Statistical Lifetime Models

Basics of MLE: Theta has asymptotic normal distribution, Information is important

$$X^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{m-k-1} \qquad \text{approximately, under } H_0,$$

1.4

cdf	$F(t) = \mathbb{P}\{L \le t\};$
survival function	$S(t) = \bar{F}(t) = 1 - F(t) = \mathbb{P}\{L > t\};$
density function	$f(t) = dF/dt;$
hazard rate	$\lambda(t) = f(t)/\bar{F}(t).$

Hazard / Force of Mortality

Exponential distribution is defined by constant hazard

$$\ell(\theta) = \sum_{i=1}^{n} \log \lambda_\theta(T_i) e^{-\Lambda_\theta(T_i)} = \sum_{i=1}^{n} \log \lambda_\theta(T_i) - \sum_{i=1}^{n} \Lambda_\theta(T_i), \tag{1.6}$$

where Λ_θ is the cumulative hazard function (depending on θ).

Likelihood for censored observations: Censored data only contributes negatively, only probability of exceeding censoring time

$$\ell(\theta) = \sum_{i=1}^{n} \log \lambda_\theta(T_i) - \sum_{i=1}^{n} \Lambda_\theta(T_i) - \sum_{i=1}^{m} \Lambda_\theta(C_i). \tag{1.7}$$

In the special case of a constant hazard μ this reduces to

$$\ell(\mu) = n \log \mu - \mu \Big(\sum_{i=1}^{n} T_i + \sum_{i=1}^{m} C_i \Big)$$

1.5 Distribution of Residual Lifetime (life remaining given aged x now)

$$\bar{F}_{T_x}(t) = \bar{F}_{T-x|T>x}(t) = \frac{\bar{F}_T(x+t)}{\bar{F}_T(x)}, \qquad f_{T_x}(t) = f_{T-x|T>x}(t) = \frac{f_T(x+t)}{\bar{F}_T(x)}, \qquad t \ge 0. \tag{1.8}$$

1.6 Force of Mortality/ Hazard Rate

$$h_T(t) = \mu_t = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \mathbb{P}(T \le t + \varepsilon | T > t) = \lim_{\varepsilon \downarrow 0} \frac{\frac{1}{\varepsilon}\mathbb{P}(t < T \le t + \varepsilon)}{\mathbb{P}(T > t)} = \frac{f_T(t)}{\bar{F}_T(t)}. \tag{1.9}$$

$$\bar{F}'_T(t) = -\mu_t \bar{F}_T(t), \quad \bar{F}(0) = 1 \qquad \Rightarrow \qquad \bar{F}_T(t) = \exp\left\{ -\int_0^t \mu_s ds \right\}, \qquad t \ge 0. \tag{1.10}$$

$$\bar{F}_{T_x}(t) = \frac{\bar{F}_T(x+t)}{\bar{F}_T(x)} = \exp\left\{ -\int_x^{x+t} \mu_s ds \right\} = \exp\left\{ -\int_0^t \mu_{x+r} dr \right\}, \qquad t \ge 0. \tag{1.11}$$

1.8 Curtate Lifespan: The floor of the age. If someone is 18 years and 6 months, we say they're 18

Chapter 2: Lifetime Distributions and Life Tables

Notation

2.3 Notation for life tables

q_x	Probability that individual aged x dies before reaching age $x+1$
p_x	Probability that individual aged x survives to age $x+1$
${}_tq_x$	Probability that individual aged x dies before reaching age $x+t$
${}_tp_x$	Probability that individual aged x survives to age $x+t$
l_x	Number of people who survive to age x. Note: This is based on starting with a fixed number l_0 of lives, called the Radix; most commonly, for human populations, the radix is 100,000
d_x	Number of individuals who die aged x (from the standard population)
${}_tm_x$	Mortality rate between exact age x and exact age $x+t$
e_x	Remaining (curtate) life expectancy at age x

$$d_x = l_x - l_{x+1}; \quad l_{x+1} = l_x p_x = l_x(1-q_x); \quad {}_tp_x = \prod_{i=0}^{t-1} p_{x+i}$$ (discrete)

q_x may be thought of as discrete mortality rate

Continuous vs Discrete: do we assume underlying lifetimes are discrete (i.e. die at x or x+1 in a binomial fashion) or continuous (e.g. die between x and x+1 and our table is showing curtate ages/ approximations). - Continuous allows you to discretise later

2.5 Lifetime Table Models

Continuous to Discrete ${}_tq_x = 1 - e^{-\int_x^{x+t} \mu_s ds}$. , $K = \lfloor T \rfloor$. , (fractional part) $S = T - K = \{T\}$

$$\mathbb{P}(K = n) = \mathbb{P}(n \le T < n+1) = \int_n^{n+1} f_T(t)dt = \bar{F}_T(n) - \bar{F}_T(n+1)$$
$$= \exp\left\{-\int_0^n \mu_T(t)\,\mathrm{d}t\right\}\left(1 - \exp\left\{-\int_n^{n+1} \mu_T(t)\,\mathrm{d}t\right\}\right)$$

$$q_k = \mathbb{P}(K = k | K \ge k) = \frac{\mathbb{P}(K=k)}{\mathbb{P}(K \ge k)} = 1 - \exp\left\{-\int_k^{k+1} \mu_T(t)\,\mathrm{d}t\right\}$$

and $p_k = 1 - q_k$, $k \in \mathbb{N}$, we obtain the probability of success after n independent Bernoulli trials with varying success probabilities q_k:

$$\mathbb{P}(K = n) = p_0 \dots p_{n-1} q_n.$$

Note that q_k only depends on the hazard rate between ages k and $k+1$. As a consequence, for $K_x = \lfloor T_x \rfloor$

$$\mathbb{P}(K_x = n) = p_x \dots p_{x+n-1} q_{x+n}$$

Discrete to Continuous

${}_1q_k = 1 - e^{-\mu_k}; \quad \mu_k = -\log p_k.$ Assuming constant hazard in interval

S has the distribution of an exponential random variable conditioned on $S < 1$, so it has density

$$f_S(s) = \mu_k \frac{\mathrm{e}^{-\mu_k s}}{1 - \mathrm{e}^{-\mu_k}}.$$

2.6 Deterministic life-table parameters

- Discrete (or *direct*) method;
- Continuous method;
- Census method.

Mortality Probability: Direct method

numbers ℓ_x alive at age x — called the Initial Exposed to Risk — and use $\hat{q}_x^{(0)} = d_x/\ell_x$, or similar quantities as an estimate for the one-year death probability q_x.

Force of Mortality: Direct method

There is no direct way to estimate the force of mortality — the mortality rate in the continuous model — from discrete demographic data.On the other hand, if we start with n individuals at exact integer age x, and observe the exact ages T1,...,Tk of deaths over the following year, it is natural, drawing on what we know of estimation for an exponential distribution to think of mx = k ` ˜ as an estimate of the constant (or average) force of mortality over that year, where

$\tilde{\ell} := (n-k) + \sum_{i=1}^{k}(T_i - x)$ is called the *total time exposed to risk.*

Mortality probability: the continuous method

$D_x \sim B(n, q_x)$ giving a maximum likelihood estimator $\widehat{q}_x = D_x/n$.

Trouble is, this assumption doesn't let us ask questions about sections of intervals, e.g. # deaths in half a year (half of an event?).

Central/Initial Exposed to Risk: Initial says everyone alive at start lived at least the full year (e.g. died on x+1), Central assumes on average all that died in the interval died halfway

Thus, the actuarial estimator for q_x is

$$E_x^0 \approx E_x^c + \frac{1}{2}d_x. \qquad \widetilde{q}_x = \frac{d_x}{E_x^c + \frac{1}{2}d_x}.$$

(E0 is initial, Ec central)]

2.7 Life Expectancy

$$\mathbb{E}[T] = \int_0^\infty t f_T(t)dt. \qquad \mathbb{E}[K] = \sum_{k=0}^{\infty} k\mathbb{P}\{K=k\} = \sum_{k=0}^{\infty}\mathbb{P}\{K>k\}.$$

Integration by parts, using the fact that $f_T = -\bar{F}_T'$, turns this into a much more useful form,

$$\mathbb{E}[T] = -t\bar{F}_T(t)\Big|_0^\infty + \int_0^\infty \bar{F}_T(t)\,\mathrm{d}t = \int_0^\infty \bar{F}_T(t)\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-\int_0^t \mu_s\,\mathrm{d}s}\,\mathrm{d}t. \tag{2.1}$$

Applying this to life tables, we see that the expected curtate lifetime is

$$\mathbb{E}[K] = \sum_{k=0}^{\infty}\mathbb{P}\{K>k\} = \sum_{k=1}^{\infty}\frac{l_k}{l_0} = \sum_{k=1}^{\infty} p_0\cdots p_{k-1}.$$

Note that expected future lifetimes can be expressed as

$$\mathring{e}_x := \mathbb{E}[T_x] = \int_x^\infty \exp\left\{-\int_x^t \mu_s\,\mathrm{d}s\right\}\mathrm{d}t \qquad \text{and} \qquad e_x := \mathbb{E}[K_x] = \sum_{k=1}^{\infty} p_x\ldots p_{x+k-1} = \sum_{k=1}^{\infty}\frac{l_{k+x}}{l_x}.$$

2.9 Comparing continuous and discrete methods

$$\mathbb{P}\{T < k+1 \,|\, T \geq k\} = q_k \approx \hat{q}_k = \frac{d_k}{E_k^0}.$$

The continuous model suggests that we estimate the same quantity by

$$\mathbb{P}\{T < k+1 \,|\, T \geq k\} = 1 - e^{-\mu_k} \approx 1 - e^{-\hat{\mu}_k} = 1 - e^{-d_k/E_k^c} \leq \frac{d_k}{E_k^c}. \tag{2.3}$$

The direct discrete method treats the curtate lifetimes as the true lifetimes; then E_k^0 is the same as E_k^c, so the continuous model gives a strictly smaller answer, unless $d_k = 0$. Why is that? The difference here is that the continuous model presumes that individuals are dying all through the year, making E_k^c somewhat smaller than E_k^0. In fact, the actuarial estimator $E_k^c \approx E_k - d_k/2$

NOTE: the above inequality is true only assuming the continuous model is correct

2.10 Statistical estimation of life-table parameters

model the deaths as Bernoulli events with probability q_x.

