MATH96048/MATH97075/MATH97183
Survival Models

Professor Axel Gandy

2022/23

These notes are based on earlier versions of the course, incorporting notes of Professor Nicholas Heard and Dr David Whitney.

1 Principles of Modelling Time to Event Data

1.1 Statistical Modelling

A stochastic or statistical model of a system is a mathematical model which represents inherent uncertainties in the system as random variables. Any model which does not make such allowances for uncertainty, on the other hand, is said to be deterministic. In actuarial science, statistical models are especially useful for dealing with uncertainty in the time until a specific event will occur, such as predicting the number of premium payments that will be received before paying out on a life assurance contract. These statistical models for time to event data also have much wider applications in fields such as medicine, biostatistics and engineering.

Mathematical models are an imperfect representation of reality. Their utility comes from being able to approximately learn the consequences of hypothetically changing certain experimental inputs or actions. Statistical models extend this utility by capturing the uncertainty surrounding unknown future outcomes, providing the possibility of searching for an optimal decision under this uncertainty.

1.2 Model Choice

1.2.1 Complexity

The complexity of a good statistical model is often constrained for several reasons. The analyst might be restricted by the inputs for which data are readily available, or by computational or statistical limitations to the models which can be reliably fitted. Additionally, it is important that the chosen model satisfies the purposes for which it will be used; usually this means that the mathematical structure of the model need be easily interpretable for purposes of understanding and communication, and that the resulting fitted model will not overfit and understate uncertainty and risk. For these reasons, the statistical analyst will often seek to fit the most parsimonious model that the data will sensibly allow.

1.2.2 Sensitivity

Model choice and fitting should be viewed as an iterative procedure. Once a particular model has been fitted to the available data, the quality of the fit should be inspected; hypothesis tests such as goodness of fit tests can determine the suitability of the selected model. If the structure of the data is not well captured by the model, this may suggest that the model chosen is inadequate and the analyst should return to the model selection stage to consider other alternatives.

Note however that performance of the chosen model will also depend heavily on the quality of the data, with poor quality inputs likely to lead to unreliable output inference (*garbage in, garbage out*). If there are questions about the accuracy of the data being used, then it becomes particularly important to carry out a sensitivity analysis. This investigates the magnitude of change in the model outputs when small perturbations are applied to the model inputs.

Finally, it should be noted that the suitability of a well fitted model may not be comfortably relied upon outside the range of the data used; for example, an exponential relationship can appear fairly linear in the short term, before ballooning away from such a fit in the longer term.

2 Distributions of Event Times

2.1 Random Variable Modelling

Consider a homogeneous population of individuals, who each have an associated event time which is initially unknown and treated as a random variable. We will often refer to the period of time until the event occurs as the lifetime of the individual.

Throughout the course, let T denote the future lifetime of a new-born individual (aged 0). Unless otherwise stated, we shall assume T is a continuous random variable which takes values on the positive part of the real line $\mathbb{R}^+ = [0, \infty)$, with associated probability measure P.

More specifically, in this chapter we might assume the existence of a limiting age ω which T cannot exceed, so $T \in [0, \omega]$. For human life calculations, typically we take $\omega \approx$120 years.

We first define the cumulative distribution function, F, the survivor function S, the density f, the hazard h and the cumulative hazard H for a lifetime, and derive relationships between them.

This will provide us with a range of tools for specifying the distribution of a lifetime, and rules for moving between them.

2.2 Cumulative Distribution and Survivor Functions

Definition: The cumulative distribution function (CDF) of T is

$$F(t) = \mathrm{P}(T \leq t),$$

the probability of death by age t.

Definition: The survivor (or reliability) function of T is

$$S(t) = \mathrm{P}(T > t) \quad = \quad 1 - F(t),$$

the probability of surviving beyond age t.

2.2.1 Criteria for a valid survivor function:

$S(t)$ is a valid survivor function iff

1. $S(0) = 1, \lim_{t \to \infty} S(t) = 0$
2. $0 \leq S(t) \leq 1, \forall t \in \mathbb{R}^+$;
3. Monotonicity: $\forall t_1, t_2 \in \mathbb{R}^+, t_1 < t_2 \implies S(t_1) \geq S(t_2)$;
4. S is càdlàg: $\forall t \in \mathbb{R}^+$,
   (a) $S(t^+) \equiv \lim_{u \downarrow t} S(u) = S(t)$ (S is right-continuous);
   (b) $S(t^-) \equiv \lim_{u \uparrow t} S(u)$ exists. (left limits exist for S).

2.2.2 Future lifetime after age x

Definition: Let T_x be the future lifetime of an individual who has survived to age x, for $0 \le x \le \omega$, so $T_x \in [0, \omega - x]$. So clearly

- $T_0 \equiv T$;
- The distribution of T_x is the same as $T - x | T > x$, in other words: for all measurable sets A,
$$\mathrm{P}(T_x \in A) = \mathrm{P}(T - x \in A | T > x).$$

Definition: The cumulative distribution and survivor functions of T_x are

$$\begin{aligned} F_x(t) &= \mathrm{P}(T_x \le t), \\ S_x(t) &= \mathrm{P}(T_x > t) = 1 - F_x(t). \end{aligned}$$

2.2.3 Relationship between T_x and T

For consistency with T, for the CDF we have

$$\begin{aligned} F_x(t) &= \mathrm{P}(T_x \le t) = \mathrm{P}(T \le x + t | T > x), \\ \Longrightarrow F_x(t) &= \frac{F(x+t) - F(x)}{S(x)} \end{aligned}$$

and for the survivor function,

$$S_x(t) \qquad = 1 - F_x(t) = \frac{S(x+t)}{S(x)}.$$

2.3 Density Function

Definition: The probability density function (PDF) of the random variable T_x is

$$f_x(t) = \frac{d}{dt} F_x(t) = \lim_{h \downarrow 0} \frac{F_x(t+h) - F_x(t)}{h}.$$

Since we are assuming T_x is a continuous random variable, by definition this density exists.

For individuals who have currently survived to age x, $f_x(t)$ is the rate of death t further units of time into the future. That is, for such an individual, the probability of death within the interval $[t, t+h]$ for a small interval width h is approximately $f_x(t)h$.

2.4 Hazard Function / Force of Mortality

The hazard function plays a central role in survival analysis. We denote the hazard function (or force of mortality) at age x, $0 \le x \le \omega$, by $h(x)$.

Definition: The hazard function of T is defined as

$$h(x) = \lim_{h \downarrow 0} \frac{\mathrm{P}(T \le x + h | T > x)}{h}.$$

We will always assume this limit exists.

Note

$$h(x) = \lim_{h \downarrow 0} \frac{F_x(h)}{h} = f_x(0).$$

The interpretation of $h(x)$ is important. It represents the instantaneous death rate for an individual who has survived to time x.

Or, approximately, for small h

$$\mathrm{P}(T \leq x + \Delta | T > x) \approx \Delta h(x). \tag{1}$$

Given an individual has reached age x, the probability of death in the next short period of time of length Δ is roughly proportional to Δ, the constant of proportionality being $h(x)$.

Important result

$$h(x+t) = \frac{f_x(t)}{S_x(t)},$$

or alternatively

$$f_x(t) = h(x+t)S_x(t).$$

So in particular, for $x = 0$ we have

$$f = hS.$$

Proof

$$\begin{aligned}
f_x(t) &= \lim_{h \downarrow 0} \frac{1}{h} \left\{ \mathrm{P}(T_x \leq t+h) - \mathrm{P}(T_x \leq t) \right\} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \left\{ \mathrm{P}(T \leq x+t+h | T > x) - \mathrm{P}(T \leq x+t | T > x) \right\} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \frac{\{F(x+t+h) - F(x)\} - \{F(x+t) - F(x)\}}{S(x)} \\
&= \lim_{h \downarrow 0} \frac{1}{h} \frac{F(x+t+h) - F(x+t)}{S(x)} \\
&= \frac{S(x+t)}{S(x)} \lim_{h \downarrow 0} \frac{1}{h} \frac{F(x+t+h) - F(x+t)}{S(x+t)} \\
&= \frac{S(x+t)}{S(x)} \lim_{h \downarrow 0} \frac{1}{h} \mathrm{P}(T \leq x+t+h | T > x+t) \\
&= \frac{S(x+t)}{S(x)} h(x+t) \\
&= S_x(t)h(x+t). \qquad \square
\end{aligned}$$

Summary

- T_x is a continuous random variable denoting the random future lifetime of an individual alive at age x.
- The distribution function $F_x(t)$is defined on $[0, \omega - x]$ with PDF $f_x(t) = F_x'(t)$.

- The survivor function and the PDF provide two methods of specifying the distribution of a continuous random variable.
- An additional way was provided via the hazard function

$$h(x+t) = \frac{f_x(t)}{S_x(t)} = \frac{-S_x'(t)}{S_x(t)}.$$

Why additionally consider the hazard rate when we have $F(t)$ and $f(t)$?

(i) It may be physically enlightening to consider the immediate risk.

(ii) Comparisons of groups of individuals are sometimes most incisively made via the hazard.

(iii) Hazard-based models are often convenient when there is censoring.

(iv) When fitting parametric models the form of the hazard function can be enlightening about the assumptions made by the model: e.g. Exponential $\Longrightarrow$ constant hazard.

(v) The hazard rate does not need to satisfy as many conditions as pdf/CDF.

2.5 Cumulative Hazard Function

Definition: The cumulative (or integrated) hazard rate of T_x, denoted $H_x(t)$, is simply

$$H_x(t) = \int_0^t h(x+s)ds.$$

This leads to another important relationship,

$$S_x(t) = \exp\{-H_x(t)\}.$$

Recall,

$$\begin{aligned} h(x+t) &= \frac{-S_x'(t)}{S_x(t)} \\ &= -\frac{d}{dt}\log S_x(t). \end{aligned}$$

Furthermore we have the condition $S_x(0) = 1$ from which $S_x(t) = \exp\{-H_x(t)\}$ follows.

Note,

$$f_x(t) = h(x+t)\exp\{-H_x(t)\}.$$

2.6 Censoring

A defining feature of survival data, which renders many standard statistical analysis methods inappropriate, is that survival times are frequently censored.

A survival time is said to be censored if the exact event time (e.g. time of death) has not been observed. Such observations provide only partial information about the value of T.

There are a number of common censoring mechanisms that prevent observation of some event times.

2.6.1 Right-censoring

An event time is right-censored if the censoring mechanism prematurely terminates observation of the individual, before the event has actually occurred. For example, if we lose track of an individual, or the study comes to an end. We only learn that $T > t$ for some $t \in (0, \omega)$.

If it is known beforehand that right-censoring is due to happen at time c for all events which have not yet occurred, then the time spent observing the individual $T_{\text{obs}} = \min(T, c)$ is a random variable of mixed type. That is, T_{obs} has continuous distribution over the interval $[0, c)$ and then a discrete atom of mass at time c.

2.6.2 Left-censoring

An event time is left-censored if we discover the event occurred before observation of the individual began; alternatively, it could be that we observe the event time but only have a lower bound on when the life of the subject began. In either case, all we can say is that the actual time to event T is less than some time t. For example, we begin a study of patients three months after an operation and find that some have died. All we know is that $T < 3$ months for those patients.

2.6.3 Interval-censoring

When all that is known is that the event occurred within an interval. For example, if we check the status of individuals every d days.

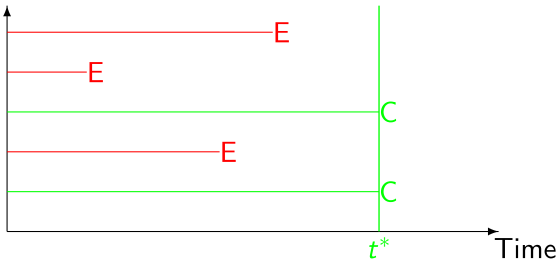
Although always undesirable, a censoring mechanism may inevitably be introduced through the design of an experiment or data collection process.

2.6.4 Truncation

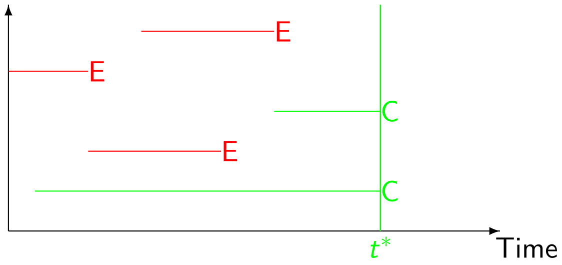
Truncation happens when we wish to estimate the distribution of T but some of our data is sampled instead from a conditional distribution such as $T | a < T \leq b$. Truncation is somewhat similar to censoring but is not a type of censoring.

2.6.5 Type I censoring

Type I censoring occurs if we take n individuals and observe them for a pre-specified time t^*. Any non-events are right censored with censoring time t^*. Commonly found in medical studies, e.g. follow 15 patients for two years after operation.



- Generalised Type I censoring: As above, although now individuals join the study at different times ⇒ even censored observations have different durations on the study.



2.6.6 Type II censoring

Where we take n individuals and observe them until the d^{th} event occurs. The remaining non-events are right-censored, although (unlike Type I) this censoring time is not known in advance. Commonly found in reliability analysis, e.g. take 20 components and keep testing until half of them fail.

2.6.7 Competing risks

Suppose we are interested in the marginal distribution of failure time T_1 due to a particular cause, but the occurrence of a separate event, a competing risk, at random time T_2 would cause the individual to exit the study. Then we will only be able to observe $\min(T_1, T_2)$, along with the type of failure. Observations where $T_2 < T_1$ will constitute right-censored observations of T_1.

One important assumption made throughout this course is that there is independence between the censoring mechanism and the event time. That is, if C denotes a (possibly deterministic) censoring time, and T is the (perhaps unobserved) event time, for simplicity of inference we assume T and C are statistically independent random variables.

3 Parametric Distributions of Random Lifetimes T

3.1 Introduction

In Chapter 2 we defined various mathematical expressions that could be used to describe the probability distribution of a future random lifetime T. But until now, we have not considered what forms these quantities might take.

In this chapter we shall meet some standard parametric distributions that are commonly used in survival analysis.

Clearly, *any* distribution over the positive half-line $\mathbb{R}^+$ is a possible candidate. Moreover, distributions on $\mathbb{R}$ can be models for $\log T$. However, a number of standard distributions have emerged as being particularly appropriate for the task of analysing survival data.

We shall consider the three most popular models, these being the exponential, the Weibull and the Gompertz-Makeham distributions. Each will make a different assumption about the nature of the hazard rate.

The parametric models in this chapter do not admit a limiting age ω, all three have support over the whole of $\mathbb{R}^+$. If we were to truncate these distributions to lie on $[0, \omega]$, we should be mindful of the induced changes in the nature of the hazard functions.

3.2 Exponential Distribution

The exponential distribution is a natural starting point as a lifetime distribution. For $\lambda > 0$, if $T \sim \text{Exp}(\lambda)$ then

$$\begin{aligned} f(t) &= \lambda \exp(-\lambda t), \\ S(t) &= \exp(-\lambda t), \\ h(t) &= \lambda, \\ H(t) &= \lambda t. \end{aligned}$$

The constant hazard function reflects a property known as *lack of memory*; for the exponential distribution,

$$S_x(t) = S(t).$$

In practice however, the assumption of a constant hazard is often untenable and when fitting the exponential distribution to data it can be sensitive to outliers.

The other two distributions we now consider can be seen to generalise the exponential distribution; that is, each contains the exponential distribution within their family as a special case.

3.3 Weibull Distribution

The Weibull distribution can be written as

$$\begin{aligned} f(t) &= \eta \alpha^{-\eta} t^{\eta-1} \exp(-(t/\alpha)^\eta), \\ S(t) &= \exp(-(t/\alpha)^\eta), \\ h(t) &= \eta \alpha^{-\eta} t^{\eta-1}, \\ H(t) &= (t/\alpha)^\eta, \end{aligned}$$

where $\alpha > 0$ and $\eta > 0$ are known as the scale and shape parameters respectively.

The Weibull generalises the exponential distribution, which is recovered when $\eta = 1$.

Figures 1 and 2 show the hazard and density functions for different values of the shape parameter.

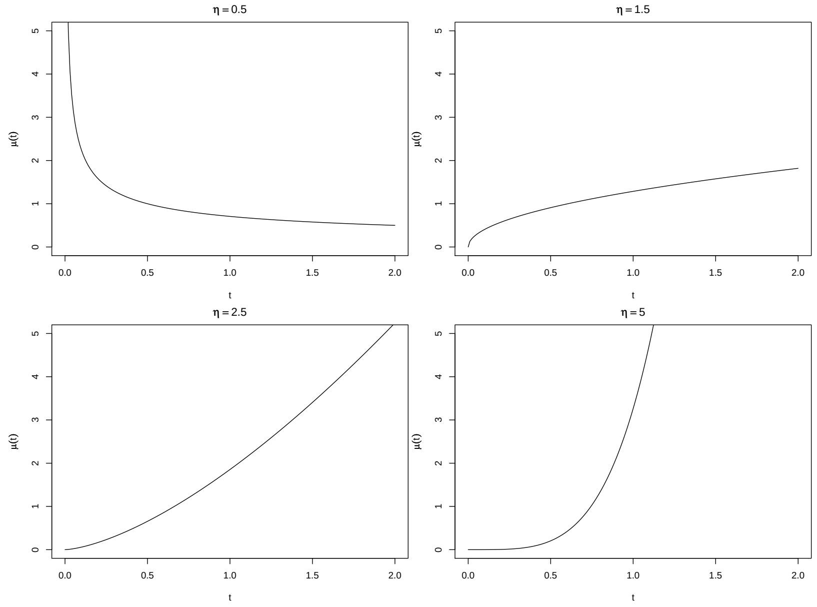


Figure 1: Hazard functions for Weibull with mean 1 and shape parameter η.

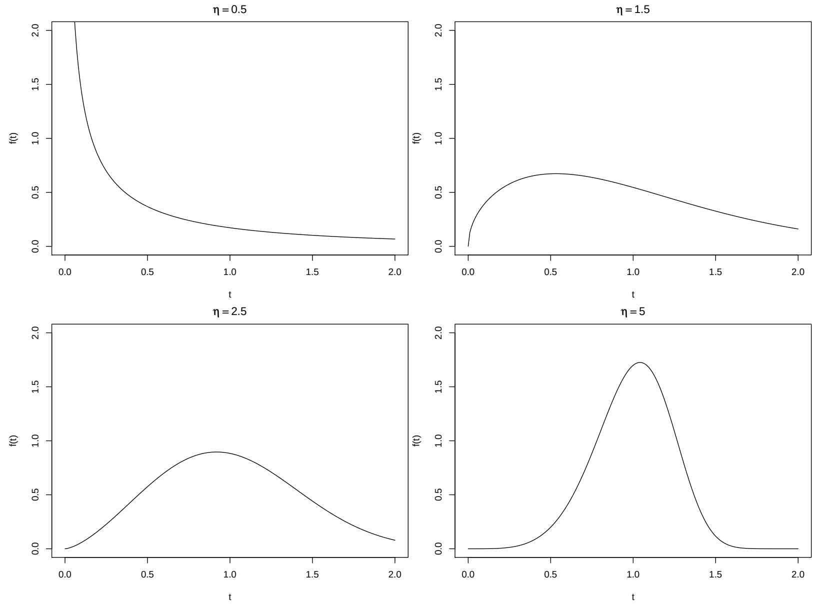


Figure 2: Density function for Weibull with mean 1 and shape parameter η.

