COURSE ORGANISATION

- Week: Video Lectures, Live Lectorials, Tutorials
- Assessments:
  - Weekly formative activities – 10%
  - Main Assignment – 25%
  - Final Exam – 65%

M0: Revision

LEAST SQUARES ESTIMATORS

- Consider a single factor model for a log transformed survival time:

$$\log T_i = \beta_0 + \beta_1 x_{i1} + \sigma\varepsilon_i$$

We have

$$\log T_i - [\beta_0 + \beta_1 x_{i1}] = \sigma\varepsilon_i$$

Now select β_0, and β_1 so that

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} [\log T_i - [\beta_0 + \beta_1 x_{i1}]]^2$$

is minimised (how?)

- Partial differentiation gives:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} [\log T_i - [\beta_0 + \beta_1 x_{i1}]]$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_{i1} [\log T_i - [\beta_0 + \beta_1 x_{i1}]]$$

Set these to zero and solve for (least squares) estimators of β_0 and β_1

Solution is

$$\tilde{\beta}_0 = \frac{\left[\begin{array}{c} \left(\sum_{i=1}^{n} x_{i1}^2\right)\left(\sum_{i=1}^{n} \log T_{i1}\right) \\ -\left(\sum_{i=1}^{n} x_{i1}\right)\left(\sum_{i=1}^{n} x_i \log T_{i1}\right) \end{array} \right]}{n\sum_{i=1}^{n} x_{i1}^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}$$

$$\tilde{\beta}_1 = \frac{\left[\begin{array}{c} n\left(\sum_{i=1}^{n} x_i \log T_{i1}\right) \\ -\left(\sum_{i=1}^{n} x_{i1}\right)\left(\sum_{i=1}^{n} \log T_{i1}\right) \end{array} \right]}{n\sum_{i=1}^{n} x_{i1}^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}$$

PROPERTIES OF LEAST SQUARES ESTIMATORS

Assuming ε_i are independent random variables with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = 1$ and x_{i1} fixed (not random)

$$E[\tilde{\beta}_0] = \beta_0$$

$$E[\tilde{\beta}_1] = \beta_1$$

$$Var\left(\tilde{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} (x_{i1})^2}{n\sum_{i=1}^{n} (x_{i1})^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}$$

$$Var\left(\tilde{\beta}_1\right) = \frac{n\sigma^2}{n\sum_{i=1}^{n} (x_{i1})^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}$$

An (unbiased) estimate of σ^2 is given by

$$s^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n} \left[\log T_i - \left[\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1}\right]\right]^2}{n-2}$$

where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are least squares estimates of β_0 and β_1

- Normal Assumption:

  If ε_i has a standard normal distribution i.e. $N(0,1)$ then T_i is log-normal

  Standard error of estimates

$$s_{\tilde{\beta}_0} = \left(\frac{s^2 \sum_{i=1}^{n} (x_{i1})^2}{n\sum_{i=1}^{n} (x_{i1})^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}\right)^{\frac{1}{2}}$$

$$s_{\tilde{\beta}_1} = \left(\frac{ns^2}{n\sum_{i=1}^{n} (x_{i1})^2 - \left(\sum_{i=1}^{n} x_{i1}\right)^2}\right)^{\frac{1}{2}}$$

TESTING

- CLT implies that:

$$\frac{\tilde{\beta}_j - \beta_j}{s_{\tilde{\beta}_j}} \sim t_{n-2}$$

- We can determine p-values for tests of $\beta_j = 0$ (i.e. probability that under the null hypothesis this sample value would be observed)
- We can apply a significance level of 0.05 or 0.1 to determine if it is significantly different from 0.

PROPERTIES OF ESTIMATORS

- $\tilde{\theta}$, estimator of, θ, is a random variable

- Important properties of estimators:
  - Unbiased Estimator:
    $$E\left(\tilde{\theta}\right) = \theta$$
  - Consistency:
    $$\lim_{n\to\infty} \Pr\left(\left|\tilde{\theta}_n - \theta\right| < \epsilon\right) = 1$$
    The estimator gets 'closer' to the true θ as the sample size increases (convergence in probability)
  - Efficiency – the estimator has the minimum variance of all estimators under consideration
  - Asymptotic distribution – Distribution of estimator in large samples
- Ideal estimator:
  - Has minimum variance of all possible estimators
  - Is unbiased
  - Has an asymptotic distribution for inference purposes

MAXIMUM LIKELIHOOD ESTIMATOR (MLE)

- Notation:
  - Likelihood for a single observation x_i
    $$f(\mu; x_i)$$
  - Likelihood for the sample
    $$L(\mu; x) = \prod_{i=1}^{n} f(\mu; x_i)$$
  - Log-likelihood for a single observation
    $$\ln f(\mu; x_i)$$
  - Log-likelihood for the sample
    $$l(\mu; x) = \sum_{i=1}^{n} \ln f(\mu; x_i)$$
- Properties of MLE:
  - $\widehat{\mu}_n$ is a consistent estimator of μ ($\widehat{\mu}_n$ converges in probability to μ as n approaches infinity)
  - $\widehat{\mu}_n$ is asymptotically unbiased
    $$\lim_{N\to\infty} E(\widehat{\mu}_n) = \mu$$
  - $\widehat{\mu}_n$ is approximately normal in large sample (asymptotically normal)
    $$\sqrt{n}(\widehat{\mu}_n - \mu)$$
    tends to a normal distribution as $n \to \infty$

- $\widehat{\mu}_n$ has asymptotic variance

$$\begin{aligned} Var\left(\widehat{\mu}_N\right) &= \frac{1}{N\ E\left[\left(\frac{\partial}{\partial\mu} l\left(\mu; x_i\right)\right)^2\right]} \\ &= \frac{1}{E\left[\left(\frac{\partial}{\partial\mu} l\left(\mu; x\right)\right)^2\right]} \\ &= \frac{1}{E\left[-\frac{\partial^2}{\partial\mu^2} l\left(\mu; x\right)\right]} \end{aligned}$$

- MLE vs Least Squares:
  - If errors are Normal than MLE is the same as Least Squares for linear regression
  - For censored data and non-normal error distributions, we need to use MLE

M1: Survival Models and the Life Table

- Lifetime: time to the occurrence of a certain event, e.g.
  - Human mortality (time to death)
  - Length of time that a surviving individual will hold an insurance policy
- Survival models are models that describe the probability distribution of such "lifetimes".

A SIMPLE MODEL OF SURVIVAL | CONTINUOUS

- Future lifetime T of a newborn:
  - T is a continuous random variable on the interval $[0, \omega]$, where ω is the limiting age (highest age recorded 122 years and 164 days - Jeanne Calment - see, e.g. this link).
  - The distribution function of T is $F(t) = \Pr[T \leq t]$.
  - The survival function of T is $S(t) = \Pr[T > t] = 1 - F(t)$.
  - The probability density function of T is $f(t) = \frac{dF(t)}{dt}$. ∎.
- Future lifetime T_X after age x:
  - This is for a life who has already survived to age x.
  - T_x is a continuous random variable taking values in $[0, \omega - x]$
  - The distribution function of T_x is $F_x(t) = \Pr[T_x \leq t]$.
  - The survival function of T_x is $S_x(t) = \Pr[T_x > t] = 1 - F_x(t)$.

The probability density function of T_x is $f_x(t) = \frac{dF_x(t)}{dt}$.

$$F_x(t) = \Pr[T_x \leq t] = \Pr[T \leq x + t | T > x] = \frac{S(x) - S(x+t)}{S(x)}$$

$$S_x(t) = \Pr[T_x > t] = \Pr[T > x + t | T > x] = \frac{S(x+t)}{S(x)}$$

ACTUARIAL NOTATION

- Probability of death:

$$ {}_tq_x = \Pr[T_x \leq t] = F_x(t) = \int_0^t f_x(s)\mathrm{d}s $$

- Probability of survival:

$$ {}_tp_x = \Pr[T_x > t] = 1 - {}_tq_x = S_x(t) $$

- Deferred probability:

$$ {}_{n|m}q_x = Pr[n < T_x < n + m] $$

- Customary notation:

$$ q_x = {}_1q_x,\ p_x = {}_1p_x,\ {}_{|m}q_x = {}_{1|m}q_x $$

- Note:

$$ {}_{s+t}p_x = ({}_tp_x)({}_sp_{x+t}) $$

$$ {}_tp_x = p_x p_{x+1} \cdots p_{x+t-1} $$

FORCE OF MORTALITY

The force of mortality at age x $(0 \leq x < \omega)$

$$ \mu_x = \lim_{h \to 0^+} \frac{1}{h} \times \Pr[T \leq x + h | T > x] = \frac{\frac{\mathrm{d}F(x)}{\mathrm{d}x}}{S(x)} $$

is an instantaneous measure of mortality at age x. It is known as hazard rate in statistics. Looking forward,

$$ \begin{aligned} \mu_{x+t} &= \lim_{h \to 0^+} \frac{1}{h} \times \Pr[T \leq x + t + h | T > x + t] \\ &= \lim_{h \to 0^+} \frac{1}{h} \times \Pr[T_x \leq t + h | T_x > t]. \end{aligned} $$

- Cumulative Hazard Function:

$$ H(x) = \int_0^x \mu_t \mathrm{d}t = -\log S(x), $$

so that

$$ S(x) = \exp\left\{ -\int_0^x \mu_t \mathrm{d}t \right\} $$

and

$$ {}_tp_x = \exp\left\{ -\int_x^{x+t} \mu_s \mathrm{d}s \right\}. $$

- Important result:

$$ f_x(t) = {}_tp_x \mu_{x+t} $$

- Proof:

$$
\begin{aligned}
f_x(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\Pr[T_x \le t] \\
&= \lim_{h\to 0^+} \frac{1}{h} \times (\Pr[T_x \le t+h] - \Pr[T_x \le t]) \\
&= \lim_{h\to 0^+} \frac{1}{h} \times (\Pr[T \le x+t+h | T > x] - \Pr[T \le x+t | T > x]) \\
&= \lim_{h\to 0^+} \frac{\frac{\Pr[T\le x+t+h]-\Pr[T\le x]}{S(x)} - \frac{\Pr[T\le x+t]-\Pr[T\le x]}{S(x)}}{h} \\
&= \lim_{h\to 0^+} \frac{\Pr[T \le x+t+h] - \Pr[T \le x+t]}{S(x)h} \\
&= \frac{S(x+t)}{S(x)} \times \lim_{h\to 0^+} \frac{1}{h} \frac{\Pr[T \le x+t+h] - \Pr[T \le x+t]}{S(x+t)} \\
&= S_x(t)\mu_{x+t} \\
&= {}_tp_x\mu_{x+t}
\end{aligned}
$$