If we write $m(x) = (1-q_0)\dots(1-q_{x-1})q_x$, the likelihood is (for curtate lifetimes $k_1 \dots k_n$)

$$\prod_{i=1}^{n} m(k_i) = \prod_{x\in\mathbb{N}} (m(x))^{d_x} = \prod_{x\in\mathbb{N}} (1-q_x)^{\ell_x - d_x} q_x^{d_x} \quad \begin{matrix} d_x = d_x(k_1,\dots,k_n) = \#\{1 \leq i \leq n : k_i = x\} \\ \ell_x = \ell_x(k_1,\dots,k_n) = \#\{1 \leq i \leq n : k_i \geq x\} \end{matrix},$$

If we treat ℓ_x as a fixed quantity, the variance of $\hat{q}_x^{(0)}$ will be

$$\mathrm{Var}\left(\frac{d_x}{\ell_x}\right) = \frac{\mathrm{Var}(d_x)}{\ell_x^2} = \frac{\ell_x q_x (1-q_x)}{\ell_x^2}$$ So the estimate we use for the variance of $\hat{q}_x^{(0)}$ is

$$\mathrm{Var}(\hat{q}_x^{(0)}) \approx \frac{\hat{q}_x^{(0)}(1-\hat{q}_x^{(0)})}{\ell_x} = \frac{d_x(\ell_x - d_x)}{\ell_x^3}.$$

Continuous

Assume that you observe $n = \ell_0$ independent lives $t_1, \dots, t_n$. Then the likelihood function is

$$\prod_{i=1}^{n} f_T(t_i) = \prod_{i=1}^{n} \mu_{t_i} \exp\left\{-\int_0^{t_i} \mu_s ds\right\} \tag{2.6}$$

Now assume that the force of mortality μ_s is constant on $[x, x+1)$, $x \in \mathbb{N}$ and denote these values by

$$\mu_x = -\log(p_x) \qquad \left(\text{remember } p_x = \exp\left\{-\int_x^{x+1} \mu_s ds\right\}\right). \tag{2.7}$$

Then, the likelihood takes the form $\prod_{x=0}^{\infty} \mu_x^{d_x} \exp\left\{-\mu_x \tilde{\ell}_x\right\}$

where only $\max\{t_1,\dots,t_n\}+1$ factors in the infinite product differ from 1, and

$$d_x = d_x(t_1,\dots,t_n) = \#\left\{1 \leq i \leq n : [t_i] = x\right\},$$
$$\tilde{\ell}_x = \tilde{\ell}_x(t_1,\dots,t_n) = \sum_{i=1}^{n} \int_x^{x+1} 1_{\{t_i > s\}} ds.$$

$\tilde{\ell}_x$ is the *total time exposed to risk.*

$$\hat{\mu}_x = \hat{\mu}_x(t_1, \ldots, t_n) = \frac{d_x(t_1, \ldots, t_n)}{\tilde{\ell}_x(t_1, \ldots, t_n)},$$

$$\hat{q}_x = \hat{q}_x(t_1, \ldots, t_n) = 1 - \hat{p}_x = 1 - \exp\{-\hat{\mu}_x\}$$ A

$$\mathrm{Var}(\hat{\mu}_x) \approx \frac{\hat{\mu}_x}{\tilde{\ell}_x} = \frac{d_x}{\tilde{\ell}_x^2}.$$ (use Fischer Info)

2.11 Central exposed to risk and the census approximation

Because of the limitations of the data, we reinterpret d_x and E_x^c as the number of deaths **observed** and number of years of life **observed**, for which we need to create estimates that obey the **Principle of Correspondence**:

An individual alive at time t should be included in the exposure at age x at time t if and only if, were that individual to die immediately, he or she would be counted in the death data d_x at age x.

The key point is that we can tolerate a substantial amount of uncertainty in the numerator and the denominator (number of events and total time at risk), but failing to satisfy the Principle of Correspondence can lead to serious error

Census Approximation

$P_{x,k}$ = Number of individuals in the population aged $[x, x+1)$ at time $k = 0, \ldots, n$.

$$E_x^c = \int_0^n P_{x,t}\,\mathrm{d}t$$ (note density 1)

$$E_x^c \approx \sum_{k=1}^{n} \frac{1}{2}(P_{x,k-1} + P_{x,k}).$$ (interpolating each interval)

d'_x = Number of deaths aged x on the birthday in the calendar year of death.

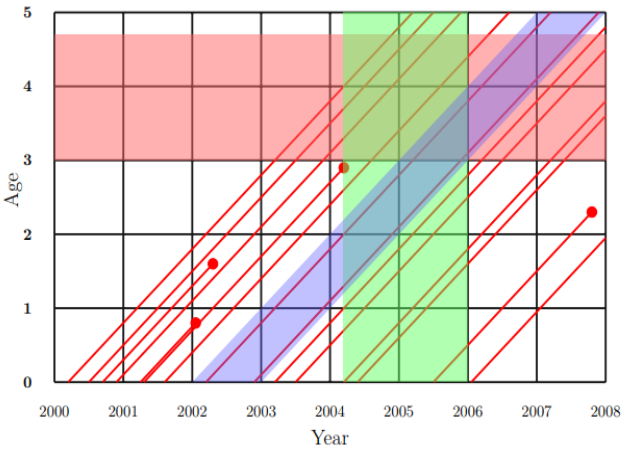
$P'_{x,t}$ = Number of individuals in the population at t with xth birthday in calendar year $\lfloor t \rfloor$.

The Principle of Correspondence requires

$$E_x^{c\prime} = \int_0^n P'_{x,t}\,\mathrm{d}t,$$

2.12 Lexis Diagrams

The red region represents the experience of all individuals (at whatever time) aged between 3 years and 4 years, 8 months.
The green region represents the experience of all individuals during the period from 15 March 2004 and 31 December 2005.
The blue region represents the experience of the cohort born in 2002, from birth to age 5.
Red diagonal lines are individuals, dying at the circles



If we assume that $P_{x,t}$ is approximately linear over such an interval, we may approximate th average over $[k, k+1]$ by $\frac{1}{2}(P_{x,k} + P_{x,k+1})$. Then we get the approximation

$$E_x^c \approx \frac{1}{2}P_{x,0} + \sum_{k=1}^{T-1} P_{x,k} + \frac{1}{2}P_{x,T}.$$

(discussion on assumptions here in notes), and on Cohort vs Period lifetime tables

Cohort vs Period Lifetime Tables

1. Cohort life table describing a real population. These make most sense in a biological context, where there is a small and short-lived population. The ℓ_x numbers are actual counts of individuals alive at each time, and the rest of the table is simply calculated from these, giving an alternative descriptions of survival and mortality.

2. Period life tables, which describe a notional cohort (usually starting with radix ℓ_0 being a nice round number) that passes through its lifetime with mortality rates given by the q_x. These q_x are estimated from data such as those of Table 2.2, giving the number of individuals alive in the age class during the period (or number of years lived in the age class) and the number of deaths.

3. Synthetic cohort life tables. These take the q_x numbers from a real cohort, but express them in terms of survival ℓ_x starting from a rounded radix. This is what is usually meant by a "cohort life table".

2.14 Graduation - Each interval has a small sample size so tends to have high variance. Want to smoothen our estimates

2.14.1 Parametric models

We may fit a formula to the data. Possible examples are $\mu_x = \mu$ (Exponential);

$\mu_x = \alpha\rho^\alpha x^{\alpha-1}$ (Weibull); $\mu_x = Be^{\theta x}$ (Gompertz); $\mu_x = A + Be^{\theta x}$ (Makeham).

2.14.2 Reference to a standard table

Here q_x^0, μ_x^0 represent the graduated estimates. We could have a linear dependence

$$q_x^0 = a + bq_x^s, \quad \mu_x^0 = a + b\mu_x^s$$

or possibly a translation of years

$$q_x^0 = q_{x+k}^s, \quad \mu_x^0 = \mu_{x+k}^s$$

2.14.3 Nonparametric smoothing

We effectively smooth our data when we impose the assumption that mortality rates are constant over a year. We may tune the strength of smoothing by requiring rates to be constant over longer intervals. This is a form of local averaging, and there are more and less sophisticated versions of

2.14.4 Methods of fitting

1. Apply the binomial model, and set $q_x = a + bq_x^s$ in the likelihood; find maximum likelihood estimators for the unknown parameters a, b. Similarly, do the same for μ_x in the Poisson model. Other parametric functions of the standard life table than linear may also be used.

2. Use weighted least squares and minimise

$$\sum_{\text{all ages } x} w_x \left(\widehat{q}_x - \overset{\circ}{q}_x\right)^2$$

$$\sum_{\text{all ages } x} w_x \left(\widehat{\mu}_x - \overset{\circ}{\mu}_x\right)^2$$

as appropriate. For the weights suitable choices are either E_x or E_x^c respectively. Alternatively we can use $1/\text{var}$, where the variance is estimated for $\widehat{q}_x$ or $\widehat{\mu}_x$, respectively.

Chapter 3: Multiple Decrements Model

Examples

- A working population insured for disability might transition into multiple different possible causes of disability, which may be associated with different costs.
- Workers may leave a company through retirement, resignation, or death.
- A model of unmarried cohabitations, which may end either by separation or marriage.

Competing Risks Assumption $${}_tq_x = 1 - \left(1 - {}_tq_x^{CAUSE1}\right)\left(1 - {}_tq_x^{CAUSE2}\right)\cdots$$

Consequently,

$$\lambda_x^{CAUSE1} = \text{fraction of deaths due to CAUSE 1} \times \lambda_x,$$

3.1.2 Multiple decrements – time-homogeneous rates

In the time-homogenous case, we can think of the multiple-decrement model as m exponential clocks C_j with parameters λ_j, $1 \le j \le m$, and when the first clock goes off, say, clock j, we observe that time as the lifetime $L = \min\{C_1, \ldots, C_m\}$, and we also record the state j

$$\prod_{i=1}^{n} \lambda_{j_i} e^{-t_i\lambda_+} = \prod_{j=1}^{m} \lambda_j^{n_j} e^{-\lambda_j(t_1+\ldots+t_n)}, \tag{3.2}$$

where n_j is the number of transitions to j. Again, this can be solved factor by factor to give

$$\hat{\lambda}_j = \frac{n_j}{t_1 + \ldots + t_n}, \qquad 1 \le j \le m. \tag{3.3}$$

$\hat{\lambda}_+ = n/(t_1 + \ldots + t_n)$, since $n_1 + \ldots + n_m = n$.

In the competing-clocks description, we can interpret the likelihood as consisting of m ingredients, namely the density $\lambda_j e^{-\lambda_j t}$ of clock j to go off at time t, and probabilities $e^{-\lambda_k t}$ of clocks C_k, $k \ne j$, to go off after time t.

3.2 Estimation for general multiple decrements

(L, J), is given by

$$\prod_{i=1}^{n} \lambda_{j_i}(t_i) \exp\left\{-\int_0^{t_i} \lambda_+(t)dt\right\}.$$

Let us assume that the forces of decrement $\lambda_j(t) = \lambda_j(x)$ are constant on $x \le t < x+1$, for all $x \in \mathbb{N}$ and $1 \le j \le m$. Then the likelihood can be given as

$$\prod_{x\in\mathbb{N}} \prod_{j=1}^{m} \left(\lambda_j(x)\right)^{d_{j,x}} \exp\left\{-\tilde{\ell}_x \lambda_j(x)\right\}, \tag{3.5}$$

where $d_{j,x}$ is the number of decrements to state j between ages x and $x+1$, and $\tilde{\ell}_x$ is the total time spent alive between ages x and $x+1$.

Now the parameters are $\lambda_j(x)$, $x \in \mathbb{N}$, $1 \le j \le m$, and they are again well separated to deduce

$$\hat{\lambda}_j(x) = \frac{d_{j,x}}{\tilde{\ell}_x}, \qquad 1 \le j \le m, \quad 0 \le x \le \max\{L_1, \ldots, L_n\}. \tag{3.6}$$

3.4 The distribution of the endpoint

The time-homogeneous multiple-decrement model makes a transition at the minimum of m exponential clocks as opposed to one clock in the single decrement model. In the same way, we can construct the time-inhomogeneous multiple-decrement model from m independent clocks C_j with hazard function $\lambda_j(t)$, $1 \le j \le m$. Then the likelihood for a transition at time t to state j is the product of $f_{C_j}(t)$ and $\bar{F}_{C_k}(t)$.

and we obtain

$$\mathbb{P}(L = C_j) = \int_0^\infty \mathbb{P}(L = C_j | L = t) f_L(t) dt = \int_0^\infty \lambda_j(t) \bar{F}_L(t) dt = \mathbb{E}(\Lambda_j(L)), \tag{3.8}$$

where $\Lambda_j(t) = \int_0^t \lambda_j(s) ds$ is the integrated hazard functior

The discrete (curtate) lifetime model: We can also split the curtate lifetime $K = [L]$ according to the type of decrement J ($J = j$ if $L = T_j$) and define

$$q_{j,x} = \mathbb{P}(L < x + 1, J = j | L > x), \qquad 1 \le j \le m, \quad x \in \mathbb{N}, \tag{3.9}$$

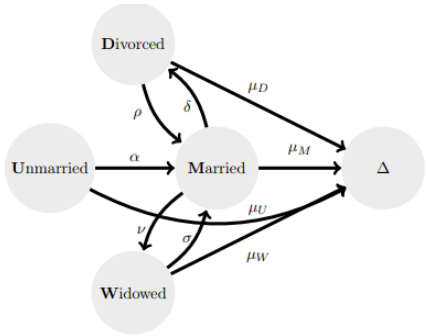
then clearly for $x \in \mathbb{N}$

$$q_{1,x} + \ldots + q_{m,x} = q_x \tag{3.10}$$

and, for $1 \le j \le m$,

$$p_{(J,K)}(j,x) = \mathbb{P}(J = j, K = x) = \mathbb{P}(L \le x + 1, J = j | L > x)\mathbb{P}(L > x) = p_0 \ldots p_{x-1} q_{j,x}. \tag{3.11}$$

Chapter 4: Multistate Models



The general rule for estimating transition rates is

$$\hat{q}_{xy} = \frac{\#\text{ transitions } x \to y}{E_x}, \qquad \text{Var}(\hat{q}_{xy}) \approx \frac{\hat{q}_{xy}}{E_x},$$

where E_x is the total time observed occupying state x.
We compute an approximate $100(1-\alpha)\%$ confidence interval as $\hat{q}_{xy} \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{q}_{xy})}$.

