Survival Data

Part III Course, Lent 2010

Revision Notes

Daniel Guetta

guetta@cantab.net

Based on notes by Peter Treasure

Introduction

Introduction

- Survival data is **time-to-event analysis**
    - At most one event per subject
    - Highly positively skewed data
    - Censoring/truncation of data
- $T \geq 0$ is a **random variable** which contains the **time-to-event**. $T = 0$ is the well-defined start.
- Types of missing data:
    - **Censoring** (left/right) refers to a situation in which we only know that an event happened before/after a certain time.
    - **Truncation** (left/right) refers to a situation in which, if an event happened before/after a certain time, we have *no information* about that event.
- It is important for missing data to be **uninformative** – in other words, *the distribution of potential times $T > t$ for uncensored individuals is the same as for an individual censored at t, all other things being equal.*

Notation and Distributions

- Notation
    - Let there be n individuals
    - Let x_i be either the observed event time or the time of censoring.
    - Let $v_i = 1$ for observed events and 0 for censored events.
    - Let a_j be only those times at which an event occurs.
- Distributions
    - **Density** $\boxed{f(t \mid \theta)}$, such that $\mathbb{P}(a < T < b) = \int_a^b f(t \mid \theta)\,\mathrm{d}t$
    - **Survivor function** $\boxed{F(t \mid \theta) = \int_t^\infty f(t \mid \theta)\ \mathrm{d}t}$, probability of suriviing *more* than t. Note that $f(t) = -F'(t)$.

- **Hazard** is given by

$$\begin{aligned} h(t \mid \theta) &= \lim_{\Delta \to 0} \frac{\mathbb{P}(t < T \le t + \Delta \mid T > t, \theta)}{\Delta} \\ &= \lim_{\Delta \to 0} \frac{1}{\mathbb{P}(T > t \mid \theta)} \cdot \frac{\mathbb{P}(t < T \le t + \Delta \mid \theta)}{\Delta} \end{aligned}$$

$$\boxed{h(t \mid \theta) = \frac{f(t \mid \theta)}{F(t \mid \theta)}}$$

- **Integrated hazard** is given by

$$\begin{aligned} H(t \mid \theta) &= \int_0^t h(u \mid \theta)\ \mathrm{d}u \\ &= \int_0^t \frac{-F'(u \mid \theta)}{F(u \mid \theta)}\ \mathrm{d}u \\ &= -\log\left(F(t \mid \theta)\right) - \log(1) \end{aligned}$$

$$\boxed{H(t \mid \theta) = -\log\left(F(t \mid \theta)\right)}$$

And so $F(t \mid \theta) = \exp\left(-H(t \mid \theta)\right)$

- Note that if $F(t)$ is a survivor function, then
  - $F(\lambda t), \lambda > 0$ is also a survivor function (accelerated-life family).
  - $F(t)^k, k > 0$ is also a survivor function (proportional hazards family).
- Two specific distributions
  - ***Exponential distribution***
    - $f(t) = \rho e^{-\rho t}$
    - $F(t) = e^{-\rho t}$
    - $h(t) = \rho$
    - $H(t) = \rho t$
  - ***Weibull Distribution***
    - $f(t) = k\rho(t\rho)^{k-1} \exp\left\{-(\rho t)^k\right\}$
    - $F(t) = \exp\left\{-(\rho t)^k\right\}$
    - $h(t) = k\rho^k t^{k-1}$
    - $H(t) = (\rho t)^k$
    - Consider that if two Weibull distributions have the same k but different ρ (say ρ and $c\rho$), then

$$F(t) = \exp\left\{-(c\rho t)^k\right\} = \exp\left\{-(\rho t')^k\right\}$$

$$F(t) = \exp\left\{-(c\rho t)^k\right\} = \exp\left\{-c^k(\rho t)^k\right\} = \left[\exp\left\{-(\rho t)^k\right\}\right]^{c^k}$$

Thus, the two distribution belong to the same proportional hazards and accelerated life family.

Inference

Parametric inference

- If an individual is observed at x_i, then $f(x_i \mid \theta)$ is contributed to the likelihood. If an individual is censored at x_i, then $F(x_i \mid \theta)$ is contributed (since all we know is that the time is *greater* than x_i).

$$L(\boldsymbol{x},\theta) = \prod_{i=1}^{n}\left\{f(x_i \mid \theta)\mathbb{I}_{\{v_i=1\}} + F\left(x_i \mid \theta\right)\mathbb{I}_{\{v_i=0\}}\right\}$$

$$\ell(\boldsymbol{x},\theta) = \sum_{i=1}^{n}\left\{v_i \log f(x_i \mid \theta) + (1-v_i)\log F(x_i \mid \theta)\right\}$$

Using $f = hF$ and $F = \exp(-H)$, we obtain

$$\boxed{\ell(\boldsymbol{x},\theta) = \sum_{i=1}^{n}\left\{v_i \log h(x_i \mid \theta) - H(x_i \mid \theta)\right\}}$$

The MLE which maximises this is denoted $\hat{\theta}$.

- Let
  - Θ be the p-dimensional space in which the MLE $\hat{\theta}$ lives
  - Let $\tilde{\theta}$ be the MLE if we constrain θ to $\Theta_0 \subseteq \Theta$, a q-dimensional subspace of Θ.