The mean and variance of the Weibull density are

$$\alpha\Gamma(\eta^{-1}+1) \quad \text{and} \quad \alpha^2\{\Gamma(2\eta^{-1}+1) - [\Gamma(\eta^{-1}+1)^2]\}$$

respectively, where Γ is the gamma function $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$.

For $\eta < 1$ we have a monotonically decreasing (antitonic) hazard rate and for $\eta > 1$ we have a monotonically increasing (isotonic) hazard rate (see Fig. 1).

In particular for $1 < \eta < 2$ the hazard increases slower than linear in t; for $\eta = 2$ the increase is linear and for $\eta > 2$, faster than linear.

The Weibull distribution is probably the most widely used parametric distribution in survival analysis. Some possible reasons:

- Simplicity of the survivor and hazard functions.
- Covers a wide variety of distributional shapes (see Fig. 2).
- Empirically it has been found to be accurate in many contexts. It is an *extreme value distribution*.

3.4 Gompertz-Makeham Distribution

3.4.1 Gompertz and Makeham laws of mortality

The Gompertz law of mortality states that in a low mortality environment where external causes of death are rare, the force of mortality increases approximately exponentially with age. Empirical data gathered from observation of insects kept in laboratory conditions support this hypothesis.

Outside of such an environment, the Makeham law of mortality considers the risk of other, external causes of death to be approximately constant with age.

So together these laws suggest a hazard rate which is a sum of constant (Makeham) and exponential (Gompertz) terms. This motivates the Gompertz-Makeham distribution.

3.4.2 Gompertz-Makeham hazard function

The Gompertz-Makeham distribution is a three parameter distribution with hazard function

$$h(t) = \theta + \beta \exp(\gamma t)$$

for non-negative parameters θ, β, γ.

This distribution again generalises the exponential distribution. The survivor function and density function can be derived in the usual way (Exercise).

3.5 Choosing a Distribution

We have listed three common parametric distributions for lifetime random variables (there are many more: gamma, log-normal, log-logistic,. . .).

Given a particular data set of, say, n realisations of T, how do we decide which distribution is appropriate?

Each distribution makes particular assumptions about the form of the hazard. Knowledge of the ‘true’ hazard rate would enable us to decide which distribution was ‘closest’. In practice we can compare to a non-parametric estimate of the hazard which converges to the truth.

3.5.1 Empirical survivor function

Suppose we have gathered some survival time data for n individuals drawn from a population with common survivor function $S(t)$. Assume, for now, that there are no censored observations. Then the empirical survivor function

$$\hat{S}(t) = \frac{\text{number of observations} > t}{n}$$

provides an estimate of $S(t)$ derived purely from the data. Notice that $\hat{S}(t)$ is an antitonic step function with jumps at the death times. ($\hat{S}(t)$ is a *non-parametric* estimator of $S(t)$. In Chapter 5 we shall see how to estimate S in the presence of censoring.)

Since $S(t) = \exp\{-H(t)\}$, the simple transformation

$$\hat{H}(t) = -\log \hat{S}(t)$$

then provides a similar, data-based estimate of the cumulative hazard.

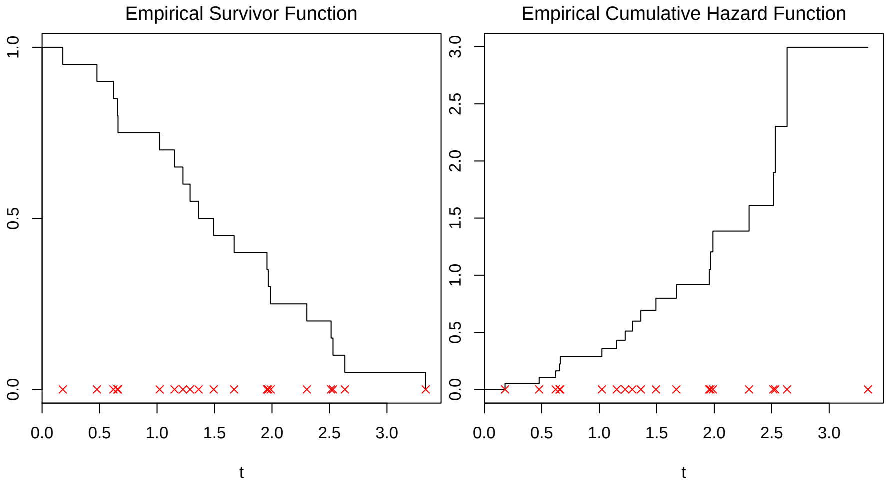


Figure 3: $\hat{S}(t)$ and $\hat{H}(t)$ for a sample of uncensored lifetime data.

We can use a plot of $\hat{H}(t)$ to see how the overall shape of the function compares with those dictated by some different distributions. In particular,

- $H(t)$ vs. t is linear for the exponential;
- $\log H(t)$ vs. $\log t$ is linear for the Weibull;
- $\log H(t)$ vs. t approaches linearity for large t for the Gompertz-Makeham.

After plotting $\hat{H}(t)$ in these ways we can decide which assumption is closest. This provides us with a crude but useful method of selecting a parametric model for a given set of data.

4 Fitting Parametric Distributions

In Chapter 3 we met some standard parametric families that might be appropriate for modelling T, and gave some simple criteria for choosing amongst them given a set of realised values.

We now assume that a distribution has been selected. What values should the parameters of the distribution take? This is a problem in **statistical inference**.

The most popular statistical approach to fitting parametric distributions to data is the method of **maximum likelihood (ML) estimation**.

In survival analysis the inference problem is often complicated by the presence of censored observations, and so we will need to consider ML estimation in this context.

4.1 Maximum Likelihood Estimation

Assume that the data $(t_1, \ldots, t_n)$ are n independent, possibly censored realisations from the distribution $F(t; \theta)$, where the form of F is known (e.g. Weibull) but the parameters θ are unknown.

Definition: The likelihood is the joint probability of the observed data, regarded as a function of the unknown parameters θ.

The **likelihood principle** states that all of the information about the parameters in a sample of data is contained in the likelihood function.

The **law of likelihood** extends this principle to state that the degree to which the data supports one parameter value over another is given by the ratio of their likelihoods.

The Maximum likelihood estimate (MLE) of θ is based solely upon the likelihood of the observed data. Maximum likelihood estimation thus provides a principled and general method of estimating parameters in parametric distributions using observed data.

As the name suggests the MLE is the value of the parameters that maximises the probability of the observed data,

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} L(\theta),$$
$$L(\theta) = \prod_{i=1}^{n} \Pr(t_i | F(; \theta)),$$

where L is the likelihood function and $\Pr(t_i|F(;\theta))$ is the probability of observing the i^{th} observation given the distribution F with parameters θ.

If none of the observations are censored we find,

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta).$$

For each censored observation the contribution to the likelihood depends on the type of censoring. Recall, an observation is left (right) censored if we only have an upper (lower) bound for T, or interval censored if we only know that T lies in a specified interval.

- $\Pr(t_i|F(;\theta)) = S(t_i;\theta)$ for an observation right-censored at t_i;
- $\Pr(t_i|F(;\theta)) = F(t_i;\theta)$ for an observation left-censored at t_i;

- $\Pr(t_i|F(;\theta)) = F(t_i^{(u)};\theta) - F(t_i^{(l)};\theta)$ for an observation interval-censored within $[t_i^{(l)}, t_i^{(u)}]$.

From now on we shall only consider right-censoring, this being the most common.

We can then split the data into two disjoint sets relating to

$$\begin{aligned} U &= \text{uncensored data} \\ C &= \text{censored data.} \end{aligned}$$

The likelihood function is then

$$L(\theta) = \prod_{i \in U} f(t_i;\theta) \prod_{i \in C} S(t_i;\theta).$$

It is almost always more convenient to work with the log-likelihood,

$$\begin{aligned} \ell(\theta) = \log\{L(\theta)\} &= \sum_{i \in U} \log f(t_i;\theta) + \sum_{i \in C} \log S(t_i;\theta) \\ &= \sum_{i \in U} \log h(t_i;\theta) + \sum_{i=1}^{n} \log S(t_i;\theta) \\ &= \sum_{i \in U} \log h(t_i;\theta) - \sum_{i=1}^{n} H(t_i;\theta). \end{aligned}$$

For an m-parameter distribution, the estimate $\hat{\theta}$ may be found by solving the likelihood equations

$$\frac{\partial}{\partial \theta_j} \ell(\theta) = 0 \quad (j = 1, \dots, m).$$

Finding the solution often involves numerical techniques such as Newton and quasi-Newton methods.

4.1.1 Newton's Method

Newton's Method for optimising $\ell(\theta)$ begins by first making an initial guess $\theta = \theta_0 \in \mathbb{R}^m$, sufficiently close to the true optimum. Provided this initial guess θ_0 is incorrect, we would wish to add an increment $\delta \in \mathbb{R}^m$ to θ_0 s.t. $\ell(\theta_0 + \delta)$ is the optimum.

The (multivariate) Taylor expansion of ℓ about a value θ yields

$$\ell(\theta + \delta) \approx \ell(\theta) + \delta' \nabla \ell(\theta) + \frac{1}{2} \delta' \nabla^2 \ell(\theta) \delta. \tag{2}$$

To solve the easier problem of maximising the Taylor expansion (2), we find the derivative of (2) wrt δ is equal to

$$\nabla \ell(\theta) + \nabla^2 \ell(\theta) \delta$$

and hence get an approximate solution for δ of

$$\delta = -\{\nabla^2 \ell(\theta)\}^{-1} \nabla \ell(\theta).$$

So the algorithm proceeds iteratively, setting

$$\theta_n = \theta_{n-1} - \{\nabla^2 \ell(\theta_{n-1})\}^{-1} \nabla \ell(\theta_{n-1})$$

until sufficient convergence in the sequence $\theta_0, \theta_1, \theta_2, \dots$ occurs.

4.1.2 Asymptotic distribution of MLE

Maximum likelihood estimation not only provides a set of optimal parameter values for θ but also allows assessment of the variance in the estimates.

Consider the $m \times m$ information matrix $I(\theta) = -\nabla^2 \ell(\theta)$, so

$$I(\theta)_{ij} = \frac{-\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta), \quad i,j = 1, \dots, m.$$

Then asymptotically for large samples, $\hat{\theta} \sim \mathrm{N}(\theta, I(\theta)^{-1})$.

Evaluating $I(\theta)$ by setting the unknown θ equal to $\hat{\theta}$, giving the observed information matrix, leads to an approximate covariance matrix for $\hat{\theta}$. That is, if $V = I(\hat{\theta})^{-1}$ then its ij^{th} entry v_{ij} is an estimate of the covariance between $\hat{\theta}_i$ and $\hat{\theta}_j$.

In particular the standard error of $\hat{\theta}_j$ is given by the square root of the j^{th} diagonal element of V,

$$\text{s.e.}\left(\hat{\theta}_j\right) \approx \sqrt{v_{jj}}.$$

4.1.3 Functional invariance of MLEs

Another key advantage of ML estimation is that the MLE of a function of the parameters $g(\theta)$ is simply

$$\widehat{g(\theta)} = g(\hat{\theta}).$$

For example, the exponential distribution has mean $\mathrm{E}(T) = 1/\lambda$, hence the MLE of $\mathrm{E}(T)$ is

$$\widehat{\mathrm{E}(T)} = \frac{1}{\hat{\lambda}} = \frac{1}{r}\sum_{i=1}^{n} t_i,$$

which is the total time on the study survived by the individuals divided by the number of deaths.

4.2 Examples of ML Estimation

We assume that we have observed lifetimes $t_1, \dots, t_n$ with r uncensored observations and $n - r$ right-censored observations.

4.2.1 Exponential distribution

The log-likelihood is

$$\ell(\lambda) = r \log \lambda - \lambda \sum_{i=1}^{n} t_i.$$

Hence

$$\frac{d\ell}{d\lambda} = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i,$$

which may be set to zero and solved immediately to give

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}.$$

Also we have,

$$\frac{-d^2\ell}{d\lambda^2} = \frac{r}{\lambda^2} \implies \text{s.e.}\left(\hat{\lambda}\right) \approx \frac{\hat{\lambda}}{r^{1/2}}.$$

So we have approximately

$$\hat{\lambda} \sim \text{Normal}(\lambda, \lambda^2/r),$$

and then more approximately

$$\hat{\lambda} \sim \text{Normal}\left(\lambda, r \Big/ \left\{\sum_{i=1}^{n} t_i\right\}^2\right).$$

4.2.2 Weibull distribution

Recall the Weibull survivor function has form,

$$S(t) = \exp(-\lambda t^{\eta})$$

with scale λ and shape η. Note, we have changed the parameterisation slightly, writing λ for $\alpha^{-\eta}$.

The log-likelihood in the presence of right-censoring is

$$\ell(\lambda, \eta) = r\log(\lambda\eta) + (\eta - 1)\sum_{i\in U}\log t_i - \lambda\sum_{i=1}^{n} t_i^{\eta}.$$

We set

$$\frac{\partial \ell}{\partial \lambda} = 0 = \frac{r}{\hat{\lambda}} - \sum_{i=1}^{n} t_i^{\hat{\eta}} \tag{3}$$

and

$$\frac{\partial \ell}{\partial \eta} = 0 = \frac{r}{\hat{\eta}} + \sum_{i\in U}\log t_i - \hat{\lambda}\sum_{i=1}^{n} t_i^{\hat{\eta}}\log t_i. \tag{4}$$

Solving (3), we obtain

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\eta}}}.$$

Substituting into (4) gives

$$\frac{r}{\hat{\eta}} + \sum_{i\in U}\log t_i - \frac{r}{\sum_{i=1}^{n} t_i^{\hat{\eta}}}\sum_{i=1}^{n} t_i^{\hat{\eta}}\log t_i = 0.$$

This is a non-linear equation in $\hat{\eta}$ which can only be solved using an iterative numerical procedure.

In practice it is common to maximise $\{\lambda, \eta\}$ simultaneously using Newton's method. An important by-product of which is an approximation of the covariance matrix from which standard errors on MLE can be obtained.

4.3 Hypothesis Testing for Nested Distributions

At the end of Chapter 3 we provided some heuristic methods for distinguishing between distributions. Here we consider asymptotic properties of the likelihood function to provide a more rigorous test that can be used for certain comparisons.

Suppose we are interested in making statements about a parameter subset $\theta^{(A)} \subseteq \theta \in \Theta$. Assume that θ is partitioned $\theta = (\theta^{(A)}, \theta^{(B)})$ and $\Theta = \Theta^{(A)} \times \Theta^{(B)}$.

- For example, in the Weibull distribution we might take $\theta^{(A)} = \eta$ (shape), $\theta^{(B)} = \lambda$ (scale).

Let $\hat{\theta} = (\hat{\theta}^{(A)}, \hat{\theta}^{(B)})$ denote the MLE of $\theta = (\theta^{(A)}, \theta^{(B)})$.

We will consider two tests for the null hypothesis $\mathrm{H}_0 : \theta^{(A)} = \theta_0^{(A)}$ vs. $\mathrm{H}_1 : \theta^{(A)} \in \Theta^{(A)}$ which make use of the MLE.

- Note that for the Weibull distribution a test for $\eta = 1$ is a test of exponentiality.

The tests lead to the derivation of a confidence region for $\theta^{(A)}$ which is the collection of parameter values in the subset $\Theta^{(A)}$ not 'rejected' at a certain significance level.

4.3.1 Likelihood ratio statistic

Let

- $\hat{\theta}_{H_1} = \hat{\theta}$ be the unconstrained MLE over Θ.
- Let $\hat{\theta}_{H_0} = \left(\theta_0^{(A)}, \hat{\theta}^{(B)}_{\theta_0^{(A)}}\right)$, where $\hat{\theta}^{(B)}_{\theta_0^{(A)}}$ is the MLE estimate of $\theta^{(B)}$ over $\Theta^{(B)}$ with $\theta^{(A)}$ fixed at $\theta^{(A)} = \theta_0^{(A)}$.

Then the likelihood ratio test statistic is

$$W\left(\theta_0^{(A)}\right) = -2\left[\ell\left(\hat{\theta}_{H_0}\right) - \ell\left(\hat{\theta}_{H_1}\right)\right].$$

Under the null hypothesis $\mathrm{H}_0 : \theta^{(A)} = \theta_0^{(A)}$, W is itself a random variable with an approximate chi-squared distribution with m_A degrees of freedom, where $m_A = \dim\left(\theta^{(A)}\right)$.

Large values of W relative to $\chi^2_{m_A}$ supply evidence against H_0.

The corresponding $1-\alpha$ confidence region for $\theta^{(A)}$ is

$$\left\{\theta^{(A)} : W\left(\theta^{(A)}\right) \leq \chi^2_{m_A,\alpha}\right\},$$

where $\chi^2_{m_A,\alpha} =$ is the upper 100α percentage point of $\chi^2_{m_A}$. Within this region we cannot reject the null hypothesis at the α significance level.

4.3.2 Wald statistic

Suppose $V = V\left(\hat{\theta}^{(A)}, \hat{\theta}^{(B)}\right)$ is the asymptotic covariance matrix for $\left(\hat{\theta}^{(A)}, \hat{\theta}^{(B)}\right)$ evaluated at the MLE.

Let V_A be the leading submatrix of V corresponding to $\theta^{(A)}$. Then,

$$W^*\left(\theta_0^{(A)}\right) = \left(\hat{\theta}^{(A)} - \theta_0^{(A)}\right)' V_A^{-1}\left(\hat{\theta}^{(A)} - \theta_0^{(A)}\right)$$

also has an approximate $\chi^2_{m_A}$ distribution under the null hypothesis $\theta^{(A)} = \theta_0^{(A)}$.

The corresponding $1-\alpha$ confidence region for $\theta^{(A)}$ is

$$\left\{\theta^{(A)} : W^*\left(\theta^{(A)}\right) \leq \chi^2_{m_A,\alpha}\right\}.$$

Both the Wald statistic and the likelihood ratio test statistic are based on large sample theory and both are asymptotically equivalent (and often give similar results in practice). However, large discrepancies are possible.

In such cases the likelihood ratio statistic is perhaps preferable as the results are invariant to reparameterisation.

Against this, the Wald statistic has the advantage that only one maximisation, over the unconstrained parameter space Θ, is required.

Note however, that the theory is for large samples and may give a poor approximation to small-sample results.

5 Nonparametric Methods

In Chapter 3 we met some parametric distributions which may be appropriate for modelling a random lifetime variable T. Once a parametric model has been chosen, the distribution of T is then constrained to take a fixed functional form and can only vary within the scope of the parameters. (In Chapter 4 we considered how best to choose those free parameters to fit a set of data.)

A limitation of this approach is that when the model is inadequate some interesting features within the data will be hidden. That is, we are constrained in how we learn about the distribution of T by the model we have selected to represent it.

In this chapter we see how this constraint can be relaxed and the search for an estimated distribution extended to the space of all distributions, leading to the name nonparametric or distribution-free inference.

So far we have been assuming T to be continuous, but in what follows we will require some results from discrete random variable distributions.