Parametric models - if we assumed a hazard rate between every state (say k states). Get k(k-1) parameters. That's wayy too many. To make it easier we:

1. Set some hazard rates to be strictly 0;
2. Require that some hazard rates be the same, or be linked by some functional identity.

Lower-dimensional parametric models

In general, we may have a model defined by parameters $\theta \in \mathbb{R}^d$. This means the hazard rates are written as some function $q_{xy}(\theta)$. Writing $\mathcal{S}$ for the set of sites, the likelihood is

$$L(\theta) = \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{S} \setminus \{x\}} q_{xy}(\theta)^{N_{xy}} \exp\left\{-E_x q_{xy}(\theta)\right\}.$$

Here N_{xy} is the observed number of transitions from x to y, and E_x is the total observed time occupying state x.

Typically no closed-form solution

Birth and Death Processes

One very common class of multistate models is one where the sites are ordered in a line — typically labelled by integers, or some subset of the integers — with transitions allowed only one step up or down in the line. This is called a *birth-and-death* process, inspired by the application

If $\mathcal{S}$ is the (possibly infinite) set of sites, the parameters for the model may be represented as $\lambda_x = q_{x,x+1}$ and $\mu_x = q_{x,x-1}$ for $x \in \mathcal{S}$. Of course, if $\mathcal{S}$ has a maximum site x, then $\lambda_x \equiv 0$, and similarly for μ_x if x is the minimum site. Note that this model could have infinitely many parameters (if $\mathcal{S}$ is unbounded), but, just as with unspecified maximal ages in the single-decrement model, we would only estimate λ_x and μ_x for sites x that have actually occurred in the data.

$$\prod_{i \in \mathbb{N}} \prod_{|j-i|=1} q_{ij}^{N_{ij}} \exp\left\{-E_i q_{ij}\right\} = \left(\prod_{i \in \mathbb{N}} (i\lambda)^{N_{i,i+1}} \exp\left\{-E_i i\lambda\right\}\right) \left(\prod_{i \in \mathbb{N}} (i\mu)^{N_{i,i-1}} \exp\left\{-E_i i\mu\right\}\right) \tag{4.1}$$

to separate the two parameters. This can best be maximised via the log likelihood, which for the μ-factor is

$$\sum_{i=1}^{\infty} \left(N_{i,i-1}(\log(i) + \log(\mu)) - E_i i\mu\right). \tag{4.2}$$

Differentiation leads to the maximum likelihood estimator $\hat{\mu} = D/W$ where $D = \sum_i N_{i,i-1}$ is the total number of deaths and $W = \sum_i iE_i$ is the weighted sum of exposure times at population

In the same way, W plays the same role of "total time" in computing asymptotic variance. Thus $\mathrm{Var}(\hat{\lambda}) \approx \hat{\lambda}/W$ and $\mathrm{Var}(\hat{\mu}) \approx \hat{\mu}/W$.

4.4 Time-varying hazard rates

We need some principle for lumping together events observed at different times, either a parametric form for the time variation of hazards or the assumption that the hazard rates are constant over some

$$\hat{q}_{xy}(t) = \frac{N_{xy}(t)}{E_x(t)}, \qquad \mathrm{Var}\left(\hat{q}_{xy}(t)\right) \approx \frac{\hat{q}_{xy}(t)}{E_x(t)}, \tag{4.3}$$

fixed time intervals.

where $N_{xy}(t)$ is the number of transitions $x \to y$ observed *during the time interval that includes* t, and $E_x(t)$ is the total amount of time *overlapping with the time interval that includes* t when the process was observed in state x.

Chapter 5: Survival Analysis

Censoring & Truncation

Left censoring is when the event of interest has already occurred before enrolment. This is less common.

Right truncation occurs when the entire study population has already experienced the event of interest (for example: a historical survey of patients on a cancer registry).

Left truncation occurs when the subjects have been at risk before entering the study (for example: life insurance policy holders where the study starts on a fixed date, event of interest is age at death).

Truncation is due to study design. It may not be obvious, since the truncated data are never observed.

Generally we deal with **right censoring** and **left truncation**.

Three types of independent **right censoring**:

Type I: All subjects start and end the study at the same fixed time. All are censored at the end of the study. If individuals have different but fixed censoring times, this is called *progressive type I* censoring.

Type II: study ends when a fixed number of events amongst the subjects has occurred.

Type III or random censoring: Individuals drop out or are lost to followup at times that are random, rather than predetermined. We generally assume that the censoring times are *non-informative*, meaning that censoring gives us no information about how long the individual would have lived had they not been been censored. That is, the distribution of the remaining lifetime of an individual still alive at time t is identical to the conditional distribution of the unobserved remaining lifetime, **conditioned on having been censored** at time t. Thus, an individual

Random Censoring

Suppose that $\tilde{T}$ is the time to event — this may sometimes be called the *latent event time* — and that C is the time to the censoring event. Assume that all subjects may have an event or be censored. That is, we get to observe for each individual i only $T_i := \min\{\tilde{T}_i, C_i\}$, and the *censoring indicator*

$$\delta_i := \begin{cases} 1 & \text{if } \tilde{T}_i \leq C_i, \\ 0 & \text{if } \tilde{T}_i > C_i. \end{cases}$$

It was shown by [32] that it is impossible to reconstruct the joint distribution of $(\tilde{T}, C)$ from these data. It is possible to reconstruct the marginal distributions under the assumption that event times $\tilde{T}$ and censoring times C are independent.

Define

$$\begin{aligned} &f(t) \text{ and } f_C(t) \text{ to be the densities,} \\ &S(t) \text{ and } S_C(t) \text{ to be the survival functions,} \\ &h(t) \text{ and } h_C(t) \text{ to be the hazard functions} \end{aligned}$$

of the event time and the censoring time respectively. If the event time and censoring time are independent then the likelihood is

$$\begin{aligned} L &= \prod_{\delta_i=1} f(\tilde{T}_i) S_C(\tilde{T}_i) \prod_{\delta_i=0} S(C_i) f_C(C_i) \\ &= \prod_i h(T_i)^{\delta_i} S(T_i) \times \prod_i h_C(T_i)^{1-\delta_i} S_C(T_i). \end{aligned}$$

Since this factors into an expression involving the distribution of $\tilde{T}$ and one involving the distribution of C, we may perform likelihood inference on the event-time distribution without

Random censoring is an unideal assumption, try instead **Non-Informative censoring**

We generally want to retain the assumption that the event times $\tilde{T}_i$ are jointly independent. The censoring process is called *non-informative* if it satisfies the following conditions:

1. For each fixed t, and each i, the distribution of $(T_i - t)\mathbf{1}_{\{T_i > t\}}$ is independent of $\{C_i > t\}$. That is, knowing that an individual was or was not censored by time t gives no information about the lifetime remaining after time t, if they were still alive at time t.

2. The event-time distribution and the censoring distribution do not depend on the same parameters.

Time on test: time someone is at risk

5.4 Non-parametric survival estimation

If there are observations $x_1, \ldots, x_n$ from a random sample then we define the empirical distribution function

$$\widehat{F}(x) = \frac{1}{n} \#\{x_i : x_i \leq x\}$$

Suppose that the observations are (T_i, δ_i) for $i = 1, 2 \ldots, n$. We consider the component of the likelihood that relates to the distribution of the survival times:

$$L = \prod_i f(T_i)^{\delta_i} S(T_i)^{1-\delta_i} \quad = \prod_i f(T_i)^{\delta_i} \left(1 - F(t_i)\right)^{1-\delta_i}$$

Kaplan-Meier Estimator

The basic idea is the following: There is no way to estimate a hazard rate from data without some kind of smoothing, so the most direct representation of the data comes from estimating the survival function directly. On any interval $[a, b)$ on which no events have been observed, it is natural to estimate $\hat{S}(b) - \hat{S}(a) = 0$. That is, the estimated probability of an event occurring in this interval is the empirical probability 0. This leads us to estimate the survival function by a step function whose jumps are at the points t_j where events have been observed.

We conventionally list the event times (not the censoring times) in order as $t_1 < \cdots < t_j < \cdots$. If there are ties — several individuals i with $T_i = t_j$ and $\delta_i = 1$ — we represent this by a count d_j, being the number of individuals whose event was at t_j. (Thus all d_j are at least 1.)

At the point t_i there is a drop in $\hat{S}$. (Like cdfs, survival functions are taken to be right-continuous: $S(t) = \mathbb{P}\{\tilde{T} > t\}$.) The size of the drop at an event time t_j is

$$S(t_j-) - S(t_j) = \mathbb{P}\{\tilde{T} = t_j\}$$

since $S(t_j-) = \mathbb{P}\{\tilde{T} \geq t_j\}$. We may rewrite this as

$$\begin{aligned} S(t_j) &= S(t_j-)\left(1 - \frac{\mathbb{P}\{\tilde{T} = t_j\}}{S(t_j-)}\right) \\ &= S(t_j-)\left(1 - \mathbb{P}\{\tilde{T} = t_j \,|\, \tilde{T} \geq t_j\}\right). \end{aligned}$$

So need to estimate discrete hazard h_j

The Kaplan-Meier estimator is the result of this process:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \hat{h}_j\right) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

$$\begin{aligned} n_j &= \#\{\text{in risk set at } t_j\}, \\ d_j &= \#\{\text{events at } t_j\}. \end{aligned}$$

$$n_{j+1} + c_j + d_j = n_j$$

Nelson Aalen Estimator

The Nelson–Aalen estimator for the cumulative hazard function is

$$\widehat{H}(t) = \sum_{t_j \le t} \frac{d_j}{n_j} \quad \left(= \sum_{t_j \le t} \widehat{h}_j \right)$$

$$\begin{aligned} \widetilde{S}(t) &= \exp\left(-\widehat{H}(t)\right) \\ &= \exp\left(-\sum_{t_i \le t} \frac{d_i}{n_i}\right) \end{aligned}$$

=>

It is not difficult to show by comparing the functions $1 - x, \exp(-x)$ on the interval $0 \le x \le 1$, that $\widetilde{S}(t) \ge \widehat{S}(t)$.

Left Truncation, example easily done

Patient ID	5	2	9	0	1	3	7	6	4	8
Event time	2	5	5	*	7	*	12	*	*	*
Censoring time	10	8	7	8	11	7	14	14	14	14
Truncation time	−2	3	6	0	1	0	6	6	−5	1
Observation	2	5	∘	8+	7	7+	12	14+	14+	14+

=>

t_j	d_j	n_j	$\hat{h}_j$	$\hat{S}(t_j)$	$\widetilde{S}(t_j)$
2	1	6	0.17	0.83	0.85
5	1	6	0.17	0.69	0.72
7	1	7	0.14	0.58	0.62
12	1	4	0.25	0.45	0.48

(Patient 9 had event time before trunc time, so was not observed)

Important note: As usual, we assume that individuals who have their event or are censored in a given year were at risk during that year. On the other hand, the default assumption is that an individual who entered the study at age x was **not** at risk in that year. (This is also the way the `survival` package in `R` interprets left-truncated data.) Why is this? Saying an

5.6 Variance estimation: Greenwood's formula

Kaplan–Meier: $\hat{S}(t) = \prod_{t_j \le t} \left(1 - \hat{h}_j\right)$, Nelson–Aalen: $\widetilde{S}(t) = \exp\left\{-\sum_{t_j \le t} \hat{h}_j\right\}$.

