to control for smoker status (a_{smoker}). Now the baseline group is non-smoking females, and although there are four combinations, only two parameters need to be estimated: one for the effect of being male and one for the effect of smoking. Here we see the power of parametric survival models over the likes of the Kaplan-Meier estimator, as stratification is not required and all data can be used in a single model. In the case of this author — a non-smoking male — in addition to the survival time, t, and status indicator, d, the additional data would be as follows:

$$z_{male} = 1 \tag{8}$$

$$z_{smoker} = 0 \tag{9}$$

The author's wife is a non-smoking female, so the additional data for her in a Cox model would be:

$$z_{male} = 0 \tag{10}$$
$$z_{smoker} = 0 \tag{11}$$

Where all the $z_{i,j}$ values are zero, then that individual belongs to the baseline group. It does not matter which group is treated as the baseline. The advantage of the approach outlined in Equations 6–10 is that it can be generalised to an unlimited number of risk factors, with no reduction in power due to stratification. However, the disadvantage of the pure Cox model of Equation 6 is that actuaries are very concerned indeed with the shape of the baseline hazard function, $\mu_0(x)$. Medical trials and research might just be concerned with whether risk factors exist and are significant, but in actuarial work the precise shape of the survival curve and the timing of mortality is critical.

9 Parametric survival models — age-varying hazards

It is relatively straightforward to extend survival models to include simultaneous estimation of the baseline hazard. A simple example would be the Gompertz (1825)[5] law for the force of mortality:

$$\mu_i(x) = e^{\alpha_i + \beta_i x} \tag{12}$$

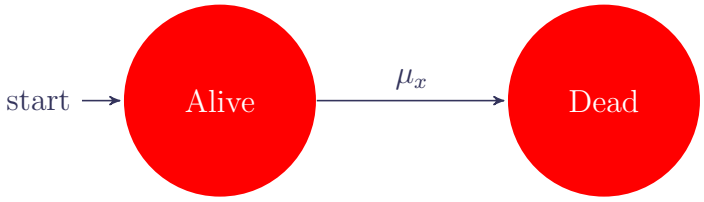
where α_i is constructed identically to Equation 7 to account for main effects which raise or lower the mortality curve. β_i is constructed similarly, and allowed for age-varying rates of mortality where the slope can also depend on the risk factors being modelled. The Gompertz model is just one of many possible alternatives, and there is a wide choice of shapes for the basic hazard function. However, estimation of the parameters in such models is trickier due to the left-truncation which accompanies almost every actuarial data set. Richards (2010)[10] details sixteen of the most commonly available parametric survival models, and compares their suitability for modelling the mortality of pensioners and annuitants. Formulae are given for both the force of mortality and the integrated hazard function for truncated data.

10 Competing risks and multiple-state models

So far we have implicitly considered modelling a single decrement whose occurrence ends the risk, e.g. a pension or annuity contract. This is a model with two-states — alive and dead — and there is no return to the first state once the event of interest has occurred. In terms of Markov models, this would be described as a "two-state model without return", and is depicted in Figure 4.

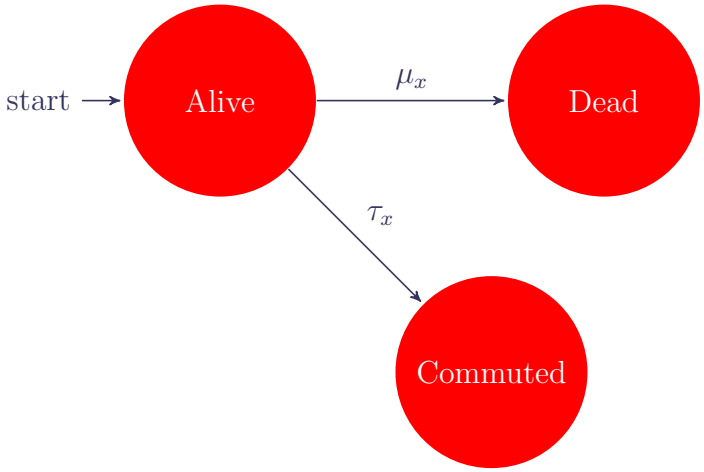
A major area of benefit from modelling μ_x in a survival model comes from the analysis of competing decrements, as illustrated in Figure 5. It is just as simple to use a survival model for μ_x in the case of Figure 5 as for Figure 4, although a q_x model would be much

Figure 4: Two-state model without return, e.g. longevity risk within a pensioner or annuitant population.