5.1 Discrete Lifetimes

Suppose that the random lifetime T could only take one of a countable set of values $\{a_1 < a_2 < \dots\}$. Then T would now be a discrete random variable, with a probability mass function (pmf)

$$\pi_j = \mathrm{P}(T = a_j), \quad j = 1, 2, \dots .$$

As usual, we would require that $\{\pi_j\}$ satisfy $\forall j$, $0 \leq \pi_j \leq 1$ and $\sum_j \pi_j = 1$.

5.1.1 Discrete hazard rate

We define the discrete hazard rate at time a_j as

$$h_j = \mathrm{P}(T = a_j | T \geq a_j) = \frac{\pi_j}{1 - \sum_{i<j} \pi_i}. \tag{5}$$

$$\pi_j = \left\{ \begin{array}{ll} h_1, & j = 1 \\ h_j \prod_{i<j}(1 - h_i), & j > 1. \end{array} \right. \tag{6}$$

Rearranging (5) gives

$$\begin{aligned} \pi_j &= h_j \left(1 - \sum_{i<j} \pi_i \right) \\ \implies \pi_j &= \pi_{j-1} \frac{h_j}{h_{j-1}} (1 - h_{j-1}) \qquad \text{(Exercise)} \end{aligned}$$

and so by induction

$$\pi_j = \begin{cases} h_1, & j = 1 \\ h_j \prod_{i<j}(1 - h_i), & j > 1. \end{cases} \tag{7}$$

Thus $\{\pi_j\}$ and $\{h_j\}$ both characterise the distribution of T, although for the latter we only need $\forall j$, $0 \leq h_j \leq 1$ (and if the range of T is finite, that $\exists j$ s.t. $h_j = 1$).

Note that if $\exists j$ s.t. $h_j = 1$, then surely $T \leq a_j$.

5.1.2 Survivor and discrete cumulative hazard rate

Since $\mathrm{P}(T > a_j) \equiv \mathrm{P}(T \geq a_{j+1})$, note that another rearrangement of (5) gives

$$\mathrm{P}(T > a_j) = \frac{\pi_{j+1}}{h_{j+1}}.$$

Substituting π_{j+1} with (7), we get

$$\begin{aligned} \mathrm{P}(T > a_j) \quad &= \textstyle\prod_{i=1}^{j}(1 - h_i) \\ \implies S(t) \quad &= \textstyle\prod_{j:a_j \leq t}(1 - h_j). \end{aligned} \tag{8}$$

So for an individual to survive to time t they must survive through all of the support points a_j up to time t, where each represents a Bernoulli trial with death probability h_j. Again, note that if $\exists j$ s.t. $h_j = 1$, then $\forall t \geq a_j, S(t) = 0$.

In an analogous definition to the continuous case (§2.5), we define the cumulative hazard to be

$$H(t) = \sum_{j:a_j \leq t} h_j. \tag{9}$$

Note that this invalidates the identity $H(t) = -\log\{S(t)\}$ from §2.5. However, we can note that for small h, $-\log(1 - h) \approx h$ and thus for low hazard rates

$$H(t) \approx -\sum_{j:a_j \leq t} \log(1 - h_j) = -\log\{S(t)\}.$$

5.2 Kaplan-Meier (Product-limit) Estimate

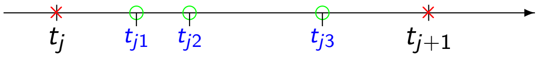
Returning to the idea of nonparametric estimation, we would like to find the maximum likelihood distribution for some observed survival data, over the space of all distributions.

Suppose we observe n independent identically distributed lives from a population with survivor function S, but that in the presence of right-censoring we only observe m deaths.

Let $t_1 < t_2 < \ldots < t_k$ be the ordered death times, with $k \leq m$ so that more than one death can occur at any one time. Let d_j denote the number of deaths at time t_j, with $\sum_{j=1}^{k} d_j = m$.

Observation of the remaining $n - m$ times is censored; Suppose that c_j lives are censored in the half open interval $[t_j, t_{j+1})$, for $j = 0, 1, 2 \ldots, k$, where $t_0 = 0$ and $t_{k+1} = \omega$.

Let $t_{j1}, t_{j2}, \ldots, t_{jc_j}$ be the censoring times in the j^{th} interval, where the $\{t_{ji}\}$ may or may not be distinct. Note that $\sum_{j=0}^{k} c_j = n - m$.



Let

$$n_j = n - \sum_{i \leq j-1} c_i - \sum_{i \leq j-1} d_i \quad \left(= \sum_{i=j}^{k}(c_i + d_i) \right) \tag{10}$$

be the total number of individuals still 'in view' as we reach time t_j. Note the convention of including in n_j any observations censored at t_j; alternative schemes may be more appropriate in some circumstances, and these naturally might lead to differing results.

It is easy to show that regardless of the nature of the true underlying distribution, the maximum likelihood distribution for T will be that of a discrete random variable with atoms of probability mass $\{\pi_j\}$ at the death times $\{t_j\}$. Informally, any distribution placing mass away from the death times would admit possibilities outside of the data observed, and so clearly give lower likelihood to the data that have been observed.

To identify the maximum likelihood discrete distribution, the problem at hand then is to identify optimal probability masses to place at each of the death times $t_1, \dots, t_k$.

In constructing a likelihood function for our unknown discrete distribution, it is most straightforward to specify the likelihood of the observed data in terms of the discrete hazard function. Proceeding through time from birth, individuals would only be at risk of death at the discrete points $t_1, \dots, t_k$. Viewing the data sequentially at these risk times, as we reach time t_j, there are n_j individuals still at risk, each subject to a Bernoulli trial with probability h_j of death and $(1 - h_j)$ of survival. Once individuals have died or been right-censored, they cease to play a part in the subsequent terms of the likelihood function.

Hence we can write the likelihood of the data given a discrete model as

$$L(h_1, \dots, h_k) = \prod_{j=1}^{k} h_j^{d_j} (1 - h_j)^{n_j - d_j},$$

which is clearly maximised by

$$\hat{h}_j = \frac{d_j}{n_j}, \quad j = 1, 2, \dots, k.$$

Hence the MLE for S is

$$\hat{S}(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

This is known as the Kaplan-Meier or product-limit estimate. It is the MLE over the space of all valid distribution/survivor functions.

Product-limit: consider, finer and finer partitions of the time axis and estimate $S(t)$ as the product of probabilities of surviving each sub-interval. We can treat this as a discrete, sequential problem using the discrete hazard definition in §5.1, and thus obtain the Kaplan-Meier estimate in the limit as the meshes of the partition tend to zero.

Example

Consider the following data on 10 individuals

- death times: 1.1, 3, 3, 7, 10, 12.4;
- right-censoring times: 0.2, 0.8, 4.5, 11.

Death time	Censoring time	n_j	d_j	$\hat{S}(t)$
	0.2			1
	0.8			1
1.1		8	1	0.875
3		7	2	0.625
	4.5			0.625
7		4	1	0.469
10		3	1	0.312
	11			0.312
12.4		1	1	0.0

Note: If the largest observation is a censored survival time, then $\hat{S}(t)$ is undefined beyond this maximum time t_{max}.

5.3 Greenwood's Formula for the Standard Error of the K-M Estimate

The Kaplan-Meier estimate is the most important and widely used estimate of the survivor function, and so it is of interest to estimate its variability.

We have

$$\log \hat{S}(t) = \sum_{j:t_j \leq t} \log(1 - \hat{h}_j)$$

and then by approximate independence the variance of $\log(\hat{S}(t))$ is

$$\text{Var}\left\{\log \hat{S}(t)\right\} \approx \sum_{j:t_j \leq t} \text{Var}\{\log(1 - \hat{h}_j)\}.$$

The variance of $(1 - \hat{h}_j)$ follows from the variance of a binomial random variable

$$\text{Var}\ (1 - \hat{h}_j) \approx \hat{h}_j(1 - \hat{h}_j)/n_j.$$

Now we make use of a general result, known as the delta method, for approximating the variance of a function of an asymptotically normal statistical estimator (like the MLE). If X is a random variable with mean h_X and $g(X)$ is a function of X with first derivative $g'(X)$, then

$$\text{Var}\{g(X)\} \approx g'(h_X)^2 \text{Var}(X), \tag{11}$$

This leads to

$$\begin{aligned} \text{Var}\{\log(1 - \hat{h}_j)\} &\approx \frac{\text{Var}(1 - \hat{h}_j)}{(1 - \hat{h}_j)^2} \approx \frac{\hat{h}_j(1 - \hat{h}_j)/n_j}{(1 - \hat{h}_j)^2} \\ &= \frac{\hat{h}_j}{n_j(1 - \hat{h}_j)} = \frac{d_j}{n_j(n_j - d_j)}. \end{aligned}$$

Hence

$$\text{Var}\{\log \hat{S}(t)\} \approx \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}$$

and further application of (11) leads to

$$\text{Var}\{\log \hat{S}(t)\} \approx \frac{1}{\{\hat{S}(t)\}^2} \text{Var}\{\hat{S}(t)\}$$

so that

$$\text{Var}\{\hat{S}(t)\} \approx \{\hat{S}(t)\}^2 \sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Finally the standard error of the Kaplan-Meier estimate is the square root of the estimated variance; that is,

$$\text{s.e.}\{\hat{S}(t)\} = \hat{S}(t) \sqrt{\sum_{j:t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}},$$

which is known as Greenwood's formula.

5.4 Nelson-Aalen Estimate of the Cumulative Hazard

We have found the MLE for S, $\hat{S}$, to be the survivor function of a discrete random variable with atoms of probability mass at the death times $\{t_j\}$. This has a discrete hazard rate $\hat{h} = \{\hat{h}_j\}$.

If our interests lay in estimating the cumulative hazard function H, as in §3.5, then following definition (9) the corresponding estimate would be

$$\hat{H}(t) = \sum_{j:t_j \leq t} \hat{h}_j = \sum_{j:t_j \leq t} \frac{d_j}{n_j}.$$

This is the Nelson-Aalen estimate. By the property of invariance of MLEs under transformation, this is the MLE for the (discrete) cumulative hazard function. The MLE for the continuous cumulative hazard function is instead $-\log(\hat{S}(t))$.

Similarly to Greenwood's formula for the Kaplan-Meier estimate we can obtain an estimate for the standard error of the Nelson-Aalen estimate,

$$\text{Var}\left\{\hat{H}(t)\right\} \approx \sum_{j:t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3} \qquad \Longrightarrow \text{s.e.}\left\{\hat{H}(t)\right\} \approx \sqrt{\sum_{j:t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3}}.$$

5.5 The Actuarial Estimate of S

The "adjusted-observed" life-table estimate or actuarial estimate of the survivor function deals with aggregated, interval censored and right censored observations of a continuous time to event variable, which are commonly found in actuarial applications. Its construction is very similar to that of the product-limit estimate.

We suppose that the time axis is divided up into intervals (usually of equal length, typically one year) $[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$ where $0 = a_0 < a_1 < \ldots < a_k$.

The data are aggregated within these intervals, so the number of deaths, d_j, and losses to censoring, c_j, are known for each interval $[a_{j-1}, a_j)$ but the actual lifetimes/censoring times are unknown. As in the product-limit model, we estimate the probabilities of surviving each subsequent interval.

If all of the deaths were known to precede all of the losses to censoring, then we could proceed with estimation as before with $\hat{h}_j = d_j/n_j$. But in general we have no reason to assume this. The actuarial assumption is that within each region the losses to censorship occur uniformly across the region. Hence by a crude approximation the adjusted, expected number of individuals at risk at any time during the interval is

$$n'_j \;=\; n - \sum_{i \leq j-1} c_i - \sum_{i \leq j-1} d_i - \frac{c_j}{2} = n_j - \frac{c_j}{2},$$

where n_j is the number still 'in view' at the start of the j^{th} interval.

So using the arguments developed in §5.1 and §5.2 we treat survival of the intervals as (approximate) binomial experiments, leading to the life-table estimate of S

$$S^*(t) = \prod_{j:a_j \leq t} \left(1 - \frac{d_j}{n'_j}\right),$$

with corresponding standard error

$$\text{s.e.}\{S^*(t)\} \approx S^*(t)\sqrt{\sum_{j:a_j \leq t} \frac{d_j}{n'_j(n'_j - d_j)}}.$$

6 Regression models

6.1 Inhomogeneous Populations

We have so far considered survival analysis under the assumption that the lifetimes of individuals in our population are independent, *identically distributed* realisations of a random variable T following a common distribution F.

This assumption of a homogeneous population is restrictive. We will often have additional information about the individuals, the knowledge of which affects our beliefs about their lifetime distribution.

In this chapter we shall consider regression methods for accommodating explanatory variables which we believe affect the distribution of T for each individual.

These variables could be:

- **demographic** - e.g. age, sex.
- **behavioural** - e.g. smoking history, exercise taken.
- **physiological** - e.g. blood pressure, cholesterol level.
- **external** - e.g. treatment group/centre, dose level.

If the information on individuals is categorical (e.g. smokes/does not smoke) and we only have a few explanatory variables, we may be able to partition the data set into smaller groups of homogeneous individuals and use the methods in Chapters 3- 5 to estimate distributions separately for each group. For example, we could split the population into smokers and non-smokers and derive Kaplan-Meier estimates $\hat{S}_{\text{smokers}}$ and $\hat{S}_{\text{non-smokers}}$ and compare.

However, often we might have many factors to take into account and if some of them are continuous then this simple approach breaks down.

Methods that can deal with these general regression situations include accelerated life models and the popular proportional hazards model.

Suppose we have decided upon p explanatory variables (or *covariates*) which we believe may affect the time until event. Let z_i be the p-vector of values of these covariates observed for individual i.

Where possible we attempt to *standardise* these covariates so that $z = 0$ relates to some standard set of conditions, such as the control group in a clinical trial.

Under both the proportional hazards and accelerated lifetime approaches, a general model for lifetime distributions is constructed from two parts:

1. A baseline model for the (hypothetical) standard individuals, corresponding to $z = 0$;
2. A scheme for modifying the baseline model as the covariates z deviate from zero.

6.2 Cox's Proportional Hazards Model

The Cox [1] model proposes that the hazard rates of individuals are related via the relationship

$$h(t; z) = h_0(t) \exp(\beta \cdot z), \qquad (12)$$

where $\beta \in \mathbb{R}^p$ is a vector of regression parameters. Equation (12) shows a multiplicative influence on the hazard of any deviation away from zero in each of

[1] `http://www.jstor.org/page/termsConfirm.jsp?redirectUri=/stable/pdfplus/2985181.pdf` Cox, D. R. (1972) Regression Models and Life Tables, *Journal of the Royal Statistical Society Series B*, **34**, 187-220.

the p covariates in z. Here $h_0(t)$ is known as the baseline hazard and represents the hazard of a (possibly hypothetical) individual with $z = 0$.

The survivor function and density follow:

$$S(t; z) = S_0(t)^{\exp(\beta \cdot z)},$$
$$f(t; z) = \exp(\beta \cdot z) S_0(t)^{\exp(\beta \cdot z) - 1} f_0(t),$$

where $S_0(t)$ and $f_0(t)$ are the baseline survivor and density functions corresponding to $h_0(t)$.

In (12), only h_0, not z or β, depends on time but the model can also be formulated with time-dependent covariates.

Under the Cox model, the hazard functions of two individuals with covariates z_1, z_2 are in constant proportion at all times; that is,

$$\frac{h(t; z_1)}{h(t; z_2)} = \frac{\exp(\beta \cdot z_1)}{\exp(\beta \cdot z_2)} = \exp\{\beta \cdot (z_1 - z_2)\},$$

giving rise to the name proportional hazards.

Note that in general we can have any positive function ψ of the covariates leading to $h(t; z) = h_0(t)\psi(z; \beta)$ and a 'proportional hazards' structure.

However, Cox's model easily ensures that the hazard is always positive, and gives a linear model for the log-hazard which is very convenient in theory and practice.

The utility of the model arises from the idea that the general *shape* of the hazard functions could be the same for all individuals in a population, with any covariate-specific differences between individuals having a constant, multiplicative effect.

So, if we are not primarily concerned with the precise form of the hazard, but with the effects of the covariates, we may wish to ignore h_0 and estimate the regression coefficient vector β from the data without reference to the shape of h_0.

This is termed a semi-parametric approach: a parametric model for covariate effects, $\exp(\beta \cdot z)$, and a distribution-free reference to h_0.

The Cox model (12) dominates the literature on regression of lifetimes, due to its simplicity and versatility.

6.2.1 Partial likelihood function

Suppose we have observation times $t_1, \dots, t_n$, some of which are right-censoring times. Let each of these individuals have associated covariate vectors $z_1, \dots, z_n$.

We consider inference about β when the baseline hazard function is completely unknown.

This raises the question: "What is the likelihood of observing the data as a function of β when h_0 is unspecified?"

That is, we would like the marginal likelihood of β. Strictly, this is not analytically available (this would require Bayesian methods). However, Cox developed the following (subjective) argument:

> In the absence of knowledge about h_0, the time intervals between deaths can provide little or no information about β, as their distribution will depend heavily on h_0.
>
> Hence it is only the event time ordering of the individuals that holds any information about β. The contribution to the likelihood function for β for an individual, with covariates z_i, that dies at time t_i is

$$\begin{aligned}
&\mathrm{P}(\text{individual } i \text{ with covariates } z_i \text{ dies at } t_i | \text{one death at } t_i) \\
&= \frac{\mathrm{P}(\text{individual } i \text{ with covariates } z_i \text{ dies at } t_i)}{\mathrm{P}(\text{one death at } t_i)} \\
&= \frac{h(t_i; z_i)}{\sum_{j \in R_{t_i}} h(t_i; z_j)} \\
&= \frac{\exp(\beta \cdot z_i)}{\sum_{j \in R_{t_i}} \exp(\beta \cdot z_j)},
\end{aligned}$$

where R_{t_i} is the risk set; that is, the set of individuals "at risk" (or "in view", so not dead or censored) at time t_i.

This leads to,

$$L(\beta) = \prod_{i \in U} \frac{\exp(\beta \cdot z_i)}{\sum_{j \in R_{t_i}} \exp(\beta \cdot z_j)}, \tag{13}$$

where the product is over the set of uncensored observations (death times).

The likelihood (13) is not a full likelihood function for the observed data as it makes no use of the actual event/censoring times. For this reason it is referred to as a partial likelihood.

Note that the baseline hazard cancels out in the partial likelihood and so $L(\beta)$ is independent of h_0 when we only consider the order in which the deaths and losses to censoring occurred.

Note that the censored observations only appear in the denominator of (13), through their presence/absence in the risk sets R_{t_i} at each of the death times.

If the event times are ordered, $t_1 < t_2 < \cdots < t_n$ we can write $j \in R_{t_i} \equiv \{t_j \geq t_i\} \iff j \geq i$.

Example

Consider the following data on 5 individuals: z_1, death at 2.8; z_2, censored at 3.1; z_3, death at 2.2; z_4, death at 4.7; z_5, censored at 2.6. So there are 3 deaths.

Writing $\psi_i = \exp(\beta \cdot z_i)$, the partial likelihood for β is constructed sequentially:

$$L(\beta) = \frac{\psi_3}{\psi_1 + \psi_2 + \psi_3 + \psi_4 + \psi_5} \times \frac{\psi_1}{\psi_1 + \psi_2 + \psi_4} \times \frac{\psi_4}{\psi_4}.$$

In general, analytically maximising $L(\beta)$ (or $\ell(\beta)$) wrt β is not possible. Most statistical packages contain numerical procedures for fitting a Cox model, such as Newton methods, which make use of $\ell'(\beta)$ and $\ell''(\beta)$.