Suppose the survival function were genuinely discrete, with conditional probability h_j (the "discrete hazard") of an event occurring at t_j for any individual i with $T_i \ge t_j$. Then $\hat{h}_j = d_j/n_j$ is an unbiased estimator for h_j. The d_j are binomially distributed with parameters (n_j, h_j), hence variance $n_j h_j (1 - h_j)$. Furthermore, while the numbers of events d_j are not independent — the underlying parameter n_j, the number of individuals at risk, depends on the outcomes of all the survival events preceding t_j — the expected value of d_j/n_j is always h_j. If the n_j are large, then, we may write

$$\hat{h}_j = h_j + Z_j \sqrt{\frac{h_j(1 - h_j)}{n_j}},$$

(Z_j iid standard normal)

$$\widehat{H}(t) - H(t) = \sum_{t_j \le t} (\hat{h}_j - h_j) = \left(\sum_{t_j \le t} \frac{h_j(1-h_j)}{n_j} \right)^{1/2} Z, \quad \approx \left(\sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3} \right)^{1/2} Z.$$

=>

$$\tilde{\sigma}^2(t) := \sum_{t_j \le t} \frac{d_j(n_j - d_j)}{n_j^3}$$

=>

as an estimator for the variance [1] of $\widehat{H}(t)$. So an approximate $100(1-\alpha)\%$ confidence interval for $H(t)$ is

$$\widehat{H}(t) \pm z_{1-\alpha/2}\tilde{\sigma}(t). \tag{5.4}$$

We can apply (5.4) directly to obtain a confidence interval for $S(t) = e^{-H(t)}$

$$\begin{aligned} &\left(\exp\left\{ -\widehat{H}(t) - z_{1-\alpha/2}\tilde{\sigma}(t) \right\}, \exp\left\{ -\widehat{H}(t) + z_{1-\alpha/2}\tilde{\sigma}(t) \right\} \right) \\ &\qquad = \left(\widetilde{S}(t) \exp\left\{ -z_{1-\alpha/2}\tilde{\sigma}(t) \right\}, \widetilde{S}(t) \exp\left\{ z_{1-\alpha/2}\tilde{\sigma}(t) \right\} \right). \end{aligned} \tag{5.5}$$

BUT this isn't centred on KP estimator: Get one via Greenwoods Formula

$$\log \hat{S}(t) = \sum_{t_j \le t} \log(1 - \hat{h}_j). \qquad (1-h_j)^{-2} \operatorname{Var}(\hat{h}_j) = \frac{h_j}{n_j(1-h_j)} \approx \frac{d_j}{n_j(n_j - d_j)}.$$

and

$$\hat{\sigma}^2(t) := \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

for the variance of $\log \hat{S}(t)$.

=> (delta method)

Traditionally we apply the delta method again to produce the

$$\sigma_G^2(t) := \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

Called Greenwoods Formula

5.8 Survival to Infinity

$$S_0(t) := \mathbb{P}\{T > t \,|\, T < \infty\};$$

$$S_0(t) = \frac{\mathbb{P}\{\infty > T > t\}}{\mathbb{P}\{\infty > T\}}.$$

has an observed event. In either case, there is no mathematical principle for distinguishing between the actual survival to ∞ — that is, the probability that the event never occurs — and simply running out of data. Nonetheless, in many cases there can be good reasons for thinking that there is a time t_∂ such that the event will never happen if it hasn't happened by that time. In that case we may use the fact that $\{T < \infty\} = \{T < t_\partial\}$ to estimate

$$S_0(t) = \frac{S(t) - S(t_\partial)}{1 - S(t_\partial)}.$$

If there is no *a priori* reason to choose a value of t_∂, we may estimate it from the data as $\max\{t_i\}$, as long as there is a significant length of time during which there is a significant number of individuals under observation, when an event *could have been* observed.

In this case, assuming that $S(t)$ is constant after $t = t_\partial$, we need to estimate the variance of $\hat{S}_0(t)$. To compute a confidence interval for $S_0(t)$ we apply the delta method again, in a slightly more complicated form. Suppose we set $Y_1 = \hat{S}(t)$ and $Y_2 = \hat{S}(t_\partial)/\hat{S}(t)$. The variance of Y_1 is approximated just by Greenwood's formula

$$\sigma_1^2 := \operatorname{Var}\left(\widehat{S}(t)\right) \approx \hat{\sigma}^2(t) = \widehat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j\left(n_j - d_j\right)},$$

Y_2 is just the estimated survival from t to t_∂, so Greenwood's formula yields

$$\sigma_2^2 := \operatorname{Var}\left(\frac{\widehat{S}(t_\partial)}{\widehat{S}(t)}\right) \approx \left(\frac{\widehat{S}(t_\partial)}{\widehat{S}(t)}\right)^2 \sum_{t < t_j \le t_\partial} \frac{d_j}{n_j\left(n_j - d_j\right)},$$

Since Y_1 and Y_2 depend on distinct survival events they are uncorrelated. We may apply the delta method (5.9) to the two-variable function $g(y_1, y_2) = y_1(1 - y_2)/(1 - y_1 y_2)$, obtaining

$$\sigma_0^2(t) := \operatorname{Var}\, \widehat{S}_0(t) = \operatorname{Var}\left(g(Y_1, Y_2)\right)$$

$$\approx \frac{\left(\widehat{S}(t) - \widehat{S}(t_\partial)\right)^2}{\left(1 - \widehat{S}(t_\partial)\right)^4} \sum_{t_j \le t} \frac{d_j}{n_j\left(n_j - d_j\right)} \quad + \frac{\left(1 - \widehat{S}(t)\right)^2 \widehat{S}(t_\partial)^2}{\left(1 - \widehat{S}(t_\partial)\right)^4} \sum_{t < t_j \le t_\partial} \frac{d_j}{n_j\left(n_j - d_j\right)}.$$

5.9 Left censoring

In pure left censoring, each individual has a censoring time C_i and an event time T_i. We get to observe $U_i := T_i \vee C_i$ (that is, the maximum of the times), and the censoring indicator $\delta_i := \mathbf{1}_{\{T_i \ge C_i\}}$. We assume that all individuals have the same hazard rate $\alpha(t)$ at time t. In this setting it is more natural to think of estimating the survival function $S(t) := e^{-A(t)}$, which is assumed (as usual) to be continuous.

If we fix τ a time that is greater than any time U_i, then $U_i^* := \tau - U_i$ is equal to $(\tau - T_i) \wedge (\tau - C_i)$, and $\delta_i = \mathbf{1}_{\{\tau - T_i \le \tau - C_i\}}$. Thus (U_i^*, δ_i) is a collection of right-censored observations of a random variable whose survival function is $F(t) = 1 - S(t)$.

The procedure is then straightforward:

1. Choose $\tau > \max\{U_i\}$, and let $U_i^* := \tau - U_i$.
2. Let $\hat{S}_<(t)$ be the Kaplan–Meier estimator for the right-censored observations (U_i^*, δ_i).
3. The estimator for the survival function is then $\widehat{S}(t) = 1 - \hat{S}_<(\tau - t)$.

By the same reasoning, we may estimate the variance of $\log \widehat{S}(t)$ by

$$\sum_{t_j \ge t} \frac{d_j}{n_j(n_j - d_j)},$$

where $d_j = \left(\#\{i : U_i = t_j\right\}$ and $n_j = \left(\#\{i : U_i \ge t_j\right\}$. Note that we are accumulating variance

5.10 Right truncation

Survival times are right truncated when individuals are excluded from the study if their time exceeds a certain threshold. This depends on a particular study design. Not surprisingly, right truncation can also be dealt with by time reversal, turning right truncation into left truncation. The situation is that there are event times T_i and truncation times R_i (assumed independent), but we only observe the subset of (T_i, R_i) such that $R_i > T_i$. (That is, we observe both times or neither. This is in contrast to censoring, where we always observe exactly one of the times.)

Chapter 6: Comparing Survival Distributions

6.1 Tests in a Parametric Setting

chi-squared test good to test # parameters (e.g. weibull vs exp), also use to see if two samples come from same set (i.e same parameters or different)

6.2 One-sample testing: Life Tables Each individual is exponential (cst hazard lambda) in year x => #deaths is poisson(lambda) over [0,E_x^c]

Suppose we treat E_x^c as though it were a constant. Then if D_x represents the numbers dying in the year the model uses

$$\mathrm{P}\{D_x = k\} = \frac{(\mu_x E_x^c)^k \, \mathrm{e}^{-\mu_x E_x^c}}{k!}, \quad k = 0, 1, 2, \cdots .$$

$$\widetilde{\mu}_x = \frac{D_x}{E_x^c}, \quad \text{with observed value } \frac{d_x}{E_x^c} \qquad \mathrm{var}\widetilde{\mu}_x = \frac{\mu_x E_x^c}{(E_x^c)^2} = \frac{\mu_x}{E_x^c} \; . \; \approx \frac{d_x}{(E_x^c)^2}$$

=>. (poisson)

6.2.2 Testing hypotheses for qx and μx

Binomial model: $D_x \sim B(E_x, q_x) \implies D_x \sim \mathcal{N}\left(E_x q_x, E_x q_x \left(1 - q_x\right)\right)$

$$\widehat{q}_x = \frac{D_x}{E_x} \sim N\left(q_x, \frac{q_x \left(1 - q_x\right)}{E_x}\right)$$

Poisson model $D_x \sim \mathcal{N}(E_x^c \mu_x, E_x^c \mu_x)$ $\qquad \widehat{\mu}_x \sim N\left(\mu_x, \frac{\mu_x}{E_x^c}\right)$

Test statistics are generally obtained from the following:

Binomial:

$$z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s \left(1 - q_x^s\right)}} \quad \left(\approx \frac{O - E}{\sqrt{V}}\right) .$$

Poisson:

$$z_x = \frac{d_x - E_x^c \mu_x^s}{\sqrt{E_x^c \mu_x^s}} \quad \left(\approx \frac{O - E}{\sqrt{V}}\right) .$$

(normal under null hypothesis)

6.2.3 The tests

Chi squared: $X = \sum_{\text{all ages } x} z_x^2 \sim \chi^2(m)$, if $m = \#$ years of study.

Disadvantages:

1. There may be a few large deviations offset by substantial agreement over part of the table. The test will not pick this up.
2. There might be bias, that is, although not necessarily large, all the deviations may be of the same sign.
3. There could be significant groups of consecutive deviations of the same sign, even if not overall.

Signs Test: $X = \#\{z_x > 0\}$ Under the null hypothesis $X \sim \mathrm{Binom}(m, \frac{1}{2})$,

ignores the size of the deviations but it will pick up small deviations of consistent sign, positive or negative, and so it addresses point 2 above.

Cumulative deviations test

This again addresses point 2 and essentially looks very similar to the logrank test between two survival curves, which we will consider later in the course. If instead of squaring $d_x - E_x q_x^s$ or $d_x - E_x^c \mu_x^s$, we simply sum then

$$\frac{\sum (d_x - E_x q_x^s)}{\sqrt{\sum E_x q_x^s (1 - q_x^s)}} \sim N(0,1), \quad \text{approximately}$$

and

$$\frac{\sum (d_x - E_x^c \mu_x^s)}{\sqrt{\sum E_x^c \mu_x^s}} \sim N(0,1) \quad \text{approximately.}$$

Other tests

There are tests to deal with consecutive bias/runs of same sign. These are called the groups of signs test and the serial correlations test. Again, a very large number of years m are required to render these tests useful.