Wilks' Lemma then tells us that

$$\text{If } \theta_{\text{real}} \in \Theta_0 \text{ then } \boxed{2\left[S(\hat{\theta}) - S(\tilde{\theta})\right] \sim \chi^2_{p-q}}$$

Thus

  - We accept the null hypothesis $\theta_{\text{real}} \in \Theta_0$ if

$$S(\hat{\theta}) - S(\tilde{\theta}) \le \frac{1}{2}C_{p-q,1-\alpha}$$

  - A confidence region for a given θ_0 (ie: if Θ_0 contains a single element) is given by

$$\left\{\theta_0 : S(\hat{\theta}) - S(\theta_0) \le \frac{1}{2}C_{p-q,1-\alpha}\right\}$$

- For example, for the exponential distribution $f(t \mid \theta) = \theta e^{-\theta t}$ and $F(t \mid \theta) = e^{-\theta t}$ so $h(t) = \theta$ and $H(t) = \theta t$. Thus

$$\ell(\boldsymbol{x},\theta) = \log\theta\sum_{i=1}^{n} v_i - \theta\sum_{i=1}^{n} x_i$$

And so

$$\hat{\theta} = \frac{\sum v_i}{\sum x_i} = \frac{\text{Number of observed events}}{\sum x_i}$$

Note also that $\ell'' = -\frac{1}{\theta^2}\sum_{i=1}^n v_i$. This is a measure of how much information is present in the sample; note that it's proportional to the number of death's we observe – *not* the number of individuals.

Non-Parametric inference

- Recall that a_j are the times at which failures actually occur.
- The **Kaplan-Meier estimate** of $F(t)$ is constructed by assuming that the number of members that survive in a time between a_{j-1} and a_j is binomial variable with probability of survival $1 - q_j$. It then estimates q_j as $\hat{q}_j = d_j / r_j$, where
  - d_j is the number of *observed* deaths at $t = X_j$ (not including censored observations)
  - r_j is the size of the "risk set" (ie: number of patients known to be still alive) just before time X_j

  The estimate of the probability of surviving longer than t (ie: not having died before t) is

$$\hat{F}(t) = \prod_{j:a_j \le t} \left(1 - \hat{q}_j\right) = \prod_{j:a_j \le t} \left(1 - \frac{d_j}{r_j}\right)$$

  (Computationally, we divide time into bands each of which contains a single observed time, and do the above).
- To estimate the error, we can use the rule for "propagation of errors" [we let $\hat{X} = \mathbb{E}(X)$]:

$$\begin{aligned}\mathbb{V}\text{ar}\left\{u(X)\right\} &\approx \mathbb{V}\text{ar}\left\{u(\hat{X}) + u'(\hat{X})(X - \hat{X})\right\} \\ &= \left[u'(\hat{X})\right]^2 \mathbb{V}\text{ar}\left(X\right)\end{aligned}$$

  Now, we perform the following steps
  - We begin by estimating the survival distribution as a sequence of binomials (ie: we ignore the r_j are random variables). We then have

$$\mathbb{V}\text{ar}(\hat{q}_j) = \mathbb{V}\text{ar}\left(\tfrac{d_j}{r_j}\right) = \frac{1}{r_j^2}\mathbb{V}\text{ar}(d_j) = \frac{1}{r_j^2} r_j q_j (1 - q_j) = \frac{q_j(1 - q_j)}{r_j}$$

  - We then use the formula for propagation of errors to get

$$\mathbb{V}\mathrm{ar}\left\{\log\left(1-\hat{q}_j\right)\right\} \approx \left(\frac{1}{\mathbb{E}(1-\hat{q}_j)}\right)^2 \mathbb{V}\mathrm{ar}\left(\hat{q}_j\right) = \frac{q_j}{r_j(1-q_j)}$$

- We then write

$$\log\left\{\hat{F}(t)\right\} = \sum_{i=1}^{n} \log(1-\hat{q}_j)$$

From which we immediately obtain

$$\mathbb{V}\mathrm{ar}\left(\log\left\{\hat{F}(t)\right\}\right) = \sum_{j:a_j \le t} \frac{q_j}{r_j(1-q_j)}$$

- Finally, we apply the rule of propagation of errors again, to find

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left\{\hat{F}(t)\right\} &= \mathbb{V}\mathrm{ar}\left\{e^{\log\left\{\hat{F}(t)\right\}}\right\} \\ &\approx \left\{F(t)\right\}^2 \sum_{j:a_j \le t} \frac{q_j}{r_j(1-q_j)}\end{aligned}$$

Greenwood's Formula for the variance is then given by

$$\begin{aligned}\hat{\mathbb{V}}\mathrm{ar}\left\{\hat{F}(t)\right\} &= \left\{\hat{F}(t)\right\}^2 \sum_{j:a_j \le t} \frac{\hat{q}_j}{r_j(1-\hat{q}_j)} \\ &= \left\{\hat{F}(t)\right\}^2 \sum_{j:a_j \le t} \frac{1}{r_j}\frac{d_j}{r_j - d_j} \\ &= s_0^2\end{aligned}$$

And a confidence interval for $F(t)$ is $\left[\hat{F}(t) - \Phi s_0, \hat{F}(t) + \Phi s_0\right]$

- Unfortunately, the confidence interval above can go beyond the interval [0,1]. Two solutions exist to this
  - **Use a transformation**. Two possibilities
    - **log transformation**