Handling ties:

In practice complications in evaluating the partial likelihood can arise if

1. some $d_j > 1$; more than one death at any time.
2. some censored observations coincide with a death time t_j.

It is usual to deal with 2. by assuming that the losses to censoring occurred just after time t_j. So individuals censored at t_j are included in the risk set R_{t_j}.

Full analytic treatment of multiple simultaneous deaths, case 1, leads to complicated expressions for the likelihood, since all possible orderings of the d_j deaths from R_{t_j} should be considered. The following approximation due to Breslow (1974) [2] is commonly used. For death times $t_1, \dots, t_k$,

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\beta \cdot s_j)}{[\sum_{i \in R_{t_j}} \exp(\beta \cdot z_i)]^{d_j}},$$

where s_j is the sum of the covariate vectors z of the d_j lives observed to die at time t_j.

The partial likelihood is a true likelihood for the order of the events, and so has all the usual properties; through maximisation it yields an estimator $\hat{\beta}$ which is asymptotically (multivariate) normally distributed and unbiased, and whose asymptotic covariance matrix can be estimated by the inverse of the observed information matrix, found from

$$I(\beta)_{ij} = -\frac{\partial^2}{\partial\beta_i \partial\beta_j}\ell(\beta), \quad i,j = 1, \dots, p,$$

evaluated at $\hat{\beta}$, where $\ell(\beta) = \log\{L(\beta)\}$ is the log partial likelihood.

$I(\hat{\beta})$ is usually provided as a by product by most computer packages when fitting the Cox model using Newton's method.

This allows standard errors for $\hat{\beta}$ to be obtained which are helpful for model checking.

6.2.2 Model checking

In a practical survival problem several possible explanatory variables might present themselves as possibly affecting the survival time.

Part of the modelling procedure is to assess which (if any) are relevant to the model. That is, which ones have a significant effect on the distribution of T.

The fact that the partial likelihood behaves like a full likelihood allows us to make use of the test statistics from §4.3.

Suppose we currently have a model with p covariates and we wish to consider the inclusion of an extra q covariates.

We can fit both models, one with p and the other with the $p + q$ covariates, by maximising the partial likelihoods. Let ℓ_p and ℓ_{p+q} be the maximised log-likelihoods of the two nested models. Note that $\ell_{p+q} \geq \ell_p$.

The likelihood ratio statistic is

$$2(\ell_{p+q} - \ell_p)$$

and has an asymptotic χ^2_q distribution under the null hypothesis that the extra q covariates have no additional effect in the presence of the original p explanatory variables.

$$H_0 : \beta_{p+1} = \dots = \beta_{p+q} = 0 \text{ vs. } H_1 : (\beta_{p+1}, \dots, \beta_{p+q}) \in \mathbb{R}^q.$$

[2] `http://www.jstor.org/page/termsConfirm.jsp?redirectUri=/stable/pdfplus/2529620.pdf`Breslow, N. (1974) Covariance Analysis of Censored Survival Data. *Biometrics*, **30**, 89-99.

Example

Suppose we are seeking a model for learning the effect of hypertension on survival.

The first model we might consider has two covariates $z_i = (z_{i1}, z_{i2})$ relating to the sex of an individual, z_{i1}, and their blood pressure, z_{i2}.

Suppose then we wanted to test the hypothesis that cigarette smoking has no effect, allowing for sex and blood pressure.

We can construct two nested models with alternative covariate vectors $z^{(1)} = \{(z_{i1}, z_{i2}) : i = 1, \dots, n\}$ and $z^{(2)} = \{(z_{i1}, z_{i2}, z_{i3}) : i = 1, \dots, n\}$ where z_{i3} is a 0-1 indicator on whether the i^{th} individual smokes.

We fit both models and note the value of the maximum log-likelihood for the two models, $\ell_{z^{(1)}}, \ell_{z^{(2)}}$. Under the assumption that smoking has no effect the test statistic $W = 2(\ell_{z^{(2)}} - \ell_{z^{(1)}})$ has an asymptotic χ_1^2 distribution.

If W is small we cannot reject the null hypothesis, that smoking has no effect.

6.2.3 Estimating the baseline distribution

Once the regression coefficients have been estimated by MLE, it is possible to obtain a nonparametric estimate of the baseline survivor function.

The following formula from Kalbfleisch and Prentice (1980) [3] is a generalisation of the product-limit (PL) estimator of §5.2 when there are no ties

$$\hat{S}_0(t) = \prod_{t_i \le t} \left(1 - \frac{\exp(\hat{\beta} \cdot z_i)}{\sum_{j \in R_{t_i}} \exp(\hat{\beta} \cdot z_j)}\right)^{\exp(-\hat{\beta} \cdot z_i)} .$$

As with the PL estimator, if the last observation is a censored value, then $\hat{S}_0(t)$ is undefined past this point. Note that a plot of $\hat{S}_0(t)$ may often suggest a suitable parametric form for S.

Of course there is no requirement that h_0 should be unspecified to perform regression modelling in survival analysis. In the case where we can assume h_0 to be from a known parametric family, then full likelihood based inference for $\{h_0, \beta\}$ can proceed. In most practical applications the form of h_0 will not be well understood beforehand; in which case, learning the effects of the covariates and from there estimating the shape of the baseline model, as above, might lead to consideration of a suitable parametric form.

6.3 Accelerated Failure Time Model

The accelerated failure time model is an alternative, general model for survival data in which the covariates are assumed to act multiplicatively on the time-scale.

The model has intuitive appeal in that the covariates are seen to 'speed up' or 'slow down' the passage of time for one individual relative to another.

Formally, the survivor function for an individual is of the form

$$S(t; z) = S_0(t\psi(z; \beta))$$

where S_0 is a baseline survivor function and $\psi(z; \beta)$ is a positive function of the covariates parameterised by β.

Assuming $\psi(0; \beta) = 1$, the corresponding random variables are related by

$$T_z = \frac{T_0}{\psi(z; \beta)}$$

[3] Kalbfleisch, J. D. and Prentice, R. L. (1980) *The statistical analysis of failure time data*. Wiley.

where T_z is the future lifetime random variable for an individual with covariates z and hence T_0 is the baseline future lifetime random variable.

It follows that

$$\begin{aligned} f(t;z) &= f_0[t\psi(z;\beta)]\psi(z;\beta) \\ h(t;z) &= h_0[t\psi(z;\beta)]\psi(z;\beta) \end{aligned}$$

As in the proportional hazards model it is convenient to take $\psi(z;\beta) = \exp(\beta \cdot z)$.

6.3.1 Relation to proportional hazards

For illustration we consider regression modelling of two groups of individuals with

- $z_i \in \{0,1\}$ (e.g. smokers/non-smokers) and
- a baseline hazard which is piecewise constant, say

$$h_0(t) = \begin{cases} 0.5 & 0 \leq t \leq 1 \\ 1 & t > 1 \end{cases}$$

Note that this is in effect a piecewise exponential model.

Consider a proportional hazards model

$$h(t;z_i) = h_0(t)\exp(\beta z_i)$$

and accelerated failure time model,

$$h(t;z_i) = h_0[t\exp(\beta z_i)]\exp(\beta z_i),$$

both with $\beta = 1$.

Under the accelerated failure time model the increase in hazard for the group $\{i : z_i = 1\}$ occurs sooner than under the proportional hazards model (see Fig 4, left).

The 'kink' in the survivor function (Fig 4, right) also occurs earlier.

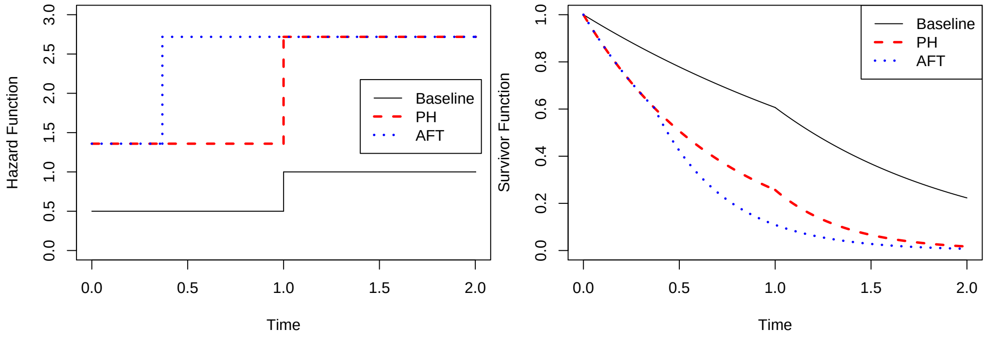


Figure 4: Hazard (left) and survivor (right) functions: Baseline (–); Proportional Hazards (-); Accelerated failure $(\cdots)$. With $\psi(z_i;\beta) = \exp(z_i)$, $z_i = 1$.

7 The Markov Model

7.1 Stochastic Processes

Previously we have modelled the future lifetime of an individual currently aged x as a *single* random variable T with an unknown distribution F. In this chapter we take a different approach and treat each lifetime as a realisation of a stochastic process; that is, an indexed *family* of random variables on a probability space.

Definition: A stochastic process on a set Ω is a collection $\{X(t)|t \in S\}$ of random variables $X(t) \in \Omega$ indexed by a set S. Let P be generic notation for the probability distribution for the random variables $X(t)$.

Usually the index set $S \subseteq \mathbb{R}$. If S is countable (e.g. $S = \mathbb{N}$), then we say $\{X(t)\}$ is a discrete-time process. If S is an interval of $\mathbb{R}$ (e.g. $S = [0, \infty)$), then $\{X(t)\}$ is a continuous-time process.

In the Markov model, the sample path of the process $\{X(t)\}$ (the values that it takes over time) will be used to define our (possibly censored) time to event T.

Example: 2-state model

The simplest example of a continuous-time stochastic process for survival analysis is the two-state Markov model illustrated by the diagram below:

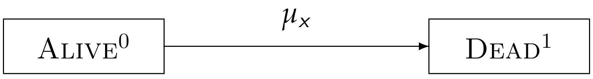


Figure 5: Alive-Dead model.

There is just an Alive state and a Dead state, with transitions in one direction only. The process takes value $X(t) = 0$ if the individual is in state Alive at time t, or $X(t) = 1$ if the individual is in state Dead, so $\Omega = \{0, 1\}$.

The state Dead is said to be an absorbing state; that is, $\forall s > 0$

$$P(X(t+s) = 1|X(t) = 1) = 1.$$

Here the event $T = t$ in previous chapters corresponds simply to a change of state from Alive to Dead at time t.

7.2 Homogeneous Continuous-Time Markov Jump Processes

7.2.1 Markov Jump Processes

If Ω is a finite set of N states (e.g. $\Omega = \{1, 2, \ldots, N\}$) and $X(t)$ is a stochastic process on the state space Ω, we say $X(t)$ is a jump process.

If the path of a continuous-time process $X = \{X(t) : t \geq 0\}$ makes only a finite number of jumps in any finite time interval, we say it is a pure jump process.

Furthermore, a pure jump process will be said to be a Markov jump process if it satisfies the Markov property

$$P\{X(t) = j|X(t_1) = j_1, \ldots, X(t_s) = j_s\} = P\{X(t) = j|X(t_s) = j_s\},$$

$\forall j_1, \ldots, j_s, j \in \Omega$ and $0 \leq t_1 < \ldots < t_s < t$.

7.2.2 Transition probabilities

Definition: The transition probability $p^{ij}(s,t)$ is defined to be probability of being in state j at time t, given that at time s the process was in state i,

$$p^{ij}(s,t) = \mathrm{P}(X(t)=j|X(s)=i), \qquad t \geq s \in \mathbb{R}^+, i,j \in \Omega.$$

7.2.3 Transition Intensities

Specifying coherent transition probabilities in continuous-time Markov models would be highly complex, and so instead these models are more typically characterised through a set of corresponding *transition intensities*.

Definition: For $t \geq 0$ and $i,j \in \Omega$, $i \neq j$ we define the force of transition or transition intensity,

$$\mu^{ij}(t) = \lim_{h \downarrow 0} \frac{p^{ij}(t,t+h)}{h} \tag{14}$$

and assume these limits exist. Transition intensities generalise the concept of the hazard function met in §2.4

When explanatory variables are available for individuals, the transition intensities for each individual can be constructed to depend on a covariate vector z in an analogous way to the regression methods of Chapter 6.

Under the assumption of existence of (14) we can write

$$p^{ij}(t,t+dt) = \mu^{ij}(t)dt + o(dt), \quad dt \geq 0$$

where $\dfrac{o(dt)}{dt} \to 0$ as $dt \downarrow 0$.

So for small dt, the transition probability $p^{ij}(t,t+dt)$ between two distinct states i and j is approximately linear in dt with constant of proportionality $\mu^{ij}(t)$,

$$p^{ij}(t,t+dt) \approx \mu^{ij}(t)dt.$$

Note that at time t, if $\mu^{ij}(t) = 0$ then transitions from i to j surely cannot occur.

For completeness, $\forall i \in \Omega$ we can define

$$\mu^{ii}(t) = \lim_{h \downarrow 0} \frac{p^{ii}(t,t+h)-1}{h}, \tag{15}$$

although note that we do **not** have *transitions* $i \to i$. This gives

$$p^{ii}(t,t+dt) = 1 + \mu^{ii}(t)dt + o(dt), \quad dt \geq 0$$
$$\implies p^{ii}(t,t+dt) \approx 1 - (-\mu^{ii}(t))dt.$$

This can be most helpfully interpreted as

$$\mathrm{P}\{X(t+dt) \neq i | X(t) = i\} \approx -\mu^{ii}(t)dt.$$

Defining $\mu^{ii}(t)$ this way leads to a useful relationship. Since $\forall i \in \Omega$ and $t,h > 0$, we know $\sum_{j=1}^{N} p^{ij}(t,t+h) = 1$, we therefore have

$$\begin{aligned} \mu^{ii}(t) &= \lim_{h \downarrow 0} \frac{\{1 - \sum_{j \neq i} p^{ij}(t,t+h) - 1\}}{h} \\ &= -\sum_{j \neq i} \lim_{h \downarrow 0} \frac{p^{ij}(t,t+h)}{h} \\ \implies \mu^{ii}(t) &= -\sum_{j \neq i} \mu^{ij}(t). \end{aligned}$$

Notice from (14) and (15) that $\forall i \neq j \in \Omega$, $\forall t \geq 0$, $\mu^{ij}(t) \geq 0$ and $\mu^{ii}(t) \leq 0$.

Definition: The $N \times N$ matrix $\mathbf{G}_t$ with ij^{th} entry $(\mathbf{G}_t)_{ij} = \mu^{ij}(t)$ is called the generator of the process.

The generator can be thought of as the continuous-time analogue of the one-step transition probability matrix of discrete-time Markov chains.

We have seen that $\mu^{ii}(t) = -\sum_{j\neq i} \mu^{ij}(t)$ and hence

$$\begin{aligned} &\sum_{j=1}^{N} \mu^{ij}(t) = 0, \qquad \forall i \in \Omega \\ \implies &\mathbf{G}_t \mathbf{1}' = \mathbf{0}' \end{aligned}$$

where $\mathbf{1}'$, $\mathbf{0}'$ are vectors of ones and zeros respectively.

7.2.4 Homogeneity

Definition: The process is said to be homogeneous if

$$p^{ij}(t, t+h) = p^{ij}(0, h), \qquad \forall i, j \in \Omega, t, h > 0,$$

in which case we can simplify notation and write $p^{ij}(h)$ for $p^{ij}(t, t+h)$.

Clearly the condition of homogeneity is equivalent to the transition intensities being constant,

$$\begin{aligned} \mu^{ij}(t) &\equiv \mu^{ij}, \\ \mathbf{G}_t &\equiv \mathbf{G}. \end{aligned}$$

7.2.5 Age rate intervals

Homogeneity is clearly a very strong condition, as it implies individuals in the process experience no effects from ageing. To make this assumption tenable, it is common to divide the time domain of interest (perhaps $\mathbb{R}^+$) into small interval domains $S_1, S_2, ...$ and fit separate Markov models for each interval S_i, within which homogeneity is more reasonable.

In studies of human mortality, these homogeneous intervals will typically be the integer age ranges $[x, x+1)$ for $x \in \mathbb{N} = \{0, 1, 2, ...\}$. Hence in these applications our implicit, hidden notation is actually $\mu^{ij}_{x+\frac{1}{2}}$, as the estimator can be thought of as corresponding most closely to the midpoint of this integer age x interval. Alternative *rate interval* definitions for age would change this; for example, if we defined age group x as $[x - \frac{1}{2}, x + \frac{1}{2}]$, i.e. nearest birthday being x years, then we would be approximately estimating μ^{ij}_x.

7.2.6 Assuming homogeneity

From now on we shall assume that $X(t)$ is a homogeneous process. We write $\mathbf{P}_t$ for the $N \times N$ matrix with entries $p^{ij}(t)$, and the transition intensities $\mu^{ij}(t)$ simply as μ^{ij}. However, it should be noted that many of the models and results that follow can be extended into the inhomogeneous, *duration-dependent* setting.

Theorem 7.1. *The family $\{\mathbf{P}_t : t \geq 0\}$, known as the transition semigroup of the process, is a stochastic monoid under the matrix multiplication operation; that is, it satisfies the following:*

1. *$\{\mathbf{P}_t : t \geq 0\}$ are right-stochastic: $\forall t \geq 0$, the entries of each row of $\mathbf{P}_t$ are non-negative and sum to 1.*

2. *Identity element: $\mathbf{P}_0 = \mathbf{I}$, the $N \times N$ identity matrix.*

3. *Semigroup property: $\mathbf{P}_{s+t} = \mathbf{P}_s\mathbf{P}_t$ if $s, t \geq 0$. These are known as the Chapman-Kolmogorov equations.*

Proof

1. For any $i \in \Omega$,

$$\begin{aligned}\sum_{j=1}^{N} (\mathbf{P}_t)_{ij} &= \sum_{j=1}^{N} p^{ij}(t) \\ &= \mathrm{P}\left(\cup_{j=1}^{N}\{X(t) = j\}|X(0) = i\right) \\ &= 1.\end{aligned}$$

2. Clearly $(\mathbf{P}_0)_{ij} = p^{ij}(0) = \delta_{ij}$.

3. Chapman-Kolmogorov:

$$\begin{aligned}p^{ij}(s+t) &= \mathrm{P}(X(s+t) = j|X(0) = i) \\ &= \sum_{k=1}^{N} \mathrm{P}(X(s) = k|X(0) = i)\mathrm{P}(X(s+t) = j|X(s) = k) \\ &= \sum_{k=1}^{N} p^{ik}(s)p^{kj}(t). \qquad \square\end{aligned}$$

The distribution of $\{X(t) : t \geq 0\}$ is completely determined by the stochastic semigroup $\{\mathbf{P}_t : t \geq 0\}$ and the initial distribution of $X(0)$. Furthermore, if we condition on $X(0)$ (i.e. assuming we can observe the initial states of our individuals) then the evolution of $X(t)$ depends only on $\mathbf{P}_t$.

7.3 A Two-State Markov Model

The two-state homogeneous Markov model in Figure 5 provides a very simple and familiar special case of these general Markov jump processes.