EXPECTATION OF LIFE

- The complete expectation of life at age x is defined by:

$$
\begin{aligned}
\mathring{e}_x &= \mathrm{E}[T_x] \\
&= \int_0^{\omega-x} t f_x(t)\mathrm{d}t \\
&= \int_0^{\omega-x} t(-\frac{\partial}{\partial t}{}_tp_x)\mathrm{d}t \\
&= -[t\,{}_tp_x]_{t=0}^{\omega-x} + \int_0^{\omega-x} {}_tp_x\mathrm{d}t \\
&= \int_0^{\omega-x} {}_tp_x\mathrm{d}t
\end{aligned}
$$

- The curtate future lifetime of a life age x is defined by:

$$K_x = [T_x] \text{ (the integer part of } T_x)$$

  - K_x is a discrete random variable taking integer values $0, 1, \cdots, [\omega - x]$.
  - $\Pr[K_x = k] = \Pr[k \le T_x < k+1] = {}_kp_xq_{x+k}$

- The curtate expectation of life is defined by:

$$\begin{aligned} e_x &= \mathrm{E}[K_x] = \sum_{k=0}^{[\omega-x]} k\, {}_kp_x q_{x+k} \\ &= {}_1p_x q_{x+1} \\ &+ {}_2p_x q_{x+2} \quad + {}_2p_x q_{x+2} \\ &+ {}_3p_x q_{x+3} \quad + {}_3p_x q_{x+3} \quad + {}_3p_x q_{x+3} \\ &+ \cdots \\ &+ {}_{[\omega-x]}p_x q_{x+[\omega-x]} + {}_{[\omega-x]}p_x q_{x+[\omega-x]} + \cdots + {}_{[\omega-x]}p_x q_{x+[\omega-x]} \\ &= \sum_{k=1}^{[\omega-x]} \sum_{j=k}^{[\omega-x]} {}_jp_x q_{x+j} = \sum_{k=1}^{[\omega-x]} \sum_{j=k}^{[\omega-x]} {}_{j|}q_x = \sum_{k=1}^{[\omega-x]} {}_kp_x, \end{aligned}$$

THE LIFE TABLE

- Radix of the table l_0: a hypothetical cohort of new-born lives
- $l_x = l_0 \times S(x)$: the expected number of survivors to age x out of the original group.
- Notation:
  - ${}_np_x = \frac{l_{x+n}}{l_x}$ conditional probability of surviving to age $x+n$, given alive at age x
  - $d_x = l_x - l_{x+1}$ the expected number who die between ages x and $x+1$
  - ${}_nd_x = l_x - l_{x+n}$ the expected number who die between ages x and $x+n$
  - ${}_nq_x = \frac{{}_nd_x}{l_x}$: conditional probability of dying within n years, given alive at age x
  - $\mu_x = \frac{-\frac{\mathrm{d}}{\mathrm{d}x}S(x)}{S(x)} = \frac{-\frac{\mathrm{d}}{\mathrm{d}x}l_x}{l_x}$
- Deferred quantities:

  Generally

$$\begin{aligned} {}_{n|m}q_x &= Pr[n < T_x < n+m] \\ &= \frac{l_{x+n} - l_{x+n+m}}{l_x} \\ &= \frac{l_{x+n}}{l_x}\frac{l_{x+n} - l_{x+n+m}}{l_{x+n}} \\ &= {}_np_x \times {}_mq_{x+n} \end{aligned}$$

  so that

$${}_{n|}q_x = {}_{n|1}q_x = {}_np_x \times q_{x+n}$$

  and

$${}_{n|}q_x = Pr[K_x = n] = Pr[n \le T_x < n+1] = \frac{d_{x+n}}{l_x}$$

INITIAL AND CENTRAL RATES OF MORTALITY

- Initial rate of mortality:
  - Probability that a life alive at age *x* (the initial time) dies before age *x+1*
  - Denoted $q_x = \frac{d_x}{l_x}$
- Central rate of mortality:
  - Probability that the population dies before age *x+1*
  - Demoted by m_x
  - Useful in population projections

$$\begin{aligned} m_x &= \frac{d_x}{\int_x^{x+1} l_z \mathrm{d}z} \\ &= \frac{q_x}{\int_0^1 {}_t p_x \mathrm{d}t}, \\ &= \frac{Pr[T_x \leq 1]}{\int_0^1 {}_t p_x \mathrm{d}t} \\ &= \frac{\int_0^1 {}_t p_x \mu_{x+t} \mathrm{d}t}{\int_0^1 {}_t p_x \mathrm{d}t} \end{aligned}$$

ASSUMPTIONS FOR FRACTIONS OF A YEAR

- All these functions ${}_s p_x$ can be determined from a lifetable for integral *x* and *s*, but not for non-integral *s*.
- The determination of functions for non-integral ages requires that values of l_{x+s} be available for all s, $0 \leq s \leq 1$, and integral *x*.
  - To this end, we will assume that l_{x+s} ($0 \leq s \leq 1$, x is any integer) has one of the following mathematical form:
    - Linear form for l_{x+s} (Uniform distribution of deaths)
    - Exponential form for l_{x+s} (Constant force of mortality)
    - Hyperbolic form for l_{x+s} (Balducci assumption/distribution)
  - We also assume that l_{x+s} is differentiable on the open interval $0 < s < 1$, which allows us to evaluate μ_{x+s} and hence the conditional density function $f(s|X > x) = {}_s p_x \mu_{x+s}$.
- Uniform distribution of deaths (UDD): Linear form

  We have

$$\begin{aligned} l_{x+s} &= l_x - s \cdot d_x = s \cdot l_{x+1} + (1-s) \cdot l_x \\ &\quad \text{(linear interpolation)} \\ {}_s p_x &= \frac{l_{x+s}}{l_x} = \frac{l_x - s \cdot d_x}{l_x} = 1 - s \cdot q_x \\ f(s|X > x) &= q_x \\ {}_s q_x &= 1 - {}_s p_x = s \cdot q_x \\ \mathring{e}_x &= e_x + \frac{1}{2} \end{aligned}$$

$${}_t p_{x+s} = \frac{l_{x+s+t}}{l_{x+s}} = \frac{l_x - (s+t) \cdot d_x}{l_x - s \cdot d_x} \text{ for } 0 \leq s, t \leq 1 \text{ and } s + t \leq 1$$

- Constant force of mortality: Exponential form

  For any integer x and $0 \leq s \leq 1$, assume that

  - l_{x+s} is a continuous exponential function with respect to s on $s \in [0,1]$, i.e. of form $a \cdot b^s$
  - in other words we assume that the log of l_x is linear:

$$\ln l_{x+s} = \ln a + \ln b \cdot s.$$

  We have then

$$\begin{aligned} \log l_{x+s} &= (1-s)\log l_x + s \log l_{x+1} \\ \mu_{x+s} &= \mu_x \text{ for } 0 < s < 1 \\ {}_sp_x &= e^{-s\mu_x} \\ {}_{t-s}p_{x+s} &= e^{-(t-s)\mu_x} \end{aligned}$$

- Balducci Assumption: Hyperbolic form

  For any integer x and $s \in [0,1]$, assume that

  - l_{x+s} is a hyperbolic function with respect to s on $s \in [0,1]$, i.e. of form $(a+bs)^{-1}$
  - alternatively, $\frac{1}{l_{x+s}} = (1-s) \cdot \frac{1}{l_x} + s \cdot \frac{1}{l_{x+1}}$ for $0 \leq s \leq 1$ (harmonic interpolation: linear interpolation on the reciprocal of the function)

  Then for any integer x and $s \in [0,1]$,

  - a convenient relationship: ${}_{1-s}q_{x+s} = (1-s)q_x$, $0 \leq s \leq 1$

- Force of mortality under constant force of mortality:

$$\mu_{x+s} = -\ln p_x, \quad 0 < s < 1$$

- Force of mortality under UDD:

$$\mu_{x+s} = \frac{q_x}{1 - sq_x}, \quad 0 < s < 1$$

- Force of mortality under Balducci:

$$\mu_{x+s} = \frac{q_x}{1 - (1-s)q_x}, \quad 0 < s < 1$$

SELECT LIFE TABLES

- Assume now mortality also depends on the date of entry in a group
  - $[x]+t$ still means age is $(x+t)$, but
    - the age at date of joining population is x; and
    - the duration from the date of joining the population is t
  - $l_{[x]+t}$: expected number of lives alive at duration t having joined the population at age $[x]$ based on some assumed radix
  - s: the length of the select period
  - After the select period the lives experience ultimate mortality, i.e. $l_{[x]+t} = l_{x+t}$ for $t \geq s$.

Example A 2-year select-and-ultimate mortality table:

$[x]$	$l_{[x]}$	$l_{[x]+1}$	l_{x+2}	$x+2$
30	9907	9905	9901	32
31	9903	9901	9897	33
32	9899	9896	9893	34

M2: Non-Parametric Methods

INTRODUCTION

- Given observations (data), the aim is to estimate the distribution of *T*.
- A simple method to estimate S(t) would be to observe a (very) large number of newborns and take the survival function as the proportion alive at each age.
- This presents a number of problems:
    - The experiment would take an extremely long time to complete
    - Lives under observation may be lost to the investigation, for one reason or another, and to exclude these from the analysis might bias the result (censoring)
    - This would be useful if all cohorts have the same mortality (which is not the case)
- Non-parametric Approach – No prior assumptions about the shape or form of the distribution
- Parametric Approach – Assume that the distribution belongs to a certain family (e.g. normal or exponential) and use the data to estimate the appropriate parameters.

CENSORING & TRUNCATION

- Type I Right Censoring –
    - Event (e.g. failure such as death) is observed only if it occurs prior to some prescribed time C_R (right censoring time).
    - The lifetime T is only known if $T \leq C_R$; the observation will be C_R if $T > C_R$.
    - *Examples of right censoring* – Investigation ends before all the lives being observed have died. Life insurance policyholders surrender their policies.
- Type II Right Censoring –
    - Observations continue until a predetermined number (say *r*) of events (failures) have occurred.
    - Data then consists of *r* smallest lifetimes in a sample of *n* (order statistics)
- Left Censoring –
    - The even of interest (such as death) has already occurred before the observation starts. So we only know that the lifetime T is less than a left censoring time C_L.
- Interval Censoring –
    - The lifetime T is only known to occur within an interval (e.g. actuarial investigations where we only know the calendar year of death).
- Truncation – Occurs when only those individuals whose event time lies within a certain observation period (Y_L, Y_R) are observed. Otherwise no information is available at all.
- Often confused with censoring – in the presence of censoring at least partial information is available (we know the event has happened, but only have partial information about it)
- Examples:
    - Any insurance claim that is not communicated because the deductible was not reached is left truncated
    - Right truncation arises in estimating the distribution of stars from the Earth because stars that are too far away are not visible and are right truncated.