      We know from above that

$$\mathbb{V}\mathrm{ar}\left(\log\left\{\hat{F}(t)\right\}\right) = \sum_{j:a_j \le t} \frac{q_j}{r_j(1-q_j)} = s_1^2$$

      and so we obtain the following confidence interval for $F(t)$:

$$\left[\hat{F}(t)e^{-\Phi s_1}, \hat{F}(t)e^{\Phi s_1}\right]$$

      This works for $F(t)$ near 0, but may get into trouble for $F(t)$ near 1.
    - **log(–log) transformation**

      Using the propagation of variance formula, we get

$$\mathbb{V}\text{ar}\left(\log\left\{-\log\left[\hat{F}(t)\right]\right\}\right) = \frac{1}{\left[\log \hat{F}(t)\right]^2} \sum_{j:a_j \le t} \frac{1}{r_j} \frac{d_j}{r_j - d_j} = s_2^2$$

And we obtain the following confidence interval

$$\left[\hat{F}(t)^{\exp[\Phi s_2]}, \hat{F}(t)^{\exp[-\Phi s_2]}\right]$$

This is guaranteed to be between 0 and 1.

- **Use the likelihood directly**: another approach is to find the likelihood L which corresponds to an arbitrary value z of $\hat{F}(t)$.

  To do this, first recall that we construct the KM estimate by assuming that our survivor function takes the form

  $$F(t) = \prod_{j:a_j \le t}(1 - q_j)$$

  Where the q_j have to be estimated. For a given set of q_j, the corresponding likelihood is

  $$L(\boldsymbol{q} \mid \boldsymbol{d}, \boldsymbol{r}) \propto \prod_{\text{all events}} q_j^{d_j}(1 - q_j)^{r_j - d_j}$$

  $$\ell(\boldsymbol{q} \mid \boldsymbol{d}, \boldsymbol{r}) = \sum_{\text{all events}} d_j \log q_j + (r_j - d_j)\log(1 - q_j)$$

  If the q are unconstrained, this likelihood is maximised by setting $\hat{q}_j = d_j / r_j$, as we do in the KM estimate.

  However, if we insist on constraining $F(t) = z$, then we need to use Lagrange multipliers to maximise the likelihood. This involves finding a λ such that

  $$\sum_{\text{all events}} d_j \log q_j + (r_j - d_j)\log(1 - q_j) - \lambda\left\{\log z - \sum_{j:a_j \le t} \log(1 - q_j)\right\}$$

  is maximised when the constraint is satisfied. This gives

  $$q_j = \begin{cases} \frac{d_j}{\lambda + r_j} & j : a_j \le t \\ d_j / r_j & j : a_j > t \end{cases}$$

  (Note that only values for $X_j \le t$ are affected, since $F(t)$ – which we are constraining – only involves these values).

  Our strategy for interval estimation is then as follows

  - Choose a λ
  - Work out the q_j
  - Using those, we can work out $F(t)$ and L

- Repeat for various values of λ, and use the values obtained to construct a likelihood graph of L against $F(t)$

Since there is effectively only one parameter here (λ), then

$$2\left[\ell\left(\hat{F}(t)\right) - \ell\left(z\right)\right] \sim \chi_1^2$$

where $\hat{F}(t)$ is the KM estimate of $F(t)$ and z as another value. Thus, to find a confidence interval for $\hat{F}(t)$, we use the graph to find the values z whose likelihoods are at most $\frac{1}{2}C_{1,1-\alpha}$ away from the maximum likelihood.

Empirical Inference

We now consider a different kind of non-parametric way to estimate the survivor function. We do this by constructing an *empirical likelihood*, and finding the F that maximise it subject to

1. $F(u) \geq F(v)$ if $u < v$
2. $1 \geq F(t) \geq 0$
3. $F(0) = 1$ (not essential, but aids exposition; implies no events at $t = 0$).

Four common kinds of contributions to the likelihood are

- i^{th} individual *right* censored at x_i adds $\boxed{F(x_i)}$
- i^{th} individual with event at x_i contributes $\boxed{F(x_i-) - F(x_i)}$, where $F(x_i-) = \lim_{\Delta \to 0} F(x_i - \Delta)$
- i^{th} individual *left* censored at x_i (ie: $T \leq x_i$) adds $\boxed{1 - F(x_i)}$
- i^{th} individual censored in the *interval* $[x_i^L, x_i^U)$ (ie: $x_i^L < T \leq x_i^U$) adds $\boxed{F(x_i^U) - F(x_i^L)}$

To find the likelihood, we multiply all the contributions together and maximise. Note that the generic term is in the form $\left[F(b) - F(a)\right]$. If there is no term involving a $-F(b)$, this means that we should increase $F(b)$ indefinitely, subject to condition 1 above. An immediate consequence, together with condition 3 above, is that $F(0) = 1$.

We can use this methodology to re-derive Kaplan Meier:

- All terms will be of the form $F(x_i-) - F(x_i)$ or $F(x_i)$.