From above, the single transition intensity $\mu = \mu^{01}$ has equation

$$\mu = \lim_{h \downarrow 0} \frac{p^{01}(x, x+h)}{h}, \quad \forall x > 0. \tag{16}$$

Noticing that in the right hand side of (16), $p^{01}(x, x+h)$ is equal to $F_x(h)$ in random variable terminology, it is then apparent that here μ must be the hazard function. Since we have that μ is constant under the assumption of homogeneity, it immediately follows that the corresponding time to event random variable has an Exponential(μ) distribution.

For making inference about this two-state Markov model in the presence of possibly censored data, the results for maximum likelihood estimation of μ immediately follow from §4.2.1.

Let v be the sum of the waiting times (i.e. the time until death or censoring) of all individuals on a study. In actuarial science, this is often called the central exposed to risk and is denoted E_x^c for age group x.

Then if d is the number of deaths in the sample, recall the MLE for μ is given by

$$\hat{\mu} = \frac{d}{v}.$$

We had approximately

$$\hat{\mu} \sim \text{Normal}(\mu, \mu^2/d),$$

and then more approximately

$$\hat{\mu} \sim \text{Normal}(\mu, d/v^2).$$

Further, we can obtain MLEs for survival probabilities $S(t)$ and hence the transition probabilities $p^{01}(t)$ via $\hat{S}(t) = \exp\{-\hat{\mu}t\}$ and $1 - \hat{S}(t)$ respectively.

7.4 Calculating Transition Probabilities

The real utility of the Markov model for survival data is that the two-state model of §7.3 can be extended to any number of states, with arbitrary transitions between them. For example, the three-state Sickness-Death model (see Fig 6).

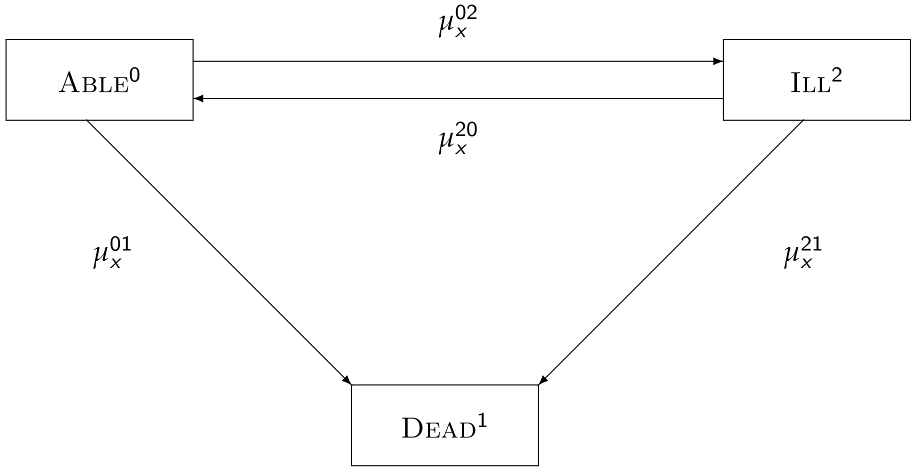


Figure 6: SICKNESS-DEATH model.

Unfortunately, analytic expressions for the transition probabilities in terms of the transition intensities are not available in the general N-state Markov model. However, we can derive differential equations for these relationships, known as the *Kolmogorov forward and backward equations*.

The forward equations are preferable if there is a single initial state of particular importance and we want to know the probabilities of being in the various other states at time t.

Conversely, the backward equations are preferred if there is a single final state of great interest and we want the probability of reaching this state at time t from various initial states.

7.4.1 Kolmogorov Forward Equations

Equations (14) and (15) can be combined and written as

$$\lim_{dt \downarrow 0} \frac{1}{dt}(\mathbf{P}_{dt} - \mathbf{I}) = \mathbf{G}.$$

Given that $\mathbf{P}_0 = \mathbf{I}$, this matrix equation amounts to saying that $\mathbf{P}_t$ is differentiable at $t = 0$.

Then, since $\mathbf{P}_{t+dt} = \mathbf{P}_t\mathbf{P}_{dt}$, we have

$$\lim_{dt\downarrow 0} \frac{1}{dt}(\mathbf{P}_{t+dt} - \mathbf{P}_t) = \mathbf{P}_t \lim_{dt\downarrow 0} \frac{1}{dt}(\mathbf{P}_{dt} - \mathbf{I}) = \mathbf{P}_t\mathbf{G}$$

and hence $\mathbf{P}_t$ is differentiable everywhere. So

$$\mathbf{P}'_t = \mathbf{P_t G},$$

where $\mathbf{P}'_t$ denotes the matrix with entries $\frac{d}{dt}p^{ij}(t)$.

Considering the individual entries of these matrices, we find

$$\frac{d}{dt}p^{ij}(t) = \sum_{k=1}^{N} p^{ik}(t)\mu^{kj}. \tag{17}$$

These are known as the Kolmogorov forward equations.

Sometimes it is more helpful to rewrite the forward equations (17) in terms of the meaningful transition intensities; recalling that $\mu^{jj} = -\sum_{k\neq j}\mu^{jk}$, (17) becomes

$$\frac{d}{dt}p^{ij}(t) = \sum_{k\neq j}\left(p^{ik}(t)\mu^{kj} - p^{ij}(t)\mu^{jk}\right).$$

7.4.2 Kolmogorov Backward Equations

Alternatively, in the above we could have conditioned on the state at dt first, i.e. $\mathbf{P}_{t+dt} = \mathbf{P}_{dt}\mathbf{P}_t$. This yields

$$\mathbf{P}'_t = \mathbf{GP_t},$$

or

$$\frac{d}{dt}p^{ij}(t) = \sum_{k=1}^{N} \mu^{ik}p^{kj}(t) \tag{18}$$

or

$$\frac{d}{dt}p^{ij}(t) = \sum_{k\neq i}\left(\mu^{ik}p^{kj}(t) - \mu^{ik}p^{ij}(t)\right).$$

These are the Kolmogorov Backward equations.

The Kolmogorov equations provide expressions for the derivatives of the transition probabilities. The power series solution is:

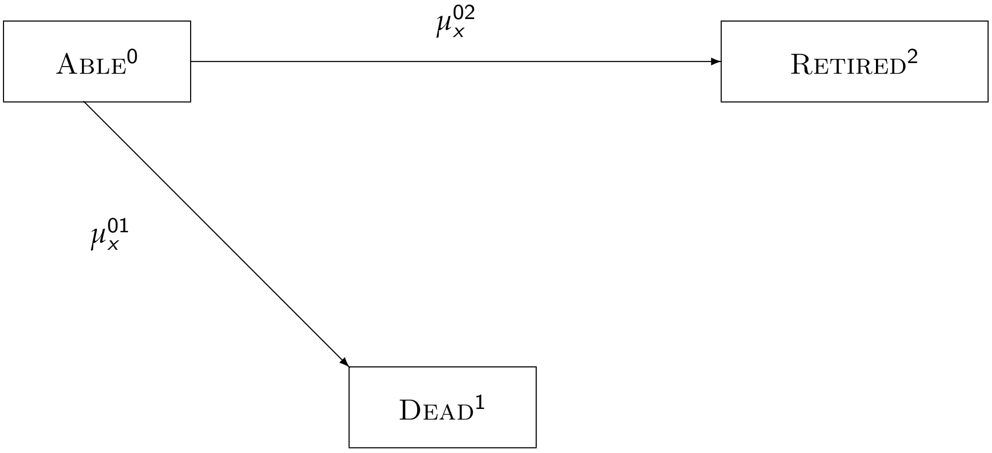
$$\mathbf{P}_t = \exp(\mathbf{G}t) = \mathbf{I} + \mathbf{G}t + \frac{\mathbf{G}^2t^2}{2!} + ...$$

Coupled with the knowledge $\mathbf{P}_0 = \mathbf{I}$, we can use numerical methods to find $\mathbf{P}_t$. Note $\mathbf{P_t G} = \mathbf{P}'_t = \mathbf{GP_t}$, and so the matrices $\mathbf{P_t}$ and $\mathbf{G}$ commute.

7.4.3 Example: The two-decrement model

In general numerical techniques are required to solve the Kolmogorov equations relating to the estimates $p^{ij}(t)$.

However in certain elementary cases the solutions can be written down in closed form. For example, consider the two-decrement model:



Then we can use, say, the Kolmogorov forward equations (17) to obtain first order linear differential equations for the transition probabilities which can be solved analytically using the general result

$$\frac{dy}{dt} = ay + b \quad \Longrightarrow \quad y = Ce^{at} - \frac{b}{a}.$$

We find

$$\begin{aligned}
p^{00}(t) &= e^{-(\mu^{01}+\mu^{02})t} \\
p^{01}(t) &= \frac{\mu^{01}}{\mu^{01}+\mu^{02}}\left[1 - e^{-(\mu^{01}+\mu^{02})t}\right] \\
p^{02}(t) &= \frac{\mu^{02}}{\mu^{01}+\mu^{02}}\left[1 - e^{-(\mu^{01}+\mu^{02})t}\right]. \qquad \text{(Exercise)}
\end{aligned}$$

7.5 Calculating Path Probabilities

7.5.1 Holding times

Clearly from our assumptions for the Markov model, $\forall t$ the path $\{X(s) : 0 \leq s \leq t\}$ is a step function in the state space Ω, making a finite number of jumps with probability 1.

It is useful to consider the waiting times, also known as *holding times*, within states for this step function.

Definition: For an individual entering state i at time s (so $X(s) = i$), we define the holding time to be

$$\inf\{t : t \geq 0, X(s+t) \neq i\}.$$

Definition: Define $p^{\overline{ii}}(t)$ to be the occupancy probability,

$$p^{\overline{ii}}(t) = \Pr(\text{remain in state } i \text{ for at least time } t).$$

Note that this has a different interpretation from $p^{ii}(t)$. $p^{\overline{ii}}(t)$ is the survivor function for the holding time in state i.

In fact, clearly

$$p^{\overline{ii}}(t) \leq p^{ii}(t).$$

By the Markov property we have

$$p^{\overline{ii}}(t+dt) = p^{\overline{ii}}(t)p^{\overline{ii}}(dt),$$

and since the probability of more than one transition in a small time window of length dt is $o(dt)$, we can write

$$p^{\overline{ii}}(t+dt) = p^{\overline{ii}}(t)(1 - \sum_{k\neq i} \mu^{ik}dt + o(dt)).$$

Hence

$$\frac{d}{dt}p^{\overline{ii}}(t) = -p^{\overline{ii}}(t)\sum_{k\neq i}\mu^{ik}.$$

From this we find

$$p^{\overline{ii}}(t) = \exp\left(-t\sum_{j\neq i}\mu^{ij}\right).$$

Alternatively we could write

$$p^{\overline{ii}}(t) = \exp\left(t\mu^{ii}\right),$$

and recall that $\mu^{ii} \leq 0$.

So the holding time for the i^{th} state in a homogeneous Markov jump process is exponentially distributed with rate $-\mu^{ii} = \sum_{j\neq i}\mu^{ij}$.

Notice the time spent in state i before jumping out therefore depends on the sum of the (non-negative) intensities for transition to the other possible states.

The exponential distribution has the 'lack of memory' property

$$\mathrm{P}(T > t+s|T > s) = \mathrm{P}(T > t).$$

That the holding time in any state should exhibit this lack of memory follows intuitively from

- the homogeneity of the process (how old you are is irrelevant)
- and the Markov property (in predicting the future, all that matters is which state you are in now).

7.5.2 Jump probabilities

We have found the distribution for the length of time spent in a state, so it remains to find the distribution for the next state the process jumps to.

Let $p_{ij} = \Pr(X$ jumps from state i to state $j|X$ jumps from state $i)$.

Since the holding time is exponentially distributed with hazard rate $-\mu^{ii}$, from the definition of the hazard function we have

$$\mathrm{P}\{X(t+dt) \neq i|X(t) = i\} = -\mu^{ii}dt + o(dt).$$

Hence for $j \neq i$, we have

$$\mathrm{P}\{X(t+dt) = j|X(t) = i\} = (-\mu^{ii}dt + o(dt))p_{ij}.$$

But also we have

$$\mathrm{P}\{X(t+dt) = j|X(t) = i\} = p^{ij}(dt) = \mu^{ij}dt + o(dt).$$

Setting these two equations equal, dividing by dt and taking limits as $dt \downarrow 0$ thus yields

$$p_{ij} = \frac{\mu^{ij}}{-\mu^{ii}} = \frac{\mu^{ij}}{\sum_{k \neq i} \mu^{ik}}.$$

So a matrix $\mathbf{P}$ with ij^{th} entry p_{ij},

$$(\mathbf{P})_{ij} = \begin{cases} \frac{\mu^{ij}}{\sum_{k \neq i} \mu^{ik}} & i \neq j \\ 0 & i = j, \end{cases}$$

represents the transition probability matrix for the embedded Markov chain of the Markov jump process, looking just at the jumps.

That is, if we were to consider only the transitions of our stochastic process $\{X(t) : t \geq 0\}$, these would form a discrete time Markov chain with transition matrix $\mathbf{P}$.

In summary then, if X is currently in state i:

- The overall magnitude of the transition intensities $\sum\limits_{j \neq i} \mu^{ij}$ controls how long our process stays in state i.
- The relative magnitude of a particular μ^{ij} gives the probability that the next state visited by the process will be j.

7.5.3 Example (continued): The two-decrement model

The transition probabilities derived in §7.4.3 are now easily interpreted:

- $p^{00}(t)$ is the survivor function for holding time in state 0, which we know to be exponentially distributed with rate parameter $\mu^{01} + \mu^{02}$.
- For the other two equations
  - the term in brackets is the probability of jumping out of state 0 by time t;
  - the fraction gives the conditional probability of each decrement having occurred, given that one of them has occurred.

7.6 Estimating Transition Intensities

7.6.1 Maximum Likelihood Estimation

We now consider the problem of making inference about the transition intensities in the presence of some data. Just as in previous chapters, we shall use a maximum likelihood approach.

Suppose we observe n individuals from this Markov model, each for a finite period of time. This will provide n independent partial realisations of the stochastic process $\{X(t) : t \geq 0\}$.

Since our processes are assumed to be Markov and homogeneous, we are able to summarise the information in those observed paths by:

1. The number of transitions $i \to j$, $\forall i \neq j$.
2. The holding times between transitions for each visit to each state i;

What is the contribution to the likelihood of each?

1. The jump probabilities for $i \to j$: $\frac{\mu^{ij}}{-\mu^{ii}}$.

2. (a) An uncensored period of length t (holding time) in state i: $-\mu^{ii} \exp(\mu^{ii} t)$.

   (b) A right-censored period of length t (holding time) in state i: $\exp(\mu^{ii} t)$.

Since these terms are combined multiplicatively, sufficient statistics are the number of transitions of each type (1) and simply the total holding time in each state (2). Let

$$D_k^{ij} = \text{Number of transitions by } k^{\text{th}} \text{ life from state } i \text{ to state } j;$$
$$V_k^{i} = \text{Holding time of the } k^{\text{th}} \text{ life in state } i.$$

Defining the totals across individuals,

$$D^{ij} = \sum_{k=1}^{n} D_k^{ij}, \qquad V^i = \sum_{k=1}^{n} V_k^i,$$

it follows that the likelihood for the transition intensities given observed data $\{v^i, d^{ij}\}$ is

$$L(G) = \prod_{i=1}^{N} \prod_{j \neq i} \exp\{-\mu^{ij} v^i\} (\mu^{ij})^{d^{ij}} \tag{19}$$

where v^i denotes the observed total holding time for state i and d^{ij} denotes the number of observed transitions $i \to j$. Maximising this function yields the maximum likelihood estimators:

$$\hat{\mu}^{ij} = \frac{D^{ij}}{V^i}.$$

The asymptotic properties of the estimators are the same as those from the 2-state model (see §7.3). Asymptotically,

$$\hat{\mu}^{ij} \sim \text{Normal}\left(\mu^{ij}, \frac{\mu^{ij2}}{d^{ij}}\right). \tag{20}$$

and then more approximately

$$\hat{\mu}^{ij} \sim \text{Normal}\left(\mu^{ij}, \frac{d^{ij}}{v^{i2}}\right). \tag{21}$$

The estimators $\hat{\mu}^{ji}$ are not independent of one another; for example, in the Sickness-Death model (Figure 6) D_i^{01} and D_i^{21} are both 0 or 1, but $D_i^{01} D_i^{21} \neq 1$, while (assuming that the i^{th} life starts in the able state) we have $D_i^{02} = D_i^{20}$ or $D_i^{02} = D_i^{20} + 1$.

The estimators are, however, asymptotically independent.

7.6.2 Interval censoring

The calculation of the estimates $\hat{\mu}^{ij}$ requires the total holding time in the i^{th} state to be calculated. When the data are interval censored and aggregated this can cause problems.

However, estimates of the total waiting time can be approximated using numerical techniques. Most commonly, we assume transitions occur half way through the interval and calculate approximate holding times accordingly.

8 Counting Processes and the Poisson Process

We shall meet the concept of *counting processes*, which provide a rich framework for the consideration of survival problems; all of the main ideas we have met in this course can be seen as being embedded within this unifying framework.

8.1 Filtrations and Martingales

Let $\{X(t)\}$ be a stochastic process (see §7.1).

For a continuous-time process $\{X(t)\}$, define

$$X(t-) = \lim_{h \downarrow 0} X(t-h),$$

and for the small time interval $[t, t+dt)$, let

$$dX(t) = X(\{t+dt\}-) - X(t-).$$

If $\{X(t)\}$ were discrete-time indexed, we would define

$$dX(t) = X(t) - X(t-1).$$

Definition: The history (or filtration) $\mathcal{H}_t$ of $\{X(t)\}$ is all that is known in relation to the process up to time t. For example, $\mathcal{H}_t$ could simply be $\{X(s) : 0 \leq s \leq t\}$, but in general it may include other information as well. $\mathcal{H}_{t-}$ is the history of the process up to, but not including, time t.

Technically, $\mathcal{H}(t)$ is defined as the smallest σ-algebra with respect to which all $X(s), s \leq t$ are measurable.

Definition: We will say that the process $\{X(t)\}$ has the Markov property wrt $\mathcal{H}(t)$ if $\forall s, t > 0$,

$$[X(t+s)|\mathcal{H}_t] \equiv [X(t+s)|X(t)].$$

Definition: We will say $\{X(t)\}$ is a martingale with respect to the filtration $\mathcal{H}_t$ if, $\forall t$,

1. $\mathrm{E}[|X(t)|] < \infty$;
2. $\mathrm{E}[X(t)|\mathcal{H}_s] = X_s \ \forall t > s$.

The second condition essentially means $\mathrm{E}[dX(t)|\mathcal{H}_{t-}] = 0$, so a martingale has zero drift.

In the definition above, If instead we have $\forall t\ \mathrm{E}[dX(t)|\mathcal{H}_{t-}] \geq 0$ and so a positive drift (or $\forall t\ \mathrm{E}[dX(t)|\mathcal{H}_{t-}] \leq 0$ and so a negative drift) we say $\{X(t)\}$ is a submartingale (or supermartingale).