The cumulative deviations test may also be generalised to a *weighted cumulative deviations test*. Arbitrary weights (chosen without reference to the observed d_x) may be applied to each term in the sum to obtain

$$\frac{\sum w_x (d_x - E_x q_x^s)}{\sqrt{\sum w_x^2 E_x q_x^s (1 - q_x^s)}} \text{ and } \frac{\sum w_x (d_x - E_x^c \mu_x^s)}{\sqrt{\sum w_x^2 E_x^c \mu_x^s}},$$

(standard under Null hypothesis)

6.3 Non-parametric testing of survival between groups (Consider only 2 groups)

Event times are $0 < t_1 < t_2 < \cdots < t_m$. $d_{ij} = \#$ events at t_j in group i,

$n_{ij} = \#$ in risk set at t_j from group i, $\quad d_j = \#$ events at t_j, $\quad n_j = \#$ in risk set at t_j.

exactly 1. More generally, the null hypothesis predicts that the group identities of the individuals whose events are at time t_j are like a sample of size d_j without replacement from a collection of n_{1j} '1's and n_{2j} '2's. The distribution of d_{1j} under such sampling is called the hypergeometric distribution. It has

$$\text{expectation} = d_j \frac{n_{1j}}{n_j}, \text{ and } \quad \text{variance} =: \sigma_j^2 = \frac{n_{1j} n_{2j} (n_j - d_j) d_j}{n_j^2 (n_j - 1)}.$$

Note that if d_j is negligible with respect to n_j, this variance formula reduces to $d_j(\frac{n_{1j}}{n_j})(\frac{n_{2j}}{n_j})$, which is just the variance of a binomial distribution.

Adding some random weights:

$$M_k := \left(\sum_{j=1}^{k} W(t_j)\big(d_{1j} - n_{1j}\frac{d_j}{n_j}\big) \right)_{k=1}^{m},$$

Called a martingale

these will be random variables with expectation 0 and variance $\sum_{i=1}^{k} W(t_j)^2 \sigma_j^2$.

Applying CLT (despite increments not being indep, using martingale CLT)

$$Z := \frac{\sum_{j=1}^{m} W(t_j)\big(d_{1j} - n_{1j}\frac{d_j}{n_j}\big)}{\sqrt{\sum_{j=1}^{m} W(t_j)^2 \frac{n_{1j} n_{2j} (n_j - d_j) d_j}{n_j^2 (n_j - 1)}}}$$

normal under null hypothesis

6.3.2 Standard Tests

1. $W(t_j) = 1, \forall j$. This is the **log rank test**, (may give too much weight to later times) under observation. **Petos' test** uses a weight dependent on a modified estimated survival function, estimated for the whole study. The modified estimator is

2.

$$\widetilde{S}(t) = \prod_{t_j \le t} \frac{n_j + 1 - d_j}{n_j + 1}$$

and the suggested weight is then

$$W(t_j) = \widetilde{S}(t_{i-1}) \frac{n_j}{n_j + 1}$$

This has the advantage of giving more weight to the early events and less to the later ones where the population remaining is smaller. Much like the cumulative-deviations test, this avoids giving extra weight to excess deaths that come later.

3. $W(t_j) = n_j$ has also been suggested (Gehan, Breslow). This again downgrades the effect of the later times.

4. D. Harrington and T. Fleming [16] proposed a class of tests that include Petos' test and the logrank test as special cases. The **Fleming–Harrington tests** use

$$W(t_j) = \left(\widehat{S}(t_{j-1})\right)^p \left(1 - \widehat{S}(t_{j-1})\right)^q$$

6.3.3 Crossing or multiple alternatives

All of these test statistics may be written in the form

$$\frac{\sum (O_{1j} - E_{1j}) W_j}{\sqrt{\sum \sigma_{1j}^2 W_j^2}}$$

so pos/neg fluctuations cancel: could use

$$X := \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Asymptotically, this should have the χ^2 distribution with $(k-1)m$ degrees of freedom. However, as with any chi-square test, the approximation is only accurate if each category has at least about 5 expected events.

Chapter 7: Regression in a survival context

Parametric Any parametric model may be turned into a regression model by imposing a functional assumption that links the covariates to one or more model parameters. As we will discuss in section 7.4, there are traditional choices of parameters to modify in this way that provide natural interpretations in a survival context.

Semiparametric In a fully parametric model the parameters defining the survival probabilities are not separated from the parameters that describe the effect of the covariates. Assumptions that we impose on the hazard rates may bias the estimate of the parameters measuring covariate effects. Semiparametric approaches split the estimation of a baseline hazard rate — which is done nonparametrically — from the estimation of a small number of parameters that describe how individuals with particular parameter values differ from that baseline. By far the most popular semiparametric model is the Cox proportional hazards regression model, described in section 7.5.

7.2: Additive Hazards, Proportional (cox) Hazards, Accelerated Lifetimes

- **Additive hazards**: It seems natural when comparing two groups to measure the difference in survival in terms of the cumulative difference in hazard — which is to say, in expected number of events — and to suppose that each increment in the covariate might produce a fixed increment in cumulative hazard. In an epidemiological context this means that the output of the model is the expected number of events caused or prevented by a change in treatment or risk factors. One disadvantage to this approach — described in detail in section 7.12 is that there is the potential for estimated hazards to become negative in some parameter regimes, which is nonsensical.

- **Proportional hazards**: Another natural approach is to suppose that the effect of a treatment or a change in risk factor is proportional to the risk. If we write $h_0(t)$ for the *baseline hazard* at time t, then we are supposing the hazard for individual i at time t

$$h_i(t) = \rho_i h_0(t)$$

  where $h_0(t)$ is the baseline hazard, and ρ_i is a function of covariates, which may themselves be changing over time. Equivalently, we have $H_i(t) = \rho_i H_0(t)$ for the cumulative hazard, and

$$S_i(t) = S_0(t)^{\rho_i}$$

  This sort of model is also called *relative-risk* regression.

- **Acccelerated lifetimes** In this approach we say that there is a standard survival function $S_0(t)$, which applies to everyone, but different individuals run through the function at different rates. So individual i with acceleration parameter ρ_i will have survival function

$$S_i(t) = S_0(\rho_i t).$$

  Equivalently, we have $H_i(t) = H_0(\rho_i t)$ for the cumulative hazard, or $h_i(t) = \rho_i h_0(\rho_i t)$ for the hazard. AL models will not be considered in this course except in the context of parametric models.

Graphical Tests

Suppose the distinct subpopulations differ by an acceleration parameter. If we could plot S_g against $\log t$, for groups $g = 1, 2, \ldots, k$, then we would see the distinct curves differing by horizontal shifts, as

$$S_g(t) = S_0(\mathrm{e}^{\log \rho_g + \log t}) \ .$$

Similarly for $H_g(t)$. Thus, the plot of $\hat{S}_g(t)$ or $\hat{H}_g(t)$ may be used as a diagnostic for AL models, where we accept the AL assumption when we see an approximate agreement between the curves when shifted horizontally.

To interrogate the PH assumption we plot the log cumulative hazard estimate (or plot the cumulative hazard on a log scale). If distinct groups differ by a proportionality constant

$$\log H_g(t) = \log \rho_g + \log H_0(t),$$

So if we plot the $\log \hat{H}_g(t)$ against either t or $\log t$ (where g is, again, a group of individuals) we expect to see a vertical shift between groups. Note that $\log \hat{H} = \log(-\log \hat{S})$ (or $\log(-\log \widetilde{S})$, as a consequence of which this plot is known as the *complementary log log plot.*

Taking both models together it is clear that we could plot $\log\left(-\log \widehat{S}_g(t)\right)$ against $\log t$ as then we can check for *AL and PH in one plot.* Generally S_g will be calculated as the Kaplan–Meier estimator for group g.

- If the accelerated life model is plausible we expect to see a horizontal shift between groups.
- If the proportional hazards model is plausible we expect to see a vertical shift between groups.

Of course, if the data came from a Weibull distribution, with differences in the ρ parameter, it is simultaneously AL and PH. We see that

$$\log\big(-\log S_g(t)\big) = \log \rho_g + \alpha \log t.$$

Thus, survival curve estimates for different groups should appear approximately as parallel lines, which of course may be viewed as vertical or as horizontal shifts of one another.

7.3 Generalised Linear Survival Models - use GLMs to determine p_i as a factor of our variables

Any parametric model may be turned into an AL model by replacing t by $\rho_i t$. And it may be turned into a PH model by replacing the hazard $h(t)$ by $\rho_i h(t)$.

The most common link function is the logarithm, producing $\log \rho_i = \beta \cdot \mathbf{x}_i$, equivalently $\rho_i = \mathrm{e}^{\beta \cdot \mathbf{x}_i}$

Response: event time, censoring status, left trunc time (possibly)
Covariate:intercept, age,sex,bp,etc.

E.g Weibull $S_i(t) = \mathrm{e}^{-(\rho t)^\alpha}$ and $\rho = \mathrm{e}^{\beta \cdot \mathbf{x}}$ $\beta \cdot \mathbf{x} = \beta_0 + \beta_1 \,\mathrm{age}_i + \beta_2 \,\mathrm{sex}_i + \beta_3 \,\mathrm{sbp}_i + \beta_4 \,\mathrm{trt}_i,$

Test via $2 \log \widehat{L}_{\mathrm{weib}} - 2 \log \widehat{L}_{\mathrm{exp}} \sim \chi^2(1)$, asymptotically. (against exp)

Relative Risk regression model/ Cox Proportional Hazards

The most general relative-risk regression model represents the hazard rate for individual i as

$$h_i(t) = h\big(t \,\big|\, \mathbf{x}_i\big) = h_0(t) r\big(\beta, \mathbf{x}_i(t); t\big). \tag{7.1}$$

generally be assuming that r is a function only of B and x (no direct dependence on t), and in that case we will drop t from the notation,

Focus mainly on choice $r(\beta, \mathbf{x}) = e^{\beta^T \mathbf{x}} = e^{\sum \beta_j x_j}$ *Cox proportional hazards* regression model.

A note about coding of covariates: In order for the baseline survival to make sense, the point where all the covariates are 0 must correspond to a plausible vector of covariates for an individual who could be in the sample. Thus, quantitative covariates should be centred, either at the mean, the median, or some approximately central value. Note that this normalisation cancels out between numerator and denominator of the partial likelihood in Cox regression, but it becomes especially important when we add interaction effects, typically introduced as products of covariates.

Our modelling assumptions don't need to be 'true', just 'good enough'

7.6 Partial Likelihood We represent the data in the following form:

1. A list of event times $t_1 < t_2 < \cdots$. (We are assuming no ties, for the moment.)
2. The identity i_j of the individual whose event is at time t_j.
3. The values of all individuals' covariates (at times t_j, if they are varying).
4. The risk sets $\mathcal{R}_j = \{i : T_i \geq t_j\}$, the set of individuals who are at risk at time t_j.

The most common way to fit a relative-risk model is to split the likelihood into two pieces: The likelihood of the event times, and the conditional likelihood of the choice of subjects given the event times. First piece assumed to have little info on parameters, and **second is used to estimate B**

$$\pi(i \mid t) := \frac{h_i(t)}{h(t)} = \frac{r(\beta, \mathbf{x}_i(t); t)\mathbf{1}_{\{i \in \mathcal{R}(t)\}}}{\sum_{j \in \mathcal{R}(t)} r(\beta, \mathbf{x}_j(t); t)},$$

, the set of individuals at risk at time t. We have the partial likelihood

$$(\beta) = \prod_{t_i} \pi(i_j \mid t_j) = \prod_{t_i} \frac{r(\beta, \mathbf{x}_{i_j}(t_j); t_j)}{\sum_{l \in \mathcal{R}_j} r(\beta, \mathbf{x}_l(t_j); t_j)}. \qquad (7.5)$$

where $\mathbb{R}(t)$ is the **risk set**, the set of individuals at risk at time t. We have the partial likelihood

$$L_P(\beta) = \prod_{t_j} \pi(i_j \mid t_j) = \prod_{t_j} \frac{r(\beta, \mathbf{x}_{i_j}(t_j); t_j)}{\sum_{l \in \mathcal{R}_j} r(\beta, \mathbf{x}_l(t_j); t_j)}.$$

The partial likelihood is useful because it involves only the parameters , isolating them from the nonparametric (and often less interesting) ↵0. The maximiser of the partial likelihood has the same essential properties as the MLE.