- Now, note that
  - There will never be a – sign in front of $F(x_i-)$, and so
$$\hat{F}(x_i-) = \hat{F}(\text{latest event})$$
$$\Rightarrow \hat{F}(a_j-) = \hat{F}(a_{j-1})$$
  - If, at a time x_i, there is no event but a censored observation, then there will never be a sign in front of $F(x_i)$, and so
$$\hat{F}(x_i) = \hat{F}\left(\text{latest event}\right)$$
$$\Rightarrow \hat{F}(x_i) = \hat{F}(a_{j:a_j \max \leq x_i})$$

  This implies that we only need to consider event times, and that the function is *constant* except at event times.
- We write $c_j = r_j - d_j - r_{j+1}$ for the number of censored events in the interval $[a_j, a_{j+1})$.
- The likelihood can then be written as
$$L = \prod_{\text{all events}} \left[F(a_{j-1}) - F(a_j)\right]^{d_j} \left[F(a_j)\right]^{c_j}$$
  Writing $F_j = F(a_j)$, we can write this as
$$L = \prod_{\text{all events}} \left[F_{j-1} - F_j\right]^{d_j} \left[F_j\right]^{c_j}$$
  The exponent on the second term is effectively the number of censored events between a_j and a_{j+1}.
- Taking logs, and then differentiating
$$\ell = \sum_{\text{all events}} d_j \log\left(F_{j-1} - F_j\right) + \sum_{\text{all events}} c_j \log F_j$$
$$\frac{\partial \ell}{\partial F_j} = -\frac{d_j}{\hat{F}_{j-1} - \hat{F}_j} + \frac{d_{j+1}}{\hat{F}_j - \hat{F}_{j+1}} + \frac{c_j}{\hat{F}_j} = 0$$
- This is a third-order recurrence relationship. It simplifies greatly if we start with $\hat{F}_g$. Consider that $d_{g+1} = 0$. The recursion relation for $\hat{F}_g$ is then
$$-\frac{d_j}{F_{g-1} - F_g} + \frac{c_j}{F_g} = 0 \Rightarrow \hat{F}_g = \frac{c_g}{c_g + d_g}\hat{F}_{g-1}$$
  (If $c_g = 0$, $\hat{F}_g = 0$. This makes sense; since the last term in the likelihood is $\left[F_{g-1} - F_g\right]^{d_g}$, we want to make $\hat{F}_g$ as small as possible).
- Now, consider that we can re-write our expression for $\hat{F}_g$ as
$$\hat{F}_g = \frac{r_g - d_j}{r_g}\hat{F}_{g-1} = \left(1 - \frac{d_j}{r_g}\right)\hat{F}_{g-1}$$

Note also that if $\hat{F}_{j+1} = \left(1 - \frac{d_{j+1}}{r_{j+1}}\right)\hat{F}_j$, then

$$-\frac{d_j}{\hat{F}_{j-1} - \hat{F}_j} + \frac{d_{j+1}}{\hat{F}_j - \hat{F}_{j+1}} + \frac{c_j}{\hat{F}_j} = 0$$

$$-\frac{d_j}{\hat{F}_{j-1} - \hat{F}_j} + \frac{d_{j+1}}{\hat{F}_j - \left(1 - \frac{d_{j+1}}{r_{j+1}}\right)\hat{F}_j} + \frac{c_j}{\hat{F}_j} = 0$$

$$-\frac{d_j}{\hat{F}_{j-1} - \hat{F}_j} + \frac{r_{j+1}}{\hat{F}_j} + \frac{c_j}{\hat{F}_j} = 0$$

$$-d_j\hat{F}_j + r_{j+1}\hat{F}_{j-1} - r_{j+1}\hat{F}_j + c_j\hat{F}_{j-1} - c_j\hat{F}_j = 0$$

$$\hat{F}_j = \frac{r_{j+1} + c_j}{d_j + r_{j+1} + c_j}\hat{F}_{j-1}$$

$$\hat{F}_j = \frac{r_j - d_j}{r_j}\hat{F}_{j-1}$$

$$\hat{F}_j = \left(1 - \frac{d_j}{r_j}\right)\hat{F}_{j-1}$$

Thus, by induction, this is true for all j. Since we have assume that $\hat{F}_0 = 1$, this is our familiar Kaplan-Meier estimate.

Note that this all ties in to our earlier discussion of deriving confidence intervals for the KM estimator. Consider that

- Constraining $F(t) = z$ is equivalent to constraining $\log F_k = \log z$, where k is chosen to be the event *just* before t.
- The quantity to maximise is then

  $$\ell = \sum_{\text{all events}} d_j \log\left(F_{j-1} - F_j\right) + \sum_{\text{all events}} c_j \log F_j + \lambda\left(\log F_k - \log z\right)$$

  Interestingly, this just looks like we've added an extra λ censored individuals at time a_k. This makes the recurrence relation easy to intuitively adapt

  $$\begin{array}{ll} j > k & \hat{F}_j = \left(1 - \frac{d_j}{r_j}\right)\hat{F}_{j-1} \\ j \le k & \hat{F}_j = \left(1 - \frac{d_j}{r_j + \lambda}\right)\hat{F}_{j-1} \end{array}$$

  We start at $\hat{F}_k = z$ to obtain those terms with $j > k$, and we start with $\hat{F}_0 = 1$ for those $j \le k$, choosing λ such that $\hat{F}_k = z$.
- As ever, the confidence intervals are then found using

  $$\left\{z : 2\left(\hat{\ell} - \tilde{\ell}(z)\right) \le C_{1,1-\alpha}\right\}$$

Where $\hat{\ell}$ is the maximum log-likelihood and $\tilde{\ell}(z)$ is the constrained likelihood.