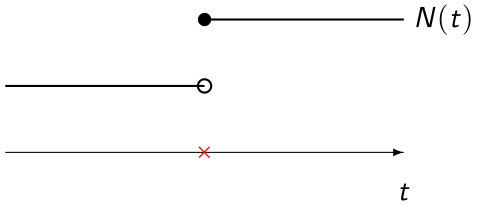
Examples of Martingales:

- Brownian Motion
- Random walk: $t \in \mathbb{N}$, $X_1 = 0$, $X_t = X_{t-1} + \epsilon_t$, $\epsilon_1, \epsilon_t, \ldots$ iid with $\mathbb{E}[\epsilon_i] = 0$.

8.2 Counting Processes

Definition: A stochastic process $\{N(t) : t \geq 0\}$ is called a counting process if

1. $N(0) = 0$;
2. $N(t) \in \mathbb{N} = \{0, 1, 2, ...\}$;
3. $N(t)$ is non-decreasing: $N(s) \leq N(t)$ for $s < t$;
4. $N(t)$ is right continuous: $\lim\limits_{h \downarrow 0} N(t + h) = N(t)$;
5. $N(t)$ has jumps of size at most 1: $N(t) - N(t-) \in \{0, 1\}$;
6. $\mathrm{E}[N(t)] < \infty \ \forall t \geq 0$.



Note that $N(t)$ is non-decreasing and hence a submartingale.

From this definition of a counting process, $N(t)$ can be thought of as recording the number of occurrences of an event from some underlying process up to time t.

Relation between event times and counting process: Let $T_0 = 0$ and $T_1 < T_2 < ...$ be the ordered observed event times of $N(t)$,

$$T_n = \inf\{t : N(t) = n\}, \quad n = 0, 1, ...$$
$$N(t) = \max\{n : T_n \leq t\}.$$

Definition: The *inter-arrival times* are random variables $X_1, X_2, ...$ given by

$$X_n = T_n - T_{n-1}$$
$$T_n = \sum_{i=1}^{n} X_i.$$

8.2.1 Intensity of a Counting Process

General definition:

Definition: Let $N(t)$ be a counting process. Let $\Lambda(t)$ be a stochastic process that is predictable wrt $\mathcal{H}_t$ [ie we know $\Lambda(t)$ given the information in $\mathcal{H}_{\sqcup -}$]. If $N(t) - \Lambda(t)$ is a martingale then $\Lambda(t)$ is called the compensator of N (alternative name: cumulative intensity).

A predctable process $\lambda(t)$ such that $\Lambda(t) = \int_0^t \lambda(s)ds$ is called the intensity of N.

Alternative definition:

Definition: For a counting process $\{N(t)\}$, define the intensity

$$\lambda(t) = \frac{\mathrm{E}[dN(t)|\mathcal{H}_{t-}]}{dt}.$$

The cumulative intensity, $\Lambda(t)$, is then defined to be

$$\Lambda(t) = \int_{s=0}^{t} \lambda(s) ds.$$

$\{\lambda(t)\}$ and $\{\Lambda(t)\}$ are again stochastic processes, but these are predictable (known) given the history process $\mathcal{H}_{t-}$.

Since $d\Lambda(t) = \mathrm{E}[dN(t)|\mathcal{H}_{t-}]$, we have

$$\mathrm{E}[d(N(t) - \Lambda(t))|\mathcal{H}_{t-}] = 0,$$

and hence the process $D(t) = N(t) - \Lambda(t)$ is a martingale, known as the counting process martingale. Writing

$$N(t) = \Lambda(t) + D(t),$$

the first part, $\Lambda(t)$, is called the compensator of the counting process $N(t)$. In contrast with the random step function $N(t)$, and hence also with the martingale $D(t)$, $\Lambda(t)$ is $\mathcal{H}_{t-}$-predictable and varies smoothly over time.

This decomposition of the right continuous submartingale $N(t)$, as the sum of a right continuous, $\mathcal{H}_{t-}$-predictable compensator process and a martingale, is unique by the *Doob-Meyer Decomposition Theorem*[4].

There is an alternative interpretation of the intensity. From the definition above, we have

$$\begin{aligned}
\lim_{dt \downarrow 0} \frac{\mathrm{E}[N(\{t+dt\}-) - N(t-)|\mathcal{H}_{t-}]}{dt} &= \lambda(t) \\
\implies \mathrm{E}[N(\{t+dt\}-) - N(t-)|\mathcal{H}_{t-}] &= \lambda(t)dt + o(dt) \\
\implies \mathrm{P}(N(\{t+dt\}-) - N(t-) = 1|\mathcal{H}_{t-}) &= \lambda(t)dt + o(dt).
\end{aligned}$$

So for a small time increment dt, we have

$$\mathrm{P}(N(\{t+dt\}-) - N(t-) = j|\mathcal{H}_{t-}) = \begin{cases} 1 - \lambda(t)dt + o(dt), & j = 0; \\ \lambda(t)dt + o(dt), & j = 1; \\ o(dt), & j > 1. \end{cases}$$

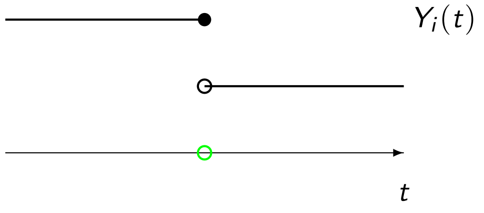
8.3 Counting processes for survival data

Suppose we are to observe the lifetimes of n homogeneous individuals, subject to right-censoring. We first consider the counting process of an individual from the group and then combine these individual processes to consider the overall counting process of the group.

8.3.1 Individual level

Let T_i be the event time of individual i.

Define the indicator $Y_i(t) \in \{0,1\}$ s.t. $Y_i(t) = 1 \iff$ observation of individual i has not been censored before time t and $T_i \geq t$. $\iff$ individual i is at risk at time t



[4] See, for example, Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1995) *Statistical Models Based on Counting Processes.* Springer, New York.

Then $Y_i(t)$ indicates whether individual i is at risk at time t.

The death status of individual i (whether or not he is known to be dead) over time can be treated as a one-jump counting process $N_i(t)$, where

$$N_i(t) = \mathbb{I}(T_i \leq t \cap Y_i(T_i) = 1).$$

The history, $\mathcal{H}_t$, is the combination of $\{N_i(s) : s \leq t\}$ and the censorship status of the individuals. Clearly $Y_i(t)$ will be contained in $\mathcal{H}_{t-}$. (If the individuals were heterogeneous, $\mathcal{H}_t$ might also include measured covariates, which might be static or time-varying.)

Recall in §8.2 we saw the general result

$$\mathrm{P}(N_i(\{t+dt\}-) - N_i(t-) = 1|\mathcal{H}_{t-}) = \lambda_i(t)dt + o(dt),$$

where $\lambda_i(t)$ is the intensity function for $N_i(t)$.

Well here the LHS is simply $\mathrm{P}(\text{Observe } t \leq T_i < t + dt|\mathcal{H}_{t-})$, which is zero if $Y_i(t) = 0$, and hence

$$\mathrm{P}(N_i(\{t+dt\}-) - N_i(t-) = 1|\mathcal{H}_{t-}) = \begin{cases} 0 & \text{if } Y_i(t) = 0, \\ h(t)dt + o(dt) & \text{if } Y_i(t) = 1. \end{cases}$$

where $h(t)$ is the hazard function for the homogeneous population.

Setting the RHS of these two equations equal yields the important relationship

$$\lambda_i(t) = Y_i(t)h(t).$$

Using the compensator representation of a counting process from 8.2,

$$N_i(t) = \int_{s=0}^{t} Y_i(s)h(s)ds + D_i(t) \qquad (22)$$

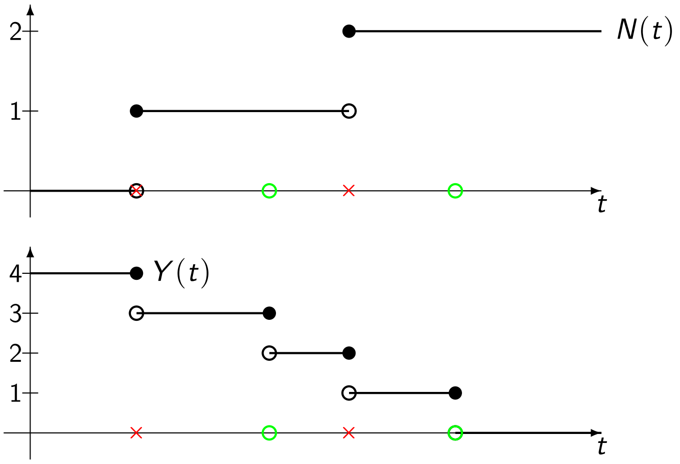
where $D_i(t)$ is a martingale and $M(t)$ is the cumulative hazard function.

8.3.2 Group level

Let

- $N(t) = \sum_{i=1}^{n} N_i(t)$ be the number of events observed up to time t;
- $Y(t) = \sum_{i=1}^{n} Y_i(t)$ be the number of individuals at risk at time t.

Example - Four individuals



Then summing (22) over all individuals gives

$$N(t) = \int_{s=0}^{t} Y(s)h(s)ds + D(t) \tag{23}$$

$$= \int_{s=0}^{t} Y(s)dM(s) + D(t), \tag{24}$$

where $D(t) = \sum_{i=1}^{n} D_i(t)$ is also a martingale.

So from (23) the intensity of the process $N(t)$ is

$$\lambda(t) = Y(t)h(t).$$

Notice that as the number of individuals who have left the study increases, the intensity of the process decreases.

Finally, consider the differential of equation (24):

$$dN(t) = Y(t)dM(t) + dD(t).$$

Then since the conditional expectation of $dD(t)$ given $\mathcal{H}_{t-}$ is zero, for $Y(t) > 0$ we can obtain a moment-based estimator

$$d\hat{M}(t) = \frac{dN(t)}{Y(t)}$$

$$\implies \hat{M}(t) = \int_{s=0}^{t} \frac{dN(s)}{Y(s)}.$$

Noting that $N(t)$ is a step function with unit steps at the death times $\{t_1, t_2, \dots, t_r\}$ say, we recover the Nelson-Aalen estimator

$$\hat{M}(t) = \sum_{t_j \le t} \frac{1}{Y(t_j)}.$$

8.4 Poisson processes

Definition: A Poisson process is a counting process $\{N(t) : t \ge 0\}$ with *independent increments*. That is, $\forall h > 0$ the number of events in the interval $(t, t+h]$,

$$N(t+h) - N(t),$$

is independent of the history $\mathcal{H}_t$.

We can immediately note that applying a process with independent increments to survival data, where necessarily $N(t) \in \{0, 1, \dots, n\}$, provides a serious limitation.

Compare with the required $\lambda(t) = Y(t)h(t)$ form we saw earlier. Now the intensity will not be allowed to depend on $Y(t)$. Instead we simply take

$$\lambda(t) = h(t).$$

This suggests the Poisson process operates as if there is always exactly one individual at risk at any stage in the study; it is as if we observe that individual until exit from the study (through death or censoring), and then immediately commence observation of the next subject, who is of identical age to the previous individual who just exited. The Poisson process is one of the simplest examples of a continuous time Markov process. As a consequence of this simplicity, we are able to derive the distribution of $N(t)$.

Recall the cumulative intensity $\Lambda(t) = \int_{s=0}^{t} \lambda(s)ds$.

Theorem 8.1. *Let $\{N(t) : t \geq 0\}$ be a Poisson process. $\forall t > 0$, $N(t)$ has a Poisson distribution with parameter $\Lambda(t)$. That is,*

$$P(N(t) = j) = e^{-\Lambda(t)} \frac{\Lambda(t)^j}{j!}, \quad j = 0, 1, 2, ...$$

Proof

$$\begin{aligned} \mathrm{P}(N(t+dt) = j) &= \sum_{i=0}^{j} \mathrm{P}(N(t) = i)\mathrm{P}(N(t+dt) = j | N(t) = i) \\ &= \sum_{i=0}^{j} \mathrm{P}(N(t) = i)\mathrm{P}(N(t+dt) - N(t) = j - i) \end{aligned}$$

by the assumption of the Poisson process.

Up to an additive term of $o(dt)$, there will be either zero or one events in the interval $[t, t+dt)$ with respective probabilities $1 - \lambda(t)dt$, $\lambda(t)dt$.

Writing $p_j(t) = \mathrm{P}(N(t) = j)$, we have

$$p_j(t+dt) = \begin{cases} (1 - \lambda(t)dt)p_j(t) + o(dt), & j = 0; \\ (\lambda(t)dt)p_{j-1}(t) + (1 - \lambda(t)dt)p_j(t) + o(dt), & j > 0. \end{cases}$$

Now, subtracting $p_j(t)$ from both side of the equations, dividing by dt and letting $dt \downarrow 0$ we obtain,

$$p_j{\prime}(t) = \begin{cases} -\lambda(t)p_j(t), & j = 0; \\ \lambda(t)\{p_{j-1}(t) - p_j(t)\}, & j > 0. \end{cases} \tag{25}$$

with the boundary condition

$$p_0(0) = 1.$$

These form a collection of differential-difference equations for $p_j(t)$.

The base case $j = 0$ can be solved immediately, to obtain

$$p_0(t) = e^{-\int_{s=0}^{t} \lambda(s)ds} = e^{-\Lambda(t)}.$$

Then for $j > 0$, assuming the inductive hypothesis that the claim is true for $p_{j-1}(t)$, we have

$$p_j{\prime}(t) = \lambda(t)e^{-\Lambda(t)} \frac{\Lambda(t)^{j-1}}{(j-1)!} - \lambda(t)p_j(t).$$

This is a *first-order ordinary differential equation* of the form

$$\frac{dy}{dt} = q(t) - p(t)y,$$

for which the general solution is

$$y = \frac{\int m(t)q(t)dt + k}{m(t)},$$

where $m(t) = \exp\{\int p(t)dt\}$.

Here, $p(t) = \lambda(t)$ giving $m(t) = e^{\Lambda(t)}$, and $q(t) = \lambda(t)e^{-\Lambda(t)}\Lambda(t)^{j-1}/(j-1)!$. Then

$$\begin{aligned} p_j(t) &= e^{-\Lambda(t)} \left\{ \int \lambda(t) \frac{\Lambda(t)^{j-1}}{(j-1)!} dt + k \right\} \\ &= e^{-\Lambda(t)} \left\{ \frac{\Lambda(t)^j}{j!} + k \right\}. \end{aligned}$$

Since $j > 0$, we require $p_j(0) = 0 \implies k = 0$, giving

$$p_j(t) = e^{-\Lambda(t)} \frac{\Lambda(t)^j}{j!}$$

as required. □

8.4.1 Homogeneous Poisson Processes

Definition: A Poisson process is said to be homogeneous if the intensity function is constant.

So when applied to survival data, we are assuming $\lambda(t) = h(t) = h$ and hence a constant hazard. Homogeneity implies the process is independent of the age of the individual in the study.

Note that the cumulative intensity of a homogeneous Poisson process (HPP) $\Lambda(t) = ht$. Hence $\forall t$,

$$N(t) \sim \text{Poisson}(ht).$$

It is easy to check that for a HPP, $\forall s, t > 0$ we also have

$$N(s + t) - N(s) \sim \text{Poisson}(ht).$$

There is an important equivalent formulation of the HPP that underlies the relationship with the homogeneous Markov models of Chapter 7.

Theorem 8.2. *The inter-arrival times $X_1, X_2, \ldots$ of a homogeneous Poisson process with intensity h are independent exponential random variables with rate parameter h.*

Proof

First consider the survivor function for X_1.

$$\text{P}(X_1 > t) = \text{P}(N(t) = 0) = e^{-ht},$$

so $X_1 \sim$Exponential(h).

Now for $j > 1$, conditioning on $\{X_1, \ldots, X_{j-1}\}$ we have survivor

$$\begin{aligned} &\text{P}\{X_j > t | X_1 = t_1, \ldots, X_{j-1} = (t_{j-1} - t_{j-2})\} = \\ &\text{P}\{N(t_{j-1} + t) = N(t_{j-1}) | X_1 = t_1, \ldots, X_{j-1} = (t_{j-1} - t_{j-2})\} \end{aligned}$$

The composite event $\{X_1 = t_1, \ldots, X_{j-1} = (t_{j-1} - t_{j-2})\}$ relates to the interval $[0, t_{j-1}]$, whereas the event $\{N(t_{j-1} + t) - N(t_{j-1}) = 0\}$ relates to the period $(t_{j-1}, t_{j-1} + t]$. By the assumption of the Poisson process, the two events must therefore be independent. And

$$\text{P}(N(t_{j-1} + t) - N(t_{j-1}) = 0) = \text{P}(N(t) = 0) = e^{-ht},$$

hence $X_j \sim$Exponential(h). □

8.4.2 Inference for HPP

We can now see the Poisson model of §10.5 was effectively assuming the death times in our study to be the event times of a HPP observed for a period of time v.

For likelihood inference, we simply have

$$P(N(t) = j) = e^{-ht}\frac{(ht)^j}{j!}, \quad j = 0, 1, 2, ...$$

For a total time on test observation period v, observing d deaths leads to the MLE

$$\hat{h} = \frac{D}{V}.$$

Under this assumed Poisson model, the estimator $\hat{h}$ has

$$E[\hat{h}] = h, \qquad \mathrm{Var}[\hat{h}] = \frac{h}{V}.$$

Note the similarity to estimating the transition intensity in the two-state Markov model of §7.3. We can think of the Poisson model as an approximation to the two state Markov model where the total waiting time (central exposed to risk) V (E_x^c) is considered fixed.

8.5 Counting processes for Markov Models

This secion give some examples that occur in Markov jump processes and their intensities.

As in Chapter 7, we now suppose that we have a sample of n independent individuals such that each individual i is a realisation of a homogeneous continuous time Markov jump process on the state space $\Omega = \{1, \ldots, N\}$.

8.5.1 Individual level

Suppose that as we reach time t individual i currently happens to be in state j; we will now write this as an indicator $Y_i(t, j) = 1$ and $Y_i(t, k) = 0 \; \forall k \neq j \in \Omega$.

If state j is an absorbing state, then we know $h^{jj} = 0$ and individual i will make no further transitions and remain in state j. Otherwise, if state j is not absorbing, then we know from §7.5.1 that the unknown time until individual i will leave state j is governed by the Exponential($-\mu^{jj}$) distribution with constant hazard rate $-\mu^{jj}$.

So by identical reasoning to §8.3.1, if we consider the counting process $N_i(t)$ of the number of transitions of any type made by individual i by time t, this has intensity $\lambda_i(t)$ at time t given by $-\mu^{jj}$. More formally,

$$\lambda_i(t) = -\sum_{j=1}^{N} Y_i(t, j)\mu^{jj}.$$

8.5.2 Group level

By identical reasoning to §8.3.2, if we consider the counting process $N(t) = \sum_{i=1}^{n} N_i(t)$ of the number of transitions of any type made by any of the individuals by time t, this has intensity $\lambda(t)$ at time t given by $\sum_{i=1}^{n} \lambda_i(t)$. More formally,

$$\lambda(t) = -\sum_{i=1}^{n}\sum_{j=1}^{N} Y_i(t, j)\mu^{jj} = -\sum_{j=1}^{N} Y(t, j)\mu^{jj}, \tag{26}$$

where $Y(t,j) = \sum_{i=1}^{n} Y_i(t,j)$ is the number of individuals in state j as we approach time t.