- Random censoring/truncation – the censoring & truncation points are also subject to randomness. E.g.
  - Other competing risk can remove the individual from the study (e.g. lapsing of policy in a mortality study of insured lives)
  - Lifetime is censored by another random event
- C_i, the time at which the *i*th observation is censored is a random variable.
- If C_i is random:
  - Censoring is non-informative if it gives no information about the lifetimes
  - For random censoring, independence of all T's and C's is sufficient for it to be non-informative.
  - *Example of informative censoring*: Withdrawal of life insurance policies because those in better health are more likely to withdraw.
  - *Example of non-informative censoring*: The end of the investigation period.
- Assuming that lifetimes and censoring times are independent, the likelihood for the observation t:
  - if it is an exact lifetime, $f(t)$
  - if it is a right-censored observation, $S(C_R)$
  - if it is a left-censored observation $1 - S(C_L)$
  - if it is an interval-censored observation, $[S(C_L) - S(C_R)]$
  - in presence of left truncation, $\frac{[\text{any of the above}]}{S(Y_L)}$
  - in presence of right truncation, $\frac{[\text{any of the above}]}{1-S(Y_R)}$
  - in presence of interval truncation, $\frac{[\text{any of the above}]}{[S(Y_L)-S(Y_R)]}$ 
- Notation:
  - Population of N lives
  - Observe m deaths; $N - m$ lives are right censored
  - Ordered times of death $t_1 < t_2 < \ldots < t_k$, $k \leq m$
  - d_j number of deaths occur at time t_j $(1 \leq j \leq k)$, (more than one death can occur at any time)
  - $d_1 + d_2 + \ldots + d_k = m$
  - c_j: the number of lives that are right censored at a time belonging to $[t_j, t_{j+1})$ $(0 \leq j \leq k)$, $t_0 = 0$ and $t_{k+1} = \infty$
  - $c_0 + c_1 + \ldots c_k = N - m$
  - The times at which observations are censored within the time interval $[t_j, t_{j+1})$ are $t_{j1}, t_{j2}, \ldots t_{jc_j}$ (need not be distinct)
  - Define n_j as the number of lives alive and at risk at time t_j^- (just before time t_j). Then $n_j = d_j + c_j + n_{j+1}$.
  - The largest observed study time $t_{max} = \max\{t_k, t_{kc_k}\}$.

DISCRETE HAZARD FUNCTION

Suppose $F(t)$ corresponds has positive probability masses at and only at the points $t_1 < t_2 < \ldots < t_k$

Define the discrete hazard function as

$$\lambda_j = \Pr[T = t_j | T \geq t_j], \qquad (1 \leq j \leq k).$$

Then

$$S(t) = 1 - F(t) = \prod_{j:t_j \leq t} (1 - \lambda_j)$$

ASSUMPTIONS FOR MODELS

- *Non-informative censoring* – Time to censoring is independent of time of death
- *Lives are independent* – Time to censoring or time to death are determined independently for each life
- Shape of the distribution:

  For each interval $(t_{j-1}, t_j]$, the likelihood of the data is

  $$\left[F(t_j) - F\left(t_j^-\right)\right]^{d_j} \prod_{l=1}^{c_j} [1 - F(t_{jl})]$$

  This is because
  - we have d_j deaths at time t_j for $j = 1, 2, \cdots, k$ and their likelihood is
    $$F(t_j) - F\left(t_j^-\right)$$
  - we have censored lives surviving to t_{jl} for $j = 0, 1, \cdots k$ and $l = 1, \cdots, c_j$ with probability
    $$1 - F(t_{jl})$$
  - Note: we can take the product due to the non-informative censoring assumption.
- To maximise the likelihood, note the following:
  - $F(t_j) > F\left(t_j^-\right)$ at each failure, otherwise the likelihood will be zero
  - $[1 - F(t_{jl})]$ will be maximised
    - if $F(t_{jl})$ is minimised
    - but $F(t_{jl})$ is non-decreasing
    - hence we assume $F(t_{jl})$ stays as low as possible over the interval, that is $F(t_{jl}) = F(t_{j-1})$ for all l

  Therefore, the maximum likelihood estimate of $F(t)$ is a càdlàg step function with jumps at the times of the observed failures (deaths).

KAPLAN-MEIER (PRODUCT LIMIT) ESTIMATOR

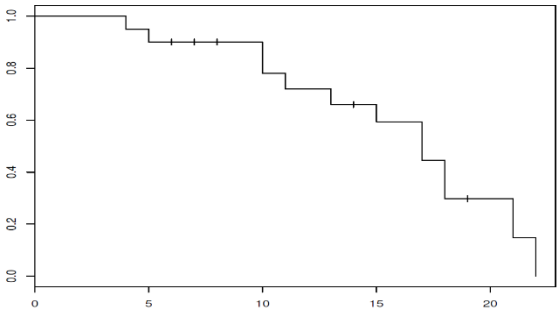
- The KM estimator is given by:

$$\widehat{F}(t) = 1 - \prod_{j:t_j \le t} \left(1 - \widehat{\lambda}_j\right)$$

- Or alternatively:

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right) & t_1 \le t \le t_{max}. \end{cases}$$

- Example of the plot of a KM estimate of a survival function:



- The KM estimator is well-defined for time points less than t_{max}:

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \widehat{S}(t_j) & \text{if } t_j \le t < t_{j+1}, j = 1, \cdots, k-1 \\ \widehat{S}(t_k) & \text{if } t_k \le t < t_{max}. \end{cases}$$

- For the estimator of the survival function beyond t_{max}:
  - If t_{max} corresponds to a death time and there is no censoring at t_{max}, the estimated survive curve is ZERO beyond t_{max}.
  - If $t_{max} = t_{kc_k}$, the value of $S(t)$ for $t > t_{max}$ is undetermined.
  - Two extreme views:
    - If assuming that the survivors at time t_{max} would have died immediately after t_{max}, $\hat{S}(t) = 0$ for $t > t_{max}$.
    - If assuming that the survivors at time t_{max} would die at ∞, $\hat{S}(t) = \hat{S}(t_{max}) = \hat{S}(t_k)$ for $t > t_{max}$.

- *Example:*

  Calculate the Kaplan-Meier estimate of $F(t)$.

j	t_j	d_j	n_j	$\hat{\lambda}_j = \frac{d_j}{n_j}$	$1 - \hat{\lambda}_j$	$1 - \prod_{k=1}^{j}(1 - \hat{\lambda}_k)$
1	3	1	10	0.1	0.9	0.1
2	4	1	9	0.11111	0.88889	0.2
3	11	2	7	0.28571	0.71429	0.42857

  From the final column, the Kaplan-Meier estimate of $F(t)$ is

$$\hat{F}(t) = \begin{cases} 0 & \text{for } 0 \le t < 3 \\ 0.1 & \text{for } 3 \le t < 4 \\ 0.2 & \text{for } 4 \le t < 11 \\ 0.42857 & \text{for } 11 \le t \le 20 \end{cases}$$

- Variance of the KM estimator – Let $\tilde{F}(t)$ denote the estimator. Greenwood's formula states:

$$Var\left(\tilde{F}(t)\right) = Var\left(\tilde{S}(t)\right) \approx \left[1 - \hat{F}(t)\right]^2 \sum_{j:t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

- Maximum likelihood estimators are asymptotically normally distributed, so we can easily construct confidence intervals:

$$\hat{S}(t) \pm Z_{1-\frac{\alpha}{2}} \sqrt{Var\left(\tilde{S}(t)\right)},$$

NELSON-AALEN ESTIMATOR

- The Nelson-Aalen estimator is based on the cumulative hazard:

$$\Lambda_t = \int_0^t \mu_s ds + \sum_{j:t_j \le t} m_j.$$

Since in our setting we do not have continuous increases in $\hat{F}(t)$ (only jumps – see previous sub-section) we focus on the second half only and use the ML estimator for λ_j to approximate the m_j's such that

$$\hat{\Lambda}_t = \sum_{j:t_j \le t} \frac{d_j}{n_j}$$

Finally,

$$\hat{S}(t) = e^{-\hat{\Lambda}_t}.$$

- Variance of the Nelson-Aalen estimator – Let $\tilde{\Lambda}(t)$ denote the estimator. Its variance is approximated as:

$$Var\left(\tilde{\Lambda}(t)\right) \approx \sum_{j:t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}$$

RELATIONSHIP BETWEEN BOTH ESTIMATORS

Denote:

- the Kaplan-Meier estimate of the survival function by $\hat{S}_{KM}(t)$
- the Nelson-Aslen estimate of the survival function by $\hat{S}_{NA}(t)$

Then

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

$$\approx \exp\left(-\sum_{j:t_j \le t} \frac{d_j}{n_j}\right) = \exp\left(-\hat{\Lambda}_t\right) = \hat{S}_{NA}(t)$$

- The KM estimator is more pessimistic.
- Furthermore,

$$\hat{S}_{KM}(t) < \hat{S}_{NA}(t) \quad \text{for} \quad t_1 \le t \le t_{max}.$$

COMPARING SURVIVAL FUNCTIONS | GENERAL APPROACH

- In many applications, one wants to compare two populations. E.g.
  - Smokers vs non-smokers
  - Effect of different treatments for a disease.
- As there is a 1-1 relationship between survival function (S) and hazard rates (h), we can test for differences between hazard rates.
- We will test the hypothesis:

$$\mathcal{H}_0 : h_1(t) = h_2(t) ; \forall t \leq \tau$$

vs

$$\mathcal{H}_1 : \text{At least one of the } h_1(t) \text{ differ from } h_2(t)$$

for some $t \leq \tau$.

To test for difference in hazard rates $h_1(t)$ and $h_2(t)$ of two different populations for all time $t \leq \tau$, the *general* form of the statistic is

$$Z_1 = \sum_{j=1}^{k} \widetilde{w}(t_j) \left[\frac{d_{1j}}{n_{1j}} - \frac{d_j}{n_j} \right],$$

where

- $\widetilde{w}(t_j)$ represent a positive weight function
- $t_1 < t_2 < \cdots < t_k$ are the distinct death times in the pooled sample
- d_{1j} is the number of deaths that occur in Group 1 at time t_j
- n_{1j} is number at risk prior to time t_j in Group 1
- n_j is the total number at risk prior to time t_j

Typically the weights are of the form

$$\widetilde{w}(t_j) = w(t_j)\, n_{1j}$$

such that the statistic becomes

$$Z_1 = \sum_{j=1}^{k} w(t_j) \left[d_{1j} - n_{1j} \left(\frac{d_j}{n_j} \right) \right] = \sum_{j=1}^{k} w(t_j) \left[d_{1j} - e_{1j} \right]$$

In this particular case, one can obtain

$$var(Z_1) = \sum_{j=1}^{k} (w(t_j))^2 \left(\frac{n_{1j}}{n_j} \right) \left(1 - \frac{n_{1j}}{n_j} \right) \left(\frac{n_j - d_j}{n_j - 1} \right) d_j$$

- An α level test:

  Under the null hypothsis, the statistic

$$\chi^2 = \frac{Z_1^2}{var(Z_1)}$$

  is a Chi-squared random variable (with 1 degree of freedom) for large samples

  This means we will reject the null hypothesis

$$\mathcal{H}_0 : h_1(t) = h_2(t) ; \forall t \leq \tau$$

  if $\frac{Z_1^2}{var(Z_1)}$ is larger than the αth upper percentage point of the Chi-squared distribution with 1 degree of freedom.