Theorem 7.1. *Let $\hat{\beta}$ maximise L_P, as given in (7.5). Then $\hat{\beta}$ is a consistent estimator of the true parameter β_0, and $\sqrt{n}(\hat{\beta} - \beta_0)$ converges to a multivariate normal distribution with mean 0 and covariance matrix consistently approximated by $\mathcal{J}(\hat{\beta})^{-1}$, where $\mathcal{J}(\hat{\beta})$ is the observed information matrix, with (i, j) component given by*

$$-\frac{\partial^2}{\partial\beta_i \partial\beta_j} \log L_P(\beta).$$

7.7 Significance Testing (B = B0) - All chi-squared (p) under H0

Wald statistic: $\xi^2_W := (\hat{\beta} - \beta_0)^T \mathcal{J}(\beta_0)(\hat{\beta} - \beta_0)$ **Score statistic**: $\xi^2_{SC} = U(\beta_0)^T \mathcal{J}(\beta_0) U(\beta_0)$;

Likelihood ratio statistic: $\xi^2_{LR} = 2\left[\ell_P(\hat{\beta}) - \ell_P(\beta_0)\right]$, where $\ell_P := \log L_P$.

7.8 Estimating baseline hazard

Breslow's Estimator

Suppose the baseline survival is given by $\widehat{S_0}(t) = \mathrm{e}^{-\widehat{H_0}(t)}$,

estimator. We start from the constraint that our MLE for the cumulative hazard must confine the points of increase (which are the times when events might be observed) to the times when events actually were observed. Conditioning on that constraint[1] we want the baseline hazard estimator to have the form

$$\widehat{H_0}(t) = \sum_{t_j \le t} \widehat{h_0}(t_j),$$

$$\widehat{h_0} = \frac{1}{\sum\limits_{i \in \mathcal{R}_j} r(\beta, \mathbf{x}_i(t_j))} \quad \text{for cox} \quad \widehat{h_0} = \frac{1}{\sum\limits_{i \in \mathcal{R}_j} \mathrm{e}^{\beta \cdot \mathbf{x}_i(t_j)}}$$

In some sense the discrete estimates for $\widehat{h_0}(t_j)$ can be thought of as the maximum likelihood estimators from the full likelihood, provided we assume that the hazard distribution is discrete (which of course it generally is not). When $\hat{\beta} = 0$ or when the covariates are all 0, this reduces simply to the Nelson–Aalen estimator. Otherwise, we see that this is equivalent to a modified Nelson–Aalen estimator, where the size of the risk set is weighted by the relative risks of the individuals. In other words, the estimate of $\widehat{h_0}$ is equivalent to the standard estimate # events/time at risk, but now time at risk is weighted by the relative risk.

$$\ell(h) = \sum_{t_j} \log(1 - e^{-h_{i_j}(t_j)}) - \sum_{\substack{i \in \mathcal{R}_j \\ k \neq i_j}} h_k \quad = \sum_{t_j} \log(1 - e^{-\hat{r}_{i_j} h_0(t_j)}) - \sum_{\substack{i \in R_j \\ j \neq [j]}} \hat{r}_i h_0(t_j).$$

We estimate $h_0(t_j)$ by

$$0 = \frac{\hat{r}_{i_j} e^{-\hat{r}_{i_j}\hat{h}_0(t_j)}}{1 - e^{-\hat{r}_{i_j}\hat{h}_0(t_j)}} - \sum_{\substack{k \in \mathcal{R}_j \\ k \neq i_j}} \hat{r}_k \approx \frac{\hat{r}_{i_j}(1 - \hat{r}_{i_j}\hat{h}_0(t_j))}{\hat{r}_{i_j}\hat{h}_0(t_j)} - \sum_{\substack{k \in \mathcal{R}_j \\ k \neq i_j}} \hat{r}_k \quad = \frac{(1 - \hat{r}_{i_j}\hat{h}_0(t_j))}{\hat{h}_0(t_j)} - \sum_{\substack{k \in \mathcal{R}_j \\ k \neq i_j}} \hat{r}_k.$$

$$1 \approx \hat{h}_0(t_j) \left(\sum_{i \in \mathcal{R}_j} \hat{r}_i \right)$$

=> (giving the h hat as before)

Individual Risk Ratios

their cumulative hazard to age t is approximated by $\hat{H}(t \,|\, \mathbf{x}) = r(\hat{\beta}, \mathbf{x})\hat{H}_0(t)$.

In case of time-varying covariates we have an individual cumulative hazard

$$H(t \,|\, \mathbf{x}) = \int_0^t r(\beta, \mathbf{x}(u)) h_0(u) du,$$

approximated by

$$\hat{H}(t \,|\, \mathbf{x}) = \sum_{t_j \leq t} \frac{r(\hat{\beta}, \mathbf{x}(t_j))}{\sum_{i \in \mathcal{R}_j} r(\hat{\beta}, \mathbf{x}_i(t_j))}.$$

In the special case of Cox regression we have

$$\hat{H}(t \,|\, \mathbf{x}) = \sum_{t_j \leq t} \frac{e^{\hat{\beta} \cdot \mathbf{x}(t_j)}}{\sum_{i \in \mathcal{R}_j} e^{\hat{\beta} \cdot \mathbf{x}(t_j)}}$$

7.9 Dealing with ties dealing with tied event times for the Cox model.

Until now in this section we have been assuming that the times of events are all distinct. In situations where event times are equal, we can carry out the same computations for Cox regression, only using a modified version of the partial likelihood. Suppose $\mathcal{R}_j$ is the set of individuals at risk at time t_j, and $\mathcal{D}_j$ the set of individuals who have their event at that time. We assume that the ties are not real ties, but only the result of discreteness in the observation. Then the probability of having precisely those individuals at time t_j will depend on the order in which they actually occurred. For example, suppose there are 5 individuals at risk at the start, and two of them have their events at time t_1. If the relative risks were $\{r_1, \ldots, r_5\}$, then the first term in the partial likelihood would be

$$\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_2}{r_2 + r_3 + r_4 + r_5} + \frac{r_2}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_1}{r_1 + r_3 + r_4 + r_5}.$$

The number of terms is $d_j!$, so it is easy to see that this computation quickly becomes intractable.

A very good alternative — accurate and easy to compute — was proposed by B. Efron.

the first contribution to the partial likelihood becomes

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4 + r_5)(\frac{1}{2}(r_1 + r_2) + r_3 + r_4 + r_5)}.$$

More generally, the log partial likelihood becomes

$$\ell_P(\beta) = \sum_{t_j} \sum_{i \in \mathcal{D}_j} \log r(\beta, \mathbf{x}_i(t_j)) - \sum_{k=0}^{d_j - 1} \log \left(\sum_{i \in \mathcal{R}_j} r(\beta, \mathbf{x}_i(t_j) - \frac{k}{d_i} \sum_{i \in \mathcal{D}_j} r(\beta, \mathbf{x}_i) \right)$$

We take the same approach to estimating the baseline cumulative hazard:

$$\hat{H}_0(t) = \sum_{t_j \le t} \sum_{k=0}^{d_j - 1} \left(\sum_{i \in \mathcal{R}_j} r\big(\beta, \mathbf{x}_i(t_j)\big) - \frac{k}{d_i} \sum_{i \in \mathcal{D}_j} r\big(\beta, \mathbf{x}_i(t_j)\big) \right)^{-1}.$$

An alternative approach, due to Breslow, makes no correction for the progressive loss of risk in the denominator:

$$\ell_P^{Breslow}(\beta) = \sum_{t_j} \sum_{i \in \mathcal{D}_j} \log r(\beta, \mathbf{x}_i(t_j)) - d_j \log \sum_{i \in \mathcal{R}_j} r\big(\beta, \mathbf{x}_i(t_j)\big).$$

This approximation is always too small, and tends to shift the estimates of β toward 0. It is widely used as a default in software packages (`SAS`, not `R`!) for purely historical reasons.

Chapter 8: Goodness-of-fit tests for survival models

For Proportional Hazards Assumption:

Log cumulative hazard plot

The simplest graphical test require that the covariate take on a few discrete values, with a substantial number of subjects observed in each category. If the covariate is continuous we stratify it, defining a new categorical covariate by the original covariate being in some fixed region.

The first approach is to consider, for categories $1, \ldots, m$, the Nelson–Aalen estimators $\widehat{H}_i(t)$ of the cumulative hazard for individuals in category i. If any relative-risk model holds then

$\widehat{H}_i(t) = r(\beta, i)\widehat{H}_0(t)$, so that $\log \widehat{H}_i(t) - \log \widehat{H}_j(t) \approx \log r(\beta, i) - \log r(\beta, j)$

should be approximately constant. (wrt t)

Andersen plot

In the Andersen plot we plot all the pairs $(\widehat{H}_i(t), \widehat{H}_j(t))$. If the proportional hazards assumption holds then each pair (i, j) should produce (approximately) a straight line through the origin. It is known (*cf.* section 11.4 of [22]) that when the ratio of hazard rates $\alpha_i(t)/\alpha_j(t)$ is increasing, the corresponding Andersen plot is a convex function; decreasing ratios produce concave Andersen plots.

8.1.3 Arjas plot

The Arjas plot is a more sophisticated graphical test, that is capable of testing the proportional hazards assumption for a single categorical covariate, within a model that includes other covariates (that might follow a different sort of model).

Suppose we have fit the model $h_i(t) = h(t|\mathbf{x}_i)$ for the hazard rate of individual i — for example, it might be $h_i(t) = e^{\hat{\beta}\cdot\mathbf{x}_i}\alpha_0(t)$, but it might be something else — and we are interested to decide whether an additional (categorical) covariate z_i (taking on values 0 and 1) ought to be included as well. For each individual we have an estimated cumulative hazard $\widehat{H}(t|\mathbf{x}_i)$. Define the weighted time on test for individual i at event time t_j as $\widehat{H}(t_j \wedge T_i|\mathbf{x}_i)$, and the total time on test for level g (of the covariate z) as

$$\mathrm{TOT}_g(t_j) = \sum_{i:z_i=g} \widehat{H}(t_j \wedge T_i|\mathbf{x}_i);$$

number of events at level g will be

$$N_g(t_j) = \sum_{i:z_i=g} \delta_i \mathbf{1}_{\{T_i \le t_j\}}.$$

$(a \wedge b = \min\{a, b\}.)$ The idea is that if the covariate z has no effect, the difference $N_g(t_j) - \mathrm{TOT}_g(t_j)$ has expectation zero for each t_j, so a plot of N_g against TOT would lie close to a straight line with 45° slope. If levels of z have proportional hazards effects, we expect to see lines of different slopes. If the effects are not proportional, we expect to see curves that are not lines.

8.2 General Principles of Model Selection

Look at deviations of the data from the best-fit model, the *residuals* (graphically or test statistics). Not same as hypothesis testing, as we never really expect data to perfectly fit our model. Asking 'Are residuals so bad that we must reject our Null?'. Also, systematic differences might lead the way to a new model. Also points out outliers we may want to remove, often detected with *Deviance residuals*

8.3 Cox-Snell Residuals

them. There is a large family of different residuals that have been defined for survival models, each of which is useful for different parts of the task of model diagnostics, including:

- Generally evaluating the appropriateness of a regression model (such as Cox proportional hazards or additive hazards);
- Specifically evaluating assumptions of the regression model (such as the proportional hazards assumption, or the log-linear action of the covariates;
- Finding specific outlier individuals in an otherwise reasonably well specified model.

The most basic version is called the *Cox–Snell residual*. It is based on the observation that if T is a sample from a distribution with cumulative hazard function H, then $H(T)$ has an exponential distribution with parameter 1.

Given a parametric model $H(T, \beta)$ we would then generate and evaluate Cox–Snell residuals as follows:

1. We use the samples (T_i, δ_i) to estimate a best fit $\hat{\beta}$;
2. Compute the residuals $r_i := H(T_i, \hat{\beta})$;
3. If the model is well specified — a good fit to the data — then (r_i, δ_i) should be like a right-censored sample from a distribution with constant hazard 1.
4. A standard way of evaluating the residuals is to compute and plot a Nelson–Aalen estimator for the cumulative hazard rate of the residuals. The null hypothesis — that the data came from the parametric model under consideration — would predict that this plot should lie close to the line $y = x$.