Note that this is a stochastic intensity function as in §8.3.2, but here even under the assumptions of a constant hazard the intensity is no longer necessarily non-increasing; individuals move in and out of different states, altering the state counts $\{Y(t,j)\}$ enabling the intensity to go up and down until individuals start to enter any absorbing states.

8.5.3 Transition level

Alternatively, rather than considering the Markov model on the individual or group level, we can think about counting transitions between particular states. For $j \neq k$, let $N^{(jk)}(t)$ be the number of transitions that have been made from state j to state k by time t by the n individuals.

As argued earlier in §7.6, clearly the collection of processes $\{N^{(jk)}(t) : j \neq k \in \Omega\}$ will be dependent one one another, and so this is an example of a *multivariate counting process*.

For an individual in state j, from (14) the hazard of leaving to state k is μ^{jk}. Alternatively, to obtain this result we have a hazard rate of leaving state j, $-\mu^{jj}$, multiplied by the probability of this jump being to state k, given by the jump probability $-\mu^{jk}/\mu^{jj}$ (see §7.5.2). Hence at time t the process $N^{(jk)}(t)$ has intensity

$$\lambda^{jk}(t) = Y(t,j)\mu^{jk}.$$

Finally, summing over all state pairs (j,k) we recover the intensity (26) for the process $N(t)$ counting the total number of all transitions,

$$\lambda(t) = \sum_{j=1}^{N} \sum_{k \neq j} \lambda^{jk}(t) = \sum_{j=1}^{N} \sum_{k \neq j} Y(t,j)\mu^{jk} = \sum_{j=1}^{N} Y(t,j) \sum_{k \neq j} \mu^{jk}.$$

9 Actuarial Considerations

9.1 Introduction

- Risk = random event with financial consequences

  Insurance = pooling of risk to allow risk sharing, averaging

- How to determine "fair" premiums?

  Cash flow vector: $(c_0, \dots, c_N)$.

  Discount rate $v(k)$

  Present value of cash flow= $\sum_{k=0}^{n} v(k)c_k$

  Example: $v(k) = 1.05^{-k}$.

  "Fair" premium: (expected) cash flow of all payments, premiums, costs, (reasonable profit by insurer) is zero.

- To determine (expected) cash flows need probabilities of insured events happening. Often: conservative estimates are being used to be "prudent".

  Will consider life insurance as example.

9.2 Notation

9.2.1 Actuarial notation

$$\begin{aligned} {}_tq_x &\equiv F_x(t), \\ {}_tp_x &\equiv S_x(t) = 1 - {}_tq_x. \end{aligned}$$

For example ${}_2q_{30} = \mathrm{P}(T_{30} \leq 2) = \mathrm{P}(T \leq 32 | T > 30)$.

When working with units of one year, $t = 1$ is often omitted, i.e. $q_x \equiv {}_1q_x$ and $p_x \equiv {}_1p_x$.

In actuarial science, the CDF ${}_tq_x$ is called the rate of mortality and q_x is called the initial rate of mortality.

9.2.2 Relationship between T_x and T - Actuarial Notation

Continuted from Section 2.2.3. In actuarial notation

$${}_tq_x = \frac{{}_{x+t}q_0 - {}_xq_0}{{}_xp_0} \quad \text{and} \quad {}_tp_x = \frac{{}_{x+t}p_0}{{}_xp_0}.$$

Hence, for any age x, and $\forall s, t > 0$

$$\begin{aligned} {}_{s+t}p_x &= \frac{{}_{x+s+t}p_0}{{}_xp_0} && \left[S_x(s+t) = \frac{S(x+s+t)}{S(x)}\right] \\ &= \frac{{}_{x+s}p_0}{{}_xp_0}\frac{{}_{x+s+t}p_0}{{}_{x+s}p_0} && \left[= \frac{S(x+s)}{S(x)}\frac{S(x+s+t)}{S(x+s)}\right] \\ \implies {}_{s+t}p_x &= {}_sp_x \ {}_tp_{x+s}. && [S_x(s+t) = S_x(s)S_{x+s}(t)] \end{aligned}$$

That is, the probability of surviving for time $s + t$ after age x is simply the probability of surviving for time s multiplied by the probability of surviving for further time t after reaching age $x + s$.

9.3 Complete and Curtate Expectations of Future Lifetime

We have now met several different methods for specifying the distribution of T_x - through the hazard, density, survivor function, etc. Given this distribution, we now consider two expectations of interest.

9.3.1 Complete future lifetime

Definition: The complete expectation of life at age x, $\mathring{e}_x$, is the expected future lifetime

$$\begin{aligned}\mathring{e}_x = \mathrm{E}(T_x) &= \int_0^{\omega-x} t f_x(t) dt \\ &= \int_0^{\omega-x} -t\left(\frac{d}{dt}{}_t p_x\right) dt \\ &= -[t\ {}_t p_x]_0^{\omega-x} + \int_0^{\omega-x} {}_t p_x dt \\ &= \int_0^{\omega-x} {}_t p_x dt.\end{aligned}$$

In the case of human mortality $\mathring{e}_x$, and in particular $\mathring{e}_0$, is used as a measure of health both for comparison between different populations and for trends within a single population.

Note that for $0 \le x_1 < x_2$, $\mathring{e}_{x_2}$ is not necessarily less than $\mathring{e}_{x_1}$. For example, there may be a high rate of infant mortality.

9.3.2 Curtate future lifetime

Suppose once more that we have an individual who has survived to age x.

As a discretised alternative to the continuous valued future lifetime T_x, we can consider the number of whole years K_x subsequently survived, so $K_x = \lfloor T_x \rfloor$; that is, the integer part of T_x (rounded down).

K_x is a discrete random variable on $\{0, 1, \dots, \lfloor \omega - x \rfloor\}$ with probability mass function (pmf)

$$\begin{aligned}\mathrm{P}(K_x = k) &= \mathrm{P}(k \le T_x < k+1) \\ &= {}_{k+1}q_x - {}_k q_x \\ \text{or } &= {}_k p_{x\,1} q_{x+k}.\end{aligned}$$

It is simple to check these are equivalent (Exercise).

We can use this pmf to calculate the expectation of K_x.

Definition: The curtate expectation of life at age x, e_x, is thus defined by

$$\begin{aligned}e_x = \mathrm{E}(K_x) &= \sum_{k=0}^{\lfloor \omega-x \rfloor} k\, {}_k p_{x\,1} q_{x+k} \\ &= \sum_{k=1}^{\lfloor \omega-x \rfloor} {}_k p_x. \qquad \text{(Exercise)}\end{aligned}$$

Again we might also be interested in the variance

$$\mathrm{Var}(K_x) = \sum_{k=0}^{\lfloor \omega-x \rfloor} k^2\, {}_k p_{x\,1} q_{x+k} - {e_x}^2.$$

9.4 Random Survivorship Group

Suppose we observe a group of ℓ_0 newborn lives sampled from our homogeneous population. That is, ℓ_0 (not necessarily independent) realisations of T.

We might then be interested in the number of individuals still alive at age x, $n(x)$, for some $x > 0$.

Since each individual will survive beyond age x with probability ${}_x p_0$, *if the lifetimes were independent* we would have

$$n(x) \sim \text{Binomial}(\ell_0, {}_x p_0).$$

In any case, the expected number of survivors at x, denoted ℓ_x, is certainly given by

$$\ell_x = \ell_0 \; {}_x p_0.$$

(ℓ_0 is referred to as the radix.)

Furthermore, it then follows that the expected number of deaths in the group between ages x and $x + t$, written ${}_t d_x$, is given by

$$\begin{aligned} {}_t d_x &= \ell_x - \ell_{x+t} \\ &= \ell_0 \{ {}_x p_0 - {}_{x+t} p_0 \}. \end{aligned}$$

Again when $t = 1$ we simply write d_x.

From the definition of the hazard function $h(x)$ and (1), for small t we also have

$$ {}_t d_x \approx \ell_x h(x) t.$$

A plot of $\ell_x \mu(x)$ $(= \ell_0 f(x))$ against x is known as the curve of deaths (!).

9.5 Life Tables

Each year the Office for National Statistics (ONS, `https://www.ons.gov.uk/`) produces a new set of estimates of the distribution of mortality in the UK, based on recorded deaths in the country over the most recent three year period. These estimates are presented in the form of life tables (The most recent numbers are `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables` for the UK. Life tables such as these which are based on large amounts of data and can therefore be considered reliable are known as standard tables.

The life data used consider only the integer age at death, and so continuous concepts such as the density and hazard functions cannot be directly estimated. Instead, the statistics are presented as those from a hypothetical survivorship group of initial size ℓ_0=100,000.

The life tables thus contain estimates for the following discrete quantities for each integer age $x \in \{0, 1, 2, \ldots, \omega\}$.

ℓ_x - as above, the expected number of survivors from the ℓ_0 newborns at x.

d_x - as above, the expected number of individuals who die between ages x and $x + 1$.

q_x - the initial rate of mortality, which is the probability of an individual who has reached age x dying before age $x + 1$. So $q_x = d_x/\ell_x$.

m_x - the central rate of mortality, is the death rate 'per year lived'. Those who die during the year can be assumed to live for half of the year, so $m_x = q_x/(1 - q_x/2)$. m_x is approximately the average force of mortality over $[x, x + 1[$ (for small q_x).

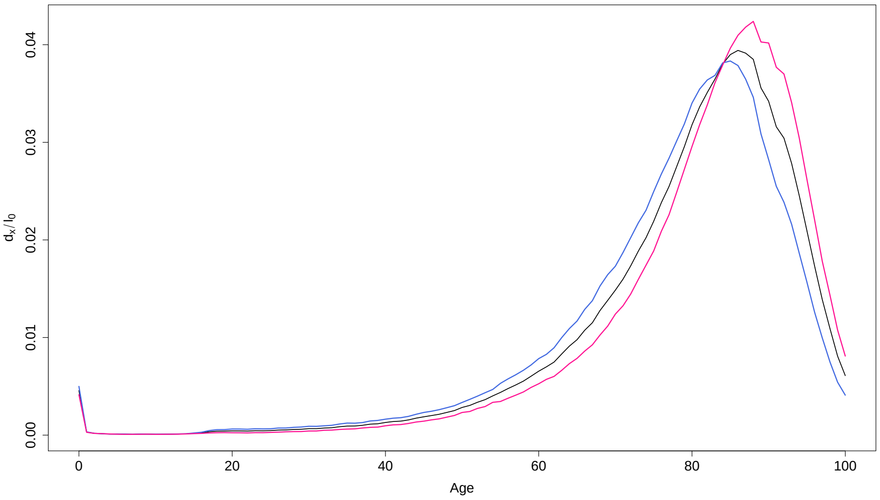
e_x - the period life expectancy for an individual who has survived to x, which is the curtate expectation of life using the estimates q_x as our probabilities.

Interim Life Tables

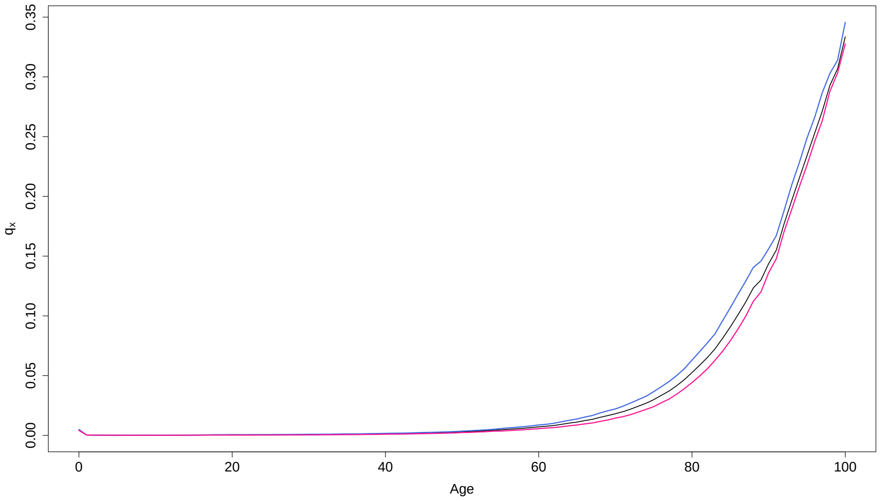
AGE x	m_x	q_x	ℓ_x	d_x	e_x
0	0.005006	0.004993	100000.0	499.3	78.05
1	0.000335	0.000335	99500.7	33.4	77.44
2	0.000189	0.000189	99467.3	18.8	76.46
3	0.000144	0.000144	99448.5	14.3	75.48
4	0.000107	0.000107	99434.2	10.6	74.49
5	0.000119	0.000119	99423.6	11.9	73.50
6	0.000112	0.000112	99411.7	11.1	72.51
7	0.000087	0.000087	99400.6	8.7	71.51
8	0.000115	0.000115	99391.9	11.4	70.52
9	0.000102	0.000102	99380.5	10.2	69.53
10	0.000098	0.000098	99370.3	9.7	68.54
$\vdots$	...	...	...	...	...

Table 1: Extract from table based on UK data for males for the years 2008-2010.

d_x/ℓ_0, 2008/10 ONS data



q_x, 2008/10 ONS data



10 The Binomial and Poisson Models of Mortality

The main uses of Binomial and Poisson models lie in actuarial science. Like our use of Markov processes in Chapter 7, these models are concerned with survival within the unit time intervals $[x, x+1)$. (This is a characteristic particular to actuarial modelling, as opposed to, say, engineering or medical survival analyses.)

The Binomial model is also valuable as it underpins the the Kaplan-Meier nonparametric estimator in Chapter 5.

10.1 Binomial model of mortality

In the analysis of mortality data we are typically presented with the following binary data: A sample of n homogeneous individuals are monitored over the age range $[x, x+1)$. The observed outcome for individual i, D_i, is either death ($D_i = 1$) or survival ($D_i = 0$). Define $D = \sum_{i=1}^{n} D_i$ as the total number of deaths.

Using such data, we would like to make inference about the probability of death for a new individual during the age range $[x, x+1)$.

We defined the cdf ${}_tq_x$ as the probability of an individual currently aged x dying within time t. For notational brevity we also took $q_x \equiv {}_1q_x$. Then q_x is precisely the quantity on which we wish to make inference.

If there is no censoring, so that the death of any individual in our sample can be observed anywhere during $[x, x+1)$, then each life indicator D_i is the outcome of a Bernoulli trial with probability of death q_x and survival probability $p_x = 1 - q_x$.

In which case, assuming the lifetimes are independent, $D \sim \text{Binomial}(n, q_x)$. That is,

$$\mathrm{P}(D = d) = \binom{n}{d} q_x^d (1 - q_x)^{(n-d)}.$$

This is the Binomial model of mortality in its simplest form. Note that $\mathrm{E}(D) = nq_x$, $\mathrm{Var}(D) = nq_x(1 - q_x) = nq_xp_x$.

The maximum likelihood estimator of q_x is

$$\hat{q_x} = \frac{D}{n},$$

which has mean and variance

$$\begin{aligned} \mathrm{E}(\hat{q_x}) &= \frac{\mathrm{E}(D)}{n} = \frac{nq_x}{n} = q_x, \\ \mathrm{Var}(\hat{q_x}) &= \frac{\mathrm{Var}(D)}{n^2} = \frac{q_x(1-q_x)}{n} = \frac{q_xp_x}{n}. \end{aligned}$$

Asymptotically, by the Central Limit Theorem we have

$$\hat{q_x} \approx \text{Normal}\left(q_x, \frac{q_x(1-q_x)}{n}\right).$$

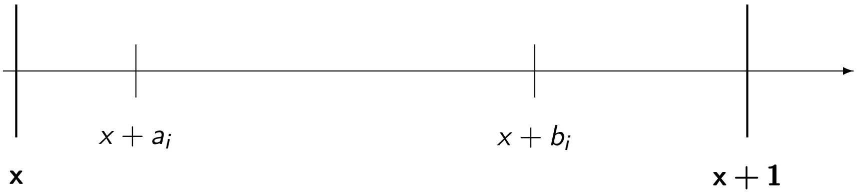
[This can also be derived in the usual MLE/information matrix way (see §4.1.2) by noting that for large n, $q_x \approx D/n$. (Exercise)]

Given a suitable set of data the probability of death in any interval can be trivially estimated in this way. However, the Binomial model can be problematic under more realistic circumstances, where we may observe individuals over different (incomplete) sub-intervals of $[x, x+1)$; there will usually be multiple decrements (individuals leaving the process) and sometimes increments (individuals joining the process) during $[x, x+1)$.

10.2 General Binomial Model - Non-uniform observation times

Again consider observing n independent lives over the interval $[x, x+1)$.

Now we assume that individual i comes into view at time $x + a_i$ and is unobserved beyond time $x + b_i$, when it is then deemed to be right-censored if still alive, for $0 \leq a_i < b_i \leq 1$.



Then for each individual i, $D_i \sim \text{Bernoulli}({}_{b_i-a_i}q_{x+a_i})$.

Defining a vector of the individual death probabilities

$$\underset{\sim}{\mathbf{q}} = ({}_{b_1-a_1}q_{x+a_1}, {}_{b_2-a_2}q_{x+a_2}, \dots, {}_{b_n-a_n}q_{x+a_n}),$$

the full likelihood function is

$$L(\underset{\sim}{\mathbf{q}}) = \prod_{i=1}^{n} ({}_{b_i-a_i}q_{x+a_i})^{d_i} (1 - {}_{b_i-a_i}q_{x+a_i})^{(1-d_i)}. \tag{27}$$

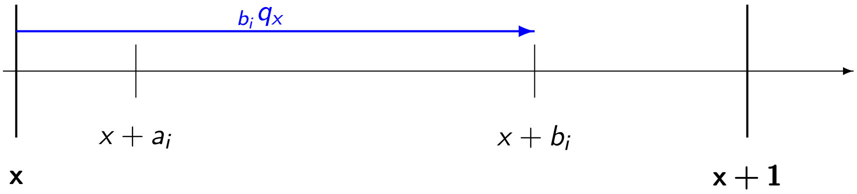
Clearly, estimating the n values of $\underset{\sim}{\mathbf{q}}$ via an unconstrained maximisation of (27) is inappropriate since if the (a_i, b_i) pairs are unique then the MLEs are degenerate:

$${}_{b_i-a_i}\hat{q}_{x+a_i} = d_i \in \{0, 1\}.$$

Furthermore, for a new individual with unique entry and exit times $(x + a_{n+1}, x + b_{n+1})$ we are no further forward in being able to make predictions.

The solution is to assume that the probabilities $\{{}_tq_x\}$ vary smoothly as a function of t, $0 \leq t \leq 1$, thus creating a dependence structure between individuals in the estimation process. To achieve this dependence, the following sections provide three popular alternative methods for *fractional age adjustment*, each reducing the inference problem to that of estimating the single parameter q_x.

But first, as a preliminary step to verify that it is sufficient to construct models for $\{{}_tq_x : 0 \leq t \leq 1\}$ to fully specify more general probabilities ${}_{b_i-a_i}q_{x+a_i}$, consider the probability ${}_{b_i}q_x$:



Clearly,

$${}_{b_i}q_x = {}_{a_i}q_x + (1 - {}_{a_i}q_x)\,{}_{b_i-a_i}q_{x+a_i}.$$

Rearranging then gives

$${}_{b_i-a_i}q_{x+a_i} = \frac{{}_{b_i}q_x - {}_{a_i}q_x}{1 - {}_{a_i}q_x},$$

an equation for the required probabilities ${}_{b_i-a_i}q_{x+a_i}$ in terms of probabilities of the type ${}_tq_x$.