- Special cases are based on the choice of weight for the test statistic:

$$Z_1 = \sum_{j=1}^{k} w(t_j) \left[d_{1j} - n_{1j} \left(\frac{d_j}{n_j} \right) \right]$$

  We have

$$\begin{aligned} w(t_j) = 1 &\quad : \quad \text{Log-rank test} \\ w(t_j) = \widehat{S}_{KM}(t_j) &\quad : \quad \text{Peto-Peto Prentice test} \\ w(t_j) = n_j &\quad : \quad \text{Wilcoxon (Breslow-Gehan) test} \end{aligned}$$

- Some other weight function may be appropriate. The choice of weight function depends on the investigators desire to give different weights to different types of errors.
  - For instance, when comparing log-rank vs Wilcoxon, the latter gives more weight to early times.
- A practical note: log-rank statistic is more powerful for detecting differences in the hazard rates when the hazard rates are proportional ($h_1(t) = rh_2(t)$), i.e.

$$S_1(t) = [S_2(t)]^r$$

  for some constant r.
- The tests can be generalised to involve more than 2 groups.

PARAMETRIC VS NON-PARAMETRIC MODELS

- **Parametric models** assume that the distribution belongs to a certain family
- Advantages –
  - Generally easy interpolation/extrapolation
  - Generally easy to draw statistical inference
  - Easy way of smoothing the data
- Disadvantages –
  - Parameter error
  - Model error
- **Nonparametric models** make no prior assumptions about the shape or form of the distribution.
- Advantage –

- Adherence to data
- Disadvantages –
  - May be hard to make inference
  - Need extra assumptions to interpolate/extrapolate the data.
- *Example of non-parametric* –
  - Estimating a probability density/mass function with a histogram
  - Modelling a survival function by using the KM method
- *Examples of semi-parametric* –
  - Modelling the mortality rate with a Cox regression model
- *Examples of parametric* –
  - Fitting data with a parametric distribution
  - Parametric regression models
- Survival analysis with parametric distributions:
  - Uniform distribution (de Moivre)

$$S(x;\omega) = \begin{cases} 1, & x \leq 0 \\ \frac{\omega - x}{\omega}, & 0 < x \leq \omega \\ 0, & x > \omega \end{cases}$$

$$\mu_x = \frac{1}{\omega - x}$$

  - Exponential distribution (constant mortality rate/memoryless)

$$S(x;\mu) = e^{-\mu x}, \ t > 0$$

$$\mu_x = \mu$$

  - Type I extreme least value:

$$S(x;a,R) = \exp\left(-\frac{R}{a}e^{ax}\right), \ -\infty < x < \infty, \ a > 0, \ R > 0$$

$$\mu_x = Re^{ax}$$

  - Gompertz distribution:
    - a truncated (from 0) modification of the type I extreme least value distribution

$$S(x;a,R) = \exp\left(\frac{R}{a}(1 - e^{ax})\right), \ x > 0$$

$$\mu_x = Re^{ax}$$

  - Makeham-Gompertz distribution:
    - allowing for accidental deaths in addition to deaths from natural causes

$$\mu_x = A + Re^{ax}$$

- Lee-Carter model:

$$\mu_x(t) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}$$

where

- α_x represents the age-pattern
- β_x represents the deviations from the averaged pattern
- κ_t represents the change in level of mortality over time
- $\epsilon_{x,t}$ represents the error term of age-specific fluctuation

- Other examples include:
  - Modelling a lifetime distribution: Weibull, log-normal, gamma, logistic
  - Modelling stochastic transitions of state: Markov chain

M3: Semi-Parametric Models | Cox Regression Model

- We are trying to model different factors that are likely to impact the observed event separately, so that we can isolate their individual effects.
    - Heterogeneity – lives with different characteristics have different levels of mortality. I.e. different factors may have different effects on the risk
    - One method is to construct a model including the effects of the covariates on survival directly: a regression model
    - The most widely used regression model is the proportional hazards model (the Cox model)
- In many survival analysis problems, covariates are of the following types
    - Demographic/Societal (e.g. age, gender, education)
    - Behavioural (e.g. smoking, physical activity, alcohol)
    - Physiological (e.g. blood pressure, heart rate)
- Covariates can be:
    - Continuous measurements (weight)
    - Discrete measurements (age last birthday)
    - Indicators (1 for smoking, 0 for non-smoking)
    - Qualitative indicators (5 severe disease to 0 no symptoms)
- Notation:

  For the i^{th} life we will denote the **covariates** (**risk factors**) by a $1 \times p$ vector

  $$Z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$$

  Example:

  - Consider the covariate vector

    $$Z_i = (\text{sex, age, weight, symptoms}).$$

  - If the 3rd life is a 68 year old male, weighing 74kg, with mild symptoms of the condition under investigation (graded as 1 on a scale of 0 to 5), then we have

    $$Z_3 = (0, 68, 74, 1).$$

THE COX REGRESSION MODEL

- Main assumptions:

  Let the hazard function for the i^{th} life with covariates Z_i be

  $$\lambda(t; Z_i) = \lambda_o(t) \exp\left(\beta Z_i^T\right) = \lambda_o(t) \exp\left(\sum_{j=1}^{p} \beta_j z_{ij}\right)$$

  where:

  - $\lambda_o(t)$ is the baseline hazard
  - $\beta_1, \beta_2, \ldots \beta_p$ are the regression parameters
  - $z_{i1}, z_{i2}, \ldots z_{ip}$ are the covariates for the i^{th} subject

- Note:
  - In this formulation only $\lambda_o(t)$ depends on time but is independent of the covariates
  - conversely, $\exp\left(\sum_{j=1}^{p} \beta_j z_{ij}\right)$ is independent of t but dependent on the covariates
- Interpretation of β
  - If β_j is positive, the hazard rate increases with the *j*th covariate
  - If β_j is negative, the hazard rate decreases with the *j*th covariate
  - The sheer magnitude of β does not say much as this depends on how the covariates have been defined
  - As such we need a hypothesis test to check that $\beta \neq 0$ at a significant level
  - If β is estimated with standard techniques then it is easy to check their level of significance

PROPORTIONALITY OF HAZARD RATES

- The ratio of hazard rates of two different lives *x* and *y*:

$$\frac{\lambda(t; Z_x)}{\lambda(t; Z_y)} = \frac{\lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j z_{xj}\right)}{\lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j z_{yj}\right)} = \exp\left(\sum_{j=1}^{p} \beta_j \{z_{xj} - z_{yj}\}\right)$$

  is constant at all times, which explains the qualification of proportional hazards model.

  - This ratio is also called the relative risk of an individual with risk factor Z_x as compared to an individual with risk factor Z_y.

- Advantages of proportionality:
  - Under the Cox model, differences of hazard rates of different groups (different covariates) are acocunted for via the exponential term, which leads to a simple expression for the relative risk
  - The Cox model is not the only model with proportional hazards. One can generalise Cox to:

$$\lambda(t; Z_i) = \lambda_0(t) g(Z_i)$$

    where $g(Z)$ is any function of Z, but not t.
  - If we are only interested in the different due to covariates and not the baseline hazard, we can ignore $\lambda_o(t)$ and concentrate only on the function $g(Z)$.

PARTIAL LIKELIHOOD ESTIMATION

To estimate $\beta = (\beta_1, \beta_2, \ldots \beta_p)$, we will select the ones that maximize the partial likelihood under the following assumptions:

- noninformative censoring
- lives are independent

Notation:

- Ordered times of observed death are $t_1 < t_2 < \cdots < t_k$
- d_j: number of deaths occuring at time t_j $(1 \leq j \leq k)$
- We will label all the lives as $1, \cdots, n$.
- $R(t_j)$ is the set of the label of all lives who are still under study at a time just prior to t_j.
- Use (j) to denote the label of the life who dies at t_j 

- We distinguish two cases:
  1. assume that only one death occurs at each t_j $(d_j = 1 \text{ for } 1 \leq j \leq k)$
  2. Relax that assumption.
     In practice there might be ties in the data, that is
     1. some $d_j > 1$; or
     2. some observations are censored at an observed lifetime.

     This can significantly complicate the partial likelihood as one needs to include the lives censored at time t_j in the risk set $R(t_j)$ and all permutations of simultaneous events.

- Case 1 – one death at each observed lifetime:

  The partial likelihood for the death at t_j is

$$\Pr\left[\text{the individual } (j) \text{ dies at } t_j \;\middle|\; \begin{array}{l}\text{the first death of the lives}\\ \text{in } R(t_j) \text{ occurs at } t_j\\ \text{and only one death at } t_j\end{array}\right]$$

$$= \frac{\lambda\left(t_j; Z_{(j)}\right)}{\sum_{i \epsilon R(t_j)} \lambda\left(t_j; Z_i\right)} = \frac{\lambda_0\left(t_j\right) \exp\left(\beta Z_{(j)}^T\right)}{\sum_{i \epsilon R(t_j)} \lambda_0\left(t_j\right) \exp\left(\beta Z_i^T\right)} = \frac{\exp\left(\beta Z_{(j)}^T\right)}{\sum_{i \epsilon R(t_j)} \exp\left(\beta Z_i^T\right)}$$

  - Multiplying the likelihood over all deaths gives the partial likelihood, whose maximum is our estimate for β:

$$\hat{\beta} = \left\{ \beta \;\middle|\; L\left(\beta\right) = \prod_{j=1}^{k} \frac{\exp\left(\beta Z_{(j)}^T\right)}{\sum_{i \epsilon R(t_j)} \exp\left(\beta Z_i^T\right)} \text{ is maximal} \right\}$$

  - This is a partial likelihood since is considers only likelihood of the deaths (censored observations contribute to the denominator) and does not depend on the times of death (just the order)

- The numerator depends on information for the individual who experiences death. The denominator utlises information about all lives who have not yet experienced death.
- $\hat{\beta}$ are asymptotically normal

- Case 2 – partial likelihood in presence of ties in the data
  - We will consider Breslow's approximation:

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp\left(\beta s_j^T\right)}{\left[\sum_{i \epsilon R(t_j)} \exp\left(\beta Z_i^T\right)\right]^{d_j}}$$

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

  - This approximation works well when the number of ties are relatively small

- Properties of the maximum partial likelihood estimator:

The efficient score function is defined by

$$u(\beta) = \left(\frac{\partial \ln L(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ln L(\beta)}{\partial \beta_p}\right)$$

$$\text{Solving } u\left(\hat{\beta}\right) = 0 \text{ gives maximum likelihood estimate } \hat{\beta}$$

The maximum partial likelihood estimator $\tilde{\beta}$ (of β) is

- asymptotically unbiased
- asymptotically (multivariate) normally distributed with mean β and variance (matrix) equal to $(I(\hat{\beta}))^{-1}$, where $I(\hat{\beta})$ is the observed information matrix given by

$$I(\hat{\beta}) = \left(-\frac{\partial^2 \ln L(\beta)}{\partial \beta_i \partial \beta_j}|_{\beta=\hat{\beta}}\right)_{i,j=1,\cdots,p}$$

HYPOTHESIS TESTS ON B'S

- Assume you want to test:

$$H_0: \quad \beta_1 = \beta_{1,0}$$

for some subset $\beta_1 \in \beta$. Often we will test $\beta_1 = 0$, that is, whether the associated covariates have a significant effect. This is also used for model building:

  - Start with the null model which includes no covariates and add possible covariates one at a time (forward selection), or
  - Start with the full model which includes all covariates, and eliminate those of no significant effect (backward selection)

- Assume you want to test whether it is worth adding a set of q β's to an existing set of p β's. To test the effect of extra covariates, the null hypothesis is:

$$H_0: \quad \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

- i.e. the associated covariates are not significant

- There are two tests, and both rely on properties of the MpLE:
  - The likelihood ratio test
  - The Wald test
- The likelihood ratio test:

  The likelihood ratio statistic is

  $$-2\left[\log L_p - \log L_{p+q}\right],$$

  where two (nested) models have been fitted one with p covariates $(z_1, z_2, \cdots, z_p)$ and another with $p+q$ covariates $(z_1, z_2, \cdots, z_p, z_{p+1}, \cdots, z_{p+q})$, and their associated likelihoods are L_p and L_{p+q}, respectively.

  - This test statistic has a χ^2 distribution with q degrees of freedom *under the null hypothesis* for large n.
  - Reject the null hypothesis H_0 at $\alpha\%$ (eg 5%) significance level if the value of the test statistic is greater than the upper $\alpha\%$ point of χ^2_q.

- The Wald test:

  The Wald statistic is

  $$(b_{p+1}, \cdots, b_{p+q})\left[Cov(\tilde{\beta}_{p+1}, \cdots, \tilde{\beta}_{p+q})\right]^{-1}(b_{p+1}, \cdots, b_{p+q})^T$$

  where $b = (b_1, \cdots, b_{p+q})$ denotes the partial maximum likelihood estimates of $\beta = (\beta_1, \cdots, \beta_{p+q})$.

  - The test statistic of the Wald test has a χ^2 distribution with q degrees of freedom under the null hypothesis for large n.
  - In general, the Likelihood ratio test and the Wald test give very similar conclusions in practice.
  - (Note: Here, the vectors are row vectors. In statistics literature, column vectors seem more popular.)

  - Covariance term:

    Recall the observed information matrix of the model with $p+q$ covariates is

    $$I(b) = \left(-\frac{\partial^2 \ln L_{p+q}(\beta)}{\partial\beta_i \partial\beta_j}|_{\beta=b}\right)_{i,j=1,\cdots,p+q}$$

    Now partition $(I(b))^{-1}$ into

    $$(I(b))^{-1} = \begin{pmatrix} I^{(11)}(b) & I^{(12)}(b) \\ I^{(21)}(b) & I^{(22)}(b) \end{pmatrix}$$

    where $I_{11}(b)$ is of dimension $p \times p$, and $I^{(22)}(b)$ is of dimension $q \times q$. We have then

    $$Cov(\tilde{\beta}_{p+1}, \cdots, \tilde{\beta}_{p+q}) = I^{(22)}(b)$$

ESTIMATION OF THE SURVIVIAL FUNCTION

To estimate the survival function $S(t; Z)$ with covariate vector Z based on a Cox model

$$\lambda(t; Z) = \lambda_o(t) \exp\{\beta Z^T\}$$

we need

- not only to fit a proportional hazards model (Cox model) to the data, and obtain the partial maximum likelihood estimates of the parameters β,
- but also to estimate the baseline cumulative hazard rate $\Lambda_o(t) = \int_0^t \lambda_o(s)ds$.

Note that

$$S(t) = e^{-\int_0^t \lambda_0(s) e^{\beta Z^T} ds} = S_0(t)^{e^{\beta Z^T}},$$

where $S_0(t)$ is the baseline survival function.

Breslow's estimator of the baseline cumulative hazard rate $\Lambda_0(t)$ is

$$\hat{\Lambda}_0(t) = \sum_{t_j \le t} \frac{d_j}{\sum_{i \in R(t_j)} \exp\{b Z_i^T\}},$$

where b is the partial maximum likelihood estimate of β.

Finally, since

$$\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\}$$

we have

$$\hat{S}(t) = \hat{S}_0(t)^{\exp(bZ^T)}.$$

COX-SNELL RESIDUALS

- The Cox-Snell residuals are defined as:

$$e_j = -\log(\hat{S}(X_j; z_j)) = \hat{\Lambda}_0(X_j) \exp(\sum_{k=1}^{p} b_k z_{jk}^T)$$

  where
  - b is the MLE of β, and
  - $\hat{\Lambda}_0(t)$ is Breslow's estimator of the cumulative baseline hazard rate.
- If the model is correct and the *b's* (MLEs) are close to the true values of β, the Cox-Snell residuals e_j's behave as a censored sample from a unit exponential distribution.
- The check the goodness of fit of a Cox regression model, we need to check whether the Cox-Snell residuals:

$$e_j = -\log(\hat{S}(X_j; z_j)) = \hat{\Lambda}_0(X_j) \exp(\sum_{k=1}^{p} b_k z_{jk}^T)$$

  behave as samples from a unit exponential random variable.

- To check whether e_j behaves as a sample from a unit exponential:
  - compute the Nelson-Aalen estimator of the cumulative hazards rate of $e_j's$: $\hat{\Lambda}_E(e_j)$
  - A plot of $\hat{\Lambda}_E(e_j)$ versus e_j should be roughly straight line through the origin with a slope of 1

PROPORTIONALITY ASSUMPTION

- To check the proportional hazards assumption for a given covariate Z_1 after adjusting for all other relevant covariates:
  - Write the covariate vector (p-dimension) as $Z = (z_1, Z_2)$, where Z_2 represents the remaining $p-1$ covariates.
  - Assume no interaction between z_1 and any of the remaining covariates.
  - Assume that z_1 has K possible values.
  - Fit a Cox model stratified on each value of z_1, and let $\hat{H}_{g0}(t)$ be the estimated cumulative baseline hazard rate in the gth $(g = 1, 2, \cdots, K)$ stratum.
- So we have K models which should be proportional for the assumption to be valid with respect to covariate z_1.
- Graphical diagnostic tools:
  1. Plot $\ln[\hat{H}_{10}(t)]$, $\ln[\hat{H}_{20}(t)], \cdots$, and $\ln[\hat{H}_{K0}(t)]$ versus t.
     - the difference of any two should be of the form

$$\ln \frac{e^{\beta_1 z_{1,g_1}}}{e^{\beta_1 z_{1,g_2}}} = \beta_1(z_{1,g_1} - z_{1,g_2}),$$

       (where z_{1,g_1} and z_{1,g_2} are the respective possible outcomes for z_1), which does not depend on t
     - Hence, if the assumption holds, these curves should be approximately parallel.
  2. Alternatively, plot $\ln[\hat{H}_{20}(t)] - \ln[\hat{H}_{10}(t)], \cdots$, and $\ln[\hat{H}_{K0}(t)] - \ln[\hat{H}_{10}(t)]$ versus t.
     - This corresponds to plotting the expression above
     - If the assumption holds, each curve should be roughly constant.

OTHER APPROACHES TO ALLOW FOR COVARIATES

- Linear model representation in log time:

  Notation:

  - T denote the time to the event of interest (e.g. lifetime)
  - $Z = (z_1, \cdots, z_p)$ is a vector of explanatory covariates

  We assume a linear relationship between $\log(T)$ and covariates Z, such that

  $$\log T = \mu + \beta Z^T + \sigma\varepsilon,$$

  where $\beta = (\beta_1, \cdots, \beta_p)$ represent the regression coefficients, and ε is the error distribution (that is where the difference is between models).

- Estimation of linear log-time model:

  - If ε is normal (so T is lognormal) and there is no censoring, we can use Ordinary Least Squares (OLS);
  - ε is usually not normal;
  - Observations of T are usually censored;
  - Need to use MLE method.

- Accelerated failure time (AFT) model:

  Here we use the idea of modifying a baseline model (with survival function S_0). This model is then scaled up or down based on the covariates. The survival function is

  $$S(t; x) = S_0\left(t \exp\left[\theta x^\top\right]\right),$$

  where $\theta = (\theta_1, \cdots, \theta_p)$ is a vector of regression coefficients.
  Note:

  - a change in covariate values changes the time scale from the baseline time scale t
  - $\exp\left[\theta x^\top\right]$ is called an acceleration factor

M4: Poisson & Binomial Models

BINOMIAL MODEL | WITHOUT CENSORING

- Assume:
  - we observe N_x independent lives at *exactly* aged x at the beginning of the year, for one whole year
  - we observe d_x deaths
  - each life has a probability q_x of death of over that year (initial rate of mortality)
- Then the random variable D_x, number of deaths, is:

$$D_x \sim \text{Binomial}\left(N_x, q_x\right),$$

that is,

$$\Pr\left[D_x = d_x\right] = \binom{N_x}{d_x} q_x^{d_x} \left(1 - q_x\right)^{N_x - d_x} \quad d_x = 0, 1, 2, \ldots N$$

- Maximum likelihood estimation – under the binomial model, the log likelihood is:

$$\ln L\left(q_x\right) = \ln \binom{N_x}{d_x} + d_x \ln q_x + N_x - d_x \ln\left(1 - q_x\right)$$

so that

$$\frac{\partial}{\partial q_x} \ln L\left(q_x\right) = \frac{d_x}{q_x} - \frac{N_x - d_x}{1 - q_x}.$$

First order condition yields

$$\hat{q}_x = \frac{d_x}{N_x}$$

Noting

$$\frac{\partial^2}{\partial q_x^2} \ln L\left(q_x\right)\Big|_{q_x = \frac{d_x}{N_x}} = -\frac{d_x}{q_x^2} - \frac{\left(N_x - d_x\right)}{\left(1 - q_x\right)^2} < 0$$

confirms that $\hat{q}_x$ is MLE.