<u>For Cox Proportional Hazards Model</u>

1. We use the samples $(T_i, \mathbf{x}_i, \delta_i)$ to estimate a best fit $\hat{\beta}$;
2. We compute the Breslow estimator $\widehat{H}_0(t)$ for the baseline hazard;
3. Compute the residuals $r_i := e^{\hat{\beta}^T \mathbf{x}_i} \widehat{H}_0(T_i)$.

After this, we proceed as above. Of course, there is nothing special about the Cox log-linear form of the relative risk. Given any relative risk function $r(\hat{\beta}, \mathbf{x})$, we may define residuals $r_i := r(\hat{\beta}, \mathbf{x}_i)\widehat{H}_0(T_i)$.

8.5 Schoenfeld Residuals - for proportional hazards assumption

as an exercise, but the definition of the j-th Schoenfeld residual for parameter β_k is

$$S_{kj}(t_j) := X_{i_j k}(t_j) - \bar{X}_k(t_j),$$

where as usual i_j is the individual with event at time t_j, and

$$\bar{X}_k(t_j) = \frac{\sum_{i \in \mathcal{R}_j} X_{ik}(t) e^{\hat{\beta}^T X_i(t)}}{\sum_{i \in \mathcal{R}_j} e^{\hat{\beta}^T X_i(t)}}$$

is the weighted mean of covariate X_k at time t. Thus the Schoenfeld residual measures the difference between the covariate at time t and the average covariate at time t. If β_k is constant this has expected value 0. If the effect of X_k is increasing we expect the estimated parameter β_k to be an overestimate early on — so the individuals with events then have lower X_k than we would have expected, producing negative Schoenfeld residuals; at later times the residuals would tend to be positive. Thus, increasing effect is associated with increasing Schoenfeld residuals. Likewise decreasing effect is associated with decreasing Schoenfeld residuals. As with the martingale residuals, we typically make a smoothed plot of the Schoenfeld residuals, to get a general picture of the time trend.

We can also make a formal test of the hypothesis β_k is constant by fitting a linear regression line to the Schoenfeld residuals as a function of time, and testing the null hypothesis of zero slope, against the alternative of nonzero slope. Of course, such a test will have little or no power to detect nonlinear deviations from the hypothesis of constant effect — for instance, threshold effects, or changing direction.

8.6 Martingale Residuals

Intuitively, the martingale residual for an individual is the difference between the number of observed events for the individual and the expected number under the model. The expected number of events from time 0 to t for individual i is $H_i(t)$, so in principle this is

$$\delta_i - \int_0^{T_i} h_0(s) \mathrm{e}^{\beta \cdot \mathbf{x}_i(s)} ds.$$

We turn this into a residual — something we can compute from the data – by replacing the integral with respect to the hazard by the differences in the estimated cumulative hazard

$$\widetilde{M}_i(t) := \delta_i - \widehat{H}_i(T_i) = \delta_i - \sum_{t_j \le T_i} \mathrm{e}^{\hat{\beta} \cdot \mathbf{x}_i(t_j)} \hat{h}_0(t_j) = \delta_i - \sum_{t_j \le T_i} \mathrm{e}^{\hat{\beta} \cdot \mathbf{x}_i(t_j)} \frac{1}{\sum_{\ell \in \mathcal{R}_j} \mathrm{e}^{\beta \cdot \mathbf{x}_\ell(t_j)}}$$

It differs from the (negative of the) Cox–Snell residual only by the addition of δ_i. When the covariates are constant in time,

$$\widetilde{M}_i = \delta_i - e^{\beta^T \mathbf{x}_i} \widehat{H}_0(T_i).$$

An individual martingale residual is a very crude measure of deviation, since for any given t the only observation here that is relevant to the survival model is δ_i, which is a binary observation.

If there are no ties, or if we use the Breslow method for resolving ties, sum of all the martingale residuals is 0.

Applying Martingale Residuals

Martingale residuals are not very useful in the way that linear-regression residuals are, because there is no natural distribution to compare them to. The main application is to estimate appropriate modifications to the proportional hazards model by way of covariate transformation: Instead of a relative risk of $e^{\beta x}$, it might be $e^{f(x)}$, where $f(x)$ could be $\mathbf{1}_{\{x<x_0\}}$ or $\sqrt{x}$, or something else.

We assume that in the population $\mathbf{x}_i$ and z_i are independent

Suppose the data $(T_i, \mathbf{x}_i(\cdot), z_i, \delta_i)$ are sampled from a relative-risk model with two covariates: a vector $\mathbf{x}_i$ and an additional one-dimensional covariate z_i, with

$$\log r(\beta, \mathbf{x}, z) = \beta^T \mathbf{x} + f(z);$$

that is, the Cox regression model holds except with regard to the last covariate, which acts as $h(z) := e^{f(z)}$. Let $\hat{\beta}$ be the p-dimensional vector corresponding to the Cox model fit without the covariate z, and let $\widetilde{M}_i$ be the corresponding martingale residuals.

Another complication is that we use $\hat{\beta}$ instead β (which we don't know). We will derive the relationship under the assumption that they are equal; again, errors in estimating β will make the conclusions less correct. For large n we may assume that β and $\hat{\beta}$ are close.

$$\bar{h}(s, \mathbf{x}) := \frac{\mathbb{E}[\mathbf{1}_{\text{at risk time } s} e^{f(z)} \,|\, \mathbf{x}]}{\mathbb{P}\{\text{at risk time } s \,|\, \mathbf{x}\}} = \mathbb{P}\big\{ h(z) \,\big|\, \text{at risk time } s; \mathbf{x}\big]; \quad \bar{h}(s) := \mathbb{E}\big[\bar{h}(s, \mathbf{x})\big].$$

We assume that $\bar{h}(s, \mathbf{x})$ is approximately the constant $\bar{h}(s)$.

Fact

$$\mathbb{E}\big[\widetilde{M} \,\big|\, z\big] \approx \frac{\sum \delta_i}{n} \left(f(z) - \log \bar{h}(\infty)\right).$$

Thus, we may estimate $f(z_0)$ by estimating the local average of $\widetilde{M}$, averaged over z close to z_0. For instance, we can compute a LOESS smooth curve fit to the scatterplot of points $(z_i, \widetilde{M}_i)$.

The basic idea is, the martingale residuals measure the excess events, the difference between the expected and observed number of events. If we compute the expectation without taking account of z, then individuals whose z value has $f(z)$ large positive will seem to have a large number of excess events; and those whose $f(z)$ is large negative will seem to have fewer events than expected.

8.7 Outliers and Influence

Deviance Residuals

The martingale residual for an individual is

$$\text{observed } \# \text{ events} - \text{expected } \# \text{ events}$$

for individual i. In principle, large values indicate outliers — results that individually are unexpected if the model is true. The problem is, these are highly-skewed variables — they take values between 1 and $-\infty$ and it is hard to determine what their distribution should look like.

We define the *deviance residual* for individual i as

$$d_i := \operatorname{sgn}(\widetilde{M}_i) \left\{ -2\big[\widetilde{M}_i + \delta_i \log(\delta_i - \widetilde{M}_i)\big] \right\}^{1/2}. \tag{8.3}$$

This choice of scaling is inspired by the intuition that each d_i should represent the contribution of that individual to the total model deviance.[1] Whereas the martingale residuals are between $-\infty$ and 1, the deviance residuals should have a similar range to a standard normal random variable. Thus, we treat values outside a range of about -2.5 to $+2.5$ as outliers, potentially

Delta-Beta Residuals Individuals with high values of $\Delta\beta$ should be looked at more closely

Recall that the *leverage* of an individual observation is a measure of the impact of that observation on the parameters of interest. The "delta–beta" residual for parameter β_k and subject i is defined

$\Delta\beta_{ki} := \hat{\beta}_k - \hat{\beta}_{k(i)}$, where $\hat{\beta}_{k(i)}$ is the estimate of $\hat{\beta}_k$ with individual i removed.

We can approximate it by a combination of the Fisher information and the Schoenfeld residual

8.9 Predictive Diagnosis; Statisticians care about model fit, practitioners care about predictive power

ROC and AUC

ROC curves[2] are a bedrock tool for evaluating diagnostic ability. Consider a family of binary classifiers — procedures for allocating individual observations into one of two categories, 0 or 1, negative or positive. Each classifier has a False Positive Rate (FPR) and a True Positive Rate (TPR), the fraction of zeros incorrectly classified as ones, and the fraction of ones correctly classified as ones. It is easy to raise the TPR by loosening the criteria for calling a case "positive", but that will also raise the FPR. There is always a tradeoff. The ROC is a plot of TPR against FPR over the whole family of classifiers.

The paradigm classifier is a thresholded score: We have a single real-valued covariate, or a linear combination x_i of covariates, and we call subjects positive when their score x_i exceeds a certain threshold K. Shifting K down increases the number of positives, true and false. If the score has nothing to do with the correct categorisation then we expect to classify approximately the same fraction of each group as positive, hence FPR = TPR, and the ROC is a the diagonal $y = x$. On the other hand, if group 0 has generally much lower values of x_i than group 1, then a high threshold will sweep up mostly true positives, and the ROC will have a high slope at first.

One useful summary of the ROC curve is the Area Under the Curve, or AUC. This is just what it says. The effective minimum AUC is 0.5, corresponding to the 45 diagonal for a useless classifier. Perfect classifier will have AUC = 1

Suppose we have random variables X_0 and X_1, representing independent draws from the distribution of predictors for negative and positive cases respectively. The basic identity for AUC, whose proof we leave as an exercise, is

$$\text{AUC} = \mathbb{P}\{X_1 > X_0\} + \frac{1}{2}\mathbb{P}\{X_1 = X_0\}. \tag{8.4}$$

(This is for a predictor for which large values are supposed to indicate a likely positive.) The empirical AUC for a data set $\{x_{01}, \ldots, x_{0n_0}; x_{11}, \ldots, x_{1n_1}\}$ (representing n_0 observations of covariates of negative cases, and n_1 observations of covariates of positive cases) is then

$$\text{empirical AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left(\mathbf{1}\{x_{1j} > x_{0i}\} + \frac{1}{2} \cdot \mathbf{1}\{x_{1j} = x_{0i}\} \right). \tag{8.5}$$

Since the theoretical distribution is not really observed, we will usually understand (8.5) to be the definition of the AUC.

In its simplest form, survival analysis is a series of classification problems. At each time we are grouping subjects into two classes: Alive or dead, did or did not already have an event. We can ask, at each time, how well a certain regression Model diagnostics 146 model helped us to make the classification. At each time we can compute an ROC. We can also compute AUC at each time, and plot the development over time, to get a picture of how the effectiveness develops over time.

The simplest approach conceptually would be to produce a *cumulative* ROC: Fix a time t_0, and classify individuals as alive or dead at time t_0. We can produce an ROC related to this classification, and the attendant AUC:

$$\text{AUC}_{\text{cum}}(t_0) = \frac{\sum_{i \in \mathcal{R}(t)} \sum_{j \in \mathcal{D}(0,t_0)} \left(\mathbf{1}\{x_j > x_i\} + \frac{1}{2} \cdot \mathbf{1}\{x_i = x_j\} \right)}{\#\mathcal{R}(t_0) \cdot \#\mathcal{D}(0, t_0)}, \tag{8.6}$$

where $\mathcal{D}(0, t)$ is defined to be the set of individuals who have event times (not censoring times) in the interval $(0, t)$

The *dynamic ROC* looks instead at the events happening at time t as the "positive" cases, comparing them to those at risk at time t. This will be unaffected by censoring and truncation as long as these are non-informative and independent of the covariates. Of course, such an ROC is generally meaningless as a graphical representation, since there will be only one, or a few, subjects dying exactly at time t, but it can be used as the basis for defining a dynamic AUC.[3] For an event time t, recalling the notation $\mathcal{D}(t)$ for the set of individuals with events at time t and with $d(t) = \#\mathcal{D}(t)$, we have

$$\mathrm{AUC}(t) = \frac{1}{d(t)(n(t) - d(t))} \sum_{i \in \mathcal{R}(t+)} \sum_{j \in \mathcal{D}(t)} \left(\mathbf{1}\{x_j > x_i\} + \frac{1}{2} \cdot \mathbf{1}\{x_i = x_j\} \right). \tag{8.7}$$

Note that $\mathcal{R}(t+) = \mathcal{R}(t) \setminus \mathcal{D}(t)$ is the set of individuals at risk **after** any events at time t have transpired. We can get information about the predictive value of the model over time by plotting a smoothed interpolation of AUC as a function of time, much as we did for the proportional hazards effect over time by smoothing the Schoenfeld residuals.