The Log-Rank Test

We now consider a situation in which we want to compare the survival performance of two groups. It is *not* a good idea to compare the groups at a particular time t because

- We are usually interested in the complete time spectrum rather than in individual points
- Comparing specific time points could lead to multiple testing problems
- We might be tempted to choose specific time points a posteriori to suit our hypothesis

The **log-rank test** is a nonparametric test that takes all observations into account. It is most powerful when used on non-overlapping survival curves (ie: where the curves belong to the same proportional hazards family, $\left(F_j^{(0)}\right)^k = F_j^{(1)}$). In fact, this test can be shown to be the score test for proportional hazards.

Consider two groups $i \in \{0,1\}$, with observed/censored times $X_j^{(i)}$. At time X_j, there are $r_j^{(i)}$ individuals at risk in group i, of which $d_j^{(i)}$ are observed to fail. The survivor function at X_j is $F_j^{(i)}$ in group i, and our null hypothesis is

$$H_0 : F_j^{(0)} = F_j^{(1)} \qquad \forall j$$

Our strategy in the log-rank test is to construct a contingency table for every time a_j at which a failure is observed, which looks like this

Time a_j	**Group 0**	**Group 1**	**Total**
Fails	$d_j^{(0)}$	$d_j^{(1)}$	d_j
Not-fails	$r_j^{(0)} - d_j^{(0)}$	$r_j^{(1)} - d_j^{(1)}$	$r_j - d_j$
# risk set	$r_j^{(0)}$	$r_j^{(1)}$	r_j

Now consider – under the null hypothesis, the probability of failing is the same for both groups. Using the hypergeometric distribution, the expectation and variance of the upper-left-hand cell should then be

$$\mathbb{E} = r_j^{(0)} \frac{d_j}{r_j}$$

$$\mathbb{V}\mathrm{ar} = \frac{d_j(r_j - d_j) r_j^{(0)} r_j^{(1)}}{r_j^2 \,/\, (r_j - 1)} = s^2$$

The *deviation from what we expect* is therefore given by

$$z_j = d_j^{(0)} - \frac{d_j}{r_j} r_j^{(0)}$$

And the log-rank statistic is given by

$$\frac{1}{s} \sum_{j=1}^{N} z_j$$

should be compared with the standard normal distribution.

Other versions weigh the z_j by r_j.

Modelling

We may be interested in modelling the effect of *explanatory variables* on the survival probabilities. The setup is as follows

- Individual i has explanatory variables $\boldsymbol{z}^{(i)} = \left(z_1^{(i)}, \cdots, z_p^{(i)}\right)$
- Our model has parameter set $\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\beta} & \boldsymbol{\psi}\end{pmatrix}^T$, where $\boldsymbol{\beta}$ are "interesting" parameters relating to the $\boldsymbol{z}$, and $\boldsymbol{\psi}$ are the nuisance parameters.

Accelerated life modelling

Here, we start with a *baseline survivor* $F_0(t, \boldsymbol{\psi})$, and we model the survivor of the i^{th} individual by

$$F\left(t, \boldsymbol{z}^{(i)}, \boldsymbol{\beta}, \boldsymbol{\psi}\right) = F_0\left(\phi(\boldsymbol{z}^{(i)}, \boldsymbol{\beta})t, \boldsymbol{\psi}\right)$$

This, however, is very rarely used.

Proportional Hazards Modelling

Here, we start with a *baseline hazard* $h_0(t, \boldsymbol{\psi})$, and we model the hazard of the i^{th} individual by

$$h(t, \boldsymbol{z}^{(i)}, \boldsymbol{\beta}, \boldsymbol{\psi}) = \phi\left(\boldsymbol{z}^{(i)}, \boldsymbol{\beta}\right) h_0(t, \boldsymbol{\psi})$$

Possible forms of the function ϕ are as follows (note that it *must* be positive)

$$\phi(\boldsymbol{z}, \boldsymbol{\beta}) = \begin{cases} e^{\boldsymbol{\beta}^T \boldsymbol{z}} & \leftarrow \text{Cox regression} \\ 1 + e^{\boldsymbol{\beta}^T \boldsymbol{z}} & \\ \log\left(1 + e^{\boldsymbol{\beta}^T \boldsymbol{z}}\right) & \end{cases}$$

We now consider the likelihood inference for $\boldsymbol{\beta}$.

- We first use an invariance argument to show that the *order* in which the events happen is sufficient for $\boldsymbol{\beta}$
  - If we transform the timescale from $t \to u$ with $t = g(u)$ and g monotonically increasing and differentiable), then
  $$h(t, \boldsymbol{z}^{(i)}, \boldsymbol{\beta}, \boldsymbol{\psi}) = \phi\left(\boldsymbol{z}^{(i)}, \boldsymbol{\beta}\right) h_0(g(u), \boldsymbol{\psi}) g'(u)$$
  - Clearly, only the baseline hazard changes. Thus, the timescale is irrelevant to the proportionality factor.