- Properties of the MLE:
  - unbiased $E\left[\hat{q}_x\right] = q_x$
  - $Var\left[\hat{q}_x\right] = \frac{q_x\left(1 - q_x\right)}{N_x}$ (minimum variance of all estimators - efficient)
  - asymptotically $\hat{q}_x \sim Normal\left(q_x, \frac{q_x(1-q_x)}{N_x}\right)$

BINOMIAL MODEL | WITH CENSORING

- Assume now that:
  - All lives are not necessarily observed over the complete year x to $x+1$, that is, there may be decrements other than death (right censoring); and also the possibility of increments (left truncation)
  - Hence, we observe a life from age $x+a_i$ to age $x+b_i$ $(0 \leq a_i < b_i \leq 1)$
  - For each life we know d_i, a_i and b_i
- Consider the random variable:

$$D_i = \begin{cases} 0 & \text{if the } i^{th} \text{ life survives from age } x+a_i \text{ to age } x+b_i, \text{ and} \\ 1 & \text{if the } i^{th} \text{ life dies} \end{cases}$$

so that

$$\begin{aligned} \Pr[D_i = 1] &= {}_{b_i-a_i}q_{x+a_i} \\ \Pr[D_i = 0] &= 1 - {}_{b_i-a_i}q_{x+a_i} \end{aligned}$$

- MLE – Assuming independence, the likelihood of the total sample will be:

$$L\left(\vec{q}; \vec{d}\right) = \prod_{i=1}^{N} \left({}_{b_i-a_i}q_{x+a_i}\right)^{d_i} \left(1 - {}_{b_i-a_i}q_{x+a_i}\right)^{1-d_i}$$

where $\vec{q} = \left({}_{b_1-a_1}q_{x+a_1}, {}_{b_2-a_2}q_{x+a_2}, \cdots, {}_{b_N-a_N}q_{x+a_N}\right)$
$\vec{d} = (d_1, d_2, \cdots, d_N)$.

- Note:
  - Unless we are using a (continuous) parametric model for the probabilities of death, there are potentially as many probabilities of death to estimate as data points
  - Another possibility is to reformulate all as a function of q_x using assumptions $(0 \leq t \leq 1)$:
    - uniform distribution of deaths: ${}_tq_x = t \cdot q_x$, or
    - Balducci assumption: ${}_{1-t}q_{x+t} = (1-t) \cdot q_x$, or
    - constant force of mortality: ${}_tq_x = 1 - e^{-\mu t}$
- The actuarial estimate –

  Let D be the number of deaths $\left(\sum_{i=1}^{N} D_i\right)$. We have then

$$E[D] = \sum_{i=1}^{N} \left({}_{b_i-a_i}q_{x+a_i}\right)$$

Now note that

$$\begin{aligned} {}_{b_i-a_i}q_{x+a_i} &= \Pr\left[\text{i}^{th} \text{ life dies between } x+a_i \text{ and } x+b_i\right] \\ &= \Pr\left[\text{i}^{th} \text{ life dies between } x+a_i \text{ and } x+1\right] \\ &\quad - \Pr\left[\text{i}^{th} \text{ life dies between } x+b_i \text{ and } x+1\right] \\ &= {}_{1-a_i}q_{x+a_i} - \left({}_{b_i-a_i}p_{x+a_i}\right)\left({}_{1-b_i}q_{x+b_i}\right) \end{aligned}$$

Furthermore, using Balducci assumption ${}_{1-a_i}q_{x+a_i} = (1-a_i)\,q_x$ and ${}_{1-b_i}q_{x+b_i} = (1-b_i)\,q_x$ we obtain

$$\begin{aligned} E[D] &= \sum_{i=1}^{N} \left({}_{b_i-a_i}q_{x+a_i}\right) \\ &= \sum_{i=1}^{N} \left[{}_{1-a_i}q_{x+a_i} - \left({}_{b_i-a_i}p_{x+a_i}\right)\left({}_{1-b_i}q_{x+b_i}\right)\right] \\ &= \sum_{i=1}^{N} \left[(1-a_i)\,q_x - \left({}_{b_i-a_i}p_{x+a_i}\right)(1-b_i)\,q_x\right] \\ &= \sum_{i=1}^{N} \left[(1-a_i)\,q_x - \left[1-E[D_i]\right](1-b_i)\,q_x\right] \end{aligned}$$

Now we substitute $E[D_i] = d_i$ and $E[D] = d$ (where $d = \sum_i^N d_i$)to get

$$d = \sum_{i=1}^{N} (1-a_i)\,q_x - \sum_{i;D_i=0} (1-b_i)\,q_x$$

which gives a moment estimate $\hat{q}_x$ of

$$\hat{q}_x = \frac{d}{\sum_{i=1}^{N}(1-a_i) - \sum_{i;D_i=0}(1-b_i)}$$

This is what is called the "actuarial estimate"

INITIAL EXPOSED TO RISK

- Recall the actuarial estimate:

$$\hat{q}_x = \frac{d}{\sum_{i=1}^{N}(1-a_i) - \sum_{i;D_i=0}(1-b_i)}$$

We can rewrite it as

$$\hat{q}_x = \frac{d_x}{E_x}$$

where the Initial exposed to risk E_x is defined as

$$E_x = \sum_{i=1}^{N}(1-a_i) - \sum_{i;D_i=0}(1-b_i) = \sum_{i:D_i=1}^{N}(1-a_i) + \sum_{i;D_i=0}(b_i-a_i)$$

Note:

- deaths contribute the period of length $(1-a_i)$ from $x+a_i$ to $x+1$
- survivors contribute the period of length (b_i-a_i) from $x+a_i$ to $x+b_i$

- Approximation to a "method of moments" estimator:
  - The actuarial estimate avoids numerical solution of equations
  - However, the actuarial estimate may not necessarily be simpler than estimates based on the multiple state model. When the exposure data are of census type, the computation of the initial exposed to risk is very complicated

CENTRAL EXPOSED TO RISK

- Sometimes we actually know when deaths occur, say $x + t_i$. We can then modify the initial exposed to risk to a central exposed to risk:

$$E_x^c = \sum_{i=1}^{N} (b_i - a_i)(1 - d_i) + \sum_{i=1} (t_i - a_i) d_i$$

where

- deaths contribute the period of length $(t_i - a_i)$ from $x + a_i$ to $x + t_i$
- survivors contribute the period of length $(b_i - a_i)$ from $x + a_i$ to $x + b_i$.

When the exact times of deaths are not available, but E_x^c is, the usual approch is to assume that deaths occur on average at age $x + \frac{1}{2}$ so that the actuarial estimate becomes

$$\hat{q}_x = \frac{d}{E_x^c + \frac{d}{2}}$$

because under that assumption

$$E_x = E_x^c + \frac{d}{2}.$$

POISSON MODEL

- Assume:
  - We observe N_x individuals over a year of age (starting from exactly age x)
  - We assume a constant force of mortality μ_x over the observed period for each individual in age interval $(x, x + 1)$
  - The sum of all of those observed periods is E_x^c, or observed waiting time v

This means that the time to death of each individual (within that year) is exponential, and hence (iff) that the number of deaths D_x is Poisson (if they can die several times):

$$\Pr[D_x = d_x] = \frac{e^{-\mu E_x^c} (\mu E_x^c)^{d_x}}{d_x!}, \quad d_x = 0, 1, 2, \ldots$$

with $E[D_x] = Var(D_x) = \mu E_x^c$.
Note that this implicitly allows for the possibility that there will be more than N_x deaths.

- MLE –

  Likelihood of observing d_x deaths is

$$L(\mu_x) = \frac{e^{-\mu_x E_x^c} (\mu_x E_x^c)^{d_x}}{d_x!}$$

  with Log-likelihood

$$\ln L(\mu_x) = -\mu_x E_x^c + d(\ln \mu_x + \ln E_x^c) - \ln(d_x!).$$

  Differentiate

$$\frac{\partial}{\partial \mu_x} \ln L(\mu_x) = \frac{d_x}{\mu_x} - E_x^c = 0$$

$$\frac{\partial^2}{\partial \mu_x^2} \ln L(\mu_x) = -\frac{d_x}{\mu_x^2} < 0$$

  leads to the estimator

$$\tilde{\mu}_x = \frac{D_x}{E_x^c}$$

- Properties of the estimator –

  Since D_x has a Poisson(μE_x^c) distribution,

$$E[D_x] = Var[D_x] = \mu_x E_x^c$$

  and we have

$$\begin{aligned} E[\tilde{\mu}_x] &= E\left[\frac{D_x}{E_x^c}\right] = \frac{\mu_x E_x^c}{E_x^c}, \text{ and} \\ Var[\tilde{\mu}_x] &= Var\left(\frac{D_x}{E_x^c}\right) = \frac{\mu_x E_x^c}{(E_x^c)^2} = \frac{\mu_x}{E_x^c}. \end{aligned}$$

  Furthermore, since $\tilde{\mu}_x$ is an MLE it is asymptotically normally distributed:

$$\tilde{\mu}_x \sim N\left(\mu_x, \frac{\mu_x}{E_x^c}\right)$$

M5: Markov Models

REVISION

- We use a Markov process with multiple states to model the random process by which a life transitions from one state to another. The states can be:
  - Alive, dead
  - Single married, divorced, widowed, dead
  - Weight
- The set that contains all possible states is called the state space
- Time index T and state space Ω can be wither continuous or discrete
- Markov Property:

  X_t is the r.v. that denotes the state at time t, and Ω denotes the state space. For $A \subset \Omega$ and $s < t$

  $$\Pr\left(X_t \in A \mid \{X_u\}_{u \leq s}\right) = \Pr\left(X_t \in A \mid X_s\right)$$

  The conditional probability does not depend on information prior to the latest available information.

- Transition probability:

  The transition probability of a continuous-time Markov Chain, $\{X_t\}_{t \geq 0}$ (defined on a countable state space Ω) is

  $$P_{i,j}(t, t+s) = \Pr(X_{t+s} = j \mid X_t = i),$$

  where $i,\ j \in \Omega$.