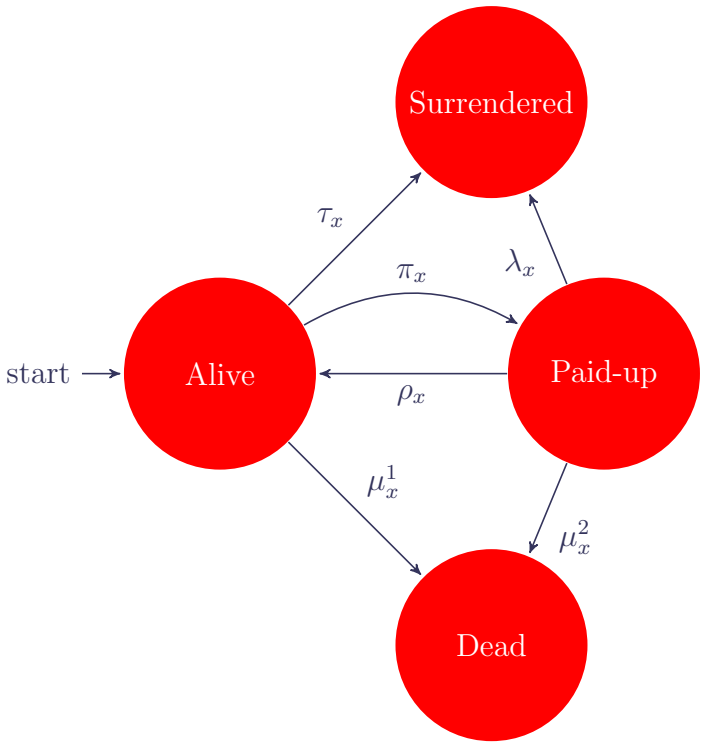
trickier and would require further artificial — and often unrealistic — assumptions regarding the distribution of competing events over the year and the independence of the risks. The complexities of q_x models increase as the number of competing risks increases.

Figure 5: Three-state model without return, e.g. longevity risk within a pensioner or annuitant population but where some beneficiaries can commute pensions on the grounds of triviality. Note that it is logically possible to move from the Commuted state to Dead. However, such transitions will not be observed after commutation, so this transition cannot be modelled and is therefore not shown.



However, the advantages of survival models — and in modelling in continuous time — become overwhelming when we consider risk models where a life may return to a previous state. This is depicted in Figure 6, which illustrates the risks for a whole-of-life insurance policy. However, there are other cases with more states and more possible transitions. The interested reader is directed to Macdonald (1996)[7] for a more detailed treatment.

Figure 6: Four-state model with return, e.g. a whole-of-life assurance policy



11 Conclusions

Survival models lend themselves particularly well to actuarial work. Payment systems and policy-administration systems contain detailed individual data, and most portfolios are essentially a longitudinal study with continuous recruitment. The role of financial liability means that data are usually updated and maintained on at least as timely a basis as the statistics collected for medical trials, where survival models are heavily used.

Survival models make the best use of available data, without the loss of information inherent in q_x models. By modelling at the level of the individual, the problem of duplicates can be addressed at source, and not in the model itself. Furthermore, survival models can more easily cope with the challenges of competing risks than q_x models.

Further mini-articles on aspects of survival models in actuarial work can be found on the author's blog at http://www.longevitas.co.uk/site/informationmatrix/.

References

[1] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall, 2003.

[2] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 24:187–220 (with discussion), 1972.

[3] R. H. Daw. Duplicate policies in mortality data. *Journal of the Institute of Actuaries*, 77:261–267, 1951.

[4] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. Wiley, New York, 1991.

[5] B. Gompertz. The nature of the function expressive of the law of human mortality. *Philosophical Transactions of the Royal Society*, 115:513–585, 1825.

[6] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[7] A. S. Macdonald. An actuarial survey of statistical models for decrement and transition data, ii: competing risks, non-parametric and regression models. *British Actuarial Journal*, 2(2):429–448, 1996.

[8] W. Nelson. Theory and applications of hazard plotting for censored failure times. *Technometrics*, 14:945–965, 1958.

[9] S. J. Richards. Applying survival models to pensioner mortality data. *British Actuarial Journal*, 14(II):257–326 (with discussion), 2008.

[10] S. J. Richards. A handbook of parametric survival models for actuarial use. *Scandinavian Actuarial Journal, (to appear)*, 2010.

[11] S. J. Richards and I. D. Currie. Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal*, 15(II) No. 65:317–365 (with discussion), 2009.