- Now, consider a situation in which we have no censoring and no ties.
  - Label the individuals $1,\cdots,n$, and let π_j be the j^{th} individual to fail, at time a_j ($\boldsymbol{\pi}$ is a permutation of the n individuals).
  - The risk set at that time is $R_j = \left\{\pi_{j'} : j' \geq j\right\}$
  - The probability of individual i failing at t is proportional to that individual's hazard at t. Now, we know that individual π_j is the *only* one from the risk set R_j to have failed at a_j. The probability of that happening was therefore

$$\frac{\phi\left(\boldsymbol{z}^{(\pi_j)},\boldsymbol{\beta}\right)\cancel{h_0(\boldsymbol{a}_j,\boldsymbol{\psi})}}{\sum_{i\in R_j}\phi(\boldsymbol{z}^{(i)},\boldsymbol{\beta})\cancel{h_0(\boldsymbol{a}_j,\boldsymbol{\psi})}} = \frac{\phi\left(\boldsymbol{z}^{(\pi_j)},\boldsymbol{\beta}\right)}{\sum_{i\in R_j}\phi(\boldsymbol{z}^{(i)},\boldsymbol{\beta})}$$

  - Thus, the probability that we observe the sequence that we did indeed observe is

$$\prod_{j=1}^{n}\frac{\phi\left(\boldsymbol{z}^{(\pi_j)},\boldsymbol{\beta}\right)}{\sum_{i\in R_j}\phi(\boldsymbol{z}^{(i)},\boldsymbol{\beta})}$$

    This *partial likelihood* for $\boldsymbol{\beta}$ can be maximised. Some software exists to do this efficiently, especially for Cox regression. (The likelihood is "partial" because it does not use all the data available – but we showed, from our invariance argument, that it is nevertheless sufficient).
- Dealing with censoring – we use exactly the same expression as above, but we only include the non-censored observations in the product. For example, if the individuals are censored in the order 3 (4) 1 2, we would obtain the following likelihood

$$\frac{\phi_1}{\phi_1+\phi_2+\phi_3+\phi_4}\frac{\phi_1}{\phi_1+\phi_2}\frac{\phi_2}{\phi_2}$$

  (Note that this is the sum of the likelihoods for 3 4 1 2, 3 1 4 2 and 3 1 2 4).
- Dealing with ties – consider the example 3, 4 = 2, 1 (ie: 4 and 2 fail at the same time). Several options:
  - Assume there is a *real* order, and sum the likelihoods.

$$\frac{\phi_3}{\phi_1+\phi_2+\phi_3+\phi_4}\left(\frac{\phi_4}{\phi_1+\phi_2+\phi_4}\frac{\phi_1}{\phi_1+\phi_2}+\frac{\phi_1}{\phi_1+\phi_2+\phi_4}\frac{\phi_4}{\phi_2+\phi_4}\right)\frac{\phi_2}{\phi_2}$$

This is called the *exact* partial likelihood, but it can get computationally very expensive.

- We consider the tie as genuine, and consider the probability of choosing this group of 2 events out of all the possible groups of two events we could have chosen

$$\frac{\phi_3}{\phi_1+\phi_2+\phi_3+\phi_4}\frac{\phi_4\phi_1}{\phi_4\phi_1+\phi_1\phi_2+\phi_2\phi_4}\frac{\phi_2}{\phi_2}$$

This should only be used if the data is truly discrete.

- We consider a mixture of individuals

$$\frac{\phi_3}{\phi_1+\phi_2+\phi_3+\phi_4}\frac{\phi_4\phi_1}{\left(\phi_1+\phi_2+\phi_4\right)\left(\phi_2+\frac{\phi_1}{2}+\frac{\phi_4}{2}\right)}\frac{\phi_2}{\phi_2}$$

This is called the *Efron approximation.*

Residuals

We define:

- The **Cox-Snell residual** is defined as follows

$$y_i = \hat{H}_i(x_i) = \phi(\boldsymbol{z}^{(i)},\boldsymbol{\beta})H_0(x_i,\boldsymbol{\psi}) = -\log \hat{S}_i(x_i)$$

where x_i is the *failure time* of individual *i*. Note that for uncensored data, $y_i \sim \text{Exp}(1)$. We can prove this as follows:

Note that if $U = H_T(T)$, and F_U and F_T are the survival functions of U and T, we have:

$$F_U(u) = \mathbb{P}\left(U > u\right) = \mathbb{P}\left(H_T(T) > u\right)$$

H_T is increasing and has an inverse, and so

$$\begin{aligned} F_U(u) &= \mathbb{P}\left(T > H_T^{-1}(u)\right) \\ &= F_T\left(H_T^{-1}(u)\right) \\ &= \exp\left(-H_T(H_T^{-1}(u))\right) \\ &= \exp\left(-u\right) \end{aligned}$$

This is the survivor function for an exponential distribution.

- If the individual is right-censored with recorded time x_i and *real* time $t_i^* > x_i$, then we obviously have that $\hat{H}_i(x_i) < \hat{H}_i(t_i^*)$. Our strategy will therefore be to add something to the Cox-Snell residuals for censored values to correct for this discrepancy.

Remember, further, that the exponential distribution is memory-less. Thus, $\hat{H}_i(x_i) \sim \text{Exp}(1) \Rightarrow \hat{H}_i(t_i^*) - \hat{H}_i(x_i) \sim \text{Exp}(1)$, and therefore $\mathbb{E}\left\{\hat{H}_i(t_i^*) - \hat{H}_i(x_i)\right\} = 1$. Thus, it seems like a sensible amount to add is 1. We define the *modified Cox-Snell Residuals*:

$$y_i' = \left(1 - v_i\right) + \hat{H}(x_i)$$

These simply add 1 to censored observations. These are therefore genuinely Exp(1) distributions, regardless of censoring and $\mathbb{E}(y_i') = 1$.