  Homogeneous Markov Chain: for any $t \geq 0$

  $$P_{i,j}(t, t+s) = P_{i,j}(0, s)$$

- Occupancy probability:

  In actuarial notation, we denote

  $${}_sp_t^{ii} = P_{i,i}(t, t+s) = \Pr(X_{t+s} = i \mid X_t = i).$$

  Occupancy probability is the probability that the individual occupies a state during a period of time. It is defined as:

  $${}_sp_t^{\overline{ii}} = \Pr\left(\{X_{t+u}\}_{0 \leq u \leq s} = i \mid X_t = i\right).$$

  In general, ${}_sp_t^{\overline{ii}} \neq {}_sp_t^{ii}$

INTRODUCTION

- Aims:
  - Present a stochastic process model
  - Estimate the transition intensities between states from data
  - Calculate transition probabilities of: Remaining in a given state for time t, Exiting a given state
  - End result – a stochastic model to be applied to survival (and other decrements), with methods to estimate the model via real world data.
- The two-state Markov model – Model the random process by which a life passes from one state (alive) to another (dead) by a Markov process.
  - States are alive a or dead d.
  - Dead state here is an absorbing state (transition rate out of this state is zero)
  - Alive state here is a transient state
  - Time spent in the alive state is the "future lifetime"
- Define:
  - A transition probability ${}_tq_x$

    ${}_tq_x = P[\text{in the dead state at age } x+t | \text{in the alive state at age } x]$

  - An occupancy or survival probability ${}_tp_x$

    ${}_tp_x = P[\text{in the alive state at age } x+t | \text{ in the alive state at age } x]$

- Assumptions underlying the mode:
  - The Markov assumption – The probabilities that a life at any given age will be found in either state at any future age depends only on the ages involved and on the state currently occupied.
  - For a short time interval dt:

    $$ {}_{dt}q_{x+t} = \mu_{x+t}dt + o(dt) \quad (t \geq 0) $$

    where a function $g(t)$ is of $o(t)$ if $\lim_{t\to 0} \frac{g(t)}{t} = 0$

- Probabilities:
  - Transition probabilities

$$
\begin{aligned}
P_{ad}(x+t, x+t+dt) &= {}_{dt}q_{x+t} = \mu_{x+t}dt + o(dt) \\
P_{aa}(x+t, x+t+dt) &= 1 - P_{ad}(x+t, x+t+dt) \\
&= 1 - \mu_{x+t}dt + o(dt) \\
P_{dd}(x+t, x+t+dt) &= 1 \\
P_{da}(x+t, x+t+dt) &= 0
\end{aligned}
$$

  - Using actuarial notation

$$
\begin{aligned}
{}_{t+dt}p_x &= {}_tp_x \Pr[\text{alive at } x+t+dt | \text{alive at } x+t] \\
&\quad + {}_tq_x \Pr[\text{alive at } x+t+dt | \text{dead at } x+t] \\
&= ({}_tp_x)_{dt}\, p_{x+t} + ({}_tq_x)\, 0 \\
&= {}_tp_x [1 - \mu_{x+t}dt + o(dt)]
\end{aligned}
$$

- Therefore:

$$ {}_tp_x = \exp\left\{-\int_0^t \mu_{x+s}ds\right\} $$

THE TWO-STATE MODEL

- The two state model consists of:
  - Alive state a and dead state d.
  - Age-dependent transition intensity $\mu_{x+t}\,(t \geq 0)$
- Kolmogorov Back Equations:

$$ \begin{bmatrix} \frac{\partial}{\partial s}P_{aa}(s,t) & \frac{\partial}{\partial s}P_{ad}(s,t) \\ \frac{\partial}{\partial s}P_{da}(s,t) & \frac{\partial}{\partial s}P_{dd}(s,t) \end{bmatrix} = \begin{bmatrix} \sigma_{aa}(s) = -\mu(s) & \sigma_{ad}(s) = \mu(s) \\ \sigma_{da}(s) = 0 & \sigma_{dd}(s) = 0 \end{bmatrix} \times \begin{bmatrix} P_{aa}(s,t) & P_{ad}(s,t) \\ P_{da}(s,t) & P_{dd}(s,t) \end{bmatrix} $$

Terminal conditions: at $s = t$

$$ \begin{bmatrix} P_{aa}(t,t) = 1 & P_{ad}(t,t) = 0 \\ P_{da}(t,t) = 0 & P_{dd}(t,t) = 1 \end{bmatrix} $$

- Kolmogorov Forward Equations:

$$ \begin{bmatrix} \frac{\partial}{\partial t}P_{aa}(s,t) & \frac{\partial}{\partial t}P_{ad}(s,t) \\ \frac{\partial}{\partial t}P_{da}(s,t) & \frac{\partial}{\partial t}P_{dd}(s,t) \end{bmatrix} = \begin{bmatrix} P_{aa}(s,t) & P_{ad}(s,t) \\ P_{da}(s,t) & P_{dd}(s,t) \end{bmatrix} \times \begin{bmatrix} -\mu(t) & \mu(t) \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} -P_{aa}(s,t)\,\mu(t) & P_{aa}(s,t)\,\mu(t) \\ 0 & 0 \end{bmatrix} $$

Initial conditions: at $t = s$

$$ \begin{bmatrix} P_{aa}(s,s) = 1 & P_{ad}(s,s) = 0 \\ P_{da}(s,s) = 0 & P_{dd}(s,s) = 1 \end{bmatrix} $$

- Definitions:
  - Assume N identical and independent lives between ages x and $x+1$
  - Observation begins at age $x + a_i$ for the i^{th} life
  - Observation ends at age $x + b_i$ for the i^{th} life
  - The death variable is represented by

$$ D_i = \begin{cases} 1 & \text{if the } i^{th} \text{ life dies} \\ 0 & \text{if the } i^{th} \text{ life does not die} \end{cases} $$

  - Random variable T_i is such that $x + T_i$ is the age at which observation of the i^{th} life ceases
    Note: D_i and T_i are not independent:
    $D_i = 0 \Leftrightarrow T_i > b_i$ (no death observed),
    $D_i = 1 \Leftrightarrow a_i < T_i \leq b_i$ (death observed)
  - Waiting time (time spent under observation)
    $V_i = \min\{T_i - a_i,\ b_i - a_i\}$

- The joint distribution of (D_i, V_i):
  - Let (d_i, v_i) denote an observed sample from the distribution of (D_i, V_i)
  - Let $f_i(d_i, v_i)$ denote the joint distribution of (D_i, V_i).
  - d_i can take two possible values,
    1. $d_i = 0$ ($\Leftrightarrow v_i = b_i - a_i$), i.e. alive over $(x + a_i, x + b_i)$

$$\begin{aligned} f_i(d_i = 0, v_i) &= {}_{b_i - a_i}p_{x+a_i} = \exp\left\{-\int_0^{b_i - a_i} \mu_{x+a_i+t}\, dt\right\} \\ &= \exp\left\{-\int_0^{v_i} \mu_{x+a_i+t}\, dt\right\} \end{aligned}$$

    2. $d_i = 1$ ($\Leftrightarrow v_i = t_i - a_i$), i.e. dies at $x + a_i + v_i$

$$\begin{aligned} f_i(d_i = 1, v_i) &= {}_{v_i}p_{x+a_i}\mu_{x+a_i+v_i} \\ &= \exp\left\{-\int_0^{v_i} \mu_{x+a_i+t}\, dt\right\} \mu_{x+a_i+v_i} \end{aligned}$$

  Taking into account the two possible scenarios, we have

$$f_i(d_i, v_i) = \exp\left\{-\int_0^{v_i} \mu_{x+a_i+t}\, dt\right\} [\mu_{x+a_i+v_i}]^{d_i}$$

- Assuming a constant force of mortality:

  Now let us consider a simplified case where μ_{x+t} is a constant μ for $0 \leq t < 1$ (piecewise constant force of mortality assumption). Then

$$f_i(d_i, v_i) = e^{-\mu v_i}\mu^{d_i}$$

  (a mixed distribution with probability mass at $v_i = b_i - a_i$)

  - If $d_i = 0$ then $v_i = b_i - a_i$ since the life survives to $x + b_i$
  - If $d_i = 1$ then $v_i = t_i - a_i$ since the life dies at age $x + t_i$.

  Note that

$$e^{-\mu(b_i - a_i)} + \int_0^{(b_i - a_i)} e^{-\mu v_i}\mu\, dv_i = 1$$

  since a life either survives to $x + b_i$ or dies between $x + a_i$ and $x + b_i$.

- Maximum likelihood estimation:

  We have the joint probability of all (D_i, V_i) $(i = 1, \cdots, N)$

  $$\prod_{i=1}^{i=N} e^{-\mu v_i} \mu^{d_i} = e^{-\mu\left(\sum_{i=1}^{i=N} v_i\right)} \mu^{\left(\sum_{i=1}^{i=N} d_i\right)} = e^{-\mu v} \mu^{d}$$

  where we denote

  - $d = \left(\sum_{i=1}^{i=N} d_i\right)$ is the total number of deaths
  - $v = \left(\sum_{i=1}^{i=N} v_i\right)$ is the total waiting time or "central exposed-to-risk"

  The likelihood can be represented by

  $$L(\mu; d, v) = e^{-\mu v} \mu^{d}.$$

  The log likelihood is then

  $$l(\mu; d, v) = -\mu v + d \ln \mu.$$

  The first order derivative of log likelihood function is

  $$\frac{\partial}{\partial \mu} l(\mu; d, v) = -v + \frac{d}{\mu}$$

  Solving $\frac{\partial}{\partial \mu} l(\mu; d, v) = -v + \frac{d}{\mu} = 0$ gives the MLE

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

  and estimator

  $$\tilde{\mu} = \frac{D}{V}$$

  where $D = \sum_{i=1}^{N} D_i$ and $V = \sum_{i=1}^{N} V_i$

  Asymptotic distribution of $\tilde{\mu}$

  $$\tilde{\mu} \sim N\left(\mu, \frac{\mu}{E[V]}\right) = N\left(\mu, \frac{\mu^2}{E[D]}\right)$$

GENERAL MARKOV MODEL

- Definitions:
  - $J = \{1, 2, \ldots, n\}$ finite set of states (state space)
  - $S(t)$ continuous time Markov process on states
  - g and h any two states with transition intensity μ_{x+t}^{gh} from state g to state h at age $x + t$
  - transition probabilities

$$
\begin{aligned}
{}_tp_x^{gh} &= \Pr(\text{in state } h \text{ at age } x+t | \text{in state } g \text{ at age } x) \\
{}_tp_x^{\overline{gg}} &= \Pr(\text{in state } g \text{ from age } x \text{ to age } x+t | \text{in state } g \text{ at age } x
\end{aligned}
$$

  - note that in general ${}_tp_x^{\overline{gg}} \neq {}_tp_x^{gg}$
- Assumptions:
  - Markov assumption: the probabilities that the process at any given time will be found in each state at any future time depend only on the times involved and on the state currently occupied
  - For any 2 distinct states, i.e. any $g \neq h$

$$
{}_{dt}p_{x+t}^{gh} = \mu_{x+t}^{gh} dt + o(dt) \quad t \geq 0
$$

  - Probability of more than one transition in time dt is $o(dt)$
- Kolmogorov Forward Equation:

$$
\frac{\partial}{\partial t} {}_tp_x^{gh} = \sum_{j \neq h} \left[{}_tp_x^{gj} \, \mu_{x+t}^{jh} - {}_tp_x^{gh} \mu_{x+t}^{hj} \right]
$$