Alternatively, given a fitted survival regression model that assigns individual hazards on the basis of covariates we may define a theoretical dynamic ROC based on the model. Suppose $h(t|x)$ is the hazard at time t for an individual with covariate x, and we make a classification based on whether $x \geq \xi$ or $x < \xi$. If an event occurs, the probability that they are correctly classified (the TPR) and the probability that they are falsely classified (the FPR) are

$$\mathrm{TPR}(\xi) = \frac{\sum_{i \in \mathcal{R}(t): x_i \geq \xi} h(t|x_i)}{\sum_{i \in \mathcal{R}(t)} h(t|x_i)}, \qquad \mathrm{FPR}(\xi) = \frac{\sum_{i \in \mathcal{R}(t): x_i < \xi} h(t|x_i)}{\sum_{i \in \mathcal{R}(t)} h(t|x_i)}.$$

The ROC plots the pairs $(\mathrm{FPR}(\xi), \mathrm{TPR}(\xi))$. Note that in the special case of a proportional hazards model these simplify to ratios of total risk functions.

Concordance Measure

With the AUC we have an embarrassment of riches — one for every time t. How can we summarise the overall value of a statistical predictor? One natural idea is to take an average of the dynamic AUC at different times, weighted by the number of individuals at risk:

$$C = \frac{\sum_{t_j} \mathcal{D}(t_j)\mathcal{R}(t_j+)\,\mathrm{AUC}(t_j)}{\sum_{t_i} \#\mathcal{D}(t_j)\#\mathcal{R}(t_j+)}.$$

This measure is called *Harrell's concordance measure*, or simply the *C-index*. The above expression simplifies to

$$C = \frac{\#\{(i,j) \,:\, T_i < T_j \text{ and } x_i > x_j\} + \frac{1}{2}\#\{(i,j) \,:\, T_i < T_j \text{ and } x_i = x_j\}}{\#\{(i,j) \,:\, T_i < T_j\}}, \tag{8.8}$$

where the pairs (i,j) are pairs of individuals from the observed population where T_i is an event time. This yields the following natural interpretation for the C-index: Pick two subjects at random from the population. Use x_i to predict which event occurs first. (That is, predict that the individual with the higher value of x has the earlier event.) Then C is the probability that the prediction is correct. (If the events are tied then pick again.)

Chapter 9: Correlated events and Repeated events

The survival models that we have considered depend upon the fundamental assumption that event times are independent. There are many settings where this assumption is unreasonable:

• **Clustered data**: • **Multiple events**: (per indiv) • **Competing events**: (different events)

9.2 Time-to-first-event analysis

When we are confronted with multiple events for a single individual, one easy approach to eliminating the complication is to throw out the correlated data. A common approach, called *time-to-first-event analysis*, is to define the event of interest to be simply the first event. Later events for the same individual are simply ignored.

This trivially produces independent observations, and will yield results that are consistent and unbiased, but at the expense of losing a substantial amount of relevant information from the data. Of course, this approach is possible only when the correlated events correspond to a single individual, so that all covariates are identical. It would not apply to an example such as the diabetetic retinopathy study described above, where the two correlated events for one individual differ in their treatment group membership.

9.3 Clustered data

9.3.1 Stratified baseline

If there are a small number of large correlated groups of survival times, we may represent the correlation within the groups by using a semiparametric model and stratifying the baseline hazard by group. Note that there is no way to distinguish between a cluster of survival times being "dependent", and times sharing a group-determined hazard function.

Suppose we have k categories of individuals, $c_i = 1, \ldots, k$, each with its own baseline hazard $h_0^{(c)}(t)$, so that individuals in category c_i with covariates $\mathbf{x}_i(t)$ have hazard

$$h_i(t) = h_0^{(c_i)}(t) r(\beta, \mathbf{x}_i(t))$$

at time t. Then we have a partial likelihood for the observation that individual i_j had the unique event at time t_j

$$L_P(\beta) = \prod_{t_j} \frac{r(\beta, \mathbf{x}_{i_j}(t_j))}{\sum_{i \in \mathcal{R}_j : c_i = c_{i_j}} r(\beta, \mathbf{x}_i(t_j))}.$$

9.3.2 The sandwich estimator for variance

interval for the parameters. In the extreme case, imagine that we had a dataset where each individual had the same survival time repeated multiple times. More observations reduces the variance estimate; but we would want to have a procedure that would be able to recognise that the duplicated observations are not actually providing additional information, and that would hence return the same variance estimate as the data set without duplication.

The standard approach in such cases is to replace the Fisher-information-based estimate for variance $J_n(\hat{\beta})^{-1}$ by the sandwich estimator

$$J_n(\hat{\beta})^{-1} V_n(\hat{\beta}) J_n(\hat{\beta})^{-1}, \tag{9.1}$$

where $V_n(\hat{\beta})$ is an estimate of the variance-covariance matrix of the score function.

Calculating Vn is hard, do for Cox regression in notes

$$U(\beta) = \sum_{j=1}^{k} \left(\mathbf{x}_{i_j} - \bar{\mathbf{x}}(t_j) \right), \qquad V_n(\hat{\beta}) = n^{-1} \sum_{i=1}^{n} U_i(\hat{\beta}) U_i(\hat{\beta})^T$$

,

$$V_n(\hat{\beta}) = \sum_{c=1}^{C} u_c(\hat{\beta}) u_c(\hat{\beta})^T,$$

For clustered

$$u_c(\hat{\beta}) = \sum_{i=1}^{n_c} \delta_{ic} \left\{ \mathbf{x}_{ic}(T_{ic}) - \bar{\mathbf{x}}(T_{ic}, \hat{\beta}) \right\} - \sum_{i=1}^{n_c} \sum_{t_j \le T_{ic}} \left\{ \mathbf{x}_{ic}(t_j) - \bar{\mathbf{x}}(t_j, \hat{\beta}) \right\} \mathrm{e}^{\hat{\beta} \cdot \mathbf{x}_{ic}(t_j)} \, \mathrm{d}\hat{H}_0(t_j).$$

9.4 Multiple Events

9.4.1 The Poisson model

events for each indiv is Poisson alpha, alpha constant

$$\ell(\alpha) = N \log \alpha - \alpha \sum T_i, \quad \hat{\alpha} = \frac{N}{\sum T_i}.$$

9.4.2 The Poisson regression model

Say each alpha$_i$ is dependent on covariates

$$\ell(\alpha) = \sum_{i=1}^{n} N_i \log \alpha(\mathbf{x}_i) - T_i \alpha(\mathbf{x}_i).$$

The most common parametric form is $\alpha = \exp\{\beta \cdot \mathbf{x}\}$, where $\beta = (\beta_0, \ldots, \beta_p)$, and we take $x_{i0} \equiv 1$. The log likelihood then becomes

$$\ell(\beta) = \sum_{k=0}^{p} \beta_k \sum_{i=1}^{n} N_i x_{ik} - \sum_{i=1}^{n} T_i e^{\beta \cdot \mathbf{x}_i}. \tag{9.3}$$

The MLE then satisfies the equations $\sum_{i=1}^{n} N_i x_{ik} = \sum_{i=1}^{n} x_{ik} T_i e^{\hat{\beta} \cdot \mathbf{x}_i}$.

This fits into the framework of GLM (generalised linear model), and may be fit in `R` using any of the standard GLM functions. Note that we are modelling $N_i \sim \mathrm{Po}(\mu_i)$, where

$\log \mu_i = \log T_i + \beta \cdot \mathbf{x}_i$. We call $\log T_i$ an *offset* in the model.

9.4.3 The Andersen–Gill model

The Poisson regression model makes sense if we believe the event intensity is constant, or if all individuals Another popular generalisation of the Poisson model, introduced in 1982 by Andersen and Gill [2], is a semi-parametric relative-risk regression model, essentially equivalent to the Cox proportional hazards regression model. The only change is that the at-risk indicator $Y_i(t)$ for an individual will not, in general, become 0 after an event. Partial likelihood is defined exactly as in (7.5), and Breslow's formula still defines an estimate of cumulative intensity (rather than cumulative hazard).

The model assumes that differences between individuals are completely described by the relative-risk function determined by their covariates. If we are unsure — as we generally will be — we can robustify the variance estimates as in section 9.3.2 by adding a `+cluster(id)` term. Alternatively, we can add a hidden frailty term to the model, as described below in section 9.5.

9.5 Shared frailty model

One way to deal with the correlation among multiple events for the same individual (or for linked individuals) is by explicitly modelling the variation in hazard rate with a random effect, generally called a *frailty* term in the survival context. The most common version is a relative risk term $e^{\omega_{\text{group}}}$, where ω_{group} is an unobserved covariate with distribution assumed to have a particular form, usually either gamma or Gaussian.

9.5.1 Negative-binomial model

The simplest version of the frailty model generalises the Poisson model: Individuals accrue events at a constant rate, but with the unknown constant dependent on the individual. For example, the Poisson model doesn't really make much sense in the example discussed in section 9.4.2. Individuals may be presumed to have differing predispositions to offend. Thus, it is not surprising that the number of offences is more spread out than you would expect under the Poisson model, which posits that everyone offended at the same rate.

We may generalise the Poisson regression model to better fit overdispersed data by adding a frailty term. That is, in place of (9.4) we represent the individual intensity by

$$\log \mu_i = \log \lambda_i + \log T_i + \beta \cdot \mathbf{x}_i. \tag{9.5}$$

The term λ_i, called a *multiplicative frailty*, represents the individual relative rate of producing events. The λ_i are treated as random effects, meaning that they are not to be estimated individually — which would not make sense — but rather, they are taken to be i.i.d. samples from a simple parametric distribution. When the frailty λ has a gamma distribution (with parameters (θ, θ), because we conventionally take the frailty distribution to have mean 1), and N is a Poisson count conditioned on λ with mean $\lambda\alpha$, then N has probability mass function

$$\mathbb{P}\{N = n\} = \frac{\Gamma(n+\theta)}{n!\Gamma(\theta)} \left(\frac{\theta}{\theta+\alpha}\right)^{\theta} \left(\frac{\alpha}{\theta+\alpha}\right)^{n},$$

which is the negative binomial distribution with parameters θ and $\alpha/(\theta+\alpha)$. (The calculation is left as an exercise.) Therefore this is called the *negative binomial regression model*. We can fit it with the `glm.nb` command in `R`. If we apply it to the same data as before, we get the following output:

9.5.2 Frailty in proportional hazards models

We can use shared frailty to account for correlated times in proportional hazards regression, whether these are unordered (clustered) times, or recurrent events. The model fitting functions numerically exactly like any other random-effects model: We treat the individual unknown frailties as unobserved data, whose expected values given the observed data may be calculated. Given the individual frailties, we may maximise the parameters, and so loop through the EM algorithm. The calculations are carried through automatically by the `coxph` function in `R`, as long as we add a `+ frailty(id)` (or whichever variable we are grouping by) term to the formula. The output will include a p-value estimate for the individual

Is the frailty term actually appropriate to the data? We may test the null hypothesis that there is no individual frailty with a likelihood ratio test. The null-hypothesis log-likelihood is simply the log partial likelihood for a traditional model without a frailty term. The alternative log likelihood is the log of the integrated partial likelihood — that is, integrated over the distribution of the frailty — called the `I-likelihood` in the `R` output.