- The *Martingale residual* is defined as

$$y_i'' = 1 - y_i' = v_i - \hat{H}(x_i)$$

This has expectation 0. It can be thought of as the number of "observed" events at x_i (1 or 0) minus the number of "expected" events.

Counting Processes

$N(t)$ is a *counting process* if

- $N(t)$ is a non-negative integer
- $N(s) < N(t)$ if $s < t$
- $\mathrm{d}N(t) = N(t) - N(t-) \in \{0,1\}$, where $N(t-) = \lim_{\delta \to 0} N(t-\delta)$
- $\mathbb{E}\left(N(t)\right) < \infty$

We write $\mathcal{H}_t$ for the *filtration* of a counting process – all that is known at time t (in particular, this includes the values of random variables known up to and including time t).

We define an *intensity* $\lambda(t)$ and *integrated intensity* $\Lambda(t)$ as follows

$$\mathbb{P}\left(N(t+\delta) - N(t-) \mid \mathcal{H}_{t-}\right) \approx \lambda(t)\delta$$

$$\mathbb{P}\left(\mathrm{d}N(t) \mid \mathcal{H}_{t-}\right) = \mathrm{d}\Lambda(t) \qquad \Lambda(t) = \int_0^t \lambda(t)\ \mathrm{d}t$$

Since $\mathrm{d}N(t)$ can only take values in $\{0,1\}$, this is equivalent to

$$\mathbb{E}\left(\mathrm{d}N(t) \mid \mathcal{H}_{t-}\right) = \mathrm{d}\Lambda(t)$$

Now, we require $\Lambda(t)$ to be *predictable* with respect to $\mathcal{H}_t$ (ie: we require it to be known given $\mathcal{H}_{t-1}$) – effectively, this means it must be continuous. That said, we can write

$$\mathbb{E}\left(\mathrm{d}N(t) - \mathrm{d}\Lambda(t) \mid \mathcal{H}_{t-}\right) = 0$$

Defining $M(t) = N(t) - \Lambda(t)$, the above clearly show that $\mathbb{E}\left(\mathrm{d}M(t) \mid \mathcal{H}_{t-}\right) = 0$. Thus, M is a *martingale*, and we can write the *Doob-Meyer decomposition*:

$$\boxed{N(t) = \Lambda(t) + M(t)}$$

In other words, the counting process can be decomposed into a martingale and an increasing compensator function.

Relation to Survival Analysis

Survival analysis can be seen as a counting process. The counting variable for individual i whose time-to-event is the random variable T_i is

$$N_i(t) = \mathbb{I}_{\{t \geq T_i\}}$$

Now

$$\mathrm{d}\Lambda_i(t) = \mathbb{E}\left(\mathrm{d}N_i(t) \mid \mathcal{H}_{t-}\right) = \text{In risk set} \times \text{Hazard} = Y_i(t)h_i(t)$$

Where $Y_i(t) = \mathbb{I}_{\{T \geq t\}}$.

When we consider all n individuals, we can construct a new counting process consisting of the sum of each individual counting process

$$N_+(t) = \Lambda_+(t) + M_+(t)$$

The summed compensator can be written as

$$\Lambda_+(t) = \int_0^t \sum_{i=1}^n Y_i(u) h_i(u) \, \mathrm{d}u$$

If all the individuals are exposed to the same hazard, this becomes

$$\Lambda_+(t) = \int_0^t Y_+(u) h(u) \, \mathrm{d}u = \int_0^t Y_+(u) \, \mathrm{d}H(u)$$

Where $H(u)$ is the integrated hazard.

The Nelson-Aalen Estimator of $H(t)$

The *Doob-Meyer decomposition* of the counting process can be written in differential form as

$$\mathrm{d}N_+(t) = \mathrm{d}\Lambda_+(t) + \mathrm{d}M_+(t)$$

We saw, however, that conditional on past history, the martingale has expectation 0. So an *estimate* of H can be obtained using

$$\mathrm{d}N_+(t) = \mathrm{d}\Lambda_+(t)$$

When the hazard is the same of every individual, this becomes

$$\mathrm{d}N_+(t) = Y_+(t) \, \mathrm{d}H(t)$$

$$\mathrm{d}\hat{H}(t) = \frac{\mathrm{d}N_+(t)}{Y_+(t)}$$

$$\hat{H}(t) = \int_0^t \frac{\mathrm{d}N_+(t)}{Y_+(t)}$$

Now, let's consider each part of this estimator

- $\mathrm{d}N_+(t)$ is 1 at any time at which an event happens, but 0 otherwise.
- $Y_+(t)$ is simply the size of the risk set at t, r_t.

Thus, the Nelson-Aalen estimator is

$$\boxed{\hat{H}(t) = \sum_{j:a_j \leq t} 1 / r_{a_j}}$$

Censored data poses no problem – if an individual fails in between times a_j and a_{j+1}, it is included in all risk sets up to a_j but none thereafter. Similarly, if failure occurs *at* a time a_j, it is included in that risk set but none thereafter.