  - Given the transition intensities (estimated from data) we can determine transition probabilities based on these equations.
  - Need to solve ordinary differential equations.
- Solving an ordinary differential equation:

  Steps to solve the following differential equation for $P(t)$:

$$
P(t)' + a(t)P(t) = b(t) \qquad (*)
$$

  - Find the integrating factor

$$
M(t) = e^{\int_0^t a(s)ds}.
$$

  - Multiply both sides of Equation (*) by $M(t)$

$$
\begin{aligned}
& P(t)'M(t) + a(t)P(t)M(t) = b(t)M(t) \\
\Longrightarrow \quad & P(t)'M(t) + P(t)M(t)' = b(t)M(t) \\
\Longrightarrow \quad & (P(t)M(t))' = b(t)M(t)
\end{aligned}
$$

- Solve $(P(t)M(t))' = b(t)M(t)$:

$$P(t)M(t) = \int_0^t b(s)M(s)ds + C$$
$$\implies \quad P(t) = \left(\int_0^t b(s)M(s)ds + C\right)/M(t)$$

- Calculate C based on initial conditions

- Computing transition probabilities:
  - In general cases, numerical methods are required to compute the transition probabilities. Consider the forward equation in matrix notation:

$$\mathbf{P}'(t) = \mathbf{P}(t)\,\mathbf{R}$$

where $\mathbf{P}$ is the transition probability matrix and $\mathbf{R}$ is the intensity matrix.
The solution is

$$\mathbf{P}(t) = e^{\mathbf{R}t}$$

where

$$e^{\mathbf{R}t} = \sum_{n=0}^{\infty} \mathbf{R}^n \frac{t^n}{n!}$$

With approximations

$$e^{\mathbf{R}t} \simeq \lim_{n\to\infty} \left(\mathbf{I} + \mathbf{R}\frac{t}{n}\right)^n$$

- Holding Times and Jump Probabilities:

In a time-homogeneous Markov jump process, let μ_{ij} be the transition rates from state i to state j for $j \neq i$ and write $\mu_i = \sum_{j\neq i} \mu_{ij}$. define

- W_i: the duration until the process leaves state i given that the current state is i.

Then: W_i is an exponential random variable,

$$S_{W_i}(t) = e^{-\mu_i t}$$
$$F_{W_i}(t) = 1 - e^{-\mu_i t}$$
$$f_{W_i}(t) = \mu_i e^{-\mu_i t}$$
$$E(W_i) = \frac{1}{\mu_i}$$
$$Var(W_i) = \left(\frac{1}{\mu_i}\right)^2$$

The probabilities that the Makov process $\{X_t : t \geq 0\}$ goes into state j when it leaves state i:

$$\Pr(X_{W_i} = j | X_0 = i) = \frac{\mu_{ij}}{\mu_i},$$

where μ_i is defined by $\mu_i = \sum_{j\neq i} \mu_{ij}$.

- Maximum likelihood estimators of the intensities:

  In a time-homogeneous Markov jump process with state space $\{1, 2, \cdots, m\}$, we observe

  - Total number of samples: N
  - $T_i = \sum_{l=1}^{N} T_i^{(l)}$: total waiting time in state i, where $T_i^{(l)}$ is the waiting time in state i of lth life.
  - $N_{ij} = \sum_{i=1}^{N} N_{ij}^{(l)}$: the total number of transitions from state i to state j, where $N_{ij}^{(l)}$ is the number of transitions from state i to state j made by lth life.

  As usual, we use lower case symbols for the observed samples. For life l, the likelihood is

$$\prod_{i=1}^{m} \left(e^{-\mu_i t_i^{(l)}} \prod_{j \neq i} \mu_{ij}^{n_{ij}^{(l)}} \right)$$

  The total likelihood is (by multiplying over all l):

$$\begin{aligned} & \prod_{l=1}^{N} \left(\prod_{i=1}^{m} \left(e^{-\mu_i t_i^{(l)}} \prod_{j \neq i} \mu_{ij}^{n_{ij}^{(l)}} \right) \right) \\ = & \prod_{i=1}^{m} \left(e^{-\mu_i t_i} \prod_{j \neq i} \mu_{ij}^{n_{ij}} \right) \end{aligned}$$

  For life i, the likelihood is proportional to

$$e^{-(\mu+\sigma)v_i} e^{-(\nu+\rho)w_i} \mu^{d_i} \nu^{u_i} \sigma^{s_i} \rho^{r_i}$$

  The total likelihood becomes

$$\begin{aligned} L(\sigma, \rho, \mu, \nu) &= \prod_{i=1}^{N} e^{-(\mu+\sigma)v_i} e^{-(\nu+\rho)w_i} \mu^{d_i} \nu^{u_i} \sigma^{s_i} \rho^{r_i} \\ &= e^{-(\mu+\sigma)v} e^{-(\nu+\rho)w} \mu^{d} \nu^{u} \sigma^{s} \rho^{r}. \end{aligned}$$

  The log-likelihood is therefore

$$\log L = -(\mu+\sigma)v - (\nu+\rho)w + d \log \mu + u \log \mu + s \log \sigma + r \log \rho.$$

We have the log-likelihood function of the able-illness-death model

$$\log L = -(\mu+\sigma)v - (\nu+\rho)w + d\log\mu + u\log\mu + s\log\sigma + r\log\rho.$$

Partial differentiation:

$$\frac{\partial \log L}{\partial \mu} = -v + \frac{d}{\mu} \quad \frac{\partial \log L}{\partial \nu} = -w + \frac{u}{\nu}$$
$$\frac{\partial \log L}{\partial \sigma} = -v + \frac{s}{\sigma} \quad \frac{\partial \log L}{\partial \rho} = -w + \frac{r}{\rho}$$

Setting these derivatives equal to 0 and solving the equations:

$$\mu = \frac{d}{v} \quad \nu = \frac{u}{w} \quad \sigma = \frac{s}{v} \quad \rho = \frac{r}{w}$$

Checking the second derivatives (negative), we have maximum likelihood estimates:

$$\hat{\mu} = \frac{d}{v} \quad \hat{\nu} = \frac{u}{w} \quad \hat{\sigma} = \frac{s}{v} \quad \hat{\rho} = \frac{r}{w}$$

Therefore, the maximum likelihood estimators are:

$$\tilde{\mu} = \frac{D}{V} \quad \tilde{\nu} = \frac{U}{W} \quad \tilde{\sigma} = \frac{S}{V} \quad \tilde{\rho} = \frac{R}{W}$$

The maximum likelihood estimates of the transition rates μ_{ij} $(i \neq j)$ from state i to state j is

$$\hat{\mu}_{ij} = \frac{n_{ij}}{t_i} \quad i \neq j$$

Here, n_{ij} is the no. of transitions from state i to j
t_i is the total waiting time in state i. As $\mu_{ii} = -\sum_{j\neq i}\mu_{ij}$, the MLE of μ_{ii} is

$$\hat{\mu}_{ii} = -\sum_{j\neq i}\hat{\mu}_{ij}$$

- Statistical properties of estimators:

  For the Able-illness-death model

  - $D_i - \mu V_i, U_i - \nu W_i, S_i - \sigma V_i, R_i - \rho W_i$ are uncorrelated (but not necessarily independent)
  - $(\tilde{\mu}, \tilde{\sigma}, \tilde{\nu}, \tilde{\rho})$ has an asymptotic multivariate Normal distribution
  - $(\tilde{\mu}, \tilde{\sigma}, \tilde{\nu}, \tilde{\rho})$ are asymptotically independent
  - Similar calculations as the simple 2 state model provides asymptotic distribution

$$\tilde{\mu} \sim Normal\left(\mu, \frac{\mu}{E[V]}\right) \quad \tilde{\sigma} \sim Normal\left(\sigma, \frac{\sigma}{E[V]}\right)$$

$$\tilde{\nu} \sim Normal\left(\nu, \frac{\nu}{E[W]}\right) \quad \tilde{\rho} \sim Normal\left(\rho, \frac{\rho}{E[W]}\right)$$

For a general Markov model, MLE $\tilde{\mu}_{ij}$ of the transition rates μ_{ij} $(j \neq i)$:

$$\tilde{\mu}_{ij} \sim Normal\left(\mu_{ij}, \frac{\mu_{ij}}{E[T_i]}\right),$$

where T_i is the total waiting time in state i.

- Estimation of transition intensities:
  - We can determine the time in each state for each individual over a study period, the number of jumps out of each state and the total waiting time in each state.
  - Hence we can estimate the transition intensities and the survival function for individuals in each state.
- *Data issue* – This calculation requires the computation of total waiting time, which may not be possible in practice.
- We can instead use the Census approach:
  - Census data is available where the number of subjects in each state is recorded only at fixed dates.
  - We can estimate waiting times by making simplifying assumption (e.g. assuming transitions occur on average half way through the time interval).

COMPARISON WITH BINOMIAL & POISSON MODELS

- Link between a Poisson model and 2-state Markov model:
  - In the Markov model you can specify that a transition from death to life is impossible, whereas with a straight Poisson RV you assume that individuals can die several times.
  - The numerical estimates of the parameter and the moments of the estimator are the same for both models
- Comparison: Underlying process
  - Binomial models allow for death or survival, but they represent only the year of death, not time of death
  - Markov models allow for the time of death (preferred if complete life histories are available)
  - Poisson model is an approximation to the Markov model where E_x^c is considered to be fixed over the year
  - When sufficient data is available (i.e. full information about the total waiting time), the two-state Markov model is preferred because using the binomial model will not make the greatest use of the information.
- Comparison: Estimating parameters
  - If exact dates of birth, entry, exit and death are known, then the MLE of the Markov model is easily calculated and the binomial model is complicated and needs further simplifying assumptions.
  - When μ is very small, the actuarial estimate provides acceptable results and the differences between the two-state Markov models and Poisson models are tiny
  - In practice, there is no difference between the two state and Poisson models, because the MLE's are the same.
- Statistical properties:

- Markov model – MLE is consistent, asymptotically unbiased and normally distributed. The variance is only available asymptotically.
- Poisson model – MLE is consistent and unbiased. The mean and variance are available exactly in terms of the true μ but are estimated from the data by the same expressions that estimate the asymptotic mean and variance in the two-state Markov model.
- Naïve Binomial model – MLE is consistent and unbiased, and the exact mean and variance can be expressed in terms of the true q_x. However, in practice, the data can rarely conform to the naïve model so only approximate results are available.
- When μ is small, none of these models is better than the other on the basis of the statistical properties of the MLEs alone.