10.2.1 Uniform distribution of deaths (UDD)

The UDD model states that any deaths are uniformly distributed over the range $[x, x+1)$. This implies a conditional cdf $\mathrm{P}(T_x \leq t | T_x < 1) = t$ for $0 \leq t < 1$. Then since $\mathrm{P}(T_x < 1) = q_x$,

$$ {}_tq_x = tq_x, \qquad 0 \leq t \leq 1. $$

Assuming UDD implies there is an *increasing* force of mortality (hazard) between integer ages. To see this, first consider the survivor function

$$ \begin{aligned} {}_tp_x &= 1 - {}_tq_x \\ &= 1 - tq_x. \end{aligned} $$

This survivor function implies a hazard rate of

$$ \begin{aligned} \implies \mu_x(t) &= -\frac{d}{dt}\log\left({}_tp_x\right) = -\frac{d}{dt}\log(1 - tq_x) \\ &= \frac{q_x}{1 - tq_x}, \end{aligned} $$

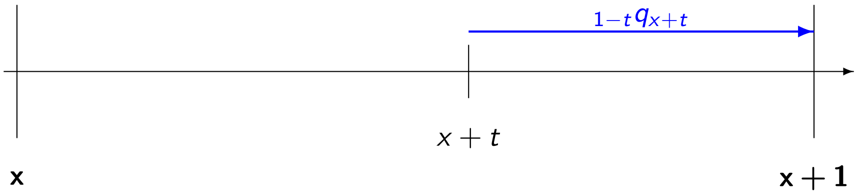
and clearly $\mu_x(t) \uparrow$ as $t \uparrow$.

10.2.2 Balducci assumption

For $q_x < 1$, the *Balducci* (or *hyperbolic*) assumption is

$$ {}_{1-t}q_{x+t} = (1-t)q_x, \qquad 0 \leq t \leq 1. $$

This model assumes a *decreasing* force of mortality. Too see this, consider partitioning the interval $[x, x+1)$ in two at time t.



Then clearly

$$ q_x = {}_tq_x + (1 - {}_tq_x)\,{}_{1-t}q_{x+t}. $$

Rearranging, we get

$$ {}_tq_x = \frac{q_x - {}_{1-t}q_{x+t}}{1 - {}_{1-t}q_{x+t}}. $$

So under the Balducci Assumption,

$$ \begin{aligned} {}_tq_x &= \frac{q_x - (1-t)q_x}{1 - (1-t)q_x} \\ &= \frac{tq_x}{1 - (1-t)q_x} = 1 - \frac{1 - q_x}{1 - (1-t)q_x}. \end{aligned} $$

(Note that, in contrast with UDD, here ${}_tq_x > tq_x$ for $t < 1$.)

Hence the survivor function

$$ {}_tp_x = \frac{1 - q_x}{1 - (1-t)q_x}, $$

and so the hazard function

$$ \begin{aligned} \mu_x(t) &= -\frac{d}{dt}\log({}_tp_x) = -\frac{d}{dt}\log(1 - q_x) + \frac{d}{dt}\log(1 - q_x + tq_x) \\ &= \frac{q_x}{1 - q_x + tq_x} \end{aligned} $$

which $\downarrow$ as $t \uparrow$.

10.2.3 Constant force of mortality

In the previous two models we have seen an increasing and then a decreasing force of mortality over the unit interval.

To achieve a constant hazard we would need

$$-\frac{d}{dt}\log({}_tp_x) = \mu_x$$

for some $\mu_x > 0$ constant w.r.t. t.

Integrating this equation, we get

$$\log({}_tp_x) = -\mu_x t$$

and hence

$${}_tq_x = 1 - \exp(-\mu_x t), \qquad 0 \leq t \leq 1. \tag{28}$$

Through the special case of (28) with $t = 1$, we see μ_x would be related to q_x by the equation $q_x = 1 - \exp(-\mu_x)$. Substituting this back into (28), we can equivalently write the constant force of mortality assumption in terms of our quantity of interest q_x by

$${}_tq_x = 1 - (1 - q_x)^t, \qquad 0 \leq t \leq 1.$$

This is the required fractional age adjustment model for ${}_tq_x$ for a constant force of mortality.

10.2.4 Comparison of the three assumptions

Under any of the three assumptions, the likelihood function (27) reduces to a (complicated) function of the single parameter q_x, $L(q_x)$. Numerical procedures are generally required to maximise $L(q_x)$ for a given data set.

Figure 7 compares the three modelling assumptions for varying values of q_x. For the small values of q_x typically found in actuarial applications at least for younger ages x, the difference between these models diminishes.

10.3 The actuarial estimate

The Balducci assumption can be used to provide a theoretical justification of the actuarial estimate of the survivor function we derived in §5.5.

For simplicity we assume that the $\{(a_i, b_i)\}$ are known. First we notice

$$\begin{aligned}
{}_{1-a_i}q_{x+a_i} &= {}_{b_i-a_i}q_{x+a_i} + (1 - {}_{b_i-a_i}q_{x+a_i})\,{}_{1-b_i}q_{x+b_i} \\
\implies {}_{b_i-a_i}q_{x+a_i} &= {}_{1-a_i}q_{x+a_i} - (1 - {}_{b_i-a_i}q_{x+a_i})\,{}_{1-b_i}q_{x+b_i} \\
\implies \mathsf{E}[D_i] &= {}_{1-a_i}q_{x+a_i} - (1 - \mathsf{E}[D_i])\,{}_{1-b_i}q_{x+b_i}
\end{aligned}$$

So under the Balducci assumption,

$$\mathsf{E}[D] = \sum_{i=1}^{n} \mathsf{E}[D_i] = \sum_{i=1}^{n}(1 - a_i)q_x - \sum_{i=1}^{n}(1 - \mathsf{E}[D_i])(1 - b_i)q_x.$$

Substituting the observed number of deaths d on the left hand side would provide an orthodox *moment estimate* of q_x. But $\mathsf{E}[D_i]$ depends on q_x and is also unknown.

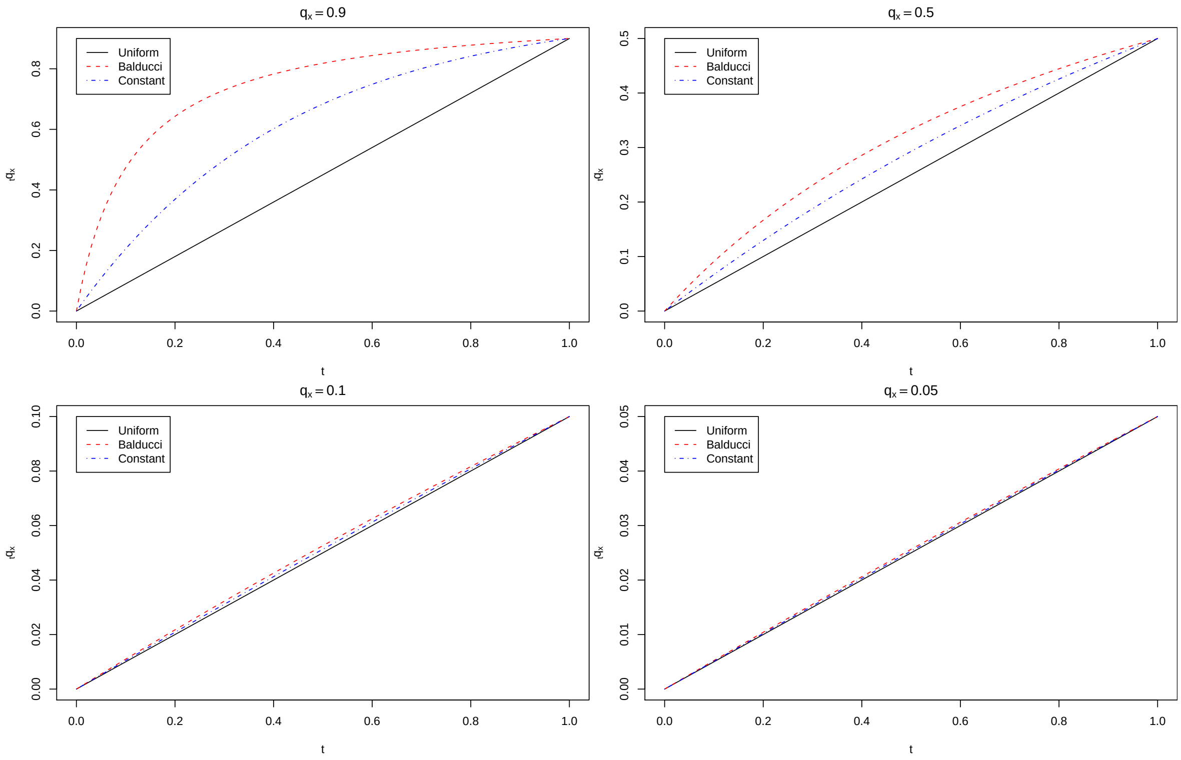


Figure 7: Comparison of uniform, Balducci and constant hazard assumptions for $q_x = 0.9, 0.5, 0.1, 0.05$.

So we also crudely estimate $\mathrm{E}[D_i]$ with the observed value d_i, leading to

$$d \approx \sum_{i=1}^{n}(1-a_i)q_x - \sum_{i=1}^{n}(1-d_i)(1-b_i)q_x.$$

Rearrangement yields the actuarial estimate (*cf.* §5.5) for q_x, now in a more general form,

$$\hat{q}_x = \frac{d}{E_x}, \tag{29}$$

where

$$\begin{aligned} E_x &= \sum_{i=1}^{n}(1-a_i) - \sum_{i:d_i=0}(1-b_i) \\ &\left(= \sum_{i:d_i=1}(1-a_i) + \sum_{i:d_i=0}(b_i-a_i)\right). \end{aligned} \tag{30}$$

So for the binomial model of mortality, the actuarial estimate (29) can be interpreted as the MLE under an approximation that

$$D \sim \mathrm{Binomial}\left(E_x, q_x\right). \tag{31}$$

10.3.1 Actuarial notation

- E_x in equation (30) is known as the *initial exposed to risk*.
  - E_x counts deaths as having been 'exposed' for a duration of $(1-a_i)$.

- Recall the total time under observation, v, is known as the *central exposed to risk*, E_x^c (see §7.3).
  - If the death times $\{x + t_i\}$ had been observed, then notice that

$$E_x = E_x^c + \sum_{i:d_i=1} (1 - t_i).$$

Finally, suppose that the individual pairs $\{(a_i, b_i)\}$ are unobserved but that we have a good estimate of the overall central exposed to risk E_x^c. Under the crude assumption that deaths occur, on average, at age $x + \frac{1}{2}$, and ignoring the awkward possibility that $a_i > \frac{1}{2}$, we obtain the formula

$$E_x \approx E_x^c + \frac{1}{2}d$$

and hence from (29),

$$\hat{q_x} = \frac{d}{v + \frac{1}{2}d}.$$

The actuarial estimate is only an approximation to a method-of-moments estimator. However, it does avoid numerical solutions of equations and for reasonably small $q_x < 0.3$ it is often in agreement with other more statistically sound procedures.

10.4 Statistical testing of estimates

From (31) we have an expected number of deaths in $[x, x+1)$ of $E_x q_x$. Furthermore, by the Central Limit Theorem we have approximately

$$D \sim \text{Normal}\left(E_x q_x, E_x q_x (1 - q_x)\right).$$

We have already seen earlier how the Binomial model can be simply extended to obtain the non-parametric product-limit estimate (§5.2) which is widely used in survival analysis.

However, perhaps the crucial weakness is that the Binomial model is difficult to generalise to settings with more than one outcome. Even the simplest three state model with two decrements (see §7.4.3) is hard to model using multinomial ideas and the general N-state models are harder still.

10.5 Poisson model of mortality

The Poisson distribution is a discrete distribution on the natural numbers $\mathbb{N} = \{0, 1, 2, ...\}$ used to model the number of rare events occurring during a unit interval of time when the risk of occurrence of these events is constant, e.g. particles emitted by a radioactive source.

For actuarial applications the Poisson distribution can be used as a model for the number of deaths among a group of n individuals.

We once again consider fitting a separate model for each unit interval age group $[x, x+1)$, with individual i being observed within a specific sub-interval $[x + a_i, x + b_i)$.

Let v denote the realisation of the total time spent observing the n individuals. (Recall $v \equiv E_x^c$, the central exposed to risk.)

As in §3.2, §7.3 and §10.2.3, we assume that there is a force of mortality (hazard rate) μ_x which is constant over $[x, x+1)$.

The Poisson model for mortality then states that conditional on v the the total number of deaths D follows a Poisson distribution with parameter $\mu_x v$. That is,

$$\mathrm{P}(D = d) = e^{-\mu_x v} \frac{(\mu_x v)^d}{d!}, \qquad d = 0, 1, 2, ...$$

An obvious flaw in this model is that $\mathrm{P}(D > n) > 0$. However, for large n and small $\mu_x v$ it can be a good approximation.

The Poisson model can be seen to approximate the Binomial model; setting $\lambda = nq_x$, under the Binomial(n, q_x) model

$$\begin{aligned}
\mathrm{P}(D = d) &= \frac{n!}{(n-d)!d!} {q_x}^d (1 - q_x)^{n-d} \\
&= \frac{1}{d!} \frac{n!}{(n-d)!n^d} (nq_x)^d \left(1 - \frac{nq_x}{n}\right)^{n-d} \\
&= \frac{\lambda^d}{d!} \frac{n!}{(n-d)!n^d} \left(1 - \frac{\lambda}{n}\right)^{n-d} \\
&= \frac{\lambda^d}{d!} \prod_{i=1}^{d} \left(\frac{n-i+1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{n-d} \\
&\to \frac{\lambda^d}{d!} e^{-\lambda} \qquad \text{as } n \to \infty,
\end{aligned}$$

where the last step makes use of the fact that

$$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.$$

Now here in the Poisson model of mortality we have taken λ to be $\mu_x v$. To see that $nq_x \approx \mu_x v$, firstly note that the LHS is the expected number of deaths from n individuals in $[x, x+1]$. Then secondly recall from §4.2.1 or §7.3 that the MLE for a constant hazard is d/v.

11 Comparison of Markov, Binomial and Poisson models

11.1 Characteristics

The Markov, Binomial and Poisson models are all used in actuarial science for estimating survival probabilities for each integer age group $[x, x+1)$.

The Binomial model assumes a restricted view where only the year of death of individuals is recorded, rather than the exact event times.

In contrast the Markov and Poisson models both require knowledge of the exact death times; the Poisson model can be seen as an approximation to the Markov model where the total observation time (central exposed to risk) is considered fixed.

11.2 Inference

If the exact death times, entry times and exit times on the study are known for all individuals, then the MLE for the Markov and Poisson models are analytically available and structurally identical; otherwise the central exposed to risk must be separately estimated.

If the total observation time were fixed, then the Poisson/Markov MLE would be unbiased; however this condition is unlikely to be satisfied in actuarial studies, in which case we have seen that this MLE is only asymptotically unbiased.

For the Binomial model we have seen the need for additional modelling techniques to make use of such rich data, such as the Balducci assumption. For low hazard rates, we have seen the differences between these models to be small, but only approximate results are available for assessing their estimates.

11.3 Generality

The Markov model extends naturally to an arbitrary number of states. The Poisson model, as an approximation of the Markov model, can also be adapted to more general problems. The binomial model does not extend easily.

Multivariate counting processes generalise the concepts of Chapter 8, with a vector of counting processes $\mathbf{N}(t) = (N_1(t), \dots, N_h(t))$ counting different event types but sharing a common filtration $\mathcal{H}_t$.

11.4 Conclusions

When the force of mortality/hazard of the population is low (such as in human mortality) the estimates we obtain are fairly robust to the different choice of models.

However, when we are to model more complicated processes or higher transition intensities, which is increasingly the case as more complex insurance products are developed, the stochastic process models appear to offer significant advantages; albeit with a higher computational burden.

12 Graduation

12.1 Requirements for Graduated Estimates

In Chapters 7-11 we have considered actuarial models for mortality over single time units (years) for individuals ageing from x to $x+1$. For each age interval $[x, x+1)$ we have separately calculated crude estimates $\{\hat{q}_x\}$ (Binomial model) or $\{\hat{\mu}^{ij}_{x+\frac{1}{2}}\}$ (Markov or Poisson models). Since these crude estimates are formed independently of one another, a plot of them is likely to be quite rough.

Intuitively, we would typically expect q_x or $\mu^{ij}_{x+\frac{1}{2}}$ to be smooth, possibly monotonic functions of age x. Reporting non-smooth estimates of such a function could be hard to justify. Furthermore, estimates of q_{x-1} or q_{x+1}, for example, should carry some information about q_x as well, and so we should be able to use this assumed smoothness to improve our estimators.

The process of smoothing crude actuarial estimates is called graduation. The resulting graduated estimates, denoted $\{\mathring{q}_x\}$ or $\{\mathring{\mu}^{ij}_{x+\frac{1}{2}}\}$, should be relatively smooth as a function of x but still sufficiently close to the crude estimates $\{\hat{q}_x\}$, $\{\hat{\mu}^{ij}_{x+\frac{1}{2}}\}$ so that they continue to adhere to the observed data. Clearly this represents a trade-off of conflicting requirements and a good balance needs to be struck.

Additionally there might be prior knowledge about the populations which the parameters should adhere to; for example, that mortality is higher in males than females, or that overall mortality is lower than it used to be. Or there might be practical considerations we wish to judge the smoothed estimates against. In actuarial science, there are two important examples:

- In life insurance, underestimating mortality leads to losses;
- In pensions, overestimating mortality leads to losses.

12.2 Graduation Methods

12.2.1 Parametric Graduation

The most common graduation methods fit a parametric model to the crude estimates or the raw data. The Gompertz-Makeham distribution (§3.4) provides a useful smoothing structure for transition intensities which in its most general form is used as

$$\mathring{\mu}^{ij}_x = r^{ij}(x) + \exp\{s^{ij}(x)\}$$

where r^{ij} and s^{ij} are polynomials. Joint likelihood analysis using (19) across age groups then becomes a question of estimating the coefficients of the polynomials r^{ij} and s^{ij}.

Often it will be difficult to find a sufficiently simple parametric model which will model the data well in all of the *age* groups. ML estimation of a parametric graduation model will typically result in a good fit where there is most data, which in practice usually means at middle ages, so at extreme ages the fit will be least reliable and may require adjustment.

12.2.2 Graduation Using Previous Estimates

Given a set of smooth, robust prior estimates $\{q_{0,x}\}$ or $\{\mu^{ij}_{0,\ x+\frac{1}{2}}\}$, perhaps obtained from standard tables (see §9.5), we might wish to bias our new estimates towards these previous values so that the new estimates are similar or related to the old ones. For example, we might stipulate that $\forall x$

$$\mathring{q}_x = a + bq_{0,x}$$

or

$$\mathring{\mu}_x^{ij} = \mu_{0,\ x+k}^{ij}$$

for constants a, b, k.

The precise parametric form for this relationship might be chosen from some exploratory plots of the crude estimates against the prior estimates. Once chosen, parameter estimates can then be found via ML estimation as in §12.2.1.

This form of graduation would not be favoured when large amounts of current data are available, but has the advantage of performing well in extreme age groups where we may have very little data.