We find the variance of the estimator as follows

- $\mathrm{d}N_+(t)$ is (locally) a Poisson variable, with mean and variance $\mathrm{d}\Lambda_+(t)$. Thus, $\hat{\mathbb{V}}\mathrm{ar}\left(\mathrm{d}N_+(t)\right) = \mathrm{d}\hat{\Lambda}_+(t) = \mathrm{d}N_+(t)$. As such
$$\hat{\mathbb{V}}\mathrm{ar}\left(\mathrm{d}\hat{H}(t)\right) = \hat{\mathbb{V}}\mathrm{ar}\left(\frac{\mathrm{d}N_+(t)}{Y_+(t)}\right) = \frac{\mathrm{d}N_+(t)}{\left[Y_+(t)\right]^2}$$
- Integrating in this case is paramount to adding lots of independent bits, so
$$\hat{\mathbb{V}}\mathrm{ar}\left(\hat{H}(t)\right) = \int_0^t \frac{\mathrm{d}N(t)}{\left[Y_+(t)\right]^2} = \sum_{j:a_j \le t} \frac{1}{r_j^2} = \left[s(t)\right]^2$$
- Confidence intervals can be worked out using the normal distribution
$$\left[\hat{H}(t) - \Phi^{-1}s(t), \hat{H}(t) + \Phi^{-1}s(t)\right]$$
But better results can be obtained by first taking log transforms, and nothing that by the propagation of variance formula,
$$\hat{\mathbb{V}}\mathrm{ar}\left\{\log \hat{H}(t)\right\} \approx \left(\frac{s(t)}{\hat{H}(t)}\right)^2$$
Thus, a confidence integral for $\log \hat{H}(t)$ is
$$\left[\log \hat{H}(t) - \Phi^{-1}\frac{s(t)}{\hat{H}(t)}, \log \hat{H}(t) + \Phi^{-1}\frac{s(t)}{\hat{H}(t)}\right]$$
Which gives
$$\left[\hat{H}(t)\exp\left\{-\Phi^{-1}\frac{s(t)}{\hat{H}(t)}\right\}, \hat{H}(t)\exp\left\{\Phi^{-1}\frac{s(t)}{\hat{H}(t)}\right\}\right]$$

There are a number of ways to handle ties at a_j

- A natural way is to simply assume $\mathrm{d}N(a_j) = 2$ at that point. The estimate is then
$$\cdots + \frac{1}{r_{j-1}} + \frac{2}{r_j} + \frac{1}{r_{j+1}} + \cdots$$
Unfortunately, the resulting estimate for $\hat{H}(t)$ for $t > a_j$ is not the same as would be obtained by substituting two distinct event times $a_j \pm \Delta$ and letting $\Delta \to 0$.
- In the second method, we actually assume that one event happens before the other. The estimate is then
$$\cdots + \frac{1}{r_{j-1}} + \frac{1}{r_j} + \frac{1}{r_j - 1} + \frac{1}{r_{j+1}} + \cdots$$

Nelson-Aalen and Kaplan-Meier

If we let $\hat{H}_{\text{NA}}(t)$ be the Nelson-Aalen estimator of $H(t)$ and $\hat{S}_{\text{KM}}(t)$, be the Kaplan-Meier estimate of $S(t)$, we can come up with a "Kaplan-Meier estimator of integrated hazard":

$$\hat{H}_{\text{KM}}(t) = -\log\left\{\hat{S}_{\text{KM}}(t)\right\}$$

and a "Nelson-Aalen estimator of the survivor function"

$$\hat{S}_{\text{NA}}(t) = \exp\left\{-\hat{H}_{\text{NA}}(t)\right\}$$

For reasonably-sized risk sets, these are close to each other.

Nelson-Aalen and Proportional Hazards

In proportional hazards modelling, we assume that

$$h_i(t) = \phi(z^{(i)}, \beta)h_0(t) \Rightarrow H_i(t) = \phi(z^{(i)}, \beta)H_0(t)$$

It sometimes helps to have an estimate of H_0. Let's use the equation for the compensator again:

$$\Lambda_+(t) = \int_0^t \sum_{i=1}^n Y_i(u)\phi(z^{(i)}, \beta)h_0(u)\ \mathrm{d}u$$

$$\Lambda_+(t) = \int_0^t \sum_{i=1}^n Y_i(u)\phi(z^{(i)}, \beta)\ \mathrm{d}H_0(u)$$

$$\mathrm{d}\Lambda_+(t) = \sum_{i=1}^n Y_i(t)\phi(z^{(i)}, \beta)\ \mathrm{d}H_0(t)$$

Once again, we assume $\mathrm{d}\Lambda_+(t) = \mathrm{d}N_+(t)$, and this gives

$$\sum_{i=1}^n Y_i(t)\phi(z^{(i)}, \beta)\ \mathrm{d}\hat{H}_0(t) = \mathrm{d}N_+(t)$$

$$\mathrm{d}\hat{H}_0(t) = \frac{\mathrm{d}N_+(t)}{\sum_{i=1}^n Y_i(t)\phi(z^{(i)}, \beta)}$$

$$\boxed{\hat{H}_0(t) = \int_0^t \frac{\mathrm{d}N_+(t)}{\sum_{i=1}^n Y_i(t)\phi(z^{(i)}, \beta)